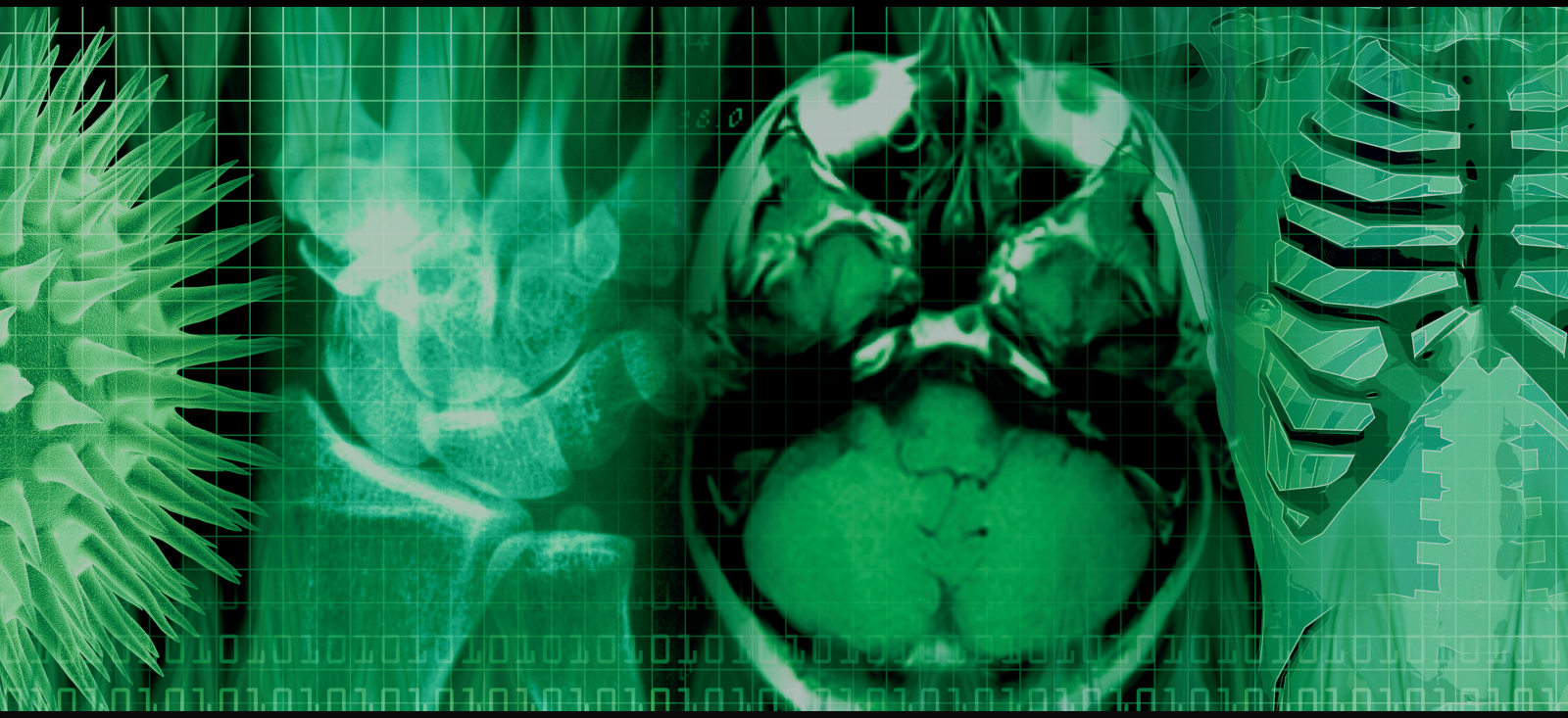
International Journal of Biomedical Imaging

Mathematical Methods for Images and Surfaces 2011

Guest Editors: Weihong Guo, Lalita Udpa, Yang Wang, Guowei Wei, and Shan Zhao



Mathematical Methods for Images and Surfaces 2011

International Journal of Biomedical Imaging

Mathematical Methods for Images and Surfaces 2011

Guest Editors: Weihong Guo, Lalita Udpa, Yang Wang, Guowei Wei, and Shan Zhao

Editorial Board

Contents

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2012, Article ID 419647, 2 pages
doi:10.1155/2012/419647

Editorial

Mathematical Methods for Images and Surfaces 2011

Weihong Guo,[1] Lalita Udpa,[2] Yang Wang,[3] Guowei Wei,[2,3] and Shan Zhao[4]

[1] *Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106, USA*
[2] *Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, USA*
[3] *Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA*
[4] *Department of Mathematics, University of Alabama, Tuscaloosa, AL 35487, USA*

Correspondence should be addressed to Guowei Wei, wei@math.msu.edu

Received 11 December 2011; Accepted 13 December 2011

The study of biomedical imaging and biological surfaces is a rapidly growing interdisciplinary field that has attracted considerable interest from mathematical, engineering, and medicine communities. Many research problems in this field are application oriented and thus the results have practical values but are challenging due to physical and biological constraints and the large scale nature of massive biomedical and biomolecular data. To efficiently solve these problems, advanced mathematical models and fast and efficient computational algorithms are indispensable tools. This special issue was called to address mathematical difficulties and challenges in image and surface analysis. We would like to share with the readers the recent advances in topics such as topological-gradients-based edge/contour detection, partial differential-equation-transform-based feature separation, blind multiple-source reconstruction in bioluminescence tomography, partial-differential-equation-based cerebral cortex reconstruction, nonlinear elasto-mammography for characterization of breast tissue properties, and protein surface characterization.

We would like to thank the authors for their excellent contributions and patience that make this special issue possible. The time, effort, and valuable work of all anonymous reviewers on these papers are also very greatly acknowledged. This special issue constitutes ten papers.

The paper entitled "*Contour detection and completion for inpainting and segmentation based on topological gradient and fast marching algorithms*" by D. Auroux et al. introduces a contour functional that is based on topological gradient and couples it with a fast marching algorithm to determine the minimal path for the purpose of generating connected contours. This offers a hybrid scheme for edge detection and contour completion. Two specific applications are considered for image processing. For image segmentation, the topological gradient is shown to be more efficient than the standard gradient approaches. For image inpainting, the hybrid scheme particularly improves the quality of the inpainted images.

The paper entitled "*Protein surface characterization using an invariant descriptor*" by Z. Abu Deeb et al. develops a new invariant descriptor for the characterization of protein surfaces. It is suitable for various analysis tasks, such as protein functional classification and search and retrieval of protein surfaces over a large database. Its novelty is the combination of the power of residue-distance cooccurrence-based local and global surface descriptors. The proposed method not only reduces the computational complexity of matching 3D structures, but also facilitates direct comparison between protein structures of different sizes. The comparison with other methods on three protein families indicates that this method is effective.

The paper entitled "*Extending local canonical correlation analysis to handle general linear contrasts for fMRI data*" by M. Jin et al. designs a novel test statistics to enable canonical correlation analysis (CCA) and to handle general linear contrasts in more complicated fMRI paradigms. This approach avoids the reparameterization of the design matrix and the reestimation of the CCA solutions for each particular contrast of interest. This test statistics is more powerful than the traditional t-test in general linear models on the inference of evoked brain regional activations from noisy fMRI data, especially for weakly evoked and localized brain activations. The method improves detection power with acceptable computation time and has potential to meet the

needs in recent fMRI where data is enormous, signal is weak, and the spatial correlation is strong.

The paper entitled "*A novel FEM-based numerical solver for interactive catheter simulation in virtual catheterization*" by S. Li et al. concerns with interactive simulation of the deformable catheters and guide-wires in virtual vascular interventional surgeries. The motion of catheters or guide-wires and their interactions with patients' vascular system are mathematically formulated in terms of a total potential energy, consisting of bending elastic energy, vessel wall deformation energy, and the work by the external forces. The minimization of the potential energy functional is numerically realized via a finite element simulation. Experimental studies indicate that the proposed method can realistically model and simulate deformable catheters and guide-wires in an interactive manner.

The paper entitled "*Cortical surface reconstruction from high-resolution MR brain images*" by S. Osechinskiy et al. presents a new PDE-based approach that readily scales with imaging resolution for reconstructing the cerebral cortex from MR images. The scalability virtue of the approach makes it promising in brain imaging research where high-resolution MRI becomes more popular. This scalability is achieved by using an implicit deformable surface model in a fast marching framework guided by a novel, computationally efficient model using potential field mapping. The method requires much lower computational resources and allows much faster computations than conventional methods.

The paper entitled "*Serial FEM/XFEM-based update of preoperative brain images using intraoperative MRI*" by L. Vigneron et al. aims to overcome the limitation of current neuronavigation systems that cannot adapt to changing intraoperative conditions over time. The authors develop a complete 3D framework for serial preoperative images updated in the presence of brain shift followed by successive resections. The key ingredient of the system is a nonrigid registration technique using a biomechanical model driven by the deformations of key surfaces tracked in successive intraoperative images. Numerical results demonstrate that the present approach significantly improves the alignment of nonrigidly registered images.

The paper entitled "*Selective extraction of entangled textures via adaptive PDE transform*" by Y. Wang et al. presents a new adaptive algorithm for selective extraction of entangled textures. Texture characterization and analysis are complicated for images with spatial entanglement, orientation mixing, and high-frequency overlapping. Based on a recently developed PDE transform method for functional mode decomposition, the statistical variance of the local variation is adaptively incorporated in the PDE transform framework for separating textures of very similar features. Successful texture separation is attained for several benchmark images.

The paper entitled "*Nonlinear elasto-mammography for characterization of breast tissue properties*" by Z. G. Wang et al. extends their previous studies by incorporating the projection of displace information obtained from the conventional X-ray mammography into a nonlinear elastography framework. In particular, projection-type displacement measurements are considered before and after breast compression, and a revised adjoint gradient method is derived for calculating the gradient of the objective function in the nonlinear elasto-mammography framework. Simulations based on a three-dimensional breast phantom involving normal and cancerous tissues are conducted to validate the feasibility and robustness of the proposed approach.

The paper entitled "*Fracture detection in traumatic pelvic CT images*" by J. Wu et al. presents an automated hierarchical algorithm for bone fracture detection in pelvic CT scans. It uses adaptive windowing, boundary tracing, and wavelet transform, while incorporating anatomical information. Fracture detection is performed based on the results of prior pelvic bone segmentation via their registered active shape model (RASM). The results are promising and show that the method is capable of detecting fractures accurately. Once verified with more data, the proposed method has the potential to be an important component of a larger modular system to extract features from CT images for a computer-assisted decision making system.

The paper entitled "*A finite element mesh aggregating approach to multiple-source reconstruction in bioluminescence tomography*" by J. Yu et al. develops a finite element mesh aggregating algorithm for blind multiple-source reconstruction in bioluminescence tomography. Without knowing the number of the sources in advance, an iterative procedure is utilized to detect multiple sources by exploiting the spatial structure of the nodes in finite element meshes and the characteristics of the energy decay. The detecting algorithm is formulated in a flexible reconstruction framework, where a variety of regularizers and inversion algorithms can be chosen by the user. Simulations using a tissue-like phantom demonstrate an improved performance of the new algorithm in terms of the automatic estimation of both locations and densities of multiple sources that differ greatly in power.

Weihong Guo
Lalita Udpa
Yang Wang
Guowei Wei
Shan Zhao

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2012, Article ID 870196, 19 pages
doi:10.1155/2012/870196

Research Article

Cortical Surface Reconstruction from High-Resolution MR Brain Images

Sergey Osechinskiy and Frithjof Kruggel

Department of Biomedical Engineering, University of California, Irvine, CA 92697, USA

Correspondence should be addressed to Sergey Osechinskiy, sosechin@uci.edu

Received 8 June 2011; Revised 22 September 2011; Accepted 28 September 2011

Academic Editor: Weihong Guo

Reconstruction of the cerebral cortex from magnetic resonance (MR) images is an important step in quantitative analysis of the human brain structure, for example, in sulcal morphometry and in studies of cortical thickness. Existing cortical reconstruction approaches are typically optimized for standard resolution (∼1 mm) data and are not directly applicable to higher resolution images. A new PDE-based method is presented for the automated cortical reconstruction that is computationally efficient and scales well with grid resolution, and thus is particularly suitable for high-resolution MR images with submillimeter voxel size. The method uses a mathematical model of a field in an inhomogeneous dielectric. This field mapping, similarly to a Laplacian mapping, has nice laminar properties in the cortical layer, and helps to identify the unresolved boundaries between cortical banks in narrow sulci. The pial cortical surface is reconstructed by advection along the field gradient as a geometric deformable model constrained by topology-preserving level set approach. The method's performance is illustrated on exvivo images with 0.25–0.35 mm isotropic voxels. The method is further evaluated by cross-comparison with results of the FreeSurfer software on standard resolution data sets from the OASIS database featuring pairs of repeated scans for 20 healthy young subjects.

1. Introduction

Cortical reconstruction, the derivation of a computerized representation of the cerebral cortical layer based on three-dimensional (3D) magnetic resonance (MR) images of the brain, is an important step in quantitative analysis of the human brain structure, for example, in the analysis of cortical folding patterns, in brain morphometry, and in cortical thickness studies. Cortical surface models typically serve as a reference basis for all further analysis and therefore must be geometrically accurate and topologically correct in order to provide valid and accurate quantitative measures of brain structure [1].

The cerebral cortex, considered at the spatial scale of MR images, is a thin layer of neural tissue, called gray matter (GM), located on the outer side of the white matter (WM), and surrounded by the cerebrospinal fluid (CSF). The cortex has a complex geometry of a highly folded layer with spatially varying curvature and thickness (thickness range 1–5 mm, average ≈2.5 mm, see [1]). The cortical layer on a brain hemisphere can be represented as the inner space between two cortical surfaces (i.e., an inner surface at the WM/GM and an outer or pial surface at the GM/CSF interface, see Figure 1). It is a useful simplification to consider each surface as topologically equivalent to a 3D sphere. In practice, limited spatial resolution of MR images, noise, intensity inhomogeneities, and partial volume effects can all be the sources of geometrical inaccuracies and topological errors in the reconstructed cortical model. In particular, the opposite banks of gray matter in deep sulci are not always resolved as separate and can appear as fused together (Figure 1), leading to invalid models of the cortical layer and propagating errors further into quantitative measurements (e.g., cortical thickness). This may present a particular challenge for an automated reconstruction algorithm, requiring specific means for an automatic detection and correction of topologically and geometrically problematic cases.

Reconstruction of cortical surface models received considerable attention in neuroimaging research. Here, we only briefly overview some state-of-the-art methods; please refer

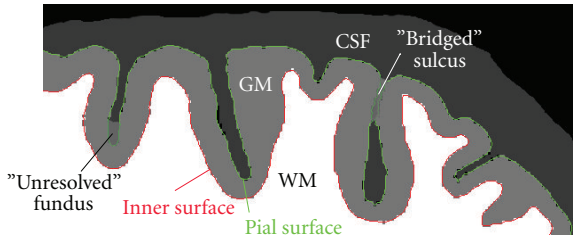


Figure 1: Schematic illustration of a fragment of brain slice. Contours of the inner and pial surface are marked in red and green. Due to partial volume effects and limited resolution, adjacent banks of gray matter in some sulci may appear as fused together, creating either a "bridged" sulcus or an unresolved sulcal fundus (a "buried" sulcus). Note that a "bridged" sulcus creates a topological defect, a handle, which may be corrected by a topology-preserving model, whereas a "buried" sulcus does not change the topology.

to Han et al. [1] and Kim et al. [2] for additional discussion. A suite of algorithms for automated cortical reconstruction is implemented in the popular and freely available FreeSurfer software [3, 4]. FreeSurfer includes an algorithm for finding and correcting the topological defects in the initial WM/GM surface [5] and a method to deform the mesh for reconstructing the inner and pial surfaces. The deformable model is constrained by a second-order smoothing term [6] and by a mesh self-intersection prevention routine [3], which both help to resolve the boundaries between adjacent banks in tight sulci. The FreeSurfer automated toolchain is optimized for standard resolution T1-weighted MR images and conforms input data to 1 mm isotropic voxel size, as a rule. This is consistent with the fact that mesh self-intersection detection and prevention is computationally expensive (see [1, 6]) and does not scale well with increasing mesh resolution. Xu et al. [7] developed a deformable mesh model for reconstruction of the central cortical surface. The model deforms the topology-corrected initial WM/GM interface by forces derived from a smoothed gradient field [8] that was computed from a GM class membership function. The model does not perform a time-consuming check of mesh self-intersections, which is arguably less critical for finding the central surface, compared to the pial surface. Kim et al. [2] presented a different deformable mesh-based approach for reconstruction of a pial surface, which is called constrained Laplacian anatomic segmentation using proximity, or CLASP. The algorithm computes a Laplacian field mapping between the GM/WM interface and the skeleton of the partial volume classification of the CSF. The Laplacian map is then integrated into the deformable model's objective function, driving mesh vertices into locations with higher values of the Laplacian field. Terms for stretch and self-proximity are included to regularize the deforming mesh and prevent from mesh self-intersection inside sulci. The method by Kim et al. depends on accurate extraction of the CSF skeleton and therefore relies on an elaborate partial volume tissue classification algorithm. However, the accuracy of the Laplacian mapping may be compromised at locations, where the fused GM sulcal banks are not resolved. In addition, the computational cost of the self-proximity term may become prohibitive for high-resolution meshes. Zeng et al. [9] used implicit surfaces in a level set framework for simultaneous reconstruction of the inner and outer cortical surfaces coupled by the minimal and maximal distance constraint. However, this approach did not gain widespread use, because it does not preserve the topology of the evolving surfaces and, in some areas, the distance coupling term may suppress the data attachment term, resulting in geometrical inaccuracies [10]. Han et al. [1] described a method for automated reconstruction of cortical surfaces, called CRUISE, which is built around a geometric deformable model using level sets. To help resolve the cortical banks in sulci, a thin digital separating barrier is constructed using the anatomically consistent enhancement algorithm ACE [1, 11], which finds a skeleton of the weighted distance function computed from the Eikonal equation with a speed function modulated by the CSF class membership. At the core of the CRUISE method is a topology-preserving geometric deformable surface model, TGDM [1, 11, 12], which models the evolution of a level set function under the influence of signed pressure forces computed from tissue class membership values and curvature forces defined by the surface geometry. The central surface of the cortex is reconstructed by a TGDM with GGVF advection forces similar to those in Xu et al. [7].

We present a method, henceforth, designated dielectric layer field mapping, or DELFMAP, for the automated reconstruction of the cortical compartment from MR images, which is based on several partial differential equation (PDE) modeling stages. Our method is inspired by the work of Han et al. and uses a similar level set framework, but introduces a different perspective, consolidating all algorithmic stages around the key mathematical model of a potential field in an inhomogeneous dielectric medium. Our method scales well with image resolution and has an advantage over other existing methods in reconstruction from high-resolution MR images with submillimeter voxel sizes, because (1) in contrast to deformable mesh models in FreeSurfer or CLASP, it avoids the computational cost of testing for mesh self-intersection and self-proximity; (2) similarly to CRUISE, it uses an efficient narrow-band algorithm for the level set evolution; (3) in contrast to CRUISE that requires solving a system of three second-order PDEs in GGVF, our method solves just one second-order PDE and does not need an intermediate step of reconstructing a central cortical surface.

Preliminary results of this work were presented in two conference publications [13, 14]. This report expands on the methodology and experimental results and adds a validation study that performs cross-comparison of our method's cortical reconstruction results with those obtained using FreeSurfer [3, 4] on standard resolution data for 20 healthy young subjects (test-retest repeated scans) from the OASIS database [15].

2. Methods

The DELFMAP method proceeds as follows. A potential field is computed using the mathematical model of an electric field in an inhomogeneous dielectric medium, where

the segmented WM poses as a charged conductive object and the classified GM poses as an inhomogeneous dielectric layer with permittivity proportional to GM class probability values. This electrostatic model serves the purpose of concentrating the flux of the mapping flow in a layer of voxels classified as GM and helps to identify the separating barriers between cortical banks in sulci, where the mapping flow collides. Correspondence trajectories following the lines of the potential field and geodesic distances from WM boundary are determined using PDEs, and a digital skeleton of the sulcal medial surface separating GM sulcal banks is derived by finding collisions in the correspondence trajectories and shocks in the distance field. The computed electric field retains the desired laminar properties of the Laplacian mapping in the bulk of the cortical layer and is used as the potential flow that maps the inner surface to the outer. The outer (pial) cortical surface is reconstructed using a geometric deformable model level set framework [16] with an advection along the gradient of the potential field, which is constrained by the identified skeleton of the sulcal medial surfaces and (optionally) by a maximal distance/proximity constraint.

2.1. Image Processing Chain. DELFMAP takes as input a set of volumetric images containing WM and GM tissue class probability/membership functions and a refined WM model, supplied either as a topology-corrected WM binary segmentation or as a WM/GM interface level set function. The overall chain of general image processing steps is outlined as follows (Figure 2) (1) A T1-weighted volumetric MR image is (optionally) aligned with the stereotaxic coordinate system, interpolated to isotropic voxel size, and is preprocessed with a brain-peeling algorithm that derives a mask of voxels related to the cerebral tissues only. (2) The brain image is corrected for intensity inhomogeneities and is classified into WM,GM,CSF/background probability images. (3) A raw WM binary segmentation is derived from the class probability images (by thresholding or a maximum-probability rule), and brain stem and cerebellum are (optionally) removed from the WM segmentation. (4) A topology-corrected WM volume is obtained from the raw WM binary segmentation by an automated algorithm or by manual editing, or a combination of both. (5) DELFMAP uses the output of step 2 and step 4 to reconstruct the inner and outer cortical surfaces. We note that steps 1–4 are common to many brain MR image processing workflows, therefore DELFMAP can be easily integrated with a wide variety of toolchains. More specifically, we used processing steps described in Yang and Kruggel [17] in our experiments with 3-Tesla in-vivo images, and we applied algorithms described in Kruggel et al. [18] for the analysis of exvivo high-resolution images. In step 4, for exvivo images, we used manual editing for filling ventricles and correcting large topological defects, and we applied a topological region-growing algorithm similar to the one in Kriegeskorte and Goebel [19] to obtain a genus zero WM binary object. In cross-validation with FreeSurfer on the OASIS data sets, we used the FreeSurfer's processing toolchain for the initial steps that are common between the two methods (i.e., steps 1–4 that lead to a topologically-corrected WM segmentation); therefore, the cross-method comparison of cortical reconstructions is not confounded by differences in preprocessing approaches. Finally, we emphasize that, in all our experiments involving DELFMAP, the tissue classification was performed by a modified version (see [18]) of the adaptive fuzzy clustering algorithm [20] augmented with a spatial regularization term [1]; this also applies to GM and WM tissue classification that was used by DELFMAP in cross-validation study on the OASIS data sets.

2.2. Inner Cortical Surface. The inner cortical surface is reconstructed by a deformable model (Figure 2, step 5.0) that smooths the initial WM/GM interface, which is determined by the corrected WM segmentation. For this purpose, we use a topology-preserving geometric deformable model (similar to [12]), which is described in detail in Section 2.6. For smoothing, we typically run 2-3 iterations of the deformable model with the mean curvature term only. We will denote the "inside" region of the level set function representing the inner cortical surface by Ω_w.

2.3. Electric Field Model. A potential field is found as a solution to the PDE modeling an electric field around a charged conductive object (WM) insulated by a dielectric layer (GM) having spatially inhomogeneous electric permittivity, which is set proportional to GM tissue class probability (Figure 2, step 5.1). In such a model, the flux of the electric field is confined in regions of higher permittivity, that is, where GM class probability is higher; therefore, trajectories following the lines of the electric field trace through the GM layer before exiting into the background space. Thus, the flux of the mapping flow is concentrated in a layer of voxels classified as GM. Let Ω denote the 3D image domain with the boundary $\Gamma(\Omega)$. We will denote WM and GM tissue class probability images by $P_w(\vec{r})$ and $P_g(\vec{r})$ ($\vec{r} \in \Omega$), where $\vec{r} = (x, y, z)$ is a 3D point. Let $\varphi(\vec{r})$ denote a potential field, a scalar function defined over Ω. Let $\varepsilon(\vec{r})$ denote another scalar function, called permittivity and computed from class probabilities as follows:

$$\varepsilon(\vec{r}) = 1 + (\varepsilon_{\max} - 1)\Big(C_d(\vec{r})P_w(\vec{r}) + P_g(\vec{r})\Big), \qquad (1)$$

where $\varepsilon_{\max}$ is the maximum permittivity of the insulating layer ($\varepsilon_{\max}$ should be $\gg 1$ in order to emphasize the inhomogeneity of the dielectric layer; $\varepsilon_{\max} = 100$ was used, and $\varepsilon_{\max} = 1000$ was tested with similar results). Thus, permittivity is close to $\varepsilon_{\max}$ when WM and/or GM class probabilities are high and is close to 1 when they are low. Note that the WM probability is included above only to ensure a proper transition of the field near the WM/GM interface, where some border voxels can be classified with low GM but high WM probability, for example, because a smoothed interface can slightly deviate from the initial WM segmentation. The inclusion of WM probability is therefore limited by the constraint field C_d, which is computed by thresholding of the WM chamfer distance transform D_{cmf} as $C_d = \{1 \text{ if } D_{\text{cmf}} < d_{\min}, 0 \text{ otherwise}\}$, where the distance threshold $d_{\min}$ can be set at the lower bound on cortical thickness ($\approx$1 mm), just enough to ensure a "high-permittivity" transition via

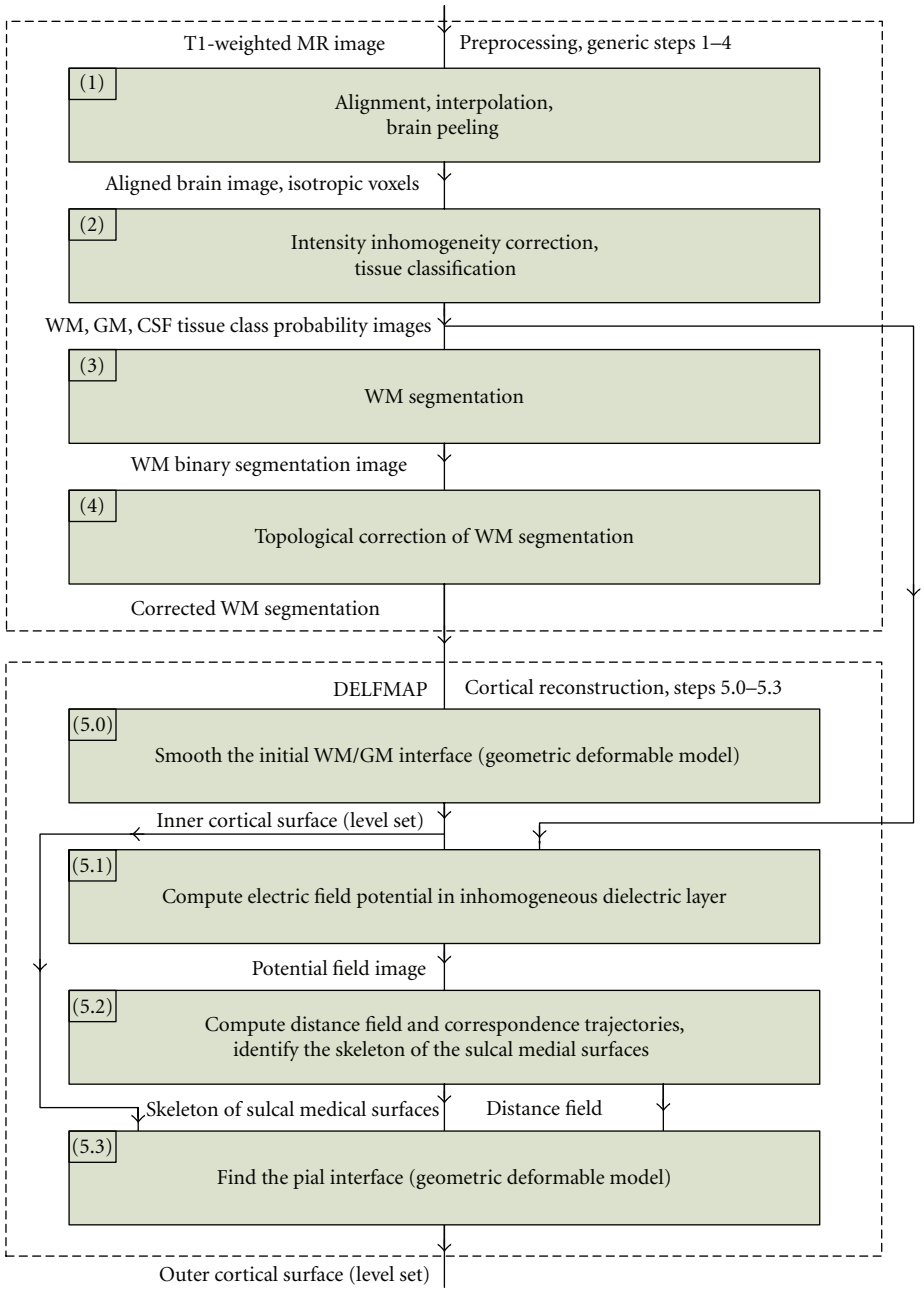


Figure 2: Block diagram of the overall image processing chain, where the DELFMAP method addresses the reconstruction of cortical surfaces (steps 5.0–5.3) after the preprocessing stage (steps 1–4).

boundary WM voxels to the layer of GM voxels. The potential field is found as a solution of Maxwell's equation for an electric field inside inhomogeneous dielectric medium in the absence of free charges:

$$\nabla\left(\varepsilon(\vec{r})\vec{E}(\vec{r})\right) = \nabla\varepsilon\nabla\varphi + \varepsilon\Delta\varphi = 0. \tag{2}$$

Equation (2) assumes that the dielectric medium has linear and isotropic properties; therefore, ε is a scalar, not a tensor. Boundary conditions are specified as $\varphi(\vec{r} \in \Omega_w) = 1$ and $\varphi(\vec{r} \in \Gamma(\Omega)) = 0$, that is, the potential is set to one in the WM core and is set to zero on the boundary of the image volume. The solution of the PDE $\varphi(\vec{r} \in \Omega \setminus \Omega_w)$ can be obtained as a steady-state solution ($\partial\varphi/\partial t \to 0$) of a corresponding nonstationary equation:

$$\frac{\partial\varphi}{\partial t} = \nabla\varepsilon\nabla\varphi + \varepsilon\Delta\varphi. \tag{3}$$

Equation (3) can also be viewed as describing the diffusion in inhomogeneous medium, where $\varepsilon(\vec{r})$ is a spatially varying but stationary diffusion coefficient and $\varphi(\vec{r}, t)$ is the concentration of the diffusing substance. This allows for a different physical interpretation of the model: we seek a steady-state spatial distribution of "particles" diffusing from

WM source into the medium with a diffusivity proportional to the GM class probability. Qualitatively, it is expected that "particles" would diffuse more freely in GM; therefore, the lines of the gradient field $\nabla\varphi$ would tend to concentrate in the GM compartment. Equation (3) can be discretized and solved iteratively as described by [21], for example, using the Jacobi method [22].

2.4. Distance Field and Correspondence Functions. Lines of the potential field φ are defined as a family of curves that are at each point tangent to the gradient $\nabla\varphi$. Let $d(\vec{s},\vec{r})$ denote the length of a line segment originating at some point in WM boundary $\vec{s} \in \Gamma(\Omega_w)$ and ending in point $\vec{r} \in \Omega \setminus \Omega_w$. If, for any point $\vec{r}$, there is one and only one streamline passing through it, then $d(\vec{r})$ defines a distance field. It is possible to compute the distance field by integrating trajectories explicitly in a Lagrangian framework. Alternatively, using the method described in Yezzi and Prince [23], the distance field can be found as a solution of a PDE in an Eulerian framework on a fixed grid. We note that $\nabla\varphi/\|\nabla\varphi\|$ is the unit tangent field of the potential field φ. Then, it can be shown that the distance field d must satisfy the following PDE:

$$\frac{\nabla\varphi}{||\nabla\varphi||} \cdot \nabla d(\vec{r}) = 1, \tag{4}$$

with the boundary condition $d(\vec{r} \in \Gamma(\Omega_w)) = 0$. Correspondences along streamline trajectories can be computed in a similar way. More specifically, let $\vec{\psi} = [\psi_1(\vec{r}), \psi_2(\vec{r}), \psi_3(\vec{r})]$ denote a vector of correspondence functions, which establishes a correspondence between a point in the field domain $\vec{r} \in \Omega \setminus \Omega_w$ and a "source" point in the WM boundary $\vec{\psi} \in \Gamma(\Omega_w)$. These correspondence functions ψ_i can be found as solutions of the following PDE (see [24]):

$$\frac{\nabla\varphi}{||\nabla\varphi||} \cdot \nabla\psi_i(\vec{r}) = 0, \tag{5}$$

with boundary conditions $\psi_i(\vec{r} = [x_1, x_2, x_3] \in \Gamma(\Omega_w)) = x_i$, where $i = 1, 2, 3$.

The first-order PDEs (4) and (5) can be solved using the numerical implementation described by Yezzi and Prince [23]. In principle, finite spatial discretization may violate the one-to-one correspondence property of the flow by clamping several streamline paths into one point on a grid, so the solutions $d(\vec{r})$ and $\vec{\psi}(\vec{r})$ may experience numerical convergence problems in some grid locations. In practice, we found that such problematic points are very sparse and do not impede numerical convergence in the computational domain at large. These points are usually detected among other "shocks" in the distance field by a skeletonization method (Figure 2, step 5.2), which is described next.

2.5. Skeleton of the Sulcal Medial Surface. Inside sulci, streamlines originating from opposite cortical banks collide (due to spatial discretization), which results into shocks in the distance field and into "discontinuities" in the correspondence functions. Shocks or singularities of a distance field d are defined as a set of points, where spatial derivatives of the field are not well-behaved, that is, the gradient ∇d is not well defined. Such shocks appear as discontinuities or sinks in the field. Note that even though the potential field in our model should be, in theory, free from the sinks (because there are no free charges), they may appear in the distance field due to spatial discretization. Let $S \subset \Omega \setminus \Omega_w$, called a skeleton of the distance field, denote a set of points on a grid, where shocks are detected by a numerical procedure. Such numerical procedure can be based on finite difference approximations to ∇d, as described by Han et al. [1]. The observation is that a centered finite difference numerical scheme will produce values of $\|\nabla d\|$ that are significantly lower than 1 on the shock points and are close to unity elsewhere. Then, the skeleton can be detected as $S = \{\vec{r} \mid (\vec{r} \in \Omega \setminus \Omega_w) \wedge (d(\vec{r}) > d_{\min}) \wedge (\|\nabla d(\vec{r})\| < T)\}$, where $d_{\min}$ is a minimum distance parameter set at the lower bound on cortical thickness and T is a specified threshold value ($T < 1$; values $d_{\min} = 1\,\text{mm}$ and $T = 0.8$ can be used, similarly to ACE in [1]). We found that the skeleton can be robustly detected by a novel algorithm based on the analysis of the correspondence function [14]. Recall that $\vec{\psi}(\vec{r}_0)$ is a vector with coordinates of the streamline's source point at WM boundary. A streamline collision can be detected if, in the neighborhood of $\vec{r}_0$, there are correspondences to source points that are "distant" between themselves. More formally, the skeleton can be determined as $S = \{\vec{r} \mid (\vec{r} \in \Omega \setminus \Omega_w) \wedge \max_i \|\vec{\psi}(\vec{r}) - \vec{\psi}(\vec{r}_i)\| > D_{\min}\}$, where $\vec{r}_i \in N_n(\vec{r})$. We used $D_{\min} = 4$ voxels and 6 adjacent points $N_6(\vec{r})$ in our computations.

2.6. Geometric Deformable Model. The geometric deformable model uses an implicit representation of a surface, embedding it into a level set function $\phi(\vec{r},t)(\vec{r} \in \Omega)$. The evolving interface is represented by the zero-level set $\Phi(t) = \{\vec{r} \mid \phi(\vec{r},t) = 0\}$ (see [16]), and it can be retrieved with subvoxel resolution by an isosurface algorithm (e.g., marching cubes). In our model, evolution of the level set function is described by the following PDE that has an advection and a mean curvature term:

$$\frac{\partial\phi(\vec{r},t)}{\partial t} + w_\alpha \vec{V}(\vec{r}) \cdot \nabla\phi(\vec{r},t) = w_\kappa \kappa(\phi) ||\nabla\phi(\vec{r},t)||, \tag{6}$$

where $\vec{V}$ is the advection velocity vector field, κ is the mean curvature, and w_κ are weights of the respective terms ($w_\alpha, w_\kappa \geq 0$). The mean curvature of the interface embedded in the level set function is [16]

$$\begin{aligned}\kappa &= \nabla \cdot \left(\frac{\nabla\phi}{||\nabla\phi||}\right) \\ &= \Big(\phi_x^2\phi_{yy} - 2\phi_x\phi_y\phi_{xy} + \phi_y^2\phi_{xx} \\ &\quad + \phi_x^2\phi_{zz} - 2\phi_x\phi_z\phi_{xz} + \phi_z^2\phi_{xx} \\ &\quad + \phi_y^2\phi_{zz} - 2\phi_y\phi_z\phi_{yz} + \phi_z^2\phi_{yy}\Big)/||\nabla\phi||^3,\end{aligned} \tag{7}$$

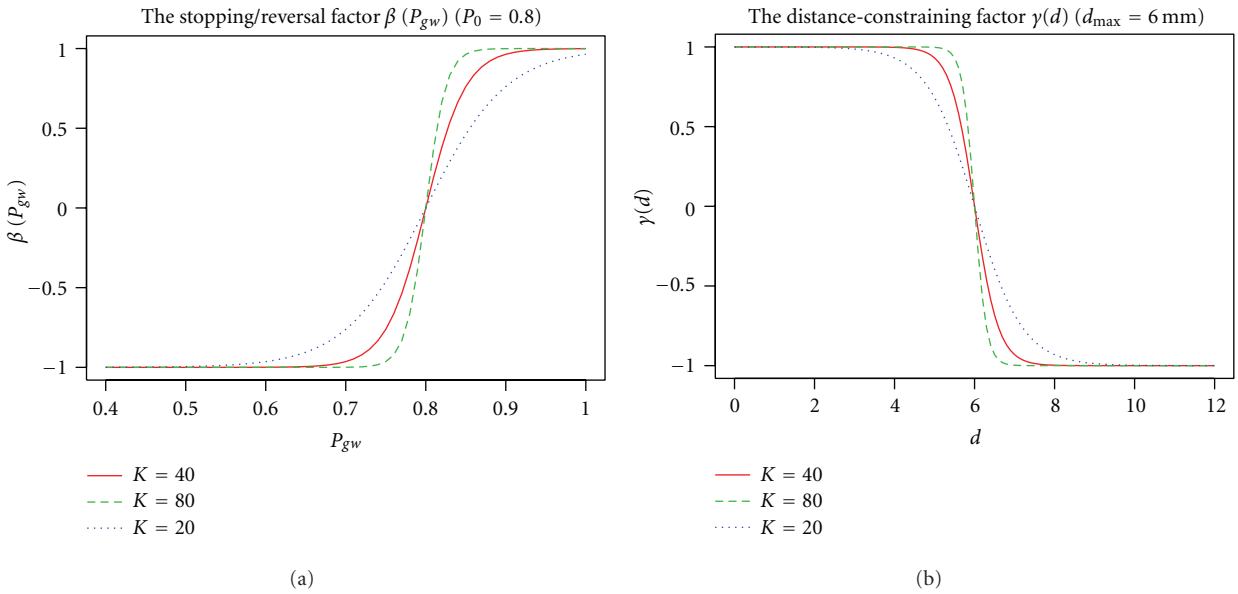


Figure 3: Plots of the stopping/reversal factor β (a) and the distance-constraining factor γ (b) at different values of the "steepness" constant K (solid red line: default $K = 40$; dashed green line: $K = 80$; dotted blue line: $K = 20$).

where the subscripts x, y, z denote partial derivatives. The advection velocity vector field $\vec{V}(\vec{r})$ is derived from the gradient of the potential φ or distance field d:

$$\vec{V}(\vec{r}) = \begin{cases} -\beta(\vec{r})\left(\dfrac{\nabla\varphi(\vec{r})}{\|\nabla\varphi\|}\right) \\ \text{or} \\ \beta(\vec{r})\left(\dfrac{\nabla d(\vec{r})}{\|\nabla d\|}\right), \end{cases} \tag{8}$$

where $\beta(\vec{r})$ is a stopping/direction-reversal factor computed from the GM/WM class probabilities. For example, this factor can have a form of a logistic function:

$$\beta(\vec{r}) = \frac{2}{1+\exp\left(-K\left[P_{gw}(\vec{r}) - P_0\right]\right)} - 1, \tag{9}$$

where K is the constant controlling the steepness of the slope of the sigmoid curve and P_0 is the GM class probability threshold value that determines the "set-point" of the deformable model. Figure 3 illustrates how the factor β depends on GM and WM probability P_{gw}. In our experiments, a moderately steep sigmoid curve with $K = 40$ and the threshold $P_0 = 0.8$ were used. For spatial regularization, the combined GM and WM class probability $P_{gw}(\vec{r}_0)$ can be calculated as a weighted sum over the (closed) neighborhood of the point $\vec{r}_0$:

$$P_{gw}(\vec{r}_0) = \sum_{\vec{r}_i \in \{\vec{r}_0, N_n(\vec{r})\}, \vec{r}_i \notin S} w_i\left(P_g(\vec{r}_i) + P_w(\vec{r}_i)\right), \tag{10}$$

where w_i are the neighborhood weights (e.g., $w_i = 0.5/n$, where $n = 18$ or 26, and for the central point $w_0 = 0.5$), and the skeleton of the sulcal medial surfaces S is used for masking of class probability values in separating barriers. As an option, the stopping factor β in (8) can be modified to include the distance-constraining factor:

$$\beta_1 = |\beta(\vec{r})|\,|\gamma(\vec{r})|\,\mathrm{sgn}(\beta, \gamma), \tag{11}$$

where the sign function is an "OR" combination of two signs:

$$\mathrm{sgn}(a, b) = \begin{cases} -1, & \text{if } a < 0 \text{ or } b < 0, \\ 1, & \text{otherwise}, \end{cases} \tag{12}$$

and the distance-constraining factor γ can also have a form of a logistic function:

$$\gamma(\vec{r}) = \frac{2}{1+\exp(-K[1/2 - \min(d(\vec{r}), 2d_{\max})/2d_{\max}])} - 1. \tag{13}$$

In (13), $d_{\max}$ is a parameter constraining the maximum distance of advection along the streamlines of the gradient field (i.e., a proximity constraint that limits the thickness of the reconstructed cortical layer). We used $d_{\max} = 6$ mm (see Figure 3) in the reported cortical reconstructions, that is, the maximum distance constraint was set above the anatomically plausible upper bound on cortical thickness and therefore was affecting only the artefactual or noncortical gray matter areas.

Our numerical implementation for solving the level set (6) is based on the narrow-band algorithm [12, 16, 25]. The initial level set function is computed as a signed-distance function (SDF) of the initial interface in the corrected WM image using the fast marching method (FMM, [16, 26]). By standard convention, "inside" points are represented by negative values of the SDF. During the evolution, the level set function $\phi(\vec{r}, t)$ is maintained close to the SDF by periodic

```
{Compute time step for each point in the narrow band}
for all r_i ∈ NarrowBand do
    {1. Compute the updated value}
    φ_new ← φ(r_i, t_k) + Δt Δφ(r_i, t_k)
    {2. Check if there is a sign change}
    if sgn(φ_new) == sgn(φ(r_i, t_k)) then
        {3.1 No sign change}
        φ(r_i, t_{k+1}) ← φ_new {apply the update}
    else if r_i ∈ S then {Check if is in the barrier}
        {3.2 Is in the barrier, do not allow sign change}
        φ(r_i, t_{k+1}) ← ε {set to a small positive value}
    else {3.3 Is clear; check for topology change}
        if IsSimple (φ(r, t_k), φ_new, r_i) then
            {3.3.1 No change in topology}
            φ(r_i, t_{k+1}) ← φ_new {apply the update}
        else {3.3.2 Do not allow topology change}
            φ(r_i, t_{k+1}) ← ε · sgn(φ(r_i, t_k)) {set to a small value of the same sign}
        end if
    end if
end for
```

Algorithm 1: The level set function update algorithm.

reinitialization with the FMM. The advection term in (6) is discretized based on the upwind differencing scheme (for details, see [16]), and the curvature term is discretized along the lines of (7) using the central differencing scheme [22]. A pseudocode outlining the narrow-band algorithm is described elsewhere (e.g., in [12, 25]). In Algorithm 1 pseudocode we focus on the core part that deals with the time-step update of the level set function. The update algorithm is novel in the way it uses the skeleton of the sulcal medial surface to create barriers for the evolving interface. In addition, the algorithm has a built-in rule preserving the digital topology of the deformed model [1, 12] that is based on the concept of simple points [27] (function IsSimple() in Algorithm 1, see details in [13]), which guarantees that the deformed surface retains the same topology as the initial WM/GM surface.

As already mentioned, the inner cortical surface is reconstructed by a few iterations of the model with the curvature term only ($w_\alpha = 0, w_\kappa = 1$) (Figure 2, step 5.0). In step 5.3 of Figure 2, the outer cortical surface is first reconstructed by a model using the advection term only ($w_\alpha = 1, w_\kappa = 0$) until convergence (i.e., until the relative amount of change in the SDF per iteration becomes small, for example, lower than 10^{-4}) or for a specified number of time steps and then smoothed by a few iterations with the curvature term, similarly to the inner surface.

3. Experiments and Results

Our algorithm was implemented in C++ in the Linux environment and ran on a PC with 2.5 GHz AMD-64 CPU and 4 GB RAM, unless otherwise noted. The algorithm's performance was evaluated on simulated test cases with a simplified geometry of a sulcus, on simulated MRI datasets, on standard resolution T1-weighted MR images of human brains, and on high-resolution (sub-mm) MR images of extracted brain hemispheres.

3.1. Simulated Data. The first test case is intended to illustrate the effect of the inhomogeneous dielectric model used in DELFMAP and shows the difference between the field produced with a nonuniform permittivity and the field computed with the uniform permittivity ($\varepsilon = 1$, the Laplacian field). Test images simulate a simplified 3D geometry of a sulcal fold and contain two WM stalks separated by the sulcal space (with a curvature radius of 10 mm); the WM is covered by a layer of GM having unequal thickness at the opposing banks and a smoothly varying thickness at the fundus (Figure 4(a)). Figure 4 shows the lines of the Laplacian field (Figure 4(b)) and the lines (Figure 4(d)) and isocontours (Figure 4(c)) of the field in the DELFMAP model. It can be seen that the "ridge" (where the field lines concentrate and the isocontours converge) of the DELFMAP field is close to the sulcal center line, whereas the "ridge" of the Laplacian field is at the geometric center.

The second test case demonstrates how the model resolves the barrier separating the two opposing cortical banks inside a sulcus. Test images simulate a fully resolved sulcus (with two banks fully separated by background), a sulcus with an unresolved fundus, and a sulcus with two banks bridged by unresolved voxels (the top row in Figure 5: left, middle, and right, resp.). The middle row in Figure 5 shows the cross-section of the sulcal medial surface (white lines) that was identified by the DELFMAP method. It can be seen that the method is capable of reconstructing the boundary surface separating the two cortical banks and finds a geometrically plausible solution in incompletely resolved cases. Side-by-side comparison of the results of our method and those of ACE (the bottom row in Figure 5) shows that skeletons produced by DELFMAP have a more regular

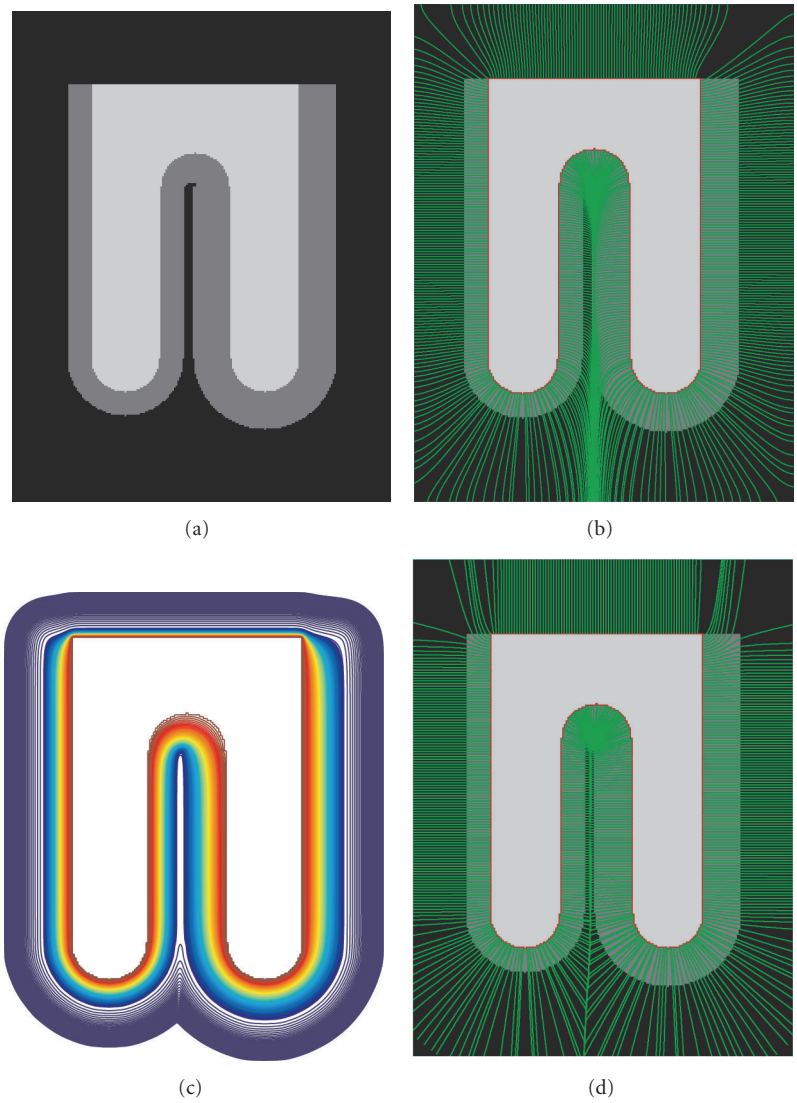


Figure 4: Cross-sections of simulated test images. (a) The input image; (b) field lines in the uniform permittivity model (Laplace equation). Bottom row: isocontours (c) and field lines (d) in the DELFMAP model with the dielectric layer (dark gray in the input image).

structure, whereas ACE skeletons can have small extraneous branches and discontinuities. Our method does not produce spurious detections very close to WM and thus does not require a minimum distance cut-off parameter, which is needed in ACE. In addition, our method is more robust with respect to noise (see [14]): skeletons produced by DELFMAP show very little degradation even at the highest noise level, while ACE skeletons are significantly affected by strong levels of noise.

Cortical reconstruction results for simulated brain phantom MR images [28] showed good reproducibility across various levels of simulated Gaussian-distributed noise and intensity inhomogeneity (see [13, 14]).

3.2. High-Resolution MR Images. Our method's performance is illustrated by results for high-resolution exvivo images, where, contrary to FreeSurfer, our method does not need to conform images to standard 1 mm isotropic voxel size. The algorithm was evaluated on three high-resolution (0.25–0.35 mm isotropic voxel size) images of explanted brain left hemispheres. DELFMAP reconstruction at 0.35 mm resolution took 67 min on a PC with 2.5 GHz AMD-64 CPU and 4 GB RAM. We tried to process the same 0.35 mm data with the recently released CRUISE plugin for MIPAV [29] on a cluster node with four Opteron 285 2.6 GHz cores and 32 GB RAM. Reconstruction of the inner surface took 28 min using 4.9 GB RAM, computation of GGVF took 32 min using 3.5 GB RAM, while reconstruction of the central and pial surfaces took 49 and 52 min using 5.3 and 5.1 GB, respectively, but did not produce adequate results with the default settings. DELFMAP computations at 0.25 mm resolution required 4.7 GB RAM and were successfully completed after 3 h 20 min. Examples of the reconstructed cortical surfaces overlaid on orthogonal cross-sections of a high-resolution MR image are shown in Figure 6.

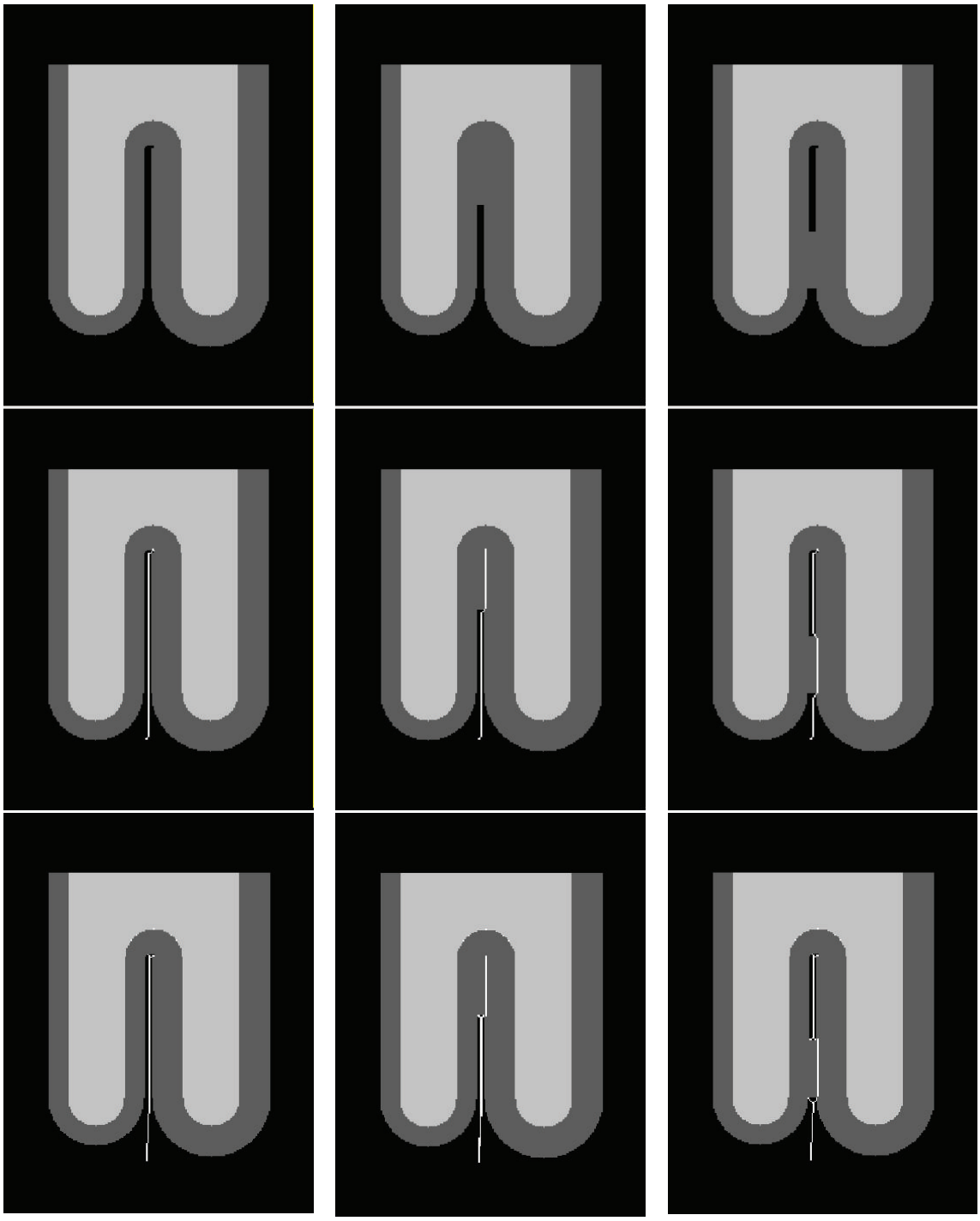


Figure 5: Cross-sections of simulated test images (left: fully resolved sulcus; middle: unresolved fundus; right: bridged sulcus). The white line shows the location of the identified sulcal medial surface skeleton. Comparison of DELFMAP (middle row) versus ACE (bottom row) shows that skeletons produced by DELFMAP have a more regular structure compared to ACE skeletons, which can have small extraneous branches and discontinuities. In the bottom row (ACE), small spurious components are visible at the fundus very close to WM, which in ACE method have to be suppressed by thresholding the distance from WM.

Lateral views of pial surfaces of three brain samples (3D rendering of thickness maps) are shown in Figure 7, left column. Measured thickness values (mean 2.2 mm; stdev 0.7 mm) are in good agreement with the literature. Inflated maps (Figure 7 middle and right column) are intended for better visualization of the surface inside sulci; they were produced with 20 iterations of Laplacian smoothing of the mesh. Maps in the right column are color-coded with convexity values that were computed as vertex travel distances during smoothing/inflation, similarly to FreeSurfer [4]. On

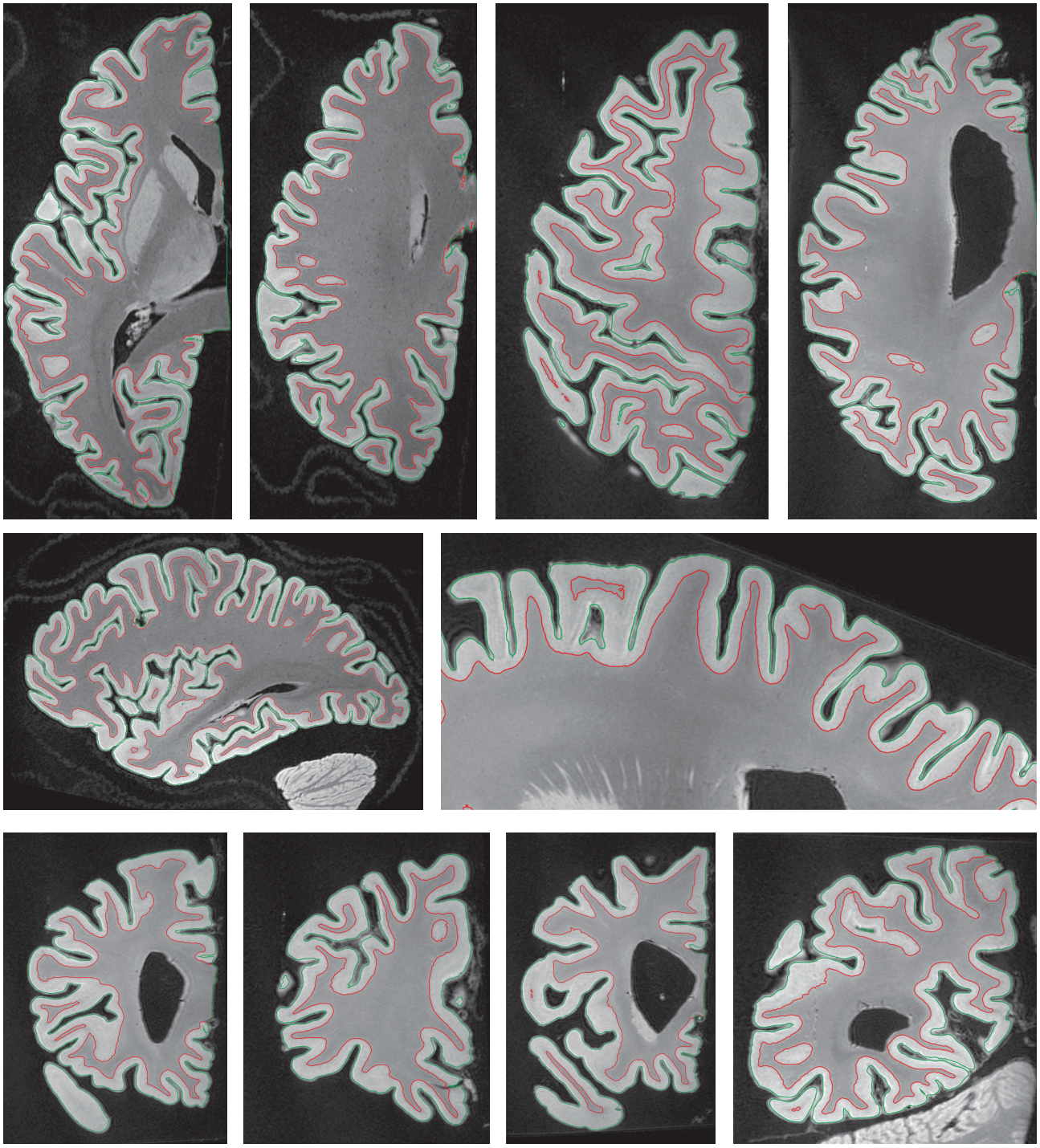


Figure 6: Isocontours of the zero level sets of reconstructed cortical surfaces overlaid on cross-sections of high-resolution MR images (red: the inner surface; green: the outer surface; top, middle, and bottom rows: examples of axial, sagittal, and coronal sections (not to scale), resp.).

convexity maps, gyral crowns appear in blue color and sulcal fundi appear in yellow-orange. Thickness and convexity maps demonstrate noticeable correlation (Pearson's correlation coefficient computed over the entire surface mesh is 0.24, 0.22, and 0.28 for the three brain samples shown, that is, significantly different from zero at the 0.05 level), which is in good agreement with the known anatomical fact that cerebral cortex is generally thicker on gyral crowns and thinner in sulcal depths.

3.3. Cross-Validation with FreeSurfer: Test-Retest Precision. Our method was validated by cross-comparison of cortical reconstruction results with those obtained using FreeSurfer. Standard resolution images for 20 right-handed healthy

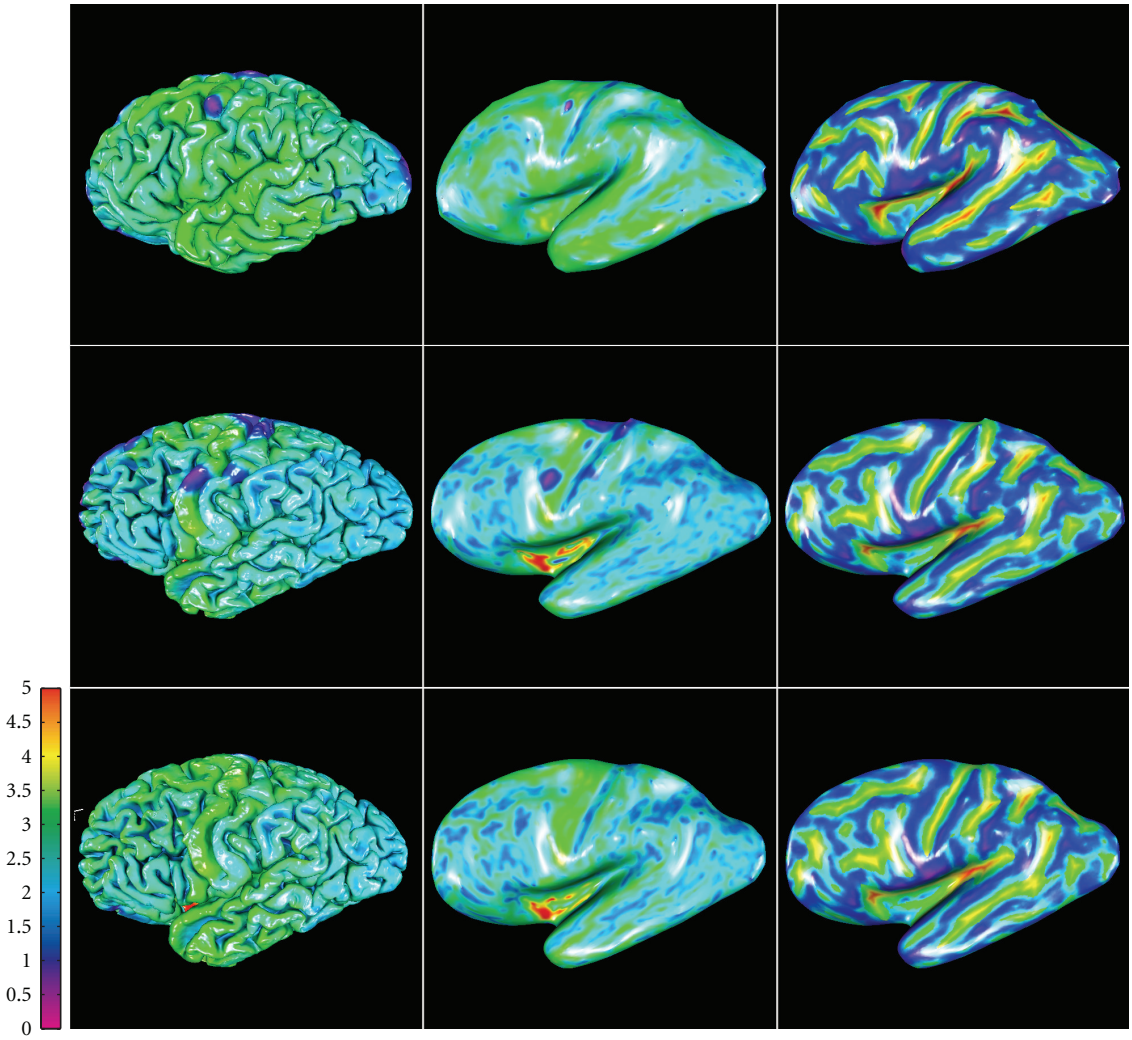


Figure 7: Lateral view of pial surfaces from three high-resolution datasets (left column: thickness maps; middle column: inflated thickness maps; right column: inflated convexity maps).

young subjects (age 19–34, average 23; 8 males/12 females) were selected from the cross-sectional OASIS database [15]. For each subject, data are available from two scan sessions (test and retest) separated by a short delay (1–89, average 21 days), with four T1-weighted standard resolution images acquired per session. This relatively short delay between two consecutive scan sessions makes data sets suitable for the assessment of test-retest reproducibility (i.e., precision) of the analysis by comparing measurements between scan sessions.

First, we analyzed data sets using the default automated pipeline in FreeSurfer and obtained 40 cortical reconstructions (two per subject), each including a pial and a white surface mesh. Next, we exported images of extracted brains (without any intensity normalization/correction) and corrected WM segmentations from FreeSurfer, ran our tissue classification algorithm on images of extracted brains, and used these results in the DELFMAP toolchain to obtain another set of 40 cortical reconstructions. For a subvoxel resolution of a digital skeleton, solutions of PDE in (3)–(6) were computed on a grid with half-voxel spacing. Implicit level set surfaces were tessellated using connectivity-consistent marching cubes algorithm [12], and triangular meshes were simplified down to 300,000 faces by a topology-preserving variant of the mesh simplification method [30]. DELFMAP processing took approximately 30 min per brain hemisphere (at half-voxel 0.5 mm res. grid) and was twice faster than FreeSurfer's deformable model step (mris_make_surfaces program, took ≈70 min at 1 mm res.). FreeSurfer computes cortical thickness at each vertex as the average of the closest-point distance (Figure 8(a)) measured between the surfaces both ways using linked vertices [6]. Since vertices on pial and white surfaces are not linked in DELFMAP, which is not based on a deformable mesh model, for the cross-method comparison, we recomputed cortical thickness using an orthogonal projection distance measure [31] (Figure 8(b) and the Appendix) that is robust and universally applicable to results from both methods. We verified that the two cortical thickness measures were in close agreement on all 40 reconstructions obtained with FreeSurfer.

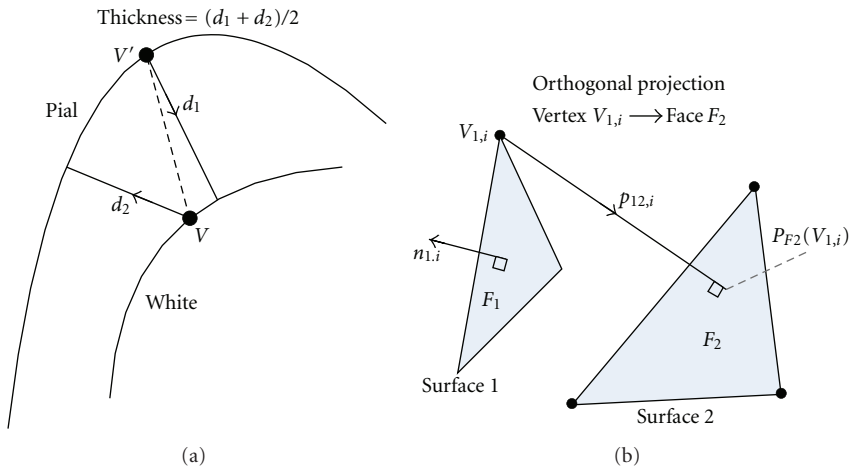


Figure 8: Illustration of two different approaches of defining a distance between two surface meshes. (a) The thickness measure defined in FreeSurfer (2D schematic drawing). (b) The (signed) distance measure defined by closest orthogonal projection.

The geometric precision or test-retest reproducibility of cortical reconstruction was evaluated independently for FreeSurfer and for DELFMAP as follows. For each subject, test and retest MR images (averages of 4 aligned scans from the first and the second session, resp.) were rigidly registered to each other using FSL FLIRT [32]. The obtained rigid transformation was applied to the first set of surface meshes, aligning the test surfaces to the retest ones. Next, signed and absolute distances (the Appendix, (A.1) and (A.2)) were measured between aligned test and retest white/pial surface meshes, and surface-wise mean and standard deviation were computed, as well as the group-wise statistics. In addition, we evaluated the test-retest precision of cortical thickness measured with FreeSurfer and with our method using the standard methodology described in the cortical thickness reproducibility study in Han et al. [33], which consists of the following four steps: (1) rigid registration of two repeated scans of each subject; (2) computation of a thickness difference map for each subject (on the first surface, using point-correspondences established according to closest Euclidean distance in registered space); (3) resampling the thickness difference map to a common template (e.g., any subject surface or the FreeSurfer's average template); (4) computing the group-wise mean and standard deviation of the differences at every vertex of the template mesh. Resampling to a common template relies on FreeSurfer's intersubject registration by nonlinear surface morphing [34].

Results of both methods, the absolute distance measure $\mathrm{AD}_{\mathrm{mean}}$ and $\mathrm{AD}_{\mathrm{stdev}}$ between test and retest cortical surfaces (the Appendix, (A.3)), per subject hemisphere, were compared statistically using a Wilcoxon signed rank test, and results are reported as P values. For FreeSurfer WM surfaces, reproducibility is characterized by mean absolute error 0.19(Δ0.06) mm (where the Δ value in parentheses indicates a statistical spread for the group, equal to two stdev). For DELFMAP WM surfaces, mean absolute error is 0.24(Δ0.06) mm ($P = 9.5 \times 10^{-5}$). For DELFMAP pial surfaces, reproducibility is characterized by a mean absolute error 0.24/0.25(Δ0.04) mm (L/R) that is similar in FreeSurfer (L: $P = 0.37$, R: $P = 0.16$, see details in Table 1). The standard deviation of the absolute distance $\mathrm{AD}_{\mathrm{stdev}}$ is much lower in DELFMAP than in FreeSurfer (L: $P = 8.2 \times 10^{-5}$, R: $P = 3.2 \times 10^{-4}$) which can be interpreted as a "tighter" reconstruction of pial surfaces in DELFMAP. Table 1 summarizes the statistics of the test-retest analysis. The mean absolute difference of the cortical thickness is similar in both methods (L: $P = 0.10$, R: $P = 0.28$), but the corresponding standard deviation is again much smaller in DELFMAP than in FreeSurfer (L: $P = 1.9 \times 10^{-6}$, R: $P = 1.0 \times 10^{-4}$). To summarize, test-retest precision of cortical thickness measurement is similar in DELFMAP and FreeSurfer in terms of the mean error, which is close to a quarter of the voxel size, but is "tighter" in DELFMAP in terms of surface-wise variance in absolute differences.

3.4. Cross-Validation with FreeSurfer: Intermethod Accuracy. The geometric accuracy of our method was evaluated by cross-comparison with FreeSurfer as follows. For each cortical reconstruction (two per subject), white (W) and pial (G) surfaces (Wf, Gf) were exported from FreeSurfer and a cortical thickness map $\mathcal{A}_{\mathrm{GfWf}}$ (A.2) was computed on pial surface. Next, maps of intermethod geometric differences ($\mathcal{D}_{\mathrm{WfWd}}$, $\mathcal{D}_{\mathrm{WdWf}}$, $\mathcal{D}_{\mathrm{GfGd}}$, $\mathcal{D}_{\mathrm{GdGf}}$) were computed as signed distances (A.1) between white or pial surfaces reconstructed with FreeSurfer and DELFMAP (Wd, Gd). On these geometric-difference maps (40 sets, four maps per set), surface-wise statistics $\mathrm{D}_{\mathrm{mean}}$, $\mathrm{D}_{\mathrm{stdev}}$, $\mathrm{AD}_{\mathrm{mean}}$, and $\mathrm{AD}_{\mathrm{stdev}}$ (A.3) were computed. In addition, maps of intermethod thickness differences were built using the cortical thickness reproducibility analysis steps 2–4 [33] as described in the previous section, except for using two pial surfaces from both methods in step 2 (we emphasize that for both FreeSurfer and DELFMAP, the compared thickness maps were measured by the same method, that is, as $\mathcal{A}_{\mathrm{GW}}$). The 40 individual maps were resampled to a common template and averaged into group-wise maps of mean difference and standard deviation. The group-wise maps of intermethod cortical thickness measurement differences allow to assess and visualize any

Table 1: Precision analysis: summary of the group-average statistics for the signed distance (SD) and absolute distance (AD) measure (in mm) between test and retest surfaces (surface: DF—DELFMAP, FS—FreeSurfer; L/R: left/right hemisphere; mean: a group average of a surface-wise mean of the distance; stdev: a group average of a surface-wise stdev of the distance; "> X mm (%)": (group-average) percentage of surface points where AD was greater than X mm; values in parentheses indicate the statistical spread within the group, measured by the group-wise stdev).

Surface	L/R	Signed distance		Absolute distance			
		Mean (mm)	stdev (mm)	Mean (mm)	stdev (mm)	>1 mm (%)	>2 mm (%)
DF pial	L	−0.02 (0.03)	0.35 (0.06)	0.24 (0.02)	0.25 (0.04)	1.4 (0.3)	0.2 (0.1)
	R	−0.01 (0.04)	0.37 (0.09)	0.25 (0.02)	0.26 (0.06)	1.5 (0.4)	0.2 (0.2)
FS pial	L	−0.02 (0.04)	0.37 (0.10)	0.24 (0.02)	0.28 (0.06)	1.9 (0.4)	0.3 (0.2)
	R	−0.03 (0.04)	0.39 (0.13)	0.24 (0.03)	0.29 (0.08)	2.0 (0.4)	0.3 (0.2)
DF white	L	−0.01 (0.07)	0.34 (0.08)	0.24 (0.02)	0.24 (0.05)	1.0 (0.3)	0.2 (0.1)
	R	+0.02 (0.06)	0.35 (0.11)	0.24 (0.03)	0.24 (0.07)	1.0 (0.3)	0.2 (0.2)
FS white	L	+0.02 (0.02)	0.31 (0.10)	0.19 (0.02)	0.23 (0.07)	1.0 (0.3)	0.2 (0.2)
	R	+0.01 (0.02)	0.31 (0.13)	0.19 (0.03)	0.23 (0.08)	0.9 (0.3)	0.2 (0.2)

Table 2: Intermethod accuracy analysis: summary of the group-average statistics for distances between DELFMAP- and FreeSurfer-generated surfaces.

Surf.	L/R	Signed distance		Absolute distance			
		Mean (mm)	stdev (mm)	Mean (mm)	stdev (mm)	>1 mm (%)	>2 mm (%)
pial	L	−0.08 (0.04)	0.49 (0.02)	0.40 (0.02)	0.37 (0.02)	6.8 (1.2)	0.6 (0.2)
	R	−0.07 (0.04)	0.53 (0.02)	0.42 (0.02)	0.38 (0.02)	7.4 (1.3)	0.6 (0.2)
white	L	0.00 (0.04)	0.28 (0.01)	0.24 (0.01)	0.17 (0.01)	0.1 (0.1)	0.0 (0.01)
	R	0.00 (0.04)	0.29 (0.01)	0.24 (0.01)	0.18 (0.01)	0.0 (0.0)	0.0 (0.01)

Table 3: Intermethod accuracy analysis: summary of the group-average statistics for difference in cortical thickness measurement between DELFMAP and FreeSurfer.

L/R	Signed difference		Absolute difference			
	Mean (mm)	stdev (mm)	Mean (mm)	stdev (mm)	>1 mm (%)	>2 mm (%)
L	0.12 (0.07)	0.47 (0.03)	0.35 (0.03)	0.34 (0.03)	4.4 (1.5)	0.4 (0.1)
R	0.11 (0.08)	0.46 (0.03)	0.34 (0.03)	0.33 (0.03)	4.1 (1.4)	0.3 (0.1)

regional patterns of agreement/disagreement between the two methods. The intermethod geometric accuracy analysis statistics is summarized in Table 2 (averaged over 40 image sets, two per subject). It can be seen from the mean signed distance SD_{mean} that, on average, DELFMAP has a very small outward bias in pial surfaces (−0.08/−0.07(Δ0.08) mm, L/R; negative sign means FreeSurfer' surface is "inside" w.r.t. DELFMAP' surface). The intermethod accuracy can be characterized by the mean absolute distance AD_{mean} (0.40/0.42(Δ0.04) mm, L/R), which is less than a half of the voxel size. The share of pial surface vertices where the AD was larger than 1 mm is less than 10%; less than 1% of pial vertices had an AD larger than 2 mm.

The intermethod accuracy analysis of cortical thickness measurements, summarized in Table 3, is in good agreement with the above observations. On average, there is a small bias towards thicker values in DELFMAP (mean signed difference: 0.12/0.11(Δ0.16) mm, L/R; positive sign here means that DELFMAP-measured thickness is larger w.r.t. FreeSurfer). The intermethod accuracy, characterized by the mean absolute difference (0.35/0.34(Δ0.06) mm, L/R), is less than a half of the voxel size. The share of pial surface vertices where the absolute difference between thickness measurements was larger than 1 mm is less than 6%, and less than 1% of pial vertices had an absolute difference larger than 2 mm. An example comparing DELFMAP and FreeSurfer pial surface reconstructions side-by-side, for one subject, is shown in Figure 9 (colored with cortical thickness; see colorbar for color map and range of values). Overall, a good correspondence is visible, but some patterns of thickness difference are noticeable: (1) for FreeSurfer, thickness is larger (indicated as yellow) in the superior region of the frontal lobe and in some temporal regions (lateral view); (2) for DELFMAP, thickness is larger (indicated as orange) in the inferior occipitotemporal region (medial view, where the cerebellum is found); (3) for FreeSurfer, thickness is smaller (indicated as blue) in the medial orbitofrontal cortex (mOFC) region (medial view). These differences can be attributed and traced to the following segmentation trends in either of the two methods: (1) oversegmentation, by FreeSurfer, into meningeal space in superior frontal region and in temporal region (see Figure 10); (2) oversegmentation, by DELFMAP, into cerebellar gray matter in the inferior occipitotemporal region; (3) too conservative segmentation,

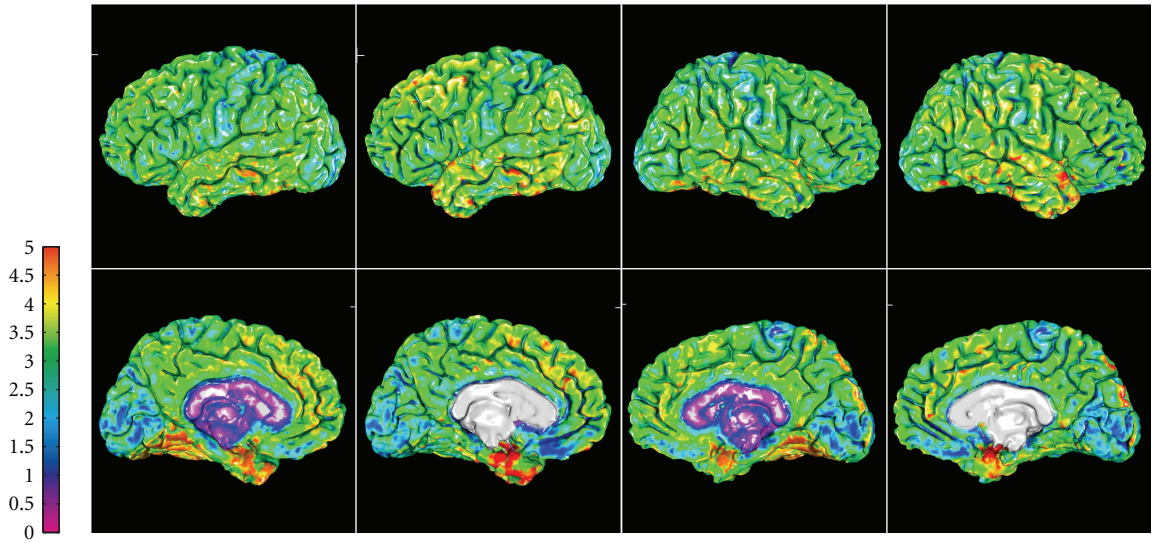


Figure 9: Example of side-by-side comparison of DELFMAP (column 1 and 3) and FreeSurfer (column 2 and 4) thickness maps (OAS1_202_1, on pial surfaces, left/right hemisphere in the left/right two columns, resp.; 1st row: lateral surface; 2nd row: medial surface; colorbar range 0–5 mm).

by FreeSurfer, in the mOFC region (too thin, less than 1.5 mm).

Regional patterns of intermethod geometric differences in pial cortical reconstructions are visible on group-average maps of geometric (Figure 11) and cortical thickness differences (Figure 12), where the above outlined three trends are also noticeable.

4. Discussion

We presented a novel PDE-based approach for reconstructing the cerebral cortex from MR images. We developed an accurate and scalable method that works on MR images with a high spatial resolution. Because high-resolution MRI begins to attract considerable attention in brain imaging research, a method that readily scales with imaging resolution is highly valuable. This scalability is achieved by using an implicit deformable surface model in a fast marching framework guided by a novel, computationally efficient model using potential field mapping. Our method requires much lower computational resources and has a much faster computation times than conventional methods. These demonstrated advantages come not only from an efficient practical implementation, but also from the design of our algorithms. For instance, other existing approaches that are based on deformable mesh models incur a significant computational cost associated with the mesh self-intersection (e.g., FreeSurfer) or mesh self-proximity (CLASP) term, which does not scale linearly with increasing mesh resolution. Although the computational cost of the straightforward mesh self-proximity term [2], which is quadratic $O(N^2/2)$ on the number of faces N, is significantly reduced in a mesh self-intersection prevention algorithm utilizing a spatial cache [3], it nevertheless remains supralinear. Similarly, the cost of another known efficient algorithm for mesh self-intersection detection, which is based on intersection of bounding boxes, is $O(N\log_2^3 N)$ [35]. In contrast to this, the computational complexity $O(Nk)$ of the narrow-band level set algorithm used in our method (and in CRUISE) is linear with respect to the size of the interface N (k is the width of the narrow band). This difference between a linear and a quadratic or supralinear algorithmic complexity, which can be tolerated when dealing with standard resolution images and meshes, becomes quite large at high resolutions. As to the comparison with the available CRUISE MIPAV software, our method's dramatic gain in speed is most likely due to differences in implementation but, at least in part, can be attributed to a smaller algorithmic cost of our method (e.g., solving one second-order PDE in DELFMAP versus a system of three second-order PDEs in GGVF, and not using an intermediate step of reconstructing a central cortical surface).

Although some algorithmic building blocks of our method were previously known to the medical image processing community (e.g., [1, 21, 23]), the central aspect of our method, that is, the use of the model of the potential field in the inhomogeneous dielectric layer introduced here, is novel and has attractive advantages. The novelty of our method is also in the newly introduced skeletonization algorithm that is based on the analysis of correspondence trajectories and in several novel aspects of the geometric deformable model (e.g., the constraint of the evolution by the medial surfaces, the maximal distance constraint of the advection, and the novel form of the advection stopping/direction-reversal factor β and the distance-constraining factor γ). We note that most of the design parameters introduced in Section 2 remain fixed, and the method is sensitive only to two settings, which can be easily tuned: the GM probability threshold P_0 (a "set-point") and the maximal distance $d_{\max}$ (which has strong influence only if set below the upper bound on cortical thickness).

The results from three high-resolution data sets demonstrate that the method is capable of reconstructing the outer

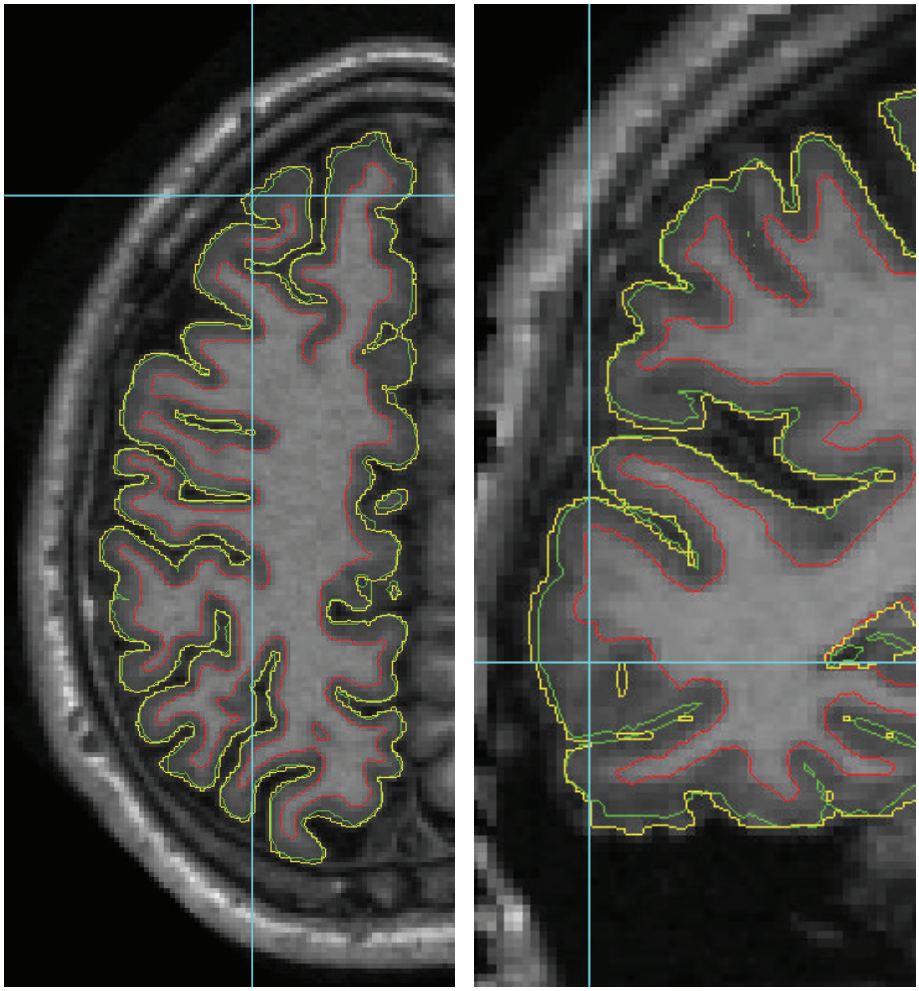


Figure 10: Contours of reconstructed cortical surfaces overlaid on the axial (left) and coronal (right) slice (red: inner surface; green: DELFMAP pial surface; yellow: FreeSurfer pial surface; note that the yellow contour appears jagged because it is displayed from FreeSurfer's volumetric signed distance function sampled at 1 mm grid, whereas red and green contours are from level set functions sampled at a finer resolution; left and right images are not to scale). On the left image at the cross-line cursor position (superior frontal region), the yellow contour of FreeSurfer's pial reconstruction oversegments into meningeal space, and a similar trend is noticeable next to cursor on the right image (temporal region).

cortical boundary with good geometric precision and accuracy, while guaranteeing the preservation of the initial surface topology. The method's performance is illustrated on synthetic images and on standard resolution MR brain images, where it compares favorably to existing methods in both quality and speed.

The precision and accuracy of our method was assessed by cross-validation in standard resolution datasets with the widely accepted approach implemented in the available FreeSurfer software. Using a database of consecutive examinations in healthy subjects, the precision of both methods was evaluated using pointwise geometric distances of reconstructed surfaces and differences in cortical thickness. Both methods are similar in terms of the mean absolute error in position and mean absolute error in cortical thickness. However, DELFMAP has a much lower variance than FreeSurfer. In a second study, we evaluated the accuracy of our method by quantifying the intermethod reproducibility of reconstructed cortical surfaces, measured by pointwise geometric distances and differences in cortical thickness measurement between the two methods. Results demonstrate that the accuracy of our method, using FreeSurfer as a reference, is better than half of a mm in terms of both mean absolute error in geometric position and mean absolute error in measured cortical thickness. Group-average analysis of the spatial distribution of geometric and thickness differences between the two methods reveals some surface regions, where one of the two methods has a tendency to systematically over- or undersegment the cortical ribbon, resulting in patterns of small (subvoxel) but measurable differences. Thus, cross-comparison of the two methods allows detection of existing regional patterns in intermethod differences, benefiting the study of accuracy of both approaches and highlighting some potentially problematic areas for further improvement of both methods.

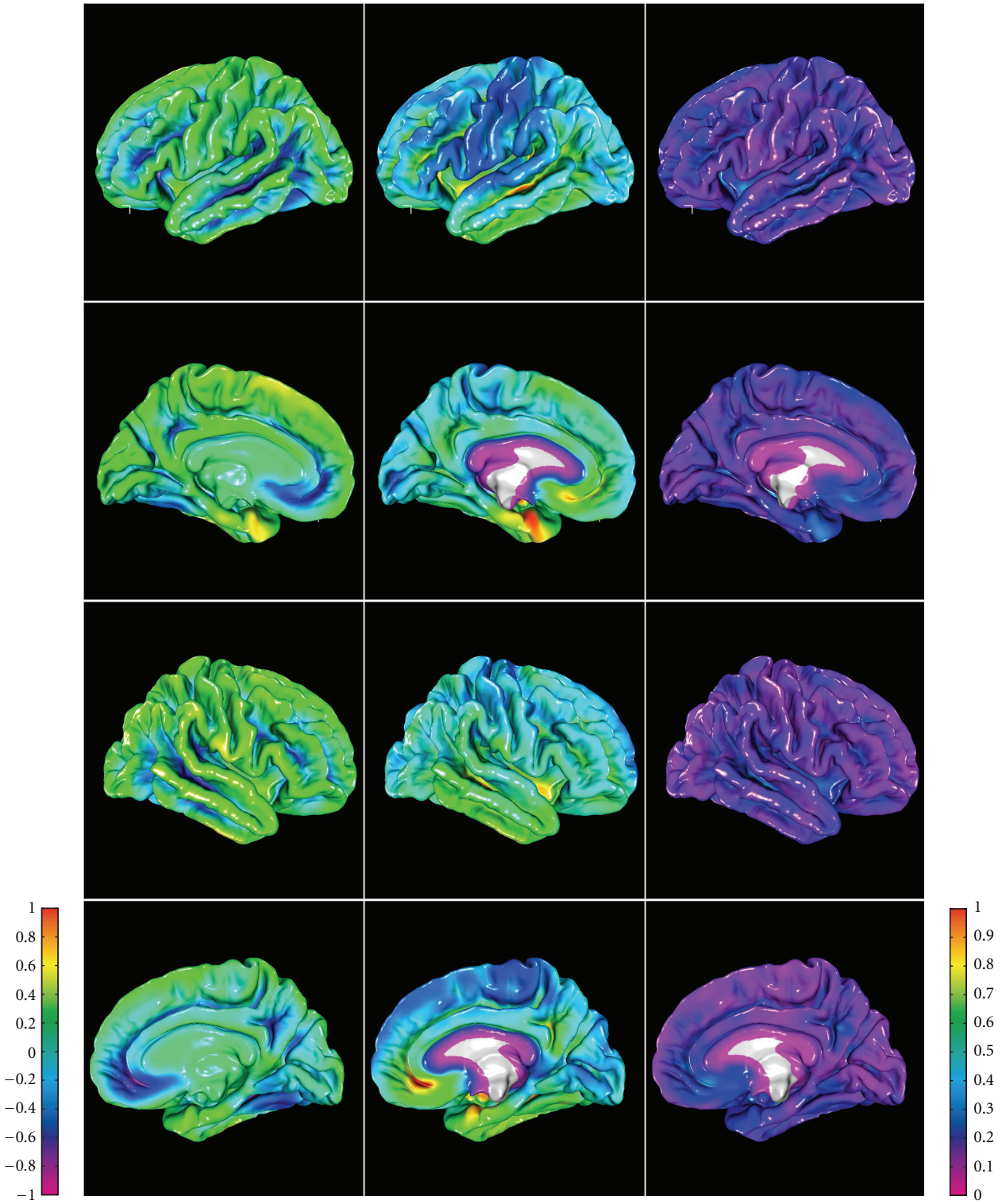


Figure 11: Group-average maps of intermethod (DELFMAP-FreeSurfer) geometric differences in pial surface reconstructions, resampled to FreeSurfer's average template (left column: signed distance mean, colorbar range ±1 mm, negative/positive values mean FreeSurfer' surface is inside/outside of DELFMAP' surface, resp.; middle column: absolute distance mean, colorbar range 0-1 mm; right column: absolute distance stdev., colorbar range 0-1 mm; rows 1–4: lateral/medial surface of left/right hemisphere, resp.).

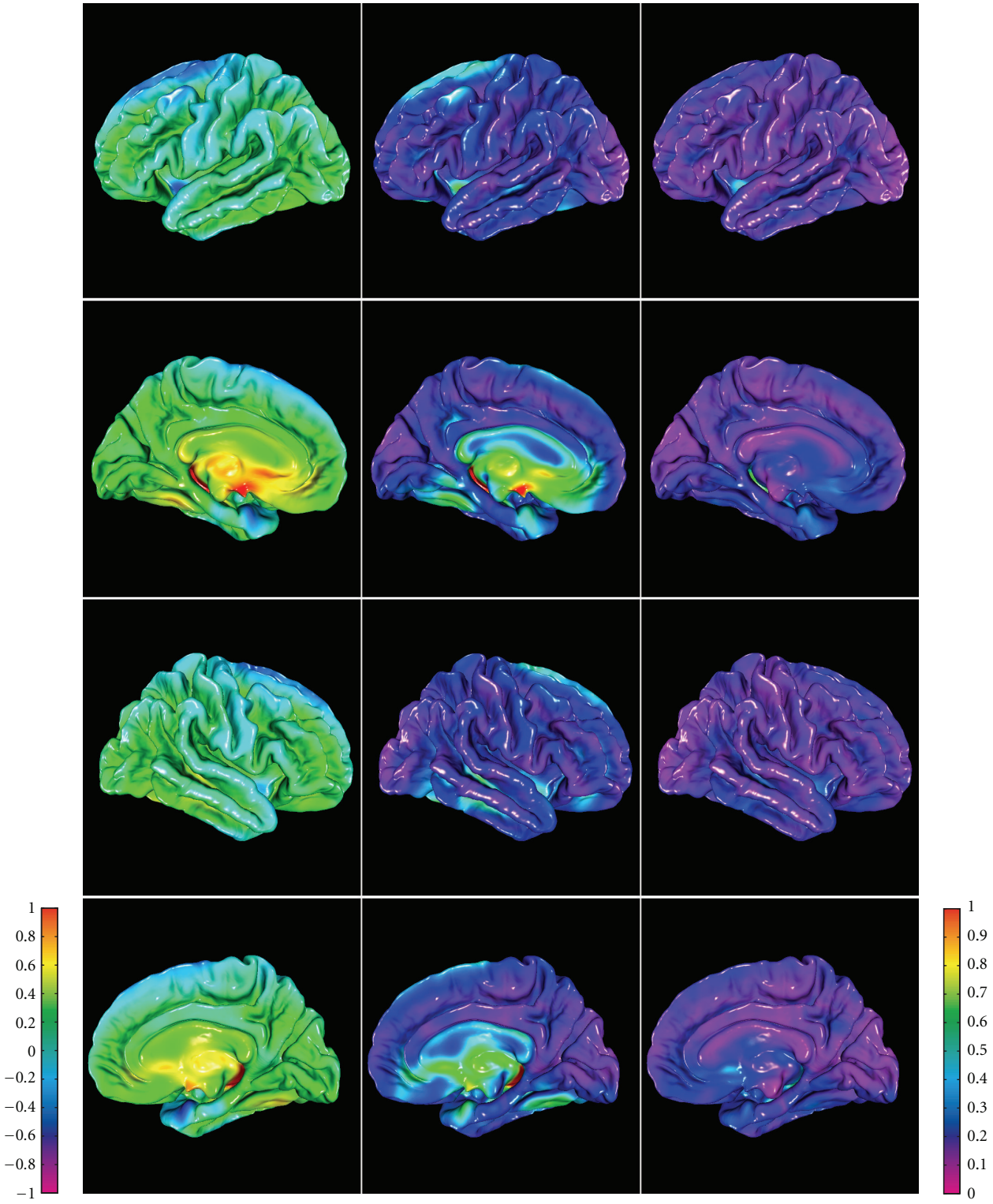


Figure 12: Group-average maps of intermethod (DELFMAP-Freesurfer) cortical thickness differences, resampled to FreeSurfer's average template (left column: signed difference mean, colorbar range ±1 mm, negative/positive values mean thickness measured with DELFMAP is smaller/larger than measured with FreeSurfer, resp.; middle column: absolute difference mean, colorbar range 0-1 mm; right column: absolute difference stdev., colorbar range 0-1 mm; rows 1–4: lateral/medial surface of left/right hemisphere, resp.).

Appendix

Distance Measure between Two Surfaces

The orthogonal projection method [31] was adapted to define a measure of geometric distance between two meshes that was used throughout our validation study. We note that a similar approach was proposed in Tosun et al. [36] for accuracy and precision analysis of cortical surface reconstructions. The signed distance between two triangulated meshes $\mathcal{M}_1 = \{\mathcal{V}_1, \mathcal{F}_1\}_{N_1}$ and $\mathcal{M}_2 = \{\mathcal{V}_2, \mathcal{F}_2\}_{N_2}$ was measured as:

$$\mathcal{D}_{12} = \left\{ d_{12,i} = \left(\vec{p}_{12,i} \cdot \vec{n}_{1,i}\right) \left\| \vec{p}_{12,i} \right\| : \vec{p}_{12,i} = \vec{v}_{1,i} - \mathcal{P}_{\mathcal{F}_2}(\vec{v}_{1,i}) \right\}_{N_1}, \tag{A.1}$$

where $\mathcal{P}_{\mathcal{F}_2}(\vec{v}_{1,i})$ is the closest orthogonal projection operator projecting a vertex $\vec{v}_{1,i}$ from the first mesh onto one of the triangles $\mathcal{F}_2$ in the second mesh, along the normal to that triangle (Figure 8(b)). The sign of the distance measure is determined by the innerproduct of the projection difference vector $\vec{p}_{12,i}$ with the first surface outward normal $\vec{n}_{1,i}$ at vertex $\vec{v}_{1,i}$; thus, a positive/negative value signifies that the second surface is outside/inside of the first surface, respectively. For the signed distance, mean and standard deviation (stdev) are computed on $d_{12,i}$ values. We define the absolute distance measure as:

$$\mathcal{A}_{12} = \left\{ a_{12,i} = \left\| \vec{p}_{12,i} \right\| : (a_{12,i} = |d_{12,i}|) \right\}_{N_1}. \tag{A.2}$$

In this notation, the cortical thickness measure of Kruggel and von Cramon [31] is defined as the absolute distance from GM to WM mesh $\mathcal{A}_{\mathrm{GW}}$; this orthogonal projection measure should not be confused with the distance along surface normal [37], which was shown to be less reliable compared to other distance measures [38]. For the absolute distance, two-way mean and standard deviation are computed (see also [36]) as:

$$\begin{aligned} \mathrm{AD}_{\mathrm{mean}} &= \frac{1}{N_1 + N_2} \left(\sum_{i=1}^{N_1} a_{12,i} + \sum_{j=1}^{N_2} a_{21,j} \right), \\ \mathrm{AD}_{\mathrm{stdev}} &= \left(\frac{1}{N_1 + N_2} \left(\sum_{i=1}^{N_1} a_{12,i}^2 + \sum_{j=1}^{N_2} a_{21,j}^2 \right) - \mathrm{AD}_{\mathrm{mean}}^2 \right)^{1/2}. \end{aligned} \tag{A.3}$$

Acknowledgments

The authors gratefully acknowledge the Athinoula A. Martinos Center for Biomedical Imaging for providing FreeSurfer software and the Open Access Series of Imaging Studies (OASIS) project for providing MRI data sets, which were used for the validation of our new method. The OASIS project is made available by Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN). The authors would like to thank Professors T. Arendt and M. K. Brückner (Paul-Flechsig-Institut für Hirnforschung, Leipzig) for providing specimens for exvivo imaging, and Professor C. J. Wiggins (MPI für Kognitions- und Neurowissenschaften, Leipzig) for high-resolution MRI acquisition, under project C15 supported by a grant from the Interdisziplinäres Zentrum für Klinische Forschung (IZKF), University of Leipzig.

References

[1] X. Han, D. L. Pham, D. Tosun, M. E. Rettmann, C. Xu, and J. L. Prince, "CRUISE: cortical reconstruction using implicit surface evolution," *NeuroImage*, vol. 23, no. 3, pp. 997–1012, 2004.

[2] J. S. Kim, V. Singh, J. K. Lee et al., "Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification," *NeuroImage*, vol. 27, no. 1, pp. 210–221, 2005.

[3] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: i. segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, 1999.

[4] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis: II. Inflation, flattening, and a surface-based coordinate system," *NeuroImage*, vol. 9, no. 2, pp. 195–207, 1999.

[5] B. Fischl, A. Liu, and A. Dale, "Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 70–80, 2001.

[6] B. Fischl and A. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 20, pp. 11044–11049, 2000.

[7] C. Xu, D. L. Pham, M. E. Rettmann, D. N. Yu, and J. L. Prince, "Reconstruction of the human cerebral cortex from magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 6, pp. 467–480, 1999.

[8] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 359–369, 1998.

[9] X. Zeng, L. Staib, R. Schultz, and J. Duncan, "Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 927–937, 1999.

[10] J. Gomes and O. Faugeras, "Level sets and distance functions," in *Proceedings of the European Conference on Computer Vision (ECCV '00)*, D. Vernon, Ed., vol. 1842 of *Lecture Notes in Computer Science*, pp. 588–602, Springer, Heidelberg, Germany, 2000.

[11] X. Han, C. Xu, D. Tosun, and J. L. Prince, "Cortical surface reconstruction using a topology preserving geometric deformable model," in *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '01)*, pp. 213–220, December 2001.

[12] X. Han, C. Xu, and J. L. Prince, "A topology preserving level set method for geometric deformable models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 755–768, 2003.

[13] S. Osechinskiy and F. Kruggel, "PDE-based reconstruction of the cerebral cortex from MR images," in *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '10)*, pp. 4278–4283, September 2010.

[14] S. Osechinskiy and F. Kruggel, "Identifying intrasulcal medial surfaces for anatomically consistent reconstruction of

the cerebral cortex," in *Proceedings of the Annual Conference of the Medical Imaging (SPIE MI '11)*, pp. 1–8, 2011.

[15] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[16] S. Osher and R. P. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, New York, NY, USA, 2002.

[17] F. Yang and F. Kruggel, "Automatic segmentation of human brain sulci," *Medical Image Analysis*, vol. 12, no. 4, pp. 442–451, 2008.

[18] F. Kruggel, M. K. Brückner, T. Arendt, C. J. Wiggins, and D. Y. von Cramon, "Analyzing the neocortical fine-structure," *Medical Image Analysis*, vol. 7, no. 3, pp. 251–264, 2003.

[19] N. Kriegeskorte and R. Goebel, "An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes," *NeuroImage*, vol. 14, no. 2, pp. 329–346, 2001.

[20] D. Pham and J. L. Prince, "Adaptive fuzzy segmentation of magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 9, pp. 737–752, 1999.

[21] G. Gerig, O. Kubler, R. Kikinis, and F. Jolesz, "Nonlinear anisotropic filtering of MRI data," *IEEE Transactions on Medical Imaging*, vol. 11, no. 2, pp. 221–232, 1992.

[22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 2002.

[23] A. Yezzi and J. L. Prince, "A PDE approach for measuring tissue thickness," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. I87–I92, December 2001.

[24] K. R. Rocha, A. J. Yezzi, and J. L. Prince, "A hybrid Eulerian-Lagrangian approach for thickness, correspondence, and gridding of annular tissues," *IEEE Transactions on Image Processing*, vol. 16, no. 3, pp. 636–648, 2007.

[25] D. Adalsteinsson and J. A. Sethian, "A fast level set method for propagating interfaces," *Journal of Computational Physics*, vol. 118, no. 2, pp. 269–277, 1995.

[26] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 4, pp. 1591–1595, 1996.

[27] G. Bertrand, "A boolean characterization of three-dimensional simple points," *Pattern Recognition Letters*, vol. 17, no. 2, pp. 115–124, 1996.

[28] D. Collins, A. Zijdenbos, V. Kollokian et al., "Design and construction of a realistic digital brain phantom," *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 463–468, 1998.

[29] P. L. Bazin, J. L. Cuzzocreo, M. A. Yassa et al., "Volumetric neuroimage analysis extensions for the MIPAV software package," *Journal of Neuroscience Methods*, vol. 165, no. 1, pp. 111–121, 2007.

[30] P. S. Heckbert and M. Garland, "Optimal triangulation and quadric-based surface simplification," *Computational Geometry*, vol. 14, no. 1–3, pp. 49–65, 1999.

[31] F. Kruggel and D. Y. von Cramon, "Measuring the cortical thickness," in *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '00)*, pp. 154–161, June 2000.

[32] M. Jenkinson and S. M. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[33] X. Han, J. Jovicich, D. Salat et al., "Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer," *NeuroImage*, vol. 32, no. 1, pp. 180–194, 2006.

[34] B. Fischl, M. I. Sereno, R. Tootell, and A. M. Dale, "High-resolution intersubject averaging and a coordinate system for the cortical surface," *Human Brain Mapping*, vol. 8, no. 4, pp. 272–284, 1999.

[35] A. Zaharescu, E. Boyer, and R. Horaud, "TransforMesh: a topology-adaptive mesh-based approach to surface evolution," in *Proceedings of the 8th Asian Conference on Computer Vision (ACCV '07)*, pp. 166–175, Springer, Heidelberg, Germany, 2007.

[36] D. Tosun, M. E. Rettmann, D. Q. Naiman, S. M. Resnick, M. A. Kraut, and J. L. Prince, "Cortical reconstruction using implicit surface evolution: accuracy and precision analysis," *NeuroImage*, vol. 29, no. 3, pp. 838–852, 2006.

[37] D. MacDonald, N. Kabani, D. Avis, and A. C. Evans, "Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI," *NeuroImage*, vol. 12, no. 3, pp. 340–356, 2000.

[38] J. P. Lerch and A. C. Evans, "Cortical thickness analysis examined through power analysis and a population simulation," *NeuroImage*, vol. 24, no. 1, pp. 163–173, 2005.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2012, Article ID 574971, 14 pages
doi:10.1155/2012/574971

Research Article

Extending Local Canonical Correlation Analysis to Handle General Linear Contrasts for fMRI Data

Mingwu Jin,[1,2] Rajesh Nandy,[3] Tim Curran,[4] and Dietmar Cordes[2]

[1] *Department of Physics, University of Texas at Arlington, Arlington, TX 76019, USA*
[2] *Department of Radiology, School of Medicine, University of Colorado Denver, Aurora, CO 80045, USA*
[3] *Departments of Biostatistics and Psychology, UCLA, Los Angeles, CA 90095, USA*
[4] *Department of Psychology and Neuroscience, University of Colorado at Boulder, Boulder, CO 80309, USA*

Correspondence should be addressed to Mingwu Jin, mingwu@uta.edu

Received 1 July 2011; Revised 25 September 2011; Accepted 28 September 2011

Academic Editor: Weihong Guo

Local canonical correlation analysis (CCA) is a multivariate method that has been proposed to more accurately determine activation patterns in fMRI data. In its conventional formulation, CCA has several drawbacks that limit its usefulness in fMRI. A major drawback is that, unlike the general linear model (GLM), a test of general linear contrasts of the temporal regressors has not been incorporated into the CCA formalism. To overcome this drawback, a novel directional test statistic was derived using the equivalence of multivariate multiple regression (MVMR) and CCA. This extension will allow CCA to be used for inference of general linear contrasts in more complicated fMRI designs without reparameterization of the design matrix and without reestimating the CCA solutions for each particular contrast of interest. With the proper constraints on the spatial coefficients of CCA, this test statistic can yield a more powerful test on the inference of evoked brain regional activations from noisy fMRI data than the conventional t-test in the GLM. The quantitative results from simulated and pseudoreal data and activation maps from fMRI data were used to demonstrate the advantage of this novel test statistic.

1. Introduction

The General Linear Model (GLM) is a widely used mass univariate analysis method to determine brain activations in functional magnetic resonance imaging (fMRI) because of its simplicity in both estimation and inference and its greater sensitivity to regional effects than global multivariate analyses [1]. The least-squares (LS) solution of the GLM is the minimum variance unbiased (MVU) estimator when Gaussian white noise assumption is satisfied, otherwise the weighted LS solution (using the inverse of the noise covariance matrix) becomes the best linear unbiased estimator (BLUE) [2]. The estimated parameters and their variances are used to construct various contrast statistics, either t or F, to test the null hypothesis of effects of interest. Another popular approach to analyze fMRI time series uses the correlation coefficient [3]. The statistical significance of the correlation coefficient is equivalent to a t-statistic testing for a regression on one single regressor [4]. The correlation coefficient is more restricted in assessing the significance of regional effects than the t-test in fMRI data analysis because the correlation coefficient does not allow more than one regressor to be included for a direct calculation. It is known, however, that the partial correlation coefficient is also equivalent to a t-test and thus could potentially be used instead. However, each contrast of interest need be constructed and the residuals, after removing effects of no interest, have to be calculated for each contrast. This process is generally less computationally efficient than the t-test used in the GLM.

While univariate (single voxel) analysis is extensively applied in fMRI, and temporal correlations are the focus of most investigations, only a few applications investigate the spatial dependence of fMRI data. Univariate analysis deals only with a uniform nonlocal spatial approach and uses fixed isotropic spatial Gaussian smoothing routinely to achieve more homogeneous regions of activation and to control the family-wise error parametrically, based on

the theory of random fields [5, 6]. These methods do not utilize local spatial information in fMRI data, and fixed spatial smoothing causes unnecessary blurring of the edges of activation. More severely, if the fixed isotropic filter kernel is larger than the activated area, it could potentially miss the detection of activated regions. Small focal regions of low contrast-to-noise ratios are rather common in episodic memory paradigms where the task is to detect activation in the medial temporal lobes (hippocampus and parahippocampus). Therefore, fixed Gaussian spatial smoothing can potentially result in missing important (but subtle) focal activations. This is especially troublesome for high-resolution fMRI data where the intrinsic point spread function of the imaging sequence is not much larger than the dimension of a voxel and there are sharp boundaries between grey matter and surrounding cerebrospinal fluid (CSF) and blood vessels (see, e.g., [7]).

A more effective method than fixed Gaussian spatial smoothing uses locally adaptive spatial filter kernels. Using the spatial dependence of fMRI data, local multivariate methods such as canonical correlation analysis (CCA) [8] and its variants [9–13] have the ability to significantly increase the detection power of fMRI activations. However, there are several drawbacks that prevent CCA methods from being widely used in fMRI analysis. First, the original unconstrained CCA method [8] increases the number of false positives due to more freedom in finding favorable linear combinations with nonactive voxel time series leading to a decrease in specificity. This drawback can be addressed by either enforcing some constraints on the spatial coefficients [10, 12, 13] or adaptively assigning the canonical correlation to the most significant voxel [11]. Second, these modified CCA methods [10, 11, 13] usually require much more computation time than the GLM and the unconstrained CCA method. Jin et al. [12] proposed a region-growing strategy to solve the constrained CCA (cCCA) problem in a much faster fashion than the traditional branch-and-bound method [10, 13, 14]. Third, in the form of previous implementations, CCA applications in fMRI data analysis were very limited because test statistics used were based on the significance of the maximum canonical correlation coefficient, thus limiting the analysis to a simple model accommodating only one temporal regressor (i.e., on-off experimental design). This drawback prevents researchers from using CCA for more complicated paradigms with multiple explanatory variables and nuisance covariates in fMRI. Though reparameterization based on the linear contrast of interest can provide a solution for this drawback [15, 24], the computational cost is high because, for each different reparameterization, the constrained CCA problem needs to be solved. The major goal of this research is to find a suitable test statistic for CCA that allows the testing of general linear contrasts and that is also fast.

In this paper, we first establish the connection between the multivariate multiple-regression (MVMR) model and CCA. Although this is not totally new in statistics, we found that there is lack of awareness for the development of CCA methods in the fMRI data analysis community. By treating the estimated spatial filter kernel of constrained CCA as a linear transformation of the original MVMR model, we further derive a novel univariate test statistic similar to a t-statistic based on general hypothesis tests of the MVMR model. This extension will allow CCA to be used for inference of general linear contrasts in more complicated fMRI designs without solving the constrained CCA problem for each particular contrast of interest.

In the following, we start from the MVMR model and its hypothesis test for general linear contrasts under a linear transformation of the original model. Then, the simultaneous estimation of spatial and temporal parameters using the LS rule in the MVMR model is derived and proved to be the same as the CCA solution. By treating the adaptive spatial smoothing as a linear transformation of the original MVMR model, a novel directional statistic for CCA similar to a t-statistic can be derived to allow for testing of general linear contrasts. Using receiver operating characteristic (ROC) techniques [16–18] on pseudoreal fMRI data [11, 19–21], we quantitatively compare the sensitivity and specificity of the proposed novel CCA statistic with the t-statistic of the GLM without and with fixed Gaussian spatial smoothing. We also apply a nonparametric approach [22] to estimate the family-wise error rate for all methods using resampled resting-state data and show the activation maps for real fMRI data for a simple visual cortex activation paradigm and also for a more complicated memory paradigm.

2. Theory

2.1. The MVMR Model. Considering a group of K local neighborhood voxels, the MVMR model can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{1}$$

where $\mathbf{X}$ is fixed (i.e., the $n \times p$ design matrix), $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K)$ is the matrix containing K neighboring voxels, $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$ is the parameter matrix to be estimated, and $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_K)$ is the error matrix. Without of loss of generality, $\mathbf{X}$ and $\mathbf{Y}$ are column centered and there is no constant column in $\mathbf{X}$. When the error matrix satisfies (i) $E(\mathbf{E}) = \mathbf{0}$, (ii) $\operatorname{cov}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma}$ for $i = 1, \ldots, n$, and (iii) $\operatorname{cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) = \mathbf{0}$ for $i \neq j$, the LS solution of the model (1) is equivalent to the BLUE, which is just the matrix form of the univariate GLM estimator leading to equivalent solutions, but a multivariate test need be adopted. Note that conditions (i)–(iii) may not be true for fMRI data, but may be reasonably satisfied using temporal whitening.

The hypothesis tests in the MVMR model can be conducted using the error matrix and the hypothesis matrix for any estimable general linear contrast matrix $\mathbf{C}'$. For a linear transformation of the original MVMR model, say $\mathbf{M}$, Wilks' Λ and other test statistics (e.g., Roy's largest root) can be used for testing the null hypothesis $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{O}$ [23]. In addition to the fixed linear transformation of the MVMR model, we will introduce estimation of the spatial filter kernel (leading to an adaptive smoothing) and treat it as a spatially variable linear transformation in the following development. This linear transformation can be estimated from the data using CCA. Utilizing the spatial and temporal

coefficients from CCA and the hypothesis test on the linear transformation of the MVMR model, a directional (one-sided) statistical test for CCA can be derived that is similar to a t statistic in the GLM. This novel statistic allows CCA to test hypothesis on general linear contrasts of an fMRI design without reparameterization of the design matrix and without reestimation of the CCA solutions for each particular contrast of interest.

2.2. Adaptive Filtering through Canonical Correlation Analysis (CCA). To increase detection power of weak activations, local spatial filtering is usually applied to decrease the noise variance. Let $\boldsymbol{\alpha}$ be the vector containing the spatial filtering coefficients, then multiplication of both sides of the MVMR model of (1) with $\boldsymbol{\alpha}$ gives

$$\mathbf{Y}\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\beta} \equiv \mathbf{B}\boldsymbol{\alpha}$, $\boldsymbol{\varepsilon} \equiv \mathbf{E}\boldsymbol{\alpha}$, and multiplication by $\boldsymbol{\alpha}$ defines a linear transformation of the original MVMR model in (1).

When both $\mathbf{Y}$ and $\boldsymbol{\alpha}$ are fixed and treated as known, such as in conventional fixed Gaussian smoothing, $\boldsymbol{\beta}$ can be easily estimated by linear regression as

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\boldsymbol{\alpha}. \tag{3}$$

Given a general linear contrast $\mathbf{C}'$, the null hypothesis of $\mathbf{C}'\mathbf{B}\boldsymbol{\alpha} = \mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$ can be tested using Wilks' Λ likelihood ratio test (assuming independent identical normal distribution of noise both spatially and temporally) by

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}, \tag{4}$$

where the error matrix is $\mathbf{E} = (\mathbf{Y}\boldsymbol{\alpha} - \mathbf{X}\widetilde{\boldsymbol{\beta}})'(\mathbf{Y}\boldsymbol{\alpha} - \mathbf{X}\widetilde{\boldsymbol{\beta}})$ and the hypothesis matrix is $\mathbf{H} = (\mathbf{C}'\widetilde{\boldsymbol{\beta}})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\widetilde{\boldsymbol{\beta}})$. Note that both matrices reduce to a scalar due to the linear transformation of the original MVMR model by vectors $\boldsymbol{\alpha}$.

A fix-sized and isotropic smoothing kernel, such as a Gaussian kernel, is not optimal, especially for weak and small activations. Our goal is to increase detection power by pooling the neighboring voxels with similar activation pattern and by determining the spatial weights $\boldsymbol{\alpha}$ from the data as well. This adaptive smoothing can be achieved by minimizing the square of fitting error (i.e., LS) for the model in (2), which leads to the equivalent solution in CCA.

Assuming that the optimal configuration of $\mathbf{Y}$ is known (please see [10, 12] for how to find this configuration), the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be estimated by LS:

$$\left(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}\right) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{Y}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|^2. \tag{5}$$

There is a trivial solution for (5): $\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\beta}} = 0$, which can be avoided by enforcing some normalization condition, such as $\widetilde{\boldsymbol{\alpha}}'\mathbf{S}_{yy}\widetilde{\boldsymbol{\alpha}} = 1$ or $\widetilde{\boldsymbol{\alpha}}'\widetilde{\boldsymbol{\alpha}} = 1$. Taking the partial derivative of the square of fitting error over $\boldsymbol{\alpha}$, we get

$$\frac{\partial\|\mathbf{Y}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|^2}{\partial\boldsymbol{\alpha}} = 2(\mathbf{Y}'\mathbf{Y}\boldsymbol{\alpha} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta}). \tag{6}$$

The solution $\widetilde{\boldsymbol{\alpha}}$ requires (6) equal to zero so that

$$\boldsymbol{\alpha} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}. \tag{7}$$

Meanwhile, the relationship in (3) is still valid. Therefore, only one vector needs to be estimated and the other can be determined by (3) or (7). Substituting (3) in (7), we get

$$\begin{aligned}\boldsymbol{\alpha} &= (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\boldsymbol{\alpha} \\ &= \mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\boldsymbol{\alpha},\end{aligned} \tag{8}$$

where the sample covariance matrices are $\mathbf{S}_{yy} = (1/(n-1))\mathbf{Y}'\mathbf{Y}$, $\mathbf{S}_{xx} = (1/(n-1))\mathbf{X}'\mathbf{X}$, and $\mathbf{S}_{xy} = \mathbf{S}_{yx}' = (1/(n-1))\mathbf{X}'\mathbf{Y}$. This is an eigenvalue problem for $\boldsymbol{\alpha}$ with eigenvalue 1, whose solution may not exist because the eigenvalue of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ is not necessarily identical to 1. Thus, a conventional method to solve (8) is to write it as an LS problem by

$$\widetilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\|\boldsymbol{\alpha} - \mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\boldsymbol{\alpha}\right\|^2. \tag{9}$$

Given that $\boldsymbol{\alpha} \neq 0$ by enforcing the normalization condition mentioned previously, the expression $\|\boldsymbol{\alpha} - \mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\boldsymbol{\alpha}\|^2$ can be minimized if $\widetilde{\boldsymbol{\alpha}}$ is the eigenvector of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ which has the eigenvalue λ_m closest to 1 (or in other words, the largest eigenvalue of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$ because its upper bound is 1), that is,

$$\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\widetilde{\boldsymbol{\alpha}} = \lambda_m\widetilde{\boldsymbol{\alpha}}. \tag{10}$$

Equation (10) results in the same solution for CCA, where $\lambda_m = r^2$ and r is the maximal canonical correlation. This is not totally unexpected because

$$\begin{aligned}\left(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}\right) &= \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{Y}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\|\frac{\mathbf{Y}\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}'\mathbf{S}_{yy}\boldsymbol{\alpha}}} - \frac{\mathbf{X}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}'\mathbf{S}_{xx}\boldsymbol{\beta}}}\right\|^2 \\ &= \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} (n-1)C_1 \\ &\quad + (n-1)C_2 - 2(n-1)\frac{\boldsymbol{\alpha}'\mathbf{S}_{yx}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}'\mathbf{S}_{yy}\boldsymbol{\alpha}}\sqrt{\boldsymbol{\beta}'\mathbf{S}_{xx}\boldsymbol{\beta}}},\end{aligned} \tag{11}$$

where $C_1 = \boldsymbol{\alpha}'\mathbf{S}_{yy}\boldsymbol{\alpha}$ and $C_2 = \boldsymbol{\beta}'\mathbf{S}_{xx}\boldsymbol{\beta}$ are nonzero constants. Therefore, we can use CCA, which maximizes the third term in the above equation, to find solutions for the model in (2). Once $\widetilde{\boldsymbol{\alpha}}$ has been determined, the temporal coefficients $\widetilde{\boldsymbol{\beta}}$ can be obtained by (3) accordingly.

Normally, we can achieve a desired filtering effect by adding constraints on the components of $\boldsymbol{\alpha}$ in a constrained CCA (cCCA) form. In this work, we constrain all components of $\boldsymbol{\alpha}$ to have the same sign. This constraint not only enforces a smoothing effect, bus also has an optimal solution through searching CCA solutions of the possible configurations of $\mathbf{Y}$ in a prescribed local region to satisfy this constraint [10, 12, 14]. In addition we add a center voxel significance constraint by requiring that the spatial

weight of the center voxel be at least 20% of the maximum spatial weight in each 3×3 neighborhood [12]. Although this introduces some nonlinearity to the optimization [24], in the current implementation, this additional constraint was found empirically to be effective in producing the best performance. A similar approach was used in [10] to increase the spatial specificity.

Generally, we suggest scaling the solution $\tilde{\boldsymbol{\alpha}}$ of cCCA to have a sum of magnitudes to be one. Although this is not required because the scaling factor will be cancelled out in calculating the novel CCA test statistic (refer to (14) in next section), this treatment can keep the error term comparable with GLM methods. A region-growing method [12] allowing a much faster implementation than the traditional branch-and-bound method [10, 14] will be used to obtain $\tilde{\boldsymbol{\alpha}}$.

Several advantages of the current implementation of cCCA over the method proposed in [10] are listed here. (1) In [10], a spatial Gaussian filter was divided into one isotropic central part and three oriented parts. The weights for these parts can be estimated using CCA to achieve anisotropic filtering (steerable spatial filtering). In our method, we search for the optimal voxel combinations and weights in a 3×3 neighborhood because the cortical layer in a typical fMRI scan is less than 5 mm and spans only a couple of voxels. Our smaller filter size can help better define activations leading to higher specificity (2) A rather slow branch-and-bound (BB) method was used in [10], which is not efficient to search optimal combinations for the center voxel in a 3×3 area (see Section 5). Our region-growing method takes 24 s for a 2D slice with 6317 in-brain pixels and is much faster than the BB method (308 s) [12] (3) The statistic used in [10] was the maximum canonical correlation coefficient, which can only be used for simple on-and-off paradigms but not for arbitrary linear functions (contrasts) in complicated paradigms. The new statistic proposed in our work can be applied for complicated paradigms without reestimating for each contrast of interest. Although it would be an interesting followup to compare different CCA methods, such a comparison is beyond the scope of the current paper.

2.3. Novel Directional Test Statistic for CCA. As a simple treatment, the estimated components of $\tilde{\boldsymbol{\alpha}}$ can be used as local spatial filter coefficients to smooth the original data. Then, the same univariate inference as the GLM can be applied to get a statistical map for any general linear contrast. However, this procedure has two drawbacks: (1) the GLM estimation of $\boldsymbol{\beta}$ on the smoothed images adds extra unnecessary computation time; (2) the resulting statistics will be biased because it does not account for the loss of degrees of freedom caused by the size of the spatial filter kernel. For example, a single voxel configuration is more significant than a multiple-voxel configuration having the same value of the test statistic. To overcome these two drawbacks, we derive the test statistic directly from the CCA coefficients $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ and account for the spatial kernel size by changing the degrees of freedom.

Given the general linear contrast $\mathbf{C}'$, the null hypothesis: $H_0 : \mathbf{C}'\mathbf{B}\tilde{\boldsymbol{\alpha}} = \mathbf{C}'\tilde{\boldsymbol{\beta}} = \mathbf{0}$ can be tested by Wilks' Λ in (4), where $\boldsymbol{\alpha}$ in the error matrix is replaced by $\tilde{\boldsymbol{\alpha}}$. In this paper, we are particularly interested in a directional test statistic when the contrast matrix $\mathbf{C}'$ reduces to a vector $\mathbf{c}'$. Thus, the test statistic on $\mathbf{c}'\tilde{\boldsymbol{\beta}}$ reduces to a univariate case with a signed value and can be defined by

$$\Lambda_{\pm} = \operatorname{sign}\left(\mathbf{c}'\tilde{\boldsymbol{\beta}}\right)\Lambda = \operatorname{sign}\left(\mathbf{c}'\tilde{\boldsymbol{\beta}}\right)\frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}, \tag{12}$$

where Λ_+ indicates the positive statistic for values $\mathbf{c}'\tilde{\boldsymbol{\beta}} > 0$ and Λ_- indicates the negative statistic for values $\mathbf{c}'\tilde{\boldsymbol{\beta}} < 0$.

Going one step further, we can define a statistic t_c bearing a similar form as the conventional t-statistic by writing

$$\Lambda_{\pm} = \operatorname{sign}\left(\mathbf{c}'\tilde{\boldsymbol{\beta}}_c\right)\frac{1}{1+t_c^2/\mathrm{DF}}, \tag{13}$$

where $\mathrm{DF} = n - p - K$ specifies the degrees of freedom (DOF) given that the number of observations is n, the number of (nonconstant) regressors is p (linear equations for $\boldsymbol{\beta}$), and the size of voxel configuration in CCA is K (constraints for $\boldsymbol{\alpha}$). As we will discuss next, t_c is not a real t-statistic, but rather using the concept of DOF to account for the voxel configuration size similar to t-statistic. Thus, a non-parametric estimation method [22] is essential to assess its statistical significance. Since the right sides of (12) and (13) are equal, this statistic can be written by using the definition of $\mathbf{E}$ and $\mathbf{H}$ as

$$t_c = \frac{\mathbf{c}'\tilde{\boldsymbol{\beta}}\sqrt{\mathrm{DF}}}{\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}\sqrt{\left(\mathbf{Y}\tilde{\boldsymbol{\alpha}} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)'\left(\mathbf{Y}\tilde{\boldsymbol{\alpha}} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)}}. \tag{14}$$

Note that the voxel configuration size has been accounted for in (14) so that the same $\mathbf{c}'\tilde{\boldsymbol{\beta}}$ values with less voxels become more significant. The new statistic reduces to a traditional t-statistic for the single voxel ($K = 1$) case (when the noise is white and Gaussian distributed) given by

$$t = \frac{\mathbf{c}'\tilde{\boldsymbol{\beta}}\sqrt{n-p-1}}{\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}\sqrt{\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)}}. \tag{15}$$

Generally, (14) will not follow a t-distribution even under the assumption of independent identical normal distribution of noise in both space and time because of the constrained CCA estimation for $\boldsymbol{\alpha}$. Without spatial correlation in the single voxel case ($K = 1$), (15) can approximate fairly well a t-distribution when prewhitening is applied to decorrelate the temporal serial correlations. Moreover, the spatial correlation of fMRI data will pose a tricky problem for approximating a true t-distribution. To deal with these difficulties, a non-parametric estimation method [22] is adapted to assess the significance of the CCA statistic of (14). The distribution of this novel statistic on null data will be shown to deviate from the true t-distribution in Section 4.

From (14), we can see the advantage of the newly developed test statistic. First, if activations exist at the center voxel and its neighbors, we get a more accurate

estimate of $\tilde{\boldsymbol{\beta}}$ (as shown in simulations in Section 4) by pooling these voxels in the estimation. Second, the error term $(\mathbf{Y}\tilde{\boldsymbol{\alpha}} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{Y}\tilde{\boldsymbol{\alpha}} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ is always smaller than $(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. Therefore, no matter what contrast vector $\mathbf{c}$ is used, t_c has a larger value than the univariate t. It would be expected that t_c values of active voxels increase more than t_c values of voxels in the null state, which will lead to an increased sensitivity. Third, the better model fitting by pooling more voxels is penalized by the degrees of freedom $\mathrm{DF} = n - p - K$. This penalty will cause considerable bias when n is comparable to $p+K$. However, this scenario is not practically meaningful because the length of the fMRI sequence is usually much greater than the sum of the number regressors and the size of the filter kernel (i.e., $n \gg p + K$).

Note that the proposed test statistic may not necessarily be the optimal test for an arbitrary contrast because we only minimize the square of fitting error in (2) that is independent of the contrast [24]. Nevertheless, the new statistic allows us to improve the detection power without reparameterization of the design matrix and without re-estimating each particular contrast of interest as shown in Section 4.

3. Methods

3.1. Imaging Data. Functional MRI (fMRI) was performed at the Brain Imaging Center of the University of Colorado Denver in a 3.0T GE HDx MRI scanner equipped with an 8-channel head coil and parallel imaging technology. Stimulus presentation was done with a rear projection system (AVOTEC, Inc.). Two different paradigms (visual paradigm and memory paradigm) were performed on two and eight healthy adult subjects, respectively, and fMRI data were collected according to local IRB approval. The pulse sequence to collect fMRI data was EPI with the following parameters: ASSET = 2, ramp sampling, TR = 2 sec, TE = 30 ms, FA = 70 deg, FOV = 22 cm × 22 cm, slice thickness = 4 mm, gap = 1 mm, 25 slices, and in-plane resolution 96 × 96. For the visual paradigm we prescribed axial slices and collected 150 volumes, whereas for the memory paradigm we prescribed coronal oblique slices perpendicular to the long axis of the hippocampus and collected 288 volumes. The first 5 volumes were discarded to establish signal equilibrium of the imaging sequence.

To obtain an accurate gray matter mask that has equivalent features of the echo-planar data (same geometry and distortions), we collected for each subject an additional coplanar IR-SE-EPI scan to get inverted T1 contrast with the following parameters: TI = 505 ms, ASSET = 2, ramp sampling, TR = 6 sec, TE = 30 ms, FOV = 22 cm × 22 cm, slice thickness = 4 mm, gap = 1 mm, 25 slices, and in-plane resolution 96 × 96. This imaging sequence yields unique high signals for gray matter so that we can easily threshold them to get accurate gray matter masks. The IR-SE-EPI images were first aligned to the mean EPI images using six-parameter affine transformation and then were thresholded to get gray matter masks. Visual inspection of masks for faithfulness was conducted before calculating the activation voxels in gray matter.

Furthermore, we acquired a coplanar standard high-resolution T2-weighted anatomical scan (FOV 22 cm, resolution 256 × 256, TR 3000 ms, TE 85 ms, NEX 2, slice thickness 4 mm, gap 1 mm). The mean EPI functional image of each individual was coregistered to its corresponding T2 image, and the same transformation was applied on all functional images. The resultant activation map shown in Section 4 was overlaid on the individual T2 image.

3.1.1. Visual Paradigm. For each subject we acquired two fMRI data sets. The first data set was collected during resting state where the subject tried to relax and refrained from executing any overt task with eyes closed. The second data set was collected while the subject was looking at a flashing checkerboard (10 Hz flashing frequency, duration 2 sec) which alternated with a fixation period of random duration (2 sec to 10 sec, uniformly distributed). During the fixation period a black screen containing in the center a small white cross (about 1 inch in size) was shown and the subject was instructed to focus on this cross. The corresponding design matrix using the canonical hemodynamic response function (HRF) model is shown in Figure 1(a). The left column in this figure represents the regressor for the fixation and the right column represents the regressor for the visual activation.

3.1.2. Memory Paradigm. Also here, we acquired two fMRI data sets for each subject. The first set contained resting-state data, and the second set was acquired while the subject performed a memory task. Behavioral responses were collected during the memory paradigm with button response pads that the subject had in each hand. The memory paradigm started with a fixation period of 16 sec followed by six identical 89 sec long cycles of "5 sec instruction," "21 sec encoding," "5 sec instruction," "11 sec control," "5 sec instruction," and "42 sec recognition". It ended with another fixation period of 16 sec. The short "instruction period" consisted of a single sentence and reminded the subject of the following task to be performed. The "encoding" task consisted of a series of novel pictures, where each picture was displayed for 3 sec, and the subject was instructed to memorize each picture. During the "control" task the subject saw the letters "Y" or "N" which appeared in random order every 100 ms on the display screen. The subject was instructed to press, as fast as possible, the right button when "Y" appears or the left button when "N" appears. The purpose of the "control" task was twofold. First, it served as a distraction task to keep attention away from the just learned pictures. Second, due to its simplicity it did not produce any activation in regions associated with the memory circuit (hippocampal complex, posterior cingulate cortex, precuneus, and fusiform gyrus). During the "recognition" task the subject saw a series of pictures where half of the pictures were novel and the other half of the pictures were identical to the pictures from the previous "encoding period." The arrangement of these pictures was random. Each picture was displayed on the screen for 3 sec. The subject was instructed to press the right button if the picture was seen before in the previous

encoding period and to press the left button if the picture was identified to be novel and not seen in the previous encoding period. The design matrix using the canonical HRF model is shown in Figure 1(b). The four conditions of "instruction," "encoding," "recognition," and "control" are denoted as "I," "E," "R," and "C," respectively.

Due to the complexity of the memory paradigm, all subjects were trained on a computer in a quiet room with the paradigm using a different set of images before fMRI scanning. The stimuli presentations were programmed in EPRIME and all button presses were recorded.

3.2. Preprocessing. All data were preprocessed in SPM5 using realignment to correct for motion artifacts, slice timing correction to correct for differences in image acquisition time between slices, and high-pass filtering using $T = 150$ sec to remove low-frequency components and signal drifts. The classic two-gamma HRF was used to construct the design matrix. In the next section, we give examples for the contrasts "Visual minus Fixation" (denoted as "V-F") for visual data and "Encoding minus Control" (denoted as "E-C") for memory data and ignore other possible contrasts of interest.

3.3. Methods of Data Analysis. Three methods were investigated using the statistics defined in (14) and (15). The first two using (15) are (i) the GLM without smoothing, denoted as "GLM-NS" and (ii) Gaussian smoothing followed by the GLM, denoted as "GLM-GS." The third one is cCCA with the region-growing method [12] using (14), denoted as "cCCA-RG." The full width at half maximum (FWHM) of Gaussian smoothing in the GLM was chosen as 2.24 pixels. This number is not only falling in the generally recommended smoothing size (2-3 times of the spatial resolution) in fMRI data analysis, but is also equal to the average size of all possible 256 configurations within a 3×3 pixel area that includes the center pixel [24].

3.4. Construction of Simulated and Pseudoreal Data. In demonstrating the estimation and detection performance of different methods, real fMRI data, where the subject performed a certain paradigm, are difficult to use since the ground truth about the activated regions is unknown. To draw any firm conclusions about the performance of a method, it is better to use simulated/pseudoreal data, where the important parameters are known and can be tested for and the data features (especially the noise characteristics) are similar to real fMRI data [11, 17]. In this work, we always use the resampled resting-state fMRI data as the noise background to preserve the noise characteristics of real data and superimpose either artificial activations or activations extracted from real activation fMRI data. Even though the difference between simulated/pseudoreal data and real data cannot be avoided, the evaluation provides a ranking of the estimation and detection performance of difference methods that is unlikely to change for real data.

To quantitatively determine the performance of different methods, we constructed both simulated and pseudoreal data by defining

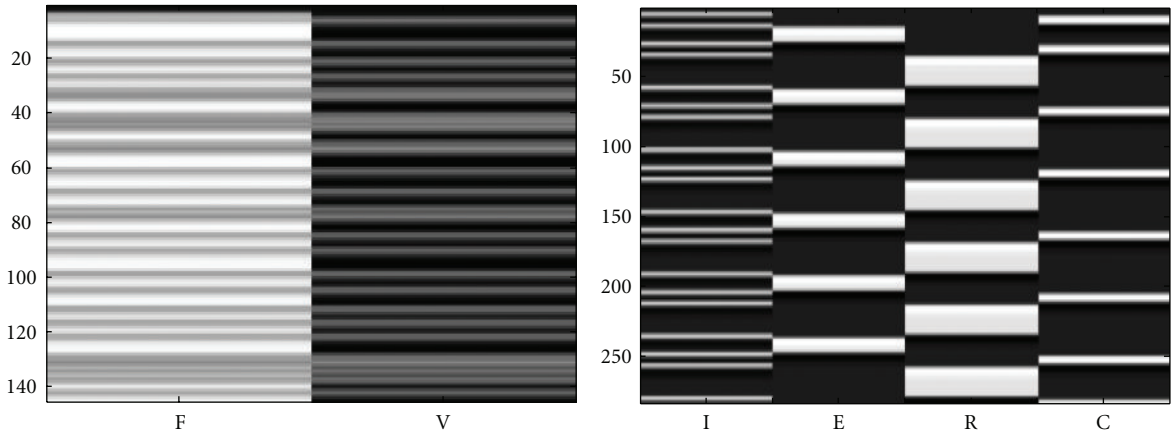
$$\mathbf{x} = \begin{cases} (1-f)\mathbf{x}_{\text{act}} + f\mathbf{x}_{\text{null}}, & \mathbf{x} \in \text{active set}, \\ \mathbf{x}_{\text{null}}, & \text{otherwise}. \end{cases} \tag{16}$$

In this equation $\mathbf{x}$ is the vector representing the time series of a voxel with activation contribution $\mathbf{x}_{\text{act}}$ and noise contribution $\mathbf{x}_{\text{null}}$. The noise fraction parameter f is a scalar number to adjust the noise level in the data vector $\mathbf{x}$ given that $\mathbf{x}_{\text{act}}$ and $\mathbf{x}_{\text{null}}$ have the same power. For null data $\mathbf{x}_{\text{null}}$, Fourier resampling [25] of resting-state fMRI data was used to randomize the phase of each time series without destroying the inherent temporal and spatial correlations in the data. Note that there are other resampling methods for fMRI data, such as wavelet resampling [26, 27] and whitening resampling [28–31], and some comparisons have been made based on different criteria [27, 31–33]. Compared to whitening resampling, both Fourier and wavelet resamplings do not assume a specific model (such as AR(p) or ARMA(p,q)) to do model fitting and are thus more general since different voxels may follow different whitening models. To avoid complicating our simulation, we chose Fourier resampling with the same phase permutation for all time series to preserve the spatial correlations of resting-state fMRI data. This resampling method is least computationally demanding and was demonstrated to have a similar ROC performance to that of wavelet resampling [33].

To define different spatial patterns of activations for simulated data, 100000 randomly shaped activations within a 3×3 grid of pixels having a size of 2 to 9 pixels were generated. The center pixel was always assigned to be active. The corresponding time courses for the activated pixels $\mathbf{x}_{\text{act}}$ were simulated to be linear combinations of the 4 random temporal regressors with random amplitudes β_1, β_2, β_3, and β_4 uniformly distributed in $[0, 1]$. Different levels of noise introduced by resampled 3×3 patches of resting-state fMRI were used for $\mathbf{x}_{\text{null}}$. Both $\mathbf{x}_{\text{null}}$ and $\mathbf{x}_{\text{act}}$ were normalized to have unit variance before the mixture.

To quantitatively evaluate both sensitivity and specificity of the novel CCA test statistic of (14) in comparison with a GLM-based t-test in a more realistic setting using ROC techniques [16–18], we constructed pseudoreal data [11] using a combination of activation data and resting-state data. First, GLM-NS was applied on the activation data. Next, the groups of highly active voxels using an unadjusted P value threshold of 10^{-8} for the t-maps of V-F in visual data and of E-C in memory data were labeled as active, that is, $\mathbf{x}_{\text{act}}$. Finally, we generated, by Fourier resampling of resting-state data, the null data $\mathbf{x}_{\text{null}}$ and constructed the final pseudoreal data according to (16).

To find the proper noise fraction parameter f in (16), we applied GLM-NS on pseudoreal data for $f \in (0, 1)$ with step size 0.01. The median of corresponding t-values of activations was compared with the median of t-values with significance level in $[10^{-8}, 10^{-3}]$ by applying GLM-NS on real (non simulated) fMRI activation data. We plotted t-values of contrasts V-F and E-C in Figures 2(a) and 2(b),



(a) Visual (b) Memory

Figure 1: Design matrices for visual (a) and memory (b) paradigms. From left to right, the regressors are fixation (F) and visual activation (V) for the visual paradigm (a), and instruction (I), encoding (E), recognition (R), and control (C) for the memory paradigm (b). The SPM-type two-gamma function was used as the HRF. Note that with centered data, (a) can be modeled by a single centered activation column and (b) can be modeled by three centered columns. The redundant presentation was used to show all experimental conditions.

respectively. As can be seen, $f = 0.6$ is a value that two medians match. Therefore, we picked two values for f : 0.55 representing the low noise case and 0.65 representing the high noise case. By normalizing the peak variances of noise and signal to be the same, these two values for f correspond to a peak signal-to-noise ratio of 67% and 29%, respectively. The logic of choosing these significance levels for determining a proper f is the following. Voxels with significance level $P < 10^{-8}$ are signals with very high SNR (which are almost certainly true activations), those with significance level in the interval $[10^{-8}, 10^{-3}]$ are the majority of signals with medium or low SNR and of interests of detection (whose median of the t-statistic was used to find a matching f), and those with significance level $P > 10^{-3}$ are dominated by noise and are therefore ignored.

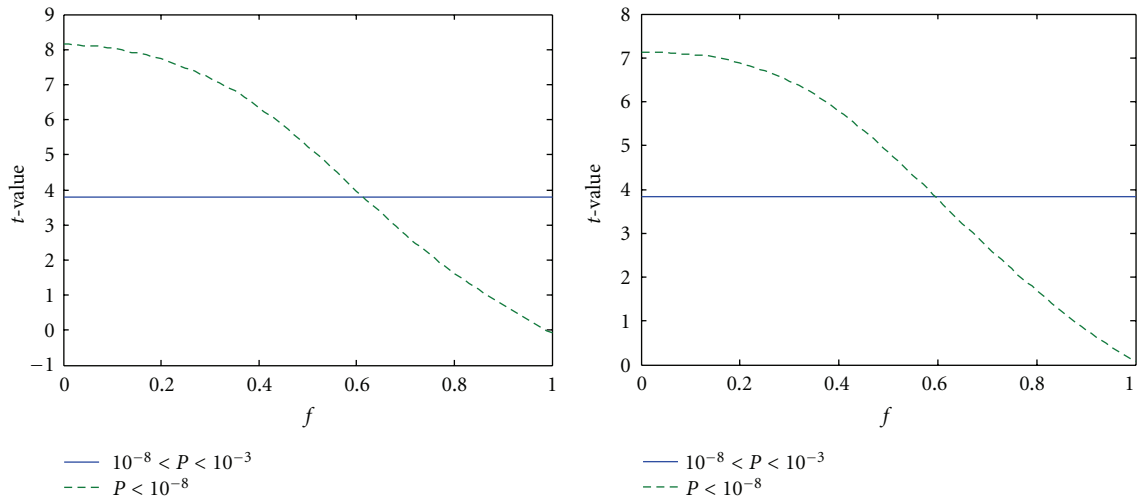
The advantage of constructing pseudoreal data using real activation data and resampled resting-state data is that the spatial and temporal correlations of both the activations and the noise are similar to real data and the locations of active and nonactive voxels are known by construction. This type of simulation then allows conventional ROC techniques to be applied.

3.5. Determination of Proper P-Value. To compare different test statistics using real visual and memory activation data, it is necessary to get the proper P-values for the corresponding t- or novel CCA statistic that is adjusted for multiple comparisons. In this work, we used a non-parametric technique [22]. A non-parametric technique is suitable for a reliable comparison between different analysis methods because the parametric distribution of the CCA statistic is intractable due to the data-adaptive spatial filtering kernel. In the following we outline how the family-wise error rate (FWE) is being calculated using Fourier resampled resting-state data using bootstrapping of the order statistics. For more details, please see the publication [22].

The multiple comparison problem is relevant when we have a family of hypotheses $\{H_\omega : \omega \in \Omega\}$ at voxel ω. Let the test statistics at voxel ω be denoted by Y_ω. Then FWE is determined by the maximum statistic $(\max Y_\omega)$, and for any threshold u, we can calculate the P-value that automatically adjusts for multiple comparisons. To estimate the null distribution of $\{\max Y_\omega\}$, we use the bootstrap method applied to the k largest order statistics $\{Y^1, \ldots, Y^k\}$ from Fourier resampled resting-state data. This method is quite general and may be applied to a broad class of test statistics in fMRI. In the present context of CCA, the relevant test statistic is given by (14) or (15). Although it is not strictly necessary, it is preferable to make a transformation of the test statistic using the known (approximate) distribution or the kernel density estimation. We calculate the negative logarithm of the P-value corresponding to the test statistic to obtain our transformed variables. Due to the monotonous nature of the transformation, without loss of generality, we can assume that Y is already transformed. Define $\{d_i = i(Y^i - Y^{i+1}),\ i = 1, \ldots, k\}$ as normalized sample spacings for the k largest order statistics. If the observed samples at the voxels are exponential i.i.d then so are the normalized sample spacings [34]. This is true since the transformed test statistic is an exponential random variable. The k largest order statistics can then be expressed as a linear function of the normalized sample spacings and Y^{k+1} as follows:

$$Y^j = Y^{k+1} + \sum_{i=j}^{k} i^{-1} d_i, \quad j = 1, \ldots, k. \tag{17}$$

Since $\{d_i, i = 1, \ldots, k\}$ are i.i.d., we can use the bootstrap method to obtain resamples of normalized spacings $\{d_i^*, i = 1, \ldots, k\}$. The latter can be used to generate resamples $\{Y^{*1}, \ldots, Y^{*k}\}$ of the k largest order statistics from which the distribution of $\{\max Y_\omega\}$ can be obtained numerically. Since Fourier resampled resting-state data are considered to be null, the obtained distribution can be considered to be the



(a) Visual (b) Memory

Figure 2: Determination of the proper value for the noise fraction f for pseudoreal data. The solid horizontal lines represent the medians of t-statistics at the significance level $10^{-8} < P < 10^{-3}$ (uncorrected) for the contrasts V-F (a) and E-C (b) by applying GLM-NS on real fMRI activation data. The dashed curves are the medians of the t-statistics of activation-defined voxels for the contrasts V-F (a) and E-C (b) by applying GLM-NS on pseudoreal data where the true activations were defined by thresholding real activation data using a very high significance level ($P < 10^{-8}$ uncorrected) and adding resampled noise according to (16) for different noise fractions $f \in [0, 1]$. The medians of the t-value matched at around $f = 0.6$. Therefore, we picked two values for $f : f = 0.55$ representing the low noise case and $f = 0.65$ representing the high noise case.

null distribution of $\{\max Y_\omega\}$. It can be shown that, under certain regularity conditions, for a suitably chosen k, the normalized spacings are i.i.d. asymptotically [35]. Due to the large number of voxels in consideration, the asymptotic result is applicable in the present context. The chosen value for k was 100 for the bootstrap method and FWE was computed for $P = 0.05$.

4. Results

4.1. Estimation of Temporal Coefficients for Simulated Data. We computed the mean square errors (MSE) between the estimated temporal coefficients of the linear combination and the original ones generating the simulated data (Table 1) for a random noise fraction parameter f uniformly distributed in $[0, 1]$. The GLM-GS method is inferior to GLM-NS due to the small and irregular defined activations. The cCCA-RG method performs best and has an improvement of more than 25% of MSE in estimating the temporal coefficients. This experiment demonstrates the superior estimation performance of temporal coefficients $\tilde{\boldsymbol{\beta}}$ by the adaptive smoothing capability of cCCA.

4.2. Null Distribution of the Proposed Test Statistic. Although the proposed novel CCA test statistic has a similar form as the t-statistic in the GLM, its null distribution deviates significantly from the theoretical t-distribution as we mentioned previously. To shed more light on this issue, we applied different methods using Fourier resampled resting-state data and the contrast vector for the memory paradigm to get the null distributions of the contrast E-C. The results were plotted in Figure 3. The theoretical t-distribution with a DOF of 278 is also plotted for reference. It can be seen that all distributions are wider than the theoretical t-distribution, even for the GLM methods. Meanwhile, since n is much greater than p and K in this case, the adjustment induced by K in (14) is almost negligible. Since the distribution of the novel CCA test statistic has a complicated structure and is difficult to parameterize, it is necessary to use non-parametric methods to determine significance values accurately.

Table 1: Mean square errors (MSEs) of estimated coefficients for different methods. To define different spatial patterns of activations, 100000 randomly shaped activations within a 3×3 grid of pixels having a size of 2 to 9 pixels were generated. The corresponding time courses for the activated voxels were simulated to be linear combinations of the 4 random temporal regressors with random amplitudes. Different levels of noise were introduced by resampling 3×3 patches of resting-state fMRI data. The mean square errors between the originally simulated amplitudes of regressors and estimated ones are shown. The cCCA-RG method achieves more than 25% less MSE compared to GLM-NS. The GLM-GS method is worse than GLM-NS due to the small and irregularly defined activation patterns.

	$\Delta\beta_1^2$	$\Delta\beta_2^2$	$\Delta\beta_3^2$	$\Delta\beta_4^2$
GLM-NS	0.1331	0.1837	0.2200	0.1488
GLM-GS	0.1627	0.2047	0.2336	0.1769
cCCA-RG	0.0979	0.1339	0.1538	0.1091

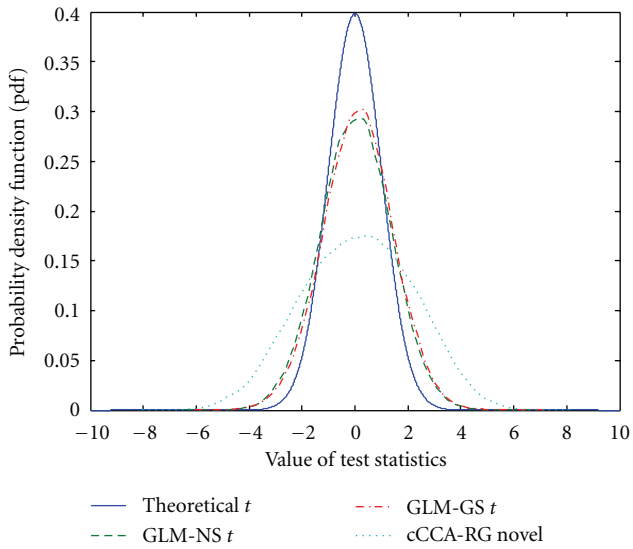


Figure 3: Distributions of the proposed CCA statistic ("cCCA-RG novel") along with the conventional t-statistic used in the GLM using resampled resting-state data and contrast E-C for the memory paradigm. The difference of the GLM-based t-statistic from a theoretical t-distribution (blue solid curve, DOF = 278) was mainly caused by the temporal correlation in fMRI signal. The novel CCA statistic has the widest profile because of the additional spatial modeling.

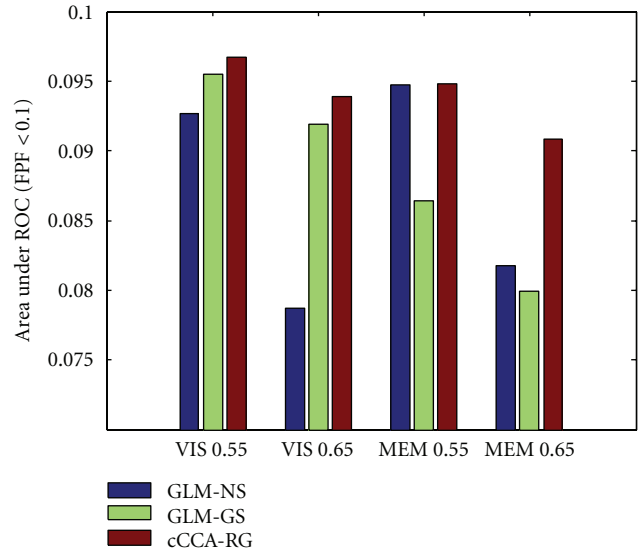


Figure 4: Detection performance of different data analysis methods showing the area under the ROC curve (AUR), integrated over FPF $\in [0, 0.1]$, using pseudoreal data. The AURs for the contrast V-F of the visual paradigm ("VIS 0.55" and "VIS 0.65") and the contrast E-C of the memory paradigm ("MEM 0.55" and "MEM 0.65") are shown for the low noise case ($f = 0.55$) and the high noise case ($f = 0.65$), respectively. The cCCA-RG achieves the greatest AUR values in all cases.

4.3. Area under the ROC Curve for Pseudoreal Data. We computed the area under the ROC curve (called "AUR") for a false positive fraction (FPF) less than 0.1 as an index of detection performance. The AUR quantity provides a weighted measure of detection power for specificities larger than 0.9 (which is the most interested range for fMRI data).

The AUR quantities for the contrast V-F of the visual data and for the contrast E-C of memory data are shown in Figure 4. Since the induced activations at the visual cortex are spatially extended, Gaussian smoothing (GLM-GS) yields better detection performance than GLM-NS. However, when activations are more irregular in shape and spatially localized as in the memory task, Gaussian smoothing produces adverse effects and GLM-GS consequently performed worse than GLM-NS. As can be seen, cCCA-RG always yields the top performance in all cases. The biggest advantage of cCCA-RG is in detecting small activations from a high noise background ("MEM 0.65").

In addition, we plotted the curves for the total false fraction (TFF) (including both false positives and false negatives) versus the false positive fraction (FPF) in Figure 5 (for the contrast V-F of the visual data) and Figure 6 (for the contrast E-C of the memory data). This measurement provides another perspective on the detection performance of different methods. For the extended activations of the contrast V-F, cCCA-RG achieves the smallest TFF at $f = 0.55$ (Figure 4(a)), followed by GLM-NS and GLM-GS. The GLM-GS method is effective in the high noise case (Figure 4(b) $f = 0.65$) and performs similar to cCCA-RG. In Figure 5, for the small activations of the contrast E-C, cCCA-RG remains the optimum and yields much more improved performance over other methods in the high noise case ($f = 0.65$). The GLM-GS method works poorly even in the high noise case. This demonstrates that it is destructive to apply fixed Gaussian spatial smoothing on the data with small activations. Constrained CCA combined with the proposed test statistic is more reliable and thus a better alternative to detect these activations.

4.4. Activation Maps Using Real Data (with Corrected $P < 0.05$). In the following, we show the activation maps with corrected $P < 0.05$ that are overlaid on their corresponding T2 images. Images in the figures are in radiological convention (left is right and vice versa). We only show them in 2D slices because the current application of cCCA-RG was in 2D, so was GLM-GS for a fair comparison and the (coregistered) activation maps were laid on each individual co-planar T2 image. In Figure 7, we show the activation maps of the contrast V-F of visual data for different methods from one representative subject. It can be seen that GLM-GS yields the smoothest activation map at the expense of loss of the visual cortex structures and GLM-NS preserves these folded structures much better but with some unappealing broken links. The activation map of cCCA-RG provides a good compromise between the smoothness of activations and preservation of fine cortical structure.

The activation maps of the contrast E-C of memory data from another representative subject are shown in Figure 8.

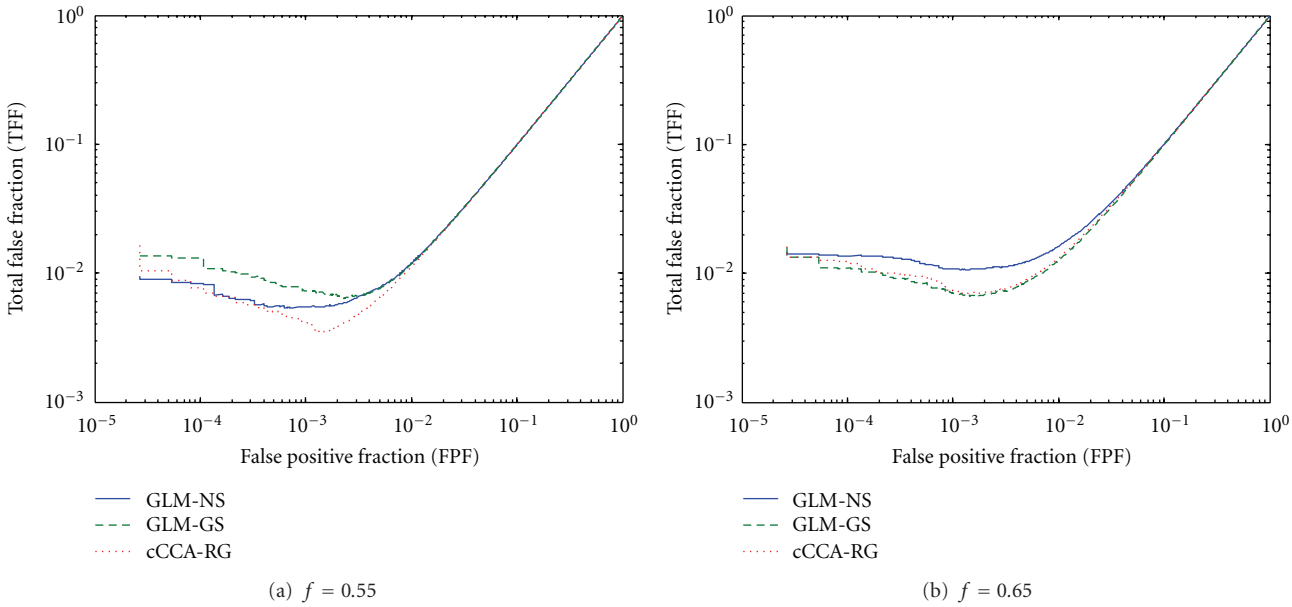


Figure 5: The total false fraction (TFF) (including false positives and false negatives) versus the false positive fraction (FPF) for the contrast V-F of the visual paradigm for pseudoreal data: (a) the low noise case ($f = 0.55$) and (b) the high noise case ($f = 0.65$). Note that all TFF curves have minima in the interval $[0.001, 0.01]$. The cCCA-RG performs nearly optimal in both cases by achieving the minimum value of TFF.

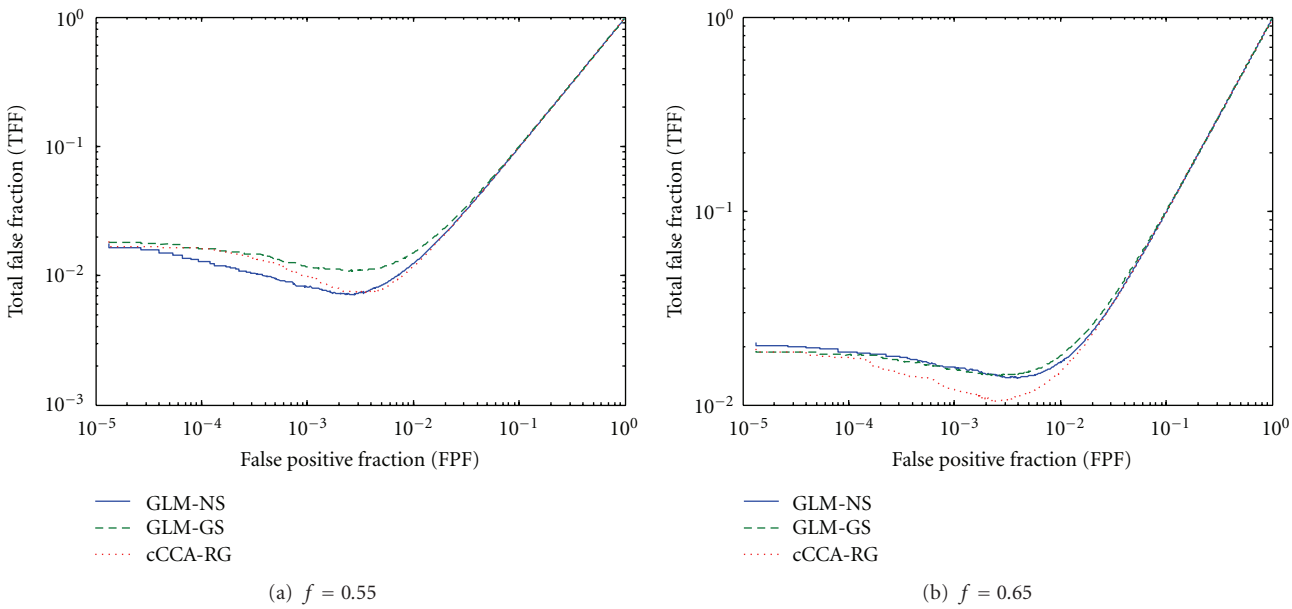
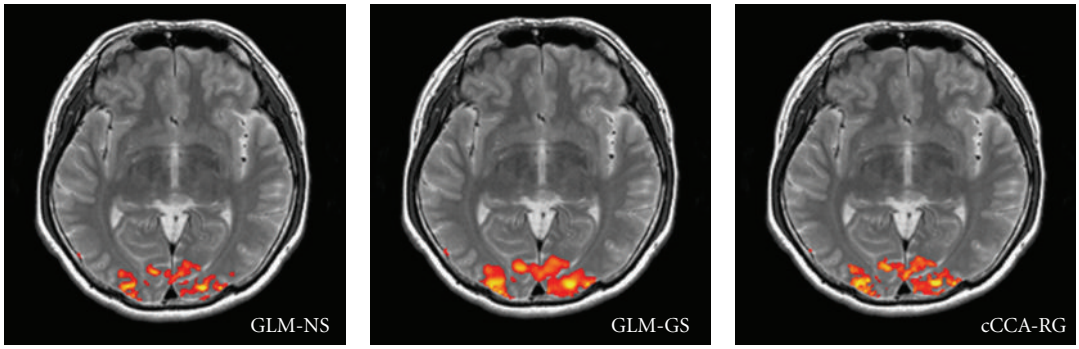


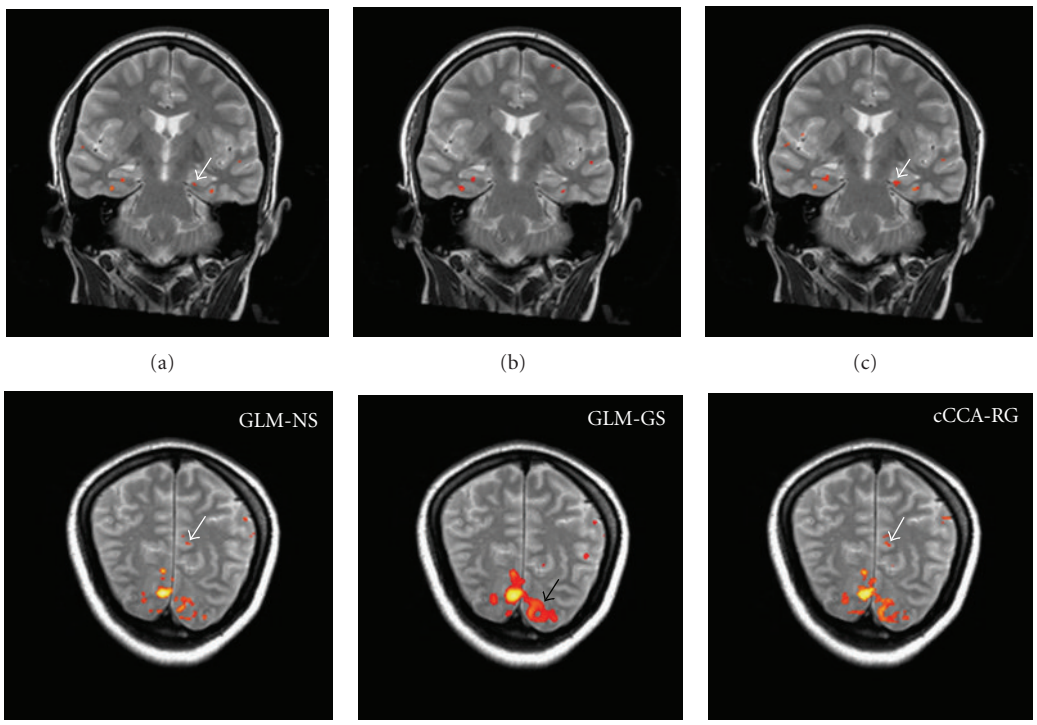
Figure 6: The total false fraction (TFF) (including false positives and false negatives) versus the false positive fraction (FPF) for the contrast E-C of the memory paradigm for pseudoreal data: (a) the low noise case ($f = 0.55$) and (b) the high noise case ($f = 0.65$). Note that all TFF curves have minima in the interval $[0.001, 0.01]$. The cCCA-RG method is optimal in both the low and high noise cases.

The slices shown in the upper row contain an anterior portion of the hippocampal complex. Symmetrical activations in hippocampus and parahippocampal gyrus are detected by GLM-NS and cCCA-RG. The missing activation at the left hippocampus (see white arrows) of GLM-GS demonstrates the undesirable effects of a fixed isotropic Gaussian spatial smoothing on localized weak activation patterns. A more posterior slice is shown in the bottom row. Memory encoding activation is obtained in the posterior cingulate cortex and precuneus. Using the GLM-GS method, activations appear overly bulgy and have some unlikely connections through white matter (shown by the black arrow). Also, small and weak activations in the posterior cingulate cortex (see white arrows) are not shown in the activation map of GLM-GS.



(a) (b) (c)

Figure 7: Activation maps for the contrast V-F of the visual paradigm using corrected P-values ($P < 0.05$). The GLM-GS method yields the smoothest activation map at the expense of showing activations reaching outside of gray matter. The GLM-NS method preserves activations in gray matter much better but with unappealing broken links among activated voxels. The activation map using cCCA-RG provides a compromise between the smooth appearance of activations and preservation of fine cortical structure.



(a) (b) (c)

(d) (e) (f)

Figure 8: Activation maps for the contrast E-C of the memory paradigm using corrected P-values ($P < 0.05$). Upper row: activations in the anterior portion of the hippocampal complex; lower row: activations in the posterior and middle cingulate cortex and in the precuneus. Note that GLM-NS and cCCA-RG lead to symmetric (left and right) activation patterns in the hippocampus and parahippocampal gyrus and also to weak and localized activations in the posterior/middle cingulate cortex (see white arrows). Using GLM-GS, strong activation patterns become overly bulgy (see black arrow). Compared to GLM-NS, cCCA-RG yields more activated voxels and better connected activations confined to gray matter.

The GLM-GS method leads not only to missing activations but also to artifactual activations where a large fraction of false activations show up in white matter and CSF regions due to the spherical (nondirectional) smoothing kernel. The cCCA-RG method yields more activated voxels and better connected activations in gray matter than GLM-NS.

To make a quantitative comparison of the locations of activations of different methods, we used the gray matter mask from the acquired IR-SE-EPI scan and calculated the ratio of the number of activated voxels detected in gray matter and the number of activated voxels detected outside of gray matter (listed in Table 2). This ratio reflects the degree

Table 2: Ratio of activations in gray matter and outside of gray matter for different analysis methods. The number in the table is the ratio of the number of voxels detected in gray matter and the number of voxels detected outside of gray matter. Note that the GLM-NS has the highest value because there is no smoothing involved. Compared to fixed Gaussian spatial smoothing (GLM-GS), cCCA-RG yields higher ratios demonstrating less blurring and better confinement to gray matter.

	GLM-NS	GLM-GS	cCCA-RG
V−F	6.21	2.84	3.35
E−C	1.81	1.32	1.53

of activations confined to gray matter. As expected, GLM-NS has the highest value because of no smoothing involved. Compared to fixed Gaussian smoothing (GLM-GS), cCCA-RG yields higher ratios, which demonstrates that the adaptive smoothing suffers less blurring outside of gray matter than fixed Gaussian spatial smoothing.

5. Discussion

Using the newly developed directional test statistic for cCCA of fMRI data, we are able to compare cCCA with traditional GLM methods for a more complicated memory paradigm. The quantitative results from the simulated and pseudoreal data and the qualitative results from real fMRI data clearly demonstrate that the proposed method (directional test with cCCA) outperformed the conventional GLM with and without Gaussian smoothing. This work paves the way for applying CCA methods for testing general linear contrasts in a more complicated fMRI experimental design.

Our comprehensive evaluation study also provides valuable insights for applying smoothing in fMRI data analysis. The pseudoreal data used in this study can be divided into four situations: (1) spatially extended and strong activation (VIS 0.55); (2) spatially extended and weak activation (VIS 0.65); (3) focal and strong activation (MEM 0.55); (4) focal and weak activation (MEM 0.65). As expected, the smoothing does not provide much benefit for detecting strong activations. The Gaussian smoothing is only effective for the second situation—spatially extended and weak activation (Figures 4 and 5) because the smoothing helps little for detection of strong signals and the isotropic smoothing adversely eliminates the small or irregular weak activation patterns. The adaptive smoothing by cCCA always performed best in all four situations and the biggest advantage takes place for the last situation—focal and weak activations (Figures 4–6). For real fMRI data, the Gaussian smoothing can yield a large block of smooth activations, which are appealing to human visual perception. However, there is a risk of overlooking important subtle activations as well as overestimating the extent of strong activations (Figure 7). As can be seen in Figures 7 and 8, the adaptive smoothing by cCCA yields activation maps that are not only visually appealing (smoothness) but also well localized (along the gyri and sulci of gray matter).

The improved detection performance of cCCA is at the expense of computation. If an exhaustive search is used for the optimization of constrained CCA, the number of CCA computations will be equal to the number of possible voxel configurations in the chosen neighborhood. This number is of the order $O(2^{N-1})$, where N is the number of voxels in the search area [12, 24]. That means 256 CCA computations for a 3×3 in-plane neighborhood and 2^{26} for a $3 \times 3 \times 3$ voxel volume. Heuristic search methods, such as the branch-and-bound algorithm [10, 14] and a region-growing algorithm [12], were used to reduce the computational cost and to maintain the detection performance. The current implementation of cCCA-RG in 2D [12] is feasible for routine fMRI data analysis. For the estimation of a 2D slice with 6317 in-brain pixels, using MATLAB on a computer equipped with Intel Core 2 2.4 GHz CPU and 4 GB memory, cCCA-RG takes about 24 seconds. Although it is about 10 times slower than GLM-NS and GLM-GS, a fully 3D brain volume sequence can be processed within 10 minutes. On the other hand, the rapid evolving computer hardware and parallel computing techniques, for example, GPU computing, can dramatically shorten the time for cCCA in future applications.

Besides CCA [8–13, 24], there exist other methods that use adaptive smoothing techniques for fMRI data analysis (e.g., [36–38]). A quite different method is used in [38], where a propagation-separation procedure is applied on contrast and residual images, obtained by the GLM, to achieve adaptive smoothing of the estimated parameters. The final activation detection is based on random field theory [6]. However, the advantage of preserved shape and geometry of the activation areas and increased signal-to-noise ratio was only demonstrated by simulated data and real motor data, thus the effectiveness of this postestimation smoothing on focal weak activations is unclear. Another test statistic similar to canonical correlation, proposed in [36], is defined as a ratio between the energy of signal space and the energy of residuals. Its power relies on the optimal spatial weighting based on different signal spaces. This method is equivalent to conventional CCA. However, the maximum energy ratio, in its current formulation, does not allow for a more general contrast design, as well as a directional test. Moreover, the estimate and inference have to be done for each signal space, which is computationally expensive. Our test statistics is more general and outperforms the GLM with or without Gaussian smoothing. In addition to its improved sensitivity, advantages are that general contrasts can be defined after the estimation and a directional test is readily available. It is worthwhile to note that adaptive smoothing can also be achieved through spatial priors defined in a Bayesian framework (e.g., [39, 40]), which produces posterior probability maps instead of statistical parametric maps as in classical inference. Though Bayesian methods hold some advantages over classical inference, such as capability of inferring an effect size and no need for multiple comparison correction, the specification of the priors and the likelihood functions may have a large impact on the final results and the computation is usually more complex and time consuming. For comparing all these adaptive smoothing methods, a thorough study needs to be conducted to evaluate their performance from detection

performance of different types of brain activations to their computational cost.

It is important to keep in mind that the advantage of adaptive smoothing may diminish in conventional group analysis, where isotropic smoothing is necessary to improve correspondence of imperfectly registered homologous areas. Nevertheless, the usefulness of adaptive smoothing will be greatly appreciated for fMRI-aided neurosurgical planning [41] and region-of-interest analysis of localized brain functions [42].

One issue that this paper has not addressed is the temporal correlation of the noise and a possible correction of the test statistic by prewhitening, as usually done in data analysis using the GLM. Based on the Gauss-Markov theorem, the LS solution of the GLM is the MVU estimator when Gaussian white noise assumption is satisfied, otherwise the weighted LS solution (using the inverse of the noise covariance matrix) becomes the BLUE. For cCCA, a BLUE does not exist because the optimization of the spatial constraints leads to a nonlinear model even though the spatial constraints can be linear [24]. Therefore, unbiasedness of constraint CCA by prewhitening is not possible and non-parametric methods need to be used to obtain accurate P-values.

The purpose of this research is to develop a simple directional test statistic for cCCA similar to a t-statistic. Given that the HRF is modeled perfectly, a t-test, as a likelihood ratio test in the univariate case, is the most sensitive test. For block designs, the canonical 2-gamma function is a good choice for the temporal modeling of the BOLD response. However, in event-related designs, more complicated temporal regressors may be useful (such as first and second derivative of the HRF function) to model the delay and dispersion of the BOLD response. In such a scenario, an unsigned test statistic, for example, F-statistic, is preferred to test for the evoked regional effects. A test statistic for CCA similar to F-statistic can be derived from Wilks' Λ as

$$F\left(\nu_{\tilde{\mathbf{H}}}, \nu_{\tilde{\mathbf{E}}}\right) = \frac{1-\Lambda}{\Lambda}\frac{\nu_{\tilde{\mathbf{E}}}}{\nu_{\tilde{\mathbf{H}}}}, \tag{18}$$

where $\nu_{\tilde{\mathbf{H}}}$ and $\nu_{\tilde{\mathbf{E}}}$ are the degrees of freedom of the hypothesis matrix and the error matrix, respectively. The delay and dispersion regressors can be included in our proposed CCA method in the same way as for the GLM since the temporal modeling of the HRF response is the same for both methods.

6. Conclusions

In this paper, we derived a novel directional test statistic for CCA so that CCA can handle general linear contrasts in more complicated fMRI paradigms. Using this novel test statistic, different contrasts can be tested after model fitting without reparameterization of the design matrix and reestimating each individual contrast of interest. With the proper constraints on the spatial coefficients of CCA, this CCA statistic can yield a more powerful test than the traditional t-test in the GLM, especially for weakly evoked and localized brain activations. This behavior was demonstrated not only by superior performance using simulations and traditional ROC techniques but also by activation maps of real fMRI applications. Since the trend in fMRI is to move toward high-resolution imaging where the signal is weak, the spatial correlation is strong, and the amount of data is enormous, we envision that our method with improved detection power and computation time will be important for future fMRI data analysis.

Acknowledgment

This research is supported by the NIH/NIA (Grant No. 1R21AG026635).

References

[1] K. J. Worsley and K. J. Friston, "Analysis of fMRI time-series revisited—again," *NeuroImage*, vol. 2, no. 3, pp. 173–181, 1995.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, PTR Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[3] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde, "Processing strategies for time-course data sets in functional MRI of the human brain," *Magnetic Resonance in Medicine*, vol. 30, no. 2, pp. 161–173, 1993.

[4] B. A. Ardekani and I. Kanno, "Statistical methods for detecting activated regions in functional MRI of the brain," *Magnetic Resonance Imaging*, vol. 16, no. 10, pp. 1217–1225, 1998.

[5] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, "A three-dimensional statistical analysis for CBF activation studies in human brain," *Journal of Cerebral Blood Flow and Metabolism*, vol. 12, no. 6, pp. 900–918, 1992.

[6] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, and A. C. Evans, "A unified statistical approach for determining significant signals in images of cerebral activation," *Human Brain Mapping*, vol. 4, no. 1, pp. 58–73, 1996.

[7] T. Q. Duong, E. Yacoub, G. Adriany et al., "High-resolution, spin-echo BOLD, and CBF fMRI at 4 and 7 T," *Magnetic Resonance in Medicine*, vol. 48, no. 4, pp. 589–593, 2002.

[8] O. Friman, J. Cedefamn, P. Lundberg, M. Borga, and H. Knutsson, "Detection of neural activity in functional MRI using canonical correlation analysis," *Magnetic Resonance in Medicine*, vol. 45, no. 2, pp. 323–330, 2001.

[9] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Detection of neural activity in fMRI using maximum correlation modeling," *NeuroImage*, vol. 15, no. 2, pp. 386–395, 2002.

[10] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Adaptive analysis of fMRI data," *NeuroImage*, vol. 19, no. 3, pp. 837–845, 2003.

[11] R. Nandy and D. Cordes, "Improving the spatial specificity of canonical correlation analysis in fMRI," *Magnetic Resonance in Medicine*, vol. 52, no. 4, pp. 947–952, 2004.

[12] M. Jin, R. Nandy, and D. Cordes, "Two local constrained canonical correlation analysis methods for fMRI," in *Proceedings of the ISMRM 17th Scientific Meeting*, p. 1708, 2009.

[13] M. Ragnehed, M. Engström, H. Knutsson, B. Söderfeldt, and P. Lundberg, "Restricted canonical correlation analysis in functional MRI-validation and a novel thresholding technique," *Journal of Magnetic Resonance Imaging*, vol. 29, no. 1, pp. 146–154, 2009.

[14] S. Das and P. K. Sen, "Restricted canonical correlations," *Linear Algebra and Its Applications*, vol. 210, no. C, pp. 29–47, 1994.

[15] S. Smith, M. Jenkinson, C. Beckmann, K. Miller, and M. Woolrich, "Meaningful design and contrast estimability in FMRI," *NeuroImage*, vol. 34, no. 1, pp. 127–136, 2007.

[16] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283–298, 1978.

[17] J. A. Sorenson and X. Wang, "ROC methods for evaluation of fMRI techniques," *Magnetic Resonance in Medicine*, vol. 36, no. 5, pp. 737–744, 1996.

[18] P. Skudlarski, R. T. Constable, and J. C. Gore, "ROC analysis of statistical methods used in functional MRI: individual subjects," *NeuroImage*, vol. 9, no. 3, pp. 311–329, 1999.

[19] R. R. Nandy and D. Cordes, "Novel ROC-type method for testing the efficiency of multivariate statistical methods in fMRI," *Magnetic Resonance in Medicine*, vol. 49, no. 6, pp. 1152–1162, 2003.

[20] R. R. Nandy and D. Cordes, "Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data," *Magnetic Resonance in Medicine*, vol. 50, no. 2, pp. 354–365, 2003.

[21] R. R. Nandy and D. Cordes, "New approaches to receiver operating characteristic methods in functional magnetic resonance imaging with real data using repeated trials," *Magnetic Resonance in Medicine*, vol. 52, no. 6, pp. 1424–1431, 2004.

[22] R. Nandy and D. Cordes, "A semi-parametric approach to estimate the family-wise error rate in fMRI using resting-state data," *NeuroImage*, vol. 34, no. 4, pp. 1562–1576, 2007.

[23] A. Rencher, *Multivariate Statistical Inference and Applications*, John Wiley & Sons, New York, NY, USA, 1998.

[24] D. Cordes, M. Jin, T. Curran, and R. Nandy, "Optimizing the performance of local canonicalcorrelation analysis in fMRI using spatial constraints," *Human Brain Mapping*. In press.

[25] D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Physical Review Letters*, vol. 73, no. 7, pp. 951–954, 1994.

[26] E. Bullmore, C. Long, J. Suckling et al., "Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains," *Human Brain Mapping*, vol. 12, no. 2, pp. 61–78, 2001.

[27] M. Breakspear, M. J. Brammer, E. T. Bullmore, P. Das, and L. M. Williams, "Spatiotemporal wavelet resampling for functional neuroimaging data," *Human Brain Mapping*, vol. 23, no. 1, pp. 1–25, 2004.

[28] E. Bullmore, M. Brammer, S. C. R. Williams et al., "Statistical methods of estimation and inference for functional MR image analysis," *Magnetic Resonance in Medicine*, vol. 35, no. 2, pp. 261–277, 1996.

[29] J. J. Locascio, P. J. Jennings, C. I. Moore, and S. Corkin, "Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging," *Human Brain Mapping*, vol. 5, no. 3, pp. 168–193, 1997.

[30] R. Nandy, C. Green, and D. Cordes, "Nonparametric analysis of fMRI data using bootstrap in autoregression," *Proceedings ISMRM*, vol. 10, p. 1438, 2002.

[31] O. Friman and C. F. Westin, "Resampling fMRI time series," *NeuroImage*, vol. 25, no. 3, pp. 859–867, 2005.

[32] A. R. Laird, B. P. Rogers, and M. E. Meyerand, "Comparison of Fourier and wavelet resampling methods," *Magnetic Resonance in Medicine*, vol. 51, no. 2, pp. 418–422, 2004.

[33] M. Jin and D. Cordes, "Validation of resampling methods for fMRI data," *Neuroimage*, vol. 41, p. S525, 2008.

[34] R. Pyke, "Spacings (with discussion)," *Journal of the Royal Statistical Society: Series B*, vol. 27, pp. 395–449, 1965.

[35] I. Weissman, "Estimation of parameters and larger quantiles based on the k largest observations," *Journal of the American Statistical Association*, vol. 73, pp. 812–815, 1978.

[36] U. E. Ruttimann, M. Unser, R. R. Rawlings et al., "Statistical analysis of functional MRI data in the wavelet domain," *IEEE Transactions on Medical Imaging*, vol. 17, no. 2, pp. 142–154, 1998.

[37] G. A. Hossein-Zadeh, B. A. Ardekani, and H. Soltanian-Zadeh, "Activation detection in fMRI using a maximum energy ratio statistic obtained by adaptive spatial filtering," *IEEE Transactions on Medical Imaging*, vol. 22, no. 7, pp. 795–805, 2003.

[38] K. Tabelow, J. Polzehl, H. U. Voss, and V. Spokoiny, "Analyzing fMRI experiments with structural adaptive smoothing procedures," *NeuroImage*, vol. 33, no. 1, pp. 55–62, 2006.

[39] W. D. Penny, N. J. Trujillo-Barreto, and K. J. Friston, "Bayesian fMRI time series analysis with spatial priors," *NeuroImage*, vol. 24, no. 2, pp. 350–362, 2005.

[40] G. Flandin and W. D. Penny, "Bayesian fMRI data analysis with sparse spatial basis function priors," *NeuroImage*, vol. 34, no. 3, pp. 1108–1125, 2007.

[41] J. R. Petrella, L. M. Shah, K. M. Harris et al., "Preoperative functional MR imaging localization of language and motor areas: effect on therapeutic decision making in patients with potentially resectable brain tumors," *Radiology*, vol. 240, no. 3, pp. 793–802, 2006.

[42] M. M. Zeineh, S. A. Engel, and S. Y. Bookheimer, "Application of cortical unfolding techniques to functional MRI of the human hippocampal region," *NeuroImage*, vol. 11, no. 6 I, pp. 668–683, 2000.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2012, Article ID 958142, 8 pages
doi:10.1155/2012/958142

Research Article

Selective Extraction of Entangled Textures via Adaptive PDE Transform

Yang Wang,[1] Guo-Wei Wei,[1,2] and Siyang Yang[1]

[1] *Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA*
[2] *Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, USA*

Correspondence should be addressed to Guo-Wei Wei, wei@math.msu.edu

Received 29 August 2011; Accepted 11 October 2011

Academic Editor: Shan Zhao

Texture and feature extraction is an important research area with a wide range of applications in science and technology. Selective extraction of entangled textures is a challenging task due to spatial entanglement, orientation mixing, and high-frequency overlapping. The partial differential equation (PDE) transform is an efficient method for functional mode decomposition. The present work introduces adaptive PDE transform algorithm to appropriately threshold the statistical variance of the local variation of functional modes. The proposed adaptive PDE transform is applied to the selective extraction of entangled textures. Successful separations of human face, clothes, background, natural landscape, text, forest, camouflaged sniper and neuron skeletons have validated the proposed method.

1. Introduction

Texture is one of the important features characterizing many natural and man-made images. Texture characterization and analysis are usually performed according to the spatial as well as frequency variations of brightness, pixel intensities, color, and texture orientation in the different regions of the image corresponding to different types of textures. For example, the roughness or bumpiness of an image usually refers to variations in the intensity values, or gray levels. Texture segmentation, recognition, and interpretation are critical for human visual perception and processing. As a result, research on texture analysis has received considerable attention in recent years. A large number of approaches has been proposed for texture classification and segmentation [1–16]. In general, texture analysis methods fall into two categories: statistical methods which analyze the Fourier power spectrum, gray level values, and various variance matrices of the input image, and structural methods which are knowledge-based algorithms with an emphasis on the structural primitives and their placement rules. Some examples of such methods include Markov random field models [17, 18], simultaneous autoregressive model [19], and fractal models [20]. Among many existing approaches, local variation minimization has been a popular and powerful technique in image analysis [21] with applications to the texture modeling [22]. Multiphase segmentation approaches are based on the structural division of gray scales [23]. More recently, multiresolution approaches have become more important in texture analysis [19, 24–26], where fixed-size neighborhood and window size are used to derive features at varying scales corresponding to the input image at different resolutions.

In general, the total texture extraction has become a mature technique in real applications. However, despite the progress in the past few decades, selective extraction of entangled textures encounters a number of difficulties. One difficulty is due to *spatial entanglement*, including orientation mixing of various textures. Another difficulty is due to *gray-scale entanglement*, especially the near-continuous merging of various textures. The other difficulty is due to *frequency entanglement* when two similar but different textures share overlapping frequency band in the frequency domain. This difficulty would especially plague texture analysis when many high-frequency textures coexist.

In this work, we propose an adaptive partial differential equation (PDE) transform approach for selective extraction of entangled textures. By using arbitrarily high-order PDEs,

the PDE transform is able to decompose signals, images, and data into functional modes, which exhibit appropriate time-frequency localizations [27–31]. Additionally, the PDE transform is able to provide a perfect reconstruction. Unlike wavelet transform or Fourier transform, the PDE transform offers results in the physical domain, which enables straightforward mode analysis and secondary processing. Based on the image mode functions generated by the PDE transform method, the adaptive PDE transform algorithm calculates the variance of the local variation of the image mode functions followed by the corresponding thresholding analysis.

2. PDE Transform Method

In the past two decades, PDE-based image processing approaches have raised a strong interest in the image processing and applied mathematical communities and have opened new approaches for image denoising, enhancement, edge detection, restoration, segmentation, and so forth. The use of PDEs for image analysis started as early as 1980s when Witkin first introduced diffusion equation for image denoising [32]. The time evolution of an image under a diffusion operator is formally equivalent to the lowpass filter. After Perona and Malik introduced anisotropic diffusion equation in 1990 [33], nonlinear PDEs have found great applications for a variety of image processing tasks such as edge detection and denoising. Two important advances in the history of image processing, namely, the Perona-Malik equation and the total variation methods [21], employ second-order nonlinear PDEs for image analysis. The Willmore flow, proposed in 1920s, is a fourth-order geometric PDE and has also been used for surface analysis. In the past decade, fourth-order nonlinear PDEs have attracted much attention in image analysis [34–36].

Arbitrarily high-order nonlinear PDEs were introduced by Wei in 1999 to more efficiently remove image noise in edge-preserving image restoration [34]:

$$\begin{aligned} u_t(\mathbf{r},t) &= \sum_q \nabla\cdot\left[d_q(u,|\nabla u|)\nabla\nabla^{2q}u\right] \\ &\quad + e(u,|\nabla u|), \quad (q=0,1,\ldots), \end{aligned} \tag{1}$$

where $u \equiv u(\mathbf{r},t)$ is the image function, $d_q(u(\mathbf{r}),|\nabla u(\mathbf{r})|,t)$ and $e(u(\mathbf{r}),|\nabla u(\mathbf{r})|,t)$ are edge-sensitive diffusion coefficients and enhancement operator, respectively. The Perona-Malik equation is recovered at $q=0$ and $e(u(\mathbf{r}),|\nabla u(\mathbf{r})|,t)=0$. As in the original Perona-Malik equation, the hyperdiffusion coefficients $d_q(u(\mathbf{r}),|\nabla u(\mathbf{r})|,t)$ in (1) can be chosen in many different ways. For instance, one can set

$$d_q(u(\mathbf{r}),|\nabla u(\mathbf{r})|,t) = d_{q0}\exp\left[-\frac{|\nabla u|^2}{2\sigma_q^2}\right], \tag{2}$$

where the values of constants d_{q0} depend on the noise level, and σ_0 and σ_1 are chosen as the local statistical variance of u and ∇u:

$$\sigma_q^2(\mathbf{r}) = \overline{\left|\nabla^q u - \overline{\nabla^q u}\right|^2} \quad (q=0,1). \tag{3}$$

The notation $\overline{Y(\mathbf{r})}$ above denotes the local average of $Y(\mathbf{r})$ centered at position $\mathbf{r}$. In this algorithm, the statistical measure based on the variance is important for discriminating image edges from noise. As such, one can bypass the image preprocessing, that is, the convolution of the noise image with a test function or smooth mask.

In general, the nonlinear PDE operators described above serve as lowpass filters. PDE-based nonlinear highpass filters were introduced by Wei and Jia [37] in 2002. They constructed two weakly coupled PDEs to act as a highpass filter. Recently, this approach has been combined with Wei's earlier arbitrarily high-order nonlinear PDE operator to give [29]

$$\partial_t\begin{pmatrix} u_m \\ v_n \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{j=0}^{m-1}\nabla\cdot d_{uj}\nabla\nabla^{2j} - \epsilon_{u_m}, \epsilon_{v_n} \\ \epsilon_{u_m}, \displaystyle\sum_{j=0}^{n-1}\nabla\cdot d_{vj}\nabla\nabla^{2j} - \epsilon_{v_n} \end{pmatrix}\begin{pmatrix} u_m \\ v_n \end{pmatrix}, \tag{4}$$

where $\epsilon_{u_m} \equiv \epsilon_{u_m}(|\nabla u_m|)$ and $\epsilon_{u_n} \equiv \epsilon_{u_n}(|\nabla v_n|)$ are made edge sensitive. As lowpass filters, both $d_{uj} \equiv d_{uj}(|\nabla u_m|) \geq 0$ and $d_{vj} \equiv d_{vj}(|\nabla v_n|) \geq 0$ when j is even. Similarly, both $d_{uj}(|\nabla u_m|) \leq 0$ and $d_{vj}(|\nabla u_m|) \leq 0$ when j is odd. We can define a PDE transform as

$$w_{m,n}(\mathbf{r},t) = u_m(\mathbf{r},t) - v_n(\mathbf{r},t) = H_{mn}(\mathbf{r},t)X(\mathbf{r}), \tag{5}$$

where $H_{mn}(\mathbf{r},t)$ can be regarded as a coupled nonlinear PDE operator. In order for (5) to work properly, we choose $|d_{vj}(|\nabla v_n|)| \gg |d_{uj}(|\nabla u_m|)|$. As shown in our earlier work, by increasing the order of the highest derivative, one can increase frequency localization and accuracy of the PDE transform for mode decomposition [29]. The frequency selection of $w_{m,n}(\mathbf{r},t)$ also depends on the evolution time. High-order PDEs are integrated by using the Fourier pseudospectral method [29].

In the PDE transform, intrinsic mode functions w^k are systematically extracted from residues X^k, that is,

$$w_{mn}^k = H_{mn}X_{mn}^k, \quad \forall k = 1,2,\ldots, \tag{6}$$

where w_{mn}^k is the kth mode function. Here, the residue function is given by

$$X_{mn}^k = X_{mn}^1 - \sum_{j=1}^{k-1} w_{mn}^j, \quad \forall k = 2,3,\ldots, \tag{7}$$

where $X_{mn}^1 = X(\mathbf{r})$. Therefore, $X = \sum_{j=1}^{k-1} w_{mn}^j + X_{mn}^k$ is a perfect reconstruction of X in terms of all the mode functions and the last residue. The mode decomposition algorithm given in (6) is inherently nonlinear, even if a linear PDE operator might be used.

The PDE transform is applied to Figure 1(a) to extract the three textures in Figures 1(b), 1(c), and 1(d). Note that only one texture is isolated at each time, which means the proposed PDE transform is able to perform a controlled or selective segmentation of textures. The PDEs of up to order 200 have been used for the selective texture segmentation. Numerically, such high-order linear PDE needs to be solved in the frequency domain [29]. Due to the ideal frequency localization, three textures are separated with clear boundary sharpness.

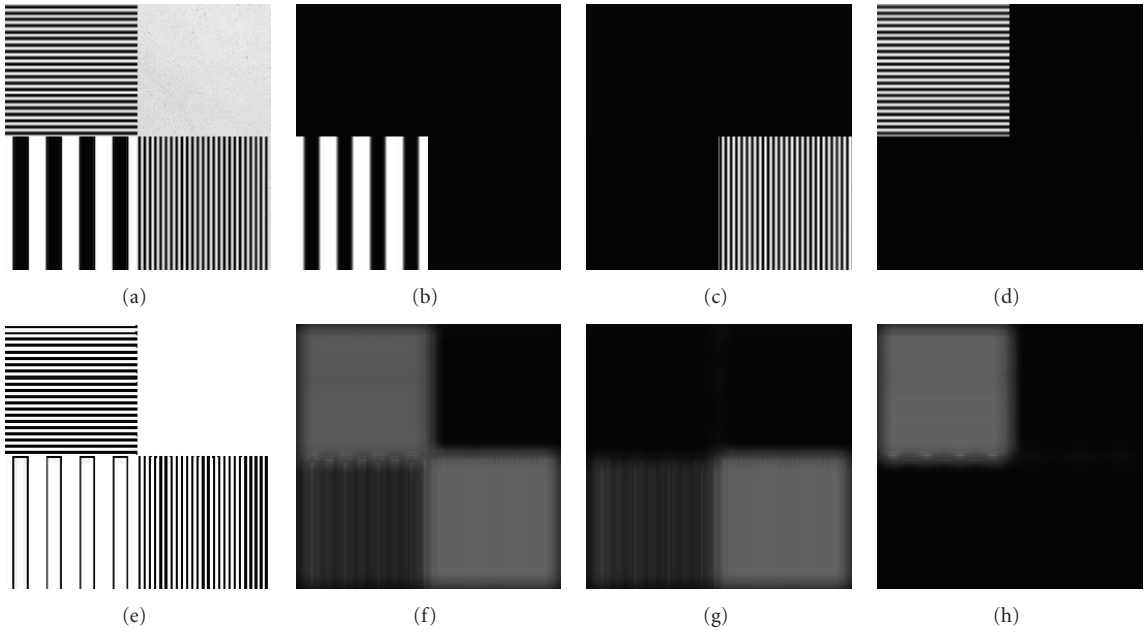


Figure 1: Extraction of various embedded textures using the PDE transform. (a) shows the original image composed of various horizontal and vertical textures. (b)–(d) show the three texture patterns extracted by applying the PDE transform, one at each time. (e) shows the edge mode obtained by applying the PDE transform to (a). (f) shows the variance of the local variation of the image mode function (e). (g) and (h) show the projection, or average, of the variance in (f) along x- and y-direction, respectively.

3. Adaptive PDE Transform Algorithm

The separation of textures that are highly entangled in spatial locations, frequency ranges, and gray scales become a challenge, and conventional segmentation techniques are in general not applicable for such cases. For example, highly oscillatory textures can be separated from slowly varying background but cannot be separated from another texture with overlapping frequency distribution purely based on frequency fingerprints. To selectively distinguish such entangled textures of high frequency, one needs a mode decomposition algorithm that is able to be highly localized in frequency. Second-order PDEs are poorly localized in the frequency domain [29]. Whereas, the PDE transform with high-order PDEs provides desirable frequency localization [29]. However, the PDE transform by itself does not perform well for the separation of entangled textures. To this end, we introduce an adaptive PDE transform algorithm for selective texture extraction. The essence of the adaptive PDE algorithm lies in the realization that features of various textures are closely correlated with both the magnitude and smoothness of the gray-scale values, or, equivalently, the local variation of the image mode functions. Similar ideas have been implemented in other methods such as total variation [21].

Nonlinear PDEs have been widely applied to detect images with noises. However, despite better image edge protection, the nonlinear anisotropic diffusion operator may still break down when the gradient generated by noise is comparable to image edges and features [38]. Application of a preconvolution with a smoothing function to the image can practically alleviate the instability and reduce gray-scale oscillation, but the image quality is often degraded. One alternative solution introduced by Wei [34] is to statistically discriminate noise from image edges by a measure based on the local statistical variance of the image or its gradient. Such a local statistical variance based edge-stopping algorithm was found to work very well for image restoration.

Similar statistical analysis can be employed to perform selective texture extraction for images containing highly entangled and overlapping textures. In the present approach, we first compute the local variation of each pixel of the image mode functions obtained by the high-order PDE transform. Unlike the total variation, the local variation is still a function, of which the variance can be calculated:

$$E(X(\mathbf{r})) = \overline{\left| \left| \nabla X^k(\mathbf{r}) \right| - \overline{\left| \nabla X^k(\mathbf{r}) \right|} \right|^2}, \tag{8}$$

where $X^k(\mathbf{r})$ is the kth mode function obtained by the PDE transform (7), and $|\nabla X^k(\mathbf{r})|$ is evaluated locally over the neighbor pixels. Equation (8) yields a statistical analysis which is used for various texture separation and segmentation with appropriate threshold values. Various threshold values need to be chosen to select the range of the variance corresponding to the particular texture of interest. All the previously classified textures are registered for sequential/recursive texture extractions. A flowchart of the adaptive algorithm of PDE transform is shown in Figure 2.

Figure 1(e) shows the edge mode obtained by applying the PDE transform to Figure 1(a). Figure 1(f) shows the variance of the local variation of gray scale calculated using

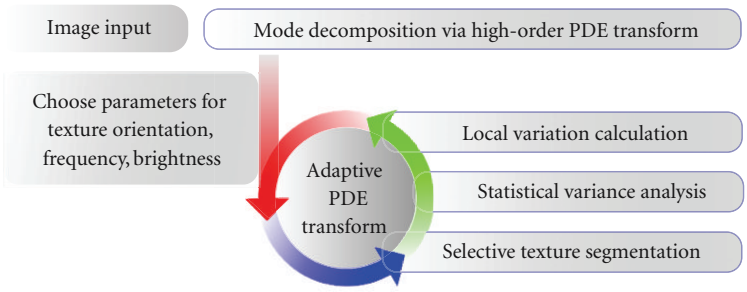


Figure 2: Algorithm of adaptive PDE transform for entangled texture separation.

the adaptive PDE transform. Figures 1(g) and 1(h) show the projection, or average, of the variance in Figure 1(f) along x- and y-direction, respectively. By slicing out different domain of the variance in Figure 1(f), three different textures in Figures 1(b)– 1(d) are then perfectly separated from each other.

4. Applications

In this section, the adaptive PDE transform is applied to three different cases to illustrate its superior capability of selective texture separation. The three images feature different types of entangled textures. Figure 3(a) contains textures overlapping in the physical space with entangled frequency fingerprints. Figures 5(a) and 6(a) contain spatially segmented textures overlapping in the frequency domain. Figure 7 contains textures with overlapping textures highly entangled in both the frequency and spatial domains.

4.1. Text-Image Separation. The adaptive PDE transform method employing the variance of the local variation of the image mode functions is applied to several benchmark test cases. In particular, separation of text and texture can be regarded as a generalized type of texture analysis. In Figure 3, texts of various fonts are imprinted on the background image. Additional background watermark in Chinese is also presented in Figure 3(a). The separation of English title from both background image and Chinese characters is a challenging task in terms of texture analysis because of the high degree of entanglement of very similar textures. Due to the font size difference in this application, high-order PDE transform plays an extremely important role in differentiating modes with slightly different frequency characteristics. In Figure 3(b), the PDE transform successfully suppresses the low-frequency parts and extracts the mode with frequency band mainly corresponding to texts. Such a procedure is similar to the edge detection in a general image processing. Statistical segmentation is then performed on the high-frequency mode. A suitable threshold value is used to cut off the region with low variance and yields only the texts as shown in Figure 3(c).

4.2. Selective Texture Extraction. The present algorithm of selective texture extraction is also tested on one of the most widely used images, the Barbara, in Figure 5. Barbara image is a benchmark test for edge detection and denoising. It contains fine details of different textures such as the table cloth, curtain behind Barbara, scarf, and clothes on her. Distinctions between all these textures and the background are much larger than those among these textures, which leads to the difficulty of selective texture separation and segmentation. Due to the tiny difference between the frequency or spectrum features of different textures mentioned above, a highly frequency-selective separation method is required. However, the conventional Fourier method is not applicable for this case since the textures are entangled in the frequency domain. Moreover, conventional statistical segmentation approaches do not perform well for this case due to the gray-scale entanglement. The present adaptive PDE transform method performs well for the selective texture extraction in the Barbara image. The total texture, or image edge, is extracted from the high-frequency mode of the PDE transform as shown in Figure 5(b). The variance of the local variation is shown in Figure 4, which is calculated and employed for selective texture extraction and separation with appropriate thresholding values. The resulting textures are shown in Figures 5(c)–5(f) which correspond to those of clothes, curtain, and table cloth, respectively. The four textures in Figure 5 are superimposed on the original image for the purpose of a clearer visualization.

In Figure 6, the present adaptive PDE transform is applied to detect a sniper hidden in the forest (Figure 6(a)). The whole image is composed of highly entangled textures. The boundaries between these textures are very challenging to be identified appropriately. In our approach, variance of the local variation is calculated and used for texture separation as in the previous examples. By appropriate thresholding, the variance can be decomposed into three regions corresponding to those of the forest, the tree trunk, and the sniper. The resulting texture modes are shown in Figures 6(b)–6(d).

4.3. Natural Neuron Skeleton Analysis. In the previous introduction to the adaptive PDE transform algorithm and applications, local variation is defined and calculated for the intensity of image mode functions to selectively extract textures beyond the total texture extraction. The selective texture extraction can be generalized to indicate any spatial

Figure 3: Extraction and separation of texts, background watermark, and textures of (a). Shown in the 3(b) and 3(c) are the image mode function and extracted texture using the proposed adaptive PDE transform.

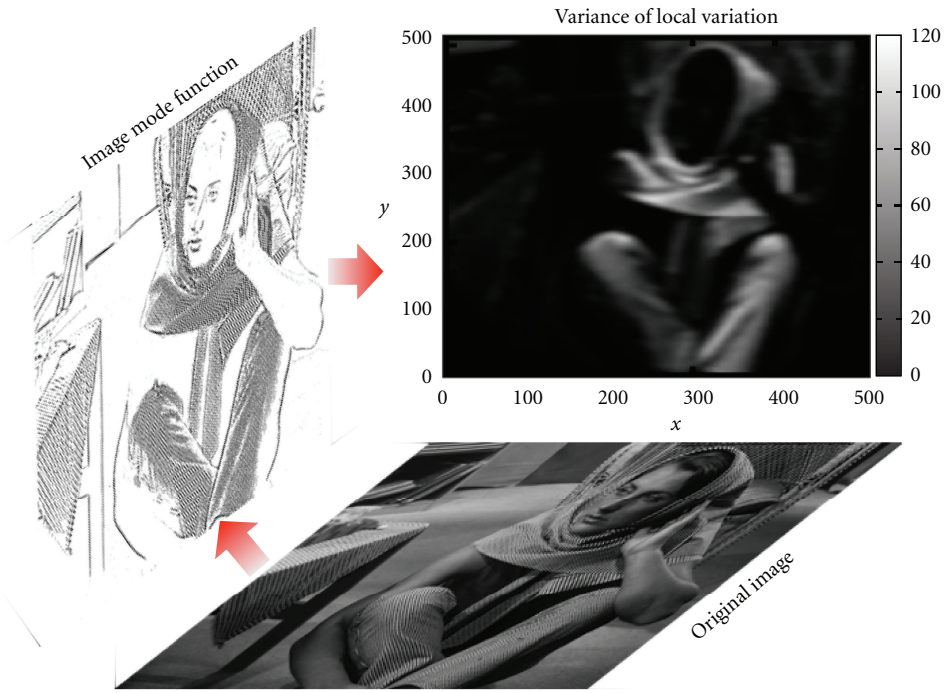
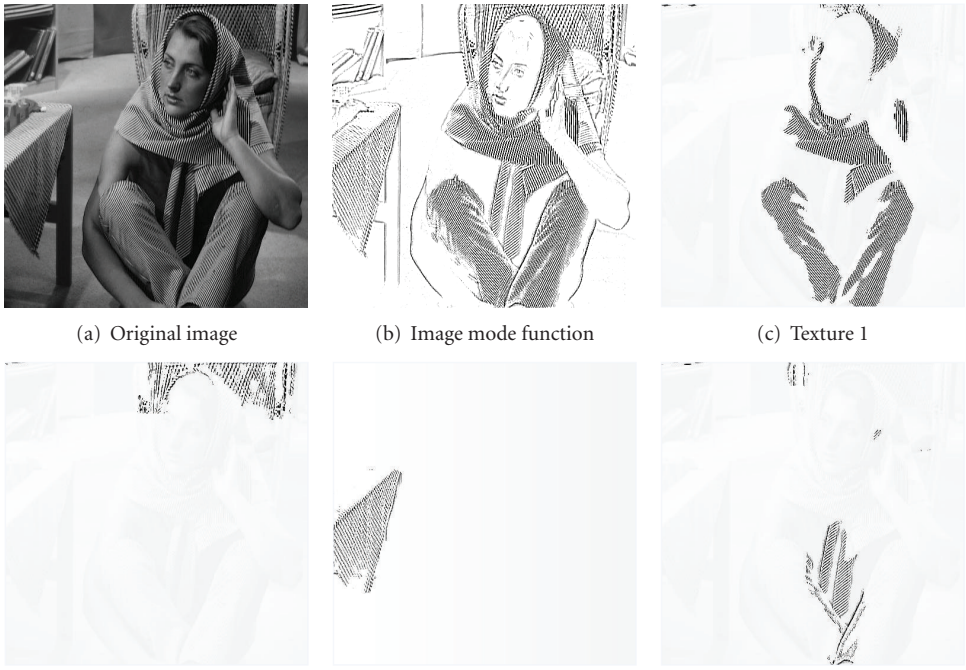


Figure 4: Adaptive PDE transform for selective texture extraction in the Barbara image. The variance of the local variation is shown in the top chart.

parts of the image characterized with specific (and usually functionally important) spatial orientation and/or frequency oscillation, such as different parts in the neuron synapses, brain cells, and retina vasculatures. In Figure 7(a), the image of a typical neuron is shown. With advanced imaging techniques made available, research scientists have been able to obtain more and clearer 2D images and 3D data of various neuron cells and networks, whose study will be important for identifying the relation between phenotype and genotype patterns in physiology and molecular biology. Closely related to the advancement in the experimental imaging techniques, various improved computational image processing techniques have been proposed to better analyze neuron images. Neuron morphology study has become more and more important since the shape and branching of dendrites in neurons are closely related to the structure and functioning of the neuron network. Advancements in both experimental imaging techniques and computational image enhancements have led to better visualization and exploration of neuron morphology [39–45]. In the study of neuron morphology, image processing and segmentation of cultured neuron skeletons provide details of how neuron grow and branches. In this work, we apply the adaptive PDE transform to the study of "natural" neuron skeleton to segment and classify neuron skeletons into desirable classes according to the spatial extension and frequency oscillation of neuron dendrites, very much like the way of dividing a total image texture into several selective fine textures. Such



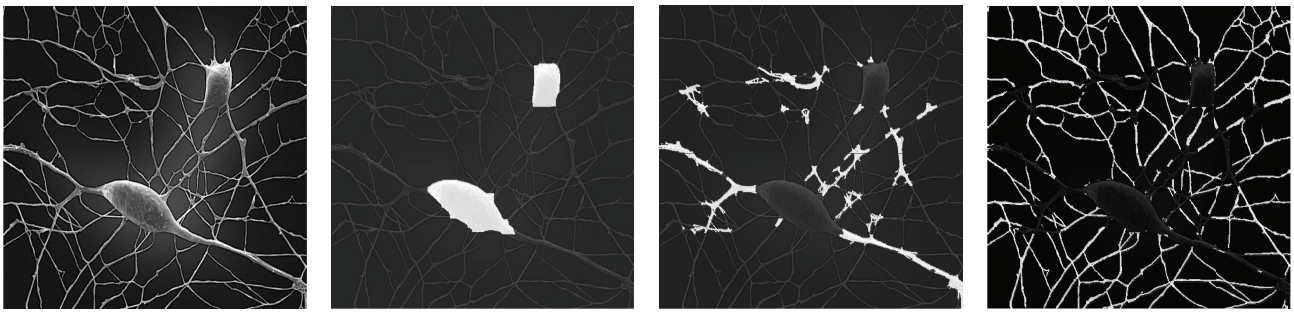
(a) Original image (b) Image mode function (c) Texture 1

(d) Texture 2 (e) Texture 3 (f) Texture 4

Figure 5: PDE transform is applied on (a) to extract edges of all textures into 5(b). Adaptive PDE transform is then applied to extract different textures from 5(b). In 5(c)–5(f), all the textures are superimposed on the original image for better viewing.

(a) Original image (b) Texture 1 (c) Texture 2 (d) Texture 3

Figure 6: Sniper detection by using adaptive PDE transform method. Textures 1, 2, and 3 are, respectively, from the forest, the tree trunk, and the sniper.



(a) Original neuron image (b) Class 1 of the selective neuron skeleton (c) Class 2 of the selective neuron skeleton (d) Class 3 of the selective neuron skeleton

Figure 7: Neuron image classification by using the adaptive PDE transform.

Table 1: Classification of natural neuron skeletons.

Neuron skeleton class	Physical meaning	Percentage of the total neuron surface area
Class 1 shown in Figure 7(b)	Soma (neuron cell body)	22%
Class 2 shown in Figure 7(c)	Major (root of) dendrite	24%
Class 3 shown in Figure 7(d)	Fine (tips of) dendrite	54%

separation and classification enable secondary processing and analysis of neuron morphology, such as the computation of surface areas (for 2D images) or volumes (for 3D data) for different classes of neuron skeletons. Specifically, we aim to separate different parts, or textures, such as soma, dendrites, axon, terminal or lobe, and numerous ramifications, from the neuron imaging as shown in Figures 7(b)–7(d), where three classes of neuron parts are separated according to the spatial extension and frequency oscillation. Surface area of each class is listed in Table 1. Ratios of these surface areas and many other geometric ratios of neuron morphology are related, on both molecular and cellular levels, to the many physiological diseases as well as the classification of neuron synapses.

5. Conclusion

Selective extraction and separation of image textures involving spatial entanglement, gray-scale mixing, and high-frequency overlapping are challenging tasks in image analysis. In this work, we introduce an appropriate adaptation to our earlier partial differential equation (PDE) transform [29] to construct an adaptive PDE transform algorithm. The adaptation is realized via a proper thresholding with the statistical variance of the local variation of image functional mode functions. The present PDE transform enables one to decompose and separate modes with entanglement in both spatial and frequency domains. The proposed method is applied to several challenging benchmark images. Textures of very similar features in the same image are successfully decomposed and separated using the present adaptive PDE transform method.

Acknowledgments

This work was supported in part by NSF Grants CCF-0936830 and DMS-1043034; NIH Grant GM-090208; MSU Competitive Discretionary Funding Program Grant 91-4600.

References

[1] F. Zhang, X. Ye, and W. Liu, "Image decomposition and texture segmentation via sparse representation," *IEEE Signal Processing Letters*, vol. 15, pp. 641–644, 2008.

[2] Y. Dong and J. Ma, "Wavelet-based image texture classification using local energy histograms," *IEEE Signal Processing Letters*, vol. 18, pp. 247–250, 2011.

[3] K. I. Kim, S. H. Park, and H. J. Kim, "Kernel principal component analysis for texture classification," *IEEE Signal Processing Letters*, vol. 8, no. 2, pp. 39–41, 2001.

[4] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[5] H. Wechsler, "Texture analysis—a survey," *Signal Processing*, vol. 2, no. 3, pp. 271–282, 1980.

[6] A. C. Bovik, "Analysis of multichannel narrow-band filters for image texture segmentation," *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 2025–2043, 1991.

[7] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 7–27, 2001.

[8] M. Elad, J. L. Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 340–358, 2005.

[9] A. Khotanzad and R. L. Kashyap, "Feature selection for texture recognition based on image synthesis," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 17, no. 6, pp. 1087–1095, 1987.

[10] V. Caselles, J. M. Morel, G. Sapiro, and A. Tannenbaum, "Introduction to the special issue on partial differential equations and geometry-driven diffusion in image processing and analysis," *IEEE Transactions on Image Processing*, vol. 7, pp. 269–273, 1998.

[11] M. Bertalmío, "Strong-continuation, contrast-invariant inpainting with a third-order optimal PDE," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1934–1938, 2006.

[12] A. Haddad and Y. Meyer, "An improvement of Rudin-Osher-Fatemi model," *Applied and Computational Harmonic Analysis*, vol. 22, no. 3, pp. 319–334, 2007.

[13] J. B. Garnett, T. M. Le, Y. Meyer, and L. A. Vese, "Image decompositions using bounded variation and generalized homogeneous Besov spaces," *Applied and Computational Harmonic Analysis*, vol. 23, no. 1, pp. 25–56, 2007.

[14] J. Gilles and Y. Meyer, "Properties of BV - G structures + textures decomposition models. Application to road detection in satellite images," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2793–2800, 2010.

[15] P. Maurel, J. F. Aujol, and G. Peyre, "Locally parallel texture modeling," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 413–447, 2011.

[16] V. Duval, J. F. Aujol, and L. A. Vese, "Mathematical modeling of textures: application to color image decomposition with a projected gradient algorithm," *Journal of Mathematical Imaging and Vision*, vol. 37, no. 3, pp. 232–248, 2010.

[17] G. R. Cross and A. K. Jain, "Markov random field texture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 1, pp. 25–39, 1983.

[18] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 609–628, 1990.

[19] J. Mao and A. K. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recognition*, vol. 25, no. 2, pp. 173–188, 1992.

[20] A. P. Pentland, "Shading into texture," *Artificial Intelligence*, vol. 29, no. 2, pp. 147–170, 1986.

[21] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1—4, pp. 259–268, 1992.

[22] L. A. Vese and S. J. Osher, "Modeling textures with total variation minimization and oscillating patterns in image processing," *Journal of Scientific Computing*, vol. 19, no. 1-3, pp. 553–572, 2003.

[23] F. Crosby and S. H. Kang, "Multiphase segmentation for 3D flash lidar images," *Journal of Pattern Recognition Research*, vol. 6, pp. 193–200, 2011.

[24] A. Khotanzad and J. Y. Chen, "Unsupervised segmentation of textured images by edge detection in multidimensional feature," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 414–421, 1989.

[25] S. Peleg, J. Naor, R. Hartley, and D. Avnir, "Multiple resolution texture analysis and classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 4, pp. 518–523, 1984.

[26] S. Krishnamachari and R. Chellappa, "Multiresolution Gauss-Markov random field models for texture segmentation," *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 251–267, 1997.

[27] Y. Wang, G. W. Wei, and S. Y. Yang, "Iterative filtering decomposition based on local spectral evolution kernel," *Journal of Scientific Computing*. In press.

[28] Y. Wang, G. W. Wei, and S. Y. Yang, "Mode decomposing evolution equations," *Journal of Scientific Computing*. In press.

[29] Y. Wang, G. W. Wei, and S. Y. Yang, "Partial differential equation transform: variational formulation and fourier analysis," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 27, no. 12, pp. 1996–2020, 2011.

[30] Q. Zheng, S. Y. Yang, and G. W. Wei, "Biomolecular surface construction by PDE transform," *International Journal for Numerical Methods in Biomedical Engineering*. In press.

[31] L. Hu, S. Y. Yang, Q. Zheng, and G. W. Wei, "PDE transform for hyperbolic conservation laws," *SIAM Journal on Scientific Computing*. In press.

[32] A. P. Witkin, "Scale-space filtering," in *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pp. 329–332, 1987.

[33] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[34] G. W. Wei, "Generalized Perona-Malik equation for image restoration," *IEEE Signal Processing Letters*, vol. 6, no. 7, pp. 165–167, 1999.

[35] Y. L. You and M. Kaveh, "Fourth-order partial differential equations for noise removal," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1723–1730, 2000.

[36] M. Lysaker, A. Lundervold, and X. C. Tai, "Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1579–1589, 2003.

[37] G. W. Wei and Y. Q. Jia, "Synchronization-based image edge detection," *Europhysics Letters*, vol. 59, no. 6, pp. 814–819, 2002.

[38] M. Nitzberg and T. Shiota, "Nonlinear image filtering with edge and corner enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 826–833, 1992.

[39] R. A. Graf and I. M. Cooke, "Outgrowth morphology and intracellular calcium of crustacean neurons displaying distinct morphologies in primary culture," *Journal of Neurobiology*, vol. 25, pp. 1558–1569, 1994.

[40] R. Yuste and T. Bonhoeffer, "Morphological changes in dendritic spines associated with long-term synaptic plasticity," *Annual Review of Neuroscience*, vol. 24, pp. 1071–1089, 2001.

[41] J. C. Fiala, B. Allwardt, and K. M. Harris, "Dendritic spines do not split during hippocampal LTP or maturation," *Nature Neuroscience*, vol. 5, no. 4, pp. 297–298, 2002.

[42] R. F. Dacheux, M. F. Chimento, and F. R. Amthor, "Synaptic input to the on-off directionally selective ganglion cell in the rabbit retina," *Journal of Comparative Neurology*, vol. 456, no. 3, pp. 267–278, 2003.

[43] P. J. Broser, R. Schulte, S. Lang et al., "Nonlinear anisotropic diffusion filtering of three-dimensional image data from two-photon microscopy," *Journal of Biomedical Optics*, vol. 9, no. 6, pp. 1253–1264, 2004.

[44] E. Jurrus, M. Hardy, T. Tasdizen et al., "Axon tracking in serial block-face scanning electron microscopy," *Medical Image Analysis*, vol. 13, no. 1, pp. 180–188, 2009.

[45] Y. Livneh and A. Mizrahi, "Long-term changes in the morphology and synaptic distributions of adultborn neurons," *The Journal of Comparative Neurology*, vol. 519, no. 11, pp. 2212–2224, 2011.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2012, Article ID 872783, 17 pages
doi:10.1155/2012/872783

Research Article

Serial FEM/XFEM-Based Update of Preoperative Brain Images Using Intraoperative MRI

Lara M. Vigneron,[1] Ludovic Noels,[2] Simon K. Warfield,[3] Jacques G. Verly,[1] and Pierre A. Robe[4]

[1] *Department of Electrical Engineering and Computer Science, University of Liège, 4000 Liège, Belgium*
[2] *Department of Aerospace and Mechanical Engineering, University of Liège, 4000 Liège, Belgium*
[3] *Computational Radiology Laboratory, Department of Radiology Children's Hospital Boston, Harvard Medical School, Boston, MA 02115, USA*
[4] *Department of Neurosurgery, University of Utrecht Medical Center, 3584 CX Utrecht, The Netherlands*

Correspondence should be addressed to Lara M. Vigneron, lara.vigneron@alumni.ulg.ac.be

Received 30 June 2011; Revised 18 September 2011; Accepted 23 September 2011

Academic Editor: Shan Zhao

Current neuronavigation systems cannot adapt to changing intraoperative conditions over time. To overcome this limitation, we present an experimental end-to-end system capable of updating 3D preoperative images in the presence of brain shift and successive resections. The heart of our system is a nonrigid registration technique using a biomechanical model, driven by the deformations of key surfaces tracked in successive intraoperative images. The biomechanical model is deformed using FEM or XFEM, depending on the type of deformation under consideration, namely, brain shift or resection. We describe the operation of our system on two patient cases, each comprising five intraoperative MR images, and we demonstrate that our approach significantly improves the alignment of nonrigidly registered images.

1. Introduction

Neurosurgery is characterized by the delicate balance between surgical success and potential for devastating side effects. Thanks to multiple technological improvements, the morbidity of neurosurgical interventions has substantially decreased over the last decades, allowing for the resection of previously inoperable lesions. In particular, image-guided neurosurgery (IGNS) devices allow the use of coregistered and fused multimodality 3D images to guide the surgeon's hand and help define preoperatively the boundaries of pathological and predefined functional structures [1]. Meanwhile, new modes of medical imaging have also improved the localization of pathological lesions and their characterization. Medical imaging nowadays includes a wealth of different techniques, such as computed tomography (CT), structural and functional magnetic resonance imaging (sMRI and fMRI), diffusion tensor imaging (DTI), and positron emission tomography (PET). Although the overall accuracy of IGNS is estimated to be 1–2 mm [2], current neuronavigation systems cannot, however, adapt to changing conditions over time. Skull-opening brain shift, brain retraction, cerebrospinal fluid suction, lesion resection, perfusion, and pharmacological manipulation during surgery indeed all alter the 3D morphology of the structures [2–5]. These changes can lead to localization errors that are one order of magnitude larger that IGNS accuracy [1, 2, 6] and may result in incomplete resections or unexpected damage to normal brain. Such inaccuracies could be reduced if one could acquire, throughout surgery, fresh images of the same modalities and quality as the preoperative ones. However, these images are still major challenges. Intraoperative images such as intraoperative MR (iMR) images are—with the exception of a handful surgical facilities—usually acquired using low-field MRI scanners that provide lower resolution and contrast than their preoperative counterparts, and, to this date, several useful imaging modalities, such as PET and possibly MEG, cannot be acquired intraoperatively. One

solution is to "bring over" the high-quality preoperative multimodality images into the intraoperative configuration of the brain using a nonrigid registration technique [7–10]. One category of nonrigid registration techniques uses physics-based models, where landmarks are tracked in successive reduced-quality intraoperative images, and their displacement fields drive the deformation of a biomechanical model. The computation is typically based on the finite element method (FEM). So far, most of the mechanical conditions of the brain cannot be estimated in the operating room, such as the volume of cerebrospinal fluid flowing out of the skull cavity, intercellular fluid volume changes that result from mannitol injection, or changes in blood volume and vessel permeability. The fact that an intraoperative image can provide the knowledge of the current state of the brain after some deformation partly eliminates the need for a complete evaluation of these mechanical conditions. The nonrigid registration technique replaces them with the landmark displacements evaluated from successive intraoperative images.

Using a nonrigid registration technique based on a biomechanical model, three types of brain deformations have been identified that require specific modeling, although they depend on common parameters, such as CSF suction, perfusion, or pharmacological manipulation. The first deformation is the brain shift, which appears at the beginning of surgery with the opening of the skull and dura. The suction or leakage of CSF, as well as the release of intracranial pressure caused by tumor growth, generally cause such shift of the brain (note that in this work, we name "brain shift" the specific shift of the brain that occurs after the opening of the skull and dura, before any other surgical act has happened). The brain also shifts with the two other deformations mentioned below. However, for these deformations, we will consider that the shift is a part of these two deformations. The second deformation is the retraction; when target tissues are located deep inside the brain, the surgeon incises brain tissues and inserts a retractor to spread out the tissues, and to create a path towards the target. The third deformation is the resection, that is, the removal, of lesion tissues. Both resection and retraction de facto imply a cut of tissues. In addition, the resection implies that part of tissues is removed. Three deformations can thus be defined in terms of the two elemental actions that change the topology of the brain: the introduction of a discontinuity and the removal of some tissues.

Most studies of brain deformation based on biomechanical models have focused on shifts (the topology of the brain is not modified), that occurs just after the opening of the skull and dura [11–25]. A good review of these different studies can be found in [24, 26–28]. Resection and retraction are more complex to model than (brain) shift. Until recently, their modeling for the specific application of preoperative image update has received much less attention. One of the difficulty for modeling resection and retraction is that both induce a topological change of the brain because some tissue are cut. A method of mesh adaptation [29–31] or remeshing [32–35] must be used in conjunction with FEM if an accurate representation of the location of the cut, for example, the resection cavity or retraction path, is needed to deform the model. Indeed, FEM cannot directly handle discontinuities that go through the FEs, and requires to realign the discontinuity with FE boundaries.

In the field of fracture mechanics, which studies the growth and propagation of cracks in mechanical parts, some methods were developed to avoid using FEM in conjunction with mesh adaptation or remeshing [36]. The extended finite element method (XFEM or X-FEM) appeared in 1999 [37] and has been the object of considerable research since then [38]. XFEM works by allowing the displacement field to be discontinuous within some FEs of the mesh. The mesh does not have to conform to the discontinuities, so that these can be arbitrarily located with respect to the underlying FE mesh. Because XFEM allows an accurate representation of the discontinuities while avoiding mesh adaption or remeshing, and because of the similarity between cracks in mechanical parts and cuts in tissue, we proposed the use of XFEM for handling cut, resection, and retraction in the updating of preoperative images. This paper presents a complete 3D framework for updating multimodal preoperative images with respect to surgical brain deformations, due to brain shift and successive resections, followed and quantified using iMR images. Our approach is modular, and is applied iteratively each time a new intraoperative image is acquired. We take into account successive deformations based on a linear elastic biomechanical model which is deformed using FEM or XFEM, depending on the type of deformation occurring between the pair of iMR images under consideration, namely, brain shift or resection. Some 2D results were presented in [39]. While some 3D results have already been presented for brain shift [40], and initial 3D results for resection [41] modelings, this paper is the first complete and detailed account of the generalization to 3D of our 2D previous work.

The structure of the paper is as follows. In Section 2, we present the state-of-the-art of resection modeling for preoperative image update. In Section 3, we describe our basic strategy for updating preoperative images based on successive intraoperative images. In Section 4, we give detail about our methods and algorithms. In Section 5, we consider two patient cases that illustrate our approach for handling brain shift followed by successive resections. In Section 6, we validate our results. In Section 7, we conclude and discuss future work.

2. State-of-the-Art

Among studies that take into account resection for preoperative image update, one should distinguish two categories. The first category of studies models brain deformation using two time-point images, the first image being acquired before surgery has started, the second image being acquired after resection. In this category, the methods that existed for a second image showing some brain shift are adapted for a second image showing some resection. However, the resection is not explicitly modeled. The second category of studies models brain deformation using more than two time-point images, and models at least two successive resections.

Among the first category of studies, Hagemann et al. [42] developed a 2D method for modeling brain deformation between a preoperative MR image and a postoperative MR image, the postoperative image showing a complete resection. The 2D mesh of the biomechanical model corresponded to the underlying pixel grid of the 2D image. The biomechanical model included four different linear elastic laws for the skull/skin region, the whole-brain region, the CSF region, and the image background. They computed the correspondence of the skull boundary, the whole-brain region boundary in the neighborhood of the tumor, and the posterior midline between the two images. They also computed the correspondence between the internal tumor region boundary visible in the preoperative image, and the resection cavity boundary visible in the postoperative image, both boundaries corresponding under the assumption that the resection is complete. The displacements fields of these landmarks drove the deformation of the biomechanical model. As a result, the biomechanical model presented high deformation in the tumor region, which is not physically plausible. However, the resection was complete, and, thus, they were not interested by the displacement field of the biomechanical model in the tumor region itself.

Clatz et al. [12] developed a 3D method for modeling the brain deformation between a preoperative MR image and an iMR image, the latter showing partial or complete resection. The biomechanical model was deformed based on a sparse volume displacement field evaluated from the two images, using a block matching algorithm. In their algorithm, blocks of voxels that presented discriminant structures were selected in the preoperative image. The blocks were then matched to blocks in the iMR image using a similarity criterion, for example, a coefficient of correlation. The value of the similarity criterion was used as a value of confidence in the displacement measured by the block matching algorithm. The biomechanical model was then deformed iteratively, driven by the sparse displacement field of the matched blocks, where a block rejection step was included for measured block displacements initially selected but considered as outliers. In the iMR image, a part, or the totality, of the tumor tissues were resected. The blocks were thus selected and matched in only the healthy-brain region of the two images. They tested their algorithm on six patient cases, and used for validation nine landmarks picked up manually in each image. They found a mean and maximum error on displacements of 0.75 mm and 2.5 mm, respectively. The error increased as one approached the tumor region. They explained this phenomenon by the fact that a substantial number of block matchings were rejected in the tumor neighborhood. The deformation of the biomechanical model in the tumor neighborhood was thus essentially governed by the linear elastic law, and the result might show the limitation of this model. Archip et al. [7] also tested the nonrigid registration method presented in [12] on eleven patient cases, and used the 95% Hausdorff distance [43] for evaluating the alignment of the nonrigidly registered images. As a result, they obtained a mean error of 1.82 mm.

Among the second category of studies, Miga et al. [44] simulated two successive resections. They built a linear poroelastic biomechanical model and preoperatively tagged the tetrahedron FEs that were going to be removed to simulate the brain deformation due to successive resections. The modeling of resection was performed in two steps. First, the preoperatively tagged FEs were removed. This consisted in duplicating the nodes at the boundary of the resection cavity. The nodes were actually not eliminated, which avoids the cost of remeshing operations. Second, a boundary condition was applied to the new boundary of the resection cavity, in order to model the relaxation of strain energy, induced by preoperative tumor growth or surgery acts, stored in the resected tissues, and released after their removal. In this approach, the tissue discontinuity was represented as best as possible with a jagged topology defined by the FE facets defining the boundary of the resection cavity. Forest et al. [45, 46] also modeled the removal of tetrahedra in order to model the action of an ultrasonic aspirator in the context of real-time surgery simulation.

Ferrant et al. [13, 47, 48] modeled successive resections based on several time-point iMR images. Between two successive images, they deformed the biomechanical model, in its current state of update, to take into account the (partial) resections(s) that took place between these two images. The modeling of resection was performed in two steps. First, the biomechanical model, in its current state of update, was deformed in accordance with the displacement field of the healthy-brain boundary between the pair of images under consideration. Second, the FEs that fell into the resection cavity in the second image of the pair were removed, while the FEs that laid across the resection-cavity boundary were cut. To ensure the link between the successive deformed configuration of the biomechanical model, their algorithm kept track of the topology modification between FEs and nodes of the mesh before and after the removal of FEs. They tested their algorithm on one patient case including five iMR images (the first two iMR images being used for brain shift modeling), and used for validation thirty-two landmarks picked up manually in each image. They found a mean and maximum error on the displacements of 0.9 ± 0.7 mm (mean $\pm$ standard deviation) and 3.7 mm, respectively. The error increased as one approached the tumor region. They explained this phenomenon by the limited accuracy in the process of picking landmarks in that region, and because the retraction occurring between the second and third images was modeled as a resection, that is, a removal of tissues, even though the tissues were not removed but simply spread out.

The methods described above have been all developed using an FEM-based biomechanical model for intraoperative image registration. Surgical simulation is another research field that broadly uses FEM-based biomechanical model. The objective of a surgical simulator is to provide an interactive manipulation with force feedback of the anatomical part to be operated using various surgical instruments. In order to model a large range surgical procedures, a real-time interactive cutting method should be included in the simulator. Jeřábková and Kuhlen [49] have applied nonlinear XFEM for simulating cut, and have shown that XFEM is successfully efficient for such purpose.

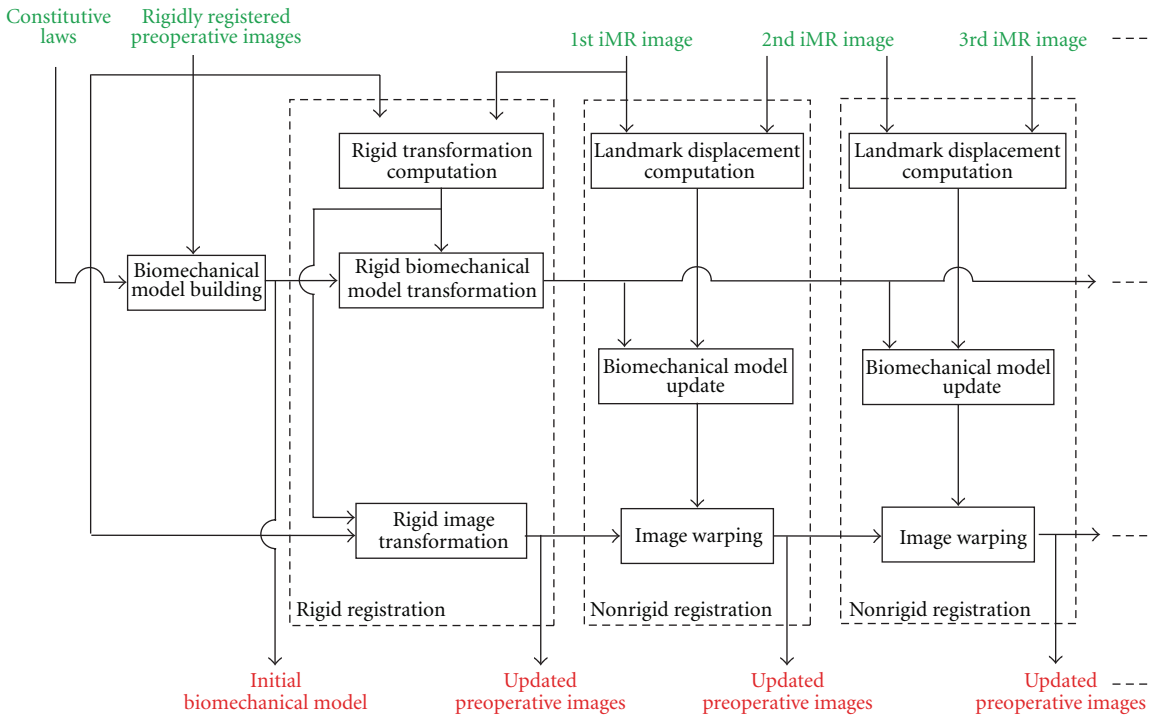


Figure 1: Block diagram of our serial preoperative image-update system dealing with successive brain deformation for a linear formulation.

3. Basic Strategy for Serial Preoperative Image Update

The block diagram of Figure 1 shows our global approach for updating preoperative images using successive iMR images acquired at different critical points during surgery. Although the principles of the approach are quite general, they are tailored for use based on images acquired with a 0.5 Tesla intraoperative GE Signa scanner, which guarantees that the full volume of brain tissues is included in the image field of view. In our present strategy, the preoperative images are updated incrementally. At the end of each update, the preoperative images should be in the best possible alignment with the last iMR image acquired. The actual algorithms and equations used to this end are described in Section 4.

Prior to surgery, a patient-specific biomechanical model is built from the set of preoperative images. Because the patient does not necessarily lie in the same position during the acquisition of each of the preoperative images, one may need to perform a rigid registration (involving translations, rotations, and scales) to bring these images into correspondence, assuming, in first approximation, there is no local, that is, nonuniform, brain deformation between preoperative images. Once the 1st iMR image has been acquired prior to the opening of the skull, the set of registered preoperative images and the biomechanical model are registered to the 1st iMR image via a rigid transformation. In the present situation, it is assumed that the patient's brain imaged in the 1st iMR image has the same physical shape as the brain imaged in the preoperative images (note that in the following, when an iMR image is defined by a number, this number is the index of the iMR image in the series for a specific patient case. The 1st iMR image thus corresponds to the very first iMR image of the series).

As each iMR image is acquired, this new image and the preceding iMR image are used to estimate the deformation of the brain. The update of the preoperative images is done incrementally with each new pair of successive iMR images. For each pair, we proceed as follows. A set of common anatomical landmarks are tracked between the two iMR images. In our approach, we use as landmarks the surfaces of key brain structures. The use of surface structures rather than volume structures [12] seems more appropriate given the reduced-quality of typical intraoperative images, and would be more easily adapted to intraoperative modalities other than iMR, such as iUS. The landmark surface displacement fields resulting from the matching are then applied to the biomechanical model, which is deformed using FEM or XFEM, depending on the type of deformation occurring between the acquisition times of the iMR images in the pair under consideration, namely, brain shift, or resection. The resulting displacement field of the biomechanical model is finally used to warp the set of preoperative images in their current state of updating. This process is repeated with each new acquisition of an iMR image. Note that, for each deformation modeling, the biomechanical model is deformed in accordance with the landmark displacements tracked between the pair of successive iMR images under consideration. Because intraoperative deformation can follow a reverse direction [5], it is important to track the landmarks between the next-to-last and the last acquired iMR images, rather than track the landmarks between the first and the last acquired iMR images.

For the patient cases treated in Section 5, we assume that the brain undergoes relatively small deformations (small strains and small displacements), and we use a linear finite-element formulation in the biomechanical model. A consequence of using this linear formulation (linear elasticity) is that the equations of solid mechanics can be solved based on the initial configuration of the solid.

Actually, knowing the displacement field increment $\Delta u_n^{n+1} = u^{n+1} - u^n$ at the anatomical landmarks between configuration n and increment $n+1$, one can apply this constrained displacement field increment Δu_n^{n+1} to the initial configuration, and the finite element analysis will lead to the deformation tensor increment $\Delta\varepsilon_n^{n+1}$ between the configuration n and $n+1$. The final deformation tensor or the body is thus simply obtained from $\varepsilon^{n+1} = \sum_{k=0}^{n}\varepsilon_k^{k+1}$. Remark that rigorously, the increment of constrained displacement field at the landmark should be applied to the balanced solution of the solid reached after increment n, but as we are using a linear elasticity model, this step can be skipped owing to the superposition principle: if $\sigma^{n+1} = C\varepsilon^{n+1}$, then $\sigma^{n+1} = \sum_{k=0}^{n}\Delta\sigma_k^{k+1} = C\sum_{k=0}^{n}\varepsilon_k^{k+1} = C\varepsilon^{n+1}$, where C is the Hooke tensor. As a summary, with this approach, the process of deformations is modeled as a succession of deformations $\Delta\varepsilon_k^{k+1}$, for example, brain deformation composed of shift followed by successive resections and the current configuration of the brain biomechanical model, after a specific deformation can then be recovered by adding the computed volume displacements for all successive incremental deformations. Remark that this is not a limitation of the method as we could easily extend it to nonlinear model by simply keeping in memory the previous deformed configuration n and adding the constrained displacement field increment Δu_n^{n+1} to compute the new deformed configuration at increment $n+1$, simply this would be less computationally efficient.

Because we use a linear formulation (and, thus, the incremental volume displacement fields can be added to recover the current configuration of the biomechanical model), we could theoretically obtain an identical deformed configuration of the biomechanical model using the two following approaches. The first one would consist of computing and adding the successive incremental deformations of the biomechanical model based on the landmarks tracked between the next-to-last and the last acquired iMR images. The second approach would consist in computing directly the deformed configuration of the biomechanical model based on the landmarks tracked between the first and the last acquired iMR images. However, the landmarks selected to drive the deformation of the biomechanical model vary depending on the type of deformation, namely, brain shift or resection. In addition, part of the biomechanical model is "cut," using XFEM, to model resection. Consequently, we would not get an identical deformed configuration of the biomechanical model by these two approaches. In order to use a maximum of information from the iMR images, we track, as explained for the first approach, the landmarks between the next-to-last and the last acquired iMR images.

The problem of updating preoperative images between more than two critical points during surgery, that is, based on more than two iMR images, is addressed in only a small number of studies. In our previous work [39, 41], and in [13], the biomechanical model was successively deformed, and this was done using a linear formulation. The framework proposed here, where the initial biomechanical model is always used, instead of using it in its successive states of deformation, has the important advantage of using a good quality mesh for each deformation modeling rather than using a mesh whose quality progressively deteriorates with each successive deformation modeling, and which would require remeshing or mesh adaptation for getting back good FE quality.

To summarize, for each deformation, the landmarks are tracked between the two successive iMR images under consideration. Because we use a linear formulation, the displacement fields of these landmarks are applied to the initial, rather than current, configuration of the biomechanical model. The resulting volume displacement field corresponds to the deformation that the brain undergoes between the two iMR images. This volume displacement field is used to deform the preoperative images in their current state of update, that is, registered (at the previous step, if any) to the first iMR image of the pair. After the deformation, the preoperative images are thus in as good as possible registration to the second iMR image of the pair.

In all the rest of this work, we make a simplification of the approach just presented, by using the 1st iMR image as a substitute for the preoperative images. The biomechanical model is thus built based on structures visible in the 1st iMR image, instead of in the preoperative images, and the structures used in the model are limited to the ones visible in the intraoperative image. Except for the rigid registration between the preoperative images, the biomechanical model, and the 1st iMR image, this simplified approach allows us to discuss, illustrate, and test all key aspects of the system. The 1st iMR image is also updated instead of the preoperative images. The above strategy allows us to focus on the main issue of this paper, that is, the estimation and handling of 3D deformations. Even though the issues involved in the update of preoperative images will need to be addressed in a operational image update system, the present strategy of deforming the iMR images remains useful for calibration purpose, even in the operating room.

4. Methods

This section details the different methods that are commonly used for updating preoperative images in presence of brain shift and resection. More specifically, the block diagram of Figure 2(a) shows the building of the biomechanical model from the preoperative images. Specific regions from the preoperative images are segmented, meshed, and assigned appropriate constitutive laws. The block diagram of Figure 2(b) shows, for any pair of successive iMR images, a detailed view of the calculation of the volume displacement field of the initial biomechanical model that corresponds to the deformation that has occurred between the acquisition time of these images.

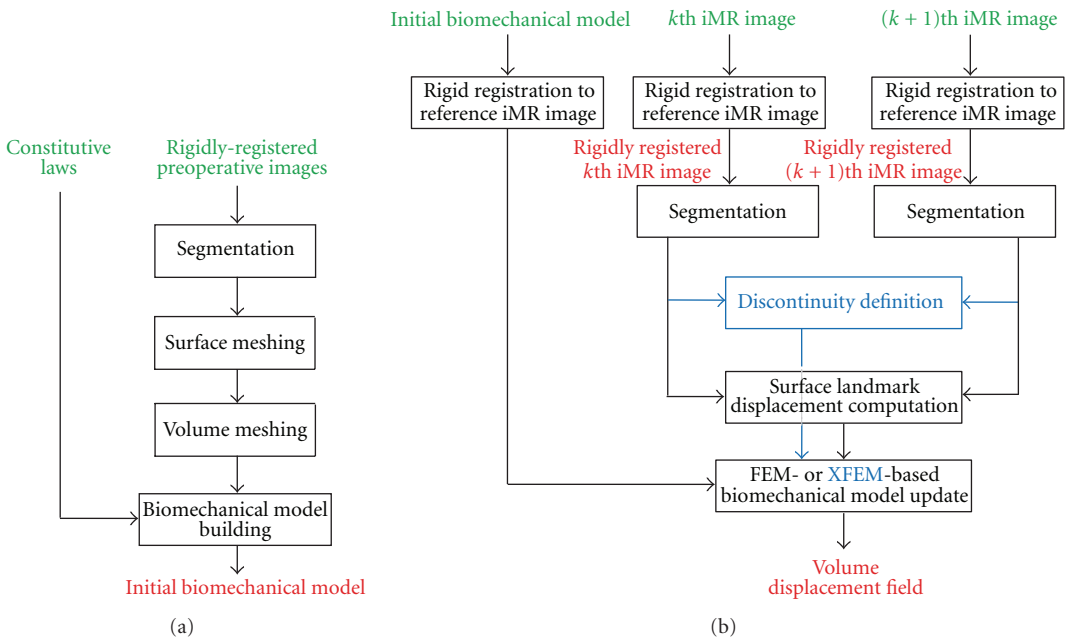


Figure 2: Detailed block diagram of the three subsystems of our serial preoperative image-update system. (a) Building of the biomechanical model from the preoperative images. (b) Calculation of the volume displacement field of the initial biomechanical model using the displacement fields of surface landmarks tracked between a pair of successive iMR images. The updated iMR images are used for validation. For each subsystem, inputs are in green, outputs are in red, and steps related to the definition and use of a discontinuity are in blue.

4.1. Rigid Registration of Intraoperative Images. All along surgery, the patient is lying inside the 0.5 Tesla intraoperative GE Signa scanner. Although the patient's head is fixed, one cannot totally rule out the possibility of slight head motion. iMR images thus have to be rigidly coregistered to take into account this potential rigid motion. The rigid registration that we use is the point-based landmark transform available in VTK (http://www.vtk.org/). The corresponding landmark points are manually selected in the successive iMR images.

4.2. Segmentation of Intraoperative Images. The segmentation of iMR images into specific regions, for example, healthy-brain and tumor regions, is first performed manually using 3D SLICER (http://www.slicer.org/) and then smoothed to minimize the dependance of the results on segmentation roughness. It is clear that performing a manual segmentation in the operating room is not acceptable, and that this process needs to be automated as completely as possible to test the feasibility of our framework online. However, while there exist sophisticated segmentation algorithms that could be used [50–52], in particular for extracting the whole-brain region (skull and external cerebrospinal fluid masked out), the segmentation of the tumor region is still challenging.

4.3. Building of Biomechanical Model. As mentioned above, the biomechanical model is built, in the present context, from the 1st iMR image rather than from the preoperative images. Thanks to the use of XFEM instead of FEM for modeling discontinuities, this biomechanical model can be built offline before the operation starts and does not need to be repeated (through remeshing) during the surgery. With respect to FEM-based approaches, the execution time thus ceases to be a limiting parameter, which is a remarkable advantage of our approach. The object to be meshed is defined as a segmented region from an image. It thus requires specific techniques, and we use the meshing software tool ISOSURF (http://svr-www.eng.cam.ac.uk/~gmt11/software/isosurf/isosurf.html). Our goal is to model the boundaries of healthy-brain and tumor regions as two connected surfaces meshes. However, ISOSURF can only mesh the boundaries of one or several separate regions, and, thus, does not allow one to mesh connected region boundaries with common nodes at their intersections. We thus start by building two separate surfaces meshes that we connect using our own routines based on VTK. We then smooth the two surface meshes using the software SIMMETRIX (http://www.simmetrix.com/). The two connected triangle surfaces are then jointly meshed into a single volume mesh of tetrahedra that conform to the two surface meshes using GMSH (http://www.geuz.org/gmsh/) [53]. Further details on the building of the biomechanical model, in particular the building of the connected surface meshes, can be found in [40]. A linear elastic law is assigned to the biomechanical model, with Young modulus $E = 3000$ Pa and Poisson ratio $\nu = 0.45$ [13]. Because displacements, rather than forces, are applied to the model using a linear formulation, the FEM or XFEM solution is independent of Young modulus E [54].

4.4. Evaluation of Surface Landmark Displacement Fields. We choose as surface landmarks the whole-brain and internal tumor region boundaries. To evaluate the surface deformations of these region boundaries between two iMR images,

we use an active surface algorithm [55, 56]. Because these region boundaries to match must be closed surfaces, we thus use as surface landmarks the whole-brain and healthy-brain region boundaries. The surface deformation of the internal tumor region boundary will be derived from the active surface algorithm of the healthy-brain region boundary. In our active surface algorithm coming from [13, 47, 48], the external forces $\mathbf{F}(\mathbf{x})$ are computed using a gradient descent on a distance map of the region boundary. With such external forces, the active surface algorithm is not able to take correctly into account local rigid motion due, as an example, to lateral or tangential movement depending on the head orientation. For the whole-brain region, any rigid transformation that could have occurred has already been taken into account by the rigid registration of the iMR images (Section 4.1). However, for the healthy-brain region, it can happen that the internal tumor region boundary moves partly in a rigid way. Therefore, the active surface, initialized from the healthy-brain region boundary in the first iMR image, is first locally transformed in a rigid way along the internal tumor region boundary using the iterative closest point transform available in VTK. Then, this resulting surface is deformed using the active surface algorithm as explained above. Further details on the local rigid registration of the healthy-brain region boundary can be found in [40]. Before applying the displacements whole-brain and internal tumor region boundaries to the biomechanical model nodes, the two surface displacement fields are smoothed based on a weighted-distance average, that is, the displacement of each node is averaged with the displacements of its N closest neighbor nodes. This smoothing will make them consistent with each other, and compatible with the volume mesh in order to avoid element flipping, in particular at the intersections between whole-brain and internal tumor region boundaries. Depending on the brain deformation modeling, five to ten neighbor nodes are used.

4.5. FEM- or XFEM-Based Biomechanical Model Deformation. The displacement fields of the surface landmarks are applied to the biomechanical model, which deforms according to the laws of solid mechanics. The equations of solid mechanics are solved using FEM or XFEM, depending upon the type of circumstances, namely, brain shift or resection. We use the FEM-software tool METAFOR (http://metafor.ltas.ulg.ac.be/) developed in our mechanical-engineering department, to which we have added an XFEM module. The initial stress state of the brain is unknown and is thus set to zero for each FEM or XFEM computation, as in [10, 13].

FEM discretizes the solid of interest into a mesh, that is, into a set of FEs interconnected by nodes, and approximates the displacement field $\mathbf{u}(\mathbf{x})$ by the FEM displacement field $\mathbf{u}^{\mathrm{FEM}}(\mathbf{x})$ defined as

$$\mathbf{u}^{\mathrm{FEM}}(\mathbf{x}) = \sum_{i\in I} \varphi_i(\mathbf{x})\mathbf{u}_i, \tag{1}$$

where I is the set of nodes, the $\varphi_i(\mathbf{x})$'s are the nodal shape functions (NSFs), and the $\mathbf{u}_i$'s are the nodal degrees of freedom (DOFs). Each $\varphi_i(\mathbf{x})$ is defined as being continuous on its compact support $\boldsymbol{\omega}_i$, which corresponds to the union of the domains of the FEs connected to node i [57]. In our approach, we use linear NSFs.

FEM requires its displacement field $\mathbf{u}^{\mathrm{FEM}}(\mathbf{x})$ to be continuous over each FE. In contrast, XFEM handles a discontinuity by allowing the displacement field to be discontinuous within FEs [37, 58–60]. Arbitrarily-shaped discontinuities can then be modeled without any remeshing. The XFEM displacement field generalises the FEM displacement field (1) with

$$\mathbf{u}^{\mathrm{XFEM}}(\mathbf{x}) = \sum_{i\in I} \varphi_i(\mathbf{x})\mathbf{u}_i + \sum_{i\in J} \varphi_i(\mathbf{x}) \sum_{j=1}^{n^{E_i}} g_j(\mathbf{x})\mathbf{a}_{ji}. \tag{2}$$

The first term corresponds to the FEM displacement field (1), where I is the set of nodes, the $\varphi_i(\mathbf{x})$'s are the FEM NSFs, and the $\mathbf{u}_i$'s are the nodal FEM DOFs. The heart of XFEM is the "enrichment" that adds a number, n^{E_i}, of DOFs $\mathbf{a}_{ji}$ to each node i of the set J, which is the subset of nodes of I whose support is intersected by the discontinuity of interest. These DOFs are multiplied by the NSFs $\varphi_i(\mathbf{x})$ and the discontinuous functions $g_j(\mathbf{x})$.

The use of specific XFEM enrichment functions $g_j(\mathbf{x})$ for a node $i \in J$ depends on the type of discontinuity, for example, crack, hole, material interface, and so forth, to be modeled. Suppose that our goal is to model a crack, characterized by a discontinuity in the displacement field (as opposed to a material interface for instance, characterized by a discontinuity in the derivative of the displacement field). When the crack fully intersects the support of the node, a simple choice is a piecewise-constant unit function that changes sign at the boundary across the crack, that is, the Heaviside function

$$H(\mathbf{x}) = \begin{cases} 1 & \text{for } (\mathbf{x}-\mathbf{x}^*)\cdot\mathbf{e}_n > 0, \\ -1 & \text{for } (\mathbf{x}-\mathbf{x}^*)\cdot\mathbf{e}_n < 0, \end{cases} \tag{3}$$

where $\mathbf{x}$ is again the position of a point of the solid, $\mathbf{x}^*$ is the position of the point on the crack that is the closest to $\mathbf{x}$, and $\mathbf{e}_n$ is the outward normal to the crack at $\mathbf{x}^*$ [37]. In case of resection deformation, the goal is to model a discontinuity such that the part of tissues corresponding to tissue removed by the resection has no influence on the deformation of the remaining part of the tissues. One is actually interested in the deformation of the remaining part of the tissues only. In that sense, the hole function [61] as the following equation:

$$V(\mathbf{x}) = \begin{cases} 1 & \text{for } (\mathbf{x}-\mathbf{x}^*)\cdot\mathbf{e}_n > 0, \\ 0 & \text{for } (\mathbf{x}-\mathbf{x}^*)\cdot\mathbf{e}_n < 0, \end{cases} \tag{4}$$

could be used as XFEM enrichment function, instead of the Heaviside function, and would be totally sufficient. The results that we would obtain on the remaining part of the tissues would be identical. However, because the Heaviside function is necessary for retraction modeling, we have used the same function for the resection modeling even if it was not strictly necessary.

When minimizing the total deformation energy, the resulting XFEM equations remain sparse and symmetric as for FEM. Whereas FEM requires a remeshing and the duplication of the nodes along the crack to take into account any discontinuity, XFEM requires the identification of the nodes whose support is intersected by the crack and the addition of DOFs: (1) any node whose support is not intersected by the discontinuity remains unaffected and thus possesses three DOFs; (2) any node whose support is fully intersected by the discontinuity is enriched with three Heaviside DOFs and thus possesses six DOFs.

4.6. Evaluation of Deformation Modeling. To qualitatively estimate the similarity between two images, we compare the edges extracted from these images using the Canny edge detector available in ITK (http://www.itk.org/). Indeed, although potentially useful for the sake of comparing methods on a mathematical basis and defining unique correspondences, landmark-based target analysis presents several relevant limitations in the present setting.

(i) Having experts picking landmarks introduces significant intra- and interobserver variability.

(ii) Picking landmark points, as Ferrant et al. [13] did, is rather difficult when it comes to define enough visible landmarks—especially in the tumor region—on the 5 different images (and not 2 images only, as majority of studies focusing on brain shift are using).

(iii) Rather than point targets, linear tumor contours, and limits between structures and potential eloquent structures matter most in the practical case of tumor ablation neurosurgery.

These are the reason why we chose to use the canny edges in order to evaluate the registration. Besides, while it is true that these edges do not necessarily physically correspond between the successive iMR images, these images have been acquired with the same image protocol (MR sequence, voxel size, grayscale value range), which should limit this problem.

To quantitatively estimate the similarity of the two edge maps, we compute the modified Hausdorff distance between the sets of edge points, that is, voxels representing the edges, in these two images. The modified Hausdorff distance $\mathcal{H}(A,B)$ [43] between two sets of points A and B is defined as

$$\mathcal{H}(A,B) = \max(h(A,B), h(B,A)) \quad \text{with}$$
$$h(A,B) = \frac{1}{N_a}\sum_{a\in A} d(a,B), \tag{5}$$

where the directed Hausdorff distance $h(A,B)$ is a measure of the distance of the point set A to the point set B, N_a is the number of points in set A, and $d(a,B)$ is the distance of point $a \in A$ to the closest point in B, that is, $d(a,B) = \min_{b\in B}\|a-b\|$, where $\|a-b\|$ is the Euclidean distance. The directed Hausdorff distance $h(A,B)$ thus computes the average distance of points of A to points of B. The averaging minimizes the effects of outlier points, for example, due to image noise. The value of the modified Hausdorff distance $\mathcal{H}(A,B)$ increases with the amount of difference between the two sets of edges points. In the following, we denote by $\mathcal{H}(I_a, I_b)$ the modified Hausdorff distance of the edges extracted from the whole-brain region of the images I_a and I_b, that is, with the skull and external cerebrospinal fluid masked out from them.

5. Results

In this section, we apply our methods, respectively, of brain shift and resection (iMR images are acquired with the 0.5 Tesla intraoperative GE Signa scanner of the Brigham and Women's Hospital, Boston, USA. iMR image size is $256 \times 256 \times 60$ voxels, and voxel size is $0.9375 \times 0.9375 \times 2.5$ mm). All computations are done off-line. Two patient cases, each including five iMR images, are treated to illustrate our modeling and brain shift followed by successive resections. In both cases, the 1st iMR image was acquired prior to the opening of the skull; the 2nd iMR image was acquired after the opening of the skull and dura, and shows some brain shift; the 3rd, 4th, and 5th iMR images were acquired after successive resections. The modelings of brain shift, 1st, 2nd, and 3rd resection are performed using different techniques, as detailed below. Except where otherwise noted, the following discussion applies to both patient cases (the result of each deformation modeling is shown for the two patient cases at the end of Section 5.2.3).

5.1. Modeling of Brain Shift. To model brain shift based on the 1st and 2nd iMR images, we estimate the surface displacement fields of the whole-brain region boundary and the internal tumor region boundary from the two iMR images. No tissue discontinuity is involved in the brain shift deformation, so the biomechanical model is deformed using FEM. This results in the volume displacement field of the biomechanical model, which is illustrated in Figure 3 for the first patient case. This volume displacement field is used to warp the part of the 1st iMR image corresponding to the whole-brain region.

5.2. Modeling of Successive Resections. In the following sections, the three successive resections are modeled separately, because they require different types of processing. Nevertheless, a common remark can be made for each resection modeling. Matching two region boundaries to get a displacement field makes sense only if they correspond to the same physical entity. Once the resection has started, we can no longer rely on the entirety of the whole-brain region boundary, since a part of it is now missing. For modeling the successive resections, we thus evaluate the displacement field for the boundary of the healthy-brain region only.

5.2.1. Modeling of 1st Resection. The 1st resection occurs between the times the 2nd and 3rd iMR images are acquired. However, since the corresponding removal of tissues is most likely accompanied by deformation, one cannot exactly determine what tissue is removed based just on the two iMR

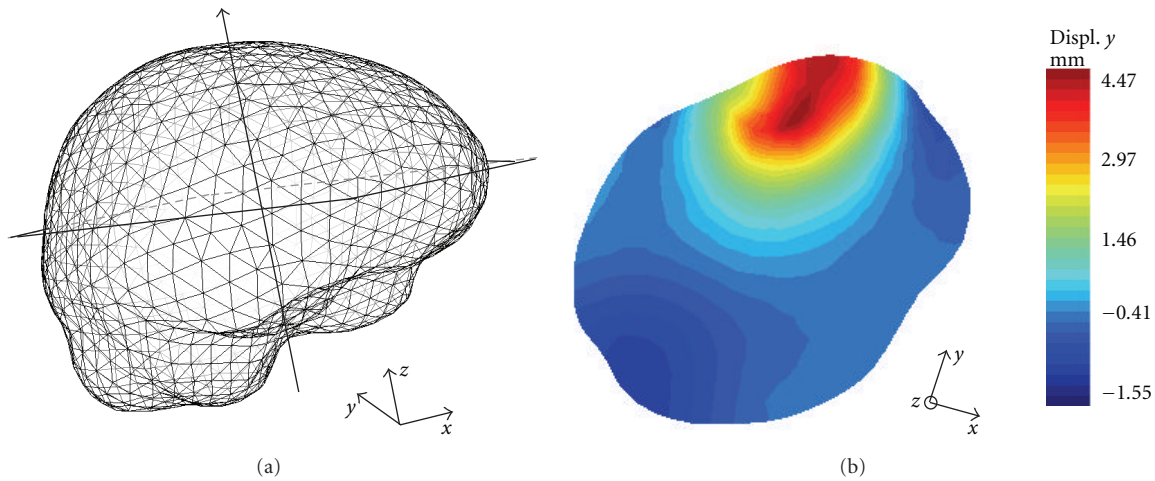


Figure 3: Result of the biomechanical model deformation for brain shift modeling (first patient case). (a) External surface mesh of the biomechanical model with the location of the slice considered in (b). (b) Selected slice of the biomechanical model with color levels corresponding to the displacements along the y-axis, which is the main direction of the brain shift for this patient case.

images. We thus decided to model the 1st resection by still relying on the displacement fields of key surfaces, here the healthy-brain region boundary, to deform the biomechanical model. This indeed appears to be the only reliable information concerning the deformation due to resection that we can extract from the 2nd and 3rd iMR images. Consequently, we do not model explicitly the removal of tissue, but we model directly the deformation resulting from it, without introducing any tissue discontinuity. Using the surface displacement field of the healthy-brain region boundary, we compute the deformation of the biomechanical model via FEM. Then, using the resulting volume displacement field, we warp the part of the 2nd iMR image corresponding to the whole-brain region, in the same way as we did in the case of for brain shift. The image resulting from the 1st resection modeling is now registered to the 3rd iMR image, except outside of the healthy-brain region boundary, that is, for the tumor region. Finally, we alter the resulting image to reflect the effect of resection. For this, we assign the background color to the voxels corresponding to the resected tissue volume "absent" in the 3rd iMR image.

5.2.2. Modeling of 2nd Resection. The significant feature of the 2nd resection is that some tissue has already been removed by the 1st resection, which means that this tissue cannot have any physical influence on subsequent brain deformations because it does not "exist" anymore. Consequently, the 1st resection must be reflected in the biomechanical model. Recall that the biomechanical model has been deformed to model the brain shift and the 1st resection and is thus registered to the 3rd iMR image. So, using the 3rd iMR image, we can define the boundary of the 1st resection, that is, the tissue discontinuity to include in the deformed biomechanical model (Figures 4(a) and 4(b)). We then enrich the nodes whose supports are intersected by the discontinuity with Heaviside DOFs. Consequently, when the XFEM-based biomechanical model deforms, the part corresponding to tissue removed by the 1st resection has no influence on the deformation of the remaining part of the brain. For the first patient case illustrated in Figure 4, the tetrahedron mesh consists of 3,317 nodes, which corresponds to 9,951 FEM DOFs. Enrichment adds 873 Heaviside DOFs.

As for the modeling of the 1st resection, the biomechanical model is deformed in accordance with the displacement field of the healthy-brain region boundary evaluated from the 3rd and 4th iMR images. Figure 4(d) shows the deformed mesh, result of the XFEM computation. The bottom part of the mesh, representing the tissue remaining after the 1st resection, has been deformed according to the displacement field of the healthy-brain region boundary, while the top part, representing the tissue removed by the 1st resection, has been subjected to a translation, but only for visualization purposes. Even though the mesh is displayed as two separate parts, it is, in fact, a single entity. Indeed, a main feature of XFEM is its ability to handle the effect of a discontinuity without modifying the underlying mesh, that is, without remeshing. For modeling the 2nd resection, the edges of FEs straddling the discontinuity have been made discontinuous and their nodes moved apart. Using the XFEM volume displacement field, we warp the part of the 3rd iMR image corresponding to the whole-brain region. The resulting image is then masked out with the whole-brain region segmented out from the 4th iMR image.

5.2.3. Modeling of 3rd Resection. One significant feature of the procedure described for modeling the 2nd resection is that it can be applied repetitively for each subsequent resection visible on successive iMR images, no matter how many there are. The modeling of the 3rd resection is thus identical to the modeling of the 2nd resection. The tissue discontinuity due to the 2nd resection is defined from the 4th iMR image, and used to appropriately enrich the nodes of the biomechanical model. Then, this biomechanical model is

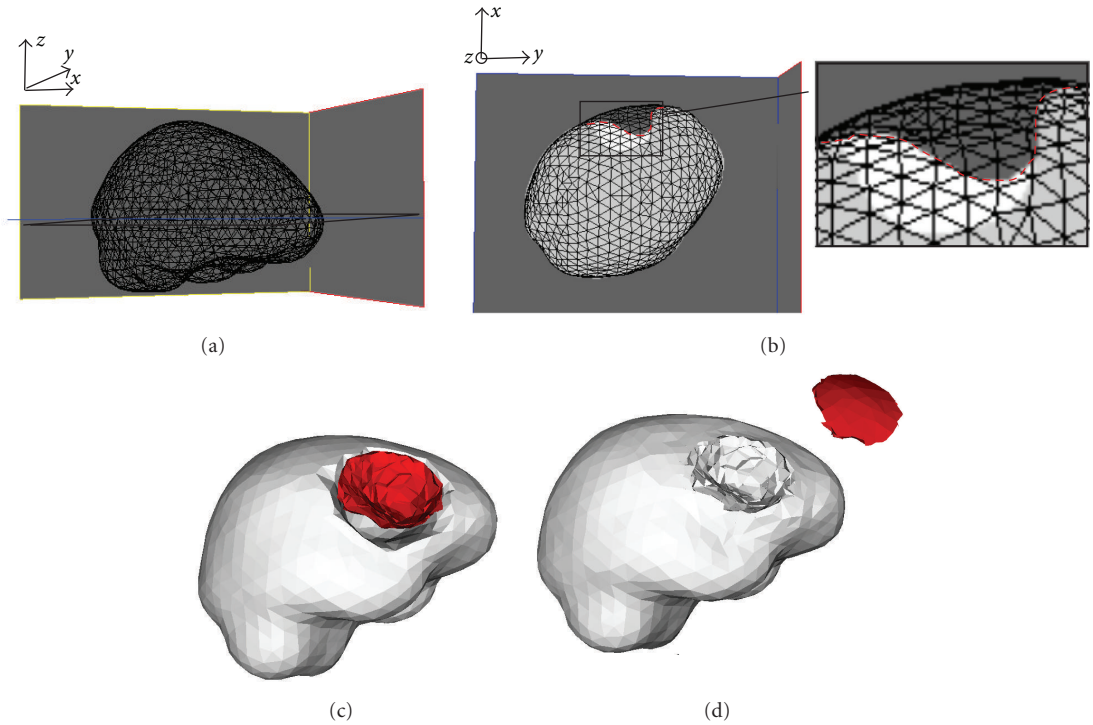


Figure 4: Definition of tissue discontinuity for 2nd resection modeling (first patient case). (a) External surface mesh (of the biomechanical model) with the location of the slice considered in (b). (b) External surface mesh superposed to the healthy-brain region (light gray) and tumor region (white) segmented out from the 3rd iMR image. This superposition allows one to define the tissue discontinuity (red boundary). (c) Surface meshes describing the healthy-brain region boundary (gray) and the tissue discontinuity (red). This tissue discontinuity gives an idea of the part of tumor tissue that was removed by the 1st resection. The gap that appears "between" the gray and red surfaces corresponds to the remaining tumor tissues. (d) Final mesh resulting from the modeling of the 2nd resection using XFEM. The tetrahedra that were added to display separately the two parts of the mesh are only for visualization purposes.

deformed using XFEM, in accordance with the displacement field of the healthy-brain region boundary evaluated from the 4th and 5th iMR images.

For the first patient case, a simplification for the modeling of the 3rd resection can be made because, by the time the 5th iMR image is acquired, the resection is complete. This means that we only need to compute the volume displacement field of the healthy-brain region. Since we apply displacements exactly to the boundary of the healthy-brain region, the results obtained with FEM and XFEM will be identical. Using the FEM (for the first patient case) or XFEM (for the second patient case) volume displacement field, we warp the part of the 4th iMR image corresponding to the whole-brain region. The resulting image is then masked out with the whole-brain region segmented out from the 5th iMR image.

Figures 5 and 6 show the results of warping the iMR images, as well as the edges extracted from them, after brain shift and each successive resection modeling for the two patient cases.

5.2.4. Comparison of FEM and XFEM for Modeling of Resection. As explained in Section 5.2.3, since we apply displacements exactly to the boundary of the healthy-brain region, the results obtained with FEM and XFEM are identical in the healthy-brain region. One can deduce that using XFEM for modeling resection is interesting when the neurosurgeon needs to have an accurate displacement field of the remaining tumor tissues. In this case, it is interesting to evaluate the impact of using FEM, instead of XFEM, to model the resection as if no resection was performed before. Using FEM for modeling resection is equivalent to ignoring the presence of resection on intraoperative images. To illustrate the comparison between FEM and XFEM results, we choose the 3rd resection modeling of the second patient case. Indeed, it is the deformation with remaining tumor tissues that shows the largest magnitude, and, thus, that is likely to give a maximum difference between the two computations. Figure 7 compares the results obtained using FEM and XFEM. The healthy-brain and tumor regions segmented out from the 4th and 5th iMR images are respectively shown in Figures 7(a) and 7(b). The volume displacement fields of the biomechanical model using XFEM and FEM are respectively shown in Figures 7(c) and 7(d). The part of the 4th iMR image corresponding to the whole-brain region is warped, first with the volume displacement field obtained via FEM, and then with that obtained via XFEM. The difference between the two warped images is shown in Figure 7(e). As expected, there is a visible difference in the remaining tumor tissue. However, the difference between the

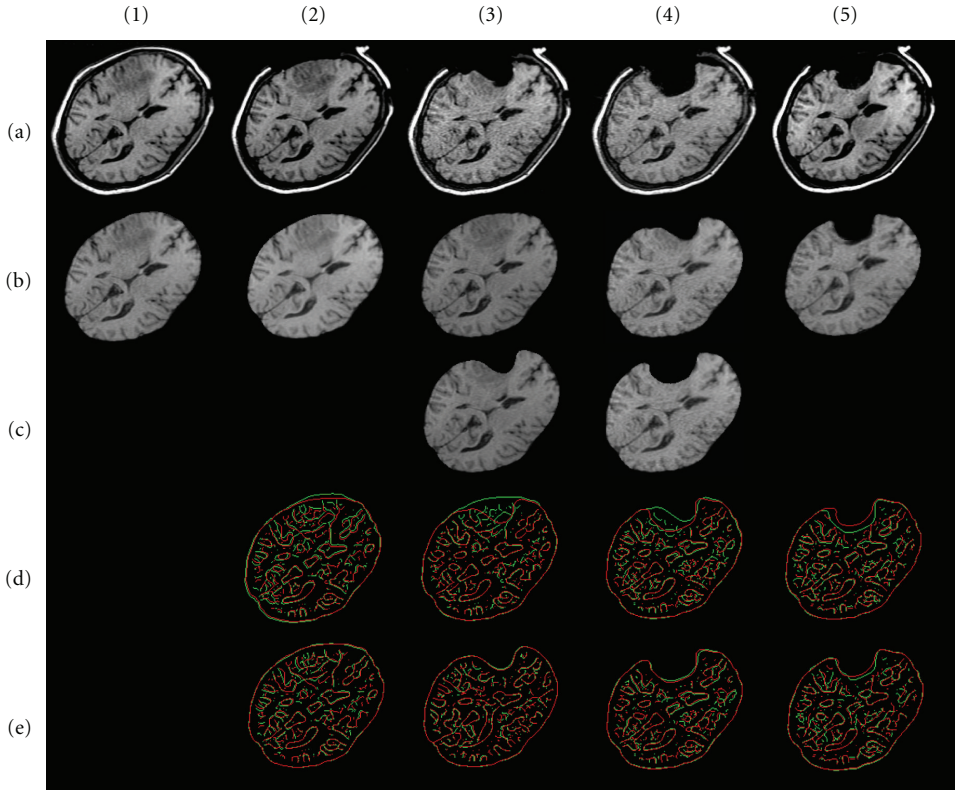


Figure 5: First patient case. (a) Sequence of five input iMR images rigidly registered to the first one. (1b) Whole-brain region extracted from (1a). (2b) Deformation of (1b) computed using FEM for brain shift modeling. (3b) Deformation of whole-brain region extracted from (2a) computed using FEM for 1st resection modeling. (3c) Masking of (3b) with whole-brain region segmented from the 3rd iMR image (3a). (4b) Deformation of whole-brain region extracted from (3a) computed using XFEM for 2nd resection modeling. (4c) Masking of (4b) with whole-brain region segmented from 4th iMR image (4a). (5b) Deformation of whole-brain region extracted from (4a) computed using FEM for 3rd resection modeling. (d) Juxtaposition of Canny edges of images rigidly registered. The edges of the first (second) image of the pair under consideration are in green (red). (e) Ditto for (d) when images are nonrigidly registered.

two volume displacement fields is smaller than the image resolution (although the difference between the two volume displacement fields is smaller than the image resolution, the difference between the images resulting of the warping using these two volume displacement fields is nonzero. This is explained by the fact that the (gray) value of each voxel of the warped image is defined as a weighted-value of voxels of the original image. The weights are defined based on the overlapping ratio of the voxel of the warped image, with voxels (determined using the volume displacement field) of the original image). In addition, the deformed 4th iMR images, using the XFEM- and the FEM-based deformations of the biomechanical model, show the same similarity, computed based on the modified Hausdorff distance, with the 5th iMR.

Two reasons explain that the differences between the FEM and XFEM results are so small. First, the brain deformation itself due to the 3rd resection is small, and, thus, it is expected to obtain small differences between the two resulting brain deformations. Second, in the case the remaining tumor tissues are close to the healthy-brain region boundary, it implies that they are close to the boundary where surface displacement fields are applied to drive the deformation of the biomechanical model. This proximity decreases the influence of the modeling of already resected tissues with XFEM. Although this comparison between FEM and XFEM should be done on more patient cases, we suggest that, in first approximation, FEM could be used for modeling resection cases with small brain deformations. Nevertheless, the presentation of the successive resections using XFEM shows the generality of our framework, and details how XFEM is implemented. Note that in Section 6 devoted to validation, the warped images are the ones deformed with XFEM.

6. Validation

For each deformation modeling based on a pair (I_k, I_{k+1}) of two successive iMR images that are already rigidly registered, we compare the similarity between these I_k and I_{k+1} images, as well as the similarity between the I_k^w and I_{k+1} images, where I_k^w is the result of warping I_k. This gives us an estimate of how well we are able to capture, and compensate for, the local deformations between I_k and I_{k+1}. The goal of the nonrigid registration is, however, to deform the preoperative images. By warping I_k for each deformation modeling, we do not take into account the fact that an error of alignment after

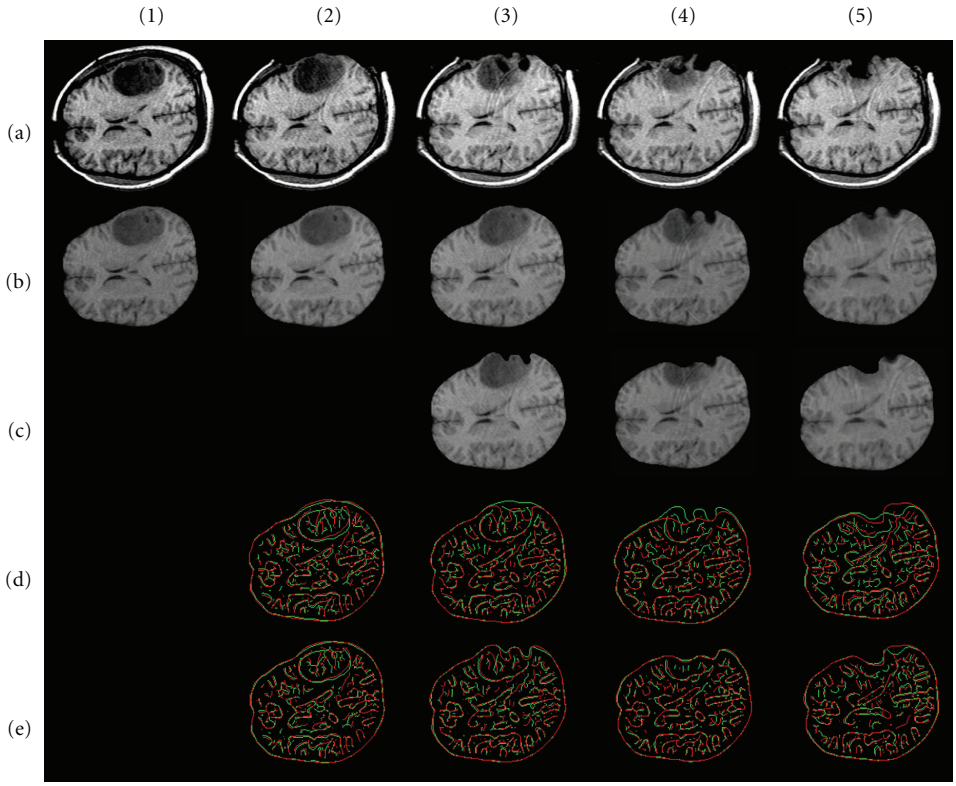


Figure 6: Second patient case. (a) Sequence of five input iMR images rigidly registered to the first one. (1b) Whole-brain region extracted from (1a). (2b) Deformation of (1b) computed using FEM for brain shift modeling. (3b) Deformation of whole-brain region extracted from (2a) computed using FEM for 1st resection modeling. (3c) Masking of (3b) with whole-brain region segmented from the 3rd iMR image (3a). (4b) Deformation of whole-brain region extracted from (3a) computed using XFEM for 2nd resection modeling. (4c) Masking of (4b) with whole-brain region segmented from 4th iMR image (4a). (5b) Deformation of whole-brain region extracted from (4a) computed using XFEM for 3rd resection modeling. (5c) Masking of (5b) with whole-brain region segmented from 5th iMR image (5a). (d) Juxtaposition of Canny edges of images rigidly registered. The edges of the first (second) image of the pair under consideration are in green (red). (e) Ditto for (d) when images are nonrigidly registered.

each deformation modeling could propagate and amplify through the successive deformation modelings. To evaluate the effect of this error amplification on the results, we also perform the required succession of warpings on I_1, and we denote the resulting image by $I^w_{1,k}$. We then compare, for each deformation modeling, the similarity between I_1 and I_{k+1}, together with the similarity between $I^w_{1,k}$ and I_{k+1}. This allows one to evaluate the propagation, that is, the amplification, of alignment error on the results. The modified Hausdorff distance computed for each pair of iMR images are given in Tables 1 and 2.

Table 1 shows, for each deformation modeling based on a pair (I_k, I_{k+1}) of two successive iMR images, the values of the modified Hausdorff distances $\mathcal{H}(I_k, I_{k+1})$ and $\mathcal{H}(I^w_k, I_{k+1})$. These values are computed using the Canny edges extracted from the pair of images (I_k, I_{k+1}) (Figures 5 (d) and 6 (d)) and (I^w_k, I_{k+1}) (Figures 5 (e) and 6 (e)). We observe that the values for the images nonrigidly registered are relatively constant, that is, ~1 mm, for each deformation modeling. Six out of eight deformation modelings give smaller modified Hausdorff distances when the iMR images are (rigidly and subsequently) nonrigidly registered. However, the modified Hausdorff distance increases for the 3rd resection modeling of the first patient case, as well as for the brain shift modeling of the second patient case. To understand if the nonrigid registration is responsible for the increase of the misalignment of the two iMR images everywhere in the whole-brain region, or if this effect is localized, we compute the modified Hausdorff distance in the region and neighborhood of the tumor only (volume region that extents by 25 mm the tumor region segmented in I_1 for both patient cases). The modified Hausdorff distance decreases from $\mathcal{H}(I_4, I_5) = 1.70$ mm to $\mathcal{H}(I^w_4, I_5) = 1.37$ mm for the first patient case, while it decreases from $\mathcal{H}(I_1, I_2) = 1.36$ mm to $\mathcal{H}(I^w_1, I_2) = 1.28$ mm for the second patient case. This indicates that the nonrigid registration enhances the alignment of the two iMR images within the tumor region and its neighborhood, which is in fact the location requiring the best modeling accuracy. This behavior could be explained by the fact that a maximum of information from the iMR images is used in this region, that is, one or two (in case of brain shift modeling) surface displacement fields are applied around it. The increase of misalignment elsewhere in the brain volume could be explained by two reasons. First,

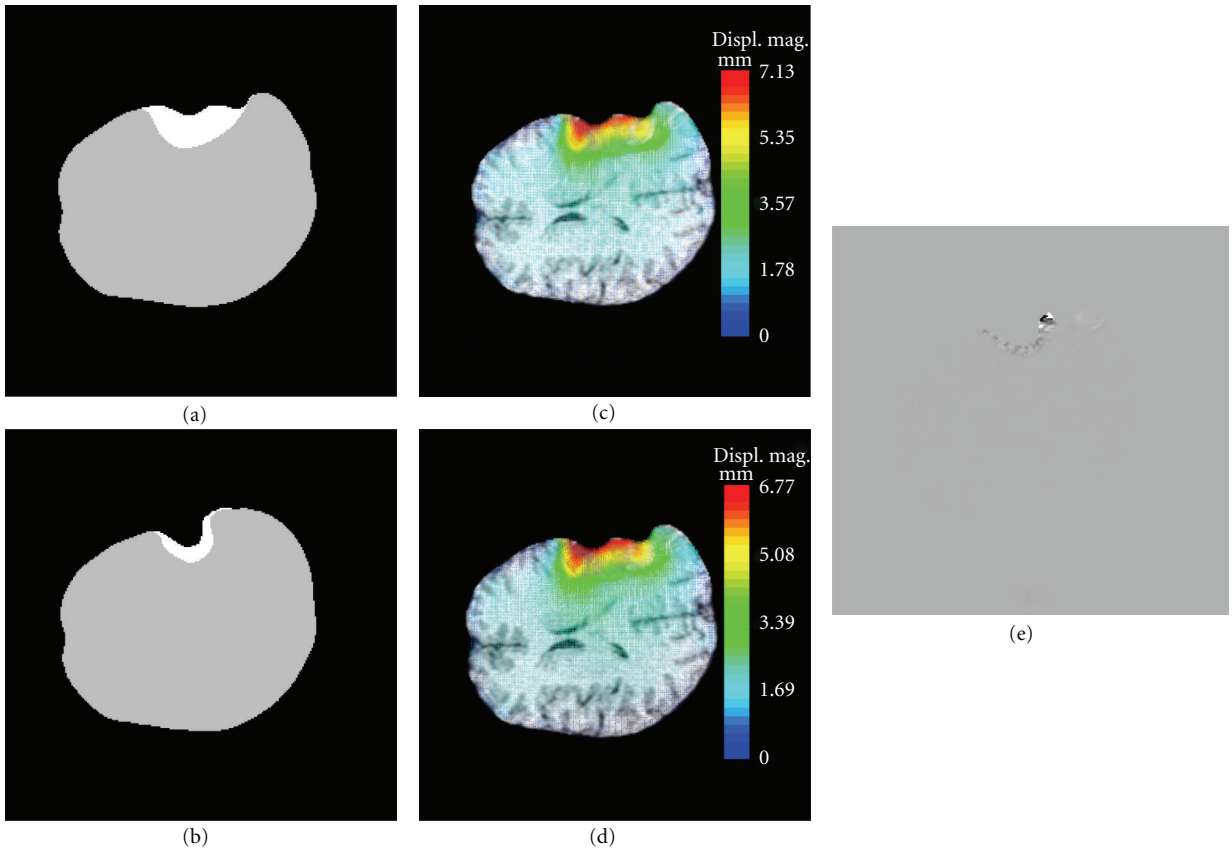


Figure 7: Difference of results using XFEM and FEM for 3rd resection modeling (second patient case). (a) Healthy-brain (gray) and tumor (white) regions segmented out from the 4th iMR image. (b) Healthy-brain and tumor regions segmented out from the 5th iMR image. (c) Volume displacement field of biomechanical model using XFEM. The part of tissue falling within the resection cavity is modeled as being removed. Color levels correspond to the magnitude of the displacement field. (d) Same as (c), but for FEM. The part of tissue falling within the resection cavity is present in the deformation modeling even though it no longer exists. Difference of magnitude between volume displacement fields using XFEM (c) and FEM (d) does not exceed 0.36 mm. (e) Difference in the warping of the part of the 4th iMR image corresponding to the whole-brain region using XFEM and FEM.

Table 1: Values of $\mathcal{H}(I_k, I_{k+1})$ and $\mathcal{H}(I_k^w, I_{k+1})$, $k = 1, \ldots, 4$, for each deformation modeling based on a pair (I_k, I_{k+1}) of two successive iMR images. First value gives measure of similarity of images rigidly registered, while second value gives measure of similarity of images both rigidly, and (subsequently) nonrigidly registered. For each I_k, only the whole-brain region is taken into account for edge extraction.

Modified Hausdorff distance (mm) between edges extracted from two iMR images		Brain shift		1st resection		2nd resection		3rd resection	
		$\mathcal{H}(I_1, I_2)$	$\mathcal{H}(I_1^w, I_2)$	$\mathcal{H}(I_2, I_3)$	$\mathcal{H}(I_2^w, I_3)$	$\mathcal{H}(I_3, I_4)$	$\mathcal{H}(I_3^w, I_4)$	$\mathcal{H}(I_4, I_5)$	$\mathcal{H}(I_4^w, I_5)$
Patient 1	Whole-brain region	1.24	1.07	0.84	0.69	1.10	0.97	0.96	0.97
	Tumor region and neighborhood							1.70	1.37
Patient 2	Whole-brain region	1.01	1.04	1.07	1.04	1.02	0.93	1.23	1.06
	Tumor region and neighborhood	1.36	1.28						

the landmarks tracked from the iMR images are surfaces. As a consequence, the nonrigid registration is expected to give better results near the tracked surfaces than far from them in the volume [13]. Second, the volume displacement field strongly depends on the constitutive laws. The volume misalignment could point out the need for better parameters values and/or other constitutive laws.

Table 2 shows, for each deformation modeling based on a pair (I_k, I_{k+1}) of two successive iMR images, the values of the modified Hausdorff distances $\mathcal{H}(I_1, I_{k+1})$ and $\mathcal{H}(I_{1,k}^w, I_{k+1})$. So far, IGNS systems allow one to rigidly register preoperative and successive iMR images. $\mathcal{H}(I_1, I_{k+1})$ thus represents the navigation accuracy that we can obtain with an IGNS system at the present time. The comparison

Table 2: Values of $\mathcal{H}(I_1, I_{k+1})$ and $\mathcal{H}(I^w_{1,k}, I_{k+1})$, $k = 1, \ldots, 4$, for each deformation modeling based on a pair (I_k, I_{k+1}) of two successive iMR images. In contrast with Table 1, I_1 is successively warped, rather than I_k, for each deformation modeling. First value gives measure of similarity of images rigidly registered, while second value gives measure of similarity of images both rigidly and (subsequently) nonrigidly registered. For each I_k, only the whole-brain region is taken into account for edge extraction.

Modified Hausdorff distance (mm) between edges extracted from two iMR images		Brain shift		1st resection		2nd resection		3rd resection	
		$\mathcal{H}(I_1, I_2)$	$\mathcal{H}(I^w_1, I_2)$	$\mathcal{H}(I_1, I_3)$	$\mathcal{H}(I^w_{1,2}, I_3)$	$\mathcal{H}(I_1, I_4)$	$\mathcal{H}(I^w_{1,3}, I_4)$	$\mathcal{H}(I_1, I_5)$	$\mathcal{H}(I^w_{1,4}, I_5)$
Patient 1	Whole-brain region	1.24	1.07	1.50	1.21	1.80	1.31	1.78	1.38
	Tumor region and neighborhood								
Patient 2	Whole-brain region	1.01	1.04	1.10	1.16	1.36	1.31	1.68	1.42
	Tumor region and neighborhood	1.36	1.28	1.76	1.44				

of $\mathcal{H}(I_1, I_{k+1})$ with $\mathcal{H}(I^w_{1,k}, I_{k+1})$ gives the improvement that could be practically achieved in the alignment with our approach. As expected, Table 2 shows that the IGNS accuracy decreases through the successive deformations. Indeed, the modified Hausdorff distance increases from $\mathcal{H}(I_1, I_2) = 1.24$ mm to $\mathcal{H}(I_1, I_5) = 1.78$ mm for the first patient case, and from $\mathcal{H}(I_1, I_2) = 1.01$ mm to $\mathcal{H}(I_1, I_5) = 1.68$ mm for the second patient case. Six out of eight deformation modelings give smaller modified Hausdorff distances when the iMR images are nonrigidly registered. To understand if the modified Hausdorff distance increases everywhere in the whole-brain region for the brain shift and 1st resection modeling of the second patient case, we compute the modified Hausdorff distance in the neighborhood of the tumor region (in the same way as explained for Table 1), and observe the improvement of the alignment within the tumor region and its neighborhood. As opposed to the values of the modified Hausdorff distances in Table 1, the values for the images nonrigidly registered in Table 2 increase through the successive resection modeling. This amplification error is due to the fact that, after having modeled brain deformation between a pair of iMR images, the deformed biomechanical model is not in perfect alignment with the second image of the pair. Since, for the subsequent deformation modeling, the surface landmarks are initialized based on the deformed biomechanical model, this can thus ampliy a misregistration error.

7. Conclusions and Future Work

We developed a complete 3D framework for serial preoperative image update in the presence of brain shift followed by successive resections. The results were presented for two patient cases, each containing five iMR images. The nonrigid registration technique used an homogeneous linear elastic biomechanical model, driven by the deformations of whole-brain and internal tumor region boundaries for brain shift modeling, and healthy-brain region boundary for resection modelings, tracked between successive iMR images. The biomechanical model was deformed using FEM for brain shift modeling, and FEM or XFEM for resection modeling, depending upon whether some brain tissues were previously resected or not. We showed that our approach was modular, and could be applied each time a new iMR image is acquired.

We used a linear formulation to characterize the deformation of the brains of both patients because the brains underwent relatively small deformations and displacements. While nonlinear biomechanical models have proven effective to decrease—yet do not abolish—the inaccuracies of FEM-based modeling methods of large brain deformations, the deformations observed in our patients during surgery remained moderate (4–7 mm), thus reducing the theoretical benefit of using nonlinear models. This allowed us to use simpler linear models and focus on the added value of XFEM to simultaneously account for surgical deformations, namely, shift and resection. Using a linear formulation implied that, for each new deformation modeling, one could use the initial configuration rather than the last-deformed configuration of the biomechanical model. This had the important advantage of using a good quality mesh for each deformation modeling rather than using a mesh whose quality progressively degraded with each successive deformation modeling. This also had the advantage that we did no longer need to reconnect the deformed mesh for each new XFEM calculation, which was one drawback of our previous method, presented in [39, 41], where the biomechanical model was successively deformed. We also showed how XFEM could handle a discontinuity for modeling resection without any remeshing or mesh adaptation while the representation of the discontinuity remained accurate, that is, the representation of the discontinuity was not based on a jagged topology using FE facets. XFEM thus also avoided making the mesh resolution richer in the neighborhood of the resection-cavity boundary for improving the accuracy of the representation of the discontinuity for that purpose only.

We showed that our nonrigid registration technique improved the alignment of the successive iMR images for most of the deformation modeling of both patient cases. When our nonrigid registration failed, it still improved the alignment locally, that is, within the tumor region and its neighborhood. We tested the explicit modeling of the lateral ventricles' region with a soft, compressible law in addition to the whole-brain region law used in the homogeneous biomechanical model. However, it did not have a significative impact on the result.

In addition to the validation that is usually performed for successive deformation modelings, that is, validation between pairs of successive intraoperative images, shown

in Table 1 of Section 6 or in the work of Ferrant et al. [13, 47, 48], we also evaluated the fact that an error of alignment after each deformation modeling could propagate and amplify through the successive deformation modelings. As a result, shown in Table 2 of Section 6, we showed that our approach suffered from the propagation of misregistration through the successive deformation modelings. We expected that this was due, at least partly, to the algorithms used to evaluate intraoperative surface displacements fields from the whole-brain and healthy-brain region boundaries. These boundaries were first manually segmented, and then smoothed. The surface displacement fields were computed using active surface algorithms, and smoothed to make them compatible with the biomechanical model. Because of these two smoothings, the deformed biomechanical model was likely to not be in a perfect alignment with the iMR image to which it was registered. Because the surface displacement fields evaluated for the next deformation modeling were initialized based on the deformed biomechanical model, we expected to observe an amplification of the misregistration, which was confirmed by our quantitative evaluation. At the present time though, commercial IGNS systems allow one to register preoperative images and successive iMR images, but in a rigid way only. Consequently, although the effect of error amplification exists, our technique still enhances the current capabilities of commercial IGNS systems.

Future work on modeling of brain shift followed by successive resection is required in five main areas. First, the effect of error amplification through the successive brain deformation modelings calls for further research. Consequently, the segmentation, and the subsequent smoothing, as well as the evaluation of surface displacement fields, should be improved to minimize the effect of error amplification. Second, further research is required to include additional structures in the biomechanical model in general, and to study the best way to include the lateral ventricles in particular. The use of a poroelastic model in order to model the cerebrospinal fluid filling the ventricles could be considered [17, 18]. Third, the fact that we use iMR images could be further exploited. Indeed, these images provide volume information (rather than surface information only), are of good quality in comparison to other intraoperative modalities, and possess a field of view that includes the full volume of brain tissues (for the 0.5 Tesla GE Signa scanner). These images thus allow one to evaluate what, and how, new structures of the brain could be used, to enhance the modeling of brain shift. Some regions, for example, the lateral ventricles' region, could be extracted from the two iMR images, and used as surface landmarks to drive the deformation of the biomechanical model [13, 62]. Indeed, the workflow presented in this paper has the advantage of being easily adaptable. In case the tumor region would not be visible (enough) on the iMR images, these new structures, easier to segment, could also adequately replace the tumor for driving the deformation. Fourth, our global approach should no longer be based on the 1st iMR image used as a substitute for preoperative images, but on the preoperative images themselves. Fifth, we should implement, for the surgery cases involving large deformations of the brain, a nonlinear formulation of FEM [63, 64], and, particularly, a nonlinear formulation of XFEM, which is the subject of recent research [65, 66].

Acknowledgments

Pierre A. Robe is a research associate of the National Fund for Scientific Research of Belgium. Simon K. Warfield is supported in part by a Translational Research Program Award from Children's Hospital Boston. This investigation was supported in part by FIRS #4319 of the University Hospital of Liège, Belgium, the "Fonds Léon Fredericq", Liège, Belgium, and NIH Grants R01 RR021885, R01 EB008015, R03 EB008680, and R01 LM010033.

References

[1] R. M. Comeau, A. F. Sadikot, A. Fenster, and T. M. Peters, "Intraoperative ultrasound for guidance and tissue shift correction in image-guided neurosurgery," *Medical Physics*, vol. 27, no. 4, pp. 787–800, 2000.

[2] P. Hastreiter, C. Rezk-Salama, G. Soza et al., "Strategies for brain shift evaluation," *Medical Image Analysis*, vol. 8, no. 4, pp. 447–464, 2004.

[3] H. Dickhaus, K. A. Ganser, A. Staubert et al., "Quantification of brain shift effects by MR-imaging," in *Proceedings of the 1997 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 491–494, November 1997.

[4] T. Hartkens, D. L. G. Hill, A. D. Castellano-Smith et al., "Measurement and analysis of brain deformation during neurosurgery," *IEEE Transactions on Medical Imaging*, vol. 22, no. 1, pp. 82–92, 2003.

[5] A. Nabavi, B. P. McL, D. T. Gering et al., "Serial intraoperative magnetic resonance imaging of brain shift," *Neurosurgery*, vol. 48, no. 4, pp. 787–798, 2001.

[6] M. A. Audette, K. Siddiqi, F. P. Ferrie, and T. M. Peters, "An integrated range-sensing, segmentation and registration framework for the characterization of intra-surgical brain deformations in image-guided surgery," *Computer Vision and Image Understanding*, vol. 89, no. 2-3, pp. 226–251, 2003.

[7] N. Archip, O. Clatz, S. Whalen et al., "Non-rigid alignment of pre-operative MRI, fMRI, and DT-MRI with intra-operative MRI for enhanced visualization and navigation in image-guided neurosurgery," *NeuroImage*, vol. 35, no. 2, pp. 609–624, 2007.

[8] P. Jannin, X. Morandi, O. J. Fleig et al., "Integration of sulcal and functional information for multimodal neuronavigation," *Journal of Neurosurgery*, vol. 96, no. 4, pp. 713–723, 2002.

[9] A. Tei, F. Talos, A. Bharatha et al., "Tracking volumetric brain deformations during image guided neurosurgery," in *VISIM: Information Retrieval and Exploration in Large Medical Image Collections, in Conjunction with (MICCAI '01)*, 2001.

[10] S. K. Warfield, F. Talos, C. Kemper et al., "Augmenting intraoperative MRI with preoperative fMRI and DTI by biomechanical simulation of brain deformation," in *Medical Imaging 2003: Visualization, Image-Guided Procedures, and Display*, vol. 5029 of *Proceedings of SPIE*, San Diego, Calif, USA, February 2003.

[11] M. Bucki and Y. Payan, "Framework and bio-mechanical model for a per-operative imageguided neuronavigator including 'brain-shift' compensation," in *Proceedings of the 2nd*

Workshop on Computer Assisted Diagnosis and Surgery, March 2006.

[12] O. Clatz, H. Delingette, I. F. Talos et al., "Robust nonrigid registration to capture brain shift from intraoperative MRI," *IEEE Transactions on Medical Imaging*, vol. 24, no. 11, pp. 1417–1427, 2005.

[13] M. Ferrant, A. Nabavi, B. Macq et al., "Serial registration of intraoperative MR images of the brain," *Medical Image Analysis*, vol. 6, no. 4, pp. 337–359, 2002.

[14] A. Hagemann, K. Rohr, and H. S. Stiehl, "Coupling of fluid and elastic models for biomechanical simulations of brain deformations using FEM," *Medical Image Analysis*, vol. 6, no. 4, pp. 375–388, 2002.

[15] C. A. Kemper, I.-F. Talos, A. Golby et al., "An anisotropic material model for image guided neurosurgery," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '04)*, vol. 3217, pp. 267–275, 2004.

[16] K. E. Lunn, K. D. Paulsen, D. W. Roberts, F. E. Kennedy, A. Hartov, and J. D. West, "Displacement estimation with co-registered ultrasound for image guided neurosurgery: a quantitative in vivo porcine study," *IEEE Transactions on Medical Imaging*, vol. 22, no. 11, pp. 1358–1368, 2003.

[17] M. I. Miga, K. D. Paulsen, P. J. Hoopes, F. E. Kennedy, A. Hartov, and D. W. Roberts, "In vivo quantification of a homogeneous brain deformation model for updating preoperative images during surgery," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 2, pp. 266–273, 2000.

[18] K. D. Paulsen, M. I. Miga, F. E. Kennedy, P. Jack Hoopes, A. Hartov, and D. W. Roberts, "A computational model for tracking subsurface tissue deformation during stereotactic neurosurgery," *IEEE Transactions on Biomedical Engineering*, vol. 46, no. 2, pp. 213–225, 1999.

[19] J. Rexilius, S. K. Warfield, C. R. G. Guttmann et al., "A novel nonrigid registration algorithm and applications," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '01)*, pp. 923–931, 2001.

[20] D. W. Roberts, M. I. Miga, A. Hartov et al., "Intraoperatively updated neuroimaging using brain modeling and sparse data," *Neurosurgery*, vol. 45, no. 5, pp. 1199–1207, 1999.

[21] O. Škrinjar, A. Nabavi, and J. Duncan, "Model-driven brain shift compensation," *Medical Image Analysis*, vol. 6, no. 4, pp. 361–373, 2002.

[22] S. Ji, A. Hartov, D. Roberts, and K. Paulsen, "Data assimilation using a gradient descent method for estimation of intraoperative brain deformation," *Medical Image Analysis*, vol. 13, no. 5, pp. 744–756, 2009.

[23] H. Sun, K. E. Lunn, H. Farid et al., "Stereopsis-guided brain shift compensation," *IEEE Transactions on Medical Imaging*, vol. 24, no. 8, pp. 1039–1052, 2005.

[24] S. K. Warfield, S. J. Haker, I. F. Talos et al., "Capturing intraoperative deformations: research experience at Brigham and Women's hospital," *Medical Image Analysis*, vol. 9, no. 2, pp. 145–162, 2005.

[25] A. Wittek, K. Miller, R. Kikinis, and S. K. Warfield, "Patient-specific model of brain deformation: application to medical image registration," *Journal of Biomechanics*, vol. 40, no. 4, pp. 919–929, 2007.

[26] J. Cohen-Adad, P. Paul, X. Morandi, and P. Jannin, "Knowledge modeling in image-guided neurosurgery: application in understanding intraoperative brain shift," in *Proceedings of the Medical Imaging: Visualization, Image-Guided Procedures and Display*, vol. 6141 of *Proceedings of SPIE*, 2006.

[27] S. K. Kyriacou, A. Mohamed, K. Miller, and S. Neff, "Brain mechanics for neurosurgery: modeling issues," *Biomechanics and Modeling in Mechanobiology*, vol. 1, no. 2, pp. 151–164, 2002.

[28] K. Miller, A. Wittek, G. Joldes et al., "Modelling brain deformations for computer-integrated neurosurgery," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 26, no. 1, pp. 117–138, 2010.

[29] H.-W. Nienhuys and F. A. van der Stappen, "A surgery simulation supporting cuts and finite element deformation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '01)*, vol. 2208, pp. 153–160, 2001.

[30] D. Serby, M. Harders, and G. Székely, "A new approach to cutting into finite element models," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '01)*, vol. 2208, pp. 425–433, 2001.

[31] D. Steinemann, M. A. Harders, M. Gross, and G. Székely, "Hybrid cutting of deformable solids," in *Proceedings of the IEEE Computer Society Conference on Virtual Reality*, pp. 35–42, 2006.

[32] D. Bielser, P. Glardon, M. Teschner, and M. Gross, "A state machine for real-time cutting of tetrahedral meshes," *Graphical Models*, vol. 66, no. 6, pp. 398–417, 2004.

[33] F. Ganovelli, P. Cignoni, C. Montani, and R. Scopigno, "Multiresolution model for soft objects supporting interactive cuts and lacerations," *Computer Graphics Forum*, vol. 19, no. 3, pp. 271–281, 2000.

[34] A. Mor and T. Kanade, "Modifying soft tissue models: progressive cutting with minimal new element creation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '00)*, vol. 1935, pp. 598–607, 2000.

[35] H.-W. Nienhuys, *Cutting in Deformable Objects*, Ph.D. thesis, Institute for Information and Computing Sciences, Utrecht University,, 2003.

[36] M. Duflot and H. Nguyen-Dang, "A meshless method with enriched weight functions for fatigue crack growth," *International Journal for Numerical Methods in Engineering*, vol. 59, no. 14, pp. 1945–1961, 2004.

[37] N. Moës, J. Dolbow, and T. Belytschko, "A finite element method for crack growth without remeshing," *International Journal for Numerical Methods in Engineering*, vol. 46, no. 1, pp. 131–150, 1999.

[38] Y. Abdelaziz and A. Hamouine, "A survey of the extended finite element," *Computers and Structures*, vol. 86, no. 11-12, pp. 1141–1151, 2008.

[39] L. M. Vigneron, M. P. Duflot, P. A. Robe, S. K. Warfield, and J. G. Verly, "2D XFEM-based modeling of retraction and successive resections for preoperative image update," *Computer Aided Surgery*, vol. 14, no. 1–3, pp. 1–20, 2009.

[40] L. M. Vigneron, R. C. Boman, J. P. Ponthot, P. A. Robe, S. K. Warfield, and J. G. Verly, "Enhanced FEM-based modeling of brain shift deformation in image-guided neurosurgery," *Journal of Computational and Applied Mathematics*, vol. 234, no. 7, pp. 2046–2053, 2010.

[41] L. M. Vigneron, R. C. Boman, J.-P. Ponthot, P. A. Robe, S. K. Warfield, and J. G. Verly, "3D FEM/XFEM-based biomechanical brain modeling for preoperative image update," in *Workshop "Computational Biomechanics for Medicine II", at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '07)*, 2007.

[42] A. Hagemann, K. Rohr, H. S. Stiehl, U. Spetzger, and J. M. Gilsbach, "Biomechanical modeling of the human head for physically based, nonrigid image registration," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 875–884, 1999.

[43] M.-P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proceedings of the 12th International Conference on Pattern Recognition (IAPR '94)*, pp. 566–568, 1994.

[44] M. I. Miga, D. W. Roberts, F. E. Kennedy et al., "Modeling of retraction and resection for intraoperative updating of images," *Neurosurgery*, vol. 49, no. 1, pp. 75–85, 2001.

[45] C. Forest, H. Delingette, and N. Ayache, "Cutting simulation of manifold volumetric meshes," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '02)*, vol. 2488, pp. 235–244, 2002.

[46] C. Forest, H. Delingette, and N. Ayache, "Removing tetrahedra from manifold tetrahedralisation: application to real-time surgical simulation," *Medical Image Analysis*, vol. 9, no. 2, pp. 113–122, 2005.

[47] M. Ferrant, *Physics-based deformable modeling of volumes and surfaces for medical image registration, segmentation and visualization*, Ph.D. thesis, Université Catholique de Louvain, Telecommunications Laboratory, Louvain-la-Neuve, Belgium, 2001.

[48] M. Ferrant, A. Nabavi, B. Macq, F. A. Jolesz, R. Kikinis, and S. K. Warfield, "Registration of 3-D intraoperative MR images of the brain using a finite-element biomechanical model," *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1384–1397, 2001.

[49] L. Jeřábková and T. Kuhlen, "Stable cutting of deformable objects in virtual environments using XFEM," *IEEE Computer Graphics and Applications*, vol. 29, no. 2, pp. 61–71, 2009.

[50] A. A. Devalkeneer, P. A. Robe, J. G. Verly, and C. L. M. Phillips, "Generalized expectation-maximization segmentation of brain MR images," in *Medical Imaging 2006: Image Processing*, vol. 6144 of *Proceedings of SPIE*, San Diego, Calif, USA, February 2006.

[51] J.-F. Mangin, V. Frouin, I. Bloch, J. Régis, and J. López-Krahe, "From 3D magnetic resonance images to structural representations of the cortex topography using topology preserving deformations," *Journal of Mathematical Imaging and Vision*, vol. 5, no. 4, pp. 297–318, 1995.

[52] S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive, template moderated, spatially varying statistical classification," *Medical Image Analysis*, vol. 4, no. 1, pp. 43–55, 2000.

[53] C. Geuzaine and J.-F. Remacle, "Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities," *International Journal for Numerical Methods in Engineering*, vol. 79, no. 11, pp. 1309–1331, 2009.

[54] K. Miller and A. Wittek, "Neuroimage registration as displacement—zero traction problem of solid mechanics," in *Workshop "Computational Biomechanics for Medicine" at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '06)*, 2006.

[55] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[56] C. Xu, D. L. Pham, and J. L. Prince, "Medical image segmentation using deformable models," in *Handbook of Medical Imaging*, J. M. Fitzpatrick and M. Sonka, Eds., vol. 2, chapter: Medical Image Processing and Analysis, pp. 129–174, SPIE Press, Bellingham, Wash, USA, 2000.

[57] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method*, Butterworth Heinemann, Woburn, Mass, USA, 2000.

[58] N. Sukumar, N. Moës, B. Moran, and T. Belytschko, "Extended finite element method for three-dimensional crack modelling," *International Journal for Numerical Methods in Engineering*, vol. 48, no. 11, pp. 1549–1570, 2000.

[59] J. E. Dolbow, *An extended finite element method with discontinuous enrichment for applied mechanics*, Ph.D. thesis, Northwestern University, 1999.

[60] N. Sukumar and J. H. Prévost, "Modeling quasi-static crack growth with the extended finite element method Part I: computer implementation," *International Journal of Solids and Structures*, vol. 40, no. 26, pp. 7513–7537, 2003.

[61] N. Sukumar, D. L. Chopp, N. Moës, and T. Belytschko, "Modeling holes and inclusions by level sets in the extended finite-element method," *Computer Methods in Applied Mechanics and Engineering*, vol. 190, no. 46-47, pp. 6183–6200, 2001.

[62] K. E. Lunn, K. D. Paulsen, D. R. Lynch, D. W. Roberts, F. E. Kennedy, and A. Hartov, "Assimilating intraoperative data with brain shift modeling using the adjoint equations," *Medical Image Analysis*, vol. 9, no. 3, pp. 281–293, 2005.

[63] T. Belytschko, W. K. Liu, and B. Moran, *Nonlinear Finite Elements for Continua and Structures*, John Wiley & Sons, New York, NY, USA, 2000.

[64] K. Miller, G. Joldes, and A. Wittek, "New finite element algorithm for surgical simulation," in *Proceedings of the 2nd Workshop on Computer Assisted Diagnosis and Surgery*, 2006.

[65] J. E. Dolbow and A. Devan, "Enrichment of enhanced assumed strain approximations for representing strong discontinuities: addressing volumetric incompressibility and the discontinuous patch test," *International Journal for Numerical Methods in Engineering*, vol. 59, no. 1, pp. 47–67, 2004.

[66] G. Legrain, N. Moës, and E. Verron, "Stress analysis around crack tips in finite strain problems using the eXtended finite element method," *International Journal for Numerical Methods in Engineering*, vol. 63, no. 2, pp. 290–314, 2005.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2012, Article ID 327198, 10 pages
doi:10.1155/2012/327198

Research Article

Fracture Detection in Traumatic Pelvic CT Images

Jie Wu,[1] Pavani Davuluri,[2] Kevin R. Ward,[3,4] Charles Cockrell,[4,5] Rosalyn Hobson,[2,4] and Kayvan Najarian[1,4]

[1] *Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA*
[2] *Department of Electrical and Computer Engineering, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA*
[3] *Department of Emergency Medicine, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA*
[4] *Virginia Commonwealth University Reanimation Engineering Science Center (VCURES), Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA*
[5] *Department of Radiology, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA*

Correspondence should be addressed to Jie Wu, wuj6@vcu.edu

Received 2 July 2011; Revised 30 September 2011; Accepted 30 September 2011

Academic Editor: Shan Zhao

Fracture detection in pelvic bones is vital for patient diagnostic decisions and treatment planning in traumatic pelvic injuries. Manual detection of bone fracture from computed tomography (CT) images is very challenging due to low resolution of the images and the complex pelvic structures. Automated fracture detection from segmented bones can significantly help physicians analyze pelvic CT images and detect the severity of injuries in a very short period. This paper presents an automated hierarchical algorithm for bone fracture detection in pelvic CT scans using adaptive windowing, boundary tracing, and wavelet transform while incorporating anatomical information. Fracture detection is performed on the basis of the results of prior pelvic bone segmentation via our registered active shape model (RASM). The results are promising and show that the method is capable of detecting fractures accurately.

1. Introduction

Pelvic fractures are high energy injuries that constitute a major cause of death in trauma patients. According to the Centers for Disease Control and Prevention (CDC), trauma injury kills more people between the ages of 1 and 44 than any other disease or illness. Among different types of trauma with a high impact on the lives of Americans, traumatic pelvic injuries, caused mainly by sports, falls, and motor vehicle accidents, contribute to a large number of mortalities every year [1, 2]. Traumatic pelvic injuries and associated complications, such as severe hemorrhage multiple organ dysfunction syndrome (MODS), result in the mortality rate from 8.6% to 50% [3]. When combined with other injuries in the body, for instance, the abdomen, the chance of mortality is even higher [4]. In general, a pelvic fracture can be associated hemorrhage, neurologic injury, vascular injury, and organ damage, as all of the vital structures run through pelvis. Pain and impaired mobility are normally the results of nerve and internal organ damage associated with the pelvic fracture [5–7].

Patient data, in particular, medical images such as computed tomography (CT) images, contain a significant amount of information, and it is crucial for physicians to make diagnostic decisions as well as treatment planning on the basis of this information and other patients' data. Currently, a large portion of the data is not optimally and comprehensively utilized, because information held in the data is inaccessible through visual observation or simple traditional computational methods. Information contained in pelvic CT images is a very important resource for the assessment of the severity and prognosis of the injuries. Each pelvic CT scan consists of several slices; each slice contains a large amount of data that may not be thoroughly and accurately analyzed via visual inspection. In addition, in the field of trauma, physicians frequently need to make quick decisions based on large amount of information. Hence, a computer-assisted pelvic trauma decision-making system is crucial and necessary for

assisting physicians in making accurate diagnostic decisions and determining treatment planning in a short period.

Automated fracture detection from segmented bones in traumatic pelvic injuries can help physicians examine the pelvic CT images and to detect the injury severity within a short period. Extraction of features such as presence and location of fracture, hemorrhage, and displacement between the fractured bones in an automated fashion is vital for such injuries. Identification of fracture alone is not sufficient to assess the injury severity. Therefore, details of the fracture such as distance and angle between the fractured bones must be taken into account. However, the task of pelvic bone segmentation and fracture detection is very challenging due to low resolution of CT images, complex pelvic structures, variations in bone shape, and size from patient to patient. Adding to these complexities, the presence of noise, partial volume effects, and in-homogeneities in the CT images make the task of fracture detection very challenging. The objective of this study is to design a computer-assisted system that helps radiologists better and further assess the bone fractures in pelvic region. It also illustrates the fracture bones in a clearer and more visible manner. In particular, mild and small fractures, while still partially visible in the CT images, are sometimes considered as "irregularities" that need further investigation by the radiologists in the first read, as radiologists may not be able to reliably label them as fractures due to the quality of the CT as well as the volume of the data to be processed. For these situations, it normally takes multiple reads to identify and determine the confirmation on the existence and/or details of fracture. A machine-based analysis can consider and process detailed information from several neighboring slices to provide radiologists with clues as to whether one particular slice contains a fracture and if so extract details such as the separation among the pieces.

While there have been few studies directly focusing on fracture detection in pelvic CT images, there are many closely related work. Moghari and Abolmaesumi [8] utilized a global registration method for multifragment fracture fixation in femur bone. However, the method suffers from initial alignment errors, and the dataset includes only femur bone generated randomly from 3D data points. Moghari and Abolmaesumi [9] proposed a technique to automatically register multiple bone fragments of a fracture using a global registration method guided by a statistical anatomical atlas model. Due to the limited number of bone models, the method is unable to capture all variations of femur. Winkelbach et al. [10] presented an which is approach based on a modified version of Hough Transformation and registration techniques for estimating the relative transformations between fragments of a broken cylindrical structure. This method is designed for computer-aided bone alignment, such as fractured long bones and fracture reduction in surgery. However, the approach is not fully automatic and requires a significant amount of human supervision. Another work, by Ryder et al. [11] explored using nonvisual methods to detect fractures. In addition, there are image processing methods for fracture detection applies to X-ray images [12–14]. Douglas et al. [12] focused on early detection of fractures with low-dose digital X-ray images in a pediatric trauma unit. Tian et al. [13] determined the presence of femoral fracture by measuring the neck-shaft angle of the femur. Lum et al. [14] used three-texture features combined with a classifier to detect radius and femur fractures. This method may suffer from the imbalanced dataset. The majority of these X-ray image processing methods may not be applicable to fracture detection in CT images because of the variation in image intensities and resolution between X-ray and CT images.

Even though few studies have been conducted on fracture detection from pelvic CT scans, several segmentation techniques have been created for medical images of various regions of human body, that is, brain, abdomen, and so forth. These methods include threshold-based techniques, region growing, classifiers, clustering, Markov random field models, artificial neural networks, deformable models, atlas-guided methods, knowledge-based methods. Thresholding techniques segment an image by creating a binary partition on the basis of the image intensities [15]. The drawback is that they cannot be effectively applied to multichannel images. The deformable model approaches start with the initial contour placement near the desired boundary, and then, the contour is improved through an iterative relaxation process [16–18]. The disadvantage is that these methods require manual interaction for the selection of initial position and appropriate parameters of the model. Atlas-guided methods utilize a standard atlas or template for segmentation [19]. The atlas used as the reference frame is generated on the basis of the previously known anatomical information. However, due to anatomical variability across individuals, accurate segmentation of complex structures remains as a challenging task. Clustering algorithms, also referred to as unsupervised methods [20, 21], while successful in some applications, they can be sensitive to noise and variations in intensity. In addition, the calculation can become computationally expensive when the clusters have a large number of pixels.

This study develops an automated hierarchical algorithm to detect fracture in pelvic bones using a hierarchical method combining several of the above-motioned methods in different steps. Fracture detection is performed using the proposed automated segmentation method, called registered active shape model (RASM), along with wavelet transformation, adaptive windowing, boundary tracing, and masking.

The rest of the paper is organized as follows. Section 2 provides the methods used for pelvic bone segmentation and fracture detection. Section 3 includes the results obtained using the proposed methods and discusses the obtained results. Section 4 concludes the proposed methods and provides the future work of the study.

2. Methods

Automated fracture detection is important for making fast and accurate decisions and treatment planning. In order to successfully detect pelvic bone fractures, utilizing the bone information contained in pelvic CT images is crucial. Figure 1 illustrates the overall process of the proposed automated fracture detection. The proposed fracture detection method involves automated bone segmentation using registered active shape model (RASM), adaptive windowing,

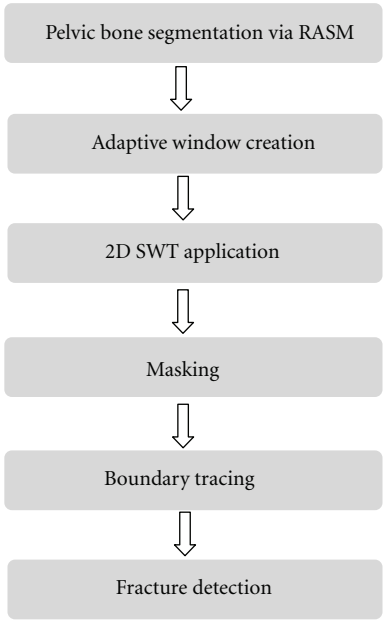


Figure 1: Schematic diagram of pelvic bone fracture detection.

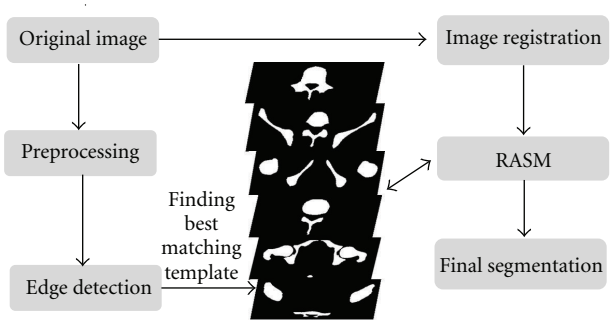


Figure 2: Schematic diagram of pelvic bone segmentation.

2D stationary wavelet transform, masking, and boundary tracing. Each step in the process is explained in detail in the following subsections.

2.1. Multilevel Segmentation of Bone in Pelvic CT Scans. Segmentation is a vital step in analyzing pelvic bones in CT images and the first step in fracture detection. Specifically, bone segmentation helps extract the bones from the images that are later used for detecting fractures. Our previous work was focused on the segmentation of pelvic bones in CT scans [22]. In this paper, a new segmentation algorithm for multilevel pelvic CT scans was developed. This is shown in Figure 2. This new segmentation technique consists of four main parts: preprocessing, edge detection, shape matching and Registered Active Shape Model (RASM) with automatic initialization.

The presence of surrounding artifacts and noise in the original pelvic CT images make bone segmentation a challenging task. Therefore, preprocessing is performed to remove the surrounding artifacts (e.g., CT table, cables, hands, and lower extremities) present in the original image. This is the first step in segmentation. The preprocessing is carried out using blob analysis. Later, high-frequency speckle noise is removed from the images using a 2D Gaussian filter. The image is then enhanced to emphasize the features of interest, that is, pelvic bones. This enhancement is done using brightness contrast stretching. Later, the bone edges are detected using Canny edge detection technique. However, some weak edges may remain unconnected, and as such, morphological operations are applied to remove spurious edges and subedge connections and removal.

The obtained preliminary segmentation results are then used to detect the best matching template using a shape matching algorithm [23]. This helps with the automation of the segmentation process and therefore contributes to minimizing human errors during the diagnostic process. 100 bone templates are created from the Visible Human Project dataset manually. These templates are then compared to each CT slice in order to determine the best-matched template. Determining best-matched template enables the application of corresponding training shape models of each best-matched template to the preprocessed image during bone segmentation phase.

The last step in the segmentation process is the extraction of pelvic bones. Standard active shape model (ASM) is one of the popular techniques that is generally used for bone segmentation. Standard ASM uses training images labeled with landmark points to generate statistical shape and intensity-level models of a desired object. The shape model can be iteratively deformed to locate the object in a test image [24]. The landmarks are points selected by an expert for the bone region in each registered image during the training phase. The pelvic bones in each original training image have different sizes, rotation angles, and locations which may lead to unstable and unreliable shape models for inaccurate bone segmentation. In addition, standard ASM is highly sensitive to initialization and requires an initial position to be correctly assigned to the training model in order to detect a target object in the image. The algorithm then attempts to fit the shape model to the object. If the shape model is not accurately placed, the standard ASM may fail to detect the target object accurately.

In order to overcome these shortcomings, a new image registration algorithm, that is, registered active shape model (RASM), is developed using enhanced homogeneity feature extraction [15], correlation coefficient calculation for similarity measure, affine transformation, and Powell algorithm application [25]. This algorithm, that is, RASM, is developed to create a set of more robust training models which will result in more accurate segmentation. This includes two stages: training stage in which registered training models are created and testing stage which includes automatic initialization. Figure 3. provides the flowchart for the RASM algorithm. After the creation of training models, segmentation is performed on the test images. As mentioned earlier, manual initialization may fail to segment the targeted objects accurately. Hence, an automated hierarchical initialization algorithm is used in the study. The proposed initialization process

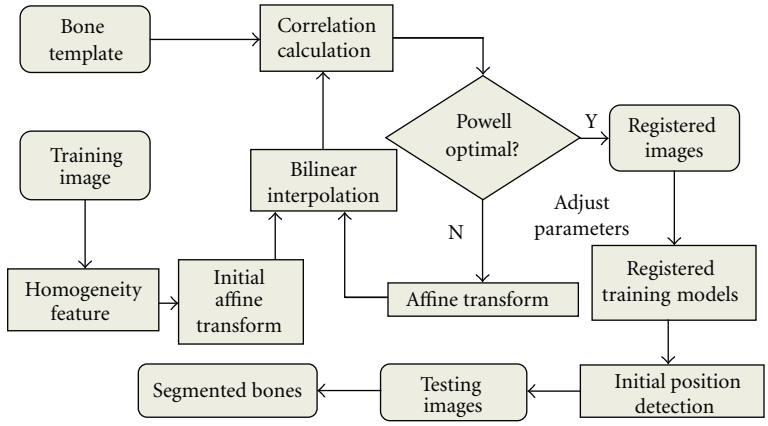


Figure 3: RASM Algorithm.

involves image registration, bone extraction, and edge detection to automatically and sequentially place the training models of each individual object for the test images to extract the bone from the background.

2.2. Fracture Detection in Pelvic CT Images. After bone segmentation, a multistage process is used for fracture detection in pelvic CT scans. Fracture detection of pelvic bones includes several steps. First, pelvic bone segmentation is conducted using the proposed RASM algorithm, as described in Section 2.1. The extracted bone boundaries are utilized to create a series of adaptive windows. Later, 2D stationary wavelet transform (SWT) is applied to each window to test the contour discontinuities in each window using boundary tracing. If there is a contour discontinuity in a window, then it is considered as a potential bone fracture.

2.2.1. Adaptive Window Creation. Discontinuities around the bone boundary help identify the presence of fracture. Therefore, a detailed view of bone boundary is required through the formation of windows around the bone whose sizes are adaptively adjusted to include the bone borders. Creation of these adaptive windows around the bone boundary will facilitate the process of identifying the discontinuities. In this study, a systematic method is proposed to form adaptive windows around the bone boundary to include and detect possible discontinuities associated with fractures. The appearance of bone fractures in a pelvic CT scan depends on the injury severity. Major fractures are usually visible, while minor fractures may not severely distort the edge of the bone; instead, they may appear as dual edges or a single subedge that is slightly blurred compared to the neighboring edges. Therefore, it is important to refine the blurred boundary of each bone in order to achieve accurate fracture detection. The refinement is done using a wavelet transform which is later described in the following subsections. However, due to local intensity variations, it may be difficult to achieve practical and desirable results by applying wavelet transform to the entire bone structure. Hence, the detected bone boundary is divided into a series of windows. The size and location of each window is determined by the area of the bone and boundary detected using the RASM. This is called adaptive windowing. The adaptive windowing algorithm is explained in detail as follows.

On the basis of the segmentation formed by the RASM algorithm, the landmarks are placed on the boundary of each segmented bone. The windows are created starting from the first segmented pelvic bone region. The adaptive window is created on the basis of each landmark placed on the segmented bone boundary.

Let $\{(x_{p1}, y_{p1}), (x_{p2}, y_{p2}), \ldots, (x_{pl}, y_{pl})\}$, $p = 1, 2, \ldots, N$, be the coordinates of the landmarks of each bone in the image. N is the number of bones, and l is the number of landmarks for each pelvic bone. The landmarks are located at the center position (C_p, D_p) of each window. The area of the window W_l is determined using

$$W_l = \frac{A_l}{6}, \tag{1}$$

where A_p is the area of the corresponding piece of bone, The determined empirical constant 1/6 has been selected to ensure that the size of the window is appropriately selected. The side length of the each leg of the cubicle (square) window is identified using

$$S_l = \sqrt{\frac{A_l}{6}}. \tag{2}$$

Since the area of each adaptive window is small, in order to obtain more suitable virtualization effects, each window is scaled to the size of 256×256 by applying the bilinear interpolation technique [14]. As shown in Figure 4, sample adaptive windows are created. Each landmark is located at the center of each window.

2.2.2. The 2D Stationary Wavelet Transform. After adaptive windowing, 2D stationary wavelet transform (SWT) is applied on each window in order to refine the blurred boundary of pelvic bone. The classic discrete wavelet transform (DWT) suffers a shortcoming that the DWT of a translated version of a signal/image is not, in general, the translated version

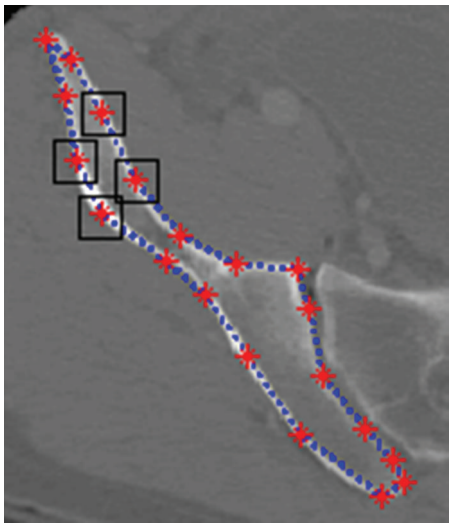


Figure 4: Example windows around the boundary of pelvic bone, positioned according to landmarks.

of the DWT of the signal/image. To overcome this, SWT is applied in our work, as it is designed to overcome any shift variation [26]. The wavelet transform algorithm is explained as follows.

The wavelet transform decomposes an input signal into different frequency components using a series of filtering operations. A wavelet $\overline{\varphi_a}(t)$ is a function with a zero average

$$\int \psi(t)dt = 0. \tag{3}$$

The wavelet generates a family of wavelets by scaling $\psi(t)$ by a and translating it by θ:

$$\varphi_{\theta,a}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-\theta}{a}\right). \tag{4}$$

The wavelet transform of a signal $s(t)$ at time θ and scale a can be represented as

$$\begin{aligned} W_s(\theta,a) &= \langle s(t), \varphi_{\theta,a}(t)\rangle\ , \\ W_s(\theta,a) &= \int_{-\infty}^{+\infty} s(t)\frac{1}{\sqrt{a}}\varphi^*\left(\frac{t-\theta}{a}\right)dt. \end{aligned} \tag{5}$$

The convolution computes the wavelet transform of the input signal with dilated band-pass filters. Two sets of coefficients are obtained through wavelet transform, one is approximation coefficients, cA_j, and the other is detail coefficients, cD_j, where j is the level of decomposition, including horizontal, vertical, and diagonal coefficients. Decimation makes wavelet transform a shift-variant process. To overcome this, a stationary discrete wavelet transform is used in this study.

The scaled window W is first decomposed using a 2D Stationary Discrete Wavelet Transform. The classical Discrete Wavelet Transform (DWT) is not a space-invariant transform. The SWT is an algorithm which does not decimate the coefficients at every level of decomposition [26]. The filters at level i are upsampled versions of those at level $(i-1)$. As with the 2D DWT, decomposition outputs approximation, horizontal, vertical, and diagonal coefficients. In this application, three levels of decomposition are applied to window W using the Haar wavelet. The level 3 detail coefficients, ${cD_{j+1}}^{(h)}$, ${cD_{j+1}}^{(v)}$, and ${cD_{j+1}}^{(d)}$, are then extracted and used to reconstruct detail arrays D_h, D_v, and D_d of horizontal, vertical, and diagonal coefficients. Figure 5 represents decomposition of 2D SWT.

The accuracy and running speed of the SWT algorithm are compared when extracting the upsampled coefficients separately at 1st, 2nd, 3rd, and 4th levels. The algorithm runs on the computer with 2.80 GHz Intel(R) Core(TM) i7 processor, 64-bit Operating System, 6.0 GB memory. For each CT slice, it takes approximately 0.15 seconds more for the 2nd level of stationary wavelet decomposition than the 1st level decomposition. While the 3rd level of decomposition is only 0.1 second slower than the 2nd level of decomposition in terms of running speed, more noise is filtered out, and edges are clearer in the 3rd level of decomposition compared to other two levels; this improves the accuracy of the fracture detection algorithm. Going to the 4th level adds another 0.15 second of additional delay while not adding much to the filtering performance. Hence, in order to achieve a suitable balance between the running speed and accuracy, the 3rd level of SWT is used in this work.

2.2.3. Masking. The next step in the fracture detection is to create a binary version of the chosen detail array W_b from the wavelet transform. This binary version not only contains the pelvic bone contour, but also includes other redundant and unnecessary edges. A mask is formed to filter these redundant edges out. The mask W_m is formed by converting the smoothed window to a binary image using Otsu's threshold [27]. The threshold is computed to minimize the intraclass variance, defined as a weighted sum of variances of two classes, black and white pixels.

$$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t). \tag{6}$$

Weights w_i are probabilities of the two classes separated by a threshold t and σ_i^2 variances of these classes. Minimizing the intraclass variance is the same as maximizing interclass variance

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = w_1(t)w_2(t)\left[\mu_1(t) - \mu_2(t)\right]^2, \tag{7}$$

where w_i are probabilities of the two classes and μ_i is the class mean.

The contour is then extracted from the binary image. The unwanted edges are removed from the binary image to create an edge window. Later, a precise edge window W_e is obtained by removing the extra edges in the image using the pelvic bone contour and the mask. The process is defined as a combination of W_b and W_m. This edge window is used for the boundary tracing as described in next step

$$W_e = W_b \times W_m. \tag{8}$$

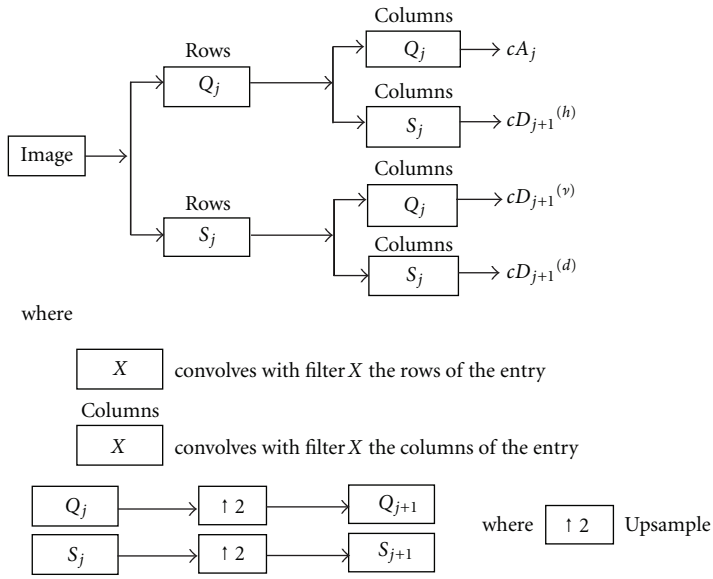


Figure 5: Decomposition steps of 2D SWT.

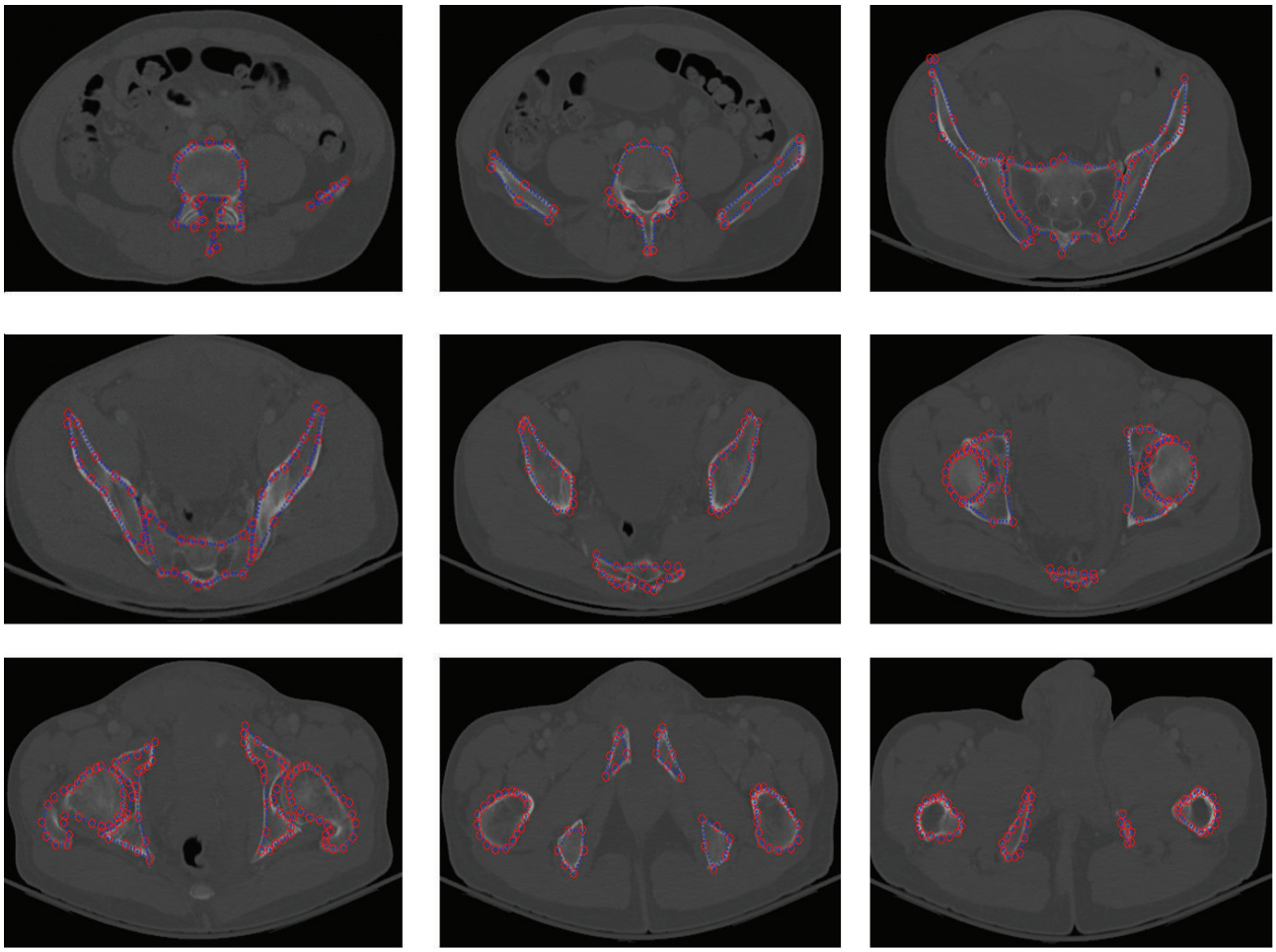


Figure 6: Example of pelvic bone segmentation results via RASM.

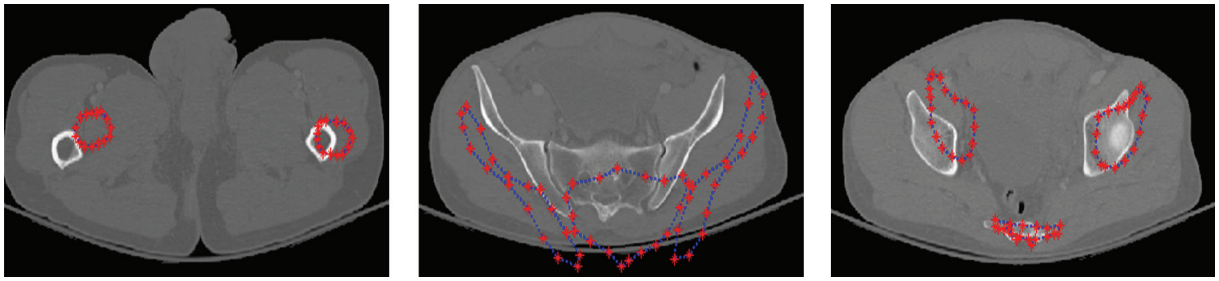
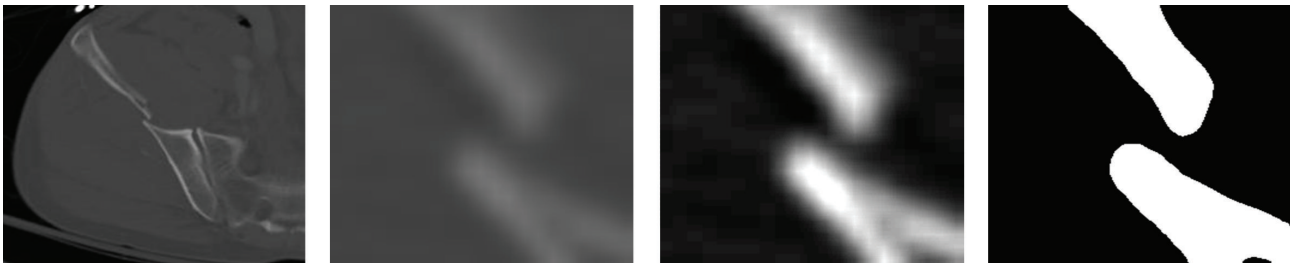


Figure 7: Example results of pelvic bone segmentation via standard ASM without initialization.



(a) (b) (c) (d)

Figure 8: Example of a detected broken boundary of pelvic bone, which may indicate a fracture.

2.2.4. Boundary Tracing. After masking, the last and final step in fracture detection is the detection of discontinuities. This is achieved by tracing the extracted bone edges. Small artifacts surrounding the extracted bone edges may interfere with the boundary tracing. Therefore, these artifacts must be removed. These are removed by applying morphologic opening to all the objects in the image with area below a specific threshold, which is predefined as 1% of the window area in the testing step. The remaining edges are then traced using the 8-neighborhood of each pixel and are returned as a matrix of pixel positions. The traced edges represent the pelvic bone contours. The window will therefore contain a single continuous boundary if there is no fracture. In the presence of fracture, multiple boundaries are present in the window, depending on the type and severity of fracture.

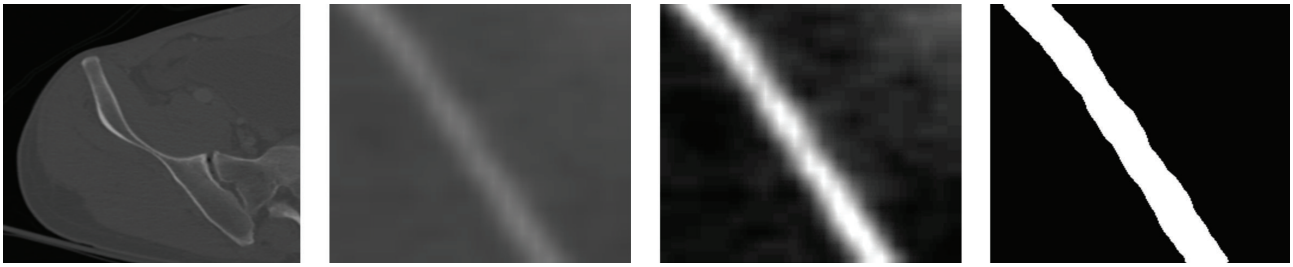
3. Results and Discussion

3.1. Dataset. The dataset has been obtained from the Virginia Commonwealth University Medical Center. Data have been collected from twelve patients with traumatic pelvic injuries. Forty-five to seventy-five images are collected from each patient. Axial CT images with five millimeter slice thickness are used for the study. Images collected from five patients are used for training, and the other seven patients' images are used for testing. For fracture detection, a total of 12 patients are used, out of which 8 patients exhibit small to very severe bone fractures.

3.2. Results of Bone Segmentation. Figure 6 shows a sample segmentation of pelvic bones using RASM. Figure 7 shows the compared results of pelvic bone segmentation via standard ASM without initialization. The main reason of inaccurate bone segmentation is that the initial positions of training models are not correctly assigned. As given in [8], total segmentation accuracy for both good and acceptable classes is 95.77%. These results were evaluated by expert radiologist as ground truth for assessment.

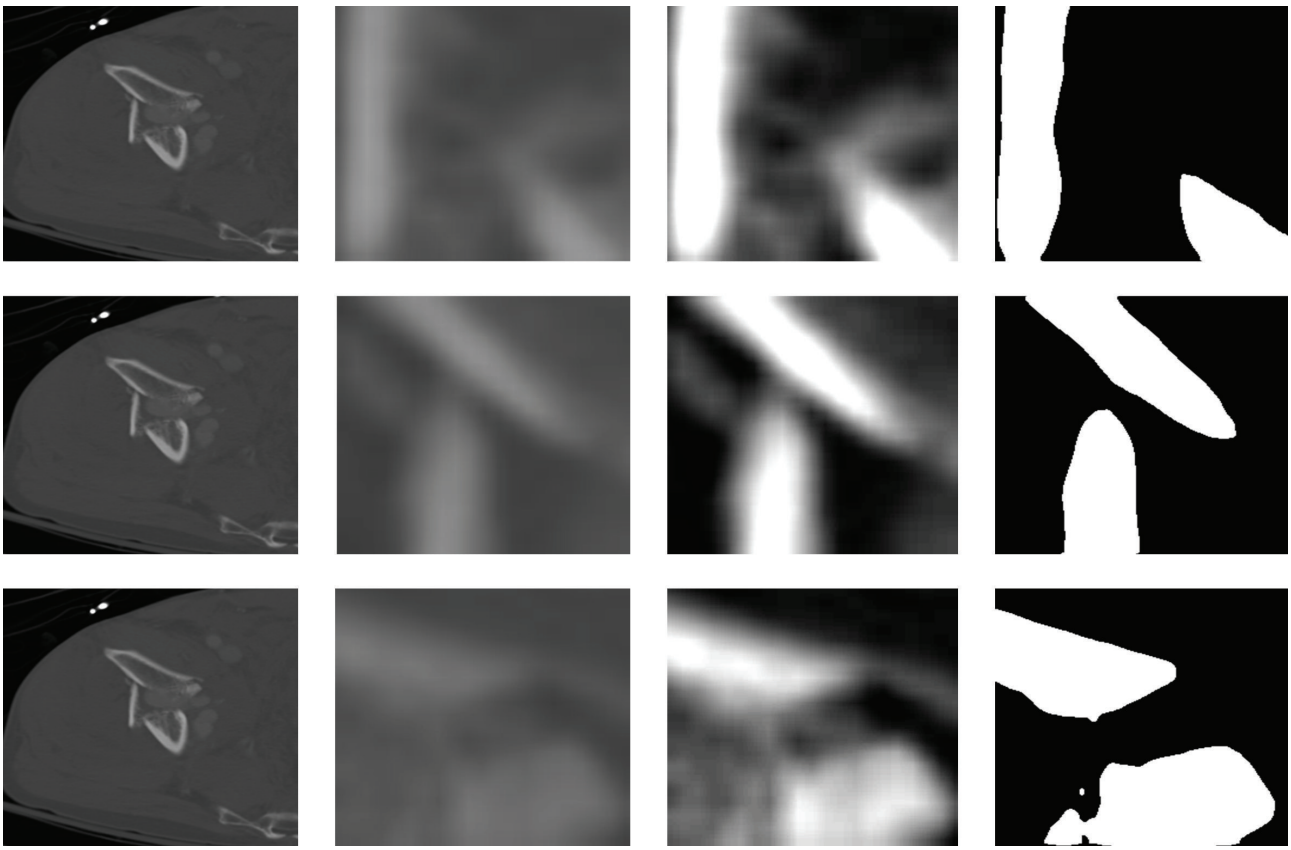
3.3. Results of Fracture Detection. Figures 8 through 10 show the results obtained at various stages of fracture detection. In these figures, (a) is the original image, (b) is the extracted adaptive window after being scaled, and (c) is the enhanced window after brightness contrast stretching. This is done for better visualization effect. And, (d) shows the final fracture detection results. In Figure 8, the patient suffers from a minor fracture in right iliac wing. Figure 8(d) indicates the fracture detected in the right iliac wing. Figure 9 is the "no fracture" case. The result in Figure 9(d) shows that the bone appears smooth with no fracture. Figure 10 illustrates a patient with a very severe fracture in the right ilium bone. Fractures are detected from the windows of this bone region. Example of detected fractures shown in Figure 10(d) indicates fractures in three different regions of the right ilium bone. These results are evaluated by an expert radiologist and are considered acceptable. For 8% of the cases, the method was unable to capture the fracture. The few cases that the algorithm gave false alarms in fracture detection may be either due to the algorithm needing further refinement or other factors such as the poor quality of these particular CT images.

The results show that the method can successfully detect bone fracture. Table 1 presents the performance of the method detecting fractures. The proposed method is highly



(a) (b) (c) (d)

Figure 9: Example of a detected nonbroken boundary of pelvic bone, which may indicate no fracture.



(a) (b) (c) (d)

Figure 10: Example of a detected broken boundary of pelvic bone, which may indicate three fractures.

Table 1: Performance of pelvic bone fracture detection.

Statistical Results	Accuracy	Sensitivity	Specificity
Rate %	91.9821	93.3333	89.2617

sensitive to the discontinuities present in the bone and is capable of detecting fractures.

3.4. Discussion. The results were validated on the basis of the assessment and evaluation made by radiologists on the CT scans in the above mentioned database. As shown in the results, the designed algorithm is able to detect the fractures relatively accurately. Using the proposed algorithm, fractured bone may be further highlighted in the processed images; this could help the radiologists better analyze the scans and increase the chances of capturing the fractures. Additionally, as it can be seen in the results, our designed method may help quantify the fracture separation distance and the angle between the broken bone pieces as well as other quantitative assessment of the fractures, which may not be easily accessible and measurable through visual inspection. The designed algorithm provides these clues and recommendations on the fracture detection in an automated fashion and with relatively high speed (the processing time is less than one second for each slice). This helps physicians reduce the decision-making

and diagnostic time, which is highly important for traumatic pelvic injuries.

4. Conclusion and Future Work

This paper presents a method for detecting fractures in pelvic bones using automated bone segmentation, adaptive windowing, boundary tracing, and 2D stationary wavelet Transform while including anatomical information. The results show that the proposed method is capable of detecting fractures in pelvic bones accurately. Automated fracture detection, once verified with more data, will be an important component of a larger modular system to extract features from CT images for a computer-assisted decision-making system. Future work will focus on the quantitative measurement of fracture on the basis of a larger dataset, for example, horizontal displacement, as well as the determination of fracture type.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant no. IIS0758410. The authors would like to thank Virginia Commonwealth University Medical Center for providing the data for the study.

References

[1] M. A. Schiff, A. F. Tencer, and C. D. Mack, "Risk factors for pelvic fractures in lateral impact motor vehicle crashes," *Accident Analysis and Prevention*, vol. 40, no. 1, pp. 387–391, 2008.

[2] A. Salim, P. G. R. Teixeira, J. DuBose et al., "Predictors of positive angiography in pelvic fractures: a prospective study," *Journal of the American College of Surgeons*, vol. 207, no. 5, pp. 656–662, 2008.

[3] University of Maryland National Study Center for Trauma/EMS, "Lower extremity injuries among restrained vehicle occupants," Tech. Rep., University of Maryland National Study Center for Trauma/EMS, 2001.

[4] G. S. Pajenda, H. Seitz, M. Mousavi, and V. Vecsei, "Concomitant intra-abdominal injuries in pelvic trauma," *Wien Klin Wochenscher*, vol. 110, no. 23, pp. 834–840, 1998.

[5] Z. Balogh, K. L. King, P. Mackay et al., "The epidemiology of pelvic ring fractures: a population-based study," *Journal of Trauma*, vol. 63, no. 5, pp. 1066–1072, 2007.

[6] P. C. Ferrera and D. A. Hill, "Good outcomes of open pelvic fractures," *Injury*, vol. 30, no. 3, pp. 187–190, 1999.

[7] F. D. Brenneman, D. Katyal, B. R. Boulanger, M. Tile, and D. A. Redelmeier, "Long-term outcomes in open pelvic fractures," *Journal of Trauma*, vol. 42, no. 5, pp. 773–777, 1997.

[8] M. H. Moghari and P. Abolmaesumi, "Global registration of multiple bone fragments using statistical atlas models: feasibility experiments," in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '08)*, pp. 5374–5377, August 2008.

[9] M. H. Moghari and P. Abolmaesumi, "Global registration of multiple point sets: feasibility and applications in multi-fragment fracture fixation," in *Proceedings of 10th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '07)*, vol. 10, pp. 943–950, Brisbane, Australia, 2007.

[10] S. Winkelbach, R. Westphal, and T. Goesling, "Pose estimation of cylindrical fragments for semi-automatic bone fracture reduction," in *Proceedings of the 25th Annual Symposium of the German Association for Pattern Recognition (DAGM '03)*, vol. 2781 of *Lecture Notes in Computer Science*, pp. 566–573, Magdeburg, Germany, 2003.

[11] D. M. Ryder, S. L. King, C. J. Olliff, and E. Davies, "Possible method of monitoring bone fracture and bone characteristics using a non-invasive acoustic technique," in *Proceedings of the International Conference on Acoustic Sensing and Imaging*, pp. 159–163, March 1993.

[12] T. S. Douglas, V. Sanders, R. Pitcher, and A. B. van As, "Early detection of fractures with low-dose digital X-ray images in a pediatric trauma unit," *Journal of Trauma*, vol. 65, no. 1, pp. E4–E7, 2008.

[13] T. P. Tian, Y. Chen, W. K. Leow, W. Hsu, T. S. Howe, and M. A. Png, "Computing neck-shaft angle of femur for X-ray fracture detection," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, vol. 2756 of *Lecture Notes in Computer Science*, pp. 82–89, Springer, 2003.

[14] V. L. F. Lum, W. K. Leow, Y. Chen, T. S. Howe, and M. A. Png, "Combining classifiers for bone fracture detection in X-ray images," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '05)*, pp. 1149–1152, September 2005.

[15] C. Lee, S. Huh, T. A. Ketter, and M. Unser, "Unsupervised connectivity-based thresholding segmentation of midsagittal brain MR images," *Computers in Biology and Medicine*, vol. 28, no. 3, pp. 309–338, 1998.

[16] J. Montagnat and H. Delingette, "4D deformable models with temporal constraints: application to 4D cardiac image segmentation," *Medical Image Analysis*, vol. 9, no. 1, pp. 87–100, 2005.

[17] J. Schmid and N. Magnenat-Thalmann, "MRI bone segmentation using deformable models and shape priors," in *Proceedings of 11th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '08)*, vol. 11, pp. 119–126, New York, NY, USA, 2008.

[18] P. C. T. Gonçalves, J. M. R. S. Tavares, and R. M. N. Jorge, "Segmentation and simulation of objects represented in images using physical principles," *Computer Modeling in Engineering and Sciences*, vol. 32, no. 1, pp. 45–55, 2008.

[19] S. Sandor and R. Leahy, "Surface-based labeling of cortical anatomy using a deformable atlas," *IEEE Transactions on Medical Imaging*, vol. 16, no. 1, pp. 41–54, 1997.

[20] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," *Pattern Recognition*, vol. 40, no. 3, pp. 825–838, 2007.

[21] H. A. Vrooman, C. A. Cocosco, R. Stokking et al., "KNN-based multi-spectral MRI brain tissue classification: manual training versus automated atlas-based training," in *Medical Imaging 2006: Image Processing*, Proceedings of the SPIE, San Diego, Calif, USA, February 2006.

[22] J. Wu, P. Davuluri, K. Ward, C. Cockrell, R. Hobson, and K. Najarian, "A new hierarchical method for multi-level segmentation of bone in pelvic CT scans," in *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '11)*, 2011.

[23] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[24] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[25] F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," *Medical Image Analysis*, vol. 3, no. 4, pp. 373–386, 1999.

[26] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," in *Wavelets and Statistics*, vol. 103 of *Lecture Notes in Statistics*, pp. 281–299, Springer, 1995.

[27] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2011, Article ID 540820, 10 pages
doi:10.1155/2011/540820

Research Article

Nonlinear Elasto-Mammography for Characterization of Breast Tissue Properties

Z. G. Wang,[1,2] Y. Liu,[3] G. Wang,[4] and L. Z. Sun[5]

[1] *Department of Civil and Environmental Engineering, The University of Iowa, Iowa City, IA 52242, USA*
[2] *Prudential Insurance Company, Newark, NJ 07102, USA*
[3] *Cooper Tire and Rubber Company, Findlay, OH 45840, USA*
[4] *School of Biomedical Engineering and Sciences, Virginia Tech, Blacksburg, VA 24061, USA*
[5] *Department of Civil and Environmental Engineering, University of California, Irvine, CA 92697-2175, USA*

Correspondence should be addressed to L. Z. Sun, lsun@uci.edu

Received 6 June 2011; Revised 23 September 2011; Accepted 23 September 2011

Academic Editor: Shan Zhao

Quantification of the mechanical behavior of normal and cancerous tissues has important implication in the diagnosis of breast tumor. The present work extends the authors' nonlinear elastography framework to incorporate the conventional X-ray mammography, where the projection of displacement information is acquired instead of full three-dimensional (3D) vector. The elastic parameters of normal and cancerous breast tissues are identified by minimizing the difference between the measurement and the corresponding computational prediction. An adjoint method is derived to calculate the gradient of the objective function. Simulations are conducted on a 3D breast phantom consisting of the fatty tissue, glandular tissue, and cancerous tumor, whose mechanical responses are hyperelastic in nature. The material parameters are identified with consideration of measurement error. The results demonstrate that the projective displacements acquired in X-ray mammography provide sufficient constitutive information of the tumor and prove the usability and robustness of the proposed method and algorithm.

1. Introduction

Breast cancer is a major threat to public health in the world. In USA and Europe, approximately 10% of women develop breast cancer during the course of their lives. While the specific causes of breast cancer are unknown, early detection and characterization of breast tumors is the key to successful treatment. Currently, X-ray mammography, a low-dose X-ray imaging modality, is the primary diagnosis method in clinics [1]. While being more efficient in detecting malignancies as age increases or the breast becomes fatty, mammography fails to identify small cancers in dense breasts. Furthermore, mammography may not be specific in terms of tumor benignity and malignancy. About 80% of suspicious masses referred by mammography for surgical breast biopsy are in fact not malignant [2–4]. These false-positive mammograms may induce patients' anxiety, distress, and intrusive thoughts.

A number of techniques have been attempted to address these problems associated with mammography. From the viewpoint of mechanics, the tissue stiffness is an important index for diagnosis of breast cancers, as tumors are stiffer than the surrounding breast tissues and malignant tumors are much stiffer than benign ones [5–7]. In other words, in vivo identification of the mechanical parameters of normal and abnormal tissues should improve the accuracy of cancer diagnosis. Correspondingly, elastography has been proposed as a method to image the tissues' elasticity in a quantitative manner. The general basis of elastography is to induce motion within tissue by mechanical stimulation. Conventional medical imaging modalities are then used to measure the spatial deformation, from which the mechanical properties can be extracted. Based on the imaging modalities used, elastography has two major classes: ultrasound elastography (USE) and magnetic resonance elastography (MRE). USE, developed in the 1990s, is the first modulus-imaging modality. It computes the lap between the pre- and postcompression radio frequency ultrasound signals to estimate the tissue's axial displacement and strain under quasistatic loading [8, 9]. While providing new information

for detecting pathological tumors, USE suffers from limited stiffness range as imposed by the minimum resolvable wavelength. The computed image in USE is also restricted by the angular resolution of the transducer and its ability to separate signals from artifacts and noise [9]. Magnetic resonance elastography (MRE) is a second-generation elastography modality that provides higher resolution images and is capable of producing sufficient 3D spatial and contrast resolution [10, 11]. MRE is, however, significantly more costly as a result of the MR imaging procedure and hence is not generally applicable for all patients. From the viewpoint of solid mechanics, the current USE and MRE are insufficient, because both are based on infinitesimal-strain linear elasticity and only very few are capable of considering anisotropic tissue properties. In other words, the large deformation, nonlinear, and anisotropic behaviors of breast tissues (fat and glandular tissues) and tumor have not yet been taken into consideration by USE or MRE. Therefore, the outcomes of USE and MRE may not be sufficiently accurate for the diagnostic purpose.

Motivated by the significance of early detection of breast tumors and the current limitations of mammography and elastography modalities, we have developed a nonlinear elasto-mammography method that takes into consideration of the finite-strain nonlinear properties of breast tissues, in combination with mammography visualization. The development has experienced two stages.

First, a *linear elasto-mammography* framework was developed to generate the elastograms of breast tissues, by combining the conventional low-dose X-ray mammography with linear elastography framework [12]. Instead of applying ultrasound or magnetic resonance as in the previous elastography research, elasto-mammography uses displacement information extracted from mammography projections before and after breast compression. Incorporating the displacement measurement, an elastography reconstruction algorithm was specifically developed to estimate the elastic moduli of heterogeneous breast tissues. Case studies with numerical breast phantoms showed that the displacement measurement obtained from mammography is sufficient to identify the material parameters of breast tissues and tumors within the framework of linear elasticity.

Then, a *nonlinear elastography* method was proposed [13]. As discussed above, the current elastography (USE or MRE) reconstruction framework is based on the assumption of linear elasticity theory. The mechanics of biological soft tissues, however, require nonlinear continuum mechanics description [14, 15]. While tissue models based on linear elasticity have been broadly used, they are reliable only when the tissue strain is less than 5% [16], which is much lower than the deformation of soft tissues. Thus, consideration of nonlinearity is essential for elastography in clinical applications. Our development of nonlinear elastography method, for the first time, enables identification of the mechanical properties of soft breast tissues and tumor. To improve the computational efficiency and enhance the stability, a nonlinear adjoint method was introduced. The phantom study demonstrated that the complex nonlinear mechanics of soft breast tissues and tumors can be quantified from 3D displacement and force measured on the surface of the breast.

The objective of the present study is to develop a *nonlinear elasto-mammography* framework that combines the simplicity of projective X-ray mammography measurement with the accuracy of nonlinear elastography. In Section 2, we present the mathematical derivation, where an adjoint gradient method is modified to consider the projective displacement measurements. Finite-element- (FE-) based numerical simulations are conducted in Section 3 to reconstruct the material parameters of a 3D heterogeneous breast phantom from mammography displacement. Two types of mammography compressive loadings are applied, and the displacements at key points on the tissue interfaces are extracted from mammography projections before and after deformation. In Section 4, the results are presented and the effect of experiential error is investigated.

2. Methods

2.1. Finite-Strain Deformation Equations. Let Ω^0 be a biological object subjected to body force $\mathbf{b}$ and surface force $\mathbf{t}$ on boundary $\Gamma^0_{\mathbf{t}}$. Here, we consider general problems that the body force $\mathbf{b}$ and surface force $\mathbf{t}$ are deformation dependent. Following the standard finite-element method, the displacement $\mathbf{u}$ is discretized as nodal displacement vector $\{u\} = \{u_1, u_2\}^T$, where u_2 corresponds to $\bar{\mathbf{u}}$ prescribed on $\Gamma^0_{\mathbf{u}}$ and u_1 is to be solved from nonlinear equations; that is, on surface $\Gamma^0_{\mathbf{u}}$ ($\Gamma^0_{\mathbf{u}} \cup \Gamma^0_{\mathbf{t}} = \partial\Omega^0$), as described in [13], the FE description of the finite-strain equilibrium equation is

$$\begin{Bmatrix} f_1^{\text{in}}(u_1, u_2; \mathbf{p}) \\ f_2^{\text{in}}(u_1, u_2; \mathbf{p}) \end{Bmatrix} - \begin{Bmatrix} f_1^{\text{out}}(u_1, u_2) \\ f_2^{\text{out}} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}. \tag{1}$$

The internal nodal force f^{in} corresponds to the stress of the tissue; that is, it changes with u_1 and material parameters $\mathbf{p}$ but not u_2 as it is prescribed. The external nodal force f_1^{out} is due to the prescribed surface force $\mathbf{t}$ and body force $\mathbf{b}$ in biological object Ω^0. It changes with the displacement in large deformation. The nodal force f_2^{out} is the unknown constraint force on $\Gamma^0_{\mathbf{u}}$.

A classic quasi-Newton method [17] is employed to solve (1) for u_1. Let $u_1^{(n)}$ be the trial solution of the unknown u_1 at the nth iterative step. An improved solution $u_1^{(n+1)} = u_1^{(n)} + \delta u_1$ can be obtained at the next step, in which δu_1 is the solution of linear equations:

$$K_{11}^{\text{eff}}\left(u_1^{(n)}\right)\delta u_1 = f_1^{\text{in}}\left(u_1^{(n)}, u_2; \mathbf{p}\right) - f_1^{\text{out}}\left(u_1^{(n)}, u_2\right) \tag{2}$$

with

$$K_{11}^{\text{eff}} = \left(K_{11}^{\text{in}} - K_{11}^{\text{out}}\right), \qquad K_{11}^{\text{in}} = \frac{\partial f_1^{\text{in}}}{\partial u_1}, \qquad K_{11}^{\text{out}} = \frac{\partial f_1^{\text{out}}}{\partial u_1}, \tag{3}$$

where the matrices are evaluated at $u_1^{(n)}$.

2.2. Nonlinear Elasto-Mammography Algorithm. We consider that the biological object Ω^0 is discretized into FE mesh, and the displacement and force are discretized consistently into nodal displacement and nodal force. Experimental measurement for elasto-mammography is displacement. We catalog the measurements as the following. (i) If the force at a node is known, it will be included into f_1^{out} which is considered "prescribed" in (1). The corresponding nodal displacement will be considered as unknown u_1 in the FE equation (1). (ii) All the other nodal displacements will be in u_2, and the corresponding unknown nodal force will be in f_2^{out}. For category (ii), u_2 must be considered "prescribed" to fulfill the requirement of the well poseness of a solid mechanics problem.

In our previous elastography method [13], displacements are also measured at some of the nodes associated with u_1 and are denoted as U_1^M. Given material parameters $\mathbf{p}$, the unknown displacement u_1 and constraint force f_2^{out} (which depends on $\mathbf{p}$) will be solved from the FE equation (1). The elastography method thus seeks $\mathbf{p}$ so that the overall difference between measured U_1^M and computed u_1 is minimum; that is, to minimize objective function:

$$\Phi(\mathbf{p}) = \left(u_1 - U_1^M\right)^T \mathbf{\Lambda}\left(u_1 - U_1^M\right), \tag{4}$$

where diagonal weight matrix $\mathbf{\Lambda} = \operatorname{diag}(a_1, a_2, \ldots, a_j, \ldots)$, with component $a_j = 1$ when the jth component of U_1^M is experimentally measured, or $a_j = 0$ otherwise.

In mammography, however, the measurement of displacement is limited by the projection; that is, only the two components perpendicular to the projection direction are obtainable. Correspondingly, the computed displacement u_1 should be *projected* in the same direction as in mammography and then compared with the mammography measurement U_1^M. As derived in Appendix A, the projection can be represented by a linear translation of u_1, as $\mathbf{R}u_1$, where $\mathbf{R}$ is a global projection matrix. The objective function for nonlinear elasto-mammography is then

$$\Phi(\mathbf{p}) = \left(\mathbf{R}u_1 - U_1^M\right)^T \mathbf{\Lambda}\left(\mathbf{R}u_1 - U_1^M\right). \tag{5}$$

2.3. Nonlinear Adjoint Method. Efficient and robust optimization-based elastography reconstruction schemes request user-supplied gradient $\partial\Phi/\partial\mathbf{p}$. Direct calculation of the gradients $\partial\Phi/\partial\mathbf{p}$ involved in the minimization-based parametric identification is difficult, because u_1 is an implicated function of $\mathbf{p}$. Recently, an adjoint method was introduced to compute the gradient analytically [18–21]. The corresponding nonlinear finite element formulas are shown in Appendix B. Briefly, given a trial $\mathbf{p}, u_1$ will be solved from FE equations (2) and (3), the objective function will calculated by (5), and the material parameters $\mathbf{p}$ will be updated by large-scale limited memory BFGS (L-BFGS) method with user supplied gradients readily obtained as:

$$\frac{\partial\Phi}{\partial\mathbf{p}} = \begin{Bmatrix} w_1 \\ w_2 \end{Bmatrix}^T \begin{Bmatrix} \dfrac{\partial f_1^{\text{in}}}{\partial\mathbf{p}} \\ \dfrac{\partial f_2^{\text{in}}}{\partial\mathbf{p}} \end{Bmatrix}, \tag{6}$$

where the virtual adjoint displacements w_1 and w_2 are solved from linear equations:

$$\begin{aligned} K_{11}^{\text{eff}} w_1 &= -2\mathbf{R}^T\mathbf{\Lambda}\left(\mathbf{R}u_1 - U_1^M\right), \\ w_2 &= 0, \end{aligned} \tag{7}$$

with the tangent stiffness matrix K_{11}^{eff} defined in (3). The most significant features of the adjoint method are the analytical formulation, high accuracy, and computational efficiency [22]. Since K_{11}^{eff} and its LU factorization have been calculated when solving the FE equation (2), the additional computational expense for w_1 in (7) is minimal. Furthermore, it only needs to solve one linear equation (7) regardless of the number of unknown parameters in $\mathbf{p}$.

The reconstruction procedure is illustrated in Figure 1. We first establish a numerical FE model of the breast tissue on which external loadings are applied. In order to measure displacement, we compare the mammography projections before and after the deformation. Then, initial guess of the distribution for material parameters (λ, μ, γ) is given. Given the external loadings and material parameters, the displacement filed u_1 is solved from (1) and is projected to $\mathbf{R}u_1$ according to the mammography direction. The difference between prediction $\mathbf{R}u_1$ and measurement U_1^M are evaluated by the objective function (5). The adjoint field w is calculated by (7), and gradients $\partial\Phi/\partial\mathbf{p}$ are obtained by (6). The material parameters could be updated by limited-memory BFGS (L-BFGS) optimization subroutine [23]. The iteration continues until a minimization is reached.

3. Numerical Simulations

3.1. Breast Phantom and Forward Problem. We establish a 3D typical breast FE phantom, shown in Figure 2, consisting of the fatty and glandular tissues and a ductal carcinoma (tumor). Boundaries of these regions are described with sets of splines. The mechanical properties of these tissues are described with Fung-type isotropic hyperelastic model [14], whose strain energy function reads

$$W(\mathbf{E}) = \frac{\gamma}{2}\left[\exp\left(\lambda(\mathbf{I}:\mathbf{E})^2 + 2\mu\mathbf{E}:\mathbf{E}\right) - 1\right], \tag{8}$$

where $\mathbf{E}$ is the Green strain and $\{\lambda, \mu, \gamma\}$ are material parameters. The parameters $\{\lambda, \mu, \gamma\}$ are previous determined [13] from ex vivo experimental data of Samani and Plewes [24] as $\lambda_d = 80$, $\mu_d = 35$, $\gamma_d = 1.5$ (λ and μ are dimensionless, γ is in kPa) of ductal carcinoma, $\lambda_f = 35$, $\mu_f = 12.5$, $\gamma_f = 0.4$ of fatty tissue, and $\lambda_g = 50$, $\mu_g = 25$, $\gamma_g = 0.25$ glandular tissue.

Motivated by the breast compression in X-ray mammography, we designed two loadings as detailed in [13]. In the FE model, the base of the breast phantom is fixed. Two paddles are used to apply displacement on the upper surface of the breast. The paddle close to tumor applies tilted compression, and another paddle is fixed to restrict the breast.

3.2. Acquisition Projection Data. For each loading, mammography projections for 3D heterogeneous breast phantom

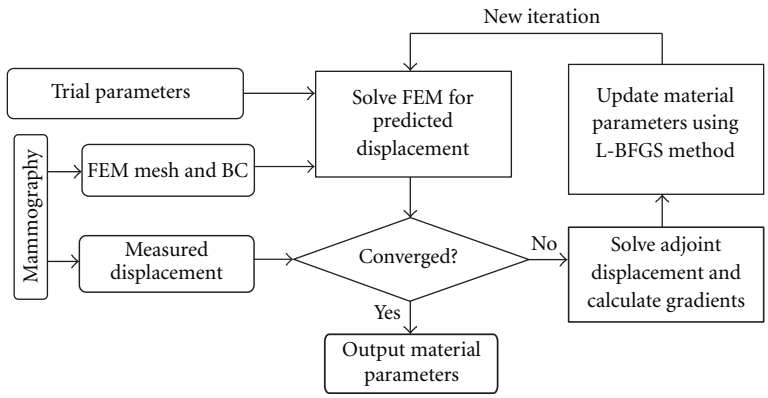


Figure 1: Overall flowchart for nonlinear reconstruction of material parameters of breast tissues.

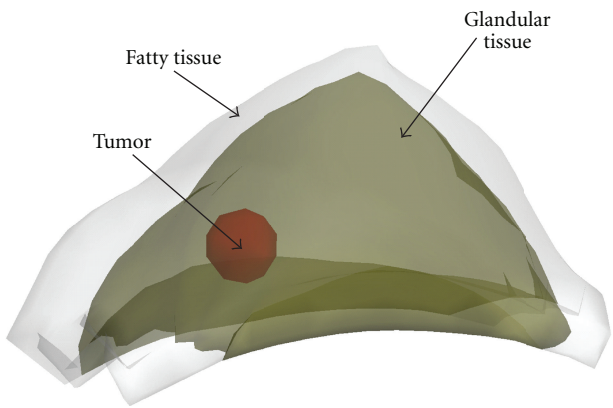


Figure 2: Mammography projections for 3D heterogeneous breast phantom after deformation. Fatty tissue, glandular tissue, and a tumor are shown.

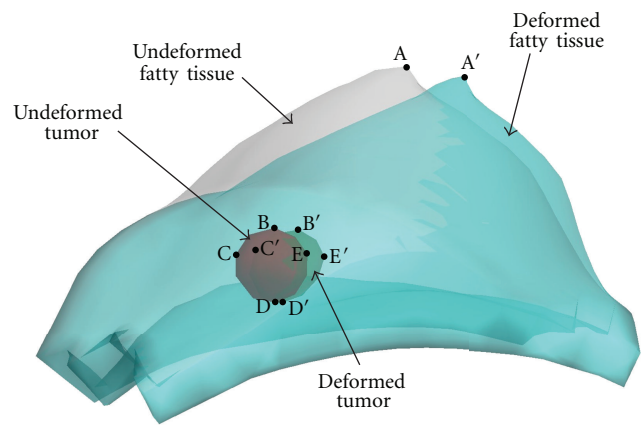


Figure 3: Overlapped mammography-type projections of the fatty tissue and tumor in deform and undeformed configuration. In the projections, vertexes A–E in underformed projection move to A′–E′ in deformed projection, respectively, giving the projected displacements of these points.

are taken before and after deformation (Figure 2). To mimic the displacement obtainable from mammography, we extract the displacement components in the projection plane at some discrete material points (Figure 3), denoted as U_1^M. We select three mammography projection directions. With each direction, one projection is made at undeformed state, and one is made at deformed configuration (Figure 2). Then, the displacement components on the projection plane are extracted from a set of landmarks in the tissues by comparison their position in undeformed and deformed projections, as shown in Figures 3 and 4. The landmarks include the top vertex on the upper breast surface (point A in Figure 3), four vertexes of the tumor surface (points B–E in Figure 3), and ten material points on the fat-glandular interface (points A–J in Figure 4). It is noted that the surfaces of tumor and glandular tissue are not smooth so that there are plenty of landmarks that can be used to track the deformation.

To explain the procedure, we use a mammography compression as example. Figure 2 shows mammography projection taken in the same direction with compression applied on the breast. The boundary of the fatty tissue, glandular tissue, and a tumor can be seen in the projection. Then, displacement components on the projection plane can be extracted by comparing the undeformed and deformed projections (Figures 3 and 4). More specifically, the undeformed and deformed projections of fatty tissue and the tumor are registered and shown together for the comparison. The top vertex of fatty tissue, point A, moves to vertex A′ after deformation. Points B–E are vertexes of the tumor in undeformed projection, and they move to vertexes B′–E′ after deformation (Figure 3). On the fat-glandular surface, we select additional ten landmarks that move from A–J to A′–J′, respectively (Figure 4). Thus, by measuring the vector from a point to its deformed position, for example, A → A′, the projective displacement components are obtained and recorded as U_1^M. In addition, it is assumed that there is no slip between the paddles and breast surface during mammography compression. Therefore, the displacement of the material points directly compressed by the paddles is considered known and is added to the measurement U_1^M.

In summary, we have obtained the following displacement measurements from mammography compression:

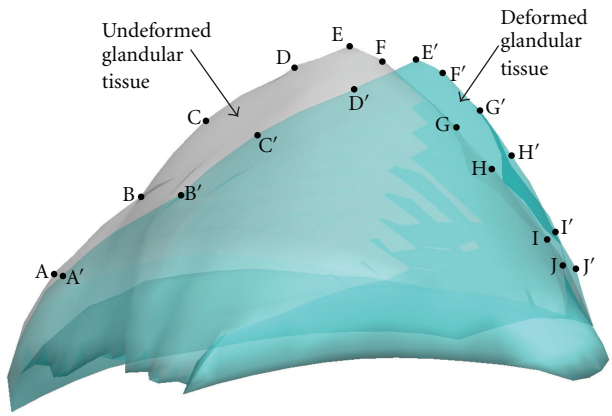


Figure 4: Overlapped mammography-type projections of deformed and undeformed glandular tissue. In the projections, ten nodes A–J on the surface of glandular in underformed projection move to A′–J′ in deformed projection, respectively, giving the projected displacements of these nodes.

(i) the top vertex on the upper breast surface and four vertexes of the tumor; (ii) ten nodes on the fat-glandular interface; (iii) material points directly compressed by the paddles. These displacement measurements are denoted as U_1^M and will be used to identify the material parameters of the tissues.

3.3. Identification of Material Parameters from Displacement Measurements. Having obtained measurement U_1^M from mammography compression, the inverse problem will be conducted to identify the material parameters $\mathbf{p} = \{\lambda_f, \mu_f, \gamma_f, \lambda_g, \mu_g, \gamma_g, \lambda_d, \mu_d, \gamma_d\}$ of the breast tissues and tumor, with use of an iterative optimization procedure (Figure 1). A homogeneous initial guess of $\lambda_0 = 20$, $\mu_0 = 10$, $\gamma_0 = 1$ (λ and μ are dimensionless, γ is in kPa) is used for all the materials. With a trial $\mathbf{p}$, the displacement field u_1 is solved from the FE equation (1) and is projected to $\mathbf{R}u_1$ according to the mammography direction. The difference between prediction $\mathbf{R}u_1$ and measurement U_1^M is evaluated by the objective function $\Phi(\mathbf{p})$ (5). The gradients $\partial\Phi/\partial\mathbf{p}$ are computed with the proposed nonlinear adjoint method. Then, a modified trial $\mathbf{p}$ will be obtained according to the present Φ and $\partial\Phi/\partial\mathbf{p}$ by using L-BFGS minimization subroutine [23]. The iteration continues until a minimization is reached, which corresponds to identified material parameters.

4. Results and Discussion

4.1. Ideal Input. Table 1 shows the initial estimate and reconstructed results, together with the real values for comparison. The results in the first part are based on the ideal input. It is demonstrated that the reconstructed results are very close to the real values. The maximum error is 0.3% (γ for tumor) since the effect of the tumor on surface force measurement is the smallest. Reconstructions using different initial estimates have been conducted and very similar results are found, which indicates the efficiency and uniqueness of the proposed nonlinear elasto-mammography using projective measurements. In our study, all numerical experiments reached convergence and had similar convergent profiles. The iteration speed is related with initial estimations. In clinical practice, the initial estimates could be selected based on data of previous patients and experiments. The more reasonable the initial estimates are, the faster the solver got convergence.

In nonlinear elastography [13] and this study, the same nonlinear material model and properties are applied. For ideal input, both frameworks can get convergence and the reconstructed results are very close to the real values. For input with noises, both frameworks could get convergence and have the similar profiles. The parameters in fatty and glandular tissues get convergence faster than these in tumors because the fatty and glandular tissues have bigger impact on surface deformation and measurement.

Convergent loci of the elastic parameters (λ, μ, γ) is plotted in Figure 5. It is observed that elastic parameters of fatty tissue and glandular tissue approach the real values rapidly. After about 50 iteration steps, their relative errors are well within the range of 5%. Then, they experience some minor adjustment. In contrast, elastic parameters of the tumor converge slower. They start to fall to the real values after 300 steps. After 350 steps, all parameters are accurately identified. Reconstructions using different initial estimates have been conducted. Very similar convergent profiles are found, and equally accurate results are obtained. This indicates uniqueness of the proposed elasto-mammography for nonlinear breast tissue properties and efficiency of the reconstruction algorithm.

The slower convergent speed of elastic parameters of the tumor is explained by the roles they play in the deformation due to the applied loadings, as discussed by Liu et al. [18]. In general, parameters with the most significant influence on the deformation are those most easy to identify. The influence of a parameter depends on size and location of the material region it belongs to, as well as characteristics of the deformation. For the present simulations, elastic parameters of fatty tissue and glandular tissue are dominant; those of tumor are much less influential, due to the small size and deep location of the tumor. So parameters of fatty tissue and glandular tissue are more accurately and easily identified than those of the tumor (Figure 5). Therefore, for successful characterization of the tumor, it is critical to apply deformation modes and acquire displacement data that are most affected by the tumor. In this elasto-mammography simulation, displacements of key points on the tumor are extracted from mammography projections, which increase the accuracy and efficiency to reconstruct the elastic parameters, especially for the tumor.

4.2. Multiple Sets of Measurements. Because of the nonuniqueness nature of most inverse problems, it is important to obtain sufficient measurements to reduce the likelihood of nonuniqueness. For 2D isotropic elastography, Barbone and Bamber [25] have shown that one set of displacement and force measurement, especially when measured only on

Table 1: Initial guess and nonlinear elasto-mammography reconstruction results of the fatty tissue, glandular tissue, and tumor in a 3D breast. The reconstructions are based on ideal mammography measurement, mammography measurement with ±5% and ±10% random noise, respectively. (λ and μ are dimensionless, γ is in kPa.)

	Fatty			Glandular			Tumor		
	λ_f	μ_f	γ_f	λ_g	μ_g	γ_g	λ_d	μ_d	γ_d
Real	35	12.5	0.4	50	25	0.25	80	35	1.5
Guess	20	10	1	20	10	1	20	10	1
				Ideal Input					
Reconstruction	35.00	12.50	0.40	50.00	25.00	0.25	79.83	34.93	1.51
				5% Noise (I)					
Reconstruction	32.95	11.76	0.44	51.82	26.15	0.23	77.12	31.10	1.69
				5% Noise (II)					
Reconstruction	34.82	12.35	0.41	51.62	26.10	0.23	66.14	29.57	1.88
				5% Noise (III)					
Reconstruction	35.9	12.69	0.39	49.67	25.08	0.25	83.75	37.27	1.40
				10% Noise (I)					
Reconstruction	35.14	12.68	0.40	48.87	24.40	0.26	107.59	35.56	1.41
				10% Noise (II)					
Reconstruction	31.89	11.69	0.46	52.17	25.39	0.24	90.29	31.01	1.69
				10% Noise (III)					
Reconstruction	36.75	12.89	0.37	48.30	24.54	0.26	107.20	48.89	0.92

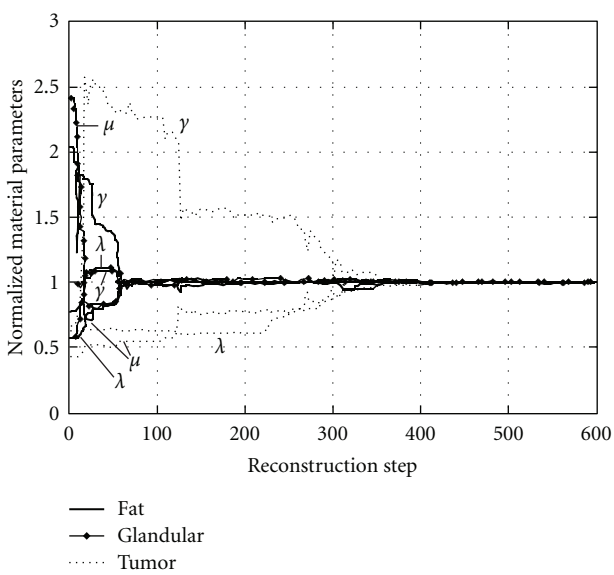


Figure 5: Convergent loci of elasto-mammography reconstruction for elastic parameters (λ, μ, γ) of fatty tissue, glandular tissue, and tumor, normalized with respect to the real values correspondingly.

the boundaries, may not provide sufficient information for reconstruction of the distribution of elastic modulus. To enhance the uniqueness of inverse problems, Barbone and Gokhale [26] proposed the feasibility of using multiple displacement fields, and Liu et al. [18] further discussed the use of multiple sets of measurements in 3D anisotropic media. In our previous nonlinear elastography study [13], measurements from four independent titled compression loadings were used to insure stable and unique material parametric reconstruction. In this work, we applied only projective measurements from two breast compression tests and found that the acquired displacement and force data are sufficient for stable parametric reconstruction, even for the small and deeply embedded tumor. This is a significant reduction, as it increases the clinical efficiency, reduces X-ray dose and operation cost, and benefits the patients.

The reduction of necessary loadings is possible because mammography projection provides displacement on the surface of the tumor, which contains direct information of the mechanics of the tumor. Our previous nonlinear elastography study [13] takes only measurement on the breast surface as input. The lack of necessary constitutive information of the tumor in the surface measurement must be compensated by increasing the number of required loadings. In case that the measurement may contain experimental errors, we must use four loadings in the elastography study, instead of two in the present elasto-mammography.

4.3. Iteration Steps. The nonlinear elasto-mammography reconstruction uses an iterative optimization procedure (Figure 1), which is controlled by user-defined criteria. This study employs more strict criteria than in our previous work [13], and it takes about 590 steps to reach the converged reconstruction results. To demonstrate the intermediate results, the uniaxial tensile strain-stress curves of the tumor predicted by the updated material parameters are plotted in Figure 6 at the 1st, 100th, 200th, 300th, and 592nd iterative steps and compared to the real one. It is observed that the reconstructed strain-stress curve approaches the real one rapidly in first 300 iterative steps. After that, the reconstruction only applies some minor adjustment.

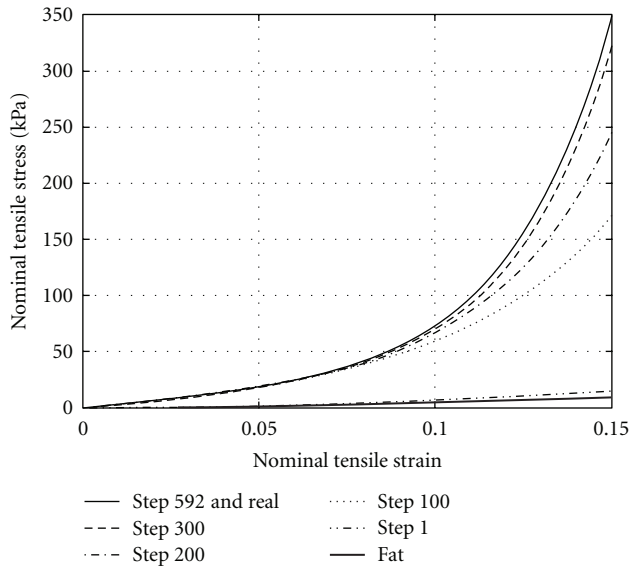


Figure 6: Nonlinear tensile strain-stress curves of the tumor as reconstructed at different iteration steps.

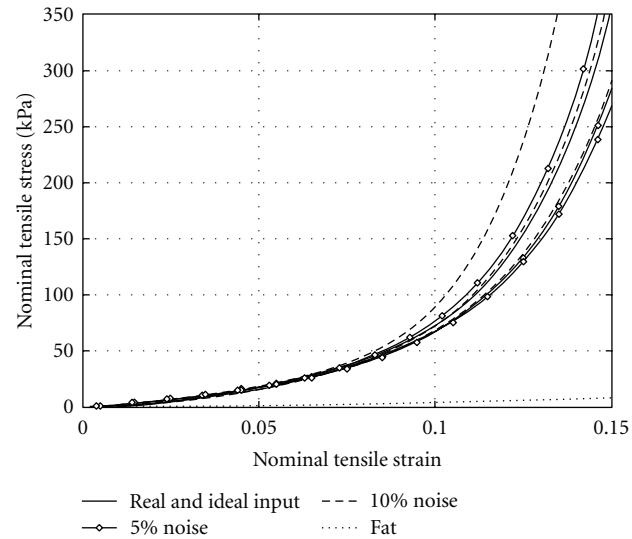


Figure 7: Nonlinear tensile strain-stress curves of the tumor as reconstructed from inputs with 5% and 10% noise.

In clinical practice, a more tolerable criterion may be applied to control the iterative reconstruction procedure to save computational expense and time. It has been recognized that tissue stiffness plays an important role for diagnosis of breast cancers, as tumors are stiffer than that surrounding breast tissues, and malignant tumors are much stiffer than benign ones [6]. In another word, the stiffness ratio between fatty tissue and tumor, instead of real material parameters, could be used to determine the character of tumors. It is observed in Figure 6 that, starting from the 100th iterative step, the stiffness ratio of tumor to fatty tissue (the lowest curve) increases rapidly, indicating that the predicted mechanical properties of the tumor are well distinguished from the normal tissues for characterizing the tumor. Therefore, from clinical point of view, the iterative reconstruction procedure could be stopped after about 100 steps.

4.4. Input with Noise. The above elasto-mammography reconstructions are conducted using ideal inputs. However, noise is unavoidable in experimental data. To investigate the capability of the proposed nonlinear elasto-mammography modality and algorithm to handle imperfect experimental data, we conduct reconstruction using noisy input, where a randomly selected relative error between ±5% or ±10% is added to each displacement data in U_1^M. For each noise level, three case studies are conducted. The results are shown as noise 5% (I)–(III) and noise 10% (I)–(III) in Table 1, and the reconstructed tensile strain-stress curves of the tumor are plotted in Figure 7.

It is observed that the strain-stress curves reconstructed with noisy input have similar shape to the ones with ideal input. It is not surprising that curves with 5% noise are closer to the real one than these curves with 10% noise. It demonstrated that, in order to get robust results, we need to make effort to decrease the noise in displacement measurements. It is noted that all the predicted strain-stress curves of tumor, with or without measurement noise, are well distinguished from the curve of fatty tissue (the lowest curve in Figure 7); that is, being much stiffer. That is, even though measurement noise exists, the tumor can be identified by recognizing the difference of stiffness between tumors and the surrounding tissues. This demonstrates that the nonlinear elasto-mammography results are accurate enough for diagnosis of tumors in clinical application.

The previous nonlinear elastography based on surface measurement [13] fails to reconstruct material parameters when ±5% random noise is added to the input. A regularization is required to provide additional constrain. In comparison, the present elasto-mammography yields accurate enough material parameters even with ±10% random noise. The reason is, as mentioned in Sections 4.1 and 4.2, that the displacements extracted on the surface of the tumor from mammography projections contain direct information of the mechanical properties of the tumor, which enhances the robustness of reconstruction and increases the accuracy, in particular of the tumor's parameters.

4.5. Advantages of Nonlinear Elasto-Mammography. In this study, a nonlinear elasto-mammography framework is developed to incorporate the conventional X-ray mammography for characterization of breast tissue properties. This work extends our previous study linear elasto-mammography [12] and nonlinear elastography [13]. Comparing with previous study, nonlinear elasto-mammography has the following three major advantages.

Imaging techniques: an imaging technique should be selected to measure deformation in elastography. In the proposed nonlinear elasto-mammography, the deformation is measured by conventional X-ray mammography while

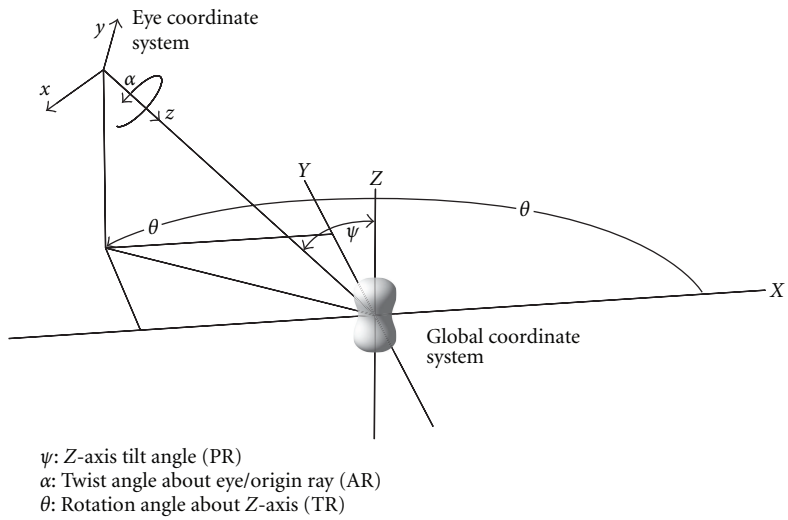


Figure 8: Illustration of global coordinate and eye coordinate. An object in global coordinates $[X, Y, Z]$ is projected in eye coordinates $[x, y, z]$. The relation between direction vectors is dependent on ψ, α, and θ.

USE or MRI is applied in nonlinear elastography. Traditional X-ray has advantages of low cost and high resolution, compared with USE and MRI.

Deformation theory: the linear elasto-mammography framework is based on infinitesimal strain deformation theory. However, it is well known that the mechanical behavior of biological soft tissue is nonlinear. In nonlinear elasto-mammography, nonlinear material model and deformation theory are applied so that more accurate results could be obtained.

Inversion techniques: once displacements are measured, an inversion technique is applied to reconstruct elastic properties. In linear elasto-mammography, an adjoint method is applied and then a nonlinear adjoint method is developed for nonlinear elastography. In this study, the nonlinear adjoint method is further improved to enhance the numerical efficiency and stability of reconstruction of elastic properties.

Therefore, the proposed nonlinear elasto-mammography framework has advantage of imaging techniques, deformation theory, and inversion techniques. It combines the simplicity of projective X-ray mammography measurement with the accuracy of nonlinear elastography.

5. Summary

This study presents a nonlinear elasto-mammography method that combines elastography reconstruction and X-ray mammography imaging for the purpose of diagnosis of breast tumors by identification of the finite-strain mechanical parameters of breast tissues and tumors. The displacement information of selected material points is extracted from mammography projections before and after breast compression. Correspondingly, the previously developed nonlinear elastography algorithm has been adjusted with a revised adjoint gradient method to incorporate projection-type displacement measurement. The simulations with heterogeneous breast phantom proved the feasibility of elasto-mammography and tested the efficiency and robustness of the reconstruction algorithm. The simulations show that the deformation of the tumor, depicted by the projected displacement on the surface of the tumor extracted from mammography images, is critical for the success of elasto-mammography reconstruction.

Appendices

A. Displacement Transition between Coordinate Systems

This appendix presents the transition of a displacement vector between global coordinates and projection coordinates. The outcome is the projection matrix $\mathbf{R}$ in formulas (5) and (7).

To be consistent with computational geometry, we call the projection coordinates as *eye coordinates*. As illustrated in Figure 8, the global coordinates are denoted as $[X, Y, Z]$ and eye coordinates are $[x, y, z]$. Their direction vectors are $[\mathbf{e}_X, \mathbf{e}_Y, \mathbf{e}_Z]^T$ and $[\mathbf{e}'_x, \mathbf{e}'_y, \mathbf{e}'_z]^T$, respectively. As shown in Figure 8, the eye coordinates rotate from global coordinates by three angles: Z-axis tilt angle ψ, twist angle about eye/original ray α, and rotation angle about Z-axis θ. It can be shown that

$$\begin{pmatrix} \mathbf{e}'_x \\ \mathbf{e}'_y \\ \mathbf{e}'_z \end{pmatrix} = [Q] \begin{pmatrix} \mathbf{e}_x \\ \mathbf{e}_y \\ \mathbf{e}_z \end{pmatrix}, \tag{A1}$$

where the rotation matrix Q is

$$[Q] = \begin{bmatrix} \cos\alpha\cdot\cos\theta\cdot\cos\psi - \sin\alpha\cdot\sin\theta & \cos\theta\cdot\sin\alpha + \cos\alpha\cdot\cos\psi\cdot\sin\theta & \cos\alpha\cdot\sin\psi \\ -\cos\theta\cdot\cos\psi\cdot\sin\alpha - \cos\alpha\cdot\sin\theta & \cos\alpha\cdot\cos\theta - \cos\psi\cdot\sin\alpha\cdot\sin\theta & -\sin\alpha\cdot\sin\psi \\ \cos\theta\cdot\sin\psi & \sin\theta\cdot\sin\psi & -\cos\psi \end{bmatrix}. \tag{A2}$$

Now, consider a displacement vector $\mathbf{u}$ of a material point from undeformed position to deformed position. The global coordinates of $\mathbf{u}$ are $\{u_X, v_Y, w_Z\}$, the eye coordinates are $\{u_x, v_y, w_z\}$, and their relationship can be derived as

$$\begin{pmatrix} u_x \\ v_y \\ w_z \end{pmatrix} = [Q] \begin{pmatrix} u_X \\ v_Y \\ w_Z \end{pmatrix}. \tag{A3}$$

In mammography projection, the displacement component in $\mathbf{e}'_z$ direction, w_z, is not obtainable, and only u_x and v_y are measured. Therefore, (A3) reduces to

$$\begin{pmatrix} u_x \\ v_y \end{pmatrix} = \underbrace{\begin{bmatrix} \cos\alpha\cdot\cos\theta\cdot\cos\psi - \sin\alpha\cdot\sin\theta & \cos\theta\cdot\sin\alpha + \cos\alpha\cdot\cos\psi\cdot\sin\theta & \cos\alpha\cdot\sin\psi \\ -\cos\theta\cdot\cos\psi\cdot\sin\alpha - \cos\alpha\cdot\sin\theta & \cos\alpha\cdot\cos\theta - \cos\psi\cdot\sin\alpha\cdot\sin\theta & -\sin\alpha\cdot\sin\psi \end{bmatrix}}_{[Q']} \begin{pmatrix} u_X \\ v_Y \\ w_Z \end{pmatrix} = [Q'] \begin{pmatrix} u_X \\ v_Y \\ w_Z \end{pmatrix}. \tag{A4}$$

Finally, the FE solution of displacement field u_1, when projected, becomes $\mathbf{R}u_1$ where $\mathbf{R}$ is the assemble of $[Q']$ according to the FE discretization and assembling methods.

B. Adjoint Method for Gradients of Objective Function

Direct calculation of the gradients $\partial\Phi/\partial\mathbf{p}$ of the objective function involved in the minimization-based parametric identification is difficult, because u_1 is an implicated function of $\mathbf{p}$. An adjoint method will be derived here for efficient and analytical calculation of the gradients. To release the implicit coupling between u_1 and $\mathbf{p}$, we introduce the constraint (1) into the objective function (5) and obtain a Lagrangian:

$$L = \left(\mathbf{R}u_1 - U_1^M\right)^T \mathbf{\Lambda}\left(\mathbf{R}u_1 - U_1^M\right) + \begin{Bmatrix} w_1 \\ w_2 \end{Bmatrix}^T \begin{Bmatrix} f_1^{\text{in}} - f_1^{\text{out}} \\ f_2^{\text{in}} - f_2^{\text{out}} \end{Bmatrix}, \tag{B1}$$

where w_1 and w_2 are arbitrary virtual displacements. In this Lagrangian, u_1 and $\mathbf{p}$ are explicit variables and are no longer coupled. It is noted that $\Phi = L$ and $\delta\Phi = \delta L$ for arbitrary w_1 and w_2 under the constraint (1). The variation δL can be expressed as

$$\begin{aligned} \delta L = {} & 2\left(\mathbf{R}u_1 - U_1^M\right)^T \mathbf{\Lambda}(\mathbf{R}\delta u_1) \\ & + \left(w_1^T K_{11}^{\text{in}} - w_1^T K_{11}^{\text{out}} + w_2^T \frac{\partial f_2^{\text{in}}}{\partial u_1} - w_2^T \frac{\partial f_2^{\text{out}}}{\partial u_1}\right)\delta u_1 \\ & + w_1^T \frac{\partial f_1^{\text{in}}}{\partial \mathbf{p}}\delta\mathbf{p} + w_2^T \frac{\partial f_2^{\text{in}}}{\partial \mathbf{p}}\delta\mathbf{p} - w_2^T \frac{\partial f_2^{\text{out}}}{\partial \mathbf{p}}\delta\mathbf{p} \end{aligned} \tag{B2}$$

for which the equality constraint (1) has been applied. Note that the prescribed external force f_1^{out} is independent of $\mathbf{p}$. Equation (B2) can be further simplified by letting the arbitrary virtual displacement $w_2 = 0$, as

$$\begin{aligned} \delta L = {} & \left\{2\left(\mathbf{R}u_1 - U_1^M\right)^T \mathbf{\Lambda}\mathbf{R} + w_1^T K_{11}^{\text{in}} - w_1^T K_{11}^{\text{out}}\right\}\delta u_1 \\ & + w_1^T \frac{\partial f_1^{\text{in}}}{\partial \mathbf{p}}\delta\mathbf{p}. \end{aligned} \tag{B3}$$

If we select a w_1 to let $\{2(\mathbf{R}u_1 - U_1^M)^T\mathbf{\Lambda}\mathbf{R} + w_1^T K_{11}^{\text{in}} - w_1^T K_{11}^{\text{out}}\}\delta u_1 = 0$ for arbitrary δu_1, we obtain a simplest form of δL, as

$$\delta L = w_1^T \frac{\partial f_1^{\text{in}}}{\partial \mathbf{p}}\delta\mathbf{p} = \left(w_1^T \frac{\partial f_1^{\text{in}}}{\partial \mathbf{p}} + w_2^T \frac{\partial f_2^{\text{in}}}{\partial \mathbf{p}}\right)\delta\mathbf{p} \quad (w_2 = 0). \tag{B4}$$

Consider that $\delta\Phi = \delta L$ for arbitrary w_1 and w_2, we obtain (6) in the text with the following selection of w_1 and w_2:

$$\begin{gathered} \left(K_{11}^{\text{in}} - K_{11}^{\text{out}}\right)w_1 = K_{11}^{\text{eff}} w_1 = -2\mathbf{R}^T\mathbf{\Lambda}\left(\mathbf{R}u_1 - U_1^M\right), \\ w_2 = 0 \end{gathered} \tag{B5}$$

which is (7) in the text.

By introducing the adjoint method, it seems that more equations (B5) and variables (w_1 and w_2) are involved. But the solution of (B5) is straightforward and the computational cost is minimal, because K_{11}^{eff} has been computed and factorized when solving for the displacement u_1 as in (3).

The gradients $\partial\Phi/\partial\mathbf{p}$ can also be calculated directly as

$$\frac{\partial\Phi}{\partial\mathbf{p}} = 2\left(\mathbf{R}u_1 - U_1^M\right)^T \mathbf{\Lambda}\mathbf{R}\frac{\partial u_1}{\partial\mathbf{p}}, \tag{B6}$$

in which $\partial u_1/\partial \mathbf{p}$ can be computed numerically using finite-different method:

$$\frac{\partial u_1}{\partial \mathbf{p}} \approx \frac{u_1(\mathbf{p}+\delta\mathbf{p}) - u_1(\mathbf{p})}{\delta\mathbf{p}} \quad (\delta\mathbf{p} \text{ is a small change of } \mathbf{p}) \tag{B7}$$

or analytically by solving linear equations:

$$K_{11}^{\text{eff}} \frac{\partial u_1}{\partial \mathbf{p}} = -\frac{\partial f_1^{\text{in}}}{\partial \mathbf{p}}. \tag{B8}$$

For finite-strain nonlinear problem, the finite-different method is unaffordable due to the high computational expense to solve (1) for u_1. Solving (B8) is straightforward and is much less expensive for K_{11}^{eff} has been computed and factorized. However, (B8) needs to be solved for every material parameters involved; for example, in the exemplar simulations in this work, it needs to be solved nine times because each material has three parameters. In comparison, the proposed adjoint method (B5), (B6) requires only one solution for w_1, regardless of the number of material parameters involved.

Acknowledgment

This work is supported by the US Army's Breast Cancer Research Program Concept Award (W81XWH-05-1-0461).

References

[1] S. J. Nass, I. C. Henderson, and J. C. Lashof, *Mammography and Beyond: Developing Technologies for the Early Detection of Breast Cancer*, National Academy Press, Washington, DC, USA, 2001.

[2] B. Boné, Z. Péntek, L. Perbeck, and B. Veress, "Diagnostic accuracy of mammography and contrast-enhanced MR imaging in 238 histologically verified breast lesions," *Acta Radiologica*, vol. 38, no. 4, pp. 489–496, 1997.

[3] M. L. Giger, Z. Huo, and C. J. Vyborny, "Computer-aided diagnosis in mammography," in *Handbook of Medical Imaging*, M. Sonka and J. M. Fitzpatrick, Eds., vol. 2, pp. 915–1004, SPIE Press, 2000.

[4] P. J. Kornguth and R. C. Bentley, "Mammographic-pathologic correlation: part 1. Benign breast lesions," *Journal of Women's Imaging*, vol. 3, no. 1, pp. 29–37, 2001.

[5] A. P. Sarvazyan, A. R. Skovoroda, S. Y. Emelianov et al., "Biophysical based of elasticity imaging," *Acoustical Imaging*, vol. 21, pp. 223–240, 1995.

[6] P. Wellman, *Tactile imaging*, Ph.D. thesis, Harvard University, 1999.

[7] A. R. Skovorda, A. N. Klishko, D. A. Gusakian et al., "Quantitative analysis of mechanical characteristics of pathologically altered soft biological tissues," *Biofizika*, vol. 40, no. 6, pp. 1335–1340, 1995.

[8] J. Ophir, I. Cespedes, H. Ponnekanti, Y. Yazdi, and X. Li, "Elastography: a quantitative method for imaging the elasticity of biological tissues," *Ultrasonic Imaging*, vol. 13, no. 2, pp. 111–134, 1991.

[9] J. Ophir, S. K. Alam, B. Garra et al., "Elastography: ultrasonic estimation and imaging of the elastic properties of tissues," *Proceedings of the Institution of Mechanical Engineers H*, vol. 213, no. 3, pp. 203–233, 1999.

[10] R. Muthupillai, D. J. Lomas, P. J. Rossman, J. F. Greenleaf, A. Manduca, and R. L. Ehman, "Magnetic resonance elastography by direct visualization of propagating acoustic strain waves," *Science*, vol. 269, no. 5232, pp. 1854–1857, 1995.

[11] A. L. McKnight, J. L. Kugel, P. J. Rossman, A. Manduca, L. C. Hartmann, and R. L. Ehman, "MR elastography of breast cancer: preliminary results," *American Journal of Roentgenology*, vol. 178, no. 6, pp. 1411–1417, 2002.

[12] Z. G. Wang, Y. Liu, L. Z. Sun, and L. L. Fajardo, "Elastomammography: theory, algorithm, and phantom study," *International Journal of Biomedical Imaging*, vol. 2006, Article ID 53050, 11 pages, 2006.

[13] Z. G. Wang, Y. Liu, G. Wang, and L. Z. Sun, "Elastography method for reconstruction of nonlinear breast tissue properties," *International Journal of Biomedical Imaging*, vol. 2009, Article ID 406854, 9 pages, 2009.

[14] Y. C. Fung, *Biomechanics-Mechanical Properties of Living Tissues*, Springer, 1993.

[15] M. A. Zulliger, P. Fridez, K. Hayashi, and N. Stergiopulos, "A strain energy function for arteries accounting for wall composition and structure," *Journal of Biomechanics*, vol. 37, no. 7, pp. 989–1000, 2004.

[16] J. J. O'Hagan and A. Samani, "Measurement of the hyperelastic properties of tissue slices with tumour inclusion," *Physics in Medicine and Biology*, vol. 53, no. 24, pp. 7087–7106, 2008.

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, Cambridge University Press, New York, NY, USA, 2nd edition, 1996.

[18] Y. Liu, L. Z. Sun, and G. Wang, "Tomography-based 3-D anisotropic elastography using boundary measurements," *IEEE Transactions on Medical Imaging*, vol. 24, no. 10, pp. 1323–1333, 2005.

[19] N. Tardieu and A. Constantinescu, "On the determination of elastic coefficients from indentation experiments," *Inverse Problems*, vol. 16, no. 3, pp. 577–588, 2000.

[20] A. A. Oberai, N. H. Gokhale, and G. R. Feijóo, "Solution of inverse problems in elasticity imaging using the adjoint method," *Inverse Problems*, vol. 19, no. 2, pp. 297–313, 2003.

[21] Y. Liu, G. Wang, and L. Z. Sun, "Anisotropic elastography for local passive properties and active contractility of myocardium from dynamic heart imaging sequence," *International Journal of Biomedical Imaging*, vol. 2006, Article ID 45957, 15 pages, 2006.

[22] A. A. Oberai, N. H. Gokhale, M. M. Doyley, and J. C. Bamber, "Evaluation of the adjoint equation based algorithm for elasticity imaging," *Physics in Medicine and Biology*, vol. 49, no. 13, pp. 2955–2974, 2004.

[23] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1–3, pp. 503–528, 1989.

[24] A. Samani and D. Plewes, "A method to measure the hyperelastic parameters of ex vivo breast tissue samples," *Physics in Medicine and Biology*, vol. 49, no. 18, pp. 4395–4405, 2004.

[25] P. E. Barbone and J. C. Bamber, "Quantitative elasticity imaging: what can and cannot be inferred from strain images," *Physics in Medicine and Biology*, vol. 47, no. 12, pp. 2147–2164, 2002.

[26] P. E. Barbone and N. H. Gokhale, "Elastic modulus imaging: on the uniqueness and nonuniqueness of the elastography inverse problem in two dimensions," *Inverse Problems*, vol. 20, no. 1, pp. 283–296, 2004.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2011, Article ID 592924, 20 pages
doi:10.1155/2011/592924

Research Article

Contour Detection and Completion for Inpainting and Segmentation Based on Topological Gradient and Fast Marching Algorithms

Didier Auroux,[1] Laurent D. Cohen,[2] and Mohamed Masmoudi[3]

[1] *Laboratoire J. A. Dieudonné, Université de Nice Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 2, France*
[2] *CEREMADE, UMR CNRS 7534, Université Paris Dauphine, Place du Marchal De Lattre De Tassigny, 75775 Paris Cedex 16, France*
[3] *Institut de Mathématiques de Toulouse, Université Paul Sabatier, 31062 Toulouse Cedex 9, France*

Correspondence should be addressed to Didier Auroux, auroux@unice.fr

Received 30 May 2011; Revised 12 September 2011; Accepted 12 September 2011

Academic Editor: Shan Zhao

We combine in this paper the topological gradient, which is a powerful method for edge detection in image processing, and a variant of the minimal path method in order to find connected contours. The topological gradient provides a more global analysis of the image than the standard gradient and identifies the main edges of an image. Several image processing problems (e.g., inpainting and segmentation) require continuous contours. For this purpose, we consider the fast marching algorithm in order to find minimal paths in the topological gradient image. This coupled algorithm quickly provides accurate and connected contours. We present then two numerical applications, to image inpainting and segmentation, of this hybrid algorithm.

1. Introduction

Contour detection is a major issue in image processing. For instance, in classification and segmentation, the goal is to split the image into several parts. This problem is strongly related to the detection of the connected contours separating these parts. It is quite easy to detect edges using local image analysis techniques, but the detection of continuous contours is more complicated and needs a global analysis of the image.

Several image processing problems like image inpainting and denoising (or enhancement) are classically solved without detecting edges and contours. The goal of image enhancement is to denoise the image without blurring it. A classical idea is to identify the edges in order to preserve them and to smooth the image outside them. In this particular case, contour completion is not prerequisite, as the quality of the result is not too much related to the completeness of the identified edges, but missing edges may lead to blurred boundaries. For most of the other image processing problems (segmentation, inpainting, classification), the detection of connected contours can drastically simplify the resolution and improve the quality of the results. For instance, the image segmentation problem is a very good example, as the goal is to split the image into its characteristic parts. In other words, one has to find connected contours, which define different subsets of the image.

For solving all these problems, various approaches have been considered in the literature. We can cite here the most commonly used models: the structural approach by region growing [1], the stochastic approaches [2–4], and the variational approaches, which are based on various strategies like level set formulations, minimizing the total variation of a quantity or the Mumford-Shah functional, active contours and geodesic active contours methods, snakes, wavelet transforms, or shape gradient [5–19, 19–24].

Another approach is based on the topological asymptotic analysis and consists of defining edges as cracks [25, 26]. The goal of topological optimization is to look for an optimal design (i.e., a subset) and its complementary. Finding the optimal subdomain is equivalent to identifying its characteristic function. At first sight, this problem is not differentiable. But the topological asymptotic expansion gives the variation of a cost function $j(\Omega)$ (see Section 2 for examples) when one switches the characteristic function from one to zero (or from zero to one) in a small region [27].

More precisely, we consider the perturbation of the main domain Ω by the insertion of a small crack (or hole) $\sigma_\rho : \Omega_\rho = \Omega \setminus \sigma_\rho$, ρ being the size of the crack. The topological sensitivity theory provides then an asymptotic expansion of the considered cost function when the size of the crack tends to zero. It takes the general form: $j(\Omega_\rho) - j(\Omega) = f(\rho)g(x) + o(f(\rho))$, where $f(\rho)$ is an explicit positive function going to zero with ρ, and $g(x)$ is the topological gradient at point x. Then, in order to minimize the criterion (or at least its first order expansion), one has to insert small cracks at points where the topological gradient is the most negative. Using this gradient type information, it is possible to build fast algorithms. In most applications, a satisfying approximation of the optimal solution is reached at the first iteration of the optimization process. A topological sensitivity framework allowing to obtain such an expansion for general cost functions has been proposed in [27].

An efficient edge detection technique, based on the topological gradient, has been introduced in [28, 29]. It is also shown that edge detection can make all these image processing problems straightforward to solve [25, 26, 30, 31]. But the identified edges are usually not connected, and the results can be degraded. Our goal is to improve these results by replacing dashed discontinuous edges by connected contours.

In the inpainting problem, we assume that there is a hidden part of the image, and our goal is to recover this part from the known part of the image. We assume that the missing part is a quite large part of the image, we do not consider the case of random sets or narrow lines. This problem has been widely studied and the most common approaches are: learning approches (neural networks, radial basis functions, ...) [32, 33], minimization of an energy cost function based on a total variation norm [34, 35], morphological component analysis methods separating texture and cartoon [36]. We also refer to [6, 8] for the description of several inpainting algorithms.

We now consider the crack detection technique, within the framework of the identification of the image edges, either in the hidden part of the image for the inpainting application, or in the whole image for the segmentation application [26]. The topological asymptotic analysis provides very quickly the location of the edges, as they are precisely defined by the most negative points of the topological gradient. The great advantage of the topological gradient in comparison with level line completion and TV-based inpainting methods (see e.g., [6, 8, 12, 13]) is that the identified edges in the unknown part of the image correspond to a regular extrapolation of the known edges, and as we will see on a numerical example, the topological gradient preserves the continuity of the edge curvature. Thus, the proposed approach is much more than simple edge detection.

The main issue of the approach based on the topological gradient is the need for connected complete contours. This can be easily understood since the hidden part of the image is filled in using the Laplace operator in each subdomain of the missing zone, and a discontinuous contour would lead to some blurred reconstruction. Up to now, one had to threshold the topological gradient with a not too small value, in order to identify connected contours, but this leads to thick identified edges, and also to consider more noisy points as potential edges. In order to overcome this limitation, we consider a minimal path technique in order to connect the edges identified by the topological gradient.

Minimal paths have been first introduced for finding the global minimum of active contour models, using the fast marching technique [37, 38]. They have then been used to find contours or tubular structures and also for perceptual grouping using a path or a set of paths minimizing a functional [38–43]. In our case, the energy to be minimized will be an increasing function of the topological gradient. As the topological gradient takes its minimal (negative) values on the edges of the image, the idea is indeed to find contours for contour completion from the various minima and small values of the topological gradient.

The energy to be minimized can be seen as a distance function. The idea is then to compute this distance function between a given starting point and all other points. For this purpose, a front propagation equation is considered. Using the fast marching propagation, the definition of the distance function is straightforward: the distance between a point x and the starting point is exactly the time at which the front reached x. Then, minimal paths between these points can be identified using a gradient descent. For perceptual grouping, a set of keypoints is considered as starting points and a set of minimal paths connecting some pairs of these keypoints is considered as a contour completion. This approach is extremely satisfactory in 2D problems, with quite few key points. It is also extremely fast. In 3D images, minimal paths find tubular structures, but in order to identify minimal surfaces, this approach is much more difficult to consider. It was dealt in the case of a surface connecting two curves in [44]. We only consider here the 2D case.

The application of the minimal path technique to the topological gradient allows us to obtain an automatic identification of the main (missing or not) edges of the image. These edges will be continuous, by construction, and will allow us to simply apply the Laplace operator to fill in the image for inpainting applications, or will directly provide the segmented image, with very good results. Another advantage of this technique is to be very fast, as it does not degrade the $\mathcal{O}(n \cdot \log(n))$ complexity of the topological gradient based algorithm introduced in [26]. We refer to [26, 45] for the inpainting and segmentation algorithms by topological asymptotic expansion, and for a detailed presentation of the topological gradient.

The paper is organized as follows. In Section 2, we present the edge detection method using the topological gradient, and the corresponding segmentation and inpainting algorithms. In Section 3, we propose an algorithm based on the minimal path and fast marching techniques in order to identify the valley lines of the topological gradient, which correspond to the main edges of the image. Then, we report the results of several numerical experiments in Section 4. We also compare this hybrid scheme with the fast marching algorithm applied to the standard gradient. Two particular image processing problems are considered: segmentation and inpainting. Finally, some conclusions are given in Section 5.

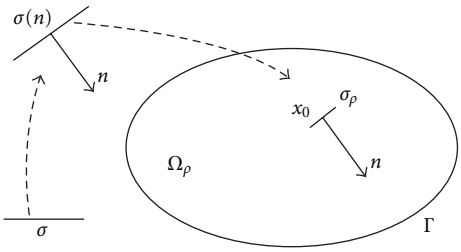


Figure 1: Example of domain and inserted crack (with its orientation).

2. Edge Detection by Topological Asymptotic Analysis and Its Application to Inpainting and Segmentation

2.1. Topological Asymptotic Analysis. Let Ω be an open bounded domain of $\mathbb{R}^2$ (note that it can easily be extended to $\mathbb{R}^n$). We consider a partial differential equation (PDE) problem defined in Ω, and we denote by u_Ω its solution (we will further see under which assumptions it can be considered). We finally consider a cost function $J(\Omega, u_\Omega)$ to be minimized, where u_Ω is the solution to the PDE in Ω. The idea of topological asymptotic analysis is to measure the impact of a perturbation of the domain Ω on the cost function.

For a small $\rho \geq 0$, let $\Omega_\rho = \Omega \setminus \sigma_\rho$ be the perturbed domain by the insertion of a crack $\sigma_\rho = x_0 + \rho\sigma(n)$, where $x_0 \in \Omega$. We denote by σ a fixed bounded straight crack containing the origin, n is a unit vector, and $\sigma(n)$ is the result of the rotation of σ so that n is the normal to $\sigma(n)$. The fixed crack σ is rotated (normal n), stretched (size ρ), and translated (center x_0) in order to get σ_ρ (see Figure 1). The topological gradient theory can also be applied in the case of arbitrary shaped holes [46–49], but we will only consider the case of crack perturbations in our applications. The small parameter ρ will represent the size of the inserted crack. Finally, we denote by $\mathcal{V}$ a Hilbert space on Ω, usually $H^1(\Omega)$ in our applications.

We now consider the variational formulation of the PDE problem on Ω

$$\begin{gathered}\text{Find } u \in \mathcal{V} \text{ such that} \\ a(u, w) = l(w), \quad \forall w \in \mathcal{V},\end{gathered} \tag{1}$$

and the corresponding variational formulation of the PDE problem on the perturbed domain

$$\begin{gathered}\text{Find } u_\rho \in \mathcal{V}_\rho \text{ such that} \\ a_\rho\left(u_\rho, w\right) = l_\rho(w), \quad \forall w \in \mathcal{V}_\rho.\end{gathered} \tag{2}$$

One should notice that for $\rho = 0$, the perturbed PDE problem becomes the original PDE problem.

We assume in the following that a_ρ is a bilinear continuous and coercive form defined on $\mathcal{V}_\rho$, a Hilbert space on Ω_ρ, and that l_ρ is a linear continuous form on $\mathcal{V}_\rho$.

We can rewrite the cost function J as a function of ρ by considering the following map:

$$\begin{aligned} j : \rho \longmapsto \Omega_\rho \longmapsto u_\rho, \text{ solution of Equation (2)} \longmapsto j(\rho) \\ := J\left(\Omega_\rho, u_\rho\right). \end{aligned} \tag{3}$$

In order to apply the topological asymptotic theory, a_ρ, l_ρ, and J have to satisfy the hypotheses of the following result [50, 51].

If there exist a linear form L_ρ defined on $\mathcal{V}_\rho$, a function $f : \mathbb{R}^+ \to \mathbb{R}^+$, and four real numbers δJ_1, δJ_2, δa, and δl such that

(1) $\lim_{\rho \to 0} f(\rho) = 0$,

(2) $J(\Omega_\rho, u_\rho) - J(\Omega_\rho, u_0) = L_\rho(u_\rho - u_0) + f(\rho)\delta J_1 + o(f(\rho))$,

(3) $J(\Omega_\rho, u_0) - J(\Omega, u_0) = f(\rho)\delta J_2 + o(f(\rho))$,

(4) $(a_\rho - a_0)(u_0, p_\rho) = f(\rho)\delta a + o(f(\rho))$,

(5) $(l_\rho - l_0)(p_\rho) = f(\rho)\delta l + o(f(\rho))$,

where the adjoint state p_ρ is solution of the adjoint equation

$$a_\rho\left(w, p_\rho\right) = -L_\rho(w) \quad \forall w \in \mathcal{V}_\rho, \tag{4}$$

and u_ρ is solution of the direct (2), then the cost function has the following asymptotic expansion:

$$j(\rho) - j(0) = f(\rho)g(x) + o(f(\rho)), \tag{5}$$

where $g(x)$ is the topological gradient, given by

$$g(x) = \delta J_1 + \delta J_2 + \delta a - \delta l. \tag{6}$$

Indeed, from second and third items, $j(\rho) - j(0) = J(\Omega_\rho, u_\rho) - J(\Omega, u_0) = L_\rho(u_\rho - u_0) + f(\rho)(\delta J_1 + \delta J_2) + o(f(\rho))$. From the definition of the adjoint state and the direct equation, $L_\rho(u_\rho - u_0) = -a_\rho(u_\rho, p_\rho) + a_\rho(u_0, p_\rho)$. From fourth item and direction (2), $-a_\rho(u_\rho, p_\rho) + a_\rho(u_0, p_\rho) = -l_\rho(p_\rho) + a_0(u_0, p_\rho) + f(\rho)\delta a + o(f(\rho)) = -l_\rho(p_\rho) + l_0(p_\rho) + f(\rho)\delta a + o(f(\rho))$. Finally, from fifth item, this term is equal to $f(\rho)(\delta a - \delta l) + o(f(\rho))$.

Then, from an asymptotic point of view, as $f(\rho) \geq 0$, the idea is to create cracks in the domain Ω, where the topological gradient g is the most negative, because

$$J\left(\Omega_\rho, u_\rho\right) = J(\Omega, u) + f(\rho)g(x) + o(f(\rho)), \tag{7}$$

and the cost function corresponding to the perturbed problem will be smaller than the original one. The main advantage of this method is that it only requires the resolution of the direct (2) and adjoint (4) problems.

2.2. Application to Edge Detection. Let Ω be an open bounded domain of $\mathbb{R}^2$, representing the image domain. For a given function v in $L^2(\Omega)$ (in our application, v represents the input image), the initial problem is defined on the unperturbed domain and reads as follows: find $u \in H^1(\Omega)$ such that

$$\begin{gathered} -\operatorname{div}(c\nabla u) + u = v \quad \text{in } \Omega, \\ \partial_n u = 0 \quad \text{on } \partial\Omega, \end{gathered} \tag{8}$$

where n denotes the outward unit normal to $\partial\Omega$ and c is a given function. Note that this problem is equivalent to linear diffusion restoration.

For a given $x_0 \in \Omega$ and a small $\rho \geq 0$, let us now consider $\Omega_\rho = \Omega \setminus \sigma_\rho$ the perturbed domain by the insertion of a crack $\sigma_\rho = x_0 + \rho\sigma(n)$, where $x_0 \in \Omega$, $\sigma(n)$ is a straight crack, and n a unit vector normal to the crack. Then, the new solution $u_\rho \in H^1(\Omega_\rho)$ satisfies

$$\begin{aligned} -\operatorname{div}\left(c\nabla u_\rho\right) + u_\rho &= v \quad \text{in } \Omega_\rho, \\ \partial_n u_\rho &= 0 \quad \text{on } \partial\Omega_\rho. \end{aligned} \tag{9}$$

Edge detection is equivalent to looking for a subdomain of Ω in which the energy is small. Indeed, we consider the image gradient energy function, and the edges correspond to high variations of the image intensity, and then to high values of the gradient. So, our goal is to find the most energetic parts of the image (in order to identify the edges), and we reformulate this problem as the minimization of the energy norm outside the edges

$$j(\rho) = J\left(\Omega_\rho, u_\rho\right) = \int_{\Omega_\rho} \left\|\nabla u_\rho\right\|^2. \tag{10}$$

Then, the cost function j has the following asymptotic expansion (see, e.g., [52] for more details):

$$j(\rho) - j(0) = \rho^2 G(x_0, n) + o(\rho^2), \tag{11}$$

with

$$\begin{aligned} G(x_0, n) = &-\pi c(\nabla u_0(x_0)\cdot n)(\nabla p_0(x_0)\cdot n) \\ &- \pi|\nabla u_0(x_0)\cdot n|^2, \end{aligned} \tag{12}$$

and where p_0 is the solution to the adjoint problem

$$\begin{aligned} -\operatorname{div}(c\nabla p_0) + p_0 &= -\partial_u J(\Omega, u_0) \quad \text{in } \Omega, \\ \partial_n p_0 &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{13}$$

The topological gradient could be written as

$$G(x, n) = (M(x)n)\cdot n, \tag{14}$$

where $M(x)$ is the 2×2 symmetric matrix defined by

$$\begin{aligned} M(x) = &-\pi c\frac{\nabla u_0(x)\nabla p_0(x)^T + \nabla p_0(x)\nabla u_0(x)^T}{2} \\ &- \pi\nabla u_0(x)\nabla u_0(x)^T. \end{aligned} \tag{15}$$

For a given x, $G(x, n)$ takes its minimal value when n is the eigenvector associated to the lowest eigenvalue $\lambda_{\min}$ of M. This value will be considered as the topological gradient associated to the optimal orientation of the crack $\sigma_\rho(n)$.

Then, we can define the identified edge set

$$\sigma = \{x \in \Omega; \lambda_{\min}(x) < \delta < 0\}, \tag{16}$$

where δ is a negative threshold.

We first illustrate this technique on a synthetic two dimensional image, in grey level, defined by a sigmoid function in x-coordinate (cumulative distribution function of a Gaussian). The image is represented in Figure 2(a). Then, the L^2 norm of its standard gradient $\|\nabla u(x)\|$ and its topological gradient $\lambda_{\min}(M(x))$ are represented in Figures 2(b) and 2(c), respectively.

One can see that the topological gradient is less sensitive to a smooth variation of the image intensity than the standard gradient. The support of the topological gradient is indeed much smaller. Thanks to the homogeneous Neumann condition on the crack, the solution of the perturbed problem is discontinuous along the crack, and the solution has a much smaller energy if one inserts a crack in the image near the middle of the x-axis.

We now apply this edge detection technique to the image represented in Figure 3(a). The opposite of the L^2 norm of its standard gradient is represented in Figure 3(b). Note that we represent its opposite in order to have comparable images with the topological gradient, which has negative values.

The topological gradient is represented on Figure 3(c). As it quantifies in a global way whether a pixel is part of an edge or not, it is much less sensitive to noise and small variations of the image than the standard gradient. For instance, the topological gradient takes much larger absolute values on the edges than outside, contrary to the standard gradient. Note also that the time required for the computation of the topological gradient is not much higher than for the standard gradient, thanks to the $\mathcal{O}(n \cdot \log n)$ complexity of the topological gradient algorithm.

However, for segmentation (or simply edge detection), the next step of topological gradient algorithms usually consists of thresholding the topological gradient in order to define the edge set. Such a threshold is represented in Figure 3(d). One can see that in order to obtain at least the main connected edge, the threshold coefficient has been set to a large value, leading to add many unwanted points to the edge set, but also to thick edges. And even in this case, the main contour is not totally continuous. This is why we need to hybridize this method with the fast marching algorithm (see Section 3.4) in order to obtain continuous edges for the segmentation and to remove the isolated unwanted pixels.

We will also see below that the fast marching algorithm needs a potential function highly related to the edges of the image, much more than the standard gradient of the image. Then, we will see that the topological gradient also improves the fast marching method within the segmentation framework, as the quality of the segmentation is directly related to the choice of the potential function.

2.3. Inpainting Algorithm by Topological Asymptotic Analysis. We also consider the inpainting application. We present here the topological gradient-based algorithm. Let $\omega \subset \Omega$ be the missing part of the image and γ its boundary. We still denote by v the input image (assumed to be known in $\Omega \setminus \omega$, and unknown in ω). The algorithm is based on the fact that two measurements are available on the boundary of the hidden part of the image: the value of the image (Dirichlet condition) and its normal derivative (Neumann condition).

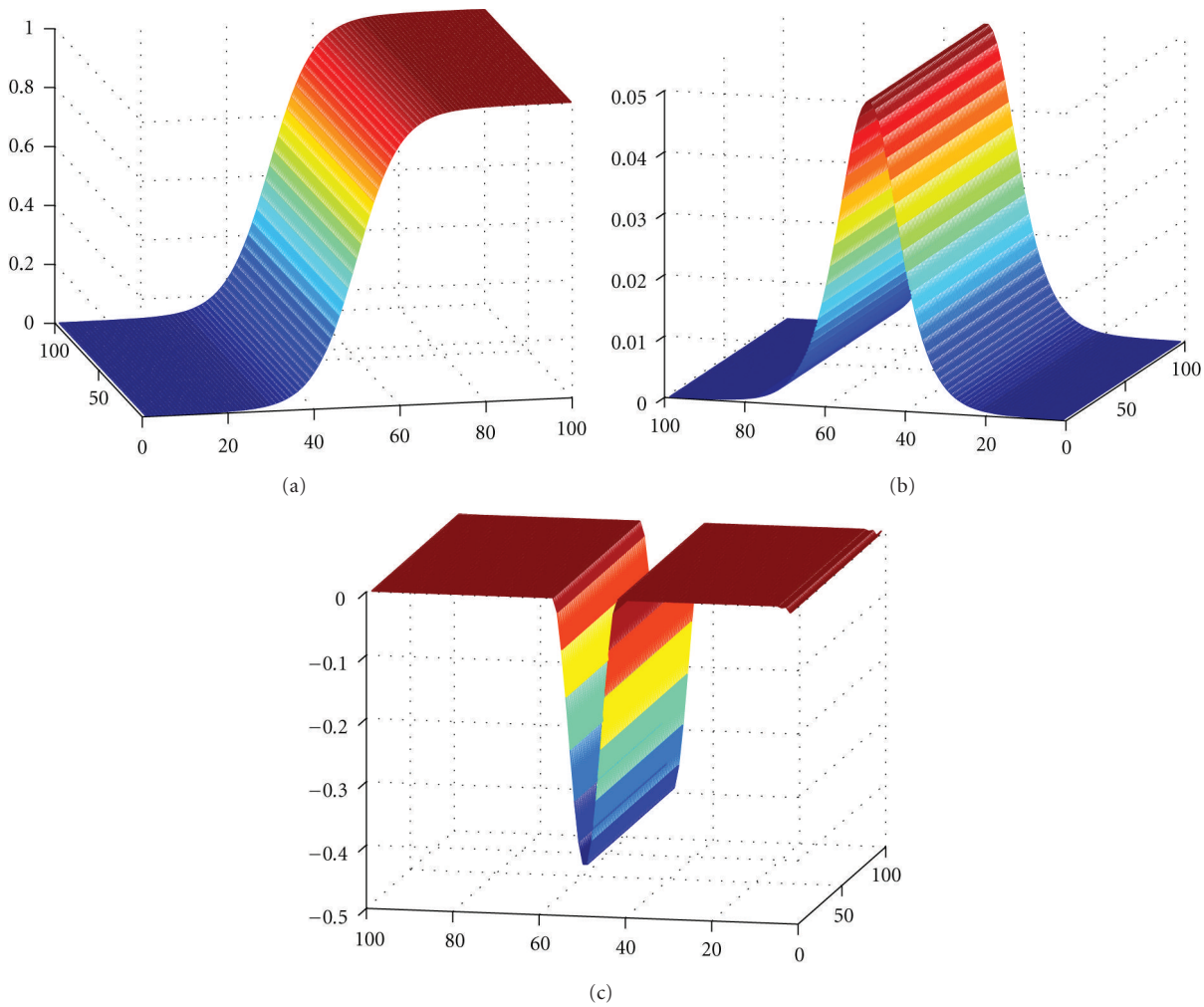


Figure 2: (a) Original image; (b) L^2 norm of the (standard) gradient of (a); (c) Topological gradient of (a).

From these two measurements, by considering the standard crack localization problem (see, e.g., [50]), it is possible to solve a Dirichlet problem and a Neumann problem for a given crack σ

$$\begin{aligned} \Delta u_D &= 0 \quad \text{in } \omega \setminus \sigma, \\ u_D &= v \quad \text{on } \gamma, \\ \partial_n u_D &= 0 \quad \text{on } \sigma, \\ u_D &= v \quad \text{in } \Omega \setminus \omega, \end{aligned} \tag{17}$$

where $u_D \in H^1(\Omega \setminus \sigma)$, and

$$\begin{aligned} \Delta u_N &= 0 \quad \text{in } \omega \setminus \sigma, \\ \partial_n u_N &= \partial_n v \quad \text{on } \gamma, \\ \partial_n u_N &= 0 \quad \text{on } \sigma, \\ u_N &= v \quad \text{in } \Omega \setminus \omega, \end{aligned} \tag{18}$$

where u_N is in $H^1(\Omega \setminus \sigma)$.

Then, in order to identify the missing edges, one has to minimize the following cost function:

$$J(\sigma) = \frac{1}{2} \| u_D - u_N \|^2_{L^2(\Omega)}. \tag{19}$$

For the actual cracks (hidden edges), the solutions u_D and u_N should be equal, as the actual solution satisfies both Neumann and Dirichlet conditions. By minimizing this cost function, one tries to find a solution that is consistent with both conditions on the boundary.

The topological gradient corresponding to this cost function is given by

$$\begin{aligned} G(x,n) = -\big[&(\nabla u_D(x) \cdot n)(\nabla p_D(x) \cdot n) \\ &+ (\nabla u_N(x) \cdot n)(\nabla p_N(x) \cdot n)\big], \end{aligned} \tag{20}$$

where p_N and p_D are the two corresponding adjoint states [25, 50]. As previously, the topological gradient can be rewritten as $G(x,n) = n^T M(x) n$, where $M(x)$ is a symmetric matrix, and G takes its minimal value when n is the eigenvector associated to the lowest eigenvalue of M.

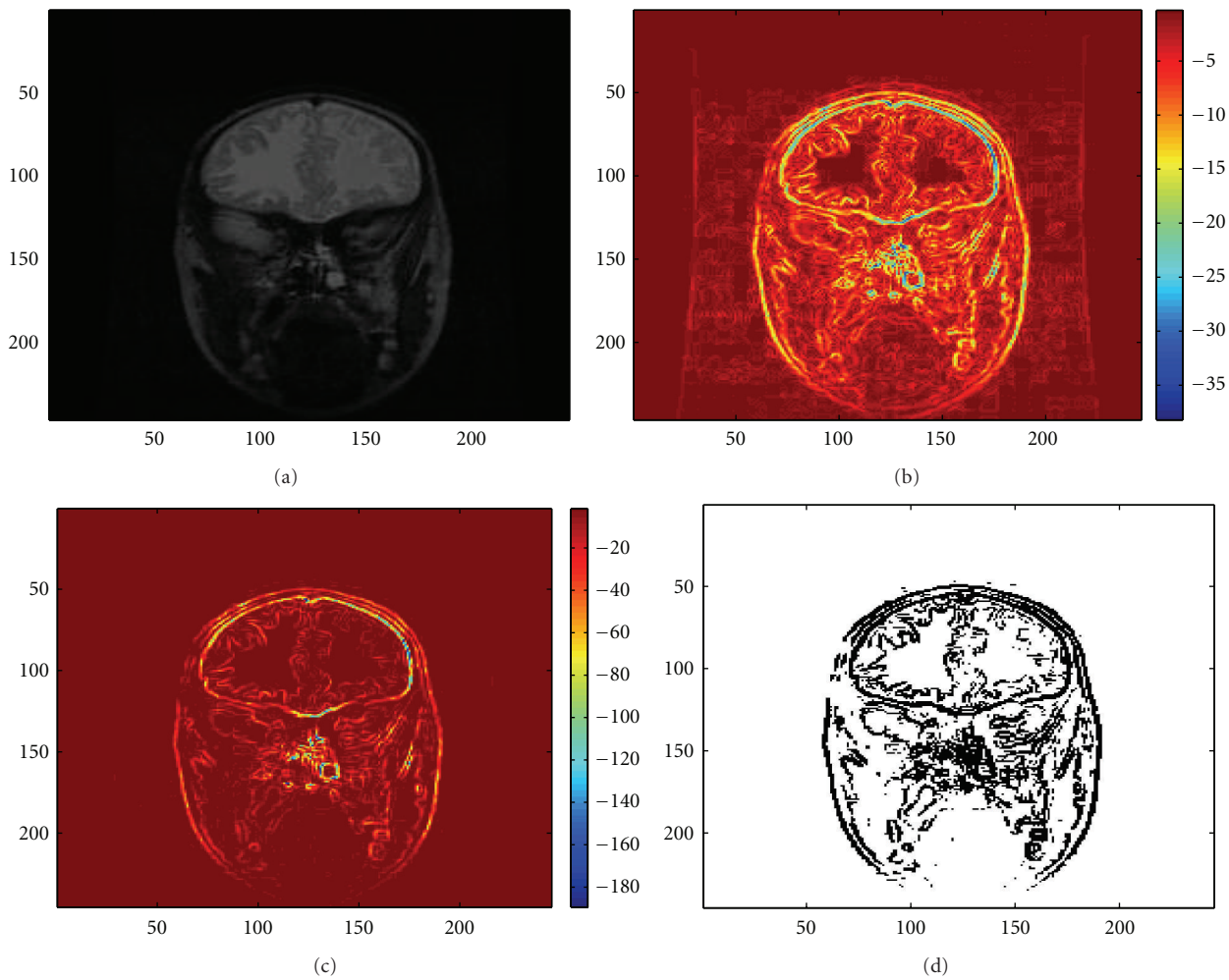


Figure 3: (a) Original image; (b) L^2 norm of the (standard) gradient of (a); (c) Topological gradient of (a); (d) Identified edges by thresholding the topological gradient.

The inpainting algorithm is then the following:

(i) calculation of u_D and u_N,

(ii) calculation of p_D and p_N,

(iii) computation of matrix $M(x)$ and its lowest eigenvalue $\lambda_{\min}$ at each point of the missing domain ω,

(iv) definition of the set of cracks: $\{x \in \omega; \lambda_{\min}(x) < \delta < 0\}$, where δ is a negative threshold,

(v) dalculation of u solution to the Neumann problem taking into account the cracks location.

This algorithm has a complexity of $\mathcal{O}(n \cdot \log(n))$, where n is the size of the image (i.e., number of pixels). We refer to [25] for more details about this algorithm.

We now illustrate this algorithm on two synthetic examples. We first want to restore a black square, partially hidden by a red square. The degraded image is represented in Figure 4(a).

If no edge is inserted in the hidden zone, then the resolution of a Poisson problem gives a blurred image, as the Laplace operator provides a smooth reconstruction between the black square and the white background, as shown in Figure 4(c). The restored image by the inpainting algorithm is represented in Figure 4(e). Using the edges identified by the topological gradient, the reconstruction by the Laplacian is much better, as there is now an insulating crack between the black and white zones.

The second synthetic example is the reconstruction of a black circle, partially hidden by a red square. The degraded image is represented in Figure 4(b), the restored image by the Laplacian without any inserted edge is shown in Figure 4(d), and the restored image the Laplacian using the edges identified by the topological gradient is represented in Figure 4(f). As one can see on these two synthetic examples, the curvature of the reconstructed edges is continuous in the neighborhood of the boundary of the occlusion. It is not common that an inpainted image has C^1 edges, and for instance, TV-based methods would connect the boundary points with a straight line.

We now explain why we also decided to hybridize the topological gradient and minimal paths methods on a more realistic case.

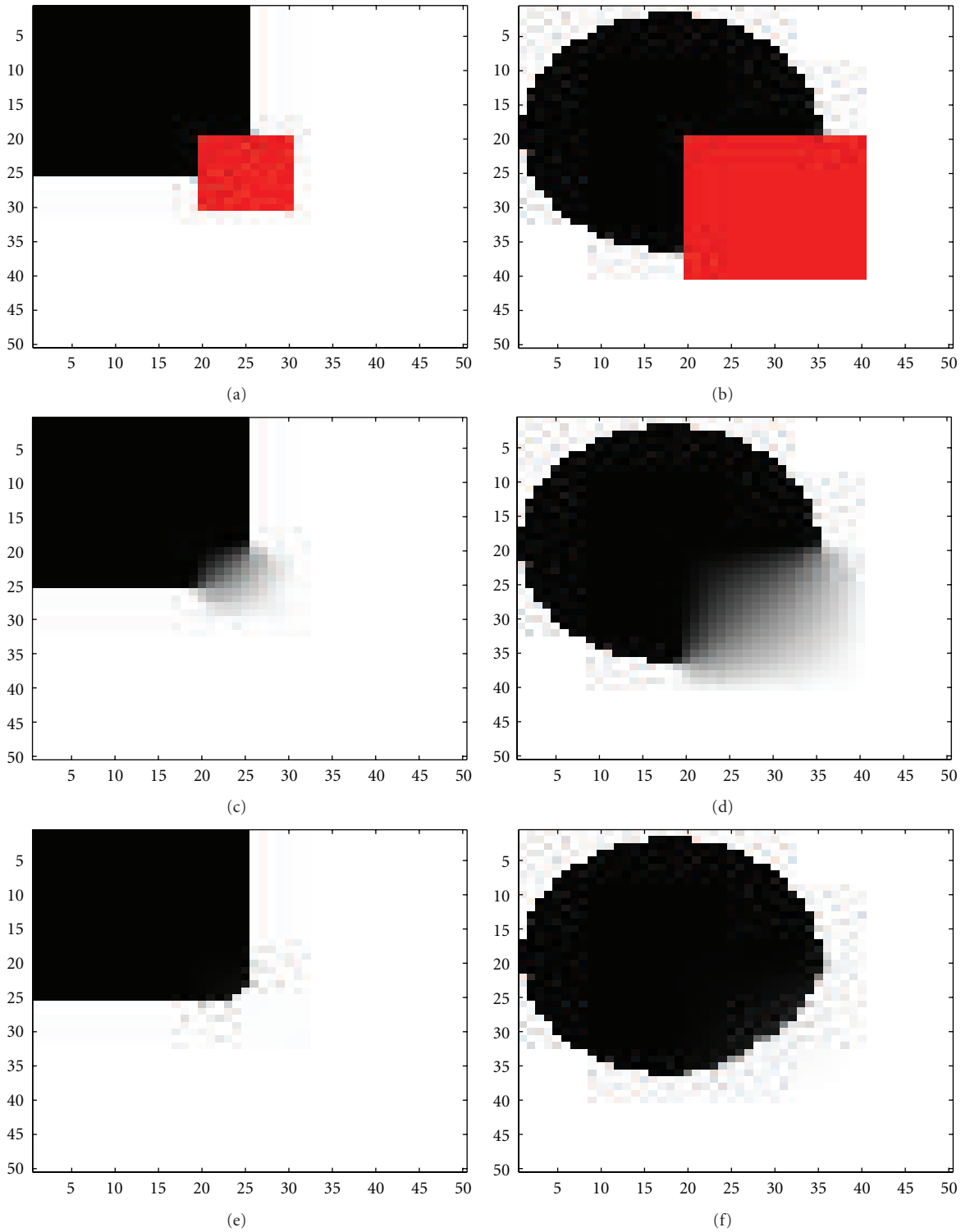


Figure 4: (a) Occluded image, defined by a black square on a white background, the occlusion being represented by a red square; (b) Occluded image, defined by a black circle on a white background, the occlusion being represented by a red square; (c) Inpainted image by diffusion (see (a) for the original degraded image), without any inserted edge in the occlusion; (d) Same as (c) in the circle case; (e) Inpainted image using the missing edges identified by the topological gradient, and then diffusion to fill in the image (see (a) for the original degraded image); (f) Same as (e) in the circle case.

Figures 5(a) and 5(b) show an example of image, in which we added a mask on a quite large part of the image ($\simeq$800 pixels). The goal of inpainting is to reconstruct as precisely as possible the original image from the occluded image. We also want the inpainted image to have sharp (unblurred) edges.

Figure 5(c) shows the corresponding topological gradient, provided by the inpainting algorithm. In this case, the topological gradient gives some information about the most probable location of the missing edges. In the inpainting algorithm presented in [25], the idea is then to threshold the topological gradient and to define the edge set of the occluded zone as being the set of points below the threshold. The main issue is that the identified missing edges must be connected in order to avoid blurry effects (due to the Laplacian) in the reconstruction. Then, the threshold is sometimes set manually in order to have connected contours. In our example, the identified edge set is represented by white points in Figure 5(d).

Figures 5(e) and 5(f) show the corresponding inpainted image. One can see that the reconstruction is not very good, particularly in the top part. This is mainly due to the fact that the missing edges identified by the topological gradient are either connected but thick with a lot of wrong identifications (if the threshold is too small) or discontinuous (otherwise).

The idea is then to apply the fast marching algorithm on the topological gradient obtained during the inpainting process in order to identify connected contours in the hidden part of the image.

3. A 2D Algorithm Based on the Minimal Paths and Fast Marching Methods

3.1. Minimal Paths. In this section, we describe the standard minimal path technique, adapted to our needs. We refer to [37, 38, 41] for more details about the minimal paths method.

In the following, let Ω be the considered image domain. We assume that Ω is a regular subset of $\mathbb{R}^2$. In order to compute some minimal paths, we need to define a potential function, measuring in some sense for any point of Ω the cost for a path to contain this point. As we want to identify paths in the topological gradient image, and considering that this potential function must be positive, we will define a potential function as follows:

$$P(x) = g(x) - \min_{y\in\Omega}\{g(y)\}, \qquad \forall x \in \Omega, \tag{21}$$

where g is the topological gradient, defined in all the domain Ω. We simply shift the topological gradient from its minimal value, in order to obtain a positive function P. We can see that the points where the topological gradient g reaches its minimal values are quite costless. This is a way to say that these points must be on the minimal paths. On the contrary, if the topological gradient takes high values, then the corresponding potential values lead to very expensive paths.

Once each point has a cost (defined by the potential function), we need to define the corresponding cost of a path. We denote by $C(s)$ a path, or curve, drawn in the image domain, where s represents the arc length. We can now define a functional, measuring the cost of such a path

$$J(C) = \int_C (P(C(s)) + \alpha)ds, \tag{22}$$

where α is a positive real coefficient that represents regularization. The first part of the cost function measures the cost itself of the path $C(s)$ simply by summing the value of the potential function on this path, and the second part is a regularization term that measures the length of this path. In our applications, α is usually very small, as the goal is to connect the most negative parts of the topological gradient, whatever the Euclidean distance is. Note also that we do not consider any regularization terms on the curvature of the contour, as the topological gradient already provides such regularity on the curvature, contrary to TV-based methods. Typically, $\alpha = 0$ would be a good choice, as we really want the minimal path along the topological gradient values, but as the minimum of P is 0 (at the minimum of the topological gradient), one has to set α to a very small value in order to avoid numerical instability (see (24)).

We now consider a key point $x_0 \in \Omega$ of the image, and x will represent any point of the image. The energy $J(C)$ of a given path C can be seen as a distance between the two endings of C, weighted by the potential function (and the regularization). The goal is to find the minimal energy integrated along the path C. We can now define the weighted distance between key point x_0 and point x by

$$D(x;x_0) = \inf_{C\in A(x,x_0)} J(C) = \inf_{C\in A(x,x_0)} \int_C (P(C(s)) + \alpha)ds, \tag{23}$$

where $A(x, x_0)$ is the set of all paths going from point x_0 to point x in the image. The idea is that finding the minimal path between points x and x_0 is now equivalent to computing the weighted distance function between these two points. If x and x_0 are on the same contour of the image, then the minimal path between these two points is obviously a continuous contour of the image, connecting these points. The minimal path has indeed the lowest cost, that is, the points on this path have low topological gradient values. The goal is now to compute the distance function given by (23).

3.2. Fast Marching. An efficient way to compute this distance function is to solve a front propagation equation:

$$\frac{\partial \mathcal{F}(s,t)}{\partial t} = \frac{1}{P(\mathcal{F}(s,t)) + \alpha}\, \mathbf{n}_{\mathcal{F}}(s,t), \tag{24}$$

where $\mathbf{n}_{\mathcal{F}}(s,t)$ is the outer normal unit vector to the front $\mathcal{F}$. We initialize the propagation with $\mathcal{F}(s,0)$, an infinitely small circle centered at key point x_0. This front evolves then with a propagation speed inversely proportional to the potential function. If for example a point in the outer part of the front has a large potential (i.e., a large cost), then the propagation speed will be nearly equal to zero, and the front will not expand much at this point. On the other hand, if the potential is small (i.e., this point is nearly costless), then

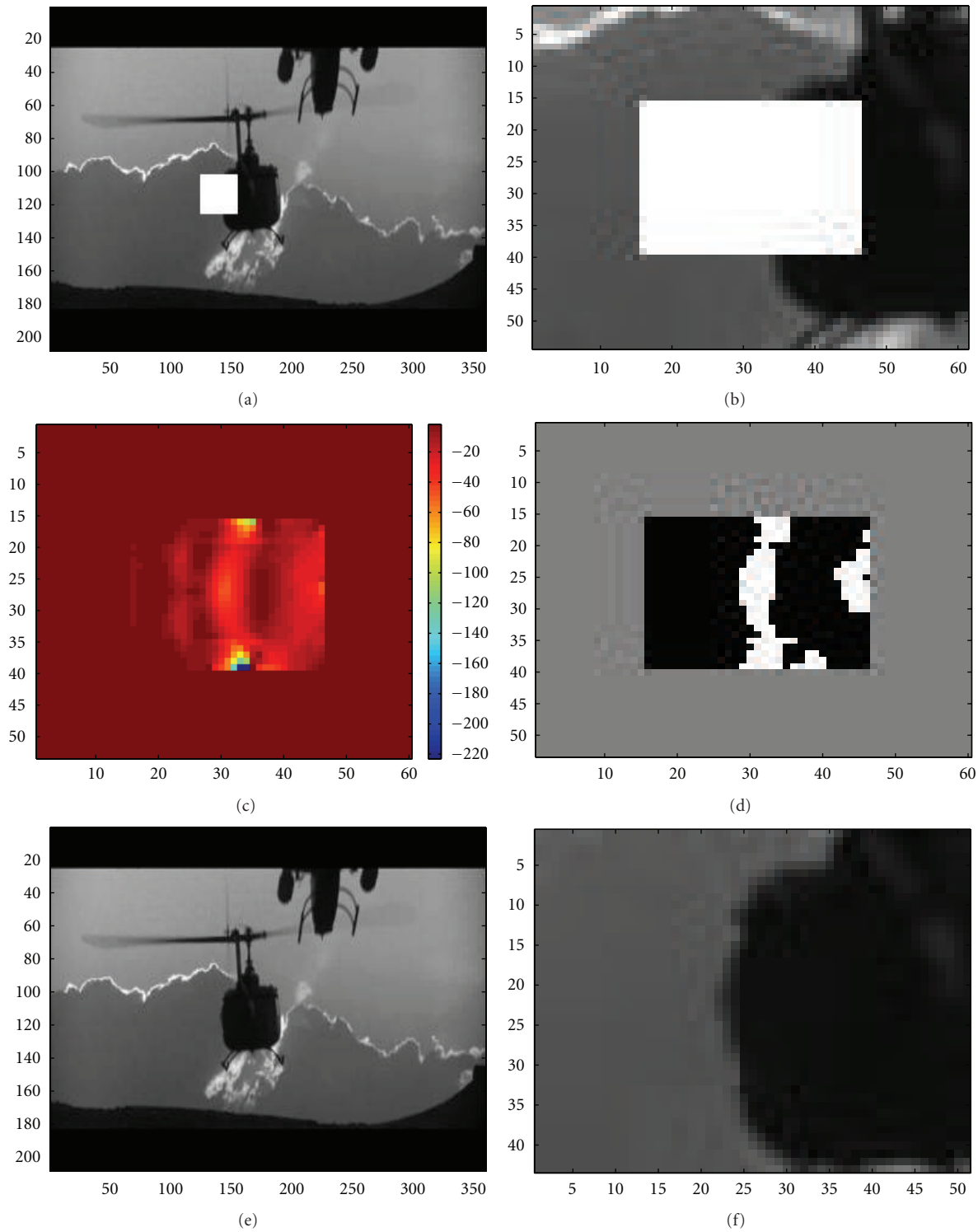


Figure 5: (a) Occluded image (by a white rectangle); (b) Zoom of the occluded zone (see (a)); (c) Topological gradient of (b); (d) Identified edges in the occluded zone by thresholding the topological gradient; (e) Inpainted image using the topological gradient; (f) Zoom of the occluded zone (see (e)).

the propagation speed is large, and the front will quickly propagate in this direction.

The distance $D(x; x_0)$ introduced in (23), between key point x_0 and point x, is then simply the instant t at which the front, initialized at key point x_0, reaches point x. The algorithm to compute the distance function is called the fast marching technique and is justified by the fact that the distance satisfies the following Eikonal equation:

$$\|\nabla_x D(x; x_0)\| = P(x) + \alpha, \tag{25}$$

with the initialization $D(x_0) = 0$. We refer to [37, 38, 44, 53, 54] for more details about the fast marching technique and the justification of (25). If n is the size of the image, the complexity of this fast marching method is bounded by $\mathcal{O}(n \cdot \log(n))$, which is also the complexity of the topological gradient algorithm.

3.3. Multiple Minimal Paths. The main issue is now to extend this minimal path technique to more than one keypoint in order to connect several points. This is exactly what we need in order to connect the identified edges by the topological gradient, as we have many identified keypoints (e.g., all negative local minima of the topological gradient) that we want to connect. As explained in [41], the first point of a multiple minimal path algorithm is to reduce the set of keypoints for computational reasons. Moreover, the selected keypoints should not be too close to each other. One usually chooses a total number N of keypoints and the first (or main) one. Then, the $N - 1$ other keypoints can be chosen for example as described in [41].

The next step consists of connecting these N points. One has to compute the distance function from each of these key points, and the common minimal paths algorithms provide then the Voronoï diagram of the distance and the corresponding saddle points (minimal distance along the edges of the diagram and maximal distance from the keypoints). The Voronoï diagram defines a partition of the image in as many subsets as the number of keypoints. Each subset is defined by the set of points that are closer to the corresponding keypoint than to all others. The saddle points minimize the distance function on the edges of the diagram: minimal distance on the edge and maximal distance to the keypoints [38]. It is useful to compute these saddle points to save computation time, since it reduces the domain of the image where the fast marching computes or updates the weighted distance map.

Finally, the idea is to consider the saddle points as initial conditions for minimizing the distance function. For each saddle point as an initial point, a minimization is performed towards each of the two corresponding keypoints (recall that the saddle points are located at the interface between two subsets of the Voronoï diagram). Each minimization produces a path between the saddle point (initial condition) and a keypoint (local minimum of the distance function). This step is usually called back propagation, as it consists of a gradient descent from the saddle point, back to the linked keypoints. The back-propagation step is straightforward, as there is no local minimum of the distance function, except the keypoints. The union of all these paths gives a continuous path, connecting the keypoints together.

The interesting part of the approach introduced in [41] is that each keypoint should not be connected to all the others, but only to at most two others, as we are looking for a set of closed connected paths. Thus, the keypoints have to be ordered in a way such that they are only connected to the other keypoints that are closest to them in the energy sense [41]. For this reason, we sort all the saddle points from smaller to larger distance, and we first try to connect the pairs of keypoints corresponding to the saddle points of smallest distance. These keypoints are indeed more likely to be connected than distant keypoints, corresponding to saddle points of large potential. Once the close keypoints are connected, we repeat the process with the new closest pairs of keypoints, provided each point remains connected to at most two other ones. At the end of the process, all the keypoints are connected to at most two other keypoints, and the union of all minimal paths between the keypoints represents one (or several) continuous contour of the image. An interesting feature of this method is that the key points are by construction widely distributed around.

If all the selected keypoints are on the same contour of the image, we are almost sure that at the end, they will all be connected together, and we will retrieve the corresponding contour, as the potential function (related to the topological gradient) is very low on this contour. If, on the contrary, one keypoint is not part of the contour, the large values of the topological gradient, and hence of the potential function, will isolate this keypoint from the other ones, and it will not disturb the contour completion process.

3.4. Algorithm. The hybrid algorithm we propose is then the following.

Fast Marching Algorithm Applied to the Topological Gradient

(i) Compute the topological gradient of the image.

(ii) Set N the number of keypoints and choose the N keypoints: the main one will be for example the global minimum of the topological gradient, the other ones being the most negative local minima of the topological gradient.

(iii) Compute the distance function (23) with all these keypoints, and the corresponding Voronoï diagram.

(iv) Compute the set of saddle points: on each edge of the Voronoï diagram, determine the point of minimal distance.

(v) Sort all these points of minimal distance, from smaller to larger distance.

(vi) For each of these saddle points, from smaller to larger distance, check if it will not be used to connect two keypoints, one of which is already connected to two other keypoints.

(vii) If this is not the case, perform the back propagation from this point: use this saddle point as an initialization for a descent type algorithm in order to connect the two corresponding keypoints.

It is straightforward to see that this algorithm converges and that at convergence, all the keypoints are connected to at most two other keypoints. This provides one or several continuous contours containing the keypoints. As the first keypoint is usually the global minimum of the topological gradient, it is on one of the main edges of the image. Consequently, using this algorithm, we can identify this edge. Then, it is possible to restart the algorithm, using other keypoints that are not on this identified edge, by initializing, for instance, the first keypoint as the minimum of the topological gradient outside the neighborhood of this edge.

Note that for inpainting applications, the number of keypoints can be set automatically, as the topological gradient takes its minimal values on the edges located on the boundary of the hidden zone, and all these minima (close to the global minimal value of the topological gradient) can be chosen as keypoints.

4. Numerical Experiments

4.1. Numerical Results for 2D Segmentation. We consider again the grey level image represented in Figure 3(a) for the segmentation application, and we now present the results corresponding to the hybrid method.

Using an automatic thresholding for identifying the most negative values of the topological gradient, Figure 6(a) shows the set of points (or admissible keypoints, in blue), in which we will choose the keypoints for the minimal path algorithm. The first keypoint is set to the minimum of the topological gradient. Then, we have set the number of keypoints to $N = 3$. From the first keypoint, we start the minimal path algorithm, and we choose the second keypoint as being the point (in the admissible set) maximizing the distance to the first keypoint. Then, we start again the minimal path algorithm from these two points, and we set the third keypoint in a similar way. These three keypoints are represented by black points in Figure 6(a). Note that the keypoints can also be (manually) provided by the user, for instance, with the aim of identifying a specific edge of the image.

From these keypoints, we run the minimal path algorithm in order to compute the distance map. Figure 6(b) shows this distance function. One can clearly see that the distance does not correspond to the Euclidean metric in the plane, as the distance remains very small on the common edge of the 3 keypoints, whereas it takes much larger values outside.

The corresponding Voronoï diagram is represented in Figure 6(c). The three keypoints are still represented by black points. Each color represents the subset Ω_i of points that are closer to keypoint i than to the others. For instance, all the points in the green zone are closer to the right keypoint than to any of the two others. This diagram is automatically provided during the distance computation by the fast marching algorithm.

For any $i \neq j$, we consider the interface $\Gamma_{ij} = \Omega_i \cap \Omega_j$ between two subsets of the Voronoï diagram. Γ_{ij} represents then the set of points equidistant from keypoints i and j. A saddle point minimizes the distance function on Γ_{ij}: same distance to keypoints i and j, minimal distance on Γ_{ij}. These saddle points are represented by blue points on Figure 6(c). These saddle points can be found during the fast marching propagation as the first meeting points of the fronts starting from each of the keypoints.

From these saddle points, the idea is finally to perform a descent-type algorithm in order to minimize the distance function from the saddle points to the keypoints. We consider a saddle point on an edge Γ_{ij} as an initial condition for two minimizations of the distance function, one towards each of the corresponding keypoints (i and j). Each of these two minimizations provides a continuous path from the saddle point to one of the two keypoints. The union of these two paths connects the two keypoints. This process is done for all pairs of keypoints.

The final set of paths is represented in green on the distance function in Figure 6(d). The three keypoints are also represented (in white). These paths correspond to the contour of the original image that contains the 3 keypoints.

The minimal path is also represented on the original image in Figure 6(e). It also confirms that the identified path perfectly matches the edge we were looking at.

By applying again this algorithm, with other keypoints (selected outside the first identified contour), it is possible to detect other contours of the image. Figure 6(f) shows, for instance, the first main contour in green and a second one in red. Contrary to the first one, we can see that this contour is not perfectly detected, as the algorithm missed some parts of the contour in the bottom left and top parts of the red zone. One should probably consider more keypoints, and maybe a different regularization coefficient, in order to avoid this phenomenon. But for the application of the topological gradient to image segmentation, the main issue was the discontinuity of the identified contours (see, e.g., [26]). With this approach, we ensure the continuity of the contours, and hence, assuming the edges are well identified, we can obtain a perfectly segmented image.

Finally, we illustrate the fact that the topological gradient provides better information about the edges of the image than the standard gradient, as previously observed (see Figures 3(b) and 3(c)). We have manually selected 3 keypoints on an edge of the image. These keypoints are represented in blue on Figure 7(a). From these keypoints, we have run the fast marching algorithm (see Section 3.4) applied to both the standard gradient and the topological gradient (hybrid scheme). The identified paths are represented in Figures 7(b) and 7(c), respectively.

The topological gradient clearly provides the best identification of the edge. This can easily be explained by the bad shape of the standard gradient in this region (see Figure 3(b)). On the contrary, the topological gradient is less sensitive to small local variations, and it is more likely to define a potential function than the standard gradient.

4.2. Numerical Results for 2D Inpainting. We now consider another application of this hybrid scheme to image inpainting. We recall that the idea of the topological gradient algorithm is to identify the missing edges in the occluded part of the image, and then to reconstruct the image from the solution of a Poisson problem with Neumann boundary conditions [25]. In this application also, it is crucial to have connected contours; otherwise, the reconstruction with the Laplacian will not be satisfactory.

We first present a comparison between the standard topological gradient approach, a TV-based inpainting method, and the new hybrid scheme. The original image is a black rectangle, and we consider various perturbations of this image. Figure 8(a) shows a first perturbation of the image, in which the missing region is represented by the red rectangle. The length of the hidden zone is 20 pixels. As previously shown, the missing zone is quite large, and as the identified

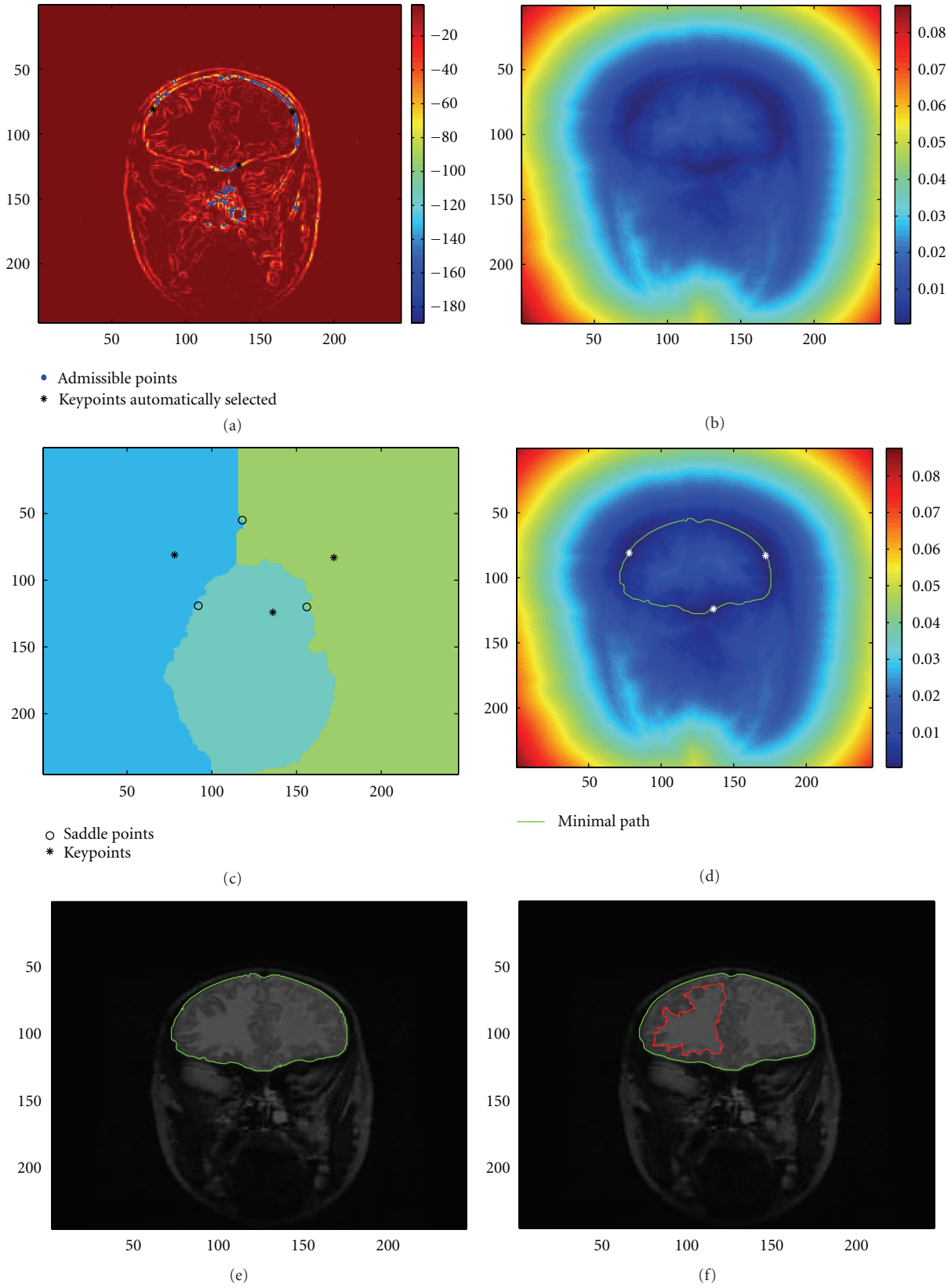


Figure 6: (a) Admissible set of points (i.e., most negative values of the topological gradient) in blue, and 3 keypoints automatically selected in black; (b) Distance function computed from these 3 keypoints with the fast marching algorithm; (c) Corresponding Voronoï diagram, with the 3 keypoints and saddle points; (d) Identified minimal path between the keypoints represented on the distance function; (e) Minimal path between the keypoints represented on the original image; (f) Another identified continuous contour from other keypoints.

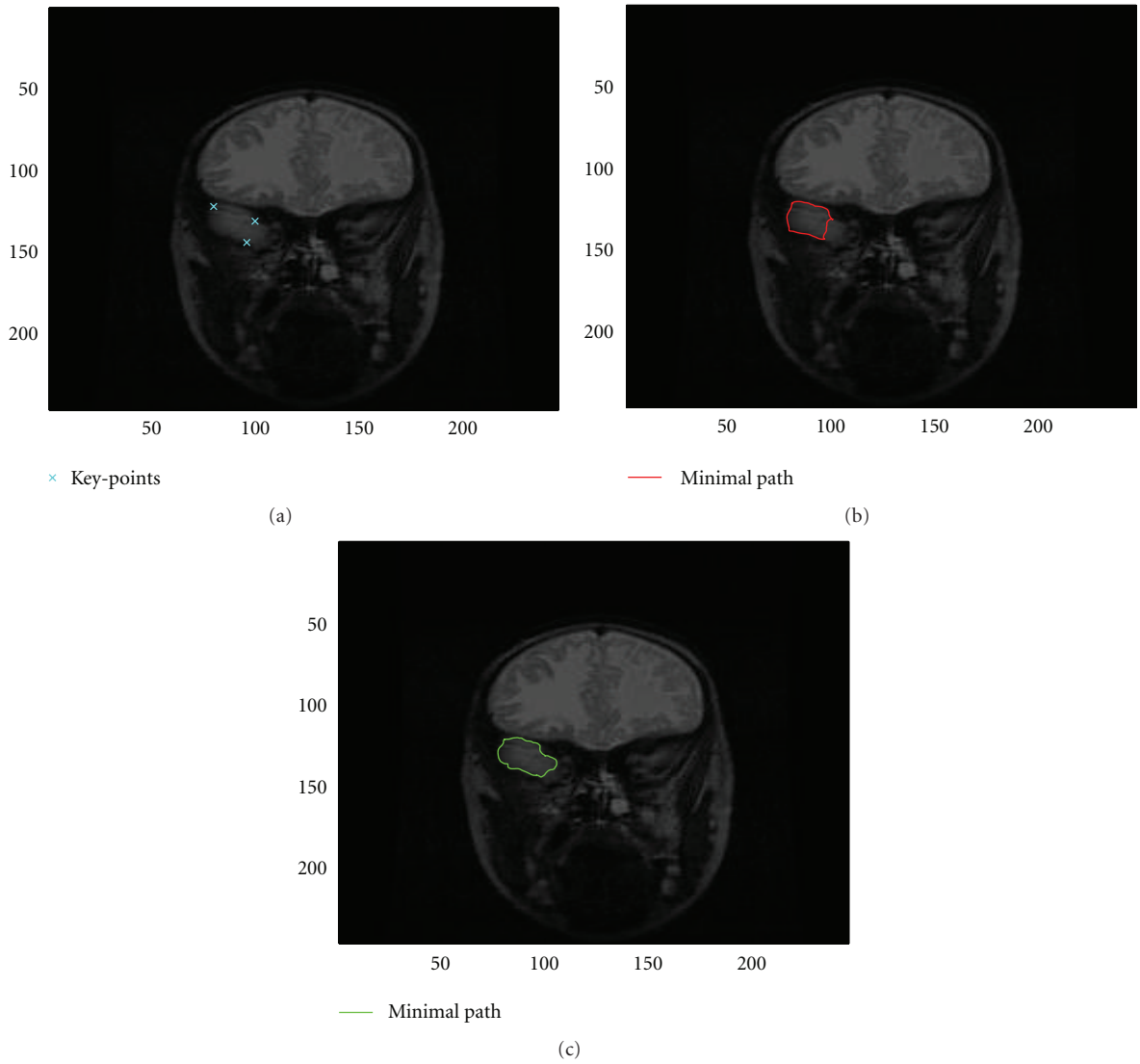


Figure 7: (a) Three selected keypoints on the original image; (b) Contours identified by the fast marching algorithm applied to the standard gradient with the three selected keypoints (see (a)); (c) Contours identified by the fast marching algorithm applied to the topological gradient with the three selected keypoints (see (a)).

edges have to be connected in order to avoid blurry effects in the reconstruction, the threshold is set manually to a quite small negative value. And then, the identified edges are then quite thick with a lot of wrong identifications. The reconstructed image by the topological gradient is shown in Figure 8(b). The reconstruction is not very good, as many wrong edges are considered in order to connect the contours. Figure 8(c) shows the identified minimal path between keypoints (that have been automatically selected, as being the main edges on the boundary of the missing zone) in green, represented on a zoom of the perturbed image. Figure 8(d) shows the corresponding inpainted image by the hybrid scheme: the image is reconstructed using the topological gradient method, with the edges identified by the minimal path technique. In this case, the reconstruction is perfectly done, and the inpainted image is identical to the original image. A TV-based inpainting method gives the same result (see Figure 8(e)), as the missing zone is not too wide (20 pixels, which is also the size of the black rectangle).

Figure 9 is similar to Figure 8 in the case of a larger perturbation. The missing zone corresponds now to 40 pixels, twice the size of the black rectangle. In this case, the topological gradient is much less negative near the middle of the hidden zone, and the threshold has to be increased to a smaller negative value in order to have closed contours. The corresponding inpainted image is not good at all. But the minimal path technique still identifies correct edges, and the inpainted image by the hybrid scheme is almost perfect, whereas a TV-based inpainting method does not connect anymore the two regions of the rectangle.

Figure 10 is similar to Figures 8 and 9, in the case of a larger perturbation. The missing zone now corresponds to 80 pixels, which is much larger than the size of the black rectangle. In this case, the topological gradient still gives

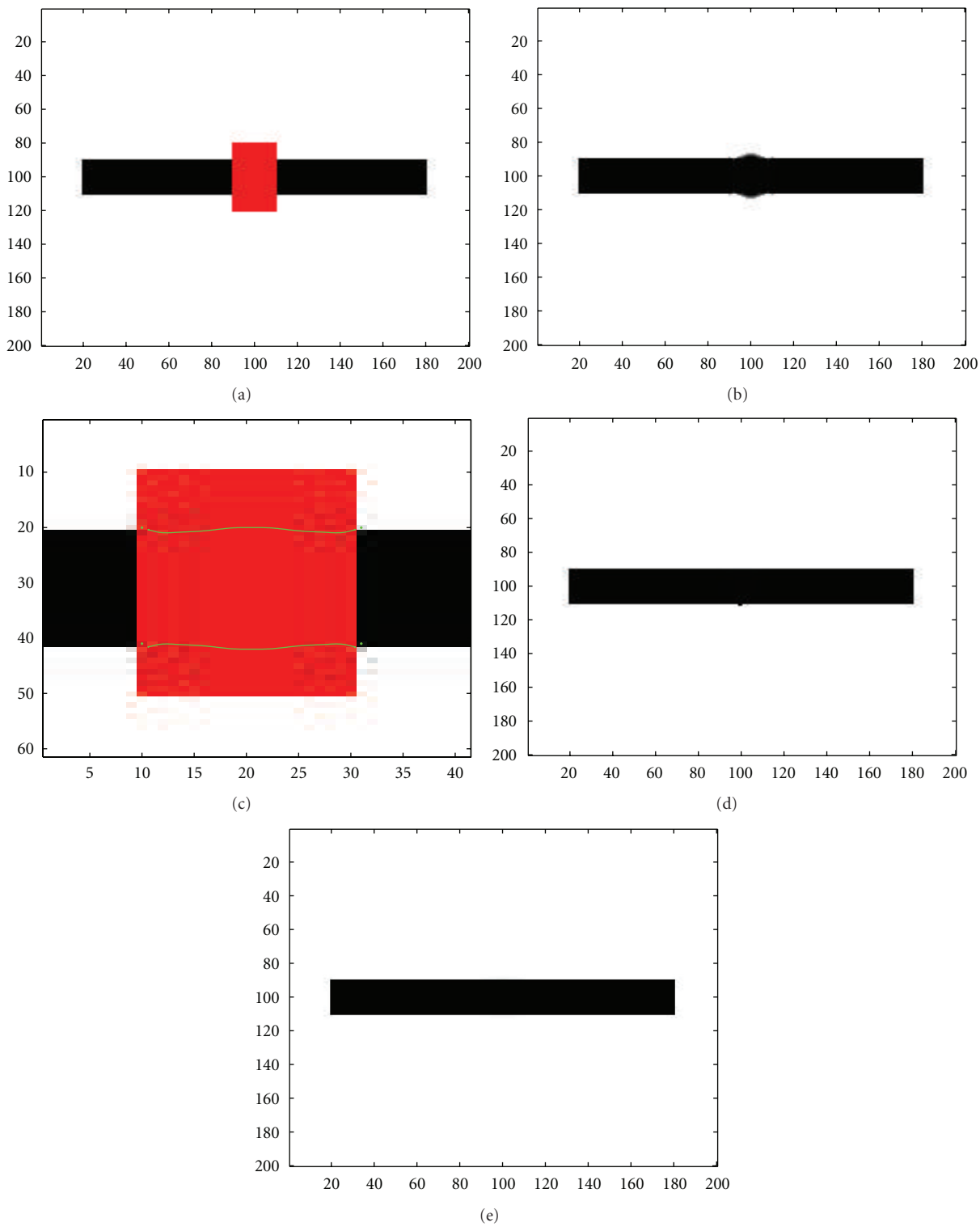


Figure 8: (a) Occluded image (black rectangle) by a red rectangle; (b) Inpainted image using the standard topological gradient; (c) Minimal path between the keypoints represented on the topological gradient; (d) Inpainted image using the hybrid scheme (fast marching algorithm for closing the contours identified by the topological gradient); (e) Inpainted image using a TV-based method.

unsatisfactory results, due to badly connected edges. Even if the topological gradient has strongly negative values along the missing edges close to the boundary of the perturbation, the missing zone is too wide, and the minimal path technique now connects wrong keypoints, and the inpainted image by the hybrid scheme is no more connected. As before, the TV-based method does not connect the two parts of the rectangle.

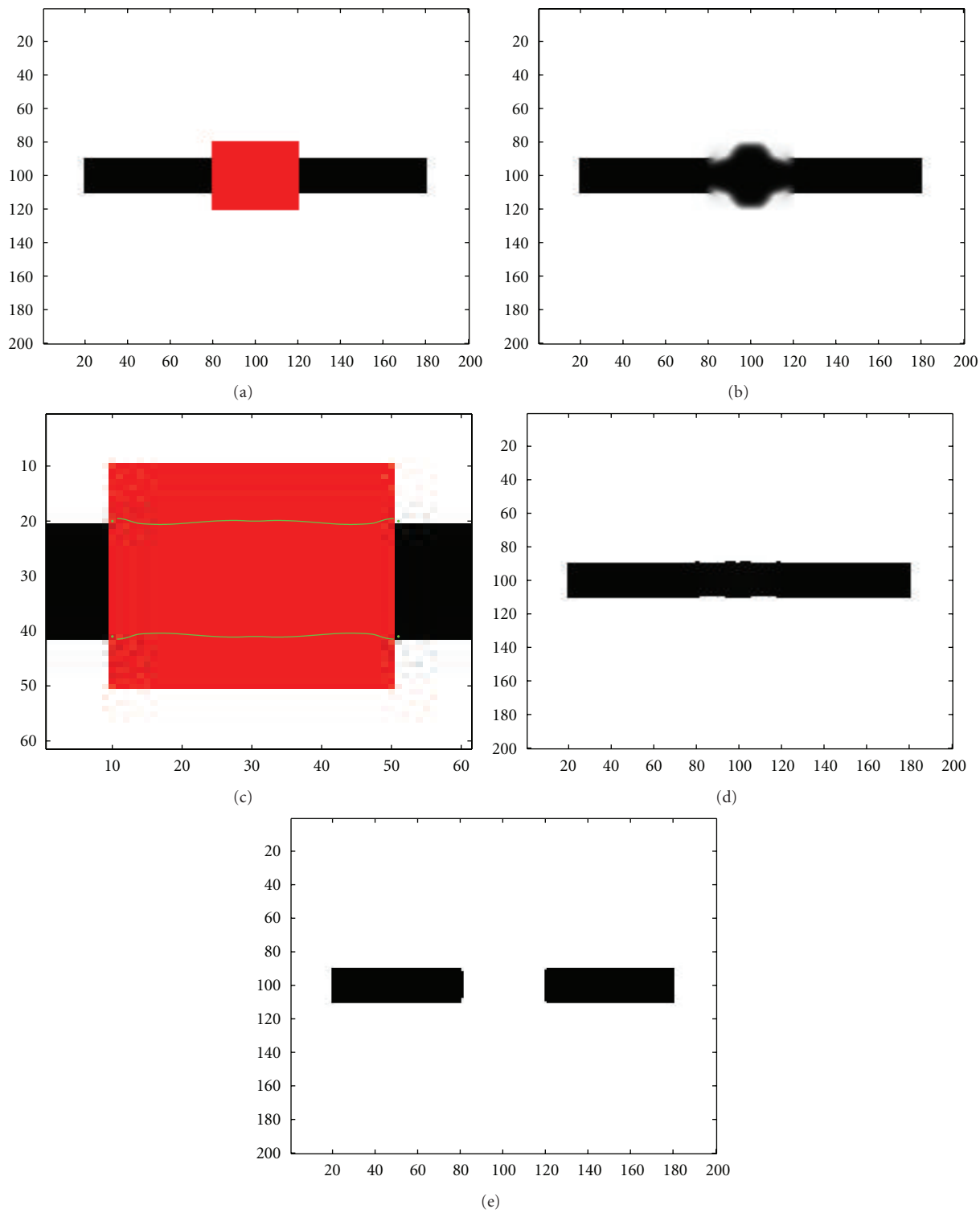


Figure 9: (a) Occluded image (black rectangle) by a red rectangle; (b) Inpainted image using the standard topological gradient; (c) Minimal path between the keypoints represented on the topological gradient; (d) Inpainted image using the hybrid scheme (fast marching algorithm for closing the contours identified by the topological gradient); (e) Inpainted image using a TV-based method.

We now consider again the occluded image given in Figure 5(a).

After thresholding the topological gradient, several points (identified by blue circles) have been identified and define the admissible set of keypoints represented in Figure 11(a). We choose then the most negative point of the topological gradient as the first keypoint and then the further admissible point as the second one. The keypoints

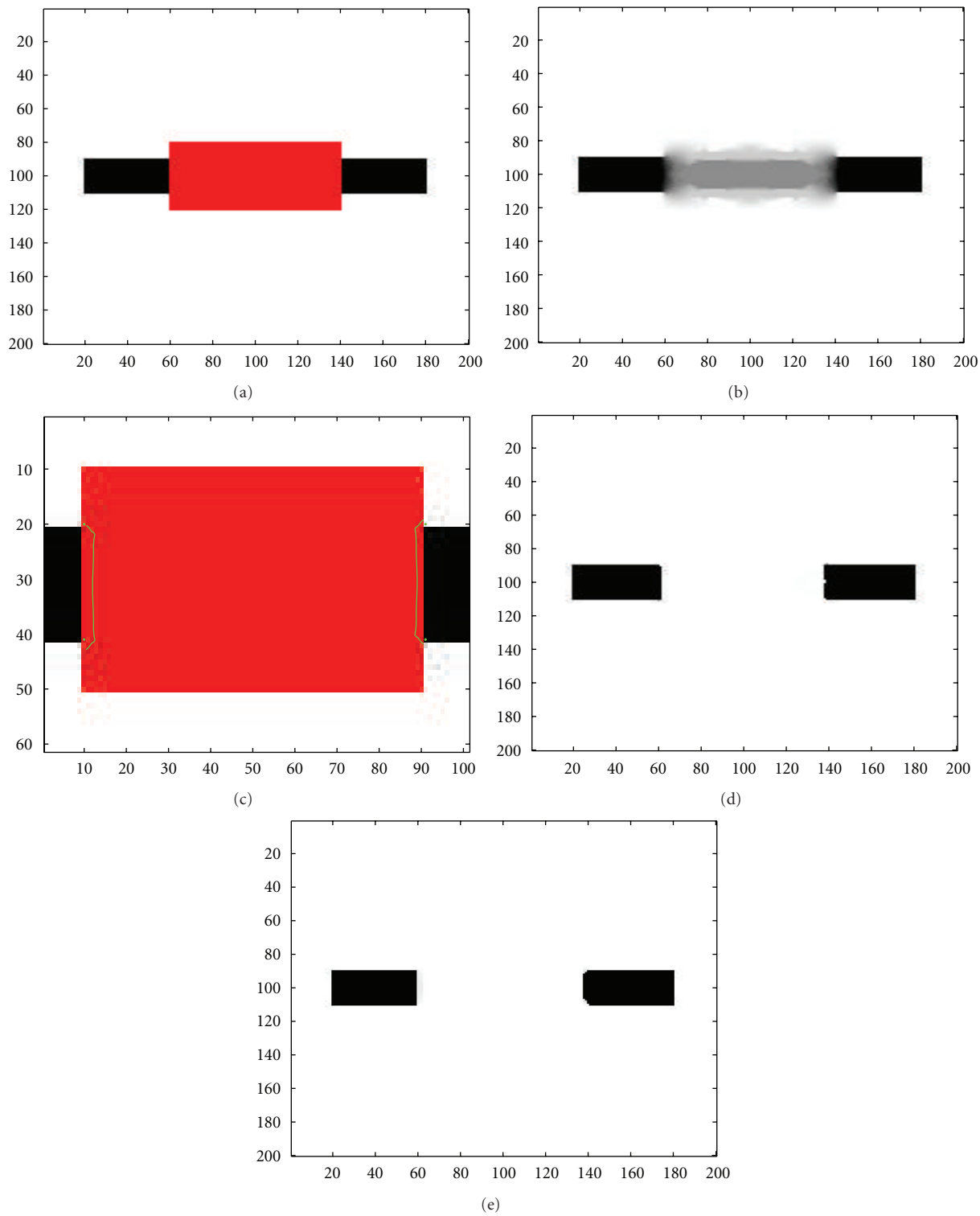


Figure 10: (a) Occluded image (black rectangle) by a red rectangle; (b) Inpainted image using the standard topological gradient; (c) Minimal path between the keypoints represented on the topological gradient; (d) Inpainted image using the hybrid scheme (fast marching algorithm for closing the contours identified by the topological gradient); (e) Inpainted image using a TV-based method.

are represented by a large black point on the same image. They are located on the edge of the domain, as the inpainting topological gradient always takes its minimal values there.

Then, the minimal path algorithm is run, and it provides a path between the keypoints, represented in green in Figure 11(b). We can see that the path follows very well the valley line of the topological gradient, from one side to the other. By choosing 3 keypoints instead of 2, there will be another keypoint on the bottom edge, near the first one, and it will simply add a small contour located all along on the

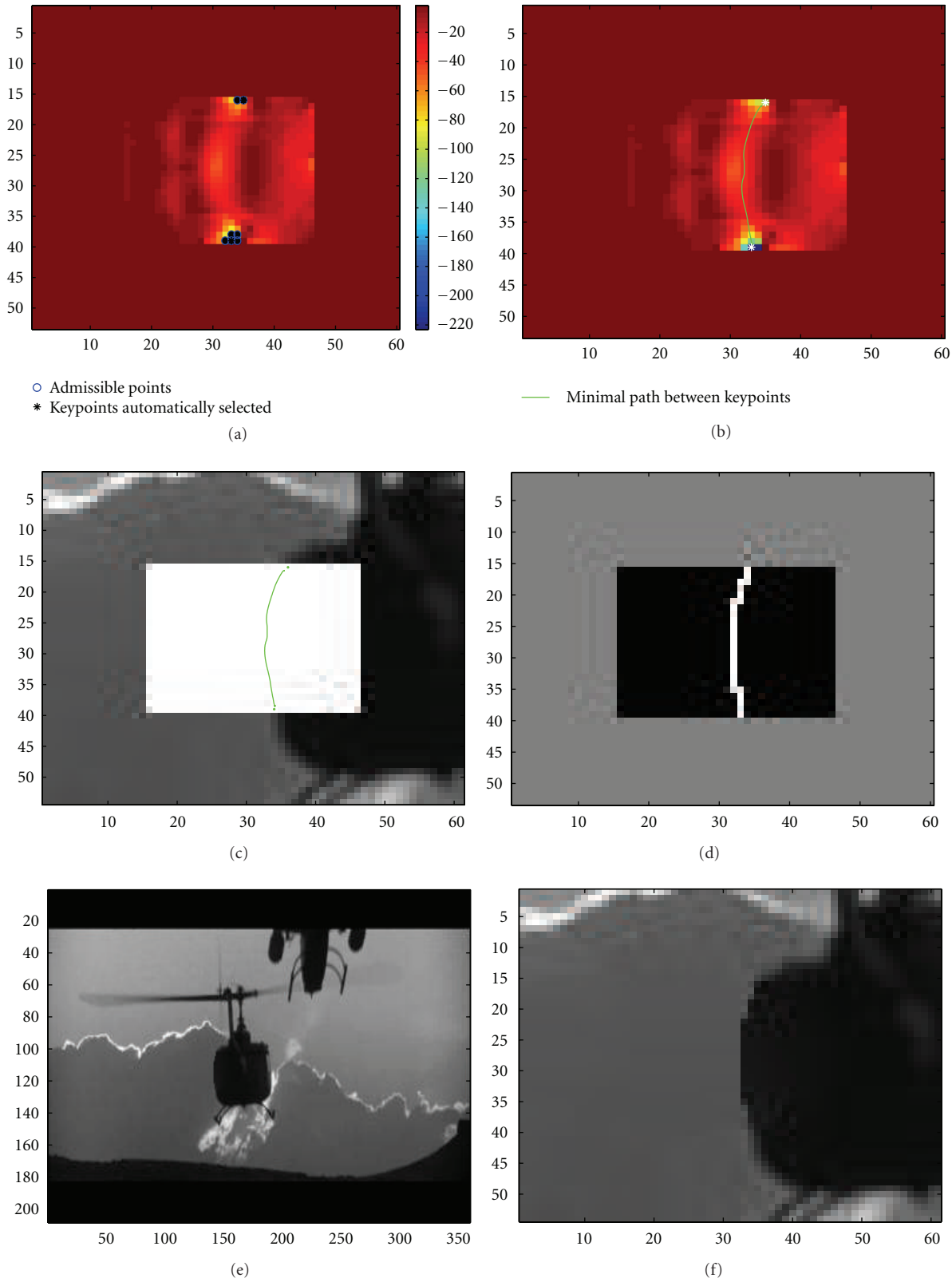


Figure 11: (a) Admissible set of keypoints and selected keypoints on the topological gradient; (b) Minimal path between the keypoints represented on the topological gradient; (c) Minimal path between the keypoints represented on the occluded image; (d) Corresponding identified missing edge in white; (e) Inpainted image using the fast marching algorithm for closing the contours identified by the topological gradient in the hidden part of the image; (f) Zoom of (e).

edge of the domain, and consequently, there is absolutely no impact on the reconstruction of the hidden part of the image.

Figure 11(c) shows the same identified path represented on the occluded image. This allows one to see that the path clearly gives a good approximation of the missing edge and also that the topological gradient is very powerful for this identification problem. The corresponding identified edge set is represented in Figure 11(d). This image should be compared with the thresholded edge set of Figure 5(d). From these two images, we can conclude that the minimal path algorithm is an excellent tool for extracting the valley lines of the topological gradient.

Finally, using this minimal path as the set of missing edges in the occluded zone, the inpainting topological gradient algorithm produces a much better reconstructed image, shown in Figures 11(e) and 11(f). The quality of the image is very good, as the missing edges used for the reconstruction are connected, and the Laplace operator will not produce any blurring effect due to a discontinuous contour. Note that there are some small discontinuities on the top left boundary due to the fact that we used the Neumann solution of the perturbed problem. The construction of a Dirichlet solution would be better, but it is also much more difficult to solve the Dirichlet problem in this case, as it is ill posed. This example confirms that the quality of all topological gradient applications in image processing can be improved by replacing a simple thresholding technique by a minimal path algorithm.

As already shown in [25], the topological gradient extrapolates the edges and their curvature in the missing part of the image (see also Figure 11(d), in which the identified edge is not a straight line), contrary to total variation-based methods. Thus, provided the identified missing edges are connected (this point is now ensured by the application of the fast marching algorithm to the topological gradient), the inpainted image has edges with continuous curvature, which is not the case with many other inpainting schemes.

5. Conclusions and Perspectives

We have introduced a hybrid scheme, based on one side on the topological gradient for edge detection, and on the other side on the fast marching and minimal paths methods for contour completion. These approaches allow us to extract connected contours in 2D images and to solve the main issue of all topological gradient-based algorithms for image processing problems (discontinuity of the edges). Moreover, the minimal path algorithm does not degrade the complexity of the topological asymptotic analysis.

We have considered two specific applications in image processing: segmentation and inpainting. In the first one (segmentation), we showed that the topological gradient is more efficient than the standard gradient for edge detection and the hybrid scheme provides better results than the fast marching method applied to the standard gradient of the image. In the second application (inpainting), we showed that the hybrid scheme particularly improves the quality of the inpainted image, as the contour completion ensures a nonblurred inpainted image and as it also helps removing the manual thresholding of the topological gradient.

The hybrid scheme is very efficient and quite automatic, as there is no more thresholding process. The topological gradient algorithm has been shown in previous inpainting articles to propagate the main edges inside the hidden zone, with some continuity of their curvature, and the use of a minimal path technique helps detect the valley lines of the topological gradient. The main drawback of the hybrid scheme is the same as for the standard topological gradient algorithm: the image is filled in with the Laplacian, and this part has to be improved in order to also recover texture information. Some preliminary results show that it is possible with the same kind of approach, thanks to higher order operators.

An interesting and natural perspective is to apply this hybrid scheme to 3D images and movies. The topological gradient can very easily be extended to 3D images. The minimal path technique has also been adapted to the identification of tubular structures in 3D [37]. Another perspective consists of dealing with the changes of topology of the edges in order to automatically detect bifurcations and T-junctions.

References

[1] T. Pavlidis and Y. T. Liow, "Integrating region growing and edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 3, pp. 225–233, 1990.

[2] M. Berthod, Z. Kato, S. Yu, and J. Zerubia, "Bayesian image classification using Markov random fields," *Image and Vision Computing*, vol. 14, no. 4, pp. 285–295, 1996.

[3] C. A. Bouman and M. Shapiro, "Multiscale random field model for Bayesian image segmentation," *IEEE Transactions on Image Processing*, vol. 3, no. 2, pp. 162–177, 1994.

[4] S. C. Zhu, "Region competition: unifying snakes, region growing, and bayes/mdl for multiband image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 884–900, 1996.

[5] G. Aubert and J.-F. Aujol, "Optimal partitions, regularized solutions, and application to image classification," *Applicable Analysis*, vol. 84, no. 1, pp. 15–35, 2005.

[6] G. Aubert and P. Kornprobst, "Mathematical Problems in Image Processing," in *Applied Mathematical Sciences*, vol. 147, Springer, New York, NY, USA, 2001.

[7] J. F. Aujol, G. Aubert, and L. Blanc-Féraud, "Wavelet-based level set evolution for classification of textured images," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1634–1641, 2003.

[8] T. F. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelets, and Stochastic Methods*, SIAM Publisher, Philadelphia, Pa, USA, 2005.

[9] L. D. Cohen, "Avoiding local minima for deformable curves in image analysis," in *Curves and Surfaces with Applications in CAGD*, A. Le Méhauté, C. Rabut, and L. L. Schumaker, Eds., pp. 77–84, 1997.

[10] S. Jehan-Besson, A. Herbulot, M. Barlaud, and G. Aubert, "Shape gradient for image and video segmentation," in *Mathematical Models in Computer Vision: The Handbook*, Springer, New York, NY, USA, 2005.

[11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[12] S. Masnou, "Disocclusion: a variational approach using level lines," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 68–76, 2002.

[13] S. Masnou and J. M. Morel, "Level lines based disocclusion," in *Proceedings of the IEEE International Conference on Image Processing*, vol. 3, pp. 259–263, 1998, 1998.

[14] T. McInerney and D. Terzopoulos, "T-snakes: topology adaptive snakes," *Medical Image Analysis*, vol. 4, no. 2, pp. 73–91, 2000.

[15] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, vol. 42, no. 5, pp. 577–685, 1989.

[16] S. Osher, *Geometric Level Set Methods in Imaging, Vision and Graphics*, Springer, New York, NY, USA, 2003.

[17] N. Paragios, N. Ayache, and J. Duncan, *Biomedical Image Analysis: Methodologies and Applications*, Springer, New York, NY, USA, 2008.

[18] N. Paragios and R. Deriche, "Geodesic active regions and level set methods for motion estimation and tracking," *Computer Vision and Image Understanding*, vol. 97, no. 3, pp. 259–282, 2005.

[19] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia, "Level set model for image classification," *International Journal of Computer Vision*, vol. 40, no. 3, pp. 187–197, 2000.

[20] C. Samson, L. Blanc-Féraud, G. Aubert, and J. Zerubia, "A variational model for image classification and restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 5, pp. 460–472, 2000.

[21] A. Tsai, A. Yezzi, and A. S. Willsky, "Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1169–1186, 2001.

[22] L. A. Vese and T. F. Chan, "A multiphase level set framework for image segmentation using the Mumford and Shah model," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 271–293, 2002.

[23] J. Weickert, "Efficient image segmentation using partial differential equations and morphology," *Pattern Recognition*, vol. 34, no. 9, pp. 1813–1824, 2001.

[24] A. Yezzi, A. Tsai, and A. Willsky, "A fully global approach to image segmentation via coupled curve evolution equations," *Journal of Visual Communication and Image Representation*, vol. 13, no. 1-2, pp. 195–216, 2002.

[25] D. Auroux and M. Masmoudi, "A one-shot inpainting algorithm based on the topological asymptotic analysis," *Computational and Applied Mathematics*, vol. 25, no. 2-3, pp. 251–267, 2006.

[26] D. Auroux and M. Masmoudi, "Image processing by topological asymptotic expansion," *Journal of Mathematical Imaging and Vision*, vol. 33, no. 2, pp. 122–134, 2009.

[27] M. Masmoudi, "The topological asymptotic," in *Computational Methods for Control Applications*, R. Glowinski, H. Karawada, and J. Périaux, Eds., vol. 16, pp. 53–72, GAKUTO International Series. Mathematical Sciences and Applications, Tokyo, Japan,, 2001.

[28] D. Auroux, M. Masmoudi, and L. Jaafar Belaid, "Image restoration and classification by topological asymptotic expansion," in *Variational Formulations in Mechanics: Theory and Applications*, E. Taroco, E. A. de Souza Neto, and A. A. Novotny, Eds., CIMNE, Barcelona, Spain, 2006.

[29] L. J. Belaid, M. Jaoua, M. Masmoudi, and L. Siala, "Image restoration and edge detection by topological asymptotic expansion," *Comptes Rendus Mathematique*, vol. 342, no. 5, pp. 313–318, 2006.

[30] L. Jaafar Belaid, "An overview of the topological gradient approach in image processing: advantages and inconveniences," *Journal of Applied Mathematics*, vol. 2010, Article ID 761783, 19 pages, 2010.

[31] H. G. Senel, "Topological gradient operators for edge detection," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '07)*, vol. 3, pp. 61–64, 2004.

[32] P. Wen, X. Wu, and C. Wu, "An interactive image inpainting method based on rbf networks," in *Proceedings of the 3rd International Symposium on Neural Networks*, pp. 629–637, 2006.

[33] T. Zhou, F. Tang, J. Wang, Z. Peng, and Q. Wang, "Digital image inpainting with radial basis functions," *Journal of Image and Graphics*, vol. 9, no. 10, pp. 1190–1196, 2004.

[34] T. Chan and J. Shen, "Mathematical models for local deterministic inpaintings," Tech. Rep. 00-11, CAM Reports—UCLA Mathematics, 2000.

[35] T. Chan and J. Shen, "Non-texture inpainting by curvature-driven diffusions (CDD)," Tech. Rep. 00-35, CAM Reports—UCLA Mathematics, 2000.

[36] M. Elad, J. L. Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 340–358, 2005.

[37] L. D. Cohen, "Minimal paths and fast marching methods for image analysis," in *Mathematical Models in Computer Vision: The Handbook*, N. Paragios, Y. Chen, and O. Faugeras, Eds., Springer, New York, NY, USA, 2005.

[38] L. D. Cohen and R. Kimmel, "Global minimum for active contour models: a minimal path approach," *International Journal of Computer Vision*, vol. 24, no. 1, pp. 57–78, 1997.

[39] F. Benmansour and L. D. Cohen, "Tubular structure segmentation based on minimal path method and anisotropic enhancement," *International Journal of Computer Vision*, vol. 92, no. 2, pp. 192–210, 2011.

[40] F. Benmansour and L. D. Cohen, "Fast object segmentation by growing minimal paths from a single point on 2D or 3D images," *Journal of Mathematical Imaging and Vision*, vol. 33, no. 2, pp. 209–221, 2009.

[41] L. D. Cohen, "Multiple contour finding and perceptual grouping using minimal paths," *Journal of Mathematical Imaging and Vision*, vol. 14, no. 3, pp. 225–236, 2001.

[42] T. Deschamps and L. D. Cohen, "Fast extraction of minimal paths in 3D images and applications to virtual endoscopy," *Medical Image Analysis*, vol. 5, no. 4, pp. 281–299, 2001.

[43] H. Li, A. Yezzi, and L. Cohen, "3D multi-branch tubular surface and centerline extraction with 4D iterative key points," in *Proceedings of the 12th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI '09)*, Imperial College, London, UK, 2009.

[44] R. Ardon, L. D. Cohen, and A. Yezzi, "A new implicit method for surface segmentation by minimal paths in 3D images," *Applied Mathematics and Optimization*, vol. 55, no. 2, pp. 127–144, 2007.

[45] D. Auroux, "From restoration by topological gradient to medical image segmentation via an asymptotic expansion," *Mathematical and Computer Modelling*, vol. 49, no. 11-12, pp. 2191–2205, 2009.

[46] H. Ammari, M. S. Vogelius, and D. Volkov, "Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter II. The full Maxwell equations," *Journal des Mathematiques Pures et Appliquees*, vol. 80, no. 8, pp. 769–814, 2001.

[47] S. Garreau, P. Guillaume, and M. Masmoudi, "The topological asymptotic for PDE systems: the elasticity case," *SIAM Journal on Control and Optimization*, vol. 39, no. 6, pp. 1756–1778, 2001.

[48] P. Guillaume and K. Sididris, "The topological asymptotic expansion for the dirichlet problem," *SIAM Journal on Control and Optimization*, vol. 41, no. 4, pp. 1042–1072, 2002.

[49] P. Guillaume and K. Sididris, "The topological sensitivity and shape optimization for the stokes equations," *SIAM Journal on Control and Optimization*, vol. 43, no. 1, pp. 1–31, 2004.

[50] S. Amstutz, I. Horchani, and M. Masmoudi, "Crack detection by the topological gradient method," *Control and Cybernetics*, vol. 34, no. 1, pp. 81–101, 2005.

[51] B. Samet, S. Amstutz, and M. Masmoudi, "The topological asymptotic for the Helmholtz equation," *SIAM Journal on Control and Optimization*, vol. 42, no. 5, pp. 1523–1544, 2004.

[52] L. Jaafar Belaid, M. Jaoua, M. Masmoudi, and L. Siala, "Application of the topological gradient to image restoration and edge detection," *Engineering Analysis with Boundary Elements*, vol. 32, no. 11, pp. 891–899, 2008.

[53] J. Dicker, *Fast marching methods and level set methods: an implementation*, Ph.D. thesis, University of British Columbia, 2006.

[54] J. A. Sethian, *Level set methods and fast marching methods*, Cambridge University Press, 1999.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2011, Article ID 815246, 8 pages
doi:10.1155/2011/815246

Research Article

A Novel FEM-Based Numerical Solver for Interactive Catheter Simulation in Virtual Catheterization

Shun Li,[1] Jing Qin,[1,2] Jixiang Guo,[1] Yim-Pan Chui,[1] and Pheng-Ann Heng[1,2]

[1] *The Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong*
[2] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Hong Kong*

Correspondence should be addressed to Shun Li, lis@cse.cuhk.edu.hk

Received 1 July 2011; Revised 12 September 2011; Accepted 13 September 2011

Academic Editor: Shan Zhao

Virtual reality-based simulators are very helpful for trainees to acquire the skills of manipulating catheters and guidewires during the vascular interventional surgeries. In the development of such a simulator, however, it is a great challenge to realistically model and simulate deformable catheters and guidewires in an interactive manner. We propose a novel method to simulate the motion of catheters or guidewires and their interactions with patients' vascular system. Our method is based on the principle of minimal total potential energy. We formulate the total potential energy in the vascular interventional circumstance by summing up the elastic energy deriving from the bending of the catheters or guidewires, the potential energy due to the deformation of vessel walls, and the work by the external forces. We propose a novel FEM-based approach to simulate the deformation of catheters and guidewires. The motion of catheters or the guidewires and their responses to every input from the interventionalist can be calculated globally. Experiments have been conducted to validate the feasibility of the proposed method, and the results demonstrate that our method can realistically simulate the complex behaviors of catheters and guidewires in an interactive manner.

1. Introduction

Vascular interventional radiology (VIR) [1] is a minimally invasive surgery (MIS) procedure. It has been widely used to cure vascular diseases, such as stroke, angiostenosis, and aneurysm. This therapy is performed by using two main kinds of instruments, catheters, and guidewires (for brevity, we use "catheters" to represent "catheter, and guidewires" hereafter) both of which are very flexible cylindrical instruments. In the procedures, they are inserted in the patient's vascular system and driven by the interventionalists to the desired point. This task is complicated because only 2D X-ray images are available, and the catheters have to be handled by the tail. It becomes a challenge to train the novices to let them acquire the skills for safe and efficient procedures [2].

There are some traditional methods for the training of catheterization skills, where the trainees practice on animals, alternative anatomic phantoms, or even actual patients. Due to distinct anatomical differences between animals' and human beings' vascular network, animals are not good substitutes of human beings for training. On the other hand, it is very difficult and expensive to produce a phantom with complex blood vessels the same as real patients. Operating on real patients directly cannot be acceptable as well, as it is very dangerous to both patients and trainees. In contrast to these traditional training methods, virtual reality- (VR-) based simulation systems provide a promising way for catheterization training with high flexibility, high realism, and low cost while without risks to patients and trainees, and there have been some research works on developing such systems in the last few years [3–6]. In order to provide a virtual training environment, a simulator would be developed to simulate the behavior of the catheters navigating inside the patient's vascular system. Therefore, the position of the catheter inside the blood vessel and its changes by the operations from the interventionalists, such as pulling or pushing, ought to be figured out by a numerical algorithm.

Several methods have been proposed to simulate the behaviors of catheters in catheterization procedures. Dawson et al. [7] firstly employed a set of rigid links connected by joints to simulate catheters where the catheter was moved by three forces, such as contact force, injection force, and forces exerted by users. However, this model cannot realistically simulate the complex behaviors of catheters in

catheterization. Later, Wang et al. [8] developed a mass-spring model for catheter simulation dynamically, but it is not consistent with physics laws of elastic thin objects. Cotin et al. [9–11] model the catheter with a set of linked deformable beams. They proposed an incremental finite element method (FEM) built on the strain-stress model of the beams for catheter simulation. Because of the local and incremental characteristics of their approach, the local errors generated when calculating the displacements of the beams can also be translated incrementally, and it is rather difficult to restrict the total error to an acceptable level. An relatively accurate model was proposed in the work of Alderliesten et al. [12, 13], which resorts to the principle of energy minimization to figure out the equilibrium of the catheters, and a semianalytic method is developed to solve this model. However, its computational cost is too expensive to be acceptable for an interactive simulator. Recently, Tang et al. [14] developed a simulating approach based on the work of Bergou et al. [15], where the virtual catheter was driven by elastic forces acted on each node of a discrete catheter. However, the stability and accuracy of this simulator is restricted by the time step used in the numerical solver.

Inspired by the methods proposed by Alderliesten et al. [12, 13], in this paper, we propose a novel method to simulate the motion of catheters and their interactions with patients' vascular system based on the principle of minimal total potential energy. We formulate the total potential energy in the vascular interventional circumstance by summing up the elastic energy deriving from the bending of the catheters, the potential energy due to the deformation of the vessel wall, and the work by the external forces. In order to overcome the shortcoming of expensive computational costs of the method by Alderliesten et al., we proposed a novel FEM-based approach to figure out the deformation of catheters while interacting with the blood vessel wall, which transforms the problem of minimizing the energies to solving a linear system. Thus, the motion of the catheter and its responses to every input from the interventionalist can be calculated globally. Our method provides a good trade-off between the accuracy and efficiency; that is, our method can achieve relatively accurate simulation while maintaining interactive performance. Comparing with other interactive simulating methods, since our method is based on the principle of total energy minimization, it can supply more realistic deformation of the catheters. In contrast to the method proposed by Alderliesten et al. our method can achieve comparable accuracy and much faster performance to make the simulator run in an interactive manner.

The rest of this paper is organized as follows. Section 2 provides the details on the physically based deformable model and the numerical algorithm for simulating catheters. Section 3 reports experiments and evaluation results. Finally, conclusions are drawn in Section 4.

2. Method

2.1. Total Potential Energy of a Catheter. During VIR interventions, a catheter is confined inside blood vessels and advanced along vasculature driven by the operations from the interventionalists. It is observed that a catheter would, regardless of what operations are performed on it, trend to reach an equilibrium state and finally be static if there are no continuing inputs, which can be well explained by the principle of minimum potential energy. That is the catheters would deform or displace to a position that minimizes the total potential energy. Therefore, we can solve the position and shape of the catheter by minimizing its potential energy.

We can define the total potential energy U of a catheter in the vascular interventional circumstance as the sum of three different components: the elastic energy U_e deriving from the bending of the catheter, the potential energy U_p generated by the interactions with the blood vessel wall, and the work W by external forces (e.g., the frictions and the forces from the users) acted on the catheter:

$$U = U_e + W + U_p. \tag{1}$$

2.2. The FEM-Based Numeric Solver for Interactive Catheter Simulation. We simulate the dynamics of a catheter during VIR procedures by employing a FEM- [16] based numerical solver, where the continuous catheter can be discretized into a set of elements (segments with two end nodes in our case), and thus the degree of freedom (DoF) (positions and tangents of nodes in our case) of the catheter can be limited. We proposed a series of methods to formulate the three aforementioned energy terms in the form of quadratic polynomial functions of the tangents of the discrete catheter. To minimize the total potential energy, we calculate the partial derivative of the quadratic polynomial functions with respect to the tangents and then build a linear system. By solving this linear system, we achieve the solution with minimal potential energy.

2.2.1. Formulation of Elastic Energy. First, we formulate the bending energy based on the Kirchhoff's theory of elastic rod [17, 18]. The Kirchhoff's theory is widely used in mechanics to formulate the elastic energy of deformed thin objects. In general, the elastic energy includes bending energy and twisting energy. However, in catheterization procedures, catheters have excellent torque controls, and it is usually assumed that the torsion constant of the catheters approaches infinity [12]. As a result, the twist is not taken into consideration when we formulate the elastic energy. Thus, the elastic energy U_{e} of the catheter can be defined as

$$U_{\mathrm{e}} = \frac{1}{2}\int_0^L \alpha(\mathbf{x}''(s))^2 ds, \tag{2}$$

where the $\mathbf{x}(s)$ is the function of the centerline curve of the catheter with respect to the arc length s, the L is the total length of the catheter, and the α is the bending constant. To avoid the difficulty in solving the second-order derivative in the energy function, we choose the tangent of the catheter's centerline to replace the function of the centerline curve. As the function of tangent $\mathbf{t}(s)$ is the derivative of the $\mathbf{x}(s)$:

$$(\mathbf{t}(s) = \mathbf{x}'(s)), \tag{3}$$

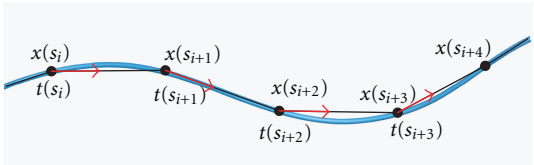


Figure 1: the discretized catheter with several elements.

the potential energy U_e can be represented as

$$U_e = \int_0^L \alpha(\mathbf{t}'(s))^2 ds. \tag{4}$$

To discretize the continuous catheter, we can apply the piecewise first-order polynomials to interpolate the function $\mathbf{t}(s)$. We divided the catheter into n elements $(\overrightarrow{s_0s_1}, \overrightarrow{s_1s_2}, \ldots, \overrightarrow{s_{n-1}s_n})$ as shown in Figure 1.

In the discretization, we make each element has the same length for computational convenience. The function of $\mathbf{t}(s)$ between nodes s_{i-1} and s_i can be interpolated on the segment by the first-order polynomial as follows:

$$\mathbf{t}(s) = \frac{s - s_i}{s_{i-1} - s_i}\mathbf{t}_{i-1} + \frac{s - s_{i-1}}{s_i - s_{i-1}}\mathbf{t}_i. \tag{5}$$

Substituting (5) into (4), we obtain the discretized energy function:

$$U_e = \sum_{i=0}^{n-1} \int_{s_i}^{s_{i+1}} \alpha(\mathbf{t}'(s))^2 ds. \tag{6}$$

After figuring out the integral equation, the U_e is actually the second-order polynomial with respect to $\mathbf{t}_i$.

2.2.2. Formulation of Potential Energy Caused by Interactions. Next, we formulate the potential energy U_p generated by the interactions with vessel walls by employing the method proposed by Alderliesten et al. [12], which is based on the Hooke's law [19]. In our simulation, blood vessels are modeled by triangular meshes [20]. During the interventional procedures, once a collision is detected, we think of it as a contact between a node of the catheter and a triangle of the vessel mesh. There should be a penetration at the contact node. We can regard the penetration as the deformation of the blood vessel wall at the contact node. Thus we can construct the formula of the U_p at the contact node according to the Hooke's Law: $U_p = (1/2)\kappa p_j^2$, where j is the number of the contact nodes, p_j is the vertical distance from the contact node $\mathbf{x}_j$ to the contact triangle, and κ is the modulus of elasticity of blood vessel wall. The plane of contact triangle can be expressed as a linear equation $g(\mathbf{x}_j)$, so the penetrating distance can be figured out as: $\mathbf{p}_j = |g(\mathbf{x}_j)|$. Summing up all contact points, the total energy of U_p can be defined as

$$U_p = \sum_j \frac{1}{2}\kappa \left| g\left(\mathbf{x}_j\right) \right|^2. \tag{7}$$

As we can represent the $\mathbf{x}(s)$ with the tangent function $\mathbf{t}(s)$ by integrating (3):

$$\mathbf{x}(s) = \int_0^s \mathbf{t}(s) ds, \tag{8}$$

if we substitute (5) into (8) and then figure out the integration, the nodal value $\mathbf{x}_i$ can be represented by a first-order polynomial with respect to the $\mathbf{t}_i$, $i \in (0, 1, \ldots, n-1)$. Thus, the U_p can be transformed into a quadratic polynomial with respect to the tangent $\mathbf{t}_i$.

2.2.3. Formulation of the Work by External Forces. The external forces in this application may include the frictions and the forces from the users. However, actually in clinical practice, in order to avoid the damage to the vessels of patients, the catheters are usually clothed by some biomedical materials to reduce the frictions with the blood vessel wall. The frictions are very small during the catheterization; therefore, we ignore them in our model. Hence we only take into account the forces from the users. It can be defined as

$$W = \sum_i \mathbf{f}_i \cdot \mathbf{d}_i, \tag{9}$$

where $\mathbf{f}_i$ is the external force exerted on the node i and $\mathbf{d}_i$ is the difference between current position of node i and its position in the last equilibrium state. So the $\mathbf{d}_i$ can be calculated by $\mathbf{x}_i - \mathbf{x}_{i0}$, where $\mathbf{x}_{i0}$ is the position of node i in the last equilibrium state. Also, in terms of (8), the W can be transformed into a quadratic polynomial with respect to the tangent $\mathbf{t}_i$.

2.2.4. The Numerical Solver. In the interventional procedure, it is necessary to insert a basic sheath into the blood vessels at first. It provides safe access to the interior of the vascular network. It can be used to prevent bleeding during the procedure and restrict the direction of the catheter inserting the vessels [2]. Therefore, in our model, we regard the constant initial tangent (i.e., $\mathbf{t}_0$) as a boundary condition. Then, to derive the conditions of energy minimization, we calculate the partial derivative of the sum of total potential energy U with respect to each $\mathbf{t}_i$, $i \in (1, 2, \ldots, n-1)$ and achieve a set of linear equations which can be expressed in matrix form as

$$\mathbf{At} = \mathbf{b}, \tag{10}$$

where $\mathbf{A}$ is a matrix by $3(n-1) \times 3(n-1)$, $\mathbf{t}$ is the vector $(\mathbf{t}_1^T, \ldots, \mathbf{t}_{n-1}^T)^T$ by $3(n-1) \times 1$, and $\mathbf{b}$ is a constant vector. During the calculation, we find that the matrix $\mathbf{A}$ can be easily transformed into an upper triangular matrix without zero element in its diagonal. As a result, it is a nonsingular matrix.

As mentioned previously, the $\mathbf{x}_i$ can be expressed as the first-order polynomial of the $\mathbf{t}_i$. This relationship can be represented in matrix form:

$$\mathbf{Bt} = \mathbf{x}, \tag{11}$$

where $\mathbf{B}$ is a lower triangular matrix by $3(n-1) \times 3(n-1)$, and there is no zero element in its diagonal, so it is a nonsingular matrix. Substituting (11) into (10), we obtain a linear system which can be expressed in matrix form:

$$\mathbf{AB}^{-1}\mathbf{x} = \mathbf{b}. \tag{12}$$

For every input from an interventionalist, we can reach the new equilibrium state of the catheter by solving (12).

(1) Initialize the positions $\mathbf{x}_i (i \in [0, \ldots, n])$ of the catheter in an equilibrium state.
(2) Translate each node of the virtual catheter a displacement by the operation from the motion sensor of our simulator $\mathbf{x}_i^{\text{new}} = \mathbf{x}_i^{\text{old}} + \mathbf{disp}$.
(3) Detect the collisions to create the set of colliding nodes and triangular meshes.
(4) Build or update the matrix A, the matrix B^{-1} and the vector $\mathbf{b}$.
(5) Solve the linear system $\mathbf{AB}^{-1}\mathbf{x} = \mathbf{b}$.
(6) Update the positions $\mathbf{x}_i$ of the virtual catheter to the new equilibrium state.

Algorithm 1: The overall algorithm of the proposed numerical solver.

Table 1: Timing performance of our method.

Number of node	Average execution time (ms)	Frames per second
50	9.2	109
100	15.4.	65
150	24.8	40
200	36.7	27
300	58.4	17

In terms of the specific shape of the matrix B, the inverse matrix of $\mathbf{B}$ can be determined easily. In order to speed up our simulation, after assembling and calculating the matrix $\mathbf{AB}^{-1}$, we employ a commercial library named CUBLAS [21], which is an implementation of BLAS (basic linear algebra subprograms) on top of the NVIDIA CUDA (compute unified device architecture) driver, to solve the linear equations in our method.

2.2.5. The Overall Algorithm. The overall algorithm of our solver is shown in Algorithm 1.

3. Experiments and Results

3.1. Implementation. We have integrated the catheters simulation into a virtual reality-based training system. It is based on a PC with a Intel Core2 6700 CPU, 4 GB memory and a NVIDIA GeForce 8800 GPU, and a hardware device made by ourself for motion sensing of the catheter. There are two views for the trainees in the system: one is the 3D navigation view, the other is the fluoroscopic view (Figure 2).

3.2. Experiment 1: Time Performance. In this experiment, we tested the time performance of our method for the virtual catheters with different number of nodes when they advanced in a virtual tubular blood vessel. We show the results in Table 1. As shown in the table, it takes about 36 milliseconds to complete a calculation of our algorithm for a catheter with 200 nodes. The FPS (frames per second) can be maintained about 30, which is suitable for an interactive system. Even when the number of the nodes increased to 300, the system can still reach a frame rate of 17 FPS.

3.3. Experiment 2: Catheters Navigation in Vascular System. In this set of experiment, we evaluated the capability of our method in simulating the catheters' navigation in various vascular structures. The virtual catheters modeled by our method are pushed or pulled by the interventionalist and constrained inside the vascular system. We employed the method reported in the work [22] to detect the collisions between a discretized catheter and blood vessels wall made up of triangular meshes. Under the acting force from the interventionalist and the reacting force from blood vessel walls, the catheter advances in various vascular structures. We show the snapshots of catheters' navigation in the different areas of vascular system in Figure 3.

3.4. Experiment 3: Behaviors of Catheters in a Vessel. We further conducted an experiment to evaluate the simulated behaviors of a catheter when moving within a blood vessel. A transparent plastic tube was used in our experiment to act like a tubular blood vessel. In this experiment, a real catheter was inserted into the plastic tube. Here, we mainly emulated a common operative situation, where the catheter would be distorted during its moving forward when its soft tip was looped back inside the blood vessel wall.

Figure 4 shows the results of the experiment comparing the real situation to the simulated one. From these four consecutive pictures, we can find that the distortion of the floppy region of the catheter becomes larger along its advancement. It was due to the fact that when the tip of the catheter collides with the tube, the tip was stopped from advancing so that a loop was formed. This is a very common situation which occurred in real operations. We can observe that our method can mimic this phenomenon well.

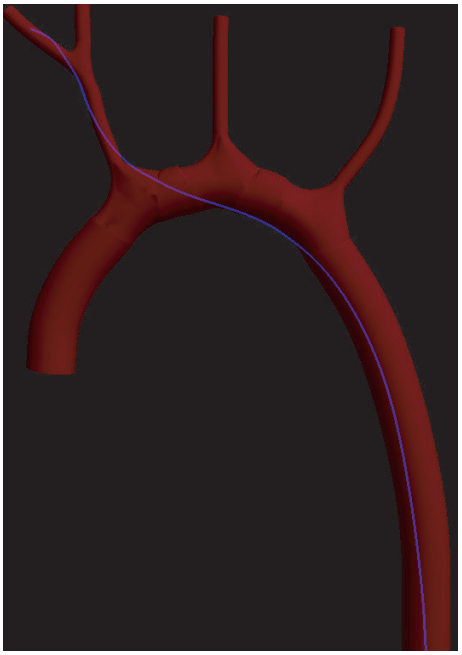
3.5. Experiment 4: Comparison of the Deformation between the Simulated Catheter and the Real One. Finally, we conducted a set of experiments to compare the simulated catheter advanced in the curved virtual vessel and the real one in the plastic tubular phantom to validate the realism of the deformation of the catheter in our method. The experiment for real catheter was performed to insert a real catheter into a curved plastic tube and advance it to a desired position as shown in Figure 5(a). The size of the curved plastic tube is also labeled in the figure. The shape and the position of the real catheter were acquired as the ground truth. In the virtual environment, we create a 3D model as the virtual blood vessel according to the size and the shape of the real plastic tube, and then the virtual catheter simulated by our method was inserted and advanced to the same position as shown in Figure 5(b). We acquire the position of the virtual catheter and compare them with the ground truth.

(a)

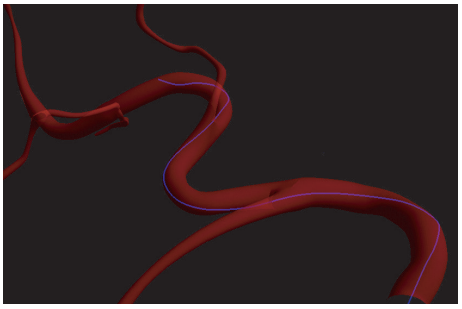
(b)

Figure 2: Visualization of our simulator: (a) fluoroscopic view to simulate the X-ray imaging (b) 3D anatomic view with the virtual vascular system, skin, and catheters.

We used the root-mean-square (RMS) error to measure the difference of deformation between the real catheter and the simulated one. The RMS is computed from the distances between the nodal positions in the simulated catheter and a set of reference nodes in the ground truth. We acquired those reference points by resampling the catheter in the ground truth with the segment length used for each specific experiment. For n nodes, the formula of RMS is $\mathrm{RMS} = \sqrt{(1/n)\sum_{i=0}^{n-1}(\mathrm{dist}_i)^2}$, where $\mathrm{dist}_i = \|x_i^s - x_i^r\|$, x_i^s is the simulated nodal position and x_i^r is the corresponding reference nodes. In addition, we also list the maximum displacement among all of the dist_i to measure the difference of deformation. In Table 2, besides RMS and the maximum displacement, we also list the total runtime in seconds of the whole procedure of the experiments.



(a)



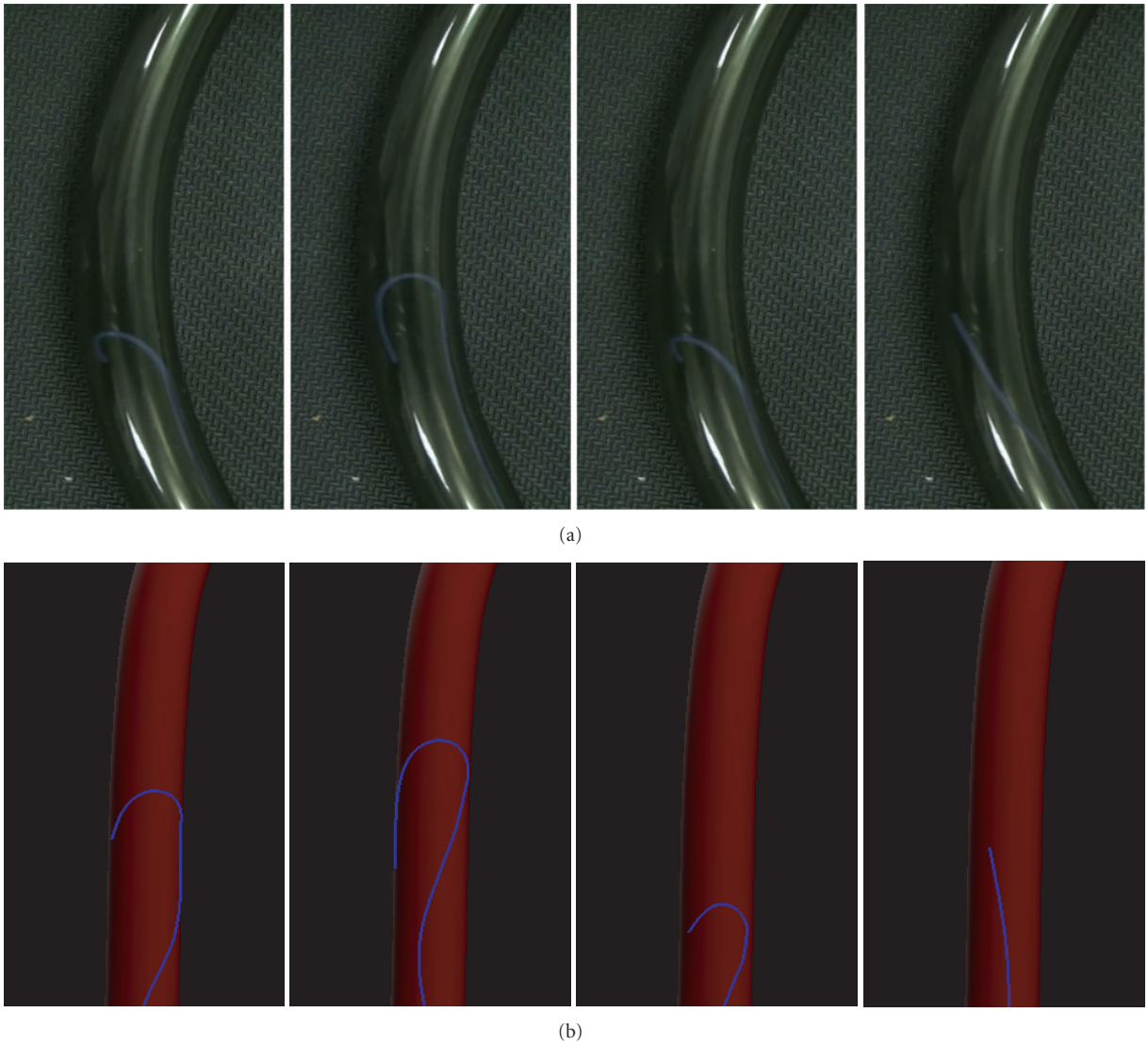
(b)

Figure 3: The virtual catheter navigating in different area of vascular system: (a) cardiovascular structure and (b) hepatic arterial structure.

The results can be compared with the experiment data in the work of Alderliesten et al. [13]. It can be observed that the error of our method is slightly bigger than their results, but the time performance is much better than their method. For example, when the segment length is 1 mm and the ratio of the stepsize to the segment length is 1/10, the runtime of our results is 41.4 seconds, while the runtime in the work of Alderliesten et al. [13] is 2117.3 seconds. According to our experimental results, We can find that the errors are becoming smaller with the reduction of the segment length, however the runtime is increasing correspondingly. Therefore, there is a trade-off between the accuracy and efficiency. We should choose the segment length as small as possible, at the same time make the simulator run in the real-time interactive manner.

4. Conclusion and Discussion

The VR-based surgical simulator is widely applied to teach and train the medical students. It is indispensable to make



(a)

(b)

Figure 4: Advancement of a catheter within a tube: (a) real situation and (b) simulated results.

Table 2: The comparison of the deformation between the simulated catheter and the real one: for different combinations of segment length l and stepsize of each input (pulling or pushing) h, the RMS error (mm) (left value), the maximum displacement error (mm) (middle value), and the total runtime of the whole procedure of the experiments in seconds (right value) are listed.

h/l	l								
	1 mm			2 mm			3 mm		
1/40	1.38	2.32	180.7	1.54	2.42	49.7	1.76	2.74	29.5
1/20	1.16	2.12	84.5	1.47	2.24	23.2	1.66	2.38	15.4
1/10	1.12	2.16	41.4	1.35	2.36	11.5	1.86	2.45	6.9
1/5	0.96	1.52	20.5	1.26	1.88	5.9	1.44	2.56	3.9
1/3	1.05	1.76	11.6	1.24	2.13	3.5	1.62	2.22	2.1

the simulator interact with the trainees with real-time response. Beyond that, the simulator should provide an as realistic virtual environment as possible in which trainees can fully immerse themselves as if they were in real operating scenarios. In the catheterization, the simulation of behaviors of catheters is a very important and relatively complex component of a VR simulator. In this paper, we are dedicated to building the physically deformable model for the simulation of the catheters and simulating the interaction between the catheters and the blood vessel wall. In our method, we regard the motion of the catheter as the transition from an equilibrium state to another; therefore, we formulate the potential energy function relevant to the elastic property of the catheter, the deformation of the blood vessel wall, and

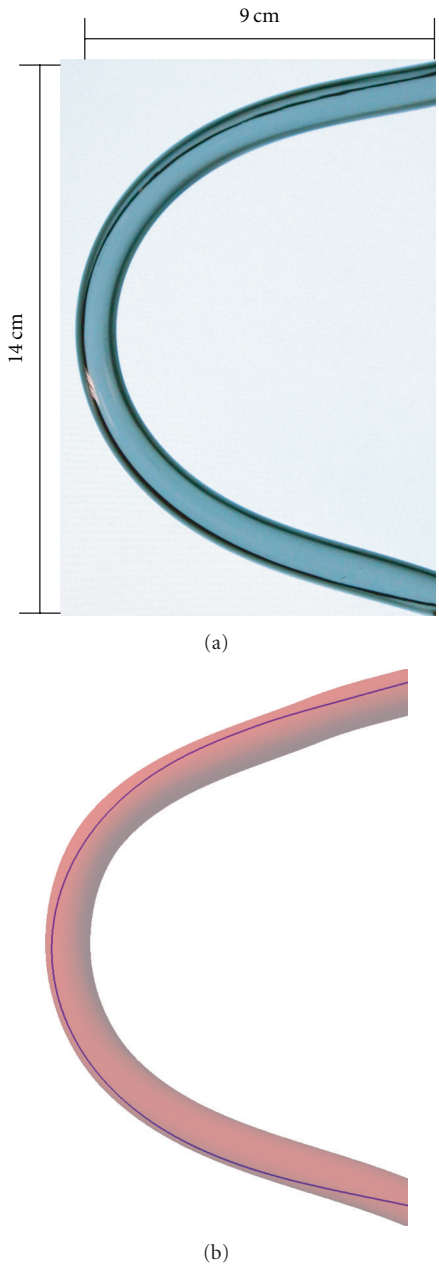


Figure 5: The comparison of the deformation between the simulated catheter and the real one.

the work by external forces. We resort to the concept of the FEM to construct and solve a linear system to achieve the new equilibrium state of the catheter with responses to each input from the interventionalist. Our method is integrated into a simulator for the training of VIR surgeries, and the behavior of the catheter simulated by our method can make the training process realistic.

However, in our proposed method we do not take into consideration the twisting problem of the catheter which is useful for the simulation of some other VIR procedure, such as embolization. Therefore, we will adopt the concept of the frames to represent the twisting state of the catheter and adapt our potential energy function for involving the twisting energy in the future. Furthermore, by improving our deformable model, we will extend our method to simulate other devices such as coils as well as the embolization process which is performed to cure the arterial aneurysm by deploying the coils in the aneurysm.

In the future, another work of us is to evaluate our virtual system by means of empirical study approach. In details, we will design a set of specific training subject in our virtual system based on the real catheterization procedure. And the virtual angiography procedure will also be integrated into our system, so that the trainees can practice the catheterization procedure under the guide of the simulated 2D X-ray imaging. We will invite the medical students and some specialists of interventionalists to participate our experiments. Then, we will let them complete a series of experiments and analyze the results to estimate the validity for training of our virtual system.

Acknowledgment

The work described in this paper was supported by the following grants: (i) General Research Fund sponsored by the Research Grants Council of Hong Kong (Project no. CUHK4121/08E), (ii) NSFC/RGC Joint Research Scheme sponsored by the Research Grants Council of Hong Kong and the National Natural Science Foundation of China (Project nos. N_CUHK409/09 and 60931160441).

References

[1] A. Bloom and R. Gordon, "Vascular interventional radiology," in *Smith's General Urology*, p. 112, 2004.

[2] P. Schneider, *Endovascular Skills: Guidewire and Catheter Skills for Endovascular Surgery*, Marcel Dekker, 2003.

[3] Y. Cai, C. Chui, X. Ye, Y. Wang, and J. H. Anderson, "VR simulated training for less invasive vascular intervention," *Computers and Graphics*, vol. 27, no. 2, pp. 215–221, 2003.

[4] Y. Y. Cai, C. K. Chui, X. Ye, J. H. Anderson, K. M. Liew, and I. Sakuma, "Simulation-based virtual prototyping of customized catheterization devices," *Journal of Computing and Information Science in Engineering*, vol. 4, no. 2, pp. 132–139, 2004.

[5] V. Luboz, C. Hughes, D. Gould, N. John, and F. Bello, "Real-time seldinger technique simulation in complex vascular models," *International Journal of Computer Assisted Radiology and Surgery*, vol. 4, no. 6, pp. 589–596, 2009.

[6] D. Zhang, T. Wang, D. Liu, and G. Lin, "Vascular deformation for vascular interventional surgery simulation," *International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 6, no. 2, pp. 170–177, 2010.

[7] S. L. Dawson, S. Cotin, D. Meglan, D. W. Shaffer, and M. A. Ferrell, "Designing a computer-based simulator for interventional cardiology training," *Catheterization and Cardiovascular Interventions*, vol. 51, no. 4, pp. 522–527, 2000.

[8] F. Wang, L. Duratti, E. Samur, U. Spaelter, and H. Bleuler, "A computer-based real-time simulation of interventional radiology," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '07)*, pp. 1742–1745, IEEE, 2007.

[9] S. Cotin, C. Duriez, J. Lenoir, P. Neumann, and S. Dawson, "New approaches to catheter navigation for interventional

radiology simulation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '05)*, vol. 2, pp. 534–542, 2005.

[10] J. Dequidt, J. Lenoir, and S. Cotin, "Interactive contacts resolution using smooth surface representation," *Medical Image Computing and Computer-Assisted Intervention*, vol. 10, no. 2, pp. 850–857, 2007.

[11] J. Lenoir, S. Cotin, C. Duriez, and P. Neumann, "Interactive physically-based simulation of catheter and guidewire," *Computers and Graphics*, vol. 30, no. 3, pp. 417–423, 2006.

[12] T. Alderliesten, M. K. Konings, and W. J. Niessen, "Simulation of guide wire propagation for minimally invasive vascular intervention," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '02)*, vol. 2, pp. 245–252, 2002.

[13] T. Alderliesten, P. A. N. Bosman, and W. J. Niessen, "Towards a real-time minimally-invasive vascular intervention simulation system," *IEEE Transactions on Medical Imaging*, vol. 26, no. 1, pp. 128–132, 2007.

[14] W. Tang, P. Lagadec, D. Gould, T. R. Wan, J. Zhai, and T. How, "A realistic elastic rod model for real-time simulation of minimally invasive vascular interventions," *Visual Computer*, vol. 26, no. 9, pp. 1157–1165, 2010.

[15] M. Bergou, M. Wardetzky, S. Robinson, B. Audoly, and E. Grinspun, "Discrete elastic rods," in *Proceedings of the ACM SIGGRAPH Asia courses*, pp. 1–12, August 2008.

[16] D. L. Logan, *A First Course in the Finite Element Method*, Thomson, 2007.

[17] C. C. Lin and H. R. Schwetlick, "On the geometric flow of kirchhoff elastic rods," *SIAM Journal on Applied Mathematics*, vol. 65, no. 2, pp. 720–736, 2005.

[18] D. Singer, "Lectures on elastic curves and rods," in *Proceedings of the Curvature and Variational Modeling in Physics and Biophysics*, vol. 1002, pp. 3–32, Citeseer, 2008.

[19] G. Arfken, H. Weber, and F. Harris, *Mathematical Methods for Physicists*, vol. 148, Academic press, New York, NY, USA, 1995.

[20] J. X. Guo, S. Li, Y. P. Chui et al., "PPU-based deformable models for catheterisation training," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI '07)*, pp. 24–32, 2007.

[21] C. Nvidia, *Cublas Library*, NVIDIA Corporation, Santa Clara, Calif, USA, 2008.

[22] S. Gottschalk, M. Lin, and D. Manocha, "OBBTree: a hierarchical structure for rapid interference detection," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 171–180, ACM, 1996.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2011, Article ID 918978, 15 pages
doi:10.1155/2011/918978

Research Article

Protein Surface Characterization Using an Invariant Descriptor

Zainab Abu Deeb,[1] Donald A. Adjeroh,[1] and Bing-Hua Jiang[2]

[1] *Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA*
[2] *Department of Pathology, Anatomy and Cell Biology Jefferson Medical College, Thomas Jefferson University, Philadelphia, PA 19107, USA*

Correspondence should be addressed to Donald A. Adjeroh, don@csee.wvu.edu

Received 6 July 2011; Accepted 14 August 2011

Academic Editor: Guowei Wei

Aim. To develop a new invariant descriptor for the characterization of protein surfaces, suitable for various analysis tasks, such as protein functional classification, and search and retrieval of protein surfaces over a large database. *Methods*. We start with a local descriptor of selected circular patches on the protein surface. The descriptor records the distance distribution between the central residue and the residues within the patch, keeping track of the number of particular pairwise residue cooccurrences in the patch. A global descriptor for the entire protein surface is then constructed by combining information from the local descriptors. Our method is novel in its focus on residue-specific distance distributions, and the use of residue-distance co-occurrences as the basis for the proposed protein surface descriptors. *Results*. Results are presented for protein classification and for retrieval for three protein families. For the three families, we obtained an area under the curve for precision and recall ranging from 0.6494 (without residue co-occurrences) to 0.6683 (with residue co-occurrences). Large-scale screening using two other protein families placed related family members at the top of the rank, with a number of uncharacterized proteins also retrieved. Comparative results with other proposed methods are included.

1. Introduction

The Protein Data Bank (http://www.pdb.org/pdb/home/home.do) (PDB) currently has more than 3000 protein structures classified as uncharacterized or as proteins of unknown function. This is about 5% of the total structures in PDB. The Pfam database was recently reported to contain over 2200 gene families with unknown function [1]. It has been argued that there are even more local regions on the protein structures that are not completely characterized, and whose functions are not known [2]. Therefore, with the increasing rate at which protein structures are being generated, the problem of protein function annotation has become a major challenge in the postgenomic era [3–5]. The function of a given protein is largely determined by its three-dimensional structure [6]. The specific shape and orientation of a protein in 3D space are key elements that determine how the protein interacts with its environment, and hence the function of the protein. Although related proteins often have similar functions, it is well known that sequence similarity between proteins does not always lead to functional similarity [7, 8]. Even different functions have been observed for structures with the same fold [9]. Conversely, sequences have been observed with low sequence similarity, but highly structural and functional similarity [10]. The trypsin-like catalytic triad [9] is one example of proteins with different folds, but similar functions. A similar argument can be made between sequence and surface, and between surface and fold. While residues on the protein surface typically make up a small percentage of the total residues in a protein, they often represent the most conserved functional elements of the protein [11]. Therefore, analyzing protein structures using information about their 3D surfaces is essential in the quest for protein function annotation, especially in the study of functional similarities between nonhomologous proteins.

At the core of most activities in the analysis of protein structures and protein function is similarity measurement between structures. Such measurements must deal with different levels of structural similarity, arbitrary mutations, deletions, and insertion of residues, local surface similarities,

and so forth. When the problem is similarity measurement between protein surfaces, a major issue becomes how the protein surface is represented, and how the representation can be used for the required similarity measurement. Another problem is that of computation. Structure alignment, the basis for most approaches to protein 3D structure analysis is known to be NP-hard [12]. A major difficulty in comparing protein surfaces locally is the problem of matching 3D structures, since structures need to undergo an exhaustive amount of rotation and translation in order to obtain an adequate structural alignment and to perform an accurate matching [8]. Clearly, a method that avoids the step of local structural alignments can have a significant advantage, especially in screening of similar surfaces over a large database.

In this paper, we introduce an invariant descriptor for the characterization of protein surfaces. We then use this characterization to study the problem of classifying proteins into their functional families based primarily on their surface characteristics. This is a challenging problem, but one that is important in the quest for functional annotation of proteins, using information from potentially nonhomologous proteins. We also show how we can use such a descriptor in various related analysis activities, such as in effective retrieval of similar protein surfaces from very large databases, such as the Protein Data Bank (PDB).

2. Background and Related Work

2.1. Protein Sequences, Structure, and Surface. Although proteins could vary significantly in their functions and 3D shapes, they also share a general common structure. Proteins are composed of 20-amino acids that are connected via peptide bonds [13]. Each protein is composed of an ordered sequence of amino acids. The order in which these amino acids are connected is called the *protein sequence*, or the *primary structure* of the protein [14]. This primary sequence determines the 3D structure of the protein. All proteins are composed of four common structural types: primary structure, secondary structure, tertiary structure, and quaternary structure. The primary structure is simply the amino acid sequence. The *secondary structure* is formed by patterns of intermolecular bonding of hydrogen and is determined primarily by the location and the directions of these patterns [14, 15]. This is often described in terms of secondary structural elements (SSEs), such as α-helixes, β-sheets, and turns. The overall 3D shape of the secondary structures determines the *tertiary structure* of the protein. When two or more chains combine to form a larger molecule, the whole structure is called the *quaternary structure*. Figure 1 shows an example of some of the common protein structural types (the sequence is not included).

A common method for protein function prediction is by annotation transfer from known homologous proteins [17]. Functions of novel proteins can be determined by sequence comparisons, for instance using sequence alignment. When proteins evolve, the protein structure remains more highly conserved when compared to the sequence. Protein sequences change more easily during evolution due to residue mutations, for instance by substitution, insertion, or deletion. Hence, proteins that belong to the same family (homologous proteins) may not be identified using sequences alone. Orengo et al. [17] reported that proteins related to the same family could share fewer than 15% identical residues. The protein structure retains a significant portion of similarity even between distant homologs. In general, the degree of structural or sequence similarity varies substantially between protein families. Some families can handle more changes than others. This so-called *structural plasticity* [17] has a considerable impact on the functionality of some proteins, or members of a protein family. A consideration of the protein structure and its variability becomes important in such situations for further analysis of functional similarity between proteins.

A classical approach for deriving the protein function is by first determining its 3D structure, which can then provide some ideas about its function [17]. Protein 3D structures provide information about the binding sites, active sites, and how proteins interact with each other, and thus could provide an insight into the function of the protein [17]. How proteins interact with each other and with other molecules (e.g., ligands) is determined primarily by the amino acids on the protein surface [18]. Therefore, knowledge of the protein surface residues could help in a better understanding of what molecules are binding together, and in some cases, why they bind [18]. The protein surface could also provide significant information about protein functions which cannot be easily detected, even in the presence of sequence or fold similarity. Therefore, the analysis of protein surfaces is important in the study of intermolecular interactions. Clearly, advances in our understanding of protein surfaces could have important implications in various biomedical fields, such as personalized medicine, drug discovery, drug design, and so forth.

2.2. Protein Surface Characterization Methods. Given the foregoing, it is not surprising that different methods have been proposed to characterize the protein surface. Popular examples include those based on surface shape distributions [19], Gauss integral [20], Fourier transform [21], spherical harmonics [22, 23], alpha-shapes [2, 24], and Zernike polynomials [7]. Contact maps between protein surfaces were studied in [25], while similarity networks between surface patches from protein binding sites were studied in [26, 27]. Protein surface similarity using varying resolutions of structural data have also been studied, for instance, using medium-resolution Cryo-EM maps in [26] and low resolution protein structure data in [28]. SHARP [29] provides a mechanism to predict protein-protein interaction by analyzing overlapping protein 3D surface regions. SURFACE [5] is a database of protein surface regions that can be useful for annotation.

Much earlier, Jones and Thornton [30] analyzed protein-protein interaction by using surface patches, where patches are defined based on the C_α atoms that have a predetermined accessible surface area, and adhere to defined constraints on the solvent vectors. Each patch is then described using

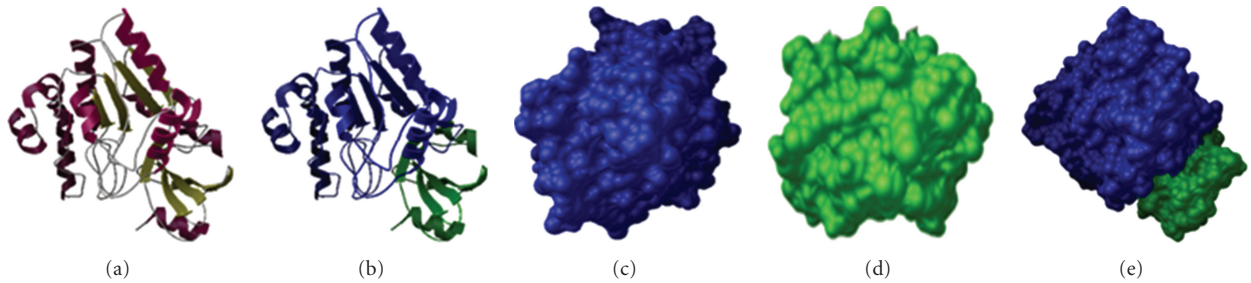


Figure 1: Protein structures for a sample protein (PDB id: 2UDI). (a) Secondary structure elements—α-helixes (magenta), β-sheets (gold), and turns (gray); (b) two chains: chain E (blue), chain I (green); (c) surface and 3D shape for chain E; (d) surface and 3D shape for chain I; (e) quaternary structure for the protein. Figures are produced using PMV [16].

six parameters, namely, solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion, and accessible surface area. Ferrè et al. [5] analyzed locally similar structures by matching surface patches composed of subsets of amino acids. Each residue on the protein surface is represented using a vector joining its C_α atom and the centroid of its side chain atoms. Surface patches are then compared for similarity by comparing the residue vectors for all possible pairs of residues from the query and target surface patch. Matches are determined based on the root mean square distance, and the residue similarity as determined using a standard substitution matrix. The results of using this method on a nonredundant list of protein chains as recorded in the SURFACE database [5], a collection of protein surface regions that can be useful for annotation. Below, we describe three approaches that are more closely related to our work. See [28, 31] for reviews on surface comparison methods.

Distance Distributions. Distances, geometry, and topology have for long been used in the analysis of general protein 3D structures [32]. Residue distances have been used in standard texture-based analysis of 2D textures (distance matrices) formed by the distances between residues in a protein structure [33]. The use of topological invariants, as captured using Gaus integrals for the automated analysis and representation of general protein 3D structures was described in [20]. Much earlier, Connolly [19] proposed the analysis of protein surfaces using the notion of surface shape distributions. Essentially, surface shapes correspond to different geometric configurations defined on the protein surface. Binkowski and Joachimiak [11] proposed the use of surface shape signatures (SSSs) as a method to describe protein surfaces by exploiting global shape and geometrical properties of the surfaces. Shape signatures are computed based on the distances measured between each unique atom pairs on the surface. Distances are then sorted based on which their distributions are generated. With the distributions, the problem of matching between two surfaces is now reduced to that of comparing their distributions. Comparison between two distributions is performed using the Kolmogorov-Smirnov (KS) test. The use of the shape distribution is fast and relatively resilient to scale, rotation, and mirroring. However, the discrimination ability is still a problem, as the SSS tends to lose important surface details.

Zernike Polynomials. Following earlier work by Canterakis [34] on the use of 3D Zernikes for the analysis of general 3D objects, Sael et al. [7] introduced 3D Zernike to the area of protein structural similarity matching. Here, the protein 3D structure is represented as a series expansion of 3D Zernike functions. The triangulated Connolly surface of the protein is computed, and subsequently the protein is placed into a 3D cubic grid and voxelized. Each voxel has a value of **1** or **0**, depending on whether the voxel is on the protein surface or in the interior. The 3D Zernike function is then applied to the voxelized 3D protein shape to obtain the 3D Zernike descriptors. Therefore, the problem of comparison of 3D surfaces is reduced to that of comparing two vectors representing the 3D Zernike descriptors for each protein surface. Several distance measures were tried, such as the Euclidean distance, Manhattan distance, and a correlation-based distance defined as the complement of the correlation coefficient between two Zernike descriptors. Venkatraman et al. [23] studied the use of both spherical harmonics and 3D Zernike descriptors in the retrieval of functionally similar proteins. In a more recent work, Sael and Kihara [28] used the Zernike descriptor to study protein surfaces in low resolution data. Computation of the required Zernike polynomials is, however, known to be a major computational huddle [35]. This problem is even worse for the 3D Zernike polynomials needed for protein surfaces. Thus, the required preprocessing before matching is performed may be a problem for indexing and real-time search of large-scale datasets.

Fingerprints. A recent work [8] used the idea of extracting invariant fingerprints from patches on the protein surface. Patches are obtained by generating the dot surface of the protein and constructing a graph to approximate the protein surface. Afterwards, circular patches are generated as a contiguous surface area from a center point, where the radius of the patch is within a predetermined cutoff. Patches are created for each single point on the surface, after which a fingerprint representation of the patch is computed as a geodesic distance-dependent distribution of directional curvature. Geodesic distances are computed from the central vertex in each patch. Comparisons between fingerprints were performed using the average fingerprint similarity score (AFSS) and the direct fingerprint similarity score (DFSS).

Final scores are computed after an alignment procedure based on the AFSS. Clearly, computational complexity will be a major problem here, especially given the computation of the patch representation for each vertex on the surface graph (number of vertices is much more than the number of surface residues). The need for a later stage of alignment for the final computation of matching scores only compounds the computational burden (see [12], for example).

The key difference in our method is the use of the local patch descriptors as defined by the distribution of distances between C_α atoms within each surface patch, conditioned on the specific residue at the center of the patch, and the particular residues found within the patch. Our method computes the residue-specific distance distributions, and residue-distance cooccurrences for the protein surface patches using only the C_α atoms on the protein surface. Residues in the interior of the protein are discarded. Unlike the approach in [8], we avoid the time complexity of generating a graph representation of the surface before the surface can be scanned to generate the patches and then compute the distance distribution. Further, ours does not depend on the time-consuming process of initial surface alignment.

3. Methods

We present an invariant descriptor for characterizing protein surfaces. We start with a local descriptor of selected circular patches on the protein surface. For a given surface patch, the local descriptor is computed based on the residue distances from the center of the patch. The descriptor records the distance distribution between the central residue and the residues within the patch, keeping track of the number of particular pairwise residue cooccurrences in the patch. A global descriptor for the entire protein surface is then constructed from the local descriptors by combining information from local descriptors with similar central residues. The proposed descriptor is invariant to rotations of the surface and mirroring.

Using a fixed patch size, we obtain a descriptor for the protein surface, independent of the size of the protein structure. Thus, the descriptor can facilitate the rapid matching of protein chains, and will eliminate the need for the exhaustive alignment of the protein 3D structures. For a given protein structure or protein chain from a database, such as the PDB, the proposed method can be summarize in the following steps:

(1) generate the Connolly surface [36] for the protein chain;

(2) generate the surface patches and compute the local invariant descriptor for each patch on the surface;

(3) compute the global invariant surface descriptor for the protein chain, by combining information from the local patch descriptors;

(4) perform surface matching and comparison using the descriptors;

(5) classify the protein into its potential functional family, or perform protein surface retrieval using the invariant descriptors.

Figure 2 shows a schematic diagram of the general approach. The method has been applied on three protein families: *uracil-DNA glycosylase*, *estrogen receptor*, and *cell division protein kinase 2*. These are the same protein families used in a recently published work [8]. We also tested on *epidermal growth factor (EGF)* and *cyclooxygenase-2 (COX-2),* two protein families that are known to play a role in cancer. Below, we provide more details on the steps enumerated above.

3.1. Surface Generation. For a given protein, we first generate its Connolly surface [36] at a given atomic radii, using the MSMS program [37], based on which the dot surface is generated. This dot surface is stored in a vertex file. We have used a probe radius of 1.4 Å in all our experiments. Next, MATLAB Bioinformatics Toolbox (Mathworks Inc, Natick, Mass, USA) was used to extract the protein chains and to generate the residue coordinates in each chain. In this step, the chains are extracted while preserving the coordinates of the C_α atoms and their respective residue types by extracting the information from the PDB and the vertex files.

3.2. The Invariant Descriptor

3.2.1. Surface Patches. To capture protein structure similarity and to avoid the computational complexity and the time-consuming problem of aligning 3D protein structures, we propose the use of a global rotational-invariant descriptor to represent overlapping patches on the protein surface. A patch is defined as a circular region with a specified radius, centered on the C_α position of a surface residue. For each residue on the surface (the central residue), we construct a surface patch by recording its residue type, and consider all residues within a certain distance threshold (τ_p) as part of the patch (see Figure 2). Thus, the proposed surface descriptor is composed of 20 distinct descriptions, one for each protein residue type. For the local descriptor, this is constructed from only information from the patch. For the global descriptor, this is constructed by combining information from patches with the same central residue.

The local invariant descriptor for the patch is created by calculating the distribution of distances between the central residue and all other surface residues within the patch. Additionally, the residue cooccurrences within the patch are also recorded as a part of the local descriptor. Each local descriptor is represented in a matrix $\mathbf{D_A}$ of size $(20 + 1) \times (b + 1)$, where the rows correspond to the 20 distinct protein residue types, plus an extra row to describe the summary distance distribution within the patch. The columns represent the individual bins used to capture the distance distributions (total of b bins), plus an extra column to represent the summary of the residue cooccurrences. To reduce the computational time and space requirement, unlike in [8] we define patches only for surface residue positions, rather than for each vertex on the dot surface (the number of vertexes is much more than the number of residues). Therefore, for

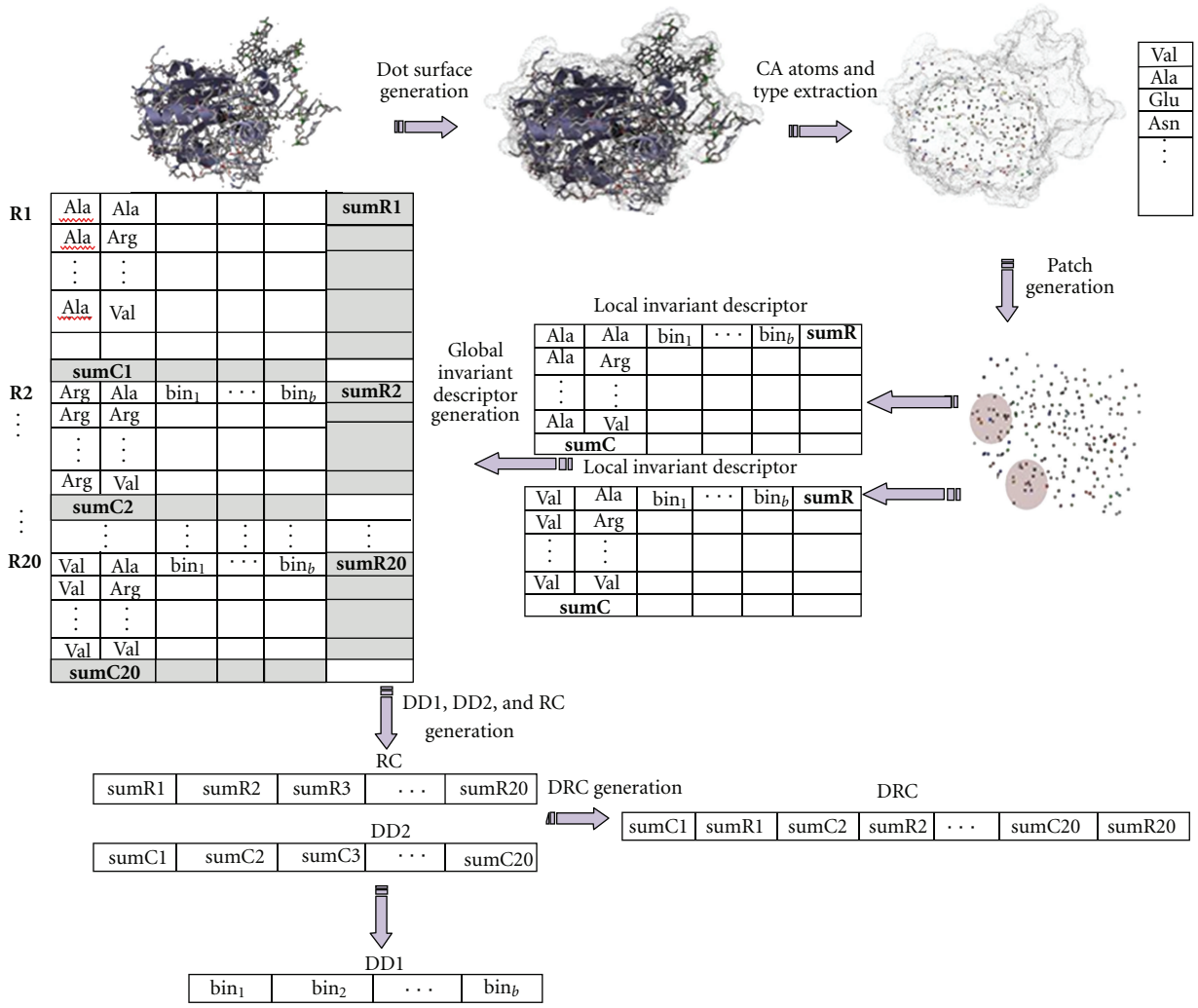


Figure 2: Schematic diagram for the protein surface characterization using an invariant descriptor. Protein structures in the figure are produced using PMV [3].

a given chain, the number of local invariant descriptors will be equal to the number of surface residues. Yet, this number can vary from tens to hundreds and sometimes to thousands of surface residues. Using a huge number of local invariant descriptors for one chain to perform matching will be very time-consuming. To further reduce the computational requirements, for a given chain, we compute a global rotational-invariant descriptor by combining the 20 distinct residue-specific descriptors. For a given residue type, the global descriptor is constructed by taking the average of all patch descriptors with a given residue type as the central residue (see Figure 2). We consider three ways to represent and use the global surface descriptor, as explained below.

3.2.2. Distance Distribution (DD2). The basic idea of using the distance distribution is that similar functional proteins should have a similar distribution of distances between the residues on their surfaces. The patch descriptor captures the distribution in two forms. The *first form* is a detailed distance distribution between the central residue in the surface patch and each of the other residues on the patch. To achieve this, a uniform distribution of the distances is assumed and the total number of bins b is used to estimate the probability distribution of finding a pair of residues at one of the b ranges. The *second form* is the global distance probability distribution. In this form we estimate the probability of observing any given residue within a patch in a particular distance range from the central residue. In this paper, we study the use of the global distance distribution in identifying similar protein surfaces, and possibly proteins with similar functions. Consequently, the question to be answered is, given a central residue of a specific type, what is the distance distribution for the residues around this central residue? That is, we seek $\Pr\{d \mid R_c\}$, the probability of observing distance d between a central residue of type R_c and any other residue. We expect that the distance distribution should be similar for surface patches from functionally similar proteins.

3.2.3. Residue Cooccurrences (RCs). Given that surface structures are more conserved than sequence over evolution [15, 17], we expect that functionally similar proteins are likely to have similar surface residues, even though the order of such residues may have changed. This intuition is captured using residue cooccurrences on the protein surface. Using the distance distribution globally provides an idea of how the distances from the central residue are distributed in the protein surface patch. However, there is no constraint on, or indication of, which residues are involved in the formation of these distributions. The co-occurrence of a given residue with the central residue is calculated as the number of times the residue occurs on a patch with the same central residue. Thus, the main problem would be to find the probability of observing residue say, R_i, given a central residue, say R_c. Again, we expect the probability $\Pr\{R_i \mid R_c\}$, to be similar for protein surfaces from functionally similar proteins. We note that the surface co-occurrence does not depend on the specific distance between the residues involved, as far as R_i is within the patch.

3.2.4. Distance-Residue Cooccurrences (DRCs). The above have considered the distance and the co-occurrence separately. The DRC combines the general distance distribution (represented as a row vector, $\mathbf{sum\,C}$ in matrix $\mathbf{D_A}$) and the residue cooccurrences (represented as a column vector, $\mathbf{sum\,R}$ in matrix $\mathbf{D_A}$) in describing the protein surface (see Figure 2). The residue-distance co-occurrence vector is defined as follows: $D_{\mathrm{RC}} = (\mathbf{sum\,C} \circ \mathbf{sum\,R}^T)$, where $\circ$ is the concatenation operator and $\mathbf{X}^T$ stands for the transpose of $\mathbf{X}$. D_{RC} is used to compute the conditional probability $\Pr\{d \mid R_c, R_i\}$, that is, the probability of observing the distance d between residues R_c and R_i given that R_c is the central residue in the patch. We expect that the residue co-occurrence ($\mathbf{sum\,R}$, or RC) should carry more distinctive functionally relevant information than the general distance distribution ($\mathbf{sum\,C}$, or DD2), since surface residue cooccurrences are likely to be more conserved over evolution. By combining both vectors, we can account for both the geometry of the protein surface and the distribution of specific residues within specific distances on the surface. Using both vectors brings in some biological relevance in the analysis and is likely to lead to improved results in the identification of functionally similar protein surfaces.

3.3. Matching and Classification. Given two proteins, say Protein 1 and Protein 2 we characterize them using their global descriptors, say D_{g1} and D_{g2} respectively. In this work, the global descriptor could be the distance distribution (DD2), residue cooccurrences (RCs), or the distance-residue cooccurrences (DRCs).

Distance Distribution. For matching using the distance distribution we create a vector D_{g1d} that is composed of the 20 global distance distributions represented by all $\mathbf{sum\,C}$ vectors from each descriptor. D_{g1d} is defined as $D_{g1d} = (D_{d1} \circ D_{d2} \circ \cdots \circ D_{d20})$, where $D_{d1}, D_{d2}, \ldots, D_{d20}$ are the distance distributions from each residue type on the surface of Protein 1. Repeat the same process for Protein 2 to create D_{g2d}. Then we perform matching using the simple Euclidean distance: $D_{12} = \sqrt{\sum_{i=1}^{n} [D_{g1d}(i) - D_{g2d}(i)]^2}$.

Residue Cooccurrences. For Protein 1 we create a vector D_{g1c} that combines the 20 residues co-occurrence vectors (denoted $\mathbf{sum\,R}$), defined as $D_{g1c} = (D_{c1}^T \circ D_{c2}^T \circ \cdots \circ D_{c20}^T)$, where $D_{c1}^T, D_{c2}^T, \ldots, D_{c20}^T$ represents $\mathbf{sum\,R_1^T}$, $\mathbf{sum\,R_2^T}$, and $\mathbf{sum\,R_{20}^T}$. Similarly, we compute D_{g2c}. Matching is performed using the Euclidean distance between D_{g1c} and D_{g2c}.

Distance-Residue Cooccurrences. Here, we create a vector **DRC** that is comprised of all of the distance distributions as well as the residue cooccurrences. For Protein 1, we have $\mathrm{DRC}_1 = (D_{d1} \circ D_{c1}^T \circ D_{d2} \circ D_{c2}^T \circ \cdots \circ D_{d1} \circ D_{c1}^T)$. Similarly we obtain DRC_2 for Protein 2. Again for simplicity, matching is performed using the Euclidean distance. Clearly, other distance measures could be used.

Classification. Having computed the surface descriptors and the distance between protein surfaces using the descriptors, one may be interested in determining whether a given unknown protein belongs to some known protein family. Using some training data, we can compute surface descriptors for the known family, and based on these perform the required classification. Classification is performed using Weka [38, 39], an open-source software for machine leaning that provides a suite of classification algorithms.

4. Results and Discussion

4.1. Datasets and Environment. We performed experiments to test the performance of the proposed protein surface descriptor in two protein structure analysis tasks, namely, classifying proteins into their most likely functional groups, and ranking and retrieval of protein surfaces. We used two datasets for the experiments. DATASET-A contained information from three protein families: *uracil-DNA glycosylase*, *cell division protein kinase 2*, and *estrogen receptor*. This was created by scanning the PDB and selecting the protein structures with protein chains belonging to one of the three families. We were able to extract 416 chains that belong to 243 proteins in the PDB. The dataset is distributed as follows: 91 chains from 46 distinct proteins for *uracil-DNA glycosylase* (Group1), 186 chains from 95 distinct proteins for *estrogen receptor* family (Group2), and 139 chains from 102 distinct proteins from *cell division protein kinase 2* (Group3). We used DATASET-A basically to train the system, and perform initial testing. DATASET-B contained protein structures from two families, namely *cyclooxygenase-2 (COX-2)* (51 proteins, 95 chains) and *epidermal growth factor (EGF)* (67 proteins, 71 chains). We then extracted protein structures from the PDB that have 10 or less chains and ignored the rest. This resulted in a total of 15,386 protein chains form 6,261 unique proteins. DATASET-B included all structures in DATASET-A. We used DATASET-B for a more comprehensive scan of the PDB, in the quest for potentially novel structures that may

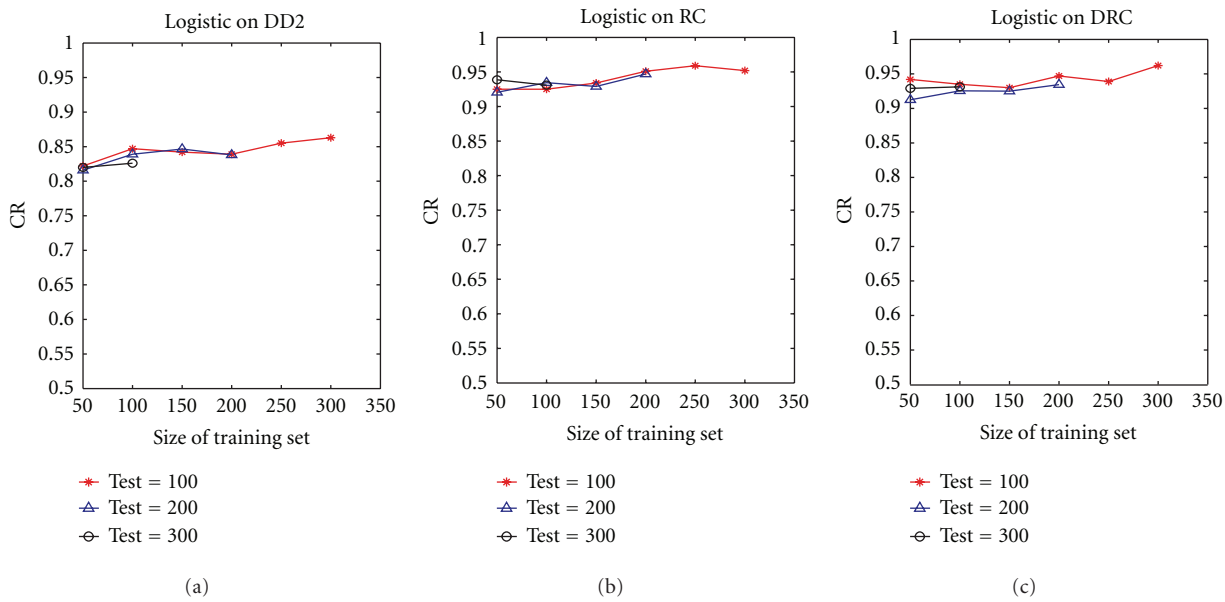


Figure 3: Variation of classification rate (CR) with size of training set using the proposed descriptors DD2 (a), RC (b), and DRC (c). Results are shown for the average over 10 runs, using logistic regression as the classifier.

be related to the two families. Experiments were performed using a SONY VAIO personal computer, with Intel Core 2 Duo Processor T8100, running at 2.10 GHz, with 2 GB of main memory. Programs were written using Matlab (Mathworks Inc, Natick, Mass, USA) with the Bioinformatics Toolbox. We set probe radius = 1.4 Å and patch distance threshold $\tau_p = 10$ Å. For distance distributions, we used a fixed number of bins, $b = 5$. Classification was performed based on algorithms implemented in Weka [38, 39] version 3-6-4.

4.2. Classification Performance. We divide DATASET-A into training and testing sets and apply different classifiers on the different descriptors proposed. In all our experiments, the training sets were kept very separate from the testing sets, with no overlap between the two. Classification performance is measured in terms of classification rate based on the three protein families in the dataset. We tested the method using various classifiers implemented in Weka, such as Naïve Bayes, logistic regression, and simple logistic classifier. We report results mainly for the logistic regression. First, we explore the impact of the size of the testing set and of the training set on the classification performance using the proposed approach. We varied the size of the training set (from 50 to 300), while keeping the size of the testing set fixed. We then checked the performance using fixed testing sets of size 100, 200, and 300. Figure 3 shows the results.

The figure shows that applying the distance distribution (DD2) alone resulted in the lowest performance accuracy as compared to using the residue cooccurrences (RCs) or distance-residue cooccurrences (DRCs). Yet, our definition of the distance distribution shows encouraging results. A steady improvement in performance with increasing training set size can be observed when using DD2 alone, peaking at about 87% with a training size of 200 and testing size of 100. The distinctiveness of our approach is the use of residue cooccurrences on the protein surface. This approach assumes that functionally similar surface proteins have similar residue cooccurrences within a small local surface region. Figure 3 (middle plot) shows that classification using residue cooccurrences (RCs) provided a significant improvement in the classification rate. A similar improvement was observed using other classifiers, such as Naïve Bayes. Using the RC descriptor, we can achieve an accuracy rate of 94% using a small training set (50 samples) and six times larger testing set (300 chains). This shows the robustness of the residue cooccurrences, even when using a few training samples. We observe that the performance using DD2 was not as robust (about 81% using small training set, peaking at about 87% using 200 training samples).

The use of distance-residue co-occurrence presents a steadier improvement in the classification rate. Using the DRC raised the accuracy rate to 99% using the simple logistic classifier on a training set of 150 and testing set of 100 (data not shown). We can observe the significant difference between the results of DD2 (which did not use information on residue cooccurrences) and RC and/or DRC (both of which used residue cooccurrences). Figure 4 shows a corresponding performance measurement with varying size of the testing set, while keeping the training set size fixed. As expected, there is a general slight decrease in performance with increasing size of the test set. The case of DRC using a training set size of 100 seemed to increase slightly with increasing testing set size. The increase is however within a small range (from 0.91 to 0.93). This shows a steady performance over increasing size of the testing set. Overall trends are similar to Figure 3, with RC and DRC performing much better than DD2. Similar trends were also observed using

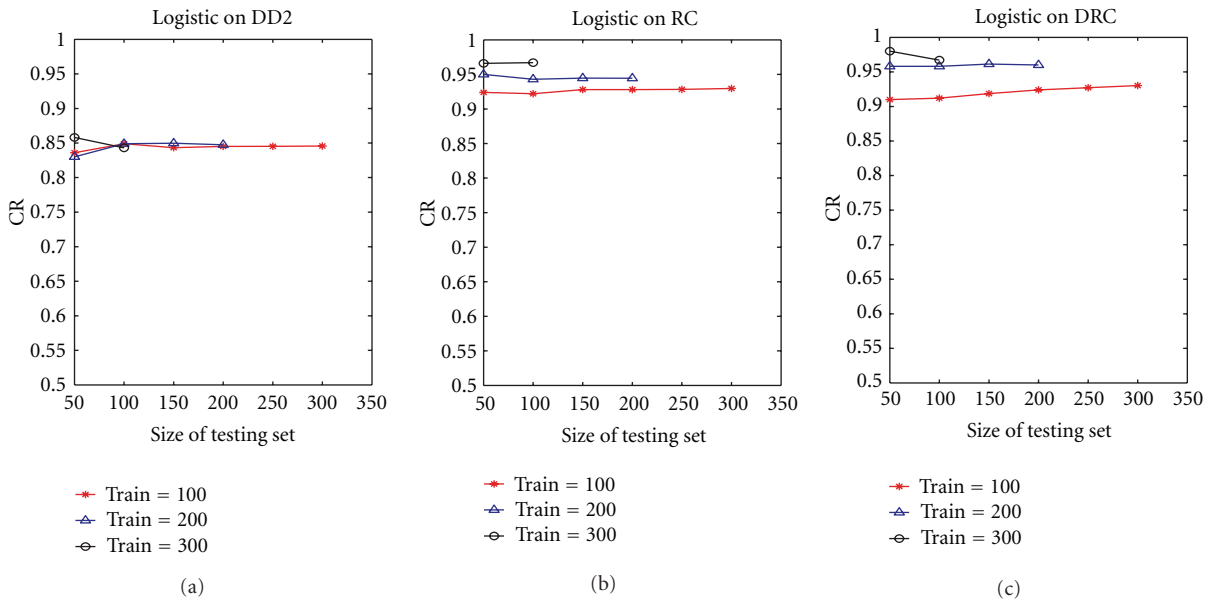


Figure 4: Variation of classification rate (CR) with size of testing set using the proposed descriptors DD2 (a), RC (b), and DRC (c). Results are shown for the average over 10 runs, using logistic regression as the classifier.

other classification algorithms. The overall classification performance is summarized in Figure 5, which shows the results of the three proposed schemes using n-fold cross validation, for different values of n.

4.3. Ranking and Retrieval. In this section, we explore the effectiveness of our approach on the problem of search and retrieval of protein surfaces. Given a query protein, we study whether our approach has the robustness to place most of the functionally similar proteins in the top hits of the retrieved surfaces. Here, a query protein from each of the three groups is used to screen the entire DATASET-A (416 samples) and provide a ranking based on the similarity. Thus, each protein structure is ranked against the query, (from 1 to 416), where a lower rank (smaller distance) implies more similarity to the query. After that, we search over the retrieved proteins to find which ranks the functionally similar proteins (i.e., proteins in the same functional group) have attained. Table 1 shows the ranking produced using the proposed descriptor, for three query samples, one for each group. Results are shown only for DRC. RC produced a slightly better ranking (especially for *uracil-DNA glycosylase* family (Group 1)), while DD2 was worse than both RC and DRC). Overall, for Group 2 and Group 3, the Top 30 ranked proteins belonged to the corresponding family, while Group 1 was more difficult.

We further measured the performance of our approach using the enrichment plot. The enrichment plot essentially measures how well a given ranking or retrieval system performs, when compared with a random selection of the data samples. At a given percentage of database screening, the enrichment factor is computed as the ratio $N_{\text{obs}}/N_{\text{exp}}$, where N_{obs} = number of functionally similar proteins observed or retrieved by the system, and N_{exp} = number of functionally similar proteins expected by random selection. For

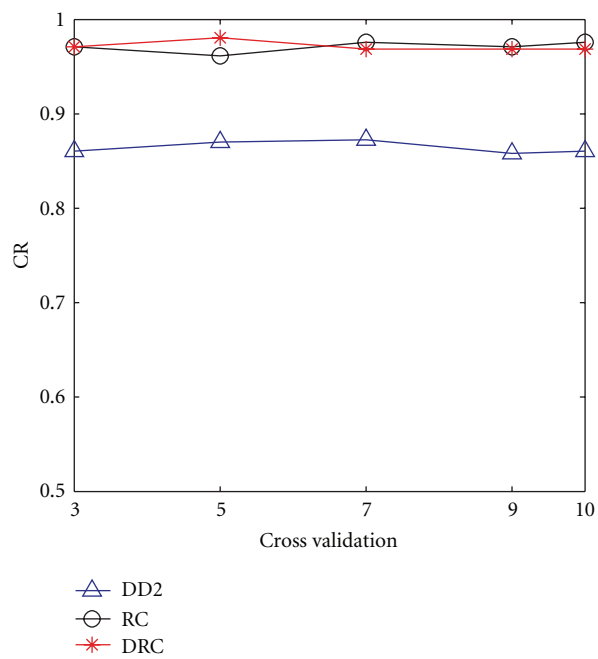


Figure 5: Summary classification performance using n-fold cross validation (the x-axis is for varying n).

an effective system, we expect that most of the functionally similar proteins should be observed after a small percentage of screening. That is, the top hits should contain mainly functionally similar proteins, and hence the enrichment factor should be high after a small percentage probe of the database, and gradually decrease towards 1 (which corresponds to random selection). Figure 6(a) shows a plot of the average enrichment factor using 5 queries from Group 3.

Table 1: Ranking the screened proteins according to their similarity to the query protein. Results are shown for the Top 30 hits for one query protein from each of the three groups, using DATASET-A (416 protein chains).

(a) DRC on query protein 1UDI chain I (Group 1)

Protein PDB ID	Chain	Rank	Distance
1UDI	I	1	0
2ZHX	B	2	2.1306
1LQM	B	3	2.3509
1LQG	C	4	2.4589
2UUG	C	5	2.5353
1EUI	C	7	2.5920
2ZHX	L	8	2.6104
1UGH	I	10	2.6349
2UGI	A	15	2.6809
1UGI	E	16	2.6872
2ZHX	H	19	2.6969
2ZHX	D	21	2.7006
2ZHX	N	22	2.7017
2ZHX	J	23	2.7141
1UGI	G	25	2.7261
1EMJ	A	42	2.7758
2BOO	A	45	2.7808
1UGI	D	47	2.7868
2OWR	B	50	2.7952
2J8X	D	61	2.8129
1LQG	D	70	2.8263
1Q3F	A	90	2.8533
2UUG	D	99	2.8675
1UGI	A	101	2.8689
2OWR	C	110	2.8749
2ZHX	F	116	2.885
2OWQ	B	129	2.8977
1SSP	E	141	2.9115
1UGI	C	142	2.9117
2ZHX	A	147	2.915

(b) DRC on query protein 1QKN chain A (Group 2)

Protein PDB ID	Chain	Rank	Distance
1QKN	A	1	0
2J7X	A	2	1.3769
2J7Y	A	3	1.5753
1QKM	A	4	1.6793
1NDE	A	5	1.7368
2GIU	A	6	1.7371
1L2I	A	7	1.7460
1U3R	B	8	1.7670
3ERD	A	9	1.7683
3OS9	A	10	1.7715

(b) Continued.

Protein PDB ID	Chain	Rank	Distance
2IOG	A	11	1.7738
3LTX	C	12	1.7742
1YIM	A	13	1.7854
3ERT	A	14	1.7966
1U3Q	D	15	1.8009
1YY4	A	16	1.8090
2OUZ	A	17	1.8126
1YIN	A	18	1.8152
1XP6	A	19	1.8260
2AYR	A	20	1.8269
3OS8	D	21	1.8311
2QH6	A	22	1.8312
3OSA	A	23	1.8367
1L2J	A	24	1.8385
2JJ3	A	25	1.8438
1G50	A	26	1.8490
3OS8	A	27	1.8509
2FSZ	A	28	1.8518
2QGW	A	29	1.8604
1UOM	A	30	1.8683

(c) DRC on query protein 1YKR chain A (Group 3)

Protein PDB ID	Chain	Rank	Distance
1YKR	A	1	0
2UZO	A	2	1.0396
2R3O	A	3	1.0810
3PXY	A	4	1.0846
3PY1	A	5	1.0940
2WMA	A	6	1.1389
2IW6	A	7	1.1609
3NS9	A	8	1.2043
3IGG	A	9	1.2141
2WFY	A	10	1.2179
2C5Y	A	11	1.2270
3DDP	A	12	1.2280
2J9M	A	13	1.2284
2R3J	A	14	1.2374
2R3L	A	15	1.2402
3PXR	A	16	1.2422
2DUV	A	17	1.2534
1W8C	A	18	1.2586
3DOG	A	19	1.2793
2V22	A	20	1.2822
2R3P	A	21	1.2883
2V22	C	22	1.2963
3IG7	A	23	1.3207
2JGZ	A	24	1.3275
2R64	A	25	1.3303

(c) Continued.

Protein PDB ID	Chain	Rank	Distance
2WHB	A	26	1.3381
2VTN	A	27	1.3458
3LFN	A	28	1.3476
2WIP	A	29	1.3514
2BKZ	A	30	1.3580

The enrichment plot shows that our proposed method provides better results as we screen a small percentage of the dataset. In most of the cases, our method retrieved about three times better than the expected random retrieval in the first 10% of screened proteins. As we increase the percent of screening, the retrieval degrades, since we are more likely to have retrieved most, if not all of the similar proteins after a small percentage of the screening. Thus, subsequent retrievals will lead to spurious results.

4.4. Screening Protein Surfaces in PDB. Encouraged by the results in classification and ranking using the proposed descriptors, we now performed a larger scale experiment, by screening the entire protein structures in PDB, using the protein chains in DATASET-B, with members of the *COX-2* and *EGF* families as the query. The main objective was to see how the proposed descriptors will perform on a large scale, and to see if the methods could predict potentially novel functional linkages between any of the families and other proteins in PDB. For this task, we used only PDB files with 10 or less chains, and ignored the rest. This resulted in a total of 15,386 protein chains from 6,261 unique proteins. Table 2(a) shows the ranking results produced by screening the PDB files based on the proposed descriptors, using a member of the *EGF* family as a query. Table 2(b) shows corresponding results using a member of *COX-2* family. Results are shown only for the DRC descriptor. Generally, similar results were obtained using RC. We can notice that some of the unknown proteins (annotated as "uncharacterized") were placed in the Top-50 rankings, implying a possible relationship with the respective families.

4.5. Comparison with Related Methods. The use of distance distributions for protein surface analysis was studied by Binkowski et al. [11]. As earlier discussed, they did not consider the specific residues in constructing the distributions. Their distance distribution (labeled as DD1 in this work) is obtained by removing the reference to the specific residue at the center of the patch (see Figure 2). Our use of surface residue cooccurrences and combining these with the residue-specific distance distributions are novel methods introduced in this paper. Tables 3(a) and 3(b) compare the overall classification performance using DD1 with those obtained with the proposed descriptors.

Figure 6 also shows the comparative performance using both the enrichment plots, and precision and recall. We define precision and recall at a given distance threshold as follows: precision = (number of correct retrievals at the threshold)/(number of total retrievals at the threshold).

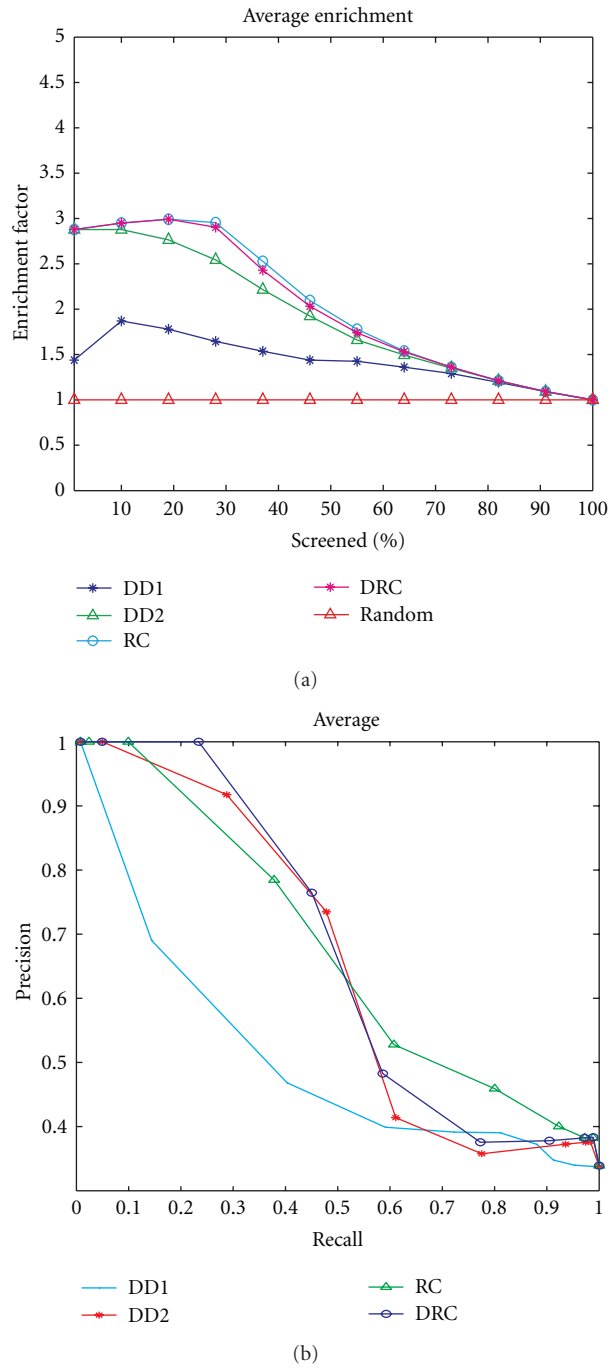


Figure 6: Ranking and retrieval performance for the proposed methods. (a) Enrichment plot for screening protein structures using the proposed descriptors. Results are average for 5 query proteins from *cell division protein kinase 2* family (Group 3), using DATASET-A (416 protein chains). (b) Average precision and recall for three queries, one for each group in DATASET-A. DD1 corresponds to the distance distribution proposed in [11], as described in Section 2 (see Section 4.5 on comparison with related methods).

Recall = (number of correct retrievals at the threshold)/(number of total true matches expected at the threshold). Here, using the ranked results, for a given query and a given

Table 2

(a) Top 50 hits using DRC for a query protein structure from the EGF family on DATASET-B. Annotations in bold correspond to members of the EGF family, predicted proteins, or uncharacterized proteins

Protein	Chain	Distance	Protein name annotation	Rank
2a2q	L	0.0000	**COAGULATION FACTOR VII**	1
2fir	L	1.4925	**COAGULATION FACTOR VII LIGHT CHAIN**	2
2zp0	L	1.5545	**FACTOR VII LIGHT CHAIN**	3
1wtg	L	1.5628	**COAGULATION FACTOR VII**	4
1wun	L	1.5844	**COAGULATION FACTOR VII**	5
2b8o	L	1.6305	**COAGULATION FACTOR VII LIGHT CHAIN**	6
2zwl	L	1.6379	**FACTOR VII LIGHT CHAIN**	7
2zzu	L	1.6536	**FACTOR VII LIGHT CHAIN**	8
1wqv	L	1.6816	**COAGULATION FACTOR VII**	9
2ec9	L	1.6832	**COAGULATION FACTOR VII**	10
1dan	L	1.7655	**BLOOD COAGULATION FACTOR VIIA**	11
1wss	L	1.7659	**COAGULATION FACTOR VII**	12
2puq	L	1.7692	**COAGULATION FACTOR VII**	13
1fak	L	1.7934	**PROTEIN (BLOOD COAGULATION FACTOR VIIA)**	14
2b7d	L	1.8024	**COAGULATION FACTOR VII**	15
6acn	A	1.8061	ACONITASE	16
2aer	L	1.8120	**COAGULATION FACTOR VII**	17
2aei	L	1.8164	**COAGULATION FACTOR VII**	18
2flr	L	1.8196	**COAGULATION FACTOR VII**	19
3ela	L	1.8668	**COAGULATION FACTOR VII LIGHT CHAIN**	20
1z6j	L	1.8859	**COAGULATION FACTOR VII**	21
2f9b	L	1.9027	**COAGULATION FACTOR VII**	22
3phs	A	1.9169	CELL WALL SURFACE ANCHOR FAMILY PROTEIN	23
3n54	B	1.9263	SPORE GERMINATION PROTEIN B3	24
3qbp	B	1.9264	FUMARASE FUM	25
3ma9	L	1.9367	TRANSMEMBRANE GLYCOPROTEIN	26
3mt0	A	1.9921	**UNCHARACTERIZED PROTEIN PA1789**	27
3lgu	A	2.0004	PROTEASE DEGS	28
3m7i	A	2.0071	TRANSKETOLASE	29
1qfk	L	2.0096	**PROTEIN (COAGULATION FACTOR VIIA (LIGHT CHAIN))**	30
3n9t	A	2.0169	PNPC	31
3pxz	A	2.0235	CELL DIVISION PROTEIN KINASE 2	32
2flb	L	2.0257	**COAGULATION FACTOR VII**	33
3lh1	A	2.0289	PROTEASE DEGS	34
3nlc	A	2.0300	**UNCHARACTERIZED PROTEIN VP0956**	35
3no5	C	2.0302	**UNCHARACTERIZED PROTEIN**	36
3msq	C	2.0347	PUTATIVE UBIQUINONE BIOSYNTHESIS PROTEIN	37
3ryk	A	2.0389	DTDP-4-DEHYDRORHSMNOSE 3,5-EPIMERASE	38
3m4a	A	2.0444	PUTATIVE TYPE I TOPOISOMERASE	39
3n3n	B	2.0444	CATALASE-PEROXIDASE	40
2R3G	A	2.0456	CELL DIVISION PROTEIN KINASE 2	41

(a) Continued.

Protein	Chain	Distance	Protein name annotation	Rank
2R3I	A	2.0465	CELL DIVISION PROTEIN KINASE 2	42
3o0r	L	2.0473	ANTIBODY FAB FRAGMENT LIGHT CHAIN	43
3n3p	B	2.0490	CATALASE-PEROXIDASE	44
3nfh	A	2.0492	DNA-DIRECTED RNA POLYMERASE I SUBUNIT RPA49	45
3qfk	A	2.0501	**UNCHARACTERIZED PROTEIN**	46
3n5h	F	2.0517	FARNESYL PYROPHOSPHATE SYNTHASE	47
3n3o	B	2.0529	CATALASE-PEROXIDASE	48
3o78	B	2.0548	CHIMERA PROTEIN OF PEPTIDE OF MYOSIN LIGHT CHAIN SMOOTH MUSCLE, GREEN FLUORESCENT PROTEIN, GREEN FLUORESCENT CALMODULIN	49
3luy	A	2.0570	PROBABLE CHORISMATE MUTASE	50

(b) Top 50 hits using DRC for a query protein structure from the COX-2 family on Dataset-B. Annotations in bold correspond to members of the COX-2 family, predicted proteins, or uncharacterized proteins

Protein	Chain	Distance	Protein name annotation	Rank
2zxw	B	0.0000	**CYTOCHROME C OXIDASE SUBUNIT 2**	1
2eil	B	1.3914	**CYTOCHROME C OXIDASE SUBUNIT 2**	2
2eij	B	1.5853	**CYTOCHROME C OXIDASE SUBUNIT 2**	3
3ag4	B	1.6147	**CYTOCHROME C OXIDASE SUBUNIT 2**	4
2dys	B	1.6991	**CYTOCHROME C OXIDASE SUBUNIT 2**	5
3ag1	B	1.8159	**CYTOCHROME C OXIDASE SUBUNIT 2**	6
2eim	B	1.8824	**CYTOCHROME C OXIDASE SUBUNIT 2**	7
3ag2	B	2.0173	**CYTOCHROME C OXIDASE SUBUNIT 2**	8
2occ	B	2.0631	**CYTOCHROME C OXIDASE**	9
3abl	B	2.1404	**CYTOCHROME C OXIDASE SUBUNIT 2**	10
2eik	B	2.1413	**CYTOCHROME C OXIDASE SUBUNIT 2**	11
3abm	B	2.1454	**CYTOCHROME C OXIDASE SUBUNIT 2**	12
1v55	B	2.1489	**CYTOCHROME C OXIDASE POLYPEPTIDE II**	13
2dyr	B	2.2044	**CYTOCHROME C OXIDASE SUBUNIT 2**	14
2ein	B	2.2628	**CYTOCHROME C OXIDASE SUBUNIT 2**	15
1v54	B	2.2976	**CYTOCHROME C OXIDASE POLYPEPTIDE II**	16
3n56	B	2.4226	INSULIN-DEGRADING ENZYME	17
3abk	B	2.4502	**CYTOCHROME C OXIDASE SUBUNIT 2**	18
3p42	A	2.4785	**PREDICTED PROTEIN**	19
3r2u	B	2.4846	METALLO-BETA-LACTAMASE FAMILY PROTEIN	20
3msu	B	2.4920	CITRATE SYNTHASE	21
3ag3	B	2.4950	**CYTOCHROME C OXIDASE SUBUNIT 2**	22
3ntd	B	2.5052	FAD-DEPENDENT PYRIDINE NUCLEOTIDE-DISULPHIDE OXIDOREDUCTASE	23
3ngi	A	2.5055	DNA POLYMERASE	24
7xim	B	2.5198	D-XYLOSE ISOMERASE	25
3mjy	A	2.5206	DIHYDROOROTATE DEHYDROGENASE	26
3nva	B	2.5391	CTP SYNTHASE	27
3lm3	A	2.5433	**UNCHARACTERIZED PROTEIN**	28
3ppn	B	2.5511	GLYCINE BETAINE/CARNITINE/CHOLINE-BINDING PROTEIN	29
3o98	B	2.5557	BIFUNCTIONAL GLUTATHIONYLSPERMIDINE SYNTHETASE/AM	30

(b) Continued.

Protein	Chain	Distance	Protein name annotation	Rank
3pom	B	2.5558	RETINOBLASTOMA-ASSOCIATED PROTEIN	31
3nt6	B	2.5656	FAD-DEPENDENT PYRIDINE NUCLEOTIDE-DISULPHIDE OXIDOREDUCTASE	32
5lym	B	2.5690	LYSOZYME	33
3n1y	B	2.5728	TOLUENE O-XYLENE MONOOXYGENASE COMPONENT	34
1occ	B	2.5772	**CYTOCHROME C OXIDASE**	35
3lxt	D	2.5860	GLUTATHIONE S TRANSFERASE	36
2q70	B	2.5922	ESTROGEN RECEPTOR	37
3l49	B	2.5930	ABC SUGAR (RIBOSE) TRANSPORTER, PERIPLASMIC SUBSTRATE-BINDING SUBUNIT	38
3pvq	A	2.5944	DIPEPTIDYL-PEPTIDASE VI	39
3puf	B	2.6032	RIBONUCLEASE H2 SUBUNIT B	40
3mve	B	2.6077	UPF0255 PROTEIN VV1_0328	41
3ld2	B	2.6143	PUTATIVE ACETYLTRANSFERASE	42
3ne6	A	2.6160	DNA POLYMERASE	43
3qae	A	2.6179	3-HYDROXY-3-METHYLGLUTARYL-COENZYME A REDUCTASE	44
3qh8	A	2.6197	BETA-LACTAMASE-LIKE	45
3m3r	A	2.6215	ALPHA-HEMOLYSIN	46
3nrb	B	2.6234	FORMYLTETRAHYDROFOLATE DEFORMYLASE	47
3n05	B	2.6240	NH(3)-DEPENDENT NAD(+) SYNTHETASE	48
3m2l	A	2.6324	ALPHA-HEMOLYSIN	49
3pns	B	2.6357	URIDINE PHOSPHORYLASE	50

Table 3

(a) Overall classification rate using different classifiers (300 training samples, 100 testing samples from DATASET-A)

Classifier	Descriptor			
	DD1	DD2	RC	DRC
Naïve bayes	58%	86%	94%	91%
Logistic	58%	85%	99%	97%
Simple logistic	58%	89%	98%	91%

(b) Overall classification rate using different classifiers (100 training samples, 300 testing samples from DATASET-A)

Classifier	Descriptor			
	DD1	DD2	RC	DRC
Naïve bayes	55%	74%	94%	93%
Logistic	62%	88%	89%	94%
Simple logistic	63%	85%	91%	90%

rank, the number of expected true matches will be **min**{rank, query group size}. This is similar to the definition used in [28]. We performed queries on DATASET-A using query proteins from each of the three groups and computed the average precision and recall for each descriptor. We then computed the area under the curve (AUC) for the average precision-recall plots. The results were as follows: DD1 (0.501052), DD2 (0.649412), RC (0.668303), and DRC (0.66759). Although the databases used are different, these results compare well with the results reported by Sael and Kihara [28], where they evaluated the retrieval performance of four surface characterization methods, based on the Zernike representation. The maximum AUC reported using standard resolution surfaces was 0.608 (without length filtering) and 0.628 (with length filtering). Yin et al. [8] proposed a fingerprint-based method, using surface alignment on selected surface patches. Their method constructs an initial patch on every vertex on the dot surface, and requires computation of geodesic distances on the surface, two very time-consuming processes. Our method neither requires surface alignment, nor expensive computations on the surface, beyond the surface generation process. Patches are generated only on positions of the surface residues, rather than over all the vertices on the generated protein surface.

4.6. Computation Time. The most time consuming part was for preprocessing, as needed to construct the protein surfaces and extract the protein chains. The construction of the protein surface from the original PDB files required about 4.065 seconds, running on *Cygwin* (a version of Linux for Microsoft Windows). Extraction of the protein chains and the C_α atoms was performed using Matlab Bioinformatics Toolbox (Mathworks Inc., Natick, Mass, USA), and required 32 seconds per PDB file. Construction of the descriptors after the above steps took an average of 0.7 seconds per PDB file.

Querying DATASET-B (15,386 chains, 6,261 unique PDB files) using the DRC descriptor required an average time of 0.28 seconds for each query PDB file.

5. Conclusion

We have introduced a novel approach to the description and characterization of protein surfaces. The proposed approach captures the surface structure of the protein by utilizing local patches defined only on the positions of surface residues, rather than over all surface vertices, or over all the surface atoms. We make residue cooccurrences on the surface a central part of the descriptor. The novelty of this approach can be observed by the ease of handling both local and global variation on the surface (using local and global descriptors). Moving from local to global not only reduces the computational problem of matching 3D structures, but also facilitates direct comparison between protein structures of different sizes. By avoiding the construction of the complete 3D surface and retaining only the surface C_α to do the analysis, the need for surface alignment of the 3D structure is eliminated. Further, we do not need to perform any geometrical transformation to insure reliable matching. This is very important for rapid analysis over a large database, such as the PDB.

We showed results on the performance of the proposed methods in functional classification of proteins into their putative families, based on the surface information. We further compared the results using enrichment plots, and the standard measures of precision and recall. For the three protein families used, we obtained an area under the curve for precision and recall of 0.6494 (DD2), 0.6683 (RC), and 0.6676 (DRC). A screening of the PDB using *COX-2* and *EGF* family members showed that the proposed methods ranked related family members in the Top-20 hits, with a number of uncharacterized proteins also retrieved. It will be interesting to perform further biological lab experiments to verify if any of the retrieved uncharacterized proteins are truly related to the respective families to which they share similar surfaces (as determined by our surface descriptors).

Acknowledgment

The first author was supported in part by the Ministry of Higher Education of Saudi Arabia.

References

[1] L. Jaroszewski, Z. Li, S. S. Krishna et al., "Exploration of uncharted regions of the protein universe," *PLoS Biology*, vol. 7, no. 9, Article ID e1000205, 2009.

[2] Y. T. Yan and W. H. Li, "Identification of protein functional surfaces by the concept of a split pocket," *Proteins*, vol. 76, no. 4, pp. 959–976, 2009.

[3] Y. Loewenstein, D. Raimondo, O. C. Redfern et al., "Protein function annotation by homology-based inference," *Genome Biology*, vol. 10, no. 2, pp. 207.1–207.8, 2009.

[4] S. Sivashankari and P. Shanmughavel, "Functional annotation of hypothetical proteins—a review," *Bioinformation*, vol. 1, no. 8, pp. 335–338, 2006.

[5] F. Ferrè, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich, "SURFACE: a database of protein surface regions for functional annotation," *Nucleic Acids Research*, vol. 32, pp. D240–D244, 2004.

[6] D. Gront, A. Kolinski, and U. H. E. Hansmann, "Protein structure prediction by tempering spatial constraints," *Journal of Computer-Aided Molecular Design*, vol. 19, no. 8, pp. 603–608, 2005.

[7] L. Sael, B. Li, D. La et al., "Fast protein tertiary structure retrieval based on global surface shape similarity," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 4, pp. 1259–1273, 2008.

[8] S. Yin, E. A. Proctor, A. A. Lugovskoy, and N. V. Dokholyan, "Fast screening of protein surfaces using geometric invariant fingerprints," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 39, pp. 16622–16626, 2009.

[9] N. Nagano, C. A. Orengo, and J. M. Thornton, "One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions," *Journal of Molecular Biology*, vol. 321, no. 5, pp. 741–765, 2002.

[10] H. H. Gan, R. A. Perlow, S. Roy et al., "Analysis of protein sequence/structure similarity relationships," *Biophysical Journal*, vol. 83, no. 5, pp. 2781–2791, 2002.

[11] T. A. Binkowski and A. Joachimiak, "Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites," *BMC Structural Biology*, vol. 8, article 45, 2008.

[12] A. Godzik, "The structural alignment between two proteins: is there a unique answer?" *Protein Science*, vol. 5, no. 7, pp. 1325–1338, 1996.

[13] L. K. Buehler and H. H. Rashidi, *Bioinformatics Basics: Applications in Biological Science and Medicine*, CRC Press, 2005.

[14] D. E. Krane and M. L. Raymer, *Fundamental Concepts of Bioinformatics*, Pearson Education, 2003.

[15] D. Whitford, *Proteins: Structure and Function*, John Wiley & Sons, West Sussex, UK, 2005.

[16] M. F. Sanner, A. J. Olson, and J. C. Spehner, "Reduced surface: an efficient way to compute molecular surfaces," *Biopolymers*, vol. 38, no. 3, pp. 305–320, 1996.

[17] C. Orengo, D. Jones, and J. Thornton, *Bioinformatics Genes, Proteins, and Computers*, BIOS Scientific, New York, NY, USA, 2003.

[18] C. Gibas and P. Jambeck, *Developing Bioinformatics Computer Skills*, O'Reilly Media, Sebastopol, Calif, USA, 2001.

[19] M. L. Connolly, "Shape distributions of protein topography," *Biopolymers*, vol. 32, no. 9, pp. 1215–1236, 1992.

[20] P. Rogen and B. Faint, "Automatic classification of protein structure by using Gauss integrals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 1, pp. 119–124, 2003.

[21] M. J. Bayley, E. J. Gardiner, P. Willett, and P. J. Artymiuk, "A fourier fingerprint-based method for protein surface representation," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 696–707, 2005.

[22] R. J. Morris, R. J. Najmanovich, A. Kahraman, and J. M. Thornton, "Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons," *Bioinformatics*, vol. 21, no. 10, pp. 2347–2355, 2005.

[23] V. Venkatraman, L. Sael, and D. Kihara, "Potential for protein surface shape analysis using spherical harmonics and 3d

zernike descriptors," *Cell Biochemistry and Biophysics*, vol. 54, no. 1–3, pp. 23–32, 2009.

[24] L. P. Albou, B. Schwarz, O. Poch, J. M. Wurtz, and D. Moras, "Defining and characterizing protein surface using alpha shapes," *Proteins*, vol. 76, no. 1, pp. 1–12, 2009.

[25] J. Gramm, "A polynomial-time algorithm for the matching of crossing contact-map patterns," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 4, pp. 171–180, 2004.

[26] K. Lasker, O. Dror, M. Shatsky, R. Nussinov, and H. J. Wolfson, "EMatch: discovery of high resolution structural homologues of protein domains in intermediate resolution Cryo-EM maps," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 28–39, 2007.

[27] K. Park and D. Kim, "Binding similarity network of ligand," *Proteins*, vol. 71, no. 2, pp. 960–971, 2007.

[28] L. Sael and D. Kihara, "Improved protein surface comparison and application to low-resolution protein structure data," *BMC Bioinformatics*, vol. 11, no. 11, article S2, 2010.

[29] Y. Murakami and S. Jones, "SHARP2: protein-protein interaction predictions using patch analysis," *Bioinformatics*, vol. 22, no. 14, pp. 1794–1795, 2006.

[30] S. Jones and J. M. Thornton, "Analysis of protein-protein interaction sites using surface patches," *Journal of Molecular Biology*, vol. 272, no. 1, pp. 121–132, 1997.

[31] A. Via, F. Ferrè, B. Brannetti, and M. Helmer-Citterich, "Protein surface similarities: a survey of methods to describe and compare protein surfaces," *Cellular and Molecular Life Sciences*, vol. 57, no. 13-14, pp. 1970–1977, 2000.

[32] W. R. Taylor, A. C. W. May, N. P. Brown, and A. Aszódi, "Protein structure: geometry, topology and classification," *Reports on Progress in Physics*, vol. 64, no. 4, pp. 517–590, 2001.

[33] I. G. Choi, J. Kwon, and S. H. Kim, "Local feature frequency profile: a method to measure structural similarity in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3797–3802, 2004.

[34] N. Canterakis, "3D zernike moments and zernike affine invariants for 3D image analysis and recognition," in *Proceedings of the 11th Scandinavian Conference on Image Analysis*, pp. 85–93, 1999.

[35] R. Mukundan and K. R. Ramakrishnan, "Fast computation of Legendre and Zernike moments," *Pattern Recognition*, vol. 28, no. 9, pp. 1433–1442, 1995.

[36] M. L. Connolly, "Solvent-accessible surfaces of proteins and nucleic acids," *American Association for the Advancement of Science*, vol. 221, no. 4612, pp. 709–713, 1983.

[37] J. Smith, "Computing a triangulated surface with MSMS. In Vanderbilt University Center for Structural Biology," 2011, http://structbio.vanderbilt.edu/comp/soft/msms/tutorial.php.

[38] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[39] R. R. Bouckaert, E. Frank, M. A. Hall et al., "WEKA—experiences with a Java open-source project," *Journal of Machine Learning Research*, vol. 11, pp. 2533–2541, 2010.

Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2011, Article ID 210428, 12 pages
doi:10.1155/2011/210428

Research Article

A Finite Element Mesh Aggregating Approach to Multiple-Source Reconstruction in Bioluminescence Tomography

Jingjing Yu,[1,2] Fang Liu,[1,2] L. C. Jiao,[2] Shuyuan Yang,[2] and Xiaowei He[3]

[1] *School of Computer Science and Technology, Xidian University, Xi'an 710071, China*
[2] *Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an 710071, China*
[3] *School of Information Sciences and Technology, Northwest University, Xi'an 710069, China*

Correspondence should be addressed to Jingjing Yu, yujj@snnu.edu.cn and Fang Liu, lf204310@163.com

Received 24 June 2011; Revised 22 August 2011; Accepted 25 August 2011

Academic Editor: Shan Zhao

A finite element mesh aggregating approach is presented to reconstruct images of multiple internal bioluminescence sources. Rather than assuming independence between mesh nodes, the proposed reconstruction strategy exploits spatial structure of nodes and aggregation feature of density distribution on the finite element mesh to adaptively determine the number of sources and to improve the quality of reconstructed images. With the proposed strategy integrated in the regularization-based reconstruction process, reconstruction algorithms need no a priori knowledge of source number; even more importantly, they can automatically reconstruct multiple sources that differ greatly in density or power.

1. Introduction

Bioluminescence tomography (BLT) is a rapidly growing field of research in optical molecular imaging, which allows for the visualization of normal and abnormal cellular processes in living subjects at the molecular or genetic level [1–4]. With BLT, we seek to recover the spatial distribution of bioluminescent light source inside a small animal from external noninvasive measurements [5]. Generally speaking, the internal source intensity is closely related to the strength of the molecular/cellular activity, such as gene expression [6]. Thus, this imaging modality can provide in-depth information of the internal biological sources concerned in longitudinal monitoring and quantitative assessment changes and efficacy and thus further facilitates our understanding of bio-molecular processes as they occur in living animals.

When using BLT technique to measure efficiency of a genic therapy or to observe the growth or migration of cancer cells, accurate detection of different sources that differ greatly in density or power is instrumental; for example, it may yield a great deal of information regarding tumor dissemination and burden in various sites before the development of gross disease [1, 7, 8]. Therefore, the emphasis of this paper is multiple-source reconstruction that has not been sufficiently considered to date in BLT.

Most reconstruction methods for BLT can be classified to model-based reconstruction [9]. In this case, given a light propagation model, the flux on the boundary can be predicted with numerical methods such as the finite element method (FEM) by combing with the structural information and optical parameters regarding different organs. And then the BLT is formulated as an optimization problem of minimizing the discrepancy between the boundary measurements and the predicted light intensities on the tissue surface [10].

In the reconstruction procedure, the ill posedness of the BLT problem does pose a challenge for determining a unique solution of the tomographic problem. Different strategies have been proposed for coping with the ill posedness of BLT inverse problems. These studies obtain stable reconstruction by increasing the amount of independent measurements with spectrally resolved approaches [11–13], or by reducing the number of unknowns [10, 14], or with regularization techniques to incorporate some *a priori* information regarding the inverse source problem [15–17]. In this paper, we focus our attention on the multiple-source reconstruction

with monochromatic boundary measurements where regularization techniques are inevitable in the reconstruction process.

The existing regularization-based reconstruction schemes in bioluminescent imaging to date can be loosely classified into three categories: l_2 regularization, l_1 regularization, and implicit regularization such as TSVD and LSQR [18, 19]. Through regularization, some constraints are applied to reconstruction and yield an approximate solution of the BLT problem. No matter which regularizer is used, source location and visualization are still needed for preclinical practice. Most source location schemes are directly based on the reconstructed density vector and the larger the density, the more probable the source center. Specifically, according to a priori knowledge of the number of sources, several nodes with larger density values are identified as the promising sources or set a global threshold by referring to the maximum density and only those nodes with a density value higher than the threshold will be displayed.

In most applications of BLT, for example, monitoring cancer metastasis, neither the sources number nor an appropriate global threshold is easy to determine. This is mainly due to the fact that bioluminescent lights are usually weak and diffuse, and consequently the number of potential sources is hard to estimate only by surface photon distributions. Moreover, the global threshold strategy is unfeasible for distinguishing multiple sources with distinct difference in power. Especially in l_2 norm regularization cases, the obtained solution is usually oversmoothing, and thus a lower threshold will incur some artifacts in the final images whereas a higher one will discard some small potential sources. Consequently, effective reconstruction scheme for multiple sources with different powers deserves further investigation.

In this paper, we develop a finite element mesh aggregating approach for multiple-source reconstruction in BLT. The contribution of this paper to BLT reconstruction includes the following. First, we propose a multiple-source detecting strategy. Rather than assuming independence between mesh nodes, the proposed reconstruction strategy exploits spatial structure of the nodes and characteristic of energy decay to adaptively determine the number of sources and to improve the quality of reconstructed images. Second, we integrate the proposed reconstruction strategy with regularization-based inverse algorithms to build a unified framework for solving BLT inverse problem. Numerical simulations and phantom experiments demonstrate the effectiveness of this framework.

The paper is organized as follows. In Section 2, we present a multiple-source reconstruction framework with the emphasis on the finite-element-mesh-aggregating-based source detection strategy. In Section 3 we evaluate the proposed method with numerical simulations. Section 4 presents a phantom experiment to further test the effectiveness of the proposed method. Short discussions and concluding remarks are given at the end of this paper.

2. Multiple-Source Reconstruction Framework

2.1. FEM-Based Inverse Model. Radiative transfer equation (RTE) plays an important role in image reconstruction by predicting the bioluminescence light intensities on the tissue boundary [20], but solving RTE remains an intractable task for biological tissue with spatially nonuniform optical properties and complex tissue geometries [21]. Instead, some approximations to RTE have been established to overcome the difficulty of directly solving RTE. Among them, the diffusion approximation (DA) model has been extensively used to describe the photon propagation in tissue where there is scattering dominant absorption [5–14]. Here, we restrict our discussion to the DA model for simplicity. The steady state diffusion equation complemented with the Robin boundary condition can be expressed as follows [10]:

$$-\nabla \cdot (D(\mathbf{r})\nabla\Phi(\mathbf{r})) + \mu_a(\mathbf{r})\Phi(\mathbf{r}) = S(\mathbf{r}), \quad (\mathbf{r} \in \Omega), \tag{1}$$

$$\Phi(\mathbf{r}) + 2A(\mathbf{r}; n, n')D(\mathbf{r})(\nu(\mathbf{r}) \cdot \nabla\Phi(\mathbf{r})) = 0, \quad (\mathbf{r} \in \partial\Omega), \tag{2}$$

where $\Phi(\mathbf{r})$ is the photon power density at $\mathbf{r} \in \Omega, S(\mathbf{r})$ is an isotropic source distribution of gene expression, and $D(\mathbf{r})$ and $\mu_a(\mathbf{r})$ are the optical diffusion and absorption coefficient, respectively. In this work, we assumed these two parameters are constant during the BLT reconstruction procedure. The term $\nu(\mathbf{r})$ in (2) denotes the unit outer normal at boundary $\partial\Omega$, $A(r; n, n') \approx (1 + R(r))/(1 - R(r))$ is the boundary mismatch factor accounting for different refractive indices across the boundary $\partial\Omega$.

Following the standard finite element analysis [22], support domain Ω is discretized into T vertex nodes $(N_1, N_2, \ldots, N_T)$ and N_e mesh elements, denoted as Ω^l $(l = 1, 2, \ldots, N_e)$; then $\Phi(r)$ and source term $S(r)$ can be approximately expressed as

$$\begin{aligned} \Phi(r) \approx \Phi^h(r) &= \sum_{k=1}^{T} \phi_k \varphi_k(r), \quad \forall r \in \Omega, \\ S(r) \approx S^h(r) &= \sum_{k=1}^{T} s_k \gamma_k(r), \quad \forall r \in \Omega, \end{aligned} \tag{3}$$

where ϕ_k is the approximate nodal value of $\Phi(r)$ on the kth node $N_k, \varphi_k(r)$ the nodal basis function with support over the elements Ω^l, s_k the discretized nodal values of $S(r)$, and $\gamma_k(r)$ the interpolation basis functions, which is usually the same with $\varphi_k(r)$.

Based on (1)–(3), a matrix equation of the linear relationship between source distribution and boundary measurements can be derived [10, Section 2]:

$$AS = \Phi^*, \tag{4}$$

where A is a typical ill-conditioned matrix and Φ^* represents measurable boundary nodal photon density. In real BLT experiments, Φ^* is computed from the surface flux image captured with a CCD camera.

2.2. General l_p-Norm-Based Regularization. As mentioned in Section 1, the flux density on the boundary can be predicted according to a forward model, thereby a natural choice for source reconstruction is to minimize the misfit between predicted data and measurements, that is,

$$S = \arg\min_{S} ||AS - \Phi^*||^2. \tag{5}$$

To deal with the ill posedness of BLT inverse problem, permissible source region is usually incorporated into the reconstruction model by spatially constraining the reconstruction domain to the area of interest [10, 14, 16, 23]. A more effective approach to reconstruction is using regularization to act as an algebraic stabilizer in estimating solutions.

Using a general l_p $(0 < p \leqslant 2)$ norm constraint, we reformulate the objective function for BLT reconstruction in (5):

$$S_{\text{reg}} = \arg\min_{S} \left\{ ||AS - \Phi^*||_2^2 + \lambda \|S\|_p \right\}, \tag{6}$$

where the first term represents reconstruction error and the second is regularization term that fuses *a priori* knowledge or constrains into reconstruction. Regularization parameter $\lambda > 0$ provides a tradeoff between data fitting and constraints regarding solutions. Obviously, Tikhonov regularization method is a special case of (6) for $p = 2$, that is, using an l_2-norm regularizer. For $p = 1$, l_1-norm-based sparse regularization methods have recently attracted considerable amount of attention in BLT [17, 23–25] and the reconstructions results therein witnessed some improvements in image quality.

2.3. Multiple-Source Detection Strategy. Based on the solution (a source density vector) obtained in Section 2.2, source localization and imaging is then performed by combining with FEM mesh information. Facing the dilemma of threshold choice mentioned in Section 1, we are hoping for an adaptive method that can avoid the difficulty of threshold selection while at the same time removing artifacts in the reconstructed images with relatively lower computational cost.

Consider that in most applications of BLT, for example, detecting events that occur during the early stages of disease progression, the bioluminescent sources we want to recover are often localized in some small subregions of the domain. On the other hand, because light intensity is heavily attenuated in biological tissue and falls off exponentially from the illumination point, the diffusion range of a bioluminescent source is limited by the source strength. Consequently, when taking the spatial structure of the mesh nodes into account, the source density vector should have a spatial aggregation on the mesh, which is also illustrated in the experiments in Section 3 (Figure 4). It is found that, in a very small local region, if a node in the mesh has a maximum density value, with a very high probability its adjacent nodes are also with larger density. It is found that in a very small local region, if a node in the mesh has a maximum density value, with a very high probability its adjacent nodes are also with a larger density. We also observe that there are some nodes with smaller density in the vicinity of nodes with the larger density. These observations are helpful for discriminating pseudosource from a cluster of mesh nodes and removing artifacts in images. On the basis of the above analysis, an iterative multiple-source detection strategy (MSDS) is proposed in the following steps.

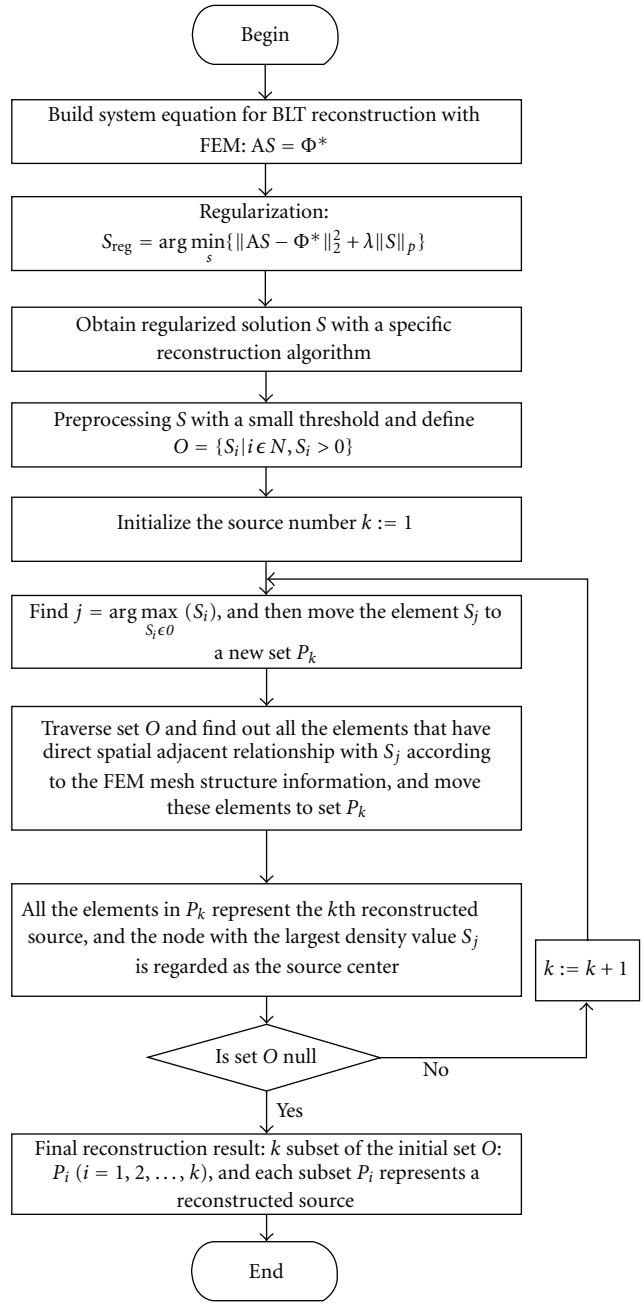


Figure 1: Flow chart of the regularization framework for multiple-source reconstruction.

Step 1. Obtain the regularized solution (the source density vector S).

Step 2. Threshold preprocessing. In the presence of inevitable noise, the solutions usually have many very small

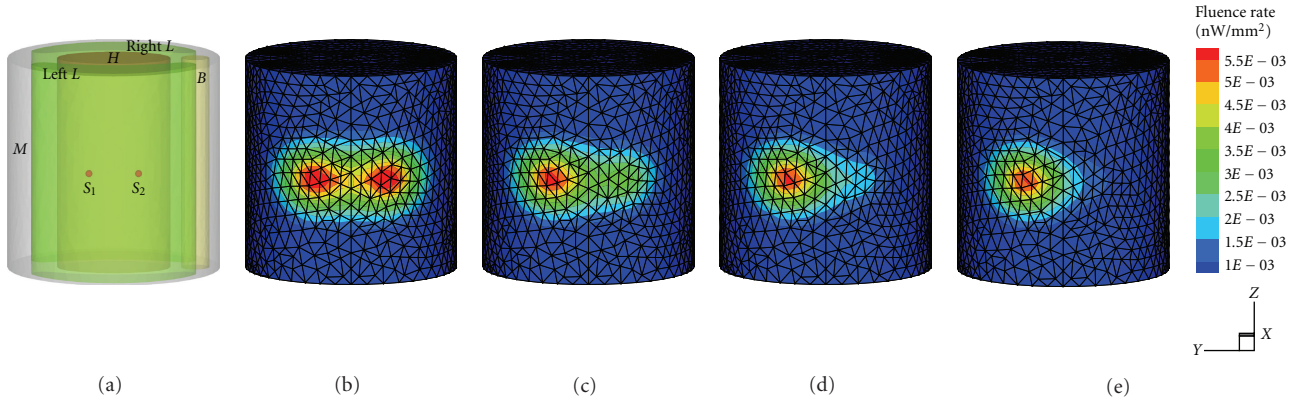


Figure 2: (a) 3D view of the heterogeneous phantom with two sphere sources in the left lung. (b)–(e) Different photon distributions generated, respectively, in power ratio of 1 : 1, 2 : 1, 4 : 1, and 8 : 1 cases.

nonzero components. Consequently, the preprocessing of solution with a small threshold of $c \max(S_i)$ is helpful to remove pseudosources and reduce the data size to be processed in the subsequent steps. For all the experiments in Section 3, the constant $c = 0.05$.

Step 3. Define a set $O = \{S_i \mid i \in N, S_i > 0\}$.

Step 4. Initial the sources number $k = 1$.

Step 5. Compute the node index $j = \arg\max_{S_i \in O}(S_i)$. We move the element S_j to a new set P_k. By traversing set O we can find out the other elements that directly adjoin the node j, if any, according to the mesh structure information. Remove these elements to P_k.

Step 6. If set O is null, stop; otherwise $k := k + 1$, and go to Step 5.

With the steps defined above, we provide an automatic method to estimate the number of sources from the reconstruction results iteratively. The final results contain k sources. Here, k is the number of subsets of the initial set O obtained at the end of the above iteration. Each subset corresponds to a reconstructed source. When P_i $(i = 1, L, k)$ has more than one member, we call this situation "overrepresentation," the nodes related to these elements will aggregate to represent a single source and the node with largest density value S_j is regarded as the source center for simplicity. Eventually, the cartesian coordinates of the reconstructed sources are obtained by their node index in the finite element mesh.

2.4. Regularization Framework for Multiple-Source Reconstruction. Based on the foregoing reconstruction scheme, we build a unified regularization framework for multiple-source reconstruction by integrating the MSDS with the general l_p-norm regularization, as shown in Figure 1.

An appealing property of this framework is its flexibility. The MSDS is a relatively independent component of the framework, and hence different regularizer and different reconstruction algorithms can be utilized according to the practice of BLT.

Table 1: Optical properties of different organs.

Material	Tissue	Lung	Heart	Bone
$\mu_a[\text{cm}^{-1}]$	0.07	0.23	0.11	0.01
$\mu_s'[\text{cm}^{-1}]$	10.31	20.00	10.96	0.60

3. Numerical Results and Analysis

In this section, we present some numerical experiments to demonstrate the utility and the effectiveness of the proposed method in multiple-source settings. Comparison is performed between the proposed MSDS and the traditional global threshold strategy (GTS). It should be pointed that the main theme of this paper is to evaluate the performance of this framework for multiple-source reconstruction in BLT, rather than the comparison between specific reconstruction algorithms. As representatives of algorithms using l_1 and l_2 regularization, Tikhonov regularization method [26] and $l1$–ls [27] are, respectively, combined with the above two strategies to recover the interior source distribution from the synthetically boundary measurements. Consequently, the reconstruction methods evaluated in the following experiments include Tikhonov + MSDS, Tikhonov + GTS, $l1$–ls + MSDS, and $l1$–ls + GTS.

It is known that regularization parameter is crucial to yield a good solution for ill-posed problems, and the choice of regularization parameter is usually nontrivial. In this paper, the regularization parameter for Tikhonov method was determined with the adaptive method proposed in [28]. As for $l1$–ls, the parameter λ was chosen as suggested in [27], that is, $\lambda = 0.1\|2A^T\Phi^*\|_\infty$.

All the experiments were performed on a cylindrical mouse chest numerical phantom as shown in Figure 2(a). The heterogeneous model is 30 mm in diameter and 30 mm high. The specific optical properties of different organs are listed in Table 1 [14].

3.1. Reconstruction for Double Sources with Different Powers. In the first study, we consider the ability to resolve sources with different powers. Two sphere sources with radius of 0.5 mm were positioned in the left lung with the centers at $S_1 = (-9, -3.5, 15)$ and $S_2 = (-9, 3.5, 15)$, respectively. They

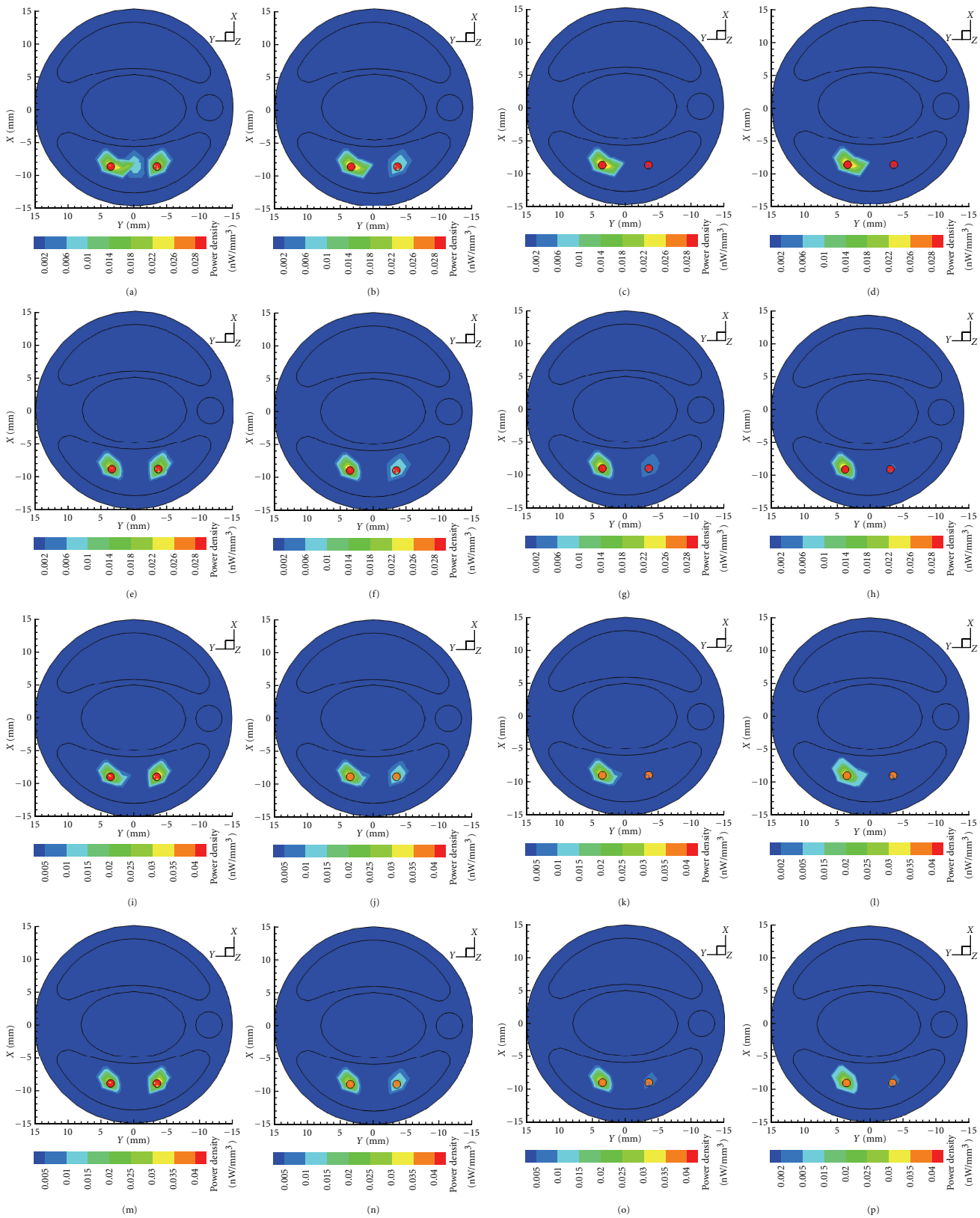


Figure 3: From left to right: transverse views of the reconstruction results at $z = 15$ mm in power ratio of 1 : 1, 2 : 1, 4 : 1, and 8 : 1. From top to bottom: final results of Tikhonov + GTS, Tikhonov + MSDS, $l1$–ls + GTS, and $l1$–ls + MSDS, respectively.

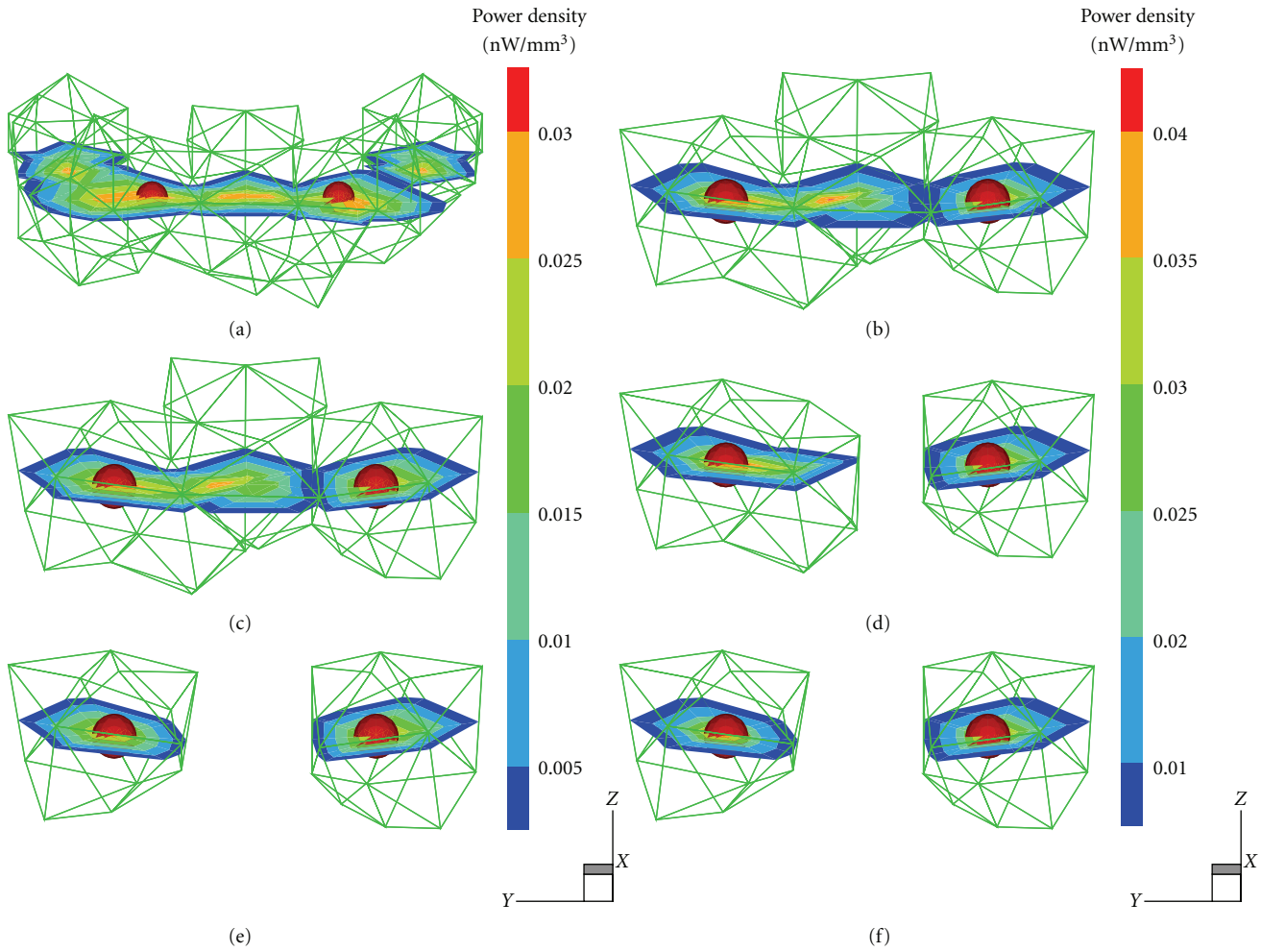


Figure 4: Top row: regularized solutions by Tikhonov regularization (left) and $l1$–ls method (right). Middle row: corresponding final reconstruction results by GTS with a threshold of 0.35 max (S_i). Bottom row: final reconstruction results by Tikhonov + MSDS (left) and $l1$–ls + MSDS (right).

were uniform in size and shape. To illustrate the point of our discussion, we consider four cases of experiment settings: (I) both of the initial source densities were 1 nW/mm^3; (II) to (IV) the densities of S_1 were still 1 nW/mm^3, but the densities of S_2 were 0.5 nW/mm^3, 0.25 nW/mm^3, and 0.125 nW/mm^3, respectively, that is, the ratios of the power of source S_2 to that of S_1 were 2 : 1, 4 : 1, and 8 : 1.

In the following experiments, the model was discretized into a fine tetrahedral element mesh and synthetic measurements were generated by solving the forward model with FEM. To simulate the noise involved in real BLT experiment, 10% Gaussian white noise was added to synthetic data. Figures 2(b)–2(e) show the forward mesh and the simulated photon distribution on the surface in the above four source settings. Obviously, it is difficult to predict the source number only according to the photon distribution especially in case (III) and case (IV).

In the reconstruction process, a permissible source region strategy was also employed as *a priori* information to decrease the ill posedness of BLT inverse problem, which was defined as $\{(x, y, z) \mid 8 < (x^2 + y^2)^{1/2} < 12, 13.5 < z < 16.5\}$ [14]. Following the proposed reconstruction framework the reconstructions were carried out with the aforementioned four methods under different source settings.

The first row and the third row of Figure 3 show the final reconstruction results by Tikhonov method and $l1$–ls method combined with the proposed MSDS. For comparison, the second row and the fourth row of Figure 3 present the corresponding reconstructed results rendered from GTS, where a global threshold (35% of the maximum density value) was used. It is obvious that the two sources are accurately detected by the proposed MSDS combined with different regularization methods in all the cases considered. On the other hand, for case (III) and case (IV), only the source with larger power is detected by Tikhonov + GTS and $l1$–ls + GTS, whereas the other weaker one is lost in the final reconstruction results.

To quantitativly assess reconstruction results in different power settings, we summarize location errors and reconstructed powers by different reconstruction schemes in Table 2, where the second column represents the actual initial power ratio of S_1 to S_2, and S_1^R and S_2^R denote

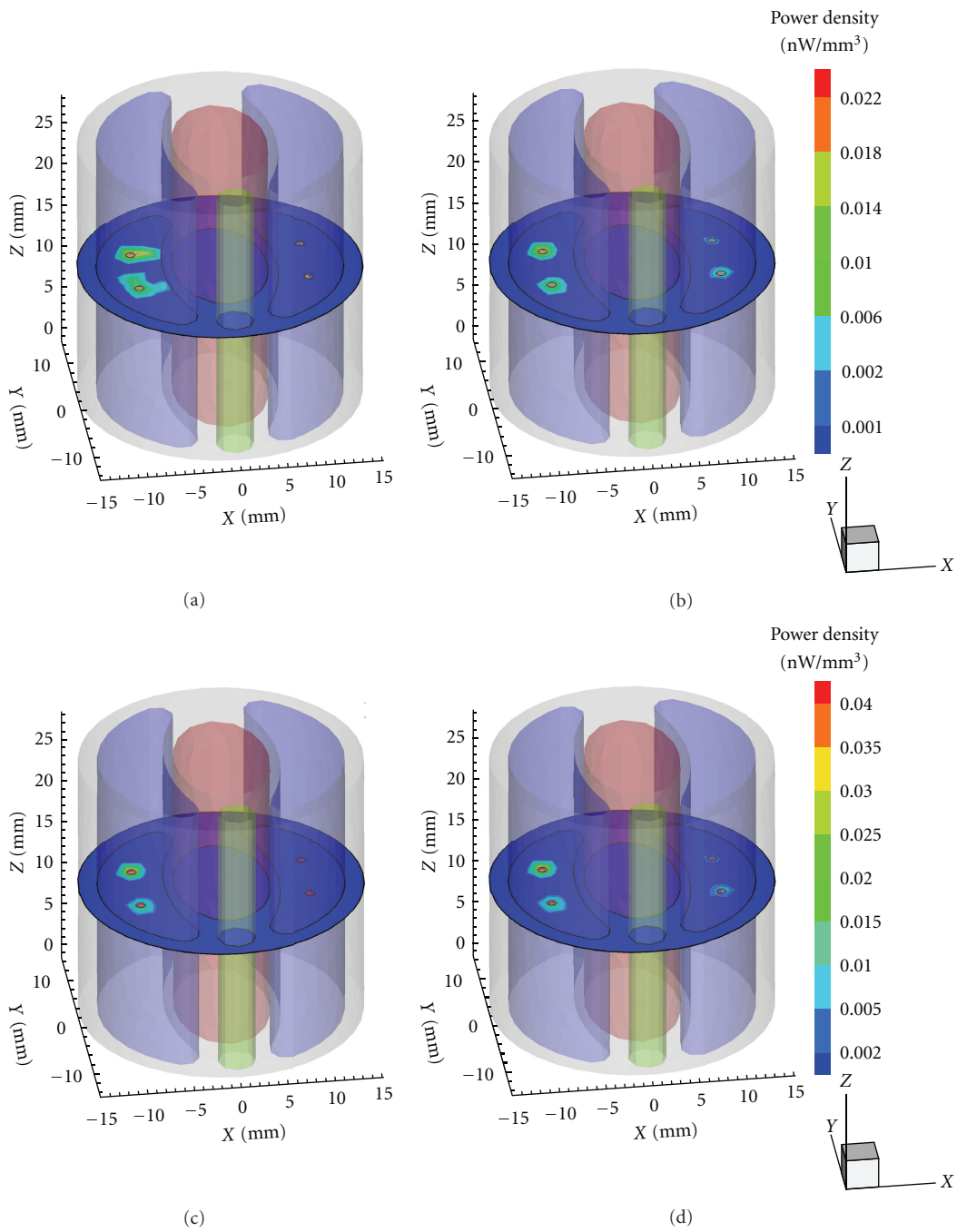


Figure 5: 3D views of reconstruction results with synthetic data generated from four scattered sources with different powers. (a)–(d) are the results of Tikhonov + GTS, $l1$–ls + GTS, Tikhonov + MSDS, and $l1$–ls + MSDS, respectively.

the corresponding reconstructed sources. N/A denotes that location information is not available.

From Table 2, it is seen that l_1-norm-based method $l1$–ls generally performs better than l_2-norm-based Tikhonov method in terms of reconstructed powers and locations.

Figure 4 illustrates the mesh aggregating process of MSDS and compares the final reconstruction results of MSDS with those of GTS in case (I). We can observe that there are some nodes with smaller density value in the vicinity of the two nodes with larger density, as shown in Figures 4(a) and 4(b). Apparently, retaining all of the nonzero components of the regularized solution will incur some artifacts in the final reconstruction image, in particular for l_2 norm solution by Tikhonov regularization method. The results in Figures 4(c) and 4(d) show that the traditional GTS directly discards those nodes with density value lower than the given threshold in the final results to improve the image quality. Usually, a higher threshold is preferred in the literature, thus a threshold of $0.35 \max(S_i)$ was used in the experiments for GTS method [16, 29]. As a result,

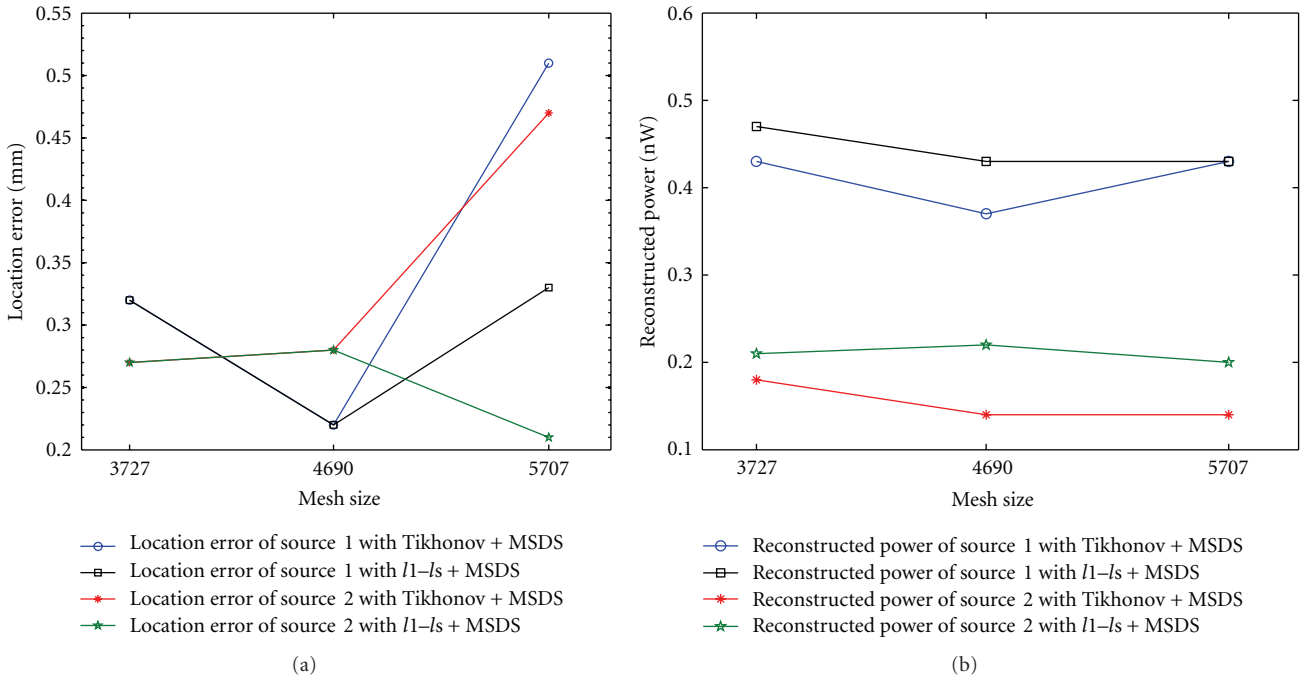


Figure 6: (a) Location error under different mesh levels. (b) Reconstructed power under different mesh levels.

Table 2: Reconstruction results in double-source case.

Case	Power ratio	Reconstruction method	Reconstructed center and location error (mm)				Reconstructed power (nW)	
			S_1^R		S_2^R		S_1^R	S_2^R
I	1 : 1	Tikhonov + GTS	−8.95, 2.13, 14.83	1.39	−8.98, −3.57, 14.73	0.28	0.65	0.28
		Tikhonov + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.35	0.28
		$l1$–ls + GTS	−8.99, 2.92, 14.77	0.62	−8.98, −3.57, 14.73	0.28	0.502	0.44
		$l1$–ls + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.41	0.44
II	2 : 1	Tikhonov + GTS	−9.03, 2.69, 14.65	0.88	−8.98, −3.57, 14.73	0.28	0.48	0.15
		Tikhonov + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.37	0.14
		$l1$–ls + GTS	−8.98, 2.98, 14.81	0.56	−8.98, −3.57, 14.73	0.28	0.51	0.22
		$l1$–ls + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.43	0.22
III	4 : 1	Tikhonov + GTS	−9.03, 2.72, 14.66	0.85	N/A	N/A	0.49	0
		Tikhonov + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.38	0.07
		$l1$–ls + GTS	−8.97, 3.00, 14.82	0.53	N/A	N/A	0.51	0
		$l1$–ls + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.4339	0.10
IV	8 : 1	Tikhonov + GTS	−9.03, 2.73, 14.67	0.84	N/A	N/A	0.49	0
		Tikhonov + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.38	0.03
		$l1$–ls + GTS	−8.97, 3.02, 14.83	0.51	N/A	N/A	0.51	0
		$l1$–ls + MSDS	−8.85, 3.59, 15.14	0.22	−8.98, −3.57, 14.73	0.28	0.43	0.04

those suspect targets with density lower than threshold will be omitted in this way. Unlike traditional methods, the proposed MSDS considers not only density value of a node but also mesh structure used in reconstruction and thus it has an ability to remove pseudosources and retain weak suspect sources in the final reconstruction results, as shown in Figures 4(e)-4(f) and 3.

3.2. Four-Source Reconstruction. In the second experiment, we attempt to reconstruct sources with synthetic data generated from four scattered sources with different initial powers, which may be a common case in tumor metastasis. Specifically, the power setup was according to ratio of 8 : 4 : 2 : 1 and the maximum power density was 1 nW/mm^3. Figure 5 shows 3D views of the results of Tikhonov regularization method and $l1$–ls method, respectively, combined with GTS and MSDS. The global threshold was the same as previous simulations. Obviously, it is hard for traditional GTS method to detect multiple sources with lower power density in such experimental setting, whereas the proposed MSDS accurately distinguishes all of the sources.

3.3. Influence of Finite Element Mesh. In view of the idea that the proposed multiple-source reconstruction approach utilizes underlying mesh structure information, it is necessary to assess the influence of different FEM discretization on the proposed method. Therefore, we conducted a set of double-source experiments under different discretization level. The results in Figure 6 (where the number of nodes in reconstruction domain denotes different discretization level or mesh size) show the influence of finite element mesh on reconstruction. For Tikhonov regularization method combined with MSDS, the location error increases slightly after a decrease along with the increasing of mesh size and the reconstructed power presents a similar variation trend. As for $l1$–ls combined with MSDS, both location error and reconstructed power vary slightly with mesh changes.

In general, finite element discretization does affect reconstructed results in the sense that the location error and the reconstructed power vary with the change of mesh. However, for all of the discretization levels considered, the proposed method is able to accurately localize and quantify light source distribution. These results demonstrate the robustness of the proposed reconstruction framework against mesh discretization.

4. Phantom Study

We further demonstrate the effectiveness of the proposed reconstruction algorithm with phantom experiments. This set of BLT experiments were conducted with a dual-modality BLT/micro-CT system [17, 30]. A backthinned, backilluminated cooled CCD camera is used to measure the signal on the phantom surface from four directions at 90-degree intervals.

The heterogeneous mouse chest phantom with 30 mm height and 15 mm diameter consists of four parts that represent muscle, lungs, heart, and bone, respectively [30]. The optical properties of different organs are listed in Table 1. Two small holes of diameter 2 mm were drilled in the phantom to place glass capillary with 1 mm inside diameter. Luminescent solutions of height 2 mm were extracted from a red luminescent light stick (Glow products, Canada) and then injected to glass capillary to serve as one testing source. The generated luminescent light had an emission peak wavelength of about 650 nm. The real center positions of the two testing sources were $(-9, 2, 16.6)$ and $(-9, -3, 16.6)$.

It is known that luminescent light intensity will decrease with the passage of time. We collected 100 gray level images of the sources, which were taken by the CCD camera every one minute. Figure 7 shows the fitted decay curve of light density. According to the decay curve, we can obtain sources with different intensities by controlling the injection time of luminescent solutions. Three groups of experiments were conducted, and the ratios of the intensity of source S_2 to that of S_1 were 1 : 1, 2 : 1, and 4 : 1, respectively. Figures 8(a)–8(c) show the front views of the corresponding measured data on CCD under different intensity settings. Subsequently, a permissible source region was roughly determined according to the surface flux density distribution, which is expressed as $\{(x, y, z) \mid 8 < (x^2 + y^2)^{1/2} < 13, 15 < z < 18\}$.

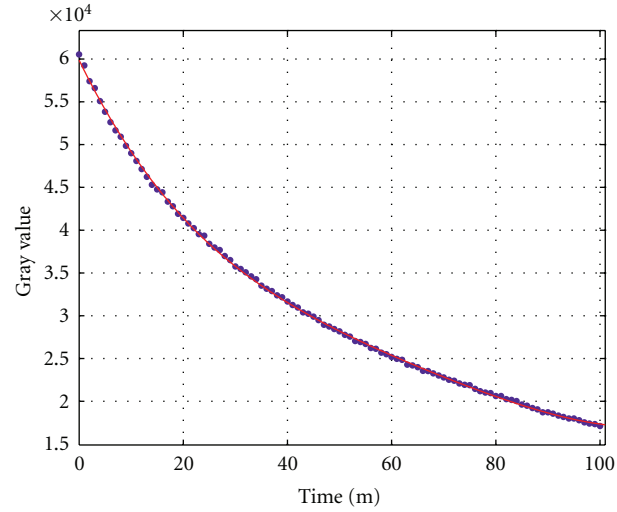


Figure 7: Decay curve of light density.

The phantom model was discretized into 4202 nodes and 21721 tetrahedra. After mapping the collected optical data on the three-dimensional phantom surface, we performed four rounds of reconstruction with Tikhonov + GTS, $l1$–ls + GTS, Tikhonov + MSDS, and $l1$–ls + MSDS under different source intensity settings. The normalized reconstruction results of Tikhonov regularization method are similar to that of $l1$–ls. To avoid interminable description, Figure 9 only presents comparison results between Tikhonov + GTS and Tikhonov + MSDS.

For all of the testing cases considered in phantom experiments, Tikhonov + MSDS and $l1$–ls + MSDS can accurately detect two sources, and the maximum location error is 1.7 mm. Even for the case of real intensity ratio 4 : 1, the reconstructed source strength ratios of them were 3.12 : 1 and 2.97 : 1. In stark contrast to the proposed methods, traditional global threshold methods failed to reconstruct the weaker of the two sources, as shown in Figure 9(c). Compared with the results of using GTS (the top row of Figure 9), the proposed MSDS methods produce fewer artifacts in the reconstructed images (the bottom row of Figure 9).

5. Discussions and Conclusion

Accurately reconstructing and distinguishing several sources with different intensities is a challenge problem in BLT, which is also an essential ability for serial observation of disease progression or response to therapy in the same animal over time. In this work, we present a unified framework for multiple-source reconstruction by integrating a novel multiple-source detection strategy with regularization-based reconstruction process. The effectiveness of this regularization framework is validated with numerical simulations and further confirmed with phantom experiments.

The advantage of this framework is twofold. First, there is no need for *a prior* knowledge regarding source number, which is automatically estimated from the reconstruction results iteratively. Second, the regularization framework is

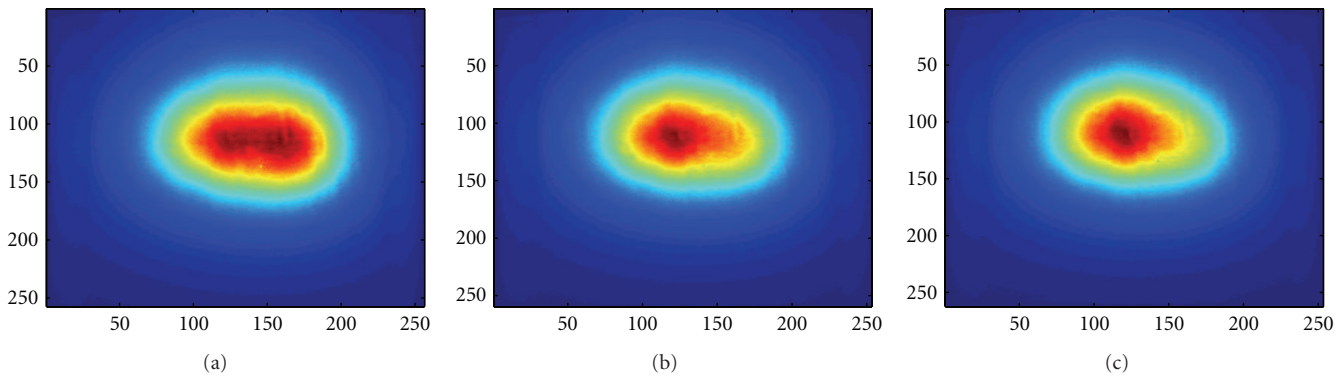


Figure 8: (a)–(c) Front views of measurements by CCD for the case of intensity ratios 1 : 1, 2 : 1, and 4 : 1.

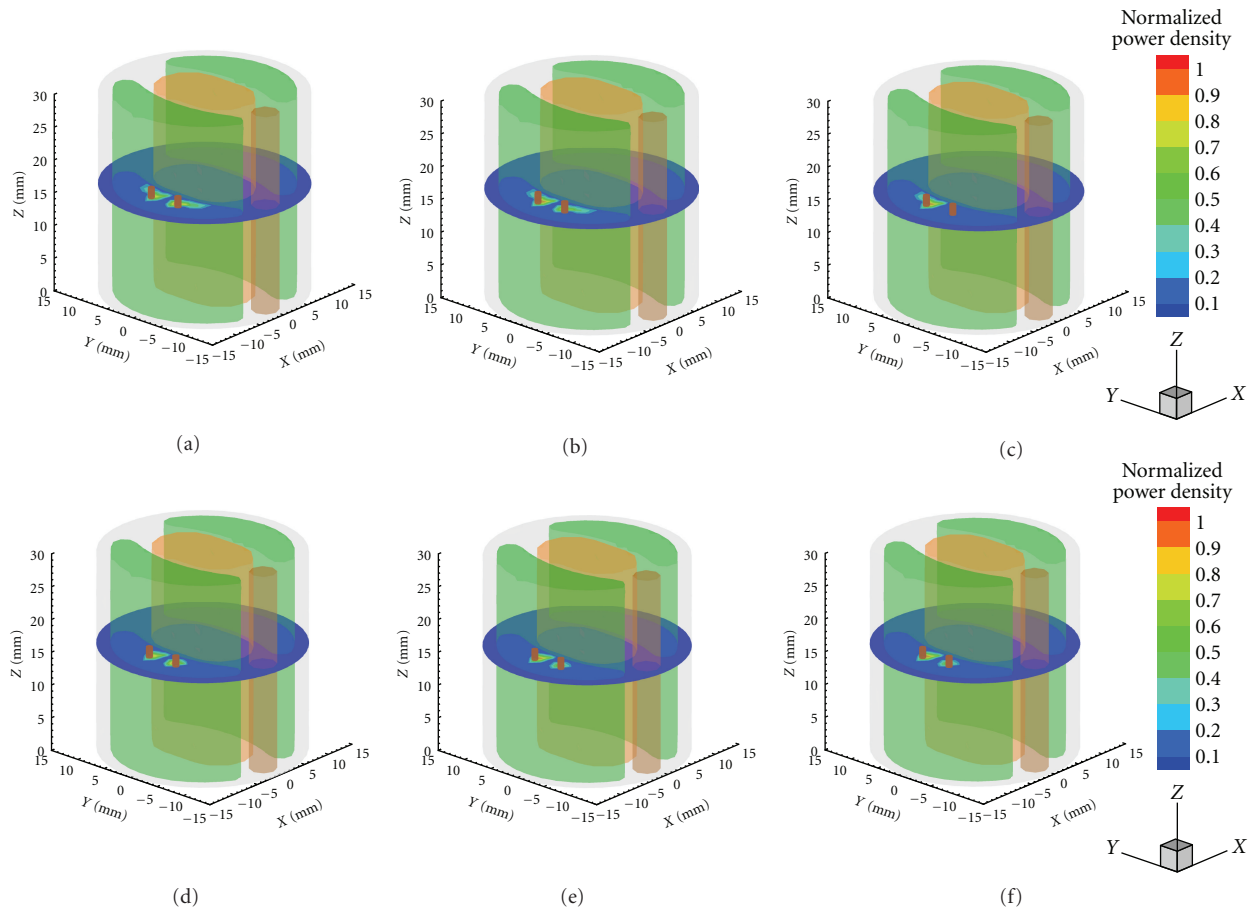


Figure 9: Normalized reconstruction results in phantom experiments. (a)–(c) are the results of Tikhonov + GTS with power ratio of 1 : 1, 2 : 1, and 4 : 1. (d)–(e) are the corresponding results of Tikhonov + MSDS.

general since it can work with different regularizers and inverse algorithms. The proposed MSDS is also easily applied to other finite-element-based reconstruction schemes to improve the final reconstruction results or image quality.

There are several limitations to the proposed method. As indicated in the experiment results, sparseness-inducing regularization method ($l1$–ls) performs better than l_2 norm method (Tikhonov). This is mainly because l_1 norm solution accords with the sparsity nature of bioluminescent source distribution in these applications. Consequently, how to select appropriate regularizer and inverse algorithm for specific BLT application is very important when using this framework.

Additionally, other regularizers can also be used in this unified framework. In fact, $l_p(0 < p < 1)$ norm regularized reconstruction has been tried for recovery of signals with weak sparsity in other image processing fields [31]. So far, related researches have not yet been reported in

BLT. Based on the proposed regularization framework, our future studies will investigate the effectiveness of other forms of regularizer for the ill-posed inverse problem of BLT.

Although only the DA model is considered for the sake of simplicity, the proposed BLT reconstruction framework has no limitation on the forward model. The performance of our framework might be improved by using more accurate forward models, which is also the direction of our further work.

Acknowledgments

The authors would like to thank Life Sciences Research Center in School of Life Sciences and Technology at Xidian University for providing experimental facilities. This work was supported in part by the National Natural Science Foundation of China (nos. 61173090, 61072108, 61072106, 60971112, 60971128, 60970067, 60970066, and 60972148), Beijing Municipal Natural Science Foundation (no. 7092020), the Fund for Foreign Scholars in University Research and Teaching Programs (no. B07048), the Shaanxi Provincial Natural Science Foundation Research Project under Grants no. 2011JQ1006 and 2011JQ8029, the Youth Foundation of Shaanxi Normal University under Grant no. 200901015, and the Fundamental Research Funds for the Central Universities (JY10000902001, K50510020001, and JY10000902045).

References

[1] R. S. Negrin and C. H. Contag, "In vivo imaging using bioluminescence: a tool for probing graft-versus-host disease," *Nature Reviews Immunology*, vol. 6, no. 6, pp. 484–490, 2006.

[2] J. Tian, J. Bai, X. P. Yan et al., "Multimodality molecular imaging: improving image quality," *IEEE Engineering in Medicine and Biology Magazine*, vol. 27, no. 5, pp. 48–57, 2008.

[3] V. Ntziachristos, J. Ripoll, L. V. Wang, and R. Weissleder, "Looking and listening to light: the evolution of whole-body photonic imaging," *Nature Biotechnology*, vol. 23, no. 3, pp. 313–320, 2005.

[4] J. K. Willmann, N. van Bruggen, L. M. Dinkelborg, and S. S. Gambhir, "Molecular imaging in drug development," *Nature Reviews Drug Discovery*, vol. 7, no. 7, pp. 591–607, 2008.

[5] G. Wang, W. Cong, H. Shen, X. Qian, M. Henry, and Y. Wang, "Overview of bioluminescence tomography-a new molecular imaging modality," *Frontiers in Bioscience*, vol. 13, no. 4, pp. 1281–1293, 2008.

[6] G. Wang, W. Cong, K. Durairaj et al., "In vivo mouse studies with bioluminescence tomography," *Optics Express*, vol. 14, no. 17, pp. 7801–7809, 2006.

[7] S. Li, Q. Zhang, and H. Jiang, "Two-dimensional bioluminescence tomography: numerical simulations and phantom experiments," *Applied Optics*, vol. 45, no. 14, pp. 3390–3394, 2006.

[8] A. Cong, W. Cong, Y. Lu, P. Santago, A. Chatziioannou, and G. Wang, "Differential evolution approach for regularized bioluminescence tomography," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 9, Article ID 5415660, pp. 2229–2238, 2010.

[9] X. Gu, Q. Zhang, L. Larcom, and H. Jiang, "Three-dimensional bioluminescence tomography with model-based reconstruction," *Optics Express*, vol. 12, no. 17, pp. 3996–4000, 2004.

[10] W. Cong, G. Wang, D. Kumar et al., "Practical reconstruction method for bioluminescence tomography," *Optics Express*, vol. 13, no. 18, pp. 6756–6771, 2005.

[11] H. Dehghani, S. C. Davis, S. Jiang, B. W. Pogue, K. D. Paulsen, and M. S. Patterson, "Spectrally resolved bioluminescence optical tomography," *Optics Letters*, vol. 31, no. 3, pp. 365–367, 2006.

[12] C. Kuo, O. Coquoz, T. L. Troy, H. Xu, and B. W. Rice, "Three-dimensional reconstruction of in vivo bioluminescent sources based on multispectral imaging," *Journal of Biomedical Optics*, vol. 12, no. 2, Article ID 024007, 2007.

[13] A. J. Chaudhari, F. Darvas, J. R. Bading et al., "Hyperspectral and multispectral bioluminescence optical tomography for small animal imaging," *Physics in Medicine and Biology*, vol. 50, no. 23, pp. 5421–5441, 2005.

[14] M. Jiang, T. Zhou, J. Cheng, W. Cong, and G. Wang, "Image reconstruction for bioluminescence tomography from partial measurement," *Optics Express*, vol. 15, no. 18, pp. 11095–11116, 2007.

[15] H. Gao, H. Zhao, W. Cong, and G. Wang, "Bioluminescence tomography with Gaussian prior," *Biomedical Optics Express*, vol. 1, p. 1259, 2010.

[16] Y. Lv, J. Tian, W. Cong et al., "A multilevel adaptive finite element algorithm for bioluminescence tomography," *Optics Express*, vol. 14, no. 18, pp. 8211–8223, 2006.

[17] X. He, J. Liang, X. Wang et al., "Sparse reconstruction for quantitative bioluminescence tomography based on the incomplete variables truncated conjugate gradient method," *Optics Express*, vol. 18, no. 24, pp. 24825–24841, 2010.

[18] P. C. Hansen, "The truncated SVD as a method for regularization," *BIT Numerical Mathematics*, vol. 27, no. 4, pp. 534–553, 1987.

[19] C. C. Paige and M. A. Saunders, "LSQR: an algorithm for sparse linear equations and sparse least squares," *ACM Transactions on Mathematical Software*, vol. 8, pp. 43–71, 1982.

[20] A. D. Klose and E. W. Larsen, "Light transport in biological tissue based on the simplified spherical harmonics equations," *Journal of Computational Physics*, vol. 220, no. 1, pp. 441–470, 2006.

[21] L. V. Wang and H. Wu, *Biomedical Optics: Principles and Imaging* , Wiley-Blackwell, 2007.

[22] S. R. Arridge and J. C. Hebden, "Optical imaging in medicine: II. Modelling and reconstruction," *Physics in Medicine and Biology*, vol. 42, no. 5, pp. 841–853, 1997.

[23] J. Yu, F. Liu, J. Wu, L. Jiao, and X. He, "Fast source reconstruction for bioluminescence tomography based on sparse regularization," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2583–2586, 2010.

[24] Y. Lu, X. Zhang, A. Douraghy et al., "Source reconstruction for spectrally-resolved bioluminescence tomography with sparse A priori information," *Optics Express*, vol. 17, no. 10, pp. 8062–8080, 2009.

[25] H. Gao and H. Zhao, "Multilevel bioluminescence tomography based on radiative transfer equation part 1: 11 regularization," *Optics Express*, vol. 18, no. 3, pp. 1854–1871, 2010.

[26] A. N. Tikhonov, V. I. A. Arsenin, and F. John, *Solutions of Ill-Posed Problems*, Winston, Washington, DC, USA, 1977.

[27] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[28] J. Yu, F. Liu, L. Jiao, and X. He, "Adaptive parameter selection for Tikhonov regularization in Bioluminescence tomography," in *Proceedings of the 3rd International Conference on BioMedical Engineering and Informatics*, pp. 86–90, October 2010.

[29] X. Song, D. Wang, N. Chen, J. Bai, and H. Wang, "Reconstruction for free-space fluorescence tomography using a novel hybrid adaptive finite element algorithm," *Optics Express*, vol. 15, no. 26, pp. 18300–18317, 2007.

[30] H. Huang, X. Qu, J. Liang et al., "A multi-phase level set framework for source reconstruction in bioluminescence tomography," *Journal of Computational Physics*, vol. 229, no. 13, pp. 5246–5256, 2010.

[31] J. Wu, F. Liu, L. C. Jiao, and X. Wang, "Compressive sensing SAR image reconstruction based on Bayesian framework and evolutionary computation," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1904–1911, 2011.