Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 513283, 20 pages
http://dx.doi.org/10.1155/2014/513283

*Research Article*

# Invariant Hough Random Ferns for Object Detection and Tracking

**Yimin Lin,[1,2] Naiguang Lu,[1,2] Xiaoping Lou,[2] Fang Zou,[3] Yanbin Yao,[3] and Zhaocai Du[3]**

[1] *Institute of Optical Communication & Optoelectronics, Beijing University of Posts & Telecommunications (BUPT), No. 10 Xitucheng Road, Haidian District, Beijing 100876, China*
[2] *School of Instrumentation Science & Optoelectronics Engineering, Beijing Information Science & Technology University (BISTU), No. 12 Qinghe Xiaoying East Road, Haidian District, Beijing 100192, China*
[3] *Beijing Aeronautical Manufacturing Technology Research Institute, Beijing 100024, China*

Correspondence should be addressed to Yimin Lin; linyimin2012@hotmail.com and Naiguang Lu; nglv2002@sina.com

Received 8 December 2013; Revised 6 February 2014; Accepted 13 February 2014; Published 8 April 2014

Academic Editor: Ilse C. Cervantes

This paper introduces an invariant Hough random ferns (IHRF) incorporating rotation and scale invariance into the local feature description, random ferns classifier training, and Hough voting stages. It is especially suited for object detection under changes in object appearance and scale, partial occlusions, and pose variations. The efficacy of this approach is validated through experiments on a large set of challenging benchmark datasets, and the results demonstrate that the proposed method outperforms state-of-the-art conventional methods such as bounding-box-based and part-based methods. Additionally, we also propose an efficient clustering scheme based on the local patches' appearance and their geometric relations that can provide pixel-accurate, top-down segmentations from IHRF back-projections. This refined segmentation can be used to improve the quality of online object tracking because it avoids the drifting problem. Thus, an online tracking framework based on IHRF, which is trained and updated in each frame to distinguish and segment the object from the background, is established. Finally, the experimental results on both object segmentation and long-term object tracking show that this method yields accurate and robust tracking performance in a variety of complex scenarios, especially in cases of severe occlusions and nonrigid deformations.

## 1. Introduction

Object detection and tracking have become central topics in computer vision research, and recent approaches have shown considerable progress under challenging situations such as changes in object appearance, scale, occlusions, and pose variations, [1, 2]. In this paper, we just focus on three themes, which are object representation, detection, and tracking, and propose a novel framework that can be used for part-based object detection in images and online object tracking through videos.

For feature description, the common local binary feature (LBF) [3] is computed in the image intensity or gradient domain yielding successful detection results for specific objects. However, the original LBF cannot be robust to rotation variations. Traditionally, the rotation problem has been addressed from a multiclass perspective by using classifiers repeatedly trained at different orientations [4]. Unfortunately, these methods suffer from two weaknesses. Firstly, the computational cost for both the training and detecting stages increases with the number of classifiers, and secondly, the use of multiple classifiers increases the number of false positives [5]. In this paper, our object representation is related to several paradigms, such as SIFT and SURF descriptor [6, 7]. They assigned one or more orientations to each patch based on image gradient directions. Thus the matching operations are simply performed on image data that has been transformed relative to the assigned orientation and achieve invariance to these transformations. We extend the same idea to LBF and propose to compute a LBF in the polar coordinates instead, since the coordinates of the descriptor are easy to rotate with a certain polar angle relative to the patch orientation. Therefore, our rotation invariant features have demonstrated remarkable results for object categorization under the challenging conditions, such as rotation variations and cluttered background.

Recently for part-based object detection, a popular way to address occlusions and nonrigid deformations is to combine the two ideas of appearance codebooks [8, 9] and generalized Hough transform [10, 11]. Such codebooks are used to classify the local appearance of interest points into scattered fragments of visual words that represent an object class [12]. The Hough transform was initially developed to detect analytically defined shapes, such as lines, circles, and ellipses. But up to now, the generalized Hough transform can be used to detect arbitrary shapes (i.e., shapes having no simple analytical form) [13]. It has been successfully adapted to the problem of part-based object detection since it is robust to partial occlusions and slightly deformed shapes [14]. Moreover it is tolerant to noise and can find multiple occurrences of a shape during the same processing pass. The main disadvantage is that it requires a lot of storage and extensive computation. Although it is inherently parallelizable, it has been reported that the Hough voting efficiency during the object categorization can be improved by a highly efficient classifier [15]. Therefore, in this paper, we apply the random ferns classifier (RFC) [3] to Hough transform for improving the search speed and reducing the need of a large space for data store. In addition, our Hough voting is performed in a rotation and scale invariant Hough space since each support point shares a stable polar angle and a scalable displacement related to the object's center.

Nowadays, visual object tracking has been formulated as online tracking by detection problem [16]. For the purpose of separating the target from background in individual frames, this method involves the continuous application of an object detection algorithm, where a target object is discriminated by a classifier. In order to handle the lack of prior knowledge and appearance changes, there is an essential need for an online learning algorithm incrementally updating the object template and retraining the classifier over time [17]. The most straightforward method, which replaces the template every frame with the image region believed to be the target, is found to suffer from gradual drift of the target out of the template, eventually resulting in the loss of the target. This phenomenon is referred to as template drifting problem [18], where template drift is due to the accumulation of small errors introduced in the location of the template each time when the template is updated. In this paper we propose a template segmentation algorithm based on a clustering scheme which groups the binary masks according to the appearances and geometric relations of object parts. Then we use back-projection to locate the hypothesis's support, which gives a rough localization of object parts. This support that has valid geometric relations guides intensity matching with the clustering patches. Therefore, the candidate template can be separated from the background pixelwise and still stays firmly attached to the original object.

In a word, the main contributions of this paper include the following.

(1) The rotation invariant LBF based on polar coordinates is shown to be invariant to image rotation. In particular, our object representation integrating with the intensity and gradient information is robust across a substantial range of rotation variations, addition of noise, and changes in illumination.

(2) The IHRF framework where Hough transform is combined with RFC provides an efficient way for object detection regardless of partial occlusions and nonrigid deformations. Specially, the Hough voting is robust to changing orientation and scale due to the stable polar angle and scalable displacement between the support point and object's center.

(3) A refined top-down segmentation algorithm based on a clustering scheme is proposed to guide a precisely segmentation process after the back-projection. The most important property of the clustering scheme is that all clusters are compact and only contain image patches that are visually similar since the clustering criterion relies on the geometric relations of object parts. As a result, this algorithm is able to separate the object precisely from a cluttered background.

(4) On the basis of IHRF, we present an online tracking approach that is able to prevent drifting problem because of the stable object detection and refined segmentation results, which enables more robust training of the classifier.

The rest of the paper is organized as follows. Firstly, Section 2 reviews the work related to Hough transform for object detection and online object tracking. Next, we introduce the IHRF framework in Section 3 and show how this could be applied for online tracking in Section 4. The experimental results including a comparison to state-of-the-art are given in Section 5. Finally, contributions and suggestions for future research are discussed in Section 6.

## 2. Related Work

The problem of object detection in images is known to be very challenging and needed to address several difficult issues such as large intraclass object variations, changes in object pose and illumination, cluttered backgrounds, and partial occlusions, [19]. Currently, there are two existing approaches which are sliding window [20–22] and part-based methods [23–25]. The latter one is more competent for solving these problems since many object categories are poorly represented by bounding boxes. Furthermore, it achieves excellent performance for occluded and deformable objects since object is represented as an assembly of local parts and flexible spatial relations between them [26, 27]. One specific subtype of this part-based detecting model is the implicit shape model (ISM) [28], which is a well-known approach based on the generalized Hough transform technique. During training, the ISM learns a model of the spatial occurrence distributions of local patches with respect to the object's center. During testing, this learned model is used to cast probabilistic votes to the location of object's center by the generalized Hough transform. Many modified approaches that are related to ISM have been proposed [29–31].

A drawback of such methods is that matching the patches with the codebooks during testing is computationally expensive due to the large number of codebooks. To overcome this, Gall et al. [1] proposed a Hough Forest for object detection that employs random forests to learn the patches in a supervised manner. Hough Forests have been shown to outperform the sliding window classifiers and are inherently capable of multiclass detection. These advantages inspired a series of applications and extensions [32, 33]. However, so far Hough Forests also have some limitations. For instance, the Hough voting step implies considerable computational effort since the computational complexity of matching a patch against a tree is logarithmic to the number of leaves. In addition, they do not include the top-down segmentation capabilities that were available in their ISM predecessor [26].

Visual object tracking, whose goal is to estimate the states (positions or regions) of a target corresponding from one frame to the next, is one of the most important issues in computer vision. It is used in a wide range of applications including automated security and surveillance, human computer interaction, augmented reality, traffic control, and vision navigation, [34, 35]. Although visual object tracking has been studied for several decades and much progress has been made in recent years, it remains a very challenging problem due to a variety of factors that affect the performance of a tracking algorithm, such as the loss of information caused by the projection of the 3D world on a 2D image, noise in images, background clutters, illumination and scale variations, partial or full occlusions, complex object motion, camera motion, and real-time processing requirements. Nowadays there exists no single tracking approach that can successfully handle all of the above scenarios [36].

Recently, visual object tracking has been formulated as an online tracking by detection problem [16]. Although state-of-the-art online approaches as [37–39] perform well in certain scenarios, the update of the error appearance degrades the model and can lead to a significant drift. Therefore several novel algorithms, such as multiple instance learning [40] or the combination of tracking and detection [41], became very robust against the drifting problem. As the author suggested, a part-based model could potentially reduce the amount of drift by better aligning the tracker location with the object [42]. In addition, inspired by works such as [9, 26, 43], Godec et al. [15] proposed an online tracker based on the GrabCut segmentation algorithm [44] to find the object boundaries. It delivers a more precise description of the object and avoids the drifting problem to some extent, while it often leads to poor segmentations of objects since GrabCut algorithm sometimes splits them into multiple regions or merges them with parts of the background.

## 3. Invariant Hough Random Ferns

Random fern descriptors, also called LBF, consist of some logical pairwise comparisons of the intensity or gradient levels of randomly selected pixels in the input images [45]. However, such comparisons are not robust to rotation and scale variations because every pairwise pixel is randomly generated offline while remaining fixed in runtime. Under these variations, we propose a rotation and scale invariant descriptor that demonstrates a high degree of stability, which we show in Sections 3.1 and 3.2. In addition, our proposed IHRF consists of random ferns that are trained to learn a mapping from a densely sampled rotation and scale invariant LBF to their corresponding votes in a Hough space (see Section 3.3). In Section 3.4, we will also show how the back-projected hypothesis's support can be used to infer a pixelwise figure-ground segmentation of the object by using a clustering scheme.

Figure 1 illustrates the procedure for object detection and segmentation based on our IHRF. We formulate the rotation invariant and multiscale object detection problem as a probabilistic Hough voting procedure. For this example, the IHRF is trained on 328 horse images and masks are obtained from the Weizmann Horse database [46]. When presented with the test image, the system extracts 100 interest points (Figure 1(a)) within the horse masks to generate the dense scanning windows (Figure 1(b)). Those local patches then cast probabilistic votes containing object centroid locations, which are collected in the voting space (Figure 1(c)). As the visualization of this space in Figure 1(d) shows, the system searches for local maxima in the voting space and returns the correct detection as the strongest hypothesis. By back-projecting the contributing votes, we retrieve the hypothesis's support in the image (Figure 1(g)) and roughly separate the object from the background. To yield precise segmentation, we cluster the local masks according to their spatial and visual similarity, as shown in Figures 1(e) and 1(f). After matching the clusters, we deliver a more precise description (Figure 1(i)) of the object based on the masks (Figure 1(h)). All the key steps are described in detail in the following sections.

### 3.1. Rotation Invariant Local Binary Feature.

Rotation invariant descriptors are useful when objects of the same class can appear in different poses. The original LBF formulation can consist of randomly selected pairwise comparisons of the image values, such as the intensity, color, and gradient of the input images. To make the illumination and intraclass variations more robust, refer to [47], we use the following 16 feature channels: 3 color channels of the Lab color space, 4 absolute values of the first- and second-order derivatives in the $x$ and $y$ directions, and 9-bin histogram of gradients as feature channels.

More specifically, suppose $I_n(x, y)$ is the $n$th ($n = 1, \ldots, 16$) feature channel obtained from an image patch centered at pixel locations $x$ and $y$. Each fern applies a series of binary tests to the pairwise pixels as follows:

$$f(i, j, n) = \begin{cases} 1, & I_n(x_i, y_i) > I_n(x_j, y_j), \\ 0, & I_n(x_i, y_i) \le I_n(x_j, y_j), \end{cases} \quad (1)$$

where $(x_i, y_i)$ and $(x_j, y_j)$ are random pairwise pixels locations and each comparison returns 0 or 1. $n$ is also randomly chosen. Generally, the number of selected pairwise pixels $S$ maps an image patch to a $2^S$-dimensional space of binary

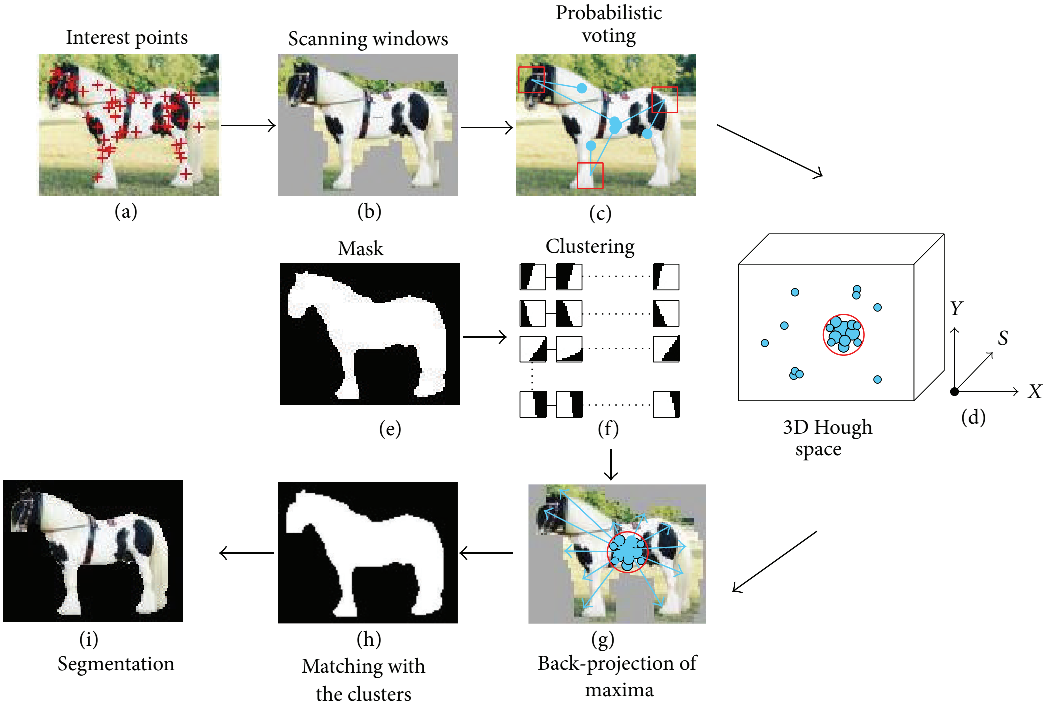


Figure 1: The part-based object detection and segmentation procedure.

descriptors in each fern. According to (1), we use the abbreviated form $f$ for $f(i, j, n)$ and LBF can be computed as

$$F_m = \{f_1, f_2, \ldots, f_S\}, \qquad (2)$$

where $F_m$ is the $m$th fern and $f_i$ means the $i$th binary feature (as (1)). Note that $m = 1, \ldots, K$, where $K$ is the number of ferns. Therefore, the entire set of random ferns can be denoted by $F = \{F_1, F_2, \ldots, F_K\}$. A trade-off between performance and memory can be made by changing the number of ferns $K$ and their sizes $S$. For example, $K = 1$ and $S = 4$; we suppose the feature $F = \{1101\}$. If $K = 2$ and $S = 4$, we get the feature $F = \{0101, 1010\}$.

Because the original LBF is not robust to rotation variations, here we present a novel rotation invariant descriptor for detecting objects in a specific category that may appear in images under different rotations. The differences are shown in Figure 2.

Figures 2(a) and 2(c) are the original image and Figures 2(b) and 2(d) are the result of a 90-degree in-plane rotation. Four random pairwise pixels are connected by the red lines shown in Figure 2. Thus, the LBF in Figure 2(a) is 1010, which is obviously different from Figure 2(b)'s LBF (1100). The reason is that the intensity distributions changed due to the rotation variations and fixed pairwise pixels, as shown in Figures 2(a) and 2(b).

Inspired by SIFT and SURF descriptors [6, 7], we propose a rotation invariant local binary feature (RILBF) based on the maximum gradient orientation (MGO) of the local image region. Therefore, an orientation histogram is formed from the gradient orientations of sample points within the region. For instance, the histograms of gradient (HoG) [21] for the original and rotational images are calculated as shown in Figure 3. Note that the orientation histogram has 72 bins covering the 360-degree range of orientations.

As can be clearly seen from Figure 3, peaks in the orientation histogram correspond to the dominant directions of the local gradients. Therefore, the MGO of the original and rotational images are 270° and 0°, respectively. The pairwise pixels can be changed according to the MGO, which is shown as the cyan arrows originating from the center of the circle in Figures 2(c) and 2(d). Considering the MGO, it is possible to precisely predict where each pairwise pixel in an original image should appear in the transformed image. To correctly measure repeatability and positional accuracy, we define the random pairwise pixels locations in a polar coordinate system, as shown in Figure 4.

Generally, the polar coordinate system is a 2D coordinate system where each point on a plane is determined by a distance $R$ from a fixed pole and a polar angle $\theta$ from a fixed polar axis. The polar coordinates $R$ and $\theta$ can be converted to the Cartesian coordinates $x$ and $y$ by using the trigonometric functions sine and cosine:

$$\begin{aligned} x &= R\cos\theta, \\ y &= R\sin\theta. \end{aligned} \qquad (3)$$

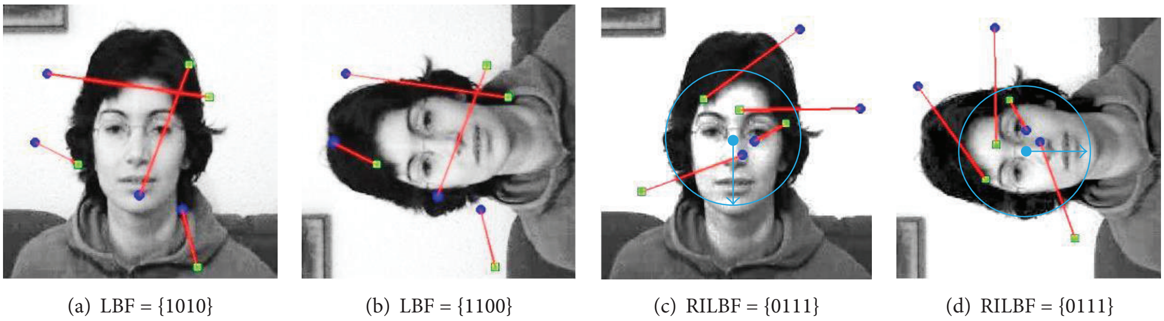

(a) LBF = {1010}
(b) LBF = {1100}
(c) RILBF = {0111}
(d) RILBF = {0111}

Figure 2: The results of LBF and RILBF on rotated images. (a) and (b) show the unequal LBF and (c) and (d) are the equal RILBF, where the orientations are indicated by the arrow from the center of the circle.

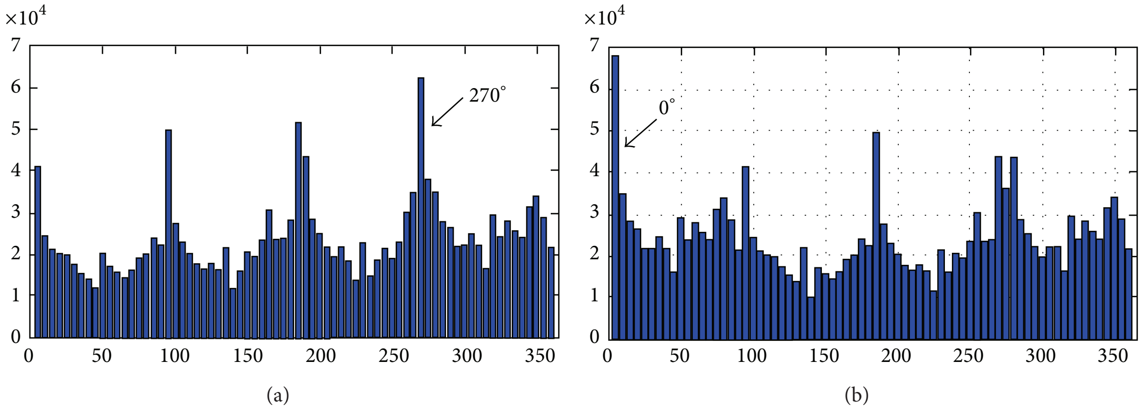


Figure 3: The HoG of the original and rotated images. (a) MGO = 270° and (b) MGO = 0°.

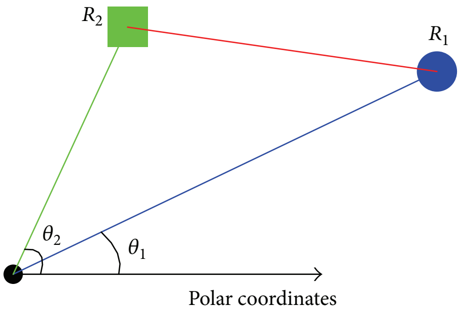


Figure 4: The polar coordinate system for pairwise pixels.

To achieve orientation invariance, the pairwise pixels used in (1) can be calculated by the polar coordinates, where gradient orientations are rotated relative to the MGO = $\theta_m$:

$$\begin{aligned} x_1 &= R_1 \cos(\theta_1 + \theta_m), \\ y_1 &= R_1 \sin(\theta_1 + \theta_m), \\ x_2 &= R_2 \cos(\theta_2 + \theta_m), \\ y_2 &= R_2 \sin(\theta_2 + \theta_m). \end{aligned} \quad (4)$$

Note that the fixed pole is located at the center of image and the fixed polar axis has the same direction as MGO, as the cyan arrows shown in Figures 2(c) and 2(d). This allows pairwise pixels to be matched correctly under arbitrary orientation change between the two images. For example, both Figures 2(c) and 2(d) result in the same representation where the RILBF $F(\theta_{m1})$ and $F(\theta_{m2})$ (here $\theta_{m1}$ and $\theta_{m2}$ are the MGO) are always equal to 0111.

Therefore, by assigning a consistent orientation to each LBF based on local image properties, the RILBF can be represented simply relative to this orientation and therefore achieve invariance to image rotation. Furthermore, by reserving typical features and reducing redundancy features, the generalization performance and training efficiency of the classifier are guaranteed.

*3.2. Scale Invariant Scanning Grid Pyramid.* To detect the position of the object, the detector scans the input image by a scanning window and, for each patch, determines the presence or absence of the object. The scanning window needs to be resized at different scales because the search often involves comparing objects that have been resized. To handle scale variations in RILBF, we maintain the same polar angle of the pairwise pixels but apply different scale ratios to radius $R$. Therefore, refer to [7]; the scale space is analyzed by changing the scanning window size rather than iteratively reducing the image size, as in the pyramid structure shown in Figure 5.

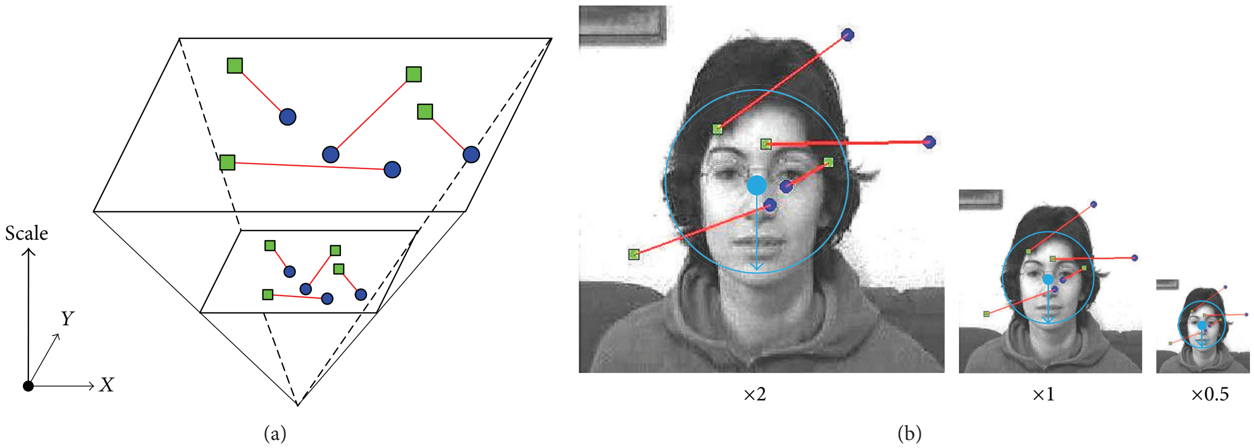


Figure 5: The pyramid structure of the scanning window for scale variations. (a) The pyramid local binary features. (b) The pyramid images for RILBF.

As shown in Figure 5, the distributions of the pairwise pixels are changed according to the different scale ratios but are preserved as the same profile. In this paper, the square region of the scanning window is 16 × 16 pixels. This is considered the initial scale layer, which we will refer to as scale $s = 1$. The following layers are obtained by gradually changing the radius $R$ according to the scales $s = 1.2^n$, where $n$ is the index range $[-2, 4]$. We generate all possible shifts of an initial bounding box with the following parameters: horizontal step = 10% of width, vertical step = 10% of height. Therefore, all these scanning windows in different scale ratios make up a scanning grid pyramid, as presented in [41]. Note that, as we do not have to downsample the image, no aliasing occurs as in Gall's work [1].

### 3.3. Part-Based Object Detection on the Hough Space.

*3.3. Part-Based Object Detection on the Hough Space.* Our IHRF consists of a set of random ferns that are trained to learn a mapping from densely sampled RILBF to their corresponding votes in a Hough space. The Hough space encodes the hypothesis $h$ for a part-based object centroid position in different scale space.

Let $\Gamma$ denote the mapping from the input appearance $I(x, y)$ of the local image patch $P_Y$, which is represented by a RILBF $F$ centered at $Y$ and where the MGO is $\theta_{md}$ (it is obtained during object detection), to the probabilistic Hough vote for the hypothesis $h$:

$$\bigcup_{x_i, y_i \in P_Y, \theta_{md}} I(x_i, y_i) \xrightarrow{\Gamma} p\left(\frac{h}{F}, \theta_{md}, Y\right). \tag{5}$$

Learning the mapping $\Gamma$ and using it for part-based object detection are described in Sections 3.3.1 and 3.3.2, respectively.

*3.3.1. Training the Random Ferns Classifier.* Random ferns are of great interest in the computer vision domain because of their speed, parallelization characteristics, and robustness to noisy training data. They are used for various tasks, such as key-point recognition [3] and image classification [48]. Özuysal et al. [3] argued that the vital element of RFC is the independence of the base ferns, which can be enforced by generating different pixel comparisons from the same image patch, which has been presented in Section 3.1. When they are applied to a large number of input vectors of the same class $C$, the output of each fern is a frequency distribution histogram, which is shown in Figure 6. In the histogram, the horizontal axis represents a $2^S$-dimensional space of binary descriptors, and the vertical axis shows the number of times the binary code appeared in a class $C$, also called class conditional probability (CCP) $p(F_i/C)$, where $i \in [1, K]$.

Random ferns replace the trees in random forests [49] by nonhierarchical ferns and pool their answers in a naive Bayesian manner to yield better results and improve classification rates in terms of the number of classes. As discussed in Section 3.1, the set of RILBF $(F, \theta_{mt})$ located in a local patch with MGO $\theta_{mt}$ (it is obtained during classifier training) is regarded as a class. Thus, a randomly selected patch detected in another image will be assigned to the most likely class $\widehat{C}$ by evaluating the posterior probability:

$$\widehat{C} = \underset{k}{\operatorname{argmax}}\, p\left(\frac{c_k}{F}, \theta_{mt}\right), \tag{6}$$

where $k = 1; 2; \ldots; H$, $c_k$ is the set of classes. and $c_k \in C$. According to seminaive Bayes [50] and (2), (6) is equivalent to a joint CCP for binary representations in each fern as

$$\widehat{C} = \underset{k}{\operatorname{argmax}} \prod_{L=1}^{K} p\left(\frac{(F_L, \theta_{mt})}{c_k}\right). \tag{7}$$

For a given test input, simply apply the binary representations accounting for the ferns and look up the corresponding probability distribution over class label, as shown in Figure 6. Finally, the RFC selects the class with the highest posterior probability as the categorized results. RFC is a remarkable classification algorithm that randomly selects and trains a collection of ferns. Then, classifying new inputs involves only simple look-up operations.

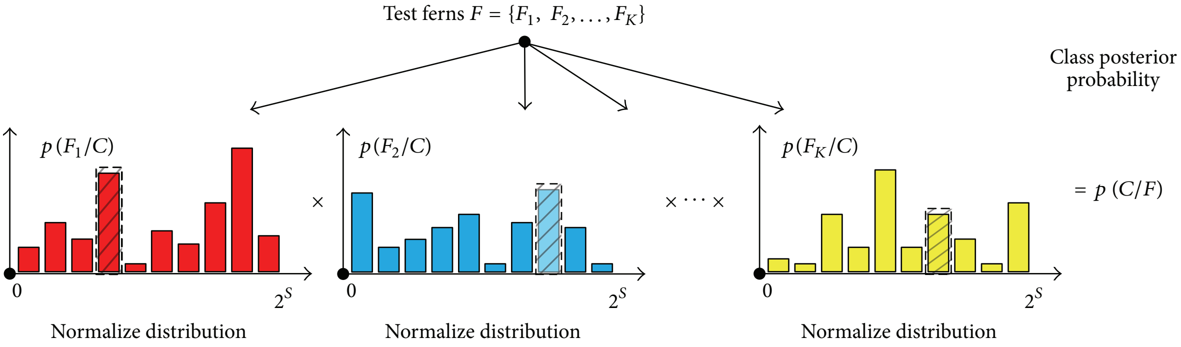


FIGURE 6: The classification using a RFC, where × is the symbol of multiplication.

For training, a set of training examples is provided for each class $C$. For each positive class, we assume that an object center within a bounding box including the whole object is provided, and we randomly divide it into several local patches. Then, we need to record the mutual geometric constraints between them. As the green rectangle in Figure 7 shows, the displacement vector $d$ from the object center to the center of a local patch in its polar coordinate system is recorded as

$$d = \{R_C, \theta_C\}, \tag{8}$$

where $R_C$ is the displacement between the two centers and $\theta_C$ is the orientation, which is rotated relative to the MGO $\theta_{mt}$ of the local patch. We refer to the ISM representation [9] as

$$\mathrm{ISM}(c_o) = \left(p\left(\frac{F}{c_o}\right), \theta_{mt}, d\right). \tag{9}$$

As a result, each fern in the IHRF consists of the ISM of each local patch belonging to the object class $c_o$. It is noted that the size of an object used for training can be represented by a scale factor $s = 1$. For the negative instances, we simply record their own class labels and the pseudodisplacements.

This allows the classifier to exploit the available training data more efficiently because image patches representing the same object but in a different configuration (i.e., rotated or scaled) can be considered the same types of information. During classification, we do not need to classify multiple scaled and rotated versions of the image, and it results in a lower complexity.

*3.3.2. Probabilistic Voting on Hough Space.* In the following, we cast the voting procedure into a probabilistic framework. For detection in a scanning patch, the extracted RILBF is passed through every fern in the random ferns, and then the potential object centroid positions stored in the ferns are used to cast votes to the Hough space. According to this, the posterior probability (5) can be simply decomposed as follows:

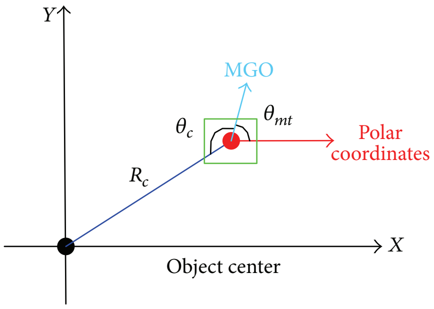


FIGURE 7: Displacement vector, where the green rectangle is the positive instance. The polar axis of the local patch has the same direction as the image $x$.

$$\begin{aligned} p\left(\frac{h}{F}, \theta_{md}, Y\right) &= p\left(\frac{h(c, X, S_V)}{F}, \theta_{md}, Y\right) \\ &= \sum_{k=1}^{H} p\left(\frac{h(c, X, S_V)}{c_k}, \theta_{md}, F, Y\right) p\left(\frac{c_k}{F}, \theta_{md}, Y\right), \end{aligned} \tag{10}$$

where $h(c, X, S_V)$ is the hypothesis for the object belonging to the class $c \in C$ with position $X$ and scale factor $S_V$. As shown in (6), we have evaluated the patch's probability independent of their location. In addition, the first term in (10) can be treated as independent of RILBF $F$ because we have mapped the unknown patch to a hypothesis $h$. Thus, (10) can be reduced to

$$\begin{aligned} p\left(\frac{h}{F}, \theta_{md}, Y\right) &= \sum_{k=1}^{H} p\left(\frac{h(c, X, S_V)}{c_k}, \theta_{md}, Y\right) p\left(\frac{c_k}{F}, \theta_{md}\right) \\ &= p\left(\frac{h(c, X, S_V)}{c_k} = c, \theta_{md}, Y\right) p\left(\frac{c_k}{F} = c, \theta_{md}\right), \end{aligned} \tag{11}$$

where the first term is the probabilistic Hough vote for an object position $X$ in different scale space $S_V$, which is based on the class label and its ISM. More specifically, the first term votes for the following object position as follows:

$$X_x = Y_x - S_V X_d, \qquad X_y = Y_y - S_V Y_d, \tag{12}$$

where the subscripts $x$ and $y$ indicate the image position in the $x$ and $y$ directions, respectively. According to the displacement vector $d$, which has been stored in the ISM, the connection vector $(X_d, Y_d)$, which is relative to the current position $(Y_x, Y_y)$, can be presented as

$$X_d = R_C \cos(\theta_C + \theta_{md}), \qquad Y_d = R_C \sin(\theta_C + \theta_{md}). \tag{13}$$

When casting votes for the object center $(X_x, Y_y)$, the object scale ratios $S_V$ are selected according to the description in Section 3.2 and are treated as a third dimension in the Hough voting space. Therefore, the distribution of the first term in (11) can be approximated by a sum of Dirac functions $\delta_d$ for all the displacement vector set $D = \{d_i\}_{i=1,2,\ldots,N}$ ($N$ is the number of displacement vectors which are obtained after training) as follows:

$$\begin{aligned} p&\left(\frac{h(c, X, S_V)}{c_k} = c, \theta_{md}, Y\right) \\ &= \frac{1}{N}\sum_{d\in D}\delta_d\left(\frac{Y_x - X_x}{S_V} - X_d, \frac{Y_y - X_y}{S_V} - Y_d\right). \end{aligned} \tag{14}$$

Thus, the vote distribution in (14) is obtained by casting a vote for each stored observation from the learned ISM. For all the ferns, the second term in (11), which is calculated by (7), is averaged as follows:

$$p\left(c_k = \frac{c}{F}, \theta_{md}\right) = \frac{1}{K}\sum_{L=1}^{K} p\left(\frac{(F_L, \theta_{md})}{c_k} = c\right). \tag{15}$$

Note that the accumulation of the probabilities in (15) is nonprobabilistic, but the results of summation are preferred over multiplication in (7) because this approach is more stable in practice [1]. To integrate the votes coming from the scanning grid pyramid of the input image $\Omega$, we accumulate them into the Hough image $H$:

$$\begin{aligned} HI &= p\left(\frac{h}{\Omega}\right) = \sum_{Y\in\Omega} p\left(\frac{h}{F}, \theta_{md}, Y\right) \\ &= \frac{1}{KN}\sum_{Y\in\Omega}\sum_{L=1}^{K}\sum_{d\in D} p\left(\frac{(F_L, \theta_{md})}{c_k} = c\right) \\ &\quad\times \delta_d\left(\frac{Y_x - X_x}{S_V} - X_d, \frac{Y_y - X_y}{S_V} - Y_d\right). \end{aligned} \tag{16}$$

As a result, the value $p(h/\Omega)$ serves as a confidence measure for the hypothesis $h$. After all the votes are cast, a global search for the local maxima obtains the position of the object center as a nonparametric probability density estimate.

*3.4. Object Segmentation.* By back-projecting the contributing votes, we retrieve the hypothesis's support in the image, which shows the rough profile on the depicted object. However, this is not a precise segmentation yet. Therefore, we propose a segmentation approach to improve recognition again by allowing the system to focus its efforts on object pixels and discard misleading influences from the background.

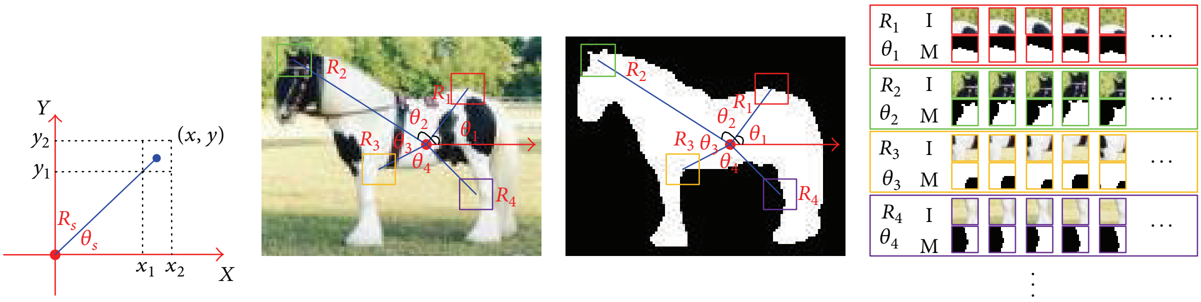
*3.4.1. Back-Projection for Object Detection.* In addition to the hypothesis voting capabilities, the IHRF can also be applied in reverse to detect the positions of their support. The location of a local maximum in Hough image $HI$ encodes scale, class $c_o$, and ISM of the object. More specifically, given a local maximum at position $S_m$, we define the support of the strongest hypothesis as the sample set

$$\begin{aligned} B(D_l, S_m) &= \bigcup_{l\in D_l}\{S_{V,l}, \theta_{md,l}, \mathrm{ISM}_l(c_o) \mid S_m\} \\ &= \bigcup_{l\in D_l}\left\{S_{V,l}, \theta_{md,l}, p_l\left(\frac{F}{c_o}\right), \theta_{mt,l}, d_l \mid S_m\right\}, \end{aligned} \tag{17}$$

which contains the patch entries of all local samples $D_l$ (they are obtained after Hough voting) that have voted for the center position $S_m$. By using their corresponding voting vectors $d_l$ and MGO $\theta_{md,l}$, we can back-project the original position of samples $l$ onto the image space. In this way, we obtain a sparse point set of positions supposedly belonging to the object that voted for the center position $S_m$.

*3.4.2. Refined Object Segmentation Based on a Clustering Scheme.* The back-projected hypothesis's support already provides a rough indication of where the object is in the image. However, the sampled local patches still contain the background structure, as shown in Figure 1(g). We can actually express the a priori known object content without background in terms of a binary mask according to the extracted image patch—for example, see Figure 1(e). Thus, we know more about the pure interpretation of the matched patches for the target object and use it to segment the object from the background by a clustering scheme.

To produce this top-down segmentation, our approach clusters the local masks according to the spatial distribution as the first constraint shown in Figure 8(a), where the coordinates' origins demonstrate the object center and we generate uniformly spaced coordinates with a special interval $\Delta$ in both the $x$ and $y$ directions. Then we group the local training patches, with positions $\{x, y\}$ that belong to the ($[x_1 x_2], [y_1 y_2]$, where $x_2 - x_1 = \Delta$, $y_2 - y_1 = \Delta$) range, as a cluster. Specifically, we use a polar coordinate $\{R, \theta\}$ to satisfy the rotation requirement in IHRF, which has a polar axis that overlaps the $x$-axis. As a result, the image and mask information can be clustered and refer to the training patches' polar parameters $\{R_i, \theta_i\}$, as shown in Figures 8(b)–8(d).



(a)

(b) Image

(c) Mask

(d)

Figure 8: The clustering scheme based on the spatial constraint and appearance similarity. (a) is the spatial constraint; (b) and (c) are the image and its mask, respectively; and (d) is the clusters for object with the two constraints.

The second constraint is the appearance similarity for matching of the support's patch, which is estimated in a similarity metric as

$$D(C_1, C_2) = \frac{\sum_{p\in C_1, q\in C_2} \mathrm{NCC}(p, q)}{|C_1| \times |C_2|} > T, \tag{18}$$

where $C_1$ and $C_2$ are the two inquiring patches, $p$ and $q$ are the color or grayscale values, $T$ is the similarity threshold, and $|*|$ represents the number of pixels in the matching patch. In addition, NCC is the similarity between two patches and is measured by a normalized correlation coefficient, defined as

$$\mathrm{NCC}(p, q) = \frac{\sum_i (p_i - \bar{p}_i)(q_i - \bar{q}_i)}{\sqrt{\sum_i (p_i - \bar{p}_i)^2 \sum_i (q_i - \bar{q}_i)^2}}. \tag{19}$$

In this paper, the object's appearance is represented by a $16 \times 16$ normalized image patch, which is resampled from an image within the original object bounding box regardless of the aspect ratio.

Our clustering scheme can be split into two stages: storing the patches and their masks using the spatial constraint and matching the similar patch using similarity metric. This approach guarantees that only the grouped patches are spatially and visually similar and provide the binary mask to segment the foreground from the background, which is evident from the refined results shown in Figures 1(h) and 1(i).

Furthermore, the spatial and appearance similarity can perform rotation invariance according to the IHRF and (17). Thus, the invariant cluster indexes of the inquiring patch can be looked up from

$$\theta_S = \begin{cases} \theta_{mt} + \theta_C - 180^\circ, & \text{in the 1th or 2th quadrant,} \\ \theta_{mt} + \theta_C + 180^\circ, & \text{in the 3th or 4th quadrant,} \end{cases} \tag{20}$$

while $R_s$ is still equal to $R_C$. Then, the local patch should be rotated by the MGO deviation angle $\Delta\theta = \theta_{md} - \theta_{mt}$, after matching by (18) and finding the local mask by the similarity, which is above $T$. Finally, we rotate the binary mask by the inverse angle $\Delta\theta$ to segment the object more precisely without the effects of rotation.

To this end, the training data is segmented by the two constraints, and the local foreground-background masks are stored at the same time. When a maximum is detected in the voting space, the local segmentation masks are used to infer a global segmentation for the detection in the image.

## 4. IHRF Based Online Object Tracking

Up to now, we have defined all the parts that are necessary to perform object detection and segmentation in an IHRF framework. We can extend this method to handle the online tracking task as well. Recently, online learning frameworks have been designed for long-term tracking of an unknown moving object. The key defining characteristic of online learning is that it can use current true label feedback to update its hypothesis for future predictions, which are close to the true labels. This framework is able to adapt and learn in difficult changing situations because of its continual feedback and update.

During long-term tracking, the main challenge is to avoid drifting problems while still being adaptive to significant occlusions, scale variations, and changes in the object's appearance and deformation. Fine segmentation delivers a more precise description of the object and is used to decrease the effect from the background in the online learning stage [15]. As long as an appropriate classifier exists, the online learning framework will learn to predict correct labels. Therefore, the key point is to use the precise segmentation of a moving object, which has been produced in Section 3.4. We then use this segmentation (essentially a binary mask) to accurately update our classifier, which allows learning of extensive object variations during tracking.

In this paper, the block diagram for online object tracking is illustrated in Figure 9. The main components of the model can be characterized as follows.

(1) The RILBF and scanning pyramid element performs full scanning of the local patches to represent object and background appearance. It provides rotation and scale invariance and discriminable features to significantly increase the classification accuracy.

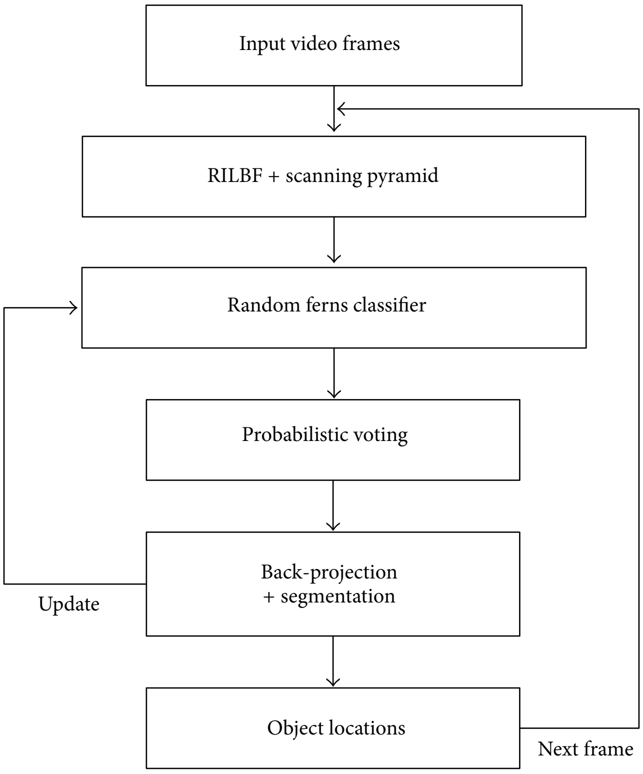


Figure 9: Block diagram of the online learning framework based on IHRF for object tracking.

(2) The RFC decides about the presence or absence of the object. Once an unknown template occurs, it will bring this new variation for retraining the classifier.

(3) The probabilistic voting model estimates the object's locations under the local maxima assumption that encodes the hypothesis's scale and ISM information.

(4) The back-projection and segmentation model integrates both of the hypothesis's support and binary mask. It provides a precise object mask without background noise to retrain the classifier and extract the object features in the next frame.

The implementation of these components was described as an IHRF in Section 3.

Based on the block diagram (see Figure 9), a processing flowchart for online object tracking can be summarized as follows.

*Step 1* (initialization in the first frame). It initializes the object RILBF as a positive sample defined by the user in the first frame and then trains the original RFC associated with some random selected background representation as a negative sample.

*Step 2* (a cyclic thread for tracking in the remaining video stream). It uses the IHRF to detect the object by a bounding box and estimate its motion between consecutive frames. In this model, a motion constraint, such as a temporal and spatial structure [41], restricts the potential states of the object.

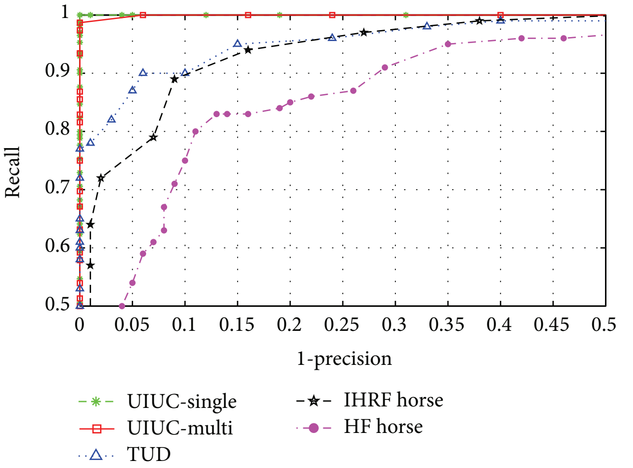


Figure 10: RPC on four datasets. All curves are generated by IHRF, except that of HF Horse, which is produced by a Hough Forest approach on Weizmann Horse.

*Step 3* (providing an update and mask procedure in each frame). According to the motion constraints and the precise binary mask, it distinguishes the object's appearance from the background and updates the classifier for all of the variances.

## 5. Experiments and Results

In this experimental section, we present two types of comparative experiments: object detection and object tracking, which are presented in Sections 5.1 and 5.2, respectively. In order to evaluate our method's performance and compare it with state-of-the-art approaches, we apply these methods to several challenging datasets. We also adhere to the experimental protocols and detection accuracy criteria established for each of the datasets in previous works. All experiments have been done on a standard 3.2 GHz PC with 2 Gigabytes of RAM.

*5.1. Object Detection.* For object detection, the settings were as follows: the RFC consists of $K = 10$ ferns and we pick $S = 13$ pairwise pixels for RILBF. In a multiscale setting, seven scale ratios $S_V = 1.2^n$, where $n$ is the index range $[-2, 4]$, were used to handle the variety of scales in the test data. The interval $\Delta = 20$ pixels and the similarity threshold $T = 0.5$ are used for clustering.

According to previous works, we evaluated the IHRF on several challenging datasets including UIUC cars, TUD pedestrians, and Weizmann Horse. The recall-precision curve (RPC) [51] (see Figure 10) is generated by changing the probability threshold on the hypotheses vote strength. In Table 1, we also provide a performance comparison with the best previously published results [1]. Obviously, IHRF outperforms the previous methods and achieves the best results.

Table 1: Performance of different methods on the four datasets at recall-precision equal error rate (EER).

| Methods | UIUC-Single | UIUC-Multi | TUD | Horse |
|---|---|---|---|---|
| ISM [9] | 97.5% | 95% | 80% | — |
| Efficient Subwindow Search [22] | 98.5% | 98.6% | — | — |
| Hough Forest (HF) [1] | 98.5% | 98.6% | 86.5% | 83% |
| Mutch and Lowe [52] | 99.9% | 90.6% | — | — |
| **IHRF** | 100% | 98.7% | 90% | 89% |

Table 2: Comparison of the two methods on the three datasets using SR and ALA.

| | Methods | UIUC-Single | UIUC-Multi | TUD |
|---|---|---|---|---|
| SR | HF | 95% | 91% | 95% |
| | IHRF | 95% | 95% | 98% |
| ALA | HF | 0.80 | 0.70 | 0.76 |
| | IHRF | 0.85 | 0.75 | 0.78 |

To define the performance more precisely, we compare the object position with ground truth using two evaluation protocols based on bounding-box overlap [36].

(1) Successful rate (SR) is equal to the number of correct positions divided by the number of test images. The correct position means that the overlap score $\mathrm{OS} = A \cap B / A \cup B$ between the bounding box of detection and its ground truth is larger than 50 percent.

(2) Average localization accuracy (ALA) is an average overlap score calculated from all the test images.

The overlap score results on the three datasets are compared for the IHRF and HF approaches, as illustrated in Figure 11, where the vertical and horizontal axis are the overlap score and corresponding image number, respectively. Note that each image in these datasets contains only one object since we only focus on single object detection in this work.

As a result, the quantitative evaluations compared with Hough Forest confirm that IHRF performs well on both ALA and SR, as demonstrated in Table 2. More details are discussed as follows.

*5.1.1. UIUC Cars.* The UIUC car dataset [53] contains two types of car images. The first is the UIUC single-scale (UIUC-Single) test set, which consists of 170 images containing side views of cars of approximately the same size. Another is the UIUC multiscale (UIUC-Multi) test set, which consists of 108 images containing car side views at multiple scales. We trained the IHRF using the available 400 positive and 400 negative training images.

Our IHRF approach achieved an impressive 100% EER for UIUC-Single and 98.7% EER for UIUC-Multi, thus exactly outperforming the state-of-the-art performance. Table 1 also shows that the IHRF considerably outperformed the Hough-based ISM approach [9] and Efficient Subwindow Search approach [22] as well as the Mutch and Lowe's method [52]. In addition, our method is both simpler and more powerful than Hough Forest [1] since naive Bayesian scheme in ferns outperforms the averaging of posteriors used to combine the output of the decision trees [3].

As the overlap score results show in Figures 11(a) and 11(b), the Hough Forest approach is sometimes equal to zero since it is more vulnerable to missing object because of the cluttered background or the multiple scales influences. However, our method can still detect the object with a tolerant error. Table 2 further confirms that our method ($\mathrm{SR} = 95\%, \mathrm{ALA} = 0.85$ for UIUC-Single and $\mathrm{SR} = 95\%, \mathrm{ALA} = 0.75$ for UIUC-Multi) slightly outperforms Hough Forest on both SR and ALA evaluation scheme.

Some examples of those detection cases are displayed in Figure 12. In this experiment, the results show that the IHRF not only detects object despite partial occlusion but also is often even able to deal with the scale variations. For an image from the UIUC car dataset, our method only requires 0.4 seconds (no less than 200 × 150 pixel resolution).

Both the datasets include samples of partially occluded cars, cars with low contrast within the cluttered background, and challenging illumination. However, the shape of the objects remains rigid, which makes the detection task easier. Therefore, we will assess the performance of our method on more challenging datasets that include highly nonrigid object transformations as follows.

*5.1.2. TUD Pedestrian.* In this section, we apply our approach to pedestrian detection in crowded street scenes using the TUD pedestrian dataset. This is a highly challenging test set that consists of 400 training images containing crowded street scenes [54]. For this experiment, we followed the experimental protocol of [47] to train our detector and tested it on 100 pedestrian images.

The performance of the different methods is shown in Table 1. For TUD pedestrians, our method performed EER = 90%, which is significantly better than both the Hough Forest [1] and ISM-based methods [9]. Figure 11(c) and Table 2 also demonstrate that our method ($\mathrm{SR} = 98\%$, $\mathrm{ALA} = 0.78$) is still competitive and performs better than Hough Forest approach.

In order to give a better impression of our method's performance, Figure 13 shows obtained detection results on example images from the test set. As can be seen from those examples, the proposed method can reliably detect and localize such deformable classes as pedestrians in crowded scenes and with severe overlaps. Our method requires no more than 1 second for a 400 × 320 pixel image from the TUD dataset.

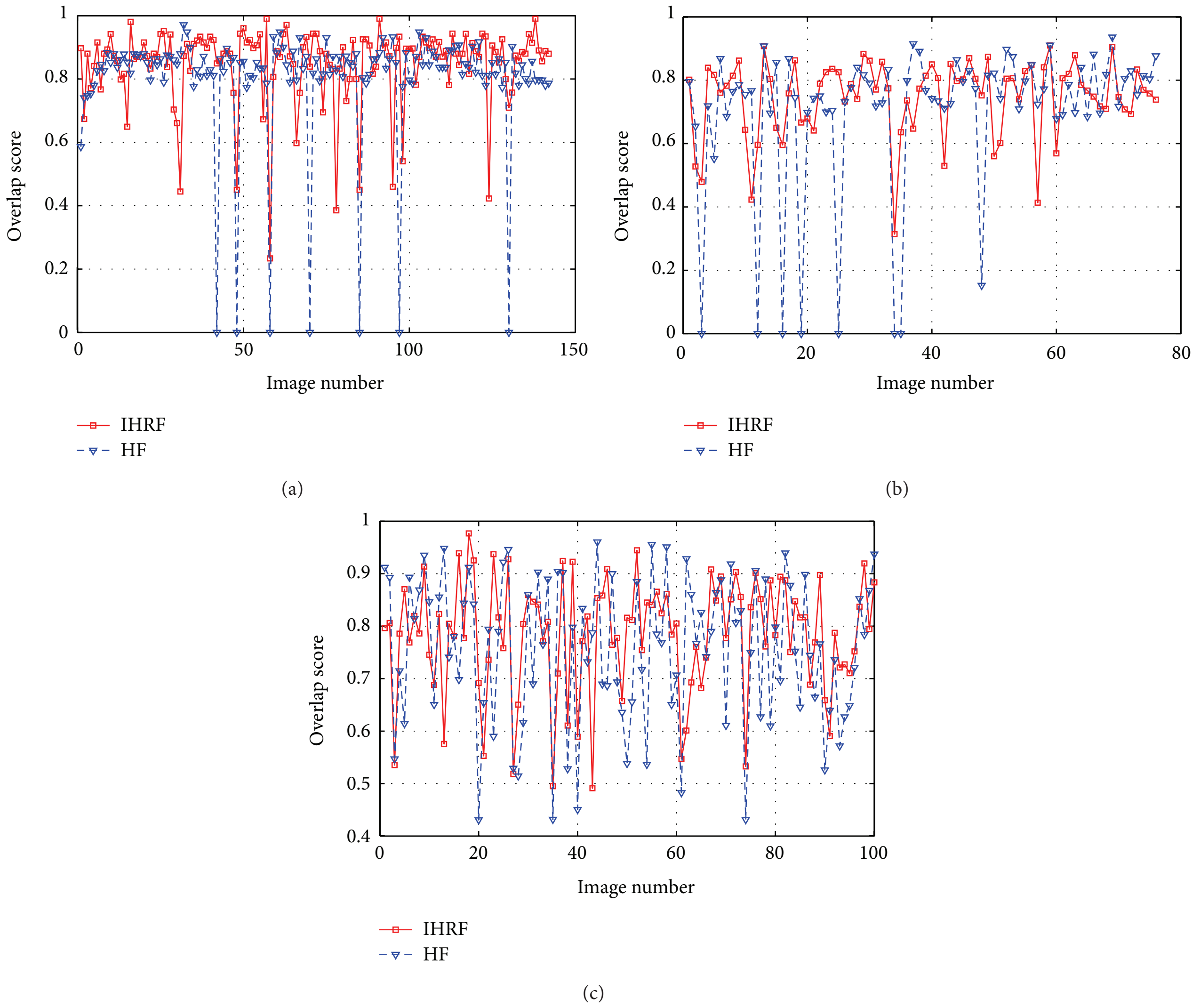


FIGURE 11: The overlap score results of detecting a single object with IHRF and HF on the three datasets. (a) UIUC-Single, (b) UIUC-Multi, and (c) TUD.

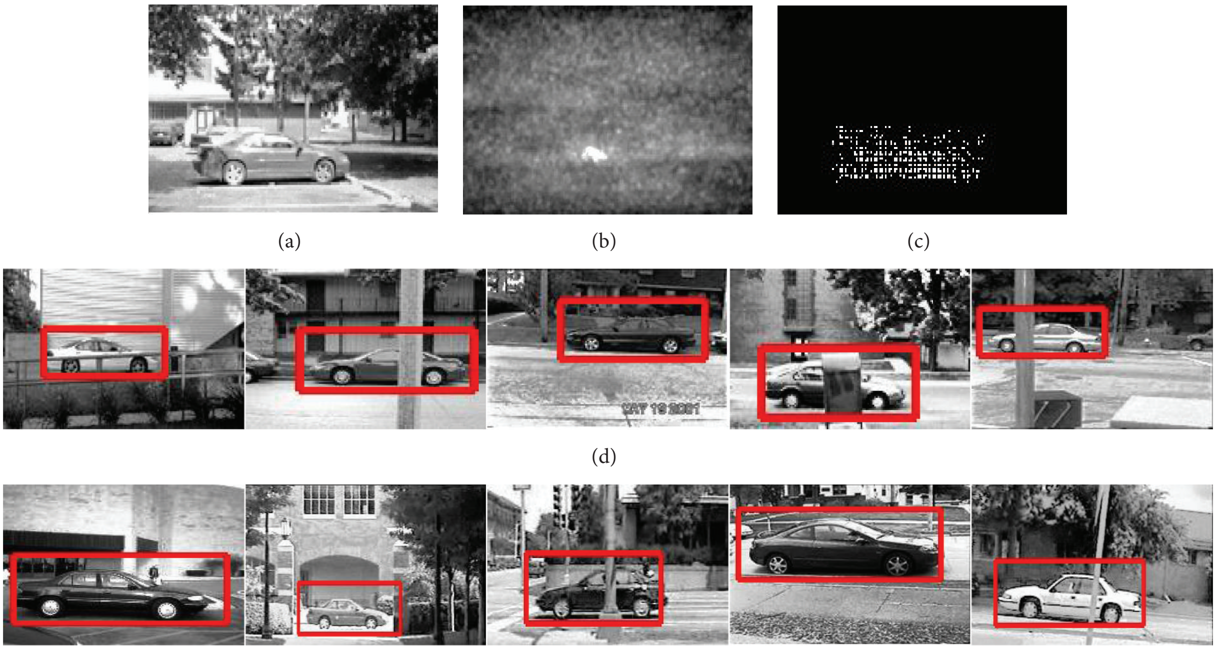
*5.1.3. Weizmann Horse.* Finally, we also assessed the IHRF and Hough Forests performance on the Weizmann Horses dataset [46], which comprises 328 multiscale side views of horses in cluttered environments under varying scale and strongly varying poses. We split the training testing as suggested in [1] by using 100 horse images and 100 background images for training and the rest of 228 horse images and 228 background images for testing.

Figure 10 shows the RPC, and Table 1 shows a comparison of our method's EER obtained by IHRF and Hough Forest on the same Horse dataset. As can be seen from those results, our method achieves good detection results with an EER performance of 89%, which presents a significant improvement over previous results. Our detection results are shown in Figure 14. Note that the training and testing stages adhere to the original horse images without two improvements, as described in [47].

*5.2. Online Object Tracking.* For the online object tracking task, we first separate the moving object from the background on a more fine-grained level to obtain more accurate training data. In fact, this is a stable and effective way to avoid the drifting problem. Therefore in this section, the experiments are divided into two parts. First, we perform segmentation experiments demonstrating three specific properties of our approach, and second we present results on public available sequences for comparison with other tracking approaches.

*5.2.1. Refined Segmentation.* The goal of this section is to illustrate three properties of the proposed segmentation approach: rotation and scale invariance, occlusion, and deformation handling capability. Therefore, multiple challenging datasets have been provided to verify our method's performance.

*(1) Rotation and Scale Invariant Performance.* We have presented a new object detection and segmentation approach based on IHRF to detect objects that may appear in the image under different orientations and scales. In contrast to other works that address this problem using multiple

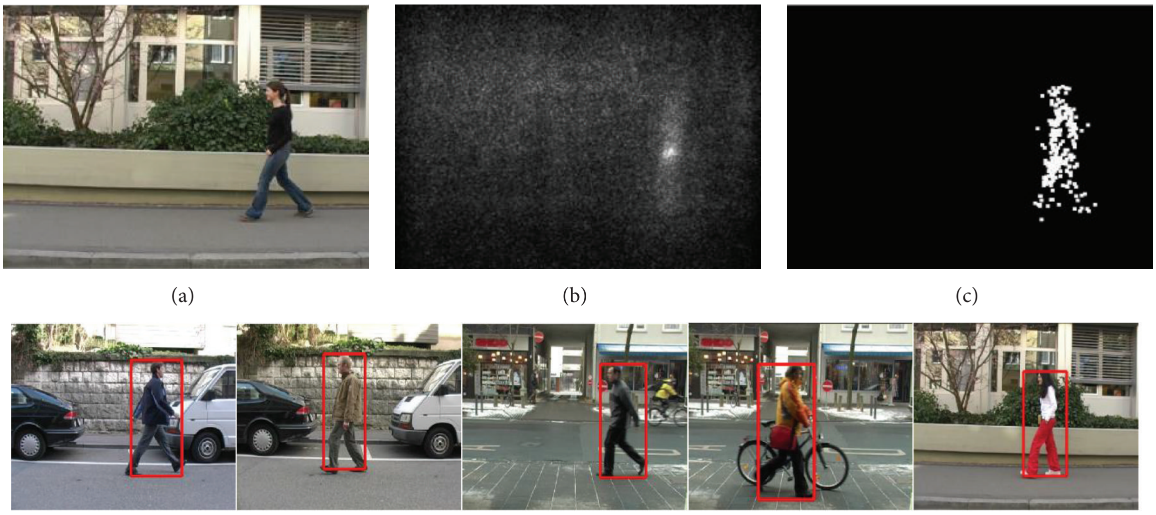


(a) (b) (c)

(d)

(e)

Figure 12: Object detection obtained by an offline trained IHRF for cars. (a) Original image; (b) Hough voting image; (c) support of the strongest hypothesis. The detection results on UIUC car datasets, such as (d) UIUC-Single and (e) UIUC-Multi.



(a) (b) (c)

(d)

Figure 13: Object detection obtained by an offline trained IHRF for pedestrians. (a) Original image; (b) Hough voting image; (c) support of the strongest hypothesis; (d) some detection results on TUD pedestrian dataset.

classifiers or multiple orientation samples, we assign a consistent orientation (MGO) to each patch based on local image properties, and then the RILBF descriptor can be represented relative to this orientation and achieve invariance to image rotation. We also use a rotation invariant back-projection to locate the support of our detection and guide a refined segmentation process that precisely separates the object from the background. In addition, the scanning windows are resized at different scales since we maintain the same polar angle of the pairwise pixels but apply different scale ratios to their radius $R$.

The following experiment is used to demonstrate the IHRF's capability of rotation invariance. We train our IHRF on our own training image and a hand-segmented face mask; see Figure 15(a). Then we sample the image under 2D rotations in 30-degree steps that leads to 12 samples. Figure 15(b) depicts their segmentation results using the IHRF. As can be seen from Figure 15(b), the IRHF approach

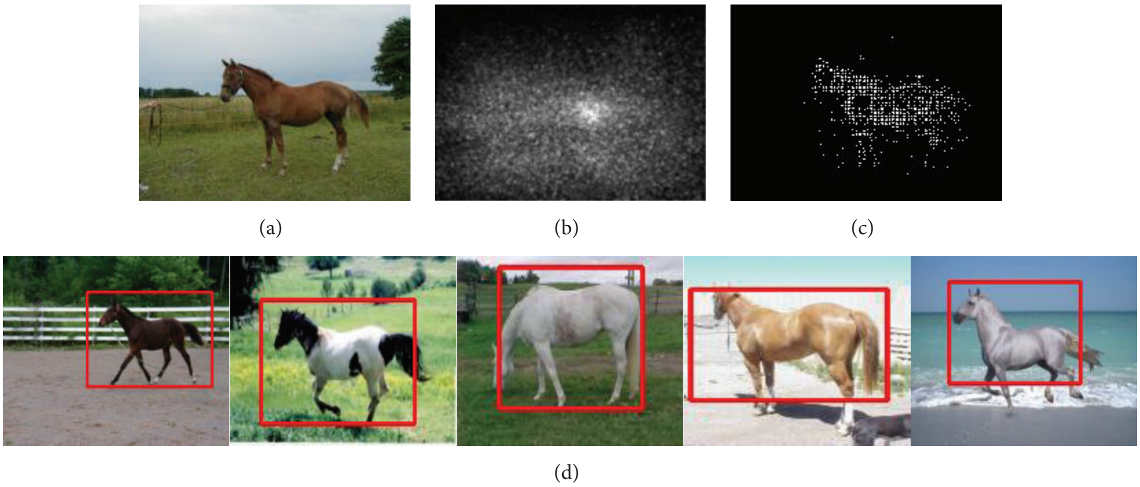


Figure 14: Object detection obtained by an offline trained IHRF for horses. (a) Original image; (b) Hough voting image; (c) support of the strongest hypothesis; (d) some detection results on Weizmann Horse dataset.

can adapt to the face rotation variation and yields a stable profile regardless of the rotation. Similarly using the IHRF, the localization of face in different scales can be observed in Figure 15(c).

*(2) Occlusion and Deformation Handling Performance.* In the domain of part-based object detection, detectors can always improve detection results for partial or self-occlusion and nonrigid deformations. We have demonstrated that the IHRF voting and segmenting models allow us to detect objects reliably even under partial occlusions and heavy nonrigid deformation. Therefore, we process some example detections in occlusion and deformation configurations and the corresponding top-down segmentations can be seen in Figure 16.

Figures 16(a) and 16(b) show human faces that are occluded by a book or a hat [42, 55]. IHRF is highly robust to facial occlusion because a large number of small patches without occlusion still have a high probability of voting for the supporting points. Those results confirm that our method still works in the presence of occlusion and cluttered backgrounds. Figure 16(c) depicts a walker with limited pose variations. In this case, several parts of the object maintain a stable geometric configuration vote for the center of the object. This clearly shows that our approach delivers reliable results even if the object undergoes heavy deformations in a complex background.

*5.2.2. Online Tracking.* This section reports on a set of quantitative experiments comparing our system (denoted as IHRFT) with other relevant algorithms, which include online boosting (OLB) [56], tracking-learning-detection (TLD) [41], online learning tracker (OLT) [57], and Hough-based GrabCut tracker (HGT) [15].

Unfortunately, no public framework is available for comparing tracking techniques. Hence, we decided to process publically available sequences, as shown in Table 3. The first two experiments (David and Girl) evaluate our system on face-tracking sequences that are commonly used in the literature [58, 59]. In both of these experiments, a human face is moving with challenging conditions such as lighting, scale, and pose changes. In addition, the Coke Can and Tiger sequences [42] contain frequent occlusions and fast motion as well as challenging out-of-plane rotations, cluttered backgrounds, changing illumination conditions, and partial occlusions. For the purpose of tracking deformed object during runtime, we have also collected three more challenging videos (Bike, Motor [15], and Diving [60]) that show different ranges of complexity and highly nonrigid deformations.

Using these datasets, each tracking task has been initialized by manually marking the target object in the first frame, and each tracker, respectively, tracks the object until the end of a sequence. In order to illustrate the performances of the trackers, the produced trajectory is then compared with ground truth using two evaluation protocols: SR and ALA, which have been adopted in Section 5.1. The parameters of our system in this experiment are the same as those defined in Section 5.1. Note that this parameter has been set empirically and its value is not critical. The overlap score results in each sequence are compared with our system and other approaches on SR (Table 3) and ALA (Table 4).

In Table 3, the results in the first two rows show that the four approaches perform well on saturated sequences, except for the OLB. However, the results for Coke Can and Tiger show that HGT may be more vulnerable to missing object in cases with cluttered backgrounds. Furthermore, the TlD and OLT show poor performance in the last three datasets because they utilize a bounding-box-based tracker that is not designed to cope with the amount of deformations in these videos. In particular, when tracking a diving woman, HGT cannot adapt to the gradual nonrigid deformations since the GrabCut segmentation algorithm fails when there are similar colors in the background [15]. In contrast, the IHRFT is not only capable of robustly tracking moving objects of interest through all these challenging sequences but also superior

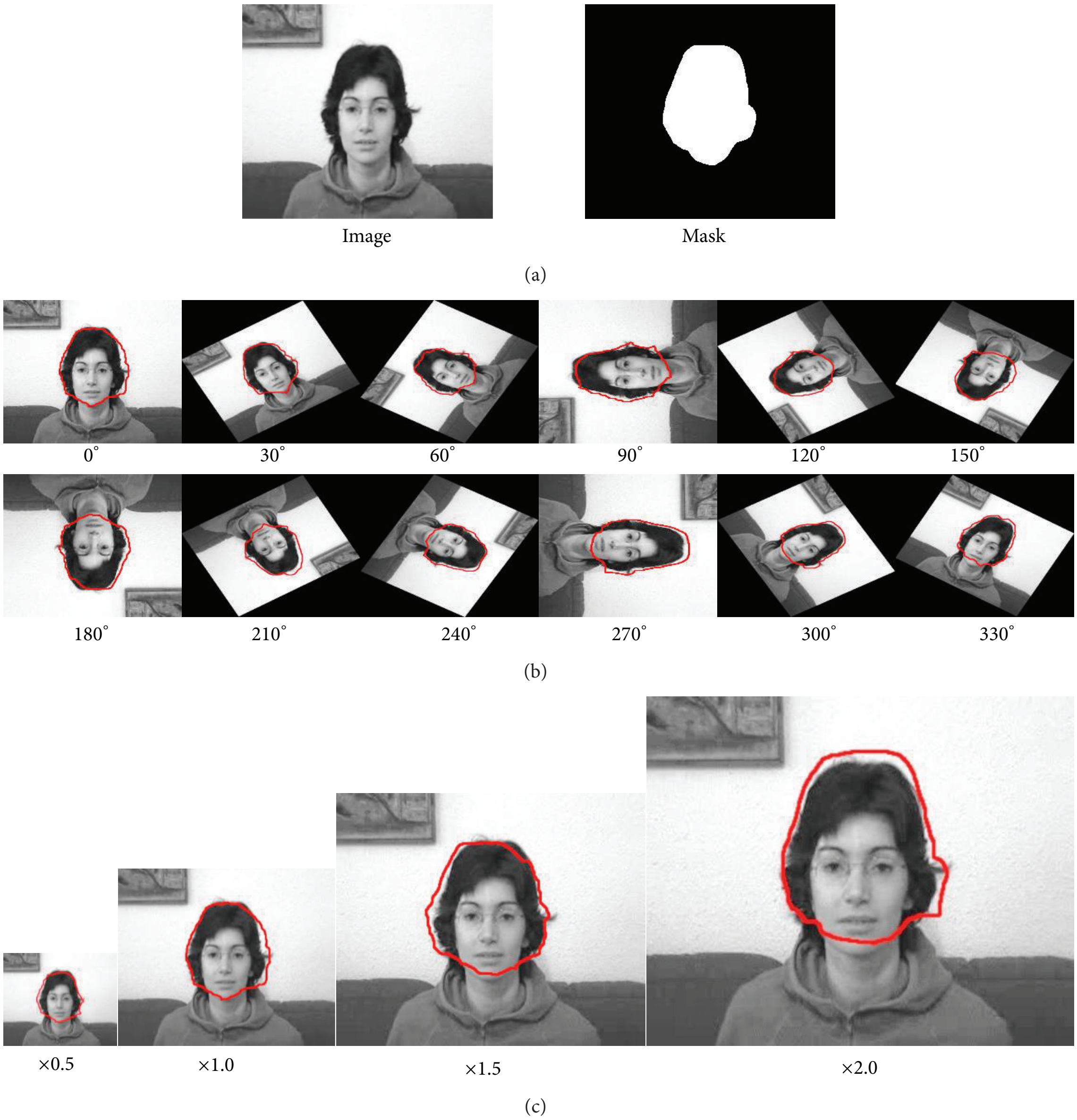


FIGURE 15: Rotation and scale invariance. (a) The training image and mask. Segmentation results under different (b) orientations and (c) scale factors.

TABLE 3: Comparison of the tracker on SR.

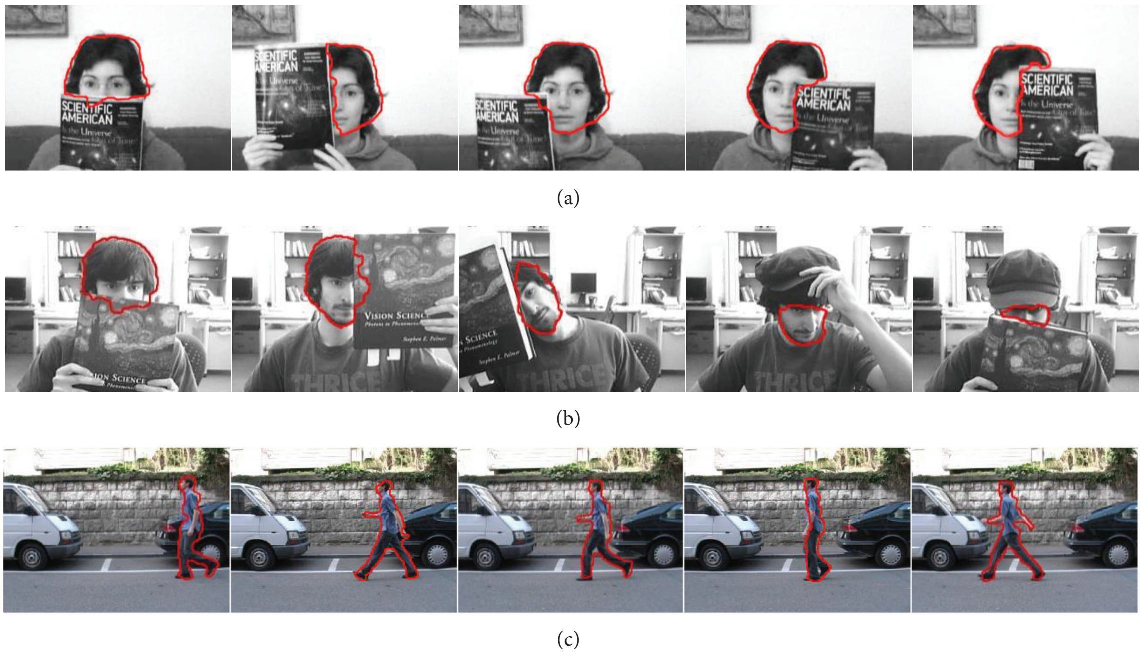
| Sequence | OLB | TLD | OLT | HGT | IHRFT |
|---|---|---|---|---|---|
| David | 24% | **100%** | 98.7% | 98.3% | **100%** |
| Girl | 24.5% | 91.7% | 86.1% | 86.6% | **97.1%** |
| Coke | 90.7% | 81.3% | 92% | 27.3% | **99%** |
| Tiger | 44.8% | 88.7% | 89.6% | 49.2% | **98.6%** |
| Bike | 94% | 74.5% | 80.2% | 99.1% | **100%** |
| Motor | 41.1% | 16% | 50.3% | **97.6%** | 93.2% |
| Diving | 16.3% | 22% | 24% | 34.9% | **71.5%** |
| Mean | 47.9% | 67.7% | 74.4% | 70.4% | **94.2%** |

Bold font indicates the best overlap score obtained by one of the trackers in each video.

Table 4: Comparison of the tracker on ALA.

| Sequence | OLB | TLD | OLT | HGT | IHRFT |
|---|---|---|---|---|---|
| David | 0.40 | **0.83** | 0.79 | 0.81 | **0.83** |
| Girl | 0.35 | 0.71 | 0.72 | 0.67 | **0.80** |
| Coke | 0.66 | 0.59 | 0.61 | 0.46 | **0.81** |
| Tiger | 0.35 | 0.68 | 0.74 | 0.36 | **0.77** |
| Bike | 0.64 | 0.53 | 0.55 | 0.82 | **0.86** |
| Motor | 0.26 | 0.60 | 0.59 | 0.61 | **0.75** |
| Diving | 0.22 | 0.27 | 0.28 | 0.21 | **0.63** |
| Mean | 0.41 | 0.60 | 0.61 | 0.56 | **0.78** |

Bold font indicates the best accuracy obtained by one of the trackers in each video.


(a)

(b)

(c)

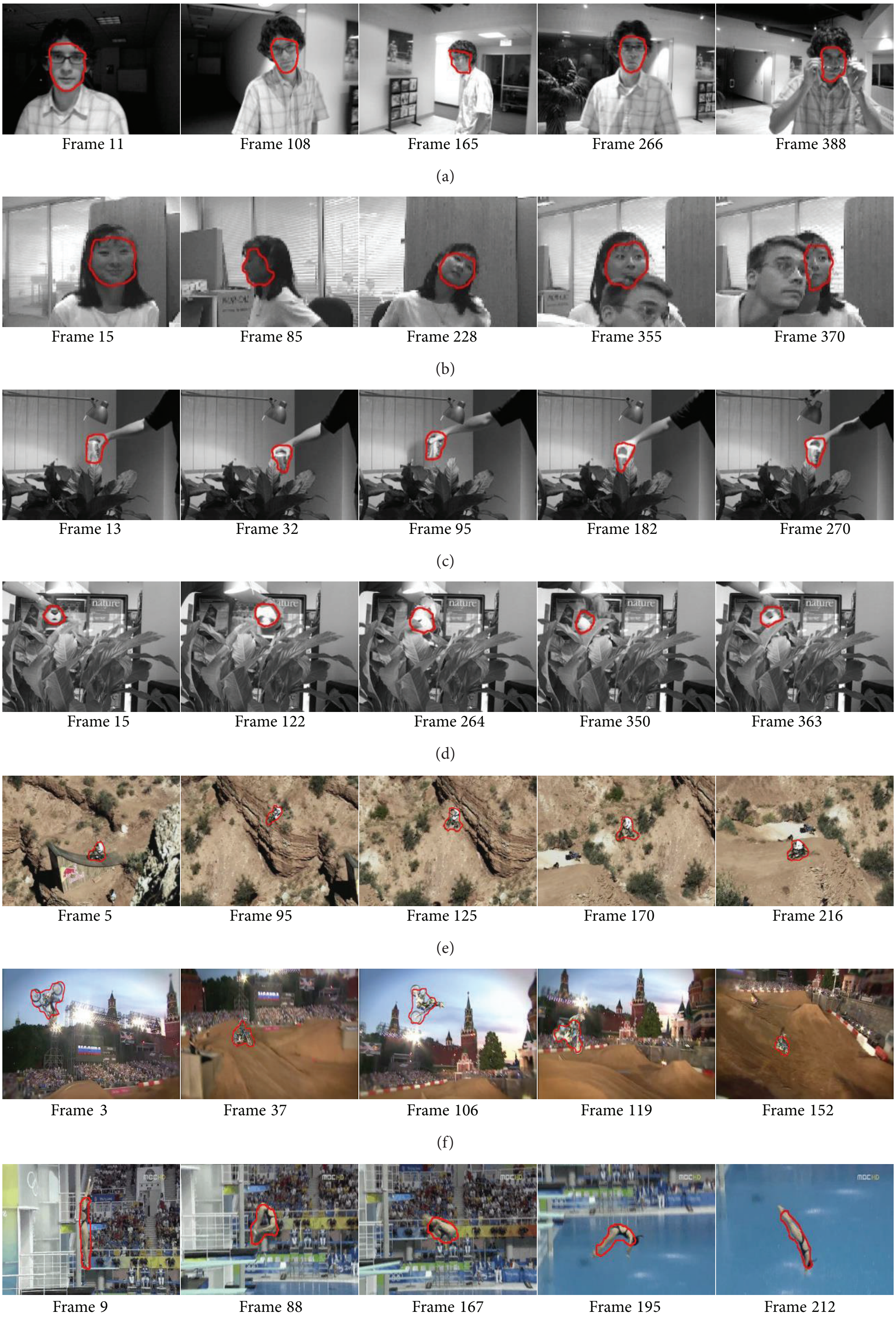
Figure 16: Object segmentation under occlusion and deformation configurations. (a) Face occlusion 1 [55]. (b) Face occlusion 2 [42]. (c) TUD pedestrian.

to the other approaches. Specifically, for the performances measured by SR, our average score outperforms the second best tracker by more than 19.8 percent.

In addition, the quantitative evaluations (in Table 4) compared with other approaches on ALA confirm that IHRFT achieves the best scores in all the sequences, where the average score is 17 percent higher than the second best results. These successful achievements rely on a highly accurate segmentation algorithm. Figure 17 shows some selected frames of the sequences and our tracking results. Even though the challenging conditions present in the sequences, such as cluttered backgrounds, changing illumination conditions, partial occlusions, and non-rigid deformations, the segmentations are still reliable and can serve as a basis for later update stages to further improve the classifier's performance.

## 6. Conclusions

We have proposed an IHRF approach for part-based object detection and online tracking. It relies on rotation and scale invariant descriptors based on RFC that are able to cast probabilistic votes within the Hough transform framework. The matching operations are simply performed on local patches that have been transformed relative to their assigned orientation, scale, and location, thereby providing invariance to these transformations. Such IHRF can be efficiently used



Frame 11 Frame 108 Frame 165 Frame 266 Frame 388

(a)

Frame 15 Frame 85 Frame 228 Frame 355 Frame 370

(b)

Frame 13 Frame 32 Frame 95 Frame 182 Frame 270

(c)

Frame 15 Frame 122 Frame 264 Frame 350 Frame 363

(d)

Frame 5 Frame 95 Frame 125 Frame 170 Frame 216

(e)

Frame 3 Frame 37 Frame 106 Frame 119 Frame 152

(f)

Frame 9 Frame 88 Frame 167 Frame 195 Frame 212

(g)

Figure 17: Snapshots of tracking results.

to detect instances of classes in large challenging images with an accuracy that is superior to previous methods. This approach also allows for time-efficient and space-saving implementation compared with related techniques.

In addition, based on the hypothesis's support determined by the back-projection, we have provided an efficient clustering scheme to guide a segmentation process which precisely separates the object from the background. This top-down segmentation delivers a more precise description of the object and is used to decrease the noise in the online learning stage for object tracking. Therefore, our online tracking method has been validated using several datasets under challenging conditions, such as cluttered background, partial occlusions, and nonrigid deformations. The tracking results show our tracker achieves good object location and segmentation performance in difficult real-world scenes and outperforms state-of-the-art methods.

Our approach applies the RFC to build IHRF which are then used for the object detection. The success of this achievement is highly conditioned on the IHRF performance in 2D images. Therefore, in future work, an interesting extension would be to apply the IHRF to the problem of 3D shape recognition and registration.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This work was supported by the Open Project of Beijing Key Laboratory on Measurement and Control of Mechanical and Electrical System (no. KF20121123206) and Key Laboratory of Modern Measurement and Control Technology (BISTU), Ministry of Education; Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT, the IRT1212); Funding Project for Academic Human Resources Development Institutions of Higher Learning under the Jurisdiction of Beijing Municipality (no. PHR201106130); and Funding Project of Beijing Municipal Science & Technology Commission (no. Z121100001612011).

## References

[1] J. Gall, A. Yao, N. Razavi, L. van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.

[2] B. Leibe, K. Schindler, N. Cornelis, and L. van Gool, "Coupled object detection and tracking from static cameras and moving vehicles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1683–1698, 2008.

[3] M. Özuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast keypoint recognition using random ferns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 448–461, 2010.

[4] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 1, pp. 446–453, October 2005.

[5] M. Villamizar, F. Moreno-Noguer, J. Andrade-Cetto, and A. Sanfeliu, "Efficient rotation invariant object detection using boosted random ferns," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 1038–1045, June 2010.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[8] R. Marée, P. Geurts, J. Piater, and L. Wehenkel, "Random subwindows for robust image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 34–40, June 2005.

[9] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 259–289, 2008.

[10] S. Maji and J. Malik, "Object detection using a max-margin hough transform," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1038–1045, June 2009.

[11] O. Barinova, V. Lempitsky, and P. Kholi, "On detection of multiple object instances using Hough Transforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1773–1784, 2012.

[12] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1270–1281, 2008.

[13] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[14] R. Okada, "Discriminative generalized hough transform for object detection," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 2000–2005, October 2009.

[15] M. Godec, P. M. Roth, and H. Bischof, "Hough-based tracking of non-rigid objects," *Computer Vision and Image Understanding*, vol. 117, pp. 1245–1256, 2012.

[16] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.

[17] Z. Kalal, J. Matas, and K. Mikolajczyk, "Pn learning: bootstrapping binary classifiers by structural constraints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 49–56, usa, June 2010.

[18] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 810–815, 2004.

[19] C. Galleguillos and S. Belongie, "Context based object categorization: a critical survey," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 712–722, 2010.

[20] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 886–893, June 2005.

[22] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: object localization by efficient subwindow search," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[23] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[24] C. Gu, J. J. Lim, P. Arbeláez, and J. Malik, "Recognition using regions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1030–1037, June 2009.

[25] P. Wohlhart, M. Donoser, P. M. Roth et al., "Detecting partially occluded objects with an implicit shape model random field," in *Asian Conference on Computer Vision*, vol. 7724, pp. 302–315, Springer, 2013.

[26] K. Rematas and B. Leibe, "Efficient object detection and segmentation with a cascaded Hough Forest ISM," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV '11)*, pp. 966–973, November 2011.

[27] J. Kwon and K. M. Lee, "Highly non-rigid object tracking via patch-based dynamic appearance modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2427–2441, 2013.

[28] B. Leibe and B. Schiele, "Interleaving object categorization and segmentation," in *Proceedings of the British Machine Vision Conference*, pp. 759–768, 2003.

[29] P. Yarlagadda, A. Monroy, and B. Ommer, "Voting by grouping dependent parts," in *Proceedings of the 11th European Conference on Computer Vision*, pp. 197–210, 2010.

[30] N. Razavi, J. Gall, P. Kohli et al., "Latent Hough transform for object detection," in *Proceedings of the 13th European Conference on Computer Vision*, pp. 312–325, 2012.

[31] A. Lehmann, B. Leibe, and L. van Gool, "Fast PRISM: branch and bound hough transform for object class detection," *International Journal of Computer Vision*, vol. 94, no. 2, pp. 175–197, 2011.

[32] V. Kumar and I. Patras, "A discriminative voting scheme for object detection using hough forests," in *Proceedings of the British Machine Vision Conference*, pp. 1–10, 2010.

[33] P. Wohlhart, S. Schulter, M. Köstinger et al., "Discriminative hough forests for object detection," in *Proceedings of the British Machine Vision Conference*, pp. 1–11, 2012.

[34] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: a survey," *ACM Computing Surveys*, vol. 38, no. 4, pp. 1–45, 2006.

[35] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: a review," *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, 2011.

[36] Y. Wu, J. Lim, and M. Yang, "Online object tracking: a benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, 2013.

[37] S. Schulter, C. Leistner, P. M. Roth et al., "On-line hough forests," in *Proceedings of the British Machine Vision Conference*, pp. 1–11, 2011.

[38] J. Gall, N. Razavi, and L. V. Gool, "On-line adaption of class-specific codebooks for instance tracking," in *Proceedings of the British Machine Vision Conference*, pp. 55. 1–55. 12, 2010.

[39] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 1, pp. 260–267, June 2006.

[40] B. Babenko, M. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[41] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[42] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 983–990, June 2009.

[43] H. Riemenschneider, S. Sternig, M. Donoser et al., "Hough regions for joining instance localization and segmentation," in *Proceedings of the 13th European Conference on Computer Vision*, pp. 258–271, 2012.

[44] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[45] M. Özuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[46] E. Borenstein and S. Ullman, "Combined top-down/bottom-up segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2109–2125, 2008.

[47] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1022–1029, June 2009.

[48] A. Bosch, A. Zisserman, and X. Muñoz, "Image classification using random forests and ferns," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.

[49] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1465–1479, 2006.

[50] F. Zheng and G. I. Webb, "A comparative study of semi-naive bayes methods in classification learning," in *Proceedings of the 4th Australasian Data Mining Conference*, pp. 141–156, 2005.

[51] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, June 2006.

[52] J. Mutch and D. G. Lowe, "Multiclass object recognition with sparse, localized features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 1, pp. 11–18, June 2006.

[53] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[54] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[55] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 1, pp. 798–805, June 2006.

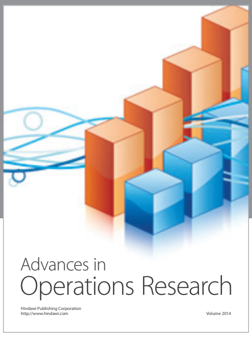
[56] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 1, 5, pp. 260–267, June 2006.

[57] Y. Lin, N. Lu, X. Lou et al., "Online learning of a cascaded classifier designed for multi-object tracking," in *Proceedings of the 11th IEEE International Conference on Electronic Measurement & Instruments*, pp. 1069–1075, 2013.

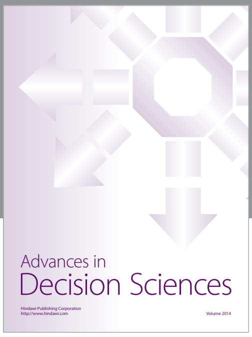
[58] D. A. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 125–141, 2008.

[59] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '98)*, pp. 232–237, June 1998.

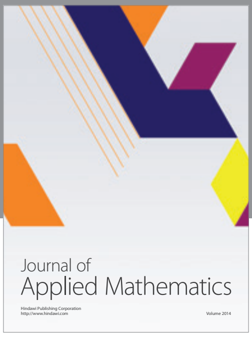
[60] J. Kwon and K. M. Lee, "Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1208–1215, June 2009.



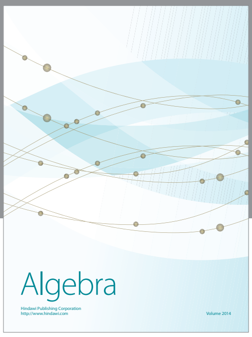
Advances in Operations Research



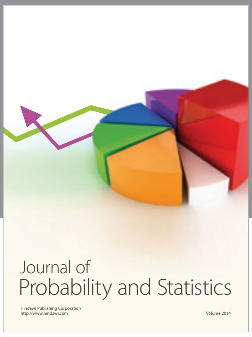
Advances in Decision Sciences



Journal of Applied Mathematics

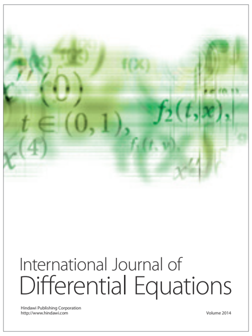


Algebra



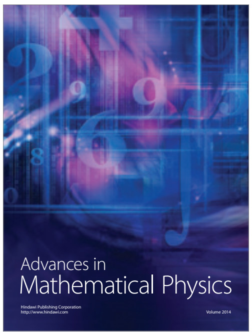
Journal of Probability and Statistics

The Scientific World Journal

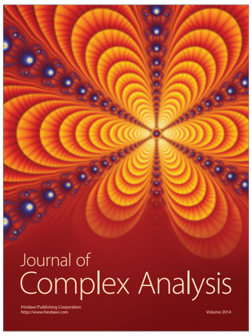


International Journal of Differential Equations

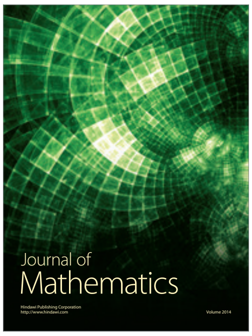
International Journal of Combinatorics



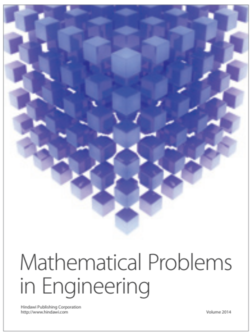
Advances in Mathematical Physics



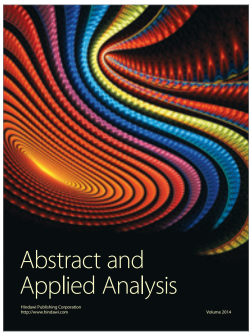
Journal of Complex Analysis



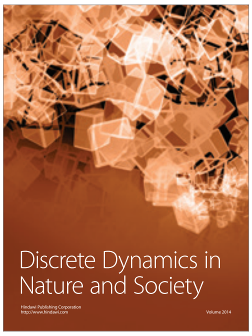
Journal of Mathematics



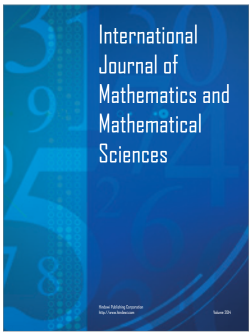
Mathematical Problems in Engineering



Abstract and Applied Analysis



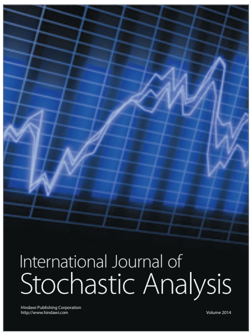
Discrete Dynamics in Nature and Society



International Journal of Mathematics and Mathematical Sciences

Journal of Discrete Mathematics

Journal of Function Spaces



International Journal of Stochastic Analysis

Journal of Optimization