Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 863732, 15 pages
http://dx.doi.org/10.1155/2015/863732

Research Article

RGBD Video Based Human Hand Trajectory Tracking and Gesture Recognition System

Weihua Liu,[1] Yangyu Fan,[1] Zuhe Li,[1] and Zhong Zhang[2]

[1]*School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China*
[2]*Computer Science and Engineering Department, University of Texas at Arlington, Arlington, TX 76019-0015, USA*

Correspondence should be addressed to Weihua Liu; lwh86117@163.com

Received 8 October 2014; Revised 30 December 2014; Accepted 31 December 2014

Academic Editor: Jyh-Hong Chou

The task of human hand trajectory tracking and gesture trajectory recognition based on synchronized color and depth video is considered. Toward this end, in the facet of hand tracking, a joint observation model with the hand cues of skin saliency, motion and depth is integrated into particle filter in order to move particles to local peak in the likelihood. The proposed hand tracking method, namely, salient skin, motion, and depth based particle filter (SSMD-PF), is capable of improving the tracking accuracy considerably, in the context of the signer performing the gesture toward the camera device and in front of moving, cluttered backgrounds. In the facet of gesture recognition, a shape-order context descriptor on the basis of shape context is introduced, which can describe the gesture in spatiotemporal domain. The efficient shape-order context descriptor can reveal the shape relationship and embed gesture sequence order information into descriptor. Moreover, the shape-order context leads to a robust score for gesture invariant. Our approach is complemented with experimental results on the settings of the challenging hand-signed digits datasets and American sign language dataset, which corroborate the performance of the novel techniques.

1. Introduction

In recent years, with the developing of the camera sensor, 3D tracking of human hands attracts considerable interest in the literature of gesture recognition, human grasping understanding, human-computer interfaces, and so forth. Due to the lack of depth information, the 2D hand tracking has to struggle with the light variations and cannot express the semantic integrity of gesture. Further, variability of hand appearance makes it challenging in detecting and tracking in 2D-dimension.

The existing alternatives utilize various sensors, including colored gloved [1] and magnetic tracker [2], to detect hands accurately. Unfortunately, such methods require a costly hardware or complex execute time. In pace with the development of the depth sensor, more and more general tracking approaches are inclined to add depth information into algorithm. Given the fact that hand movements generally occur in front of human bodies in 3D space, the depth information can thus significantly distinguish hand part from complex background. Without utilizing the color information, many papers demonstrate that the single depth data suffices for hand tracking [3, 4], while such approaches are not robust enough and can easily lose tracking since they get rid of using color information. Zhang et al. [5] recently proposed an efficient and fast hand detector by just using the hand color and motion information. The experiment result shows that such detector outperforms the Mittal et al. [6] detector with multiply features and Karlinsky et al. [7] detector by using chain model in sign language video. Nevertheless, this sign language video is captured with pure background, which does not take the more complex real environment into consideration.

Hand tracking can be thought of as a nonlinear and non-Gaussian problem because of the presence of background clutter, complex dynamics of hand motion, and varying illumination. Thus, particle filter is well designed and implemented in this field. In 2D methods, a hand is represented by its geometric features and appearance features such as contours, fingertips, and skin color. Color-based methods often use skin color to localize and track hands in single camera [8, 9]. Spruyt et al. [10] propose a complete

tracking system that is capable of long-term, real-time hand tracking with unsupervised initialization. Shan et al. [11] proposed a general tracking approach by incorporating mean shift optimization into particle filter, which can improve the sampling efficiency by moving particles to local peaks in the likelihood. However, this method is also vulnerable to other hand skin-likelihood cues and may fail to track hand continuously. Also, it does not consider the scale variation of the tracking object. To overcome the scale problem, based on mean shift and particle filter, Wang et al. [12] proposed to combine Camshift based on mean shift and particle filter, to reduce the time complexity and introduce the adaptive scale adjustment factor on each particles to evaluate the optimize object size. For articulated hand tracking, Chang et al. [13] introduced an appearance-guided particles filter for high degree-of-freedom state space. They design a set of attractors which can affect the state parameter of target and thus achieve Bayesian optimal solutions. Another PF-improved approach was proposed by Morshidi and Tjahjadi [14]; they compute the gravitational force of each particle as its weight to attract nearby particles towards the peak of the particles distribution, thus improving the sampling efficiency.

As for 3D hand tracking, Manders et al. [15] use commercially stereo camera to acquire depth information after calibrating camera pair and then depth joint probability model is established according to distance between face and camera and used as input to a Camshift tracker. Similarly, Van Den Bergh and Van Gool [16] use the time-of-flight camera to capture hand shape and location. Nonetheless, all of these color-depth information based methods are needed to be well camera pair calibrated, which are rather time consuming and of lower efficiency.

In this paper, for the hand tracking part, we extend Zhang et al.'s work [5] by incorporating additional depth feature, salient hand skin feature, and motion feature into particle filter, yielding a robust, efficient detector with high tracking accuracy. To start the tracking system, an automatic tracking initializing method is proposed. Finally, the hand center is extracted by using Kernel density estimation on particles distribution.

As for gesture recognition part, there are four main aspects, static human pose, static hand pose, activity human body gesture, and activity hand gesture. Since hand is one of the most dominant and visual sensitive body part to express the gesture meaning, we focus on studying the hand gesture recognition based on hand trajectory. Due to the variety of habits of users, hand gesture is often subtle and vulnerable to various changes, such as position, orientation, and distance of people performing gesture with respect to camera. The main trend methods for classifying gestures are related to indirect approach, such as dynamic programming (DP) methods; for example, an exemplar-based approach like DTW [17] and CTW [18] relies on aligning temporal trajectory between query and model sequence for similarity comparison. Similarly, Wobbrock et al. [19] propose a simple enough method, one dollar gesture recognizer to classify the gesture. In such approaches, trajectory locations are usually adopted as an important match feature for gesture classification, while they cannot be greatly invariant to scale and translation with respect to different sign position, orientation, and distance to camera. This restriction makes the traditional methods difficult to accommodate various ways of acting behaves. In Liu et al.'s paper [20], they consider the gesture as multi-small segmentations and build decision trees by evaluating those directions of segmentations. This method can greatly be invariant to scale and translation, but the classification accuracy still needs to be improved.

Inspired by the shape matching work [21] that develops a local shape context descriptor on each of shape points, we extend this method to spatiotemporal domain by embedding the gesture sequence order to shape context descriptor and give a name as shape-order context. Given the gesture shape and sequence order, the dissimilarity of the two-gesture trajectory can be measured as sum of matching errors among corresponding shape points. Such local descriptor can also be invariant to gesture translation and scale and with high recognition accuracy.

This paper is organized as follows. The framework of particle filter is described in Section 2. In Section 3, we discuss the proposed hand tracking algorithm based on particle filter, detailing with dynamical model, and observation model and also refer to tracking initialization model, hand center localization model. Hand gesture recognition based upon proposed shape-order context feature is then presented in Section 4. In Section 5, we finally evaluate our approach and compare the results with the existing state-of-art hand tracking and hand gesture recognition algorithms. The flow diagram of our proposed work is shown in Figure 1, which consists of two main phases: gesture trajectory tracking phase and gesture recognition phase. In the output of tracking phase, gestures are represented as a set of hand points which can be visualized in gesture sequence representation model. Meanwhile, it is treated as the input of recognition phase as well for gesture classification.

2. Probabilistic Tracking

2.1. Overview Particle Filter. Particle filter is a nonparametric system offering probabilistic framework for dynamic state estimation. It defines finite current object state $\mathbf{s}_t$ conditioned on all observations $\mathbf{z}_{1:t}$ up to time t by computing Bayes filter posterior $\mathbf{s}_t \sim \Gamma(\mathbf{s}_t \mid \mathbf{z}_{1:t})$. The posterior $\Gamma(\mathbf{s}_t \mid \mathbf{z}_{1:t})$ is approximated by sampling N particles at time t from its distribution. All particles at time $t-1$ move independently to the time t by sampling from the state transition distribution $\Gamma(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{z}_{t-1})$. Each i_th sample state $\mathbf{s}_t^{(i)}$ associates with a weight $w_t^{(i)}$, which depends on the probability of the observation state $\mathbf{z}_t$, given by $w_t^{(i)} \propto \Gamma(\mathbf{z}_t \mid \mathbf{s}_t^{(i)})$, called important factor. The probability of drawing each particle is given by its important weights. Hence, the fair particles $\mathbf{s}_t^{(i)}$ will transit to the new particles $\widehat{\mathbf{s}}_{t+1}^{(i)}$ after resampling.

2.2. Motion Model. In the pixel coordinate, we define the motion state, the position, and the velocity of the hand, at

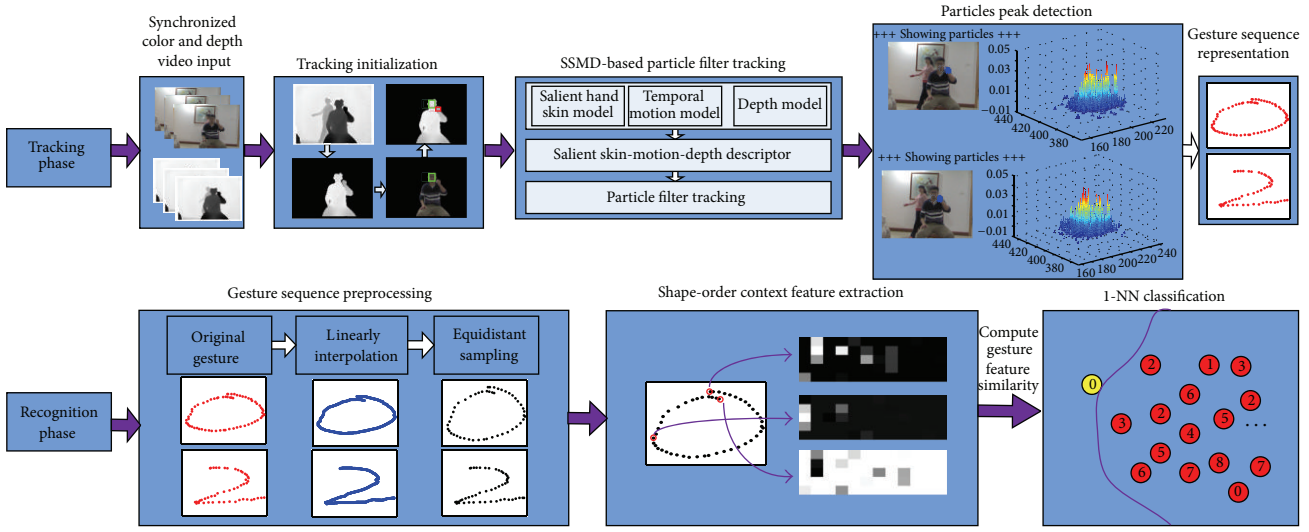


FIGURE 1: The flow diagram of RGBD based hand trajectory tracking and gesture recognition system.

time instant t as $\mathbf{d}_t = (p_t, v_t)^T$, $p_t = (x_t, y_t)^T$, and $v_t = (\hat{x}_t, \hat{y}_t)^T$, respectively. The motion state of bare hand can be established as a first order autoregressive dynamics model:

$$\mathbf{d}_t = f(\mathbf{d}_{t-1}) + \mathbf{e}_t; \quad \mathbf{e}_t \sim N(0,1), \tag{1}$$

where $\mathbf{e}_t$ denotes white processing noise with zero-mean, for each state variable.

3. Observation Model

3.1. Salient Hand Skin Model. Since human skin is a relatively uniform and discriminable cue in color space, a statistical color model can be employed to compute the probability of every pixel being skin color. A skin color distribution and a nonskin color distribution are denoted as $P(r, g, b \mid \text{skin})$ and $P(r, g, b \mid\sim \text{skin})$, respectively, in RGB color space, which is quantized to $32 \times 32 \times 32$ values. According to these two distributions, the probability of a pixel-level particle $\mathbf{s}_t^{(i)}$, which corresponds to pixel in color vector $[rgb]$, can be defined as skin particle by using Bayes rule:

$$P\left(\text{skin} \mid \mathbf{s}_t^{(i)}\right) = \frac{P\left(\mathbf{s}_t^{(i)} \mid \text{skin}\right) P(\text{skin})}{P(r, g, b)}, \tag{2}$$

where $P(r, g, b) = P(\mathbf{s}_t^{(i)} \mid \text{skin})P(\text{skin}) + P(\mathbf{s}_t^{(i)} \mid\sim \text{skin})P(\sim \text{skin})$ and $P(\text{skin})$ always equals 0.5.

A normalized skin probability of particle $\mathbf{s}_t^{(i)}$ is defined as

$$\Gamma_{\text{skin}}\left(\mathbf{s}_t^{(i)}\right) = \frac{P\left(\text{skin} \mid \mathbf{s}_t^{(i)}\right)}{\max\left\{P\left(\text{skin} \mid \mathbf{s}_t^{(i)}\right)\right\}_{i=1:K}}, \tag{3}$$

where K is particles number in each frame and we set $K = 2000$ overall our experiments.

To improve the effectiveness of the skin color detector, an optimization method based on skin saliency is proposed in order to enhance the particles. As for a local-level image region distributed by propagating particles, the contrast and the color of the tracking object are relatively unique and distinctive than the local background. Since the target hand can be concerned as the dominant object in the context of the local image, a particle $\mathbf{s}_t^{(i)}$ can be viewed as a salient particle if its appearance patch with centered at pixel location of $\mathbf{s}_t^{(i)}$ is relatively distinctive with respect to all other particle patches. Specifically, we represent each particles by cropping a patch (5×5 size) surrounding it. Let $d_{\text{color}}(\cdot)$ and $d_{\text{position}}(\cdot)$ be, respectively, the Euclidean distance of the vectorized patches of colors (in RGB space) and pixel location between two particle patches centered at $p_t(\mathbf{s}_t^{(i)})$, $p_t(\mathbf{s}_t^{(j)})$, respectively, where $p_t(\mathbf{s}_t^{(i)})$ and $p_t(\mathbf{s}_t^{(j)})$ are the pixel locations of the particles $\mathbf{s}_t^{(i)}$, $\mathbf{s}_t^{(j)}$. Based on the hand skin model, let us define a dissimilarity measure between a pair of particles as

$$d\left(\mathbf{s}_t^{(i)}, \mathbf{s}_t^{(j)}\right) = \frac{d_{\text{color}}\left(\mathbf{s}_t^{(i)}, \mathbf{s}_t^{(j)}\right)}{1 + c \cdot d_{\text{position}}\left(\mathbf{s}_t^{(i)}, \mathbf{s}_t^{(j)}\right)}. \tag{4}$$

This implies that a single pixel-level particle is treated as a salient hand skin particle when the other particles similar to it are nearby, and it is less salient when the resembling particles are far away. The parameter c set as 0.2 in our experiments. Hence, the salient hand skin factor of particle $\mathbf{s}_t^{(i)}$ is defined as

$$\omega^i\left(\mathbf{s}_t^{(i)}\right) = 1 - \exp\left\{-\frac{1}{K}\sum_{k=1}^{K} d\left(\mathbf{s}_t^{(i)}, \mathbf{s}_t^{(k)}\right)\right\}, \tag{5}$$

where K most similar patches of the particles $\{\mathbf{s}_t^{(k)}\}_{k=1:K}$ of each frame are extracted according to $d_{\text{color}}(\mathbf{s}_t^{(i)}, \mathbf{s}_t^{(k)})$, and we set $K = 100$ in the experiment. As mentioned above, a particle is identified as hand skin saliency when ω^i is high and vice versa.

3.2. Temporal Motion Model. Motion information is another useful cue for hand detection in gesture video. Since the hand is performed as the most salient object in front of the camera, hand motion changes more frequently than other human parts and objects background, such as arm and face. Typically, people incline to use the frame difference method to detect hand motion, whereas it is arduous to discriminate the objects with skin-like color close to that of the signed hand region. To improve the robustness of our tracking system, we use velocity of optical flow [22] as one of the motion cues to represent the direction and the magnitude of moving hand.

Denote $\mathbf{M}_u(p_t)$ and $\mathbf{M}_v(p_t)$ as the velocity of the x and y direction optical flow map at position p_t, respectively. And we define a patch associate with the location of a particle as patch center to represent the motion cue of each particle $\mathbf{s}_t^{(i)}$ on both directions of optical flow map:

$$M_u\left(\mathbf{s}_t^{(i)}\right) = \sum_x \sum_y \mathbf{M}_u\left(p\left(\mathbf{s}_t^{(i)}\right) + r_t W_{u\_\text{patch}}\right), \tag{6}$$

where W_{u_patch} is a priori fixed window patch through the definition of a 0-centered window with size of 50×60. Similarly, we can compute the motion cue of $M_v(\mathbf{s}_t^{(i)})$. Such patch size is empirically defined according to the face depth of performer and considered as the largest patch during tracking. The scale r_t denotes the patch size control factor needed to be estimated according to the depth value of window center. We set $r_t = \min(\text{hand\_depth}/\text{face\_depth}, 1)$ with the depth images values ranging from [0–2046] and value 2047 denotes an invalid depth pixel. It means that we control the maximum distance of hand to the camera not larger than the distance of face to the camera in our experiments. The similarity of the motion indicator between the current particle motion patches summation $M_{u,v}(\mathbf{s}_t^{(i)})$ and the pervious average motion velocity of particles $M^*_{(u,v),t-1}$ is defined as

$$\text{Dis}\left(\mathbf{s}_t^{(i)}\right) = \left\| \sqrt{M_{u,v}\left(\mathbf{s}_t^{(i)}\right)} - \sqrt{M^*_{(u,v),t-1}} \right\| \tag{7}$$

in which $M_{u,v}(\mathbf{s}_t^{(i)}) = M_u(\mathbf{s}_t^{(i)})^2 + M_v(\mathbf{s}_t^{(i)})^2$ and $M^*_{(u,v),t-1} = {M^*_{u,t-1}}^2 + {M^*_{v,t-1}}^2$.

The motion likelihood of the particle $\mathbf{s}_t^{(i)}$ is given by

$$\Gamma\left(M\left(\mathbf{s}_t^{(i)}\right)\right) = \frac{-1}{\sqrt{2\pi\sigma}} e^{-\text{Dis}(\mathbf{s}_t^{(i)})^2/2\sigma^2}, \tag{8}$$

where σ is a standard deviation and we empirically set it as constant 1.

The motion cue of each pixel is normalized as

$$\Gamma_{\text{motion}}\left(M\left(\mathbf{s}_t^{(i)}\right)\right) = \frac{\Gamma\left(M\left(\mathbf{s}_t^{(i)}\right)\right)}{\max\left\{\Gamma\left(M\left(\mathbf{s}_t^{(i)}\right)\right)\right\}_{i=1:K}}. \tag{9}$$

3.3. Depth Model. The probability $\Gamma_N(z_t^{(i)}, z^*_{t-1})$ explains the depth relationship between particle $\mathbf{s}_t^{(i)}$ with depth $z_t^{(i)}$ at current frame t and the minimum depth z^*_{t-1} across all particles at last frame $t-1$. Only if the depth variation involved in a small threshold will give a high probability of sampling that particle, we assume that the performed hand is closer to camera device than the connected forearm; hence, the depth variation of hand between two consecutive frames would not change too much and can be expressed by Gaussian distribution:

$$\Gamma_N\left(z_t^{(i)}, z^*_{t-1}\right) = \exp\left(-\kappa \left\| z_t^{(i)} - z^*_{t-1} \right\|^2\right), \tag{10}$$

in which κ is the speed controlling coefficient, which controls the speed of probability decreases with increasing depth difference. ($\kappa = 0.5$ in the experiment).

3.4. Salient Skin-Motion-Depth (SSMD) Descriptor. For each particle $\mathbf{s}_t^{(i)}$, the probability of observation $\Gamma(\mathbf{z}_t \mid \mathbf{s}_t^{(i)})$ can be computed by combining the salient skin indicator and motion indicator with depth indicator to detect the most hand-like pixel. The combined indicator probability of a particle is defined as follows:

$$\begin{aligned}\Gamma\left(\mathbf{z}_t \mid \mathbf{s}_t^{(i)}\right) = {} & \alpha\Gamma_{\text{skin}}\left(\mathbf{s}_t^{(i)}\right)\omega^i\left(\mathbf{s}_t^{(i)}\right) \\ & + \beta\Gamma_{\text{motion}}\left(M\left(\mathbf{s}_t^{(i)}\right)\right) + \gamma\Gamma_N\left(z_t^{(i)}, z^*_{t-1}\right).\end{aligned} \tag{11}$$

To find the empirical constants α, β, and γ, the sum of squares of the position errors between particles and annotated hand center of each frame is minimized:

$$\min \sum_{t=1}^{M} \sum_{i=1}^{N} \left\| p_t\left(\mathbf{s}_t^{(i)}\right) - \widehat{p}_t \right\|. \tag{12}$$

Since the particles positions are indirectly determined by the constants α, β, γ, we consider that the optimal parameters values can lead to the minimization value of (12). The optimal α, β, and γ are chosen iteratively in training gesture video with step-size 0.5 under the constrain that $\alpha + \beta + \gamma = 1$. In this gesture training video, people sit in front of the camera and randomly move hand involved in the region between the performer's face and camera. Thus, we set the empirical constant as 0.4, 0.3, and 0.3, respectively.

The weight $w_t^{(i)}$ of each particle $\mathbf{s}_t^{(i)}$ is calculated from the combined observation model $\Gamma(\mathbf{z}_t \mid \mathbf{s}_t^{(i)})$ based on (11). Thus, the desired posterior distribution $\Gamma(\mathbf{s}_t \mid \mathbf{z}_t)$ can be represented by the set of weighted particles $\{\mathbf{s}_t^{(i)}, \pi_t^i\}_{n=1}^N$. We summarize the SSMD-PF algorithm as follows.

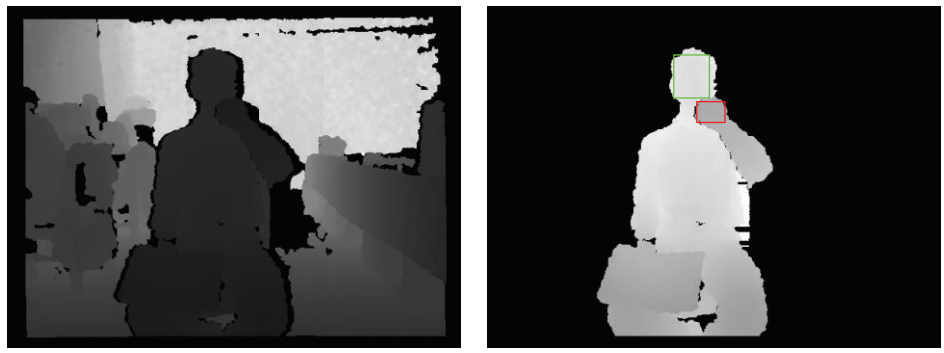
SSMD-PF Algorithm for Hand Tracking. Given the particle set $\{\mathbf{s}_{t-1}^{(i)}, w_{t-1}^i\}_{i=1}^N$, z^*_{t-1}, $M^*_{(u,v),t-1}$ at time $t-1$, perform the following steps.

(1) Resample a particle set $\{\mathbf{s}_t'^{(i)}, N^{-1}\}_{i=1}^N$ from the particles $\{\mathbf{s}_{t-1}^{(i)}, w_{t-1}^i\}_{i=1}^N$.

(2) Propagate each particle $\mathbf{s}_t'^{(i)}$ by the dynamic model in (1) to give $\{\mathbf{s}_t^i, N^{-1}\}_{i=1}^N$.

(3) Weight each particle by combining saliency skin, motion, and depth cues in (11): $w_t^{(i)} \propto \Gamma(\mathbf{z}_t \mid \mathbf{s}_t^{(i)})$ which are normalized so that $\sum_{i=1}^N w_t^{(i)} = 1$.

(4) Estimate the target state of depth and motion cues z_t^* and $M^*_{(u,v),t}$ from $\{\mathbf{s}_t^{(i)}, w_t^i\}_{i=1}^N$.



(a) A depth frame with signer ready for signing gesture

(b) The salient signer body with detected face in green rectangle and hand in red rectangle

Figure 2: Hand tracking initialization.

3.5. Tracking Initialization. A major shortcoming of pervious tracking methods is the need for supervised initialization. Furthermore, manual reinitialization is needed if the tracking drifts or fails. To overcome this limitation, we also proposed an efficient and fast tracking initialization method which can automatically localize the hand, irrespective of their pose, scale, and rotation.

In our hand tracking system, in order to accurately and explicitly express the hand, we assume that the performed hand is comparatively closer to the camera than the corresponding arm. Since the whole signer body can be considered as the most salient object in the foreground of each frame, the connected component of human body can be computed as follow:

$$C_{\max} = \arg\max_{\{C_i\}_{i=1:N}} \left\{ \frac{C_i}{C_i^Z} \right\}, \tag{13}$$

where C_i denotes the i_th depth connected component from all sets $\{C_i\}_{i=1:N}$ and C_i^Z denotes the average depth of C_i. A depth connected component which represents the performer body can be considered as the one with larger connected component size and with smaller average depth, as shown in Figure 2. Before using (13), each C_i should be roughly divided into the background or foreground by comparing C_i^Z with a depth threshold $T_r = 2000$. When $C_i^Z > T_r$, the corresponding connected component C_i is considered as background object and removed from $\{C_i\}_{i=1:N}$; otherwise it will be considered as the foreground object and wait for computing in (13).

Before performing a gesture, an actor is asked to posture the hand in the upper part of the actor body, as shown in Figure 2(a). According to this starting state, we divide the extracted $C_{\max}$ into upper and lower parts and select the upper part pixels by comparing each vertical location of pixel in $C_{\max}$ with the threshold T. Finally, the pixel with minimum depth value in the upper part is selected as the initial hand position, as shown in Figure 2(b). In our experiments, we set $T = v\_face\_center + 0.72 * (480 - v\_face\_center)$, where v_face_center is the vertical positon of the face center and the actor face can be easily detected by using Viola-Jones method [23] in color image.

3.6. Hand Center Localization. To create the hand trajectory, for simplicity, hand center should be localized to represent the entire hand. Kernel density estimation (KDE) is a nonparametric density estimation method which can be used to estimate the probability density function of particles. Given the particles distribution of each frame, the KDE is

$$P_{\mathrm{KDE}}\left(p_t\left(\mathbf{s}_t^{(n)}\right)\right) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{p_t\left(\mathbf{s}_t^{(n)}\right) - p_t\left(\mathbf{s}_t^{(i)}\right)}{h}\right), \tag{14}$$

where $K(\cdot)$ is the Gaussian kernel function and $p_t(\mathbf{s}_t^{(i)})$, $p_t(\mathbf{s}_t^{(n)})$ are the locations of particles $\mathbf{s}_t^{(i)}$, $\mathbf{s}_t^{(n)}$; h is bandwidth parameter to be set as $h = 1.06\sigma N^{-1/5}$, where $\sigma = 1$ is the standard deviation of particles and N is the number of particles. Then we choose the 2D position corresponding to the peak value of the KDE as the hand center, as shown in Figure 3.

4. Hand Gesture Recognition

After hand tracking phase, each gesture can be represented by a set of hand locations frame by frame. The start and end frames of the gesture are manually annotated in this paper. The top line of Figure 4 shows the representation of the original spatiotemporal gestures. In this section, we first describe how to preprocess the original gesture sequence. And then a naïve shape descriptor, which is called shape-order context, is introduced and we also talk about how to match two gestures using the proposed shape descriptor. At last, we compare the computational complexity between the proposed shape descriptor and the well-known shape descriptor: shape context.

4.1. Gesture Sequence Normalization. Typically, different gesture sequences have different lengths. However, for the same gesture sequences, the span of every two adjacent gesture points may be different due to the variation of the sign speed. For the peer to peer points matching, we need to normalize each trajectory sequence so that they share the same point numbers. Specifically, gesture trajectories are

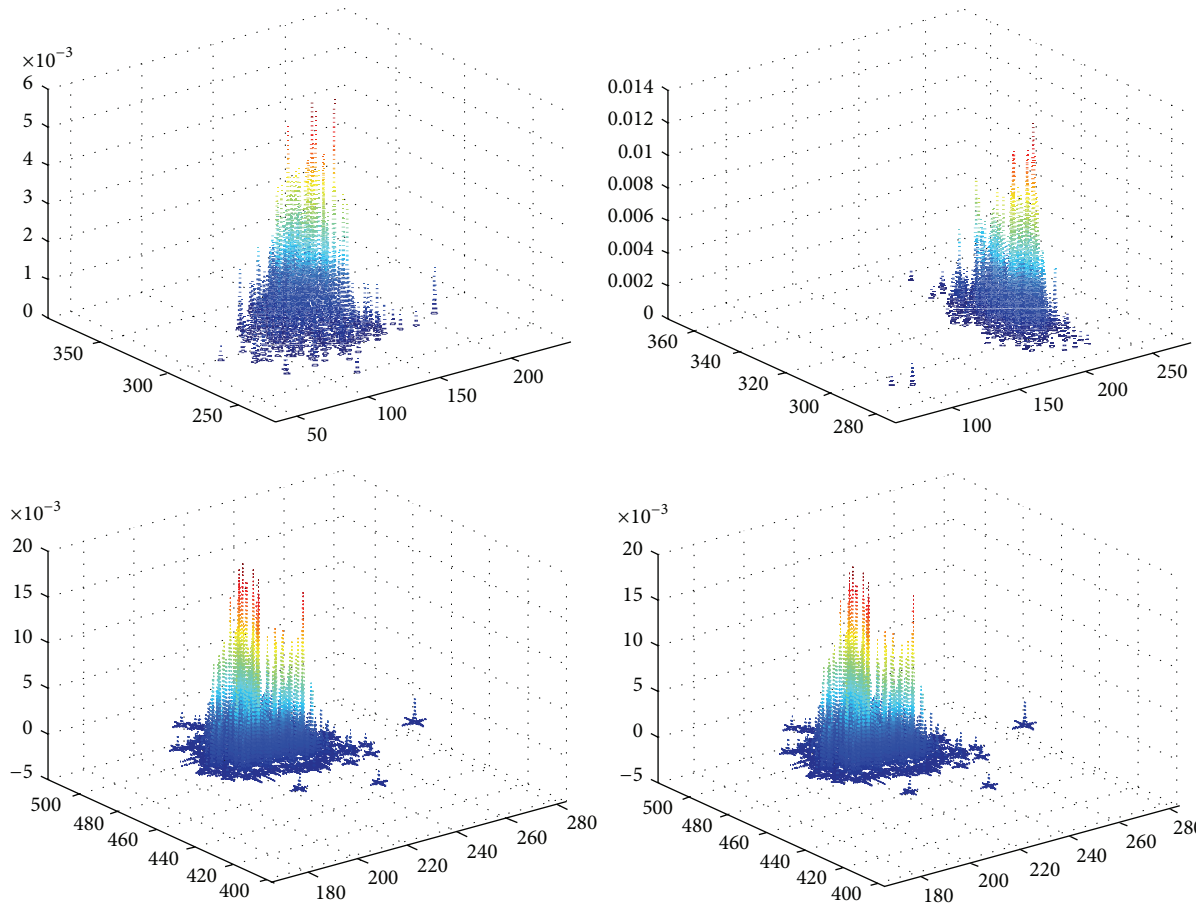


Figure 3: The Kernel density estimation of particles distribution in frames 90, 100, 130, and 180 of digital gesture "0."

linearly interpolated between every two adjacent original hand locations; then M locations are equidistantly extracted from all gesture points, so that each sequence has the same length, as shown in the middle and bottom row of Figure 4, respectively. In our method, this sequences normalization step is significantly important because the shape-order context descriptors matching requires that each gesture sample has the same length, and we need to match the corresponding shape-order context descriptors of pairwise points in both training and testing data.

4.2. Shape Context. Shape context [21] is a descriptor which expresses the object shape by considering the relationship of the set of vectors originating from a point to all other sample points on a shape. Obviously, the full set of vectors as shape descriptors contains much too details since it configures the entire shape relative to the reference points. This set of vectors is identified as a highly discriminative descriptor which can represent the shape distribution over relative positions. The shape context of p_i is defined as a coarse histogram h_i of the relative coordinates of the remaining $n-1$ points:

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in \text{bin}(k)\}. \quad (15)$$

The bins are uniform in log-polar space, making the descriptor more sensitive to positions of nearby sample points than to those of points farther away.

4.3. Our Approach: Shape-Order Context. A human hand-dominated gesture trajectory can be simplified as a set of hand location features in spatial-temporal domain. For the traditional shape context method, the object is represented only in spatial domain. As a normalized gesture sequence with length L, we improve the shape context method by computing the relationship of a trajectory point to all other trajectory points in spatial-temporal domain. Specifically, each k bin of the coarse histogram $h_i(k)$ of the relative log-polar coordinate is established by accumulating sequence order difference between gesture points p_i and remaining gesture points $q = \cup_j p_j$ which is involved in the relative region:

$$h_i(k) = \sum_{j \in J; i \neq j} (j - i) \quad \text{subject to } (q - p_i) \in \text{bin}(k), \quad q \neq p_i, \quad (16)$$

where J contains the subscripts of gesture points that belong to q. In this way, the temporal information of gesture trajectory is embedded into log-polar space bins, as shown

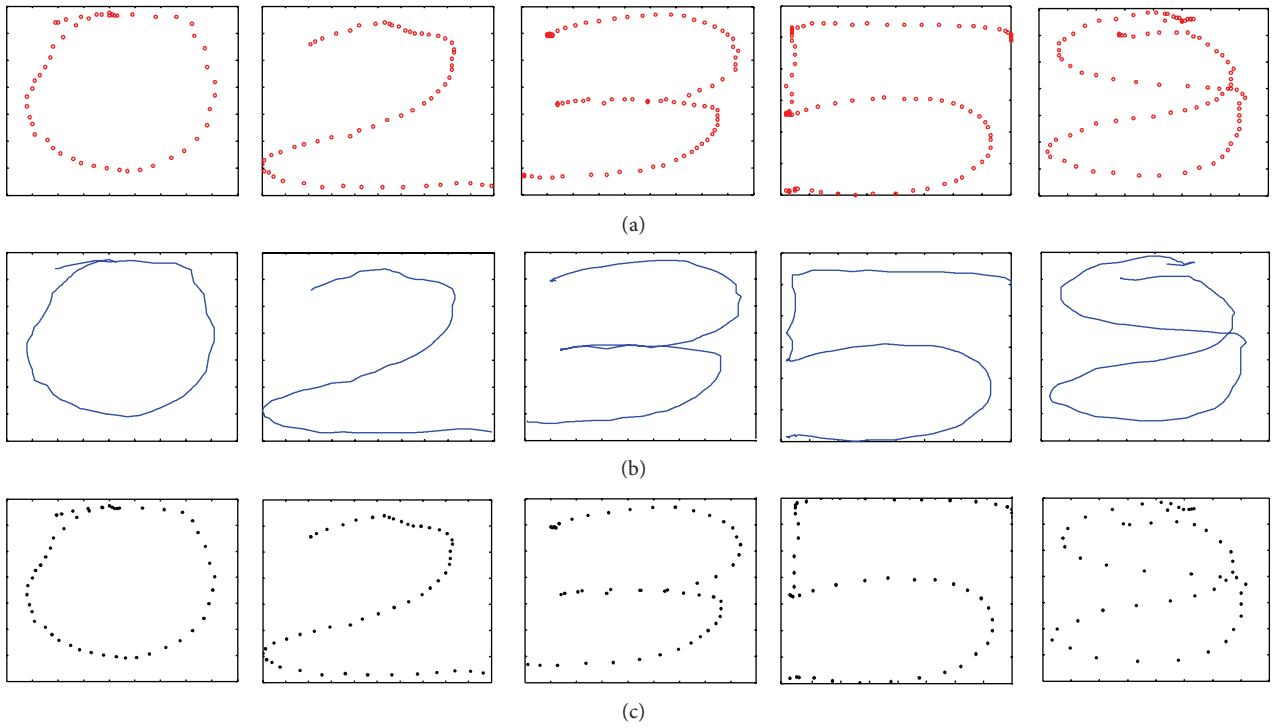


Figure 4: Hand digital gesture sequence normalization. (a) Original gesture sequence. (b) Gesture sequence with linear interpolation. (c) Gesture sequence with equidistant sampling.

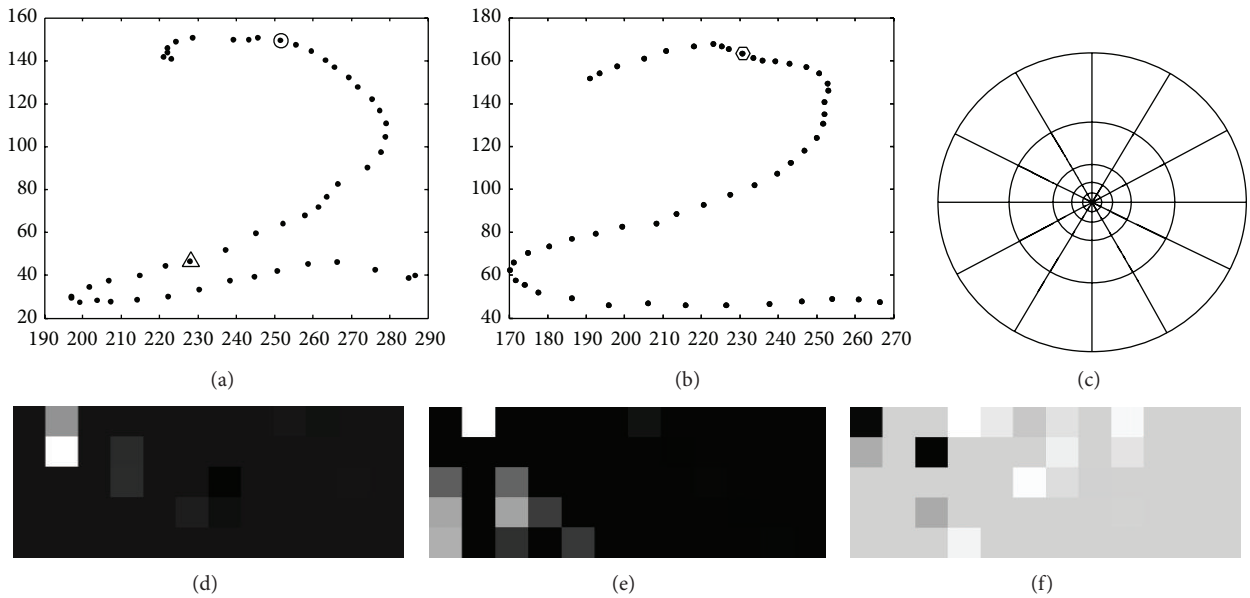


Figure 5: Shape-order context descriptor computation and matching. (a) and (b) Normalized trajectory of two gestures. (c) Diagram of log-polar histogram bins used in computing the shape-order contexts. (d), (e), and (f) Example of shape-order contexts for three reference samples. Note the visual similarity of (d) and (e) is much higher than (f).

in Figure 5. Therefore, the value of each bin is dominated not only by the distance and the angle relation in gesture spatial domain, but also by impact by the sequence order relation in gesture temporal domain. Since the descriptors of shape-order context contain rich spatial-temporal information for each point, they are inherently insensitive to small perturbations which are produced by different performers as shown in Figure 6.

To build a robust gesture recognition system, translation and scale invariance are highly desirable. For our shape-order

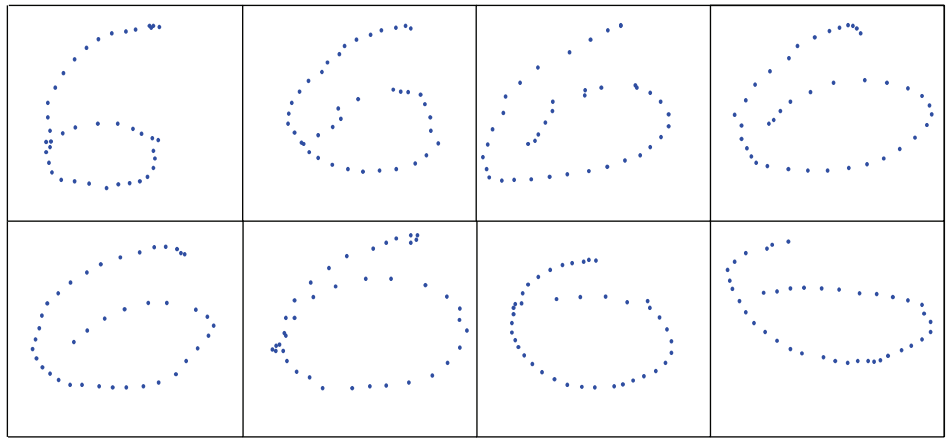


Figure 6: Gesture digital "6" under different performers.

context approach, invariance under translation is intrinsically existing since all measurements are taken with respect to points on the gesture trajectory.

To achieve scale invariance, we normalize all the gesture trajectory sets by means of calculating minimum enclosing circle (MEC) of those sets and then resize each set to the same circle's radius. As for rotation invariance, since the shape-order contexts are extremely rich descriptors and inherently insensitive to small rotation and perturbations of the shape, we use the absolute image coordinates to compute the shape-order context descriptor for each point. Another reason for using the absolute image coordinates is that for two gesture trajectories even it achieves the same appearance after rotating; however, the complete rotation invariance impedes expressing the original meaning of the gesture. Hence, it should be emphasized that the completely rotation invariance of the gesture trajectory shape is not suitable for gesture recognition.

4.4. Final Classification. Due to the fact that each of the gesture sequence point is represented as histogram based on shape-order context, it is natural to use the χ^2 test statistic to compute the histogram similarity of pairwise (p_i, q_i) in the two corresponding sequence positions. Let $C_{ij} = C(p_i, q_i)$ denote the cost of matching two corresponding points:

$$C_{ij} = \frac{1}{2}\sum_{k=1}^{K}\frac{\left[h_i(k) - h_j(k)\right]^2}{h_i(k) + h_j(k)}, \tag{17}$$

where $h_i(k)$ and $h_j(k)$ denote the K-bin normalized histogram at p_i and q_i, respectively. Thus, the cost C_{ij} of matching pairwise points not only includes the local appearance similarity but also contains local gesture order similarity, which is particularly useful when comparing the trajectory shapes. Finally, the similarity between two gestures is computed by accumulating each matching cost and using 1-NN nearest neighbor classification to determine which class the gesture belongs to.

4.5. Computational Complexity. The time complexity of the shape-order context matching algorithm can be measured as shown here. Let m be the number of gesture points after gesture normalization. Let r be the number of radial bins and let a be the number of angular bins. The time complexity of computing a gesture point histogram is $n = r * a$. The complexity of matching histogram in our shape-order context method is $O(Pmn)$, where P is the number of gestures in the training dataset and mn is the feature vector size containing m points in each gesture trajectory. Our proposed gesture recognition method has the same computational complexity as the original shape-order context algorithm. However, it has relatively higher recognition rate as shown in the following section.

Figure 7: "0–9" digital gesture exemplar in training sets.

5. Experiments and Results

In this section we describe the details of our proposed hand tracking and gesture recognition experimental results and analysis. All experiments were conducted on a 3.30 GHz based Windows PC with 8 GB of RAM.

5.1. Datasets. Hand-signed digit datasets (HSD) [24] is a commonly used benchmark hand gesture dataset for gesture recognition with 10 categories performed by 12 different people. In more detail these datasets have been organized as follow.

Training Examples. 300 digit exemplars with 30 per class were stored in the database. A total number of 10 users wearing gloves have been employed to collect all training data. These video clips are captured by using a typical color camera at 30 Hz with an image size of 240 × 320. Each digit gesture video clip depicts a user gesturing ten digits in sequence, as shown in Figure 7.

Test Examples. We recapture 440 digit exemplars with 44 per class used as queries. 400 exemplars contain distracters, in

Table 1: Frame length of the selected ASL video clip.

	DS1	DS2	DS3
# of frame	1315	3300	3200

Table 2: Comparison accuracy by varying parameters values of (11).

(α, β, γ)	(1, 0, 0)	(0, 1, 0)	(0, 0, 1)	SSMD-PF (0.4, 0.3, 0.3)
HSD easy data	95.0	96.6	98.5	100
HSD hard data	63.1	85.4	97.9	100

the form of a human moving back and forth in the background and a skin-like object closing to the sign hand. The rest of video clips are captured in prune background without dramatic motion. These video clips are captured by using kinect camera at 30 Hz with an image size of 480 × 640. The corresponding depth and color frame are already well calibrated by using OpenNI platform.

American sign language dataset (ASL) [25] is an ongoing work which contains video sequences of thousands of distinct ASL signs with prune background. We conducted our hand tracking experiments in a user independent manner using three of the sign language video datasets. The detail of these videos can be found in Table 1.

5.2. Hand Tracking Evaluation. To verify the effectiveness of the proposed observation model, we run the algorithm on some sequences by varying the ratio of (α, β, γ) in (11). The tracking on each frame is considered correct if

$$\frac{\text{area}\,(B \cap G)}{\text{area}\,(B \cup G)} > 0.5, \tag{18}$$

where B is the detection hand area and G is the ground truth hand area which is acquired by manual annotation. For comparison, the tracking is implemented by using saliency skin model, temporal motion model, and depth model, respectively. The variation of parameters values is shown in Table 2. It is observed that, by using each single model on the HSD easy dataset, the tracker has a relatively convincing tracking accuracy as SSMD-PF method does. However, on the HSD hard dataset, the accuracy of trackers with $(1, 0, 0)$ and $(0, 1, 0)$ parameters are dropped rapidly due to the confusion caused by moving people in the background, whereas the trackers with depth model or proposed SSMD-PF model successfully handle the distraction and our proposed model with parameters $(0.4, 0.3, 0.3)$ is much more robust and reliable than other three independent models.

We then assess the proposed SSMD-PF algorithm's performance in comparison to the several conventional PF methods. Evaluation is done on the HSD datasets and the selected ASL video clips, respectively, as shown in Tables 3, 4, and 5. In Table 3, our method achieves 100% accuracy on the HSD easy datasets, which is the same with the state-of-the-art methods in [11, 16]. It indicates that the pure and stable background can greatly improve the tracking accuracy. In Table 4, as expected, the tracking accuracy of SSMD-PF method achieves 100% on the HSD hard datasets, which is higher than [16, 26], which also uses the color and depth cue as tracking feature. The main reason of our method outperforming others is that of the use of the local minima hand depth cue as one of observation models in particle filter, which is strong enough to get rid of disturbing by other skin-like objects. The visualization of SSMD-PF based tracking algorithm in clean and complex background is shown in Figure 8. Figure 9 shows the tracking of digital gesture "3" with skin-like object close to that of the signed hand in the relatively same depth level. As we can see, the sign hand is not disturbed by such skin-like closing object because our system can discriminate it by taking advantage of the velocity as motion cue in observation model. Due to the lack of using the depth information, the other methods [11, 27] easily lose tracking, thus without showing tracking accuracy. In view of the fact that the ASL datasets only contain color information; we combine our salient skin and motion model (SSM-PF) to compare with other method in order to evaluate our color based on tracking method, as shown in Figure 10. In Table 5, we observe that our proposed method achieves better or comparable tracking accuracy with the Shan's method [11] and has much better performance than Zhang's [5], Perez's [26], and Mittal's method [6].

To make an intuitive comparison, we also provide a visualization of gesture digital tracking from HSD hard dataset as shown in Figure 11. Each 2D hand points in the pixel coordinate are transformed to 3D depth camera coordinate in advance based on compute camera intrinsic parameter [28]. Since the digital gestures are signed without dramatically depth variation, the gesture trajectory can be well fitted with ground truth.

Table 3: HSD easy datasets.

Methods	Accuracy (%)
SSMD-PF	100
SSM-PF	100
Van Den Bergh and Van Gool [16]	100
Shan et al. [11]	100
Pérez et al. [27]	95.2

Table 4: HSD hard datasets.

Methods	Accuracy (%)
SSMD-PF	100
Doliotis et al. [26]	95.32
Van Den Bergh and Van Gool [16]	95.38
Shan et al. [11]	—
Pérez et al. [27]	—

Table 5: ASL datasets.

Methods	Accuracy (%)
SSM-PF	91.7
Shan et al. [11]	90.3
Zhang et al. [5]	79.3
Mittal et al. [6]	40.2
Pérez et al. [27]	65.9

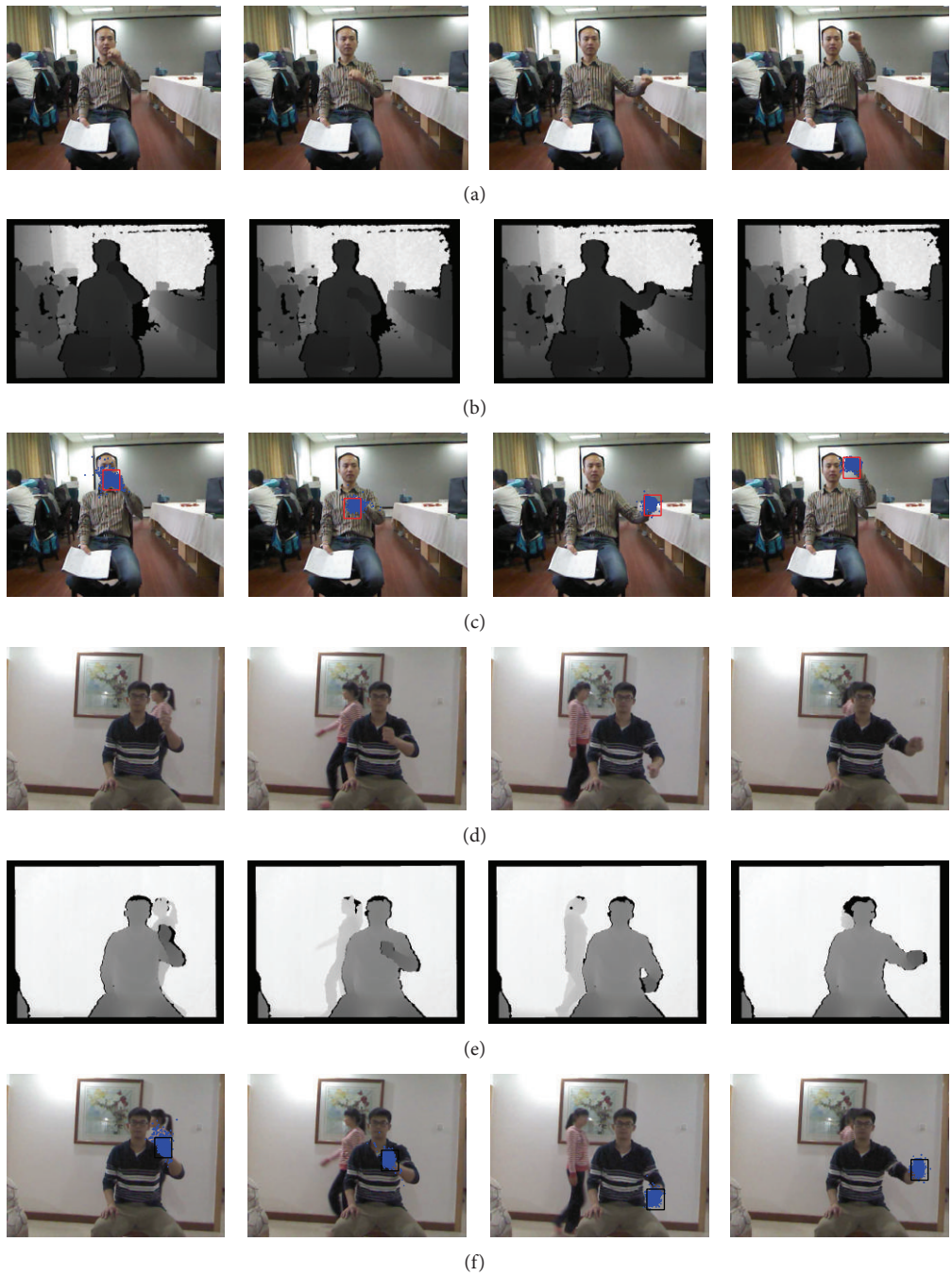


Figure 8: Digital gesture sequence tracked by SSMD-PF in clean and complex background, respectively. (a) and (d) Original frame (frames 90, 100, 130, and 160 in the left; frames 19, 40, 85, and 101 in the right). (b) and (e) The corresponding depth frame. (c) and (f) The tracking result with particles distribution in blue color and the hand bounding box in red (gray) color.

5.3. Gesture Recognition Evaluation. For gesture recognition, we evaluate our proposed method on the HSD datasets. The shape-order context method is firstly evaluated by changing the sampling point number from each interpolated gesture sequence. As shown in Table 6, with varying the sampling number, the proposed method remains stable and high classification accuracy over 98.4% with varying the sampling number from 20 to 80, and achieves 98.6% with 40 sampling numbers. It shows overall better performance than one dollar method [19] and Liu's method [20] which also implement the sequence normalization with varying sequence length.

The results of the false positive rate (FPR) and the false negative rate (FNR) on the HSD datasets (40 sampling numbers, 440 testing gesture sequences, and 300 training gesture sequences) are illustrated in Table 7. It shows that both the average FPR (with 0.0526) and FNR (with 0.0066) of

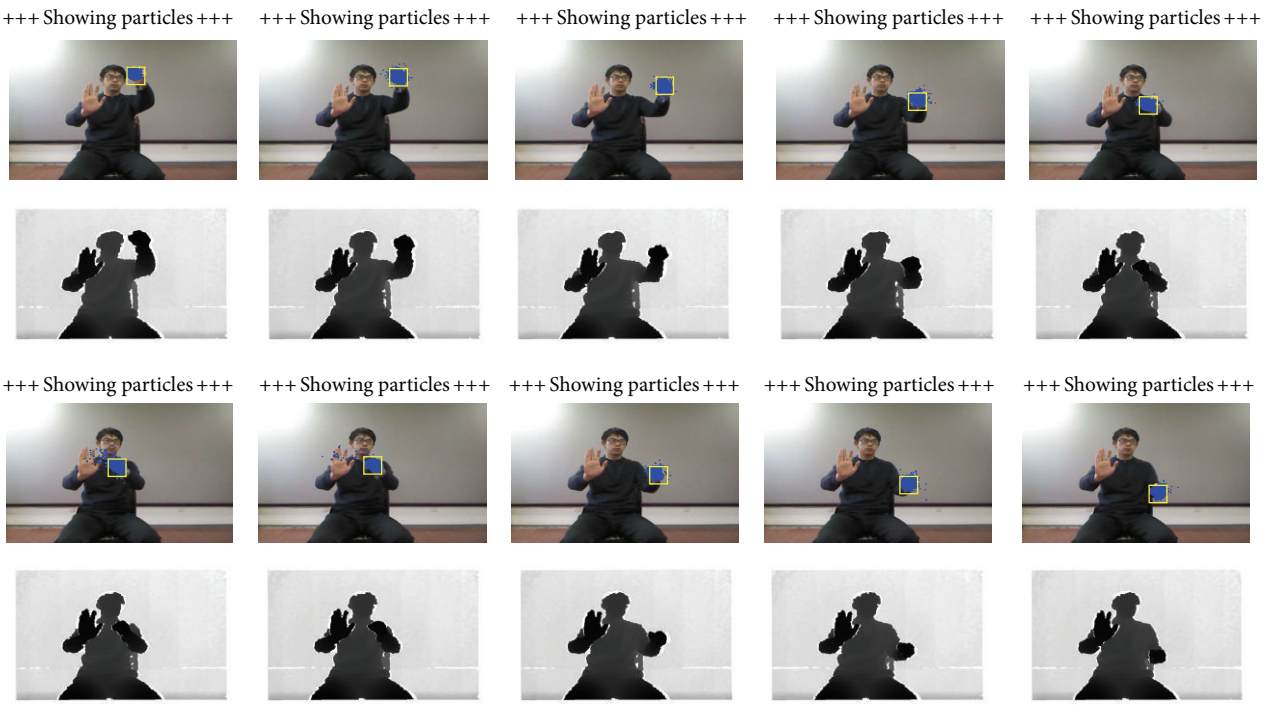


Figure 9: SSMD-PF based tracking on digital "3" with skin-like object close to that of signed hand.

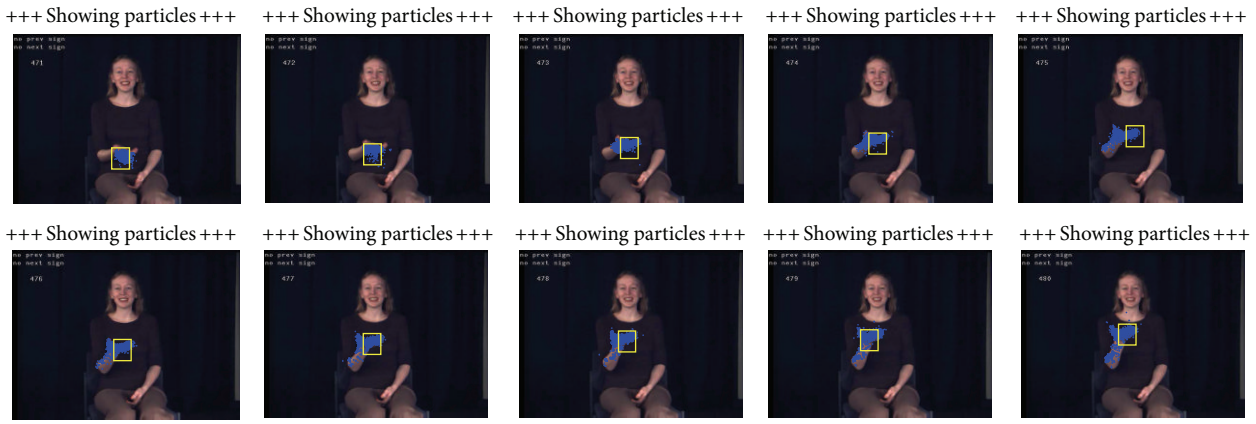


Figure 10: Tracking result of the consecutive sequence on the ASL dataset.

Table 6: Classification accuracy versus sequence normalization length on 440 HSD datasets.

Sampling number	20	30	40	50	60	70	80
Classification rate (%)							
Proposed method	98.4	98.4	98.6	98.4	98.4	98.4	98.4
Liu's method [20]	94.6	96.8	95.5	93.2	93.2	91.6	89.8
One dollar [19]	88.2	89.1	89.1	90.2	91.6	90.2	91.4

proposed method are lower than Liu's method and one dollar method. Moreover, both FPR and FNR of our method are not much fluctuating on each digital sign.

We also compare our approach to other state-of-the-art methods with varying the training size on the HSD datasets. We use the testing set with 440 numbers of gesture sequences which are captured by using our proposed SSMD-PF method. The total training size contains 300 meaningful gesture sequences with 30 gesture sequences per class. There are 10 classes in total. To vary the training size, we increase the amount of the digital gesture examples in each class as $[4, 5, 10, 15, 20, 25, 30]$. It is worth noticing that each of the subsequent training set contains all pervious training set and gradually increases until covering all 300 training data. As shown in Figure 12, all the classification accuracy can be greatly impacted by increasing training data and the performance of our approach outperforms other method in overall training size, which again validates the effectiveness of the shape-order context method in temporal-spatial domain.

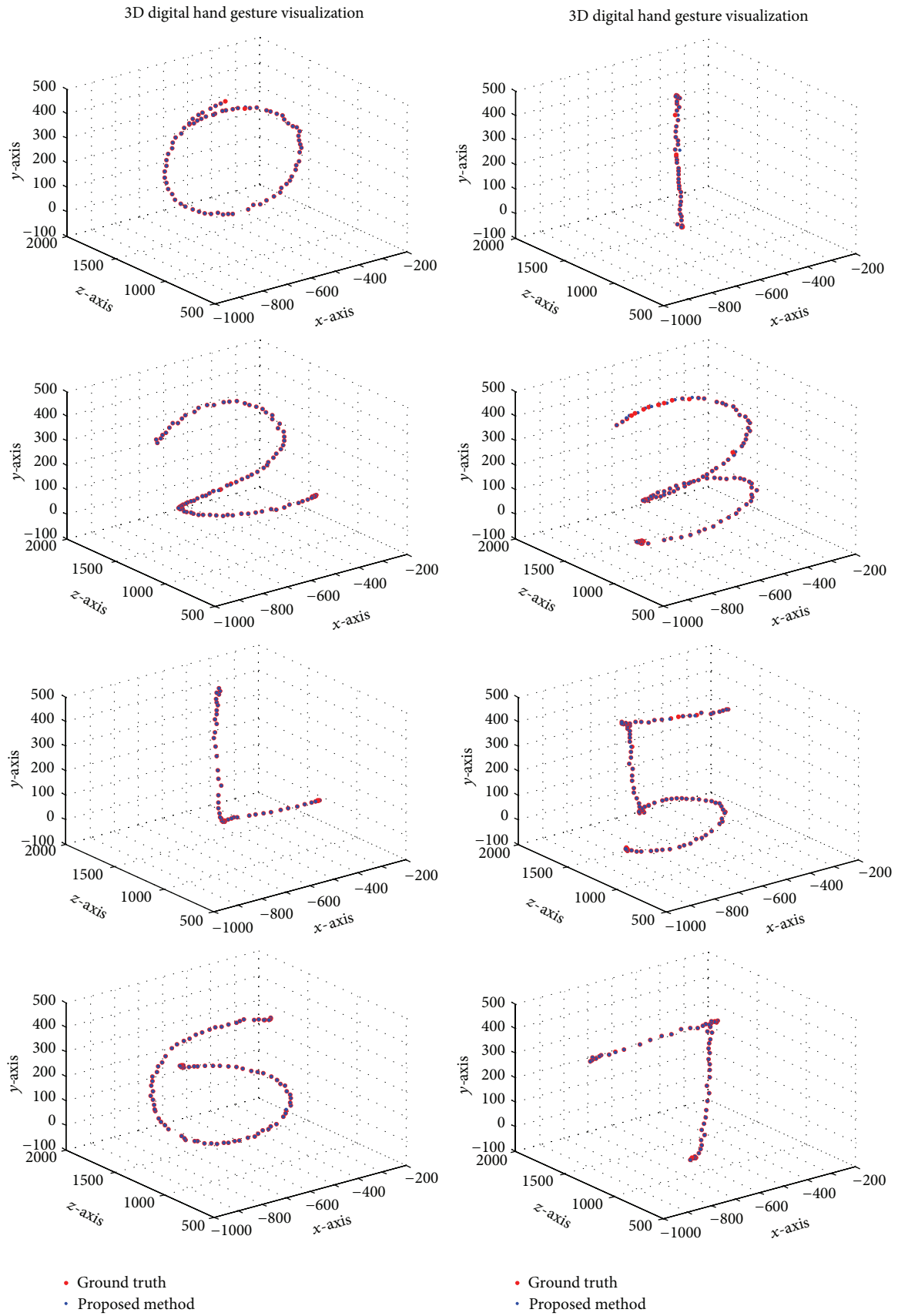


Figure 11: Continued.

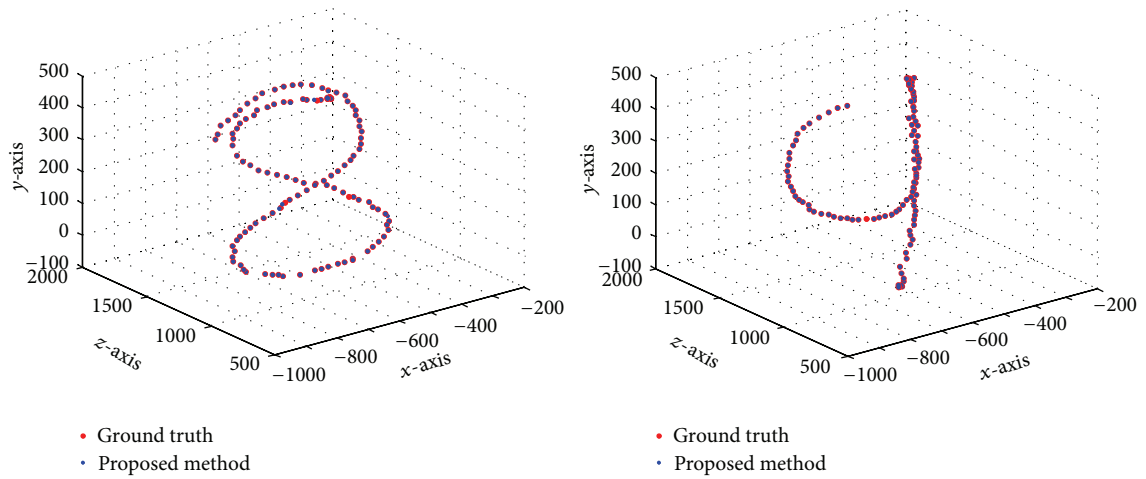


Figure 11: Visualization trajectory matching based on the HSD hard dataset.

Table 7: Comparison of the false positive rate and the false negative rate on each hand digital sign.

Digital gesture	0	1	2	3	4	5	6	7	8	9
Proposed method										
FP (%)	0.052	0.074	0.052	0.0519	0.04	0.052	0.052	0.048	0.052	0.052
FN (%)	0	0	0	0	0.033	0	0	0.033	0	0
Liu's method [20]										
FP (%)	0.51	0.51	0.51	0	0	0.25	0.51	1.26	0	0
FN (%)	2.27	9.09	0	0	6.82	0	4.55	4.55	0	4.55
One dollar [19]										
FP (%)	0.048	0.170	0.048	0.033	0.070	0.056	0.052	0.033	0.052	0.048
FN (%)	0.033	0.033	0.067	0.467	0	0.033	0	0.133	0	0.067

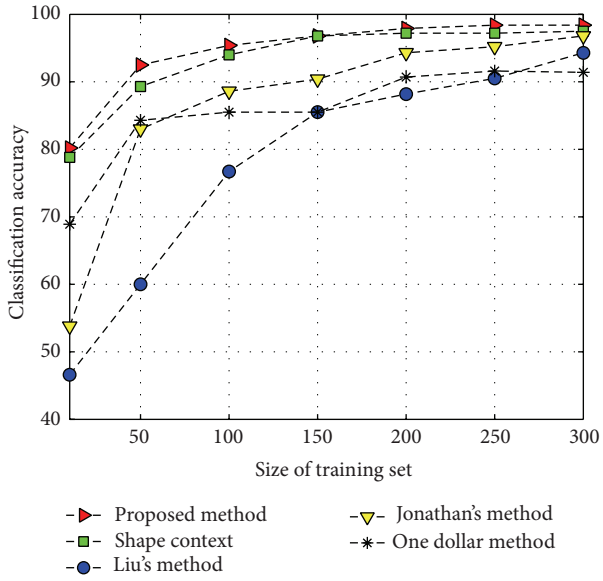


Figure 12: The classification accuracy versus training size.

Figures 13(a) and 13(b) represent a considerable improvement of recognition accuracy with changing gesture scale and translation, respectively. To apply a certain amount of translation to the input trajectory sequences, we add a set of small increments in the pixel unit ([10, 20, 30, 40, 50, 60, 70, 80, 90]) to the x and y coordinates of the position of each gesture point. To apply a certain amount of scaling to the input gestures, we multiply the x and y coordinates of each gesture point by a set of small increments in the pixel unit ([1.1, 1.2, 1.3, 1.4, 1.5, 1.6]). With gradually increasing scale and translating factors, the classification accuracy of dynamic time warping (DTW), Zhou's method, and Jonathon's method is rapidly dropped below 70%; however, the proposed method, Liu's method, and one dollar method still remain a stable accuracy without change. Moreover, our method achieves better performance with 1.6% higher than Liu's method. This advantage owes to the shape-order context descriptor which is invariant to several common transformations, which make it possible to recognize similarity gesture even with slight appearance variation.

6. Conclusion

We present a novel state-of-the-art hand tracking and gesture recognition system, respectively. By combining the enhanced skin, motion, and depth feature in particle filter model, the performing hand can be well localized and tracked in

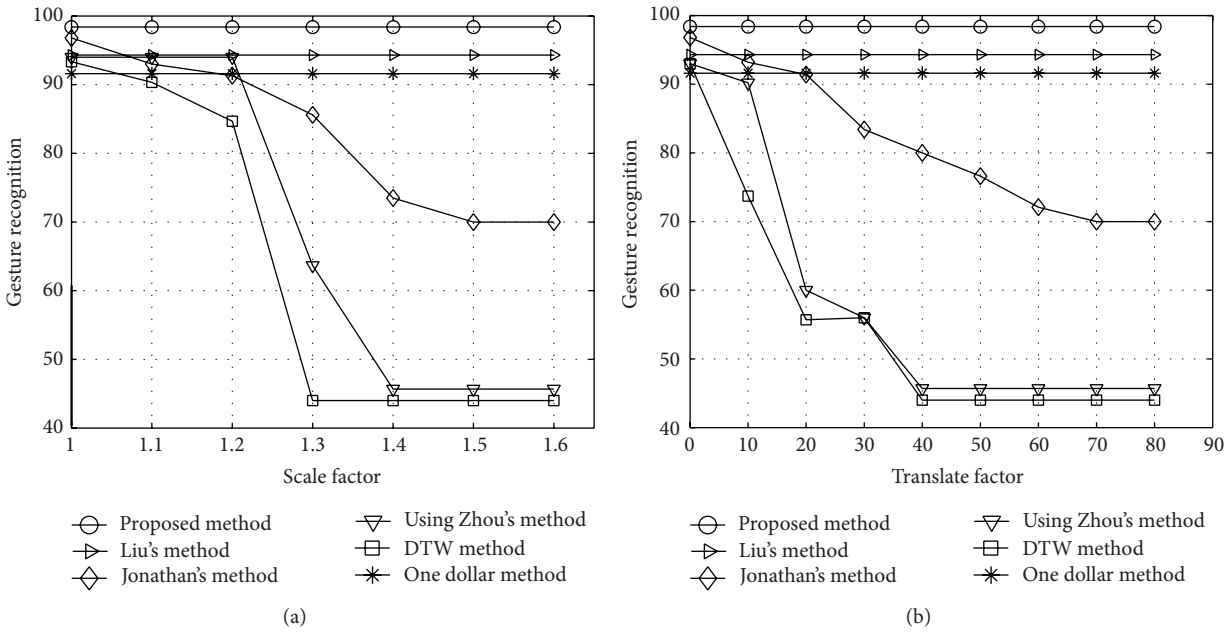


Figure 13: Classification accuracy versus gesture scale and translation. (a) The x-axis is the scale factor, which expresses the increase of multiple of hand location in the x and y coordinates, applied to both the x and y dimensions of the test sequences. (b) The x-axis is the translation factor, which expresses the pixel translation increment of hand location, applied to both the x and y dimensions of the test sequences.

every frame. We also introduce a fast and simple tracking initializing method for fully automatically tracking. A shape-order context descriptor is then proposed for gesture sequence matching in temporal spatial domain. Such a rich descriptor can greatly improve the gesture recognition rate and be invariant to gesture to translate and scale. In the future work, we will explore more sophisticated features for more advanced tracking, such as the hand attaching on object and hand-arm parallel moving. Also, the hand appearance will be considered to embed into shape-order context descriptor for more robust gesture recognition.

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 National Natural Science Foundation of China under Grant 61202314 and by the National Science Foundation for Post-Doctoral Scientists of China under Grant 2012M521801.

References

[1] R. Y. Wang and J. Popović, "Real-time hand-tracking with a color glove," *ACM Transactions on Graphics*, vol. 28, no. 3, article 63, 2009.

[2] K. Mitobe, T. Kaiga, T. Yukawa et al., "Development of a motion capture system for a hand using a magnetic three dimensional position sensor," in *Proceedings of the ACM SIGGRAPH Research Posters*, vol. 102, Boston, Mass, USA, Augast 2006.

[3] C. P. Chen, Y. T. Chen, P. H. Lee, Y.-P. Tsai, and S. Lei, "Real-time hand tracking on depth images," in *Proceedings of the IEEE Visual Communications and Image Processing (VCIP '11)*, pp. 1–4, Tainan City, Taiwan, November 2011.

[4] H. Nanda and K. Fujimura, "Visual tracking using depth data," U.S. Patent No. 7590262, September 2009.

[5] Z. Zhang, R. Alonzo, and V. Athitsos, "Hand detection on sign language videos," in *Proceedings of the International Conference on PErvasive Technologies Related to Assistive Environments.(PETRA'14)*, p. 21, Rhodes, Greece, May 2014.

[6] A. Mittal, A. Zisserman, and P. H. S. Torr, "Hand detection using multiple proposals," in *Proceedings of the British Machine Vision Conference (BMVC '11)*, 2011.

[7] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman, "The chains model for detecting parts by their context," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 25–32, San Francisco, Calif, June 2010.

[8] H. Cooper and R. Bowden, "Large lexicon detection of sign language," in *Proceedings of the International Conference on Human-Computer Interaction (HCI '07)*, pp. 88–97, Beijing, China, 2007.

[9] A. Farhadi, D. Forsyth, and R. White, "Transfer learning in sign language," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[10] V. Spruyt, A. Ledda, and W. Philips, "Real-time, long-term hand tracking with unsupervised initialization," in *Proceedings of the*

20th IEEE International Conference on Image Processing (ICIP '13), pp. 3730–3734, Melbourne, Australia, September 2013.

[11] C. F. Shan, T. N. Tan, and Y. C. Wei, "Real-time hand tracking using a mean shift embedded particle filter," *Pattern Recognition*, vol. 40, no. 7, pp. 1958–1970, 2007.

[12] Z. W. Wang, X. K. Yang, Y. Xu, and S. Yu, "Camshift guided particle filter for visual tracking," in *Proceedings of the international conference on Signal Processing Systems*, pp. 301–306, Shanghai, China, October 2007.

[13] W.-Y. Chang, C.-S. Chen, and Y.-D. Jian, "Visual tracking in high-dimensional state space by appearance-guided particle filtering," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1154–1167, 2008.

[14] M. Morshidi and T. Tjahjadi, "Gravity optimised particle filter for hand tracking," *Pattern Recognition*, vol. 47, no. 1, pp. 194–207, 2014.

[15] C. Manders, F. Farbiz, J. H. Chong et al., "Robust hand tracking using a skin tone and depth joint probability model," in *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG '08)*, pp. 1–6, Amsterdam, The Netherlands, September 2008.

[16] M. Van Den Bergh and L. Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV '11)*, pp. 66–72, Kona, Hawaii, USA, January 2011.

[17] A. Jonathan, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1685–1699, 2009.

[18] F. Zhou and F. De la Torre Frade, "Canonical time warping for alignment of human behavior," in *Advances in Neural Information Processing Systems Conference (NIPS)*, December 2009.

[19] J. O. Wobbrock, A. D. Wilson, and Y. Li, "Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes," in *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*, pp. 159–168, 2007.

[20] W. Liu, Y. Fan, T. Lei, and Z. Zhang, "Human gesture recognition using orientation segmentation feature on random forest," in *Proceedings of IEEE China Summit & International Conference on Signal and Information Processing (SIP '14)*, pp. 480–484, Xi'an, China, July 2014.

[21] 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.

[22] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, 1981.

[23] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[24] http://vlm1.uta.edu/~athitsos/projects/digits/.

[25] V. Athitsos, C. Neidle, S. Sclaroff et al., "The American sign language lexicon video dataset," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[26] P. Doliotis, A. Stefan, C. McMurrough et al., "Comparing gesture recognition accuracy using color and depth information," in *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 20–22, ACM, 2011.

[27] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Computer Vision—ECCV 2002*, vol. 2350 of *Lecture Notes in Computer Science*, pp. 661–675, Springer, Berlin, Germany, 2002.

[28] W. Liu, Y. Fan, Z. Zhong, and T. Lei, "A new method for calibrating depth and color camera pair based on Kinect," in *Proceedings of the International Conference on Audio, Language and Image Processing (ICALIP '12)*, pp. 212–217, July 2012.