Superpixel-Based Hand Gesture Recognition With Kinect Depth Camera

Chong Wang, *Member, IEEE*, Zhong Liu, and Shing-Chow Chan, *Member, IEEE*

***Abstract*—This paper presents a new superpixel-based hand gesture recognition system based on a novel superpixel earth mover's distance metric, together with Kinect depth camera. The depth and skeleton information from Kinect are effectively utilized to produce markerless hand extraction. The hand shapes, corresponding textures and depths are represented in the form of superpixels, which effectively retain the overall shapes and color of the gestures to be recognized. Based on this representation, a novel distance metric, superpixel earth mover's distance (SP-EMD), is proposed to measure the dissimilarity between the hand gestures. This measurement is not only robust to distortion and articulation, but also invariant to scaling, translation and rotation with proper preprocessing. The effectiveness of the proposed distance metric and recognition algorithm are illustrated by extensive experiments with our own gesture dataset as well as two other public datasets. Simulation results show that the proposed system is able to achieve high mean accuracy and fast recognition speed. Its superiority is further demonstrated by comparisons with other conventional techniques and two real-life applications.**

***Index Terms*—Hand gesture recognition, human-computer interaction, Kinect, superpixel earth mover's distance.**

I. Introduction

HAND GESTURE recognition has received great attention due to its potential applications in contactless human-computer interaction (HCI). There is considerable progress in this area and a number of algorithms addressing different aspects of the problem have been proposed [1]. While image-based techniques have been widely studied, it may be affected by lighting conditions, large variations of the hand gesture and textures, etc. In particular, reliable hand detection is essential to gesture recognition and many hand detection techniques have been developed for tracking and recognizing various hand gestures [2], [3]. Some of them require users to wear an electronic glove so that the key features of hand can be accurately measured but the device is somewhat costly and inconvenient for domestic applications [4]. Another class of methods employs optical markers to replace electronic gloves but it requires rather complex configuration [5]. Methods based on skin color model [6] and hand shape model [7] have also been proposed. However, they are not robust in the dynamic environment and rely significantly on the models.

Once the hands have been located, hand gesture recognition aims to interpret predetermined gestures into independent sign, mostly static, which can be classified by pattern classifiers such as k-Nearest Neighbors (kNN) [8], Hidden Markov Models [9], Principal Component Analysis (PCA) [10] and Support Vector Machine (SVM) [11]. Another effective gesture recognition algorithm [12] proposed recently is based on the Finger Earth Mover's Distance (FEMD) and Template Matching Method (TMM), which shows promising performance.

In this work, we propose a superpixel-based hand gesture recognition system to be used with Kinect depth camera. Because of the good recognition rate of segmented hand gesture, our system adopts segmentation based methods for hand detection. As current image-based techniques [1]–[3] rely heavily on the prior knowledge of hand models, it is usually problematic to separate the user's hand from cluttered backgrounds in practical applications, especially when they have similar color and textures such as the face. To obtain improved segmentation, the depth information from Kinect depth camera can be utilized to better locate and extract hand segments [13]. Moreover, Kinect can capture the color image and depth map at 30 FPS with 640×480 resolution [14] which is sufficient for our purpose. To reliably acquire the hand gesture images, we make use of both skeleton tracking and the depth map from Kinect, which help to locate the user's hands and extract their shapes, respectively. Experimental results show that the human posture tracking can greatly facilitate gesture recognition.

For efficient hand gesture recognition, a superpixel-based representation of the hand shape, inspired by the widely used concept of superpixel [15], [16], is introduced in this work, from which a new distance metric, Superpixel Earth Mover's Distance (SP-EMD), is proposed to measure the dissimilarity between the hand gestures. Moreover, a novel concept of virtual superpixels is proposed to represent folded fingers which addressed the issue of partial matching. The proposed SP-EMD naturally incorporates both the color texture and depth map into the hand shapes as well. As a result, it leads to a highly accurate and efficient hand gesture recognition system, without the requirement of specific markers. Finally, the proposed system has moderate complexity and can be implemented in real-time. Fig. 1 summarized the major steps in the proposed hand gesture recognition framework based on the SP-EMD and Kinect.

Manuscript received July 15, 2014; revised November 10, 2014; accepted November 12, 2014. Date of publication November 24, 2014; date of current version December 23, 2014. This work was supported in part by the General Research Fund (GRF) of the Hong Kong Research Grant Council (RGC). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Enrico Magli.

The authors are with the Department of EEE, The University of Hong Kong, Pokfulam, Hong Kong (e-mail: scchan@eee.hku.hk; cwang@eee.hku.hk; liuzhong@eee.hku.hk).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes a video file, which is a demo video of the proposed hand gesture recognition system. This material is 14.4 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2374357

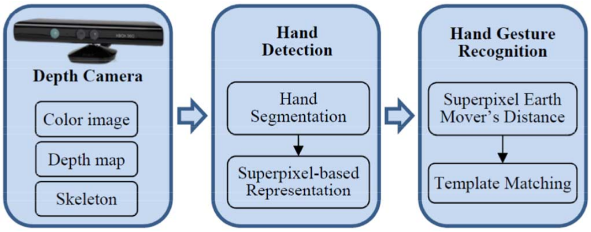


Fig. 1. Framework of the proposed superpixel-based hand gesture recognition system.

The rest of the paper is organized as follows. Related works are first reviewed in Section II. The proposed methods for hand localization, segmentation, superpixel-based representation and preprocessing are described in Section III. The novel distance metric, SP-EMD, for hand gesture classification is then proposed in Section IV. The performance of the SP-EMD-based hand gesture recognition is evaluated and compared with other state-of-the-art algorithms in Section V. We also build two real-life HCI applications to further illustrate the effectiveness of our system in Section VI. Finally, we conclude the paper in Section VII.

II. Related Works

Many vision based hand gesture recognition algorithms have been proposed in the past years and comprehensive reviews can be found in [1]–[3]. The recent development of depth cameras, such as Microsoft Kinect, Creative Senz3D or Mesa SwissRanger etc., opens up new avenues for hand gesture recognition, thanks to the extra depth information. Therefore, how the depth information can be efficiently utilized and how the depth camera can be incorporated in the hand gesture recognition system is an active topic of research [17].

One of the greatest advantages of using depth cameras for hand gesture recognition is in hand detection. In early studies, hand detection mainly relies on vision-based features. For instance, chromatic distribution of hand is used to build the hand model in [18]. However, such method is sensitive to variations of skin colors. An appearance-based detection framework was also proposed in [19]. The complicated and unpredictable hand features, however, pose great challenges to its reliability. In comparison, model-based method [20] is better suited to real-life interaction, but a dark background is usually assumed so that the hand gesture can be segmented reliably.

On the other hand, depth cameras offer a much simpler but effective way to isolate the hands by using a depth threshold. This method is widely used in many researches [12], [21]. But the selection of the depth threshold is empirical and prone to errors. To better locate and track the hands, some researches [22] use the body skeleton provided by Microsoft Kinect, which shows promising performance.

After the hand localization and segmentation, various hand features can be extracted from either the depth maps, e.g. Histogram of 3D Facets (H3DF) [23] and 3D point distribution histogram [21], or the corresponding color images such as Histogram of Oriented Gradients (HOG) [24], which will then be used for hand gesture recognition. Alternatively, it is also quite common to use the feature of hand contour, which is, however, usually noisy and distorted due to the low resolution and accuracy of the current depth cameras. Thus, contour based algorithms [25] are not robust when the contour is locally distorted, whereas the skeleton-based methods [26] may have difficulties to extract correct skeleton from the noisy or distorted hand contour. Correspondence-based methods such as shape context [27] and inner distance [28] also suffer from ambiguity due to the orientation, distortion and articulation of the hand gestures. Recently, Finger-Earth Mover's Distance (FEMD) is proposed in [12] to provide a robust way for hand gesture recognition.

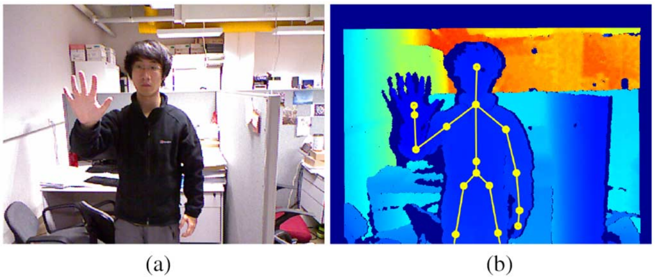


Fig. 2. Frames from color and depth camera. (a) The color texture captured by Kinect. (b) Registered depth map with the skeleton.

III. Hand Detection

Hand detection, including hand localization, segmentation and representation, is a non-trivial problem in gesture recognition. Previous studies reveal many challenges, such as low accuracy of predicting user's body movement and the weakness against cluttered backgrounds or various lighting conditions. In this work, we utilize the depth information and skeleton tracking provided by Kinect to address these problems.

A. Calibration Between Color and Depth Camera

Although the color texture and depth map are captured by Kinect simultaneously, they are not registered accurately. Therefore, a recalibration procedure for Kinect is required in order to jointly utilize the color and depth information. In this work, Heikkila's method [29] is used to recalibrate Kinect's color and infrared (IR) sensors. Since Kinect depth map is generated from the IR observations, the estimated color-IR camera parameters can be applied to register the depth map and color image. More details can be found in [14]. Fig. 2 shows an example after the recalibration, it can be seen that the color texture and depth map are precisely registered.

B. Hand Localization and Segmentation

In previous depth camera-based approaches [12], [21], [22], the hand is required to be the front-most object from the depth camera. Moreover, a black belt on the gesturing hand's wrist is also required in [12], which is rather inconvenient for real world applications. In our system, we relax these restrictions by utilizing the rather stable joints from Kinect's skeleton tracking. The Kinect joints are directly used to locate the hands, wrists and elbows. An example of the skeleton is given in Fig. 2(b), which shows that the hand location can be detected accurately.

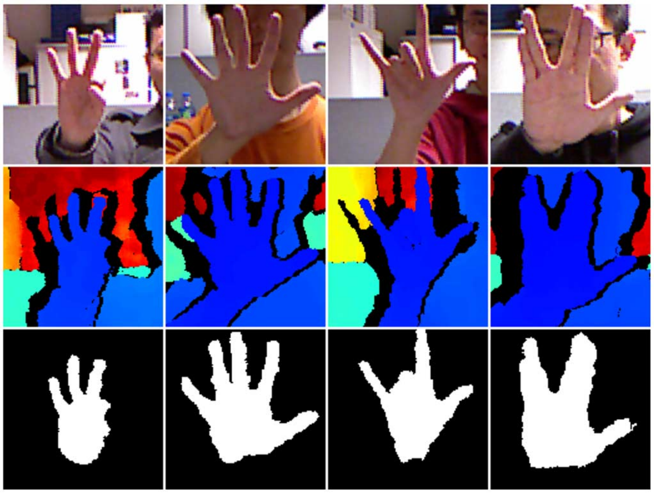


Fig. 3. Hand localization and segmentation. From top to bottom: extracted color texture blocks of the hands; depth map blocks of the hands; segmented hand shapes.

By assuming that the hand is visible to the camera without any occlusion, it allows us to quickly separate the hands from background objects using depth information alone. Using the hand joint point as the center, a pair of color texture and depth map blocks is extracted first as shown in the top two rows of Fig. 3. Based on the joint points of hand, wrist and elbow, we can also estimate the orientation of the arm in order to select an appropriate depth threshold such as the depth value of the wrist joint point which is denoted by d_{wrist}. Then the hand shape is segmented quickly by simply comparing the depth values as shown in the last row of Fig. 3. It can be seen that the hand shapes from different persons are correctly segmented, even when the hands are cluttered by the face or background.

C. Shape Representation Using Joint Color-Depth Superpixel

Instead of representing the hand shape in contour [12], [27] or skeleton [26], we propose to use superpixels to simplify the hand shape but retain as much information as possible. Therefore, not only the 2D shape but also the corresponding texture and depth can be jointly utilized in hand gesture recognition. This representation is a key ingredient of the proposed SP-EMD recognition system, which will be introduced later in Section IV.

Although various kinds of superpixel algorithms have been proposed, compact and efficient representations are preferred in the proposed superpixel-based recognition system for real-time implementation. We adopt and modify the Simple Linear Iterative Clustering (SLIC) algorithm [15] in our system, because it can enforce the compactness of superpixels. In this work, the clustering is performed in a 6-D space including the (l, a, b) values of the CIELAB color space and the (x, y, d) pixel coordinates, where d is the depth value at the pixel location (x, y). Assume an image with N pixels is segmented into K superpixels. Then each superpixel should have nearly equal size of N/K pixels. Let $[\mathbf{h}_i^T, \mathbf{u}_i^T]^T$ be the centroid of the i-th cluster, where $\mathbf{h}_i = [l_i, a_i, b_i]^T$ and $\mathbf{u}_i = [x_i, y_i, d_i]^T$. The following metric can be defined to measure the pixel-to-pixel distance

$$D_s = d_{lab} + \frac{c_s}{N/K} d_{xyd} \tag{1}$$

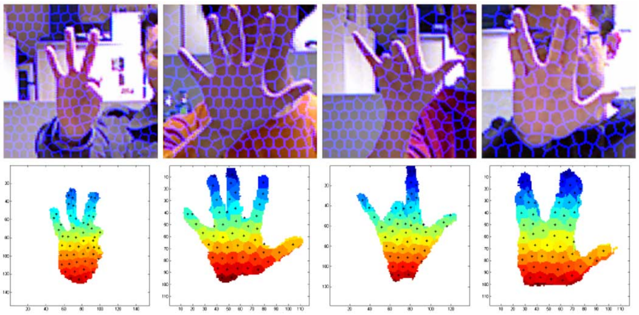


Fig. 4. Shape representation using superpixels. First row: superpixels on color textures. Last row: corresponding shapes represented by superpixels. Black dots indicate the centers of superpixels.

where c_s is the compactness coefficient of superpixels, $d_{lab} = \|\mathbf{h}_i - \mathbf{h}_j\|$ and $d_{xyd} = \|\mathbf{u}_i - \mathbf{u}_j\|$ are respectively the distances in the (l, a, b) and (x, y, d) spaces, and $||.||$ denotes the Euclidean norm. Thus D_s is a weighted sum of the *lab* distance and *xyd* distance, controlled by the compactness coefficient c_s By initializing cluster centers at regular grid steps, K-Means clustering is then used based on D_s to generate the superpixels iteratively.

Utilizing this modified SLIC, the corresponding hand shape is segmented into small clusters, i.e. superpixels. Some examples are given in Fig. 4. It shows that the hand shapes are successfully segmented into a small group of superpixels, which are denoted in different colors. The hand shape can thus be represented by the properties of those superpixels, such as the 2D centroids (black dots in Fig. 4), the color distribution and the mean depth values.

It can be seen that these joint color-depth superpixels are of similar size and regular shapes, but containing edge information from color textures as well as depth maps. This superpixel-based representation tries to group similar pixels together with reduced variables by minimizing the information loss. In other words, the key features, including the color, depth and shape, of the hands are well preserved after dimension reduction. Another merit of this representation is that it is robust to distortions of the hand contours, since the centroids of the superpixels are not determined by the contour alone. This superpixel representation will be used in the next section to define the proposed SP-EMD for hand gesture recognition.

D. Preprocessing for Scaling, Translation, In-Plane, and Out-of-Plane Rotations

In practical applications, the extracted hand gestures usually have different scales due to various distances from the camera to hand, or different rotations caused by the body postures. Moreover, different people's hands always have distinct characteristics even for the same gesture. Hence it is necessary to perform some preprocessing to normalize and align the proposed superpixel shape representation before recognition.

First of all, the hand images are scaled by a scaling factor $d_{\min}/d_{ref}$, where $d_{\min}$ is the minimum depth value on the hand and d_{ref} is a reference depth value. Considering the hand size and the range of Kinect, d_{ref} is set as 1 meter for all the experiments in this paper. However, the palm size varies from one person to another, which affects the recognition between different subjects. To address this issue, the 2D pixel coordinates (x, y) are translated and normalized according to the center and radius of the maximum inscribed circle, i.e. the palm, of the obtained hand, respectively.

To address the in-plane rotation caused by the body posture, the angle determined by the joint points of wrist and elbow is used to rotate the hands to a similar orientation. To further refine in-plane rotation, 2D Iterative Closest Point (ICP) [30] is adopted to align the superpixels of two given gestures before recognition. In our system, the initial rotation matrix and translation vector are simply set as an identical matrix and [0,0], respectively. Comparing with the thin plate spline (TPS) model used in Shape Context [27], ICP has lower complexity and thus is more suitable for real-time implementation. Moreover, the number of the superpixels is small, typically from 20 to 80, which means ICP will be very efficient in our system.

Meanwhile, moderate out-of-plane rotation of the hand is compensated with the aid of the depth map. More precisely, the palm plane is first estimated from the 3D point cloud of the hand, which is then rotated to ensure that the palm plane is parallel to the image plane. Finally, the rotated point cloud is projected back to image plane, which is then utilized for the later steps of superpixel generation and hand gesture recognition.

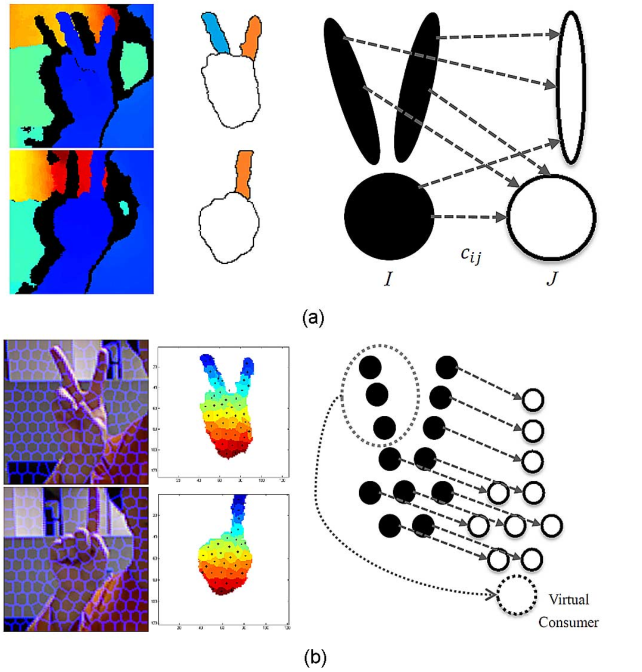


Fig. 5. Illustration of superpixel earth mover's distance. (a) is based on finger extraction and shape only. Left is the hand shapes, and right is the corresponding transportation problem. (b) is an example of joint color-depth SP-EMD. Left is the superpixel representation on both shapes and textures. Right is the corresponding transportation problem.

IV. Hand Gesture Recognition

The Earth Mover's Distance (EMD) [31] is a measure of the distance between two probability distributions over a region. It is widely used in image retrieval and pattern recognition. Inspired by the work of Finger-Earth Mover's Distance (FEMD) [12], we now propose a novel distance metric, Superpixel Earth Mover's Distance (SP-EMD), to recognize the hand gesture. The proposed SP-EMD jointly measures the distance between two hand gestures based on the shape, texture and depth information, while FEMD only uses the contour.

A. Transportation Problem in EMD

The original EMD is based on a solution to the conventional transportation problem. Given a set of suppliers I, a set of consumers J, and the cost c_{ij} to ship a unit of supply from $i \in I$ to $j \in J$, the aim is to find an optimal set F of flow f_{ij}, i.e. the amount of supply shipped from i-th supplier to j-th consumer, to minimize the overall cost,

$$\sum_{i\in I} \sum_{j\in J} c_{ij} f_{ij}, \tag{2}$$

subject to the constraints

$$f_{ij} \geq 0, \quad i \in I, j \in J \tag{3}$$

$$\sum_{i\in I} f_{ij} = t_j, \quad j \in J \tag{4}$$

$$\sum_{j\in J} f_{ij} \leq s_i, \quad i \in I \tag{5}$$

where s_j is the supply of i-th supplier and t_j is the capacity of j-th consumer.

B. Superpixel Earth Mover's Distance

For hand gesture recognition, we can use this transportation problem to mimic the motion of fingers between different gestures by defining the suppliers/consumers from the hand shapes. If the fingers can be segmented correctly, these segments can be defined as the suppliers/consumers as shown in Fig. 5(a). Two hand shapes are decomposed to two finger parts plus the palm and a single finger plus the palm, respectively. It is formed as the transportation problem with three suppliers (black nodes) and two consumers (white nodes). The finger features are used to determine the cost c_{ij}, and the area of the finger is intuitively the amount of supply (or capacity). Actually, it is a general form of FEMD, while FEMD discards the palm and uses the angle intervals of the fingers to define the cost.

However, there are some major limitations. First of all, accurate finger extraction is challenging. The spatial resolution of the shape also becomes lower, since the whole finger is considered as one cluster. Moreover, the texture is neglected, which will potentially lose rich details. Therefore, we propose a novel joint color-depth superpixel earth mover's distance (SP-EMD) to address these issues. An example is shown in Fig. 5(b), where the hand shapes, color textures and depths are represented in superpixels. Each superpixel is either a supplier p_i (black nodes) or a consumer q_j (white nodes). Its 2D location (x, y) and depth value d of p_i or q_j are then used to calculate the cost c_{ij}. The number of pixels m_i (or n_j) within p_i (or q_j) is the amount of supply (or capacity). It should be noticed that the total supply, $\sum_{i\in I} m_i$, and capacity, $\sum_{j\in J} n_j$, may be not equal due to the different shapes of hand gestures. It will lead to partial matches, which results in false matching in gesture recognition. Thus a novel virtual superpixel (consumer) is proposed here to receive exceeded supplies as the dotted circle indicated in Fig. 5(b). This virtual consumer will be located at the palm center. It is worth noting that it is not just a dummy, but representing the folded fingers, which naturally solves the partial matching problem.

For simplicity, the suppliers and consumers are denoted as two signatures. Each one is defined as a set of superpixels $p_i, i = 1, \ldots, k$ with corresponding weight w_{p_i}. Formally, let $P = \{(p_1, w_{p_1}), \ldots, (p_k, w_{p_k})\}$ be the first hand signature with k superpixels, and $Q = \{(q_1, w_{q_1}), \ldots, (q_l, w_{q_l})\}$ be the second hand signature with l clusters. The centroid $[x_{p_i}, y_{p_i}]$ and the average depth value d_{p_i} are used to define the superpixel $p_i = [x_{p_i}, y_{p_i}, d_{p_i}]^T$. Comparing with the texture, the depth is insensitive to illumination changes. The number of pixels m_i within the superpixel is used to denote the cluster weigh w_{p_i}. Then the cost c_{ij} from superpixel p_i to q_j is defined as the weighted 3D distance

$$c_{ij} = [(x_{p_i} - x_{q_j})^2 + (y_{p_i} - y_{q_j})^2 + \alpha(d_{p_i} - d_{q_j})^2]^\beta \quad (6)$$

where α is the depth weight that balances the significance between the 2D shape and depth, and β is a nonlinear fingertip coefficient to give more penalties on fingertips, considering the fact that fingertips have more impact on the gestures than the palm does. A typical choice of α and β is 1.0 and 2.0, respectively. Detailed discussion about the parameter sensitivity can be found in Section V-C.

Given two signatures P and Q, their SP-EMD distance is the least work needed to move the pixels between two sets of superpixels. As mentioned above, the virtual superpixels $p_0 = [0, 0, 0]^T$ and $q_0 = [0, 0, 0]^T$ are created with the weights

$$w_{p_0} = \begin{cases} 0, & w_p \geq w_q \\ w_q - w_p, & otherwise \end{cases} \quad (7)$$

$$w_{q_0} = \begin{cases} 0, & w_p \leq w_q \\ w_p - w_q, & otherwise \end{cases} \quad (8)$$

where $w_p = \sum_{i=1}^{k} w_{p_i}$ and $w_q = \sum_{j=1}^{l} w_{q_j}$ are respectively the total weights, i.e. total number of pixels, of signatures P and Q. Then the SP-EMD shares a similar formulation of conventional EMD as

$$SEPMD(P, Q) = \frac{\sum_{i=0}^{k} \sum_{j=0}^{l} c_{ij} f_{ij}}{\sum_{i=0}^{k} \sum_{j=0}^{l} f_{ij}} \quad (9)$$

where *SPEMD*(.) denotes the SP-EMD distance function and f_{ij} is the flow from superpixel p_i to q_j, which is calculated by minimizing (2) subject to the constrains (3-5). For the proposed SP-EMD, it can be rewritten as

$$\mathbf{F} = \arg\min \sum_{i=0}^{k} \sum_{j=0}^{l} c_{ij} f_{ij}$$
$$s.t. \begin{cases} f_{ij} \geq 0, & 0 \leq i \leq k, 0 \leq j \leq l \\ \sum_{i=0}^{k} f_{ij} = w_{q_j}, & 0 \leq j \leq l \\ \sum_{j=0}^{l} f_{ij} = w_{p_i}, & 0 \leq i \leq k \end{cases} \quad (10)$$

where $\mathbf{F}$ is the matrix form of the flow f_{ij}. The first constraint only allows moving flow along one direction. Fig. 6 shows two cases of the moving flow calculated between similar hand gestures from different subjects and different gestures from the same subject. It can be seen that the flow is short and organized between the similar gestures in Fig. 6(d), while it turns to be long and disordered between different ones in Fig. 6(e). It is noted that the length of the flow is partially proportional to the moving cost. Hence it is obvious that the flow in Fig. 6(e) will lead to a larger SP-EMD distance, which shows the nature of the proposed distance metric.

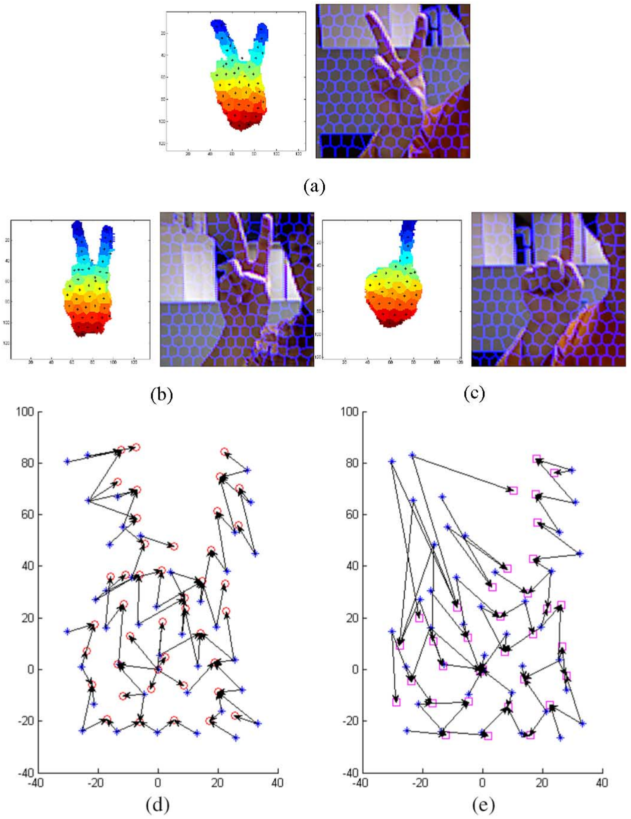


Fig. 6. The moving flow of SP-EMD between hand gestures. (a) A given hand gesture in superpixel representation. (b) A similar gesture to (a) from a different subject. (c) A different gesture given by the same subject as (a). (d) The optimal flow from (a) to (b). (e) The optimal flow from (a) to (c). Blue stars, red circles and magenta squares denote gesture (a), (b) and (c), respectively. Black arrows indicate the moving flow directions.

C. Template Matching

Template matching is utilized for hand gesture recognition based on the proposed SP-EMD. In particular, the input hand gesture is recognized as a certain class, namely g, with the minimum dissimilarity distance as

$$g = \arg\min_{g} \ SPEMD\,(H, T_g) \quad (11)$$

where H is the input gesture and T_g is the template of class g. It is obvious that the selection of templates will significantly affect the recognition rate.

For a template based approach, the performance is closely related to the selected templates, i.e. training data. In our experiments, leave-p-out (LpO) cross-validation (CV) is conducted

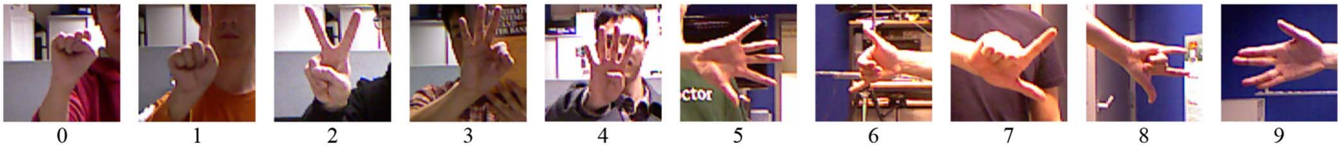


Fig. 7. Gesture samples (0-9) captured in two different environments.

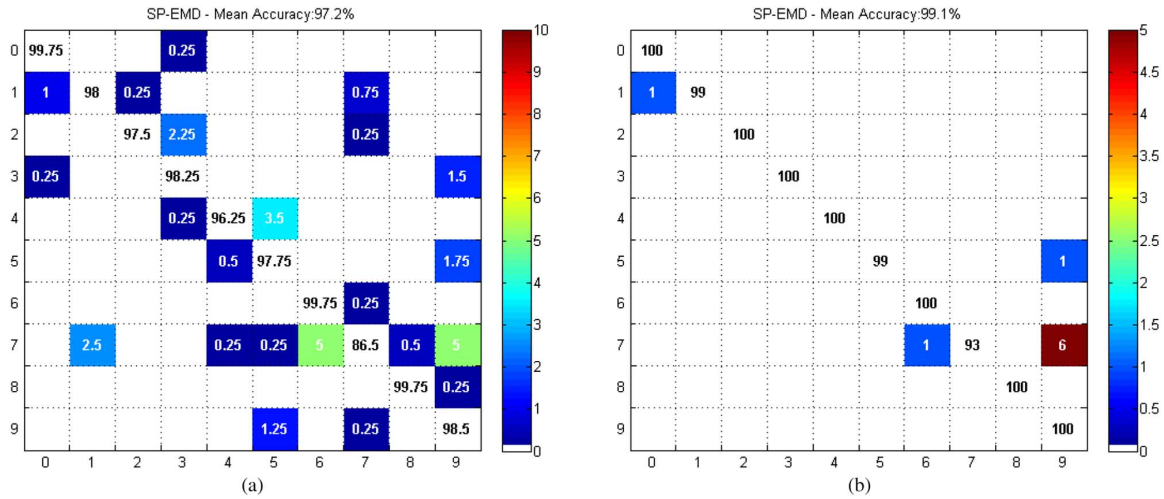


Fig. 8. Confusion matrices of hand gesture recognition using SP-EMD (unit:%). (a) L4O CV. (b) LOO CV.

to evaluate the recognition performance. For a dataset with M subjects, M–p subjects are used for training and the remaining p for testing in LpO CV. This process is repeated for every combination of p subjects so that the average accuracy can be computed. In our dataset, two values of p (1 and 4) are considered, which are respectively referred to as leave-one-out CV (LOO CV) and leave-4-out CV (L4O CV). Experiments based on these two CVs are presented in next section.

V. Experimental Evaluations

We now evaluate and compare the proposed hand gesture recognition system with various state-of-the-art recognition algorithms including Shape Context [27], Skeleton Matching [26], FEMD [12], Random Forest (RF) [32], HOG [24] and H3DF [23], using three different real world datasets, namely our joint color-depth hand gesture dataset, NTU hand digit dataset [12] and American Sign Language (ASL) finger spelling dataset [32]. Two different CV schemes are tested to show the effectiveness of the proposed hand representation and SP-EMD distance metric. The confusing cases and system sensitivity will also be discussed.

A. Datasets

A joint color-depth hand gesture dataset (available in our project homepage[1]) is collected using Kinect. It contains 10 gestures with 20 different poses from 5 subjects. Therefore, there are a total of 1,000 cases for testing, each of which consists of a pair of color texture and depth map with corresponding skeleton information used in our experiment. Gesture samples are shown in Fig. 7, which are labeled from 0 to 9. It should be noted that this dataset is a challenging real-life dataset, which is collected in two different rooms with different illumination conditions using different Kinects. Moreover, the hand motion is not very restrictive including large in-plane rotation and moderate out-of-plane rotation. In the experiments, the hand shapes are extracted and preprocessed using the methods described in Sections III-B and III-D. The evaluation strategies LOO CV and L4O CV mentioned in Section IV-C are applied in order to give a comprehensive test of the proposed hand gesture recognition system.

We also evaluate our algorithm using two public Kinect gesture datasets, namely NTU Hand Digit Dataset [12] and ASL Finger Spelling Dataset [32]. The NTU hand digit dataset [12] contains 1,000 cases of 10 hand gestures from 10 subjects. The ASL Finger Spelling dataset [32] captures about 65,000 samples of 24 hand gestures (English letters from $\boldsymbol{a}$ to $\boldsymbol{y}$ except $\boldsymbol{j}$) from 5 subjects. The hands are located and segmented using the hand-wrist belt and depth thresholding, respectively. For fair comparison with the reported accuracy in [12], [23], [32], only LOO CV is applied to these two datasets.

B. Performance Evaluation

All experiments were done on an Intel Core i7-920 2.66 GHz CPU with 6 GB of RAM. Now we evaluate the performance of the proposed system from mean accuracy, time efficiency and comparisons with other methods.

Mean Accuracy: In all the experiments, the depth weight α and fingertip coefficient β are fixed as 1.0 and 2.0, respectively. The average size of superpixels N/K is set as 81, i.e. roughly equal to a size of 9×9. The confusion matrices for L4O CV and LOO CV on our dataset are shown in Fig. 8. We can see that the most confusing case is between gestures 7 and 9, and 7 and 6. Two examples of confusing cases are given in Fig. 9. It shows

[1] Project homepage of SP-EMD, https://sites.google.com/site/spemdkinect.

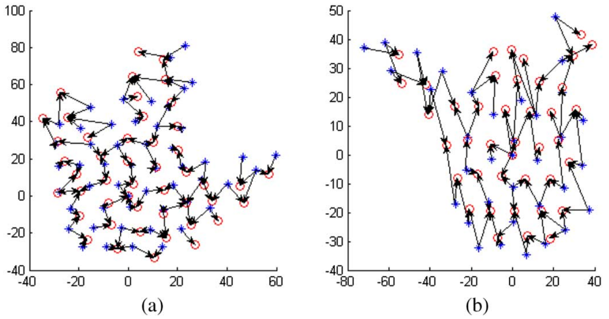


Fig. 9. Confusing cases of SP-EMD. (a) Gestures 7 (blue stars) and 9 (red circles). (b) Gestures 7 (blue stars) and 6 (red circles).

TABLE I
THE MEAN ACCURACY AND MEAN RUNNING TIME OF FEMD, SHAPE CONTEXT, SKELETON MATCHING, AND OUR PROPOSED SP-EMD ON OUR DATASET

Algorithms	Mean Accuracy		Running Time
	L4O CV	LOO CV	
FEMD [12] (Thresholding)	91.025%	95.0%	1.155 s (Matlab) 0.075 s (C/C++)
Shape Context [27] (without bending cost)	92.200%	97.5%	7.608 s (Matlab)
Shape Context [27] (with bending cost)	85.375%	95.7%	7.608 s (Matlab)
Skeleton Matching [26] (DCE [33])	89.575%	96.0%	2.262 s (Matlab)
Skeleton Matching [26] (DSE [34])	90.475%	96.0%	2.417 s (Matlab)
SP-EMD (Shape Only, $\beta = 2.0$)	96.500%	98.3%	0.067 s (C/C++)
SP-EMD ($\alpha = 1.0, \beta = 2.0$)	**97.200%**	**99.1%**	0.067 s (C/C++)

that the hand shapes become very similar due to the distortion or the habits of different people in performing the gestures.

As presented in Table I, the mean accuracies are 97.2% and 99.1% for L4O CV and LOO CV, respectively. It can be seen that LOO CV achieves better recognition rates than L4O CV because the number of training data (or templates) in the former is 4 times larger. Also, we note that our recognition rate is at least 5% higher than other algorithms for the L4O CV considered in Table I. In other words, the performance of our hand gesture recognition system is relatively insensitive to the number of training data used, which is very convenient in real-life applications.

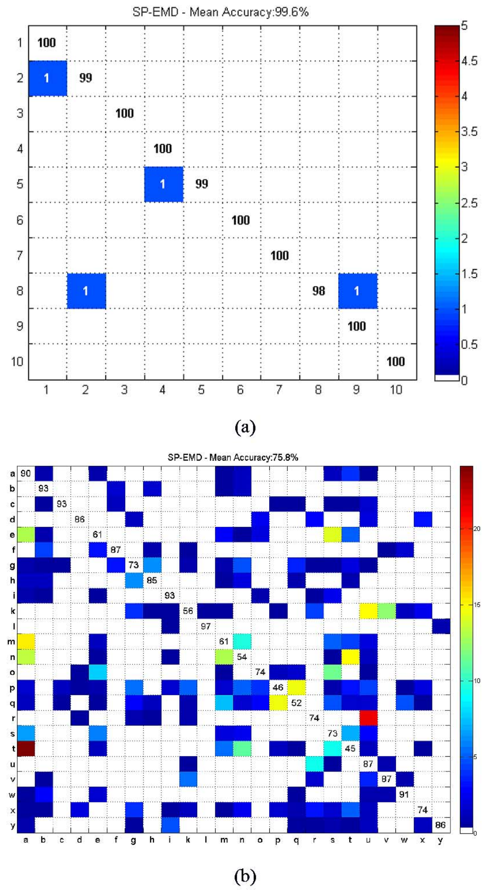
To illustrate the significance of the joint color-depth information from the texture and depth map, we perform another experiment that only the shape information is considered. To be specific, the color distance d_{lab}, depth d and depth weight α are all set to 0 for the generation of superpixels in (1) and calculation of the cost in (6). As shown in Table I, the mean accuracies are slightly degraded to 96.500% for L4O CV and 98.3% for LOO CV. This suggests that the proposed SP-EMD is very effective even without the texture and depth map, while the additional joint color-depth information do help to improve the recognition rate.

TABLE II
COMPARISON ON THE NTU HAND DIGIT DATASET

Algorithms	FEMD [12] (Near Convex)	HOG [23]	H3DF [23]	SP-EMD
Mean Accuracy (LOO CV)	93.9%	93.1%	95.5%	**99.6%**

TABLE III
COMPARISON ON THE ASL FINGER SPELLING DATASET

Algorithms	RF [32]	HOG [23]	H3DF [23]	SP-EMD
Mean Accuracy (LOO CV)	49.0%	65.4%	73.3%	**75.8%**



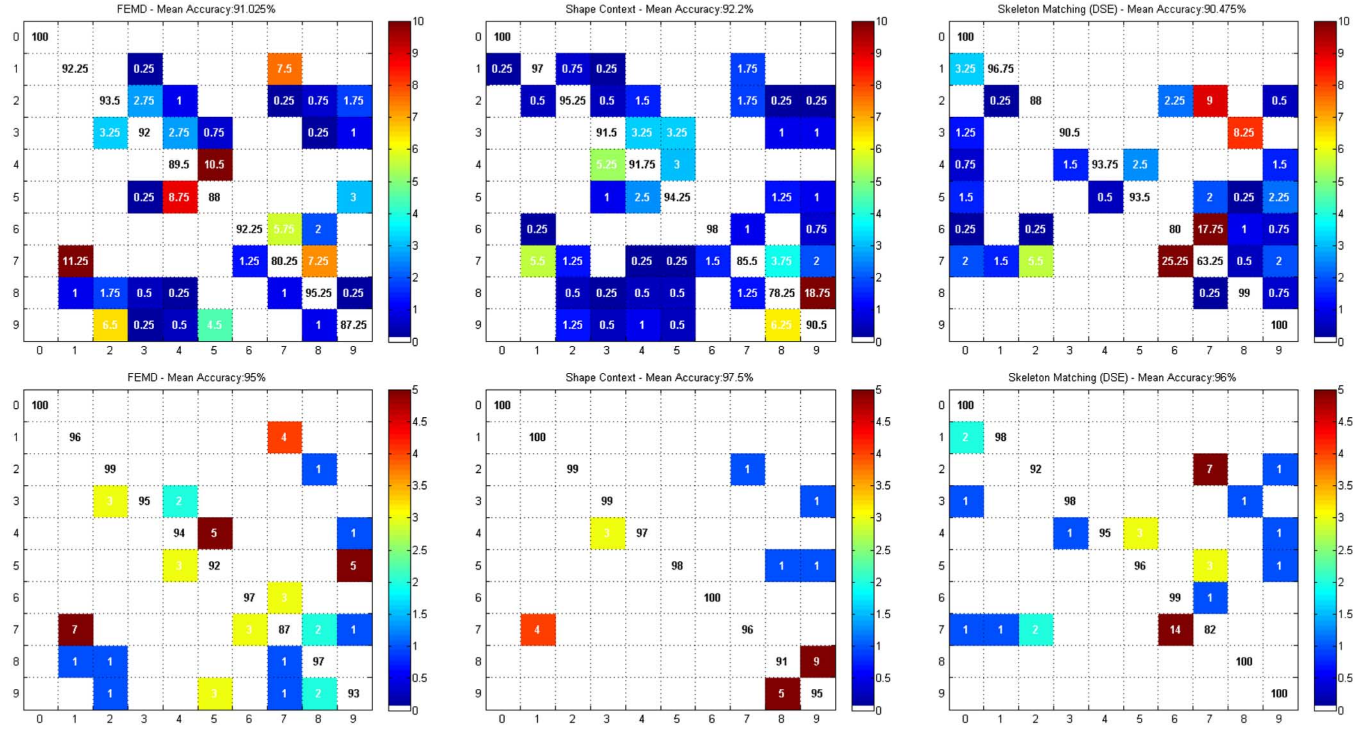
(a)

(b)

Fig. 10. The confusion matrices of hand gesture recognition (LOO CV) using SP-EMD on (a) NTU hand digit dataset [12] and (b) ASL finger spelling dataset [32] (unit: %).

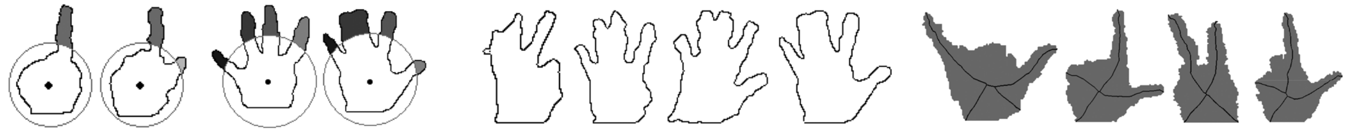
Apart from our dataset, we also applied the proposed system to two other public datasets, NTU hand digit dataset [12] and ASL finger spelling dataset [32], for which mean accuracies (99.6% and 75.8%) are respectively summarized in Tables II and III. The confusion matrices on these two datasets are given in Fig. 10.

Time Efficiency: Table I gives the average running time (0.067 seconds) for the recognition process with our $\mathrm{C/C++}$ implementation. The whole process includes generating superpixels, extracting features and calculating SP-EMD between the input gesture and 10 templates. Thanks to the superpixel-based representation, the ICP is very efficient (around 1 millisecond) in the proposed system. The most time consuming step is superpixel generation, which costs about 27 milliseconds. That means computing SP-EMD once only needs less than 4 milliseconds, which is very quick. As a result, the proposed SP-EMD is not only accurate but also capable of running in real-time.



(a) (b) (c)

Fig. 11. The confusion matrices of hand gesture recognition using (a) FEMD [12], (b) shape context [27], and (c) skeleton matching [26] with DSE [34] (unit: %). The upper and lower rows are the results of L4O CV and LOO CV, respectively.



(a) (b) (c)

Fig. 12. Confusing cases of (a) FEMD [12], (b) shape context [27], and (c) skeleton matching [26] with DSE [34].

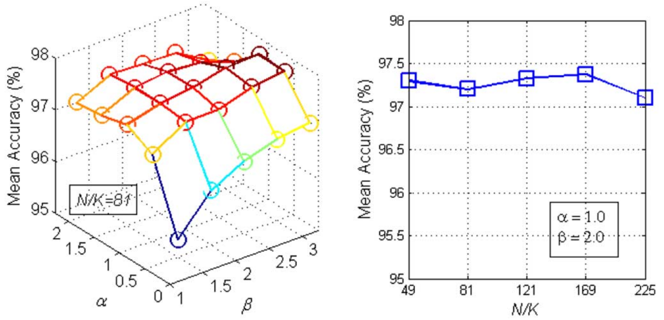
Comparisons With Other Methods: To further illustrate the advantage of our system, we first compare it with other three state-of-the-art recognition algorithms, Shape Context [27], Skeleton Matching [26] and FEMD [12] on our dataset. Their mean accuracies and running time are shown in Table I. The hand shapes are segmented and preprocessed using the same method described in Sections III-B and III-D. It can be seen that the proposed hand recognition system achieves the highest mean accuracy.

It is worth noting that FEMD is particularly designed for depth-camera based hand gesture recognition. Two different finger decomposition methods are proposed for FEMD in [12], and the reported mean accuracies are very close. In particular, we apply the thresholding decomposition based method to our dataset. Note that the lower left corner of the hand shape is chosen as the initial point of the contour in our dataset, instead of the wrist belt used in [12]. The confusion matrices of FEMD for L4O CV and LOO CV on our dataset are shown in Fig. 11(a). It can be seen that the most confusing cases are between gestures 4 and 5, and 1 and 7. The main reason is that the fingers in those gestures are not correctly segmented due to the distortion. Moreover, different hand sizes also change the weight of each finger in FEMD, which leads to mismatching. Fig. 12(a) presents some examples of confusing cases for FEMD. As it shows, the palm size and relatively the length of the finger have great impact on the recognition performance.

Fig. 11(b) shows the confusion matrices of Shape Context [27]. The results are generated based on the Matlab demo code provided at the project homepage of Shape Context.[2] The most confusing cases are between gesture 1 and 7, and 8 and 9, since they have similar contours. Sometimes two fingers may fuse into one due to the distortion. That is why gesture 4 is also mistakenly recognized as gesture 3. Fig. 12(b) gives some confusing examples for Shape Context. It can be seen that the distortion changes the contours severely.

Two different skeleton pruning methods, discrete curve evolution (DCE) [33] and discrete skeleton evolution (DSE) [34], are tested for Skeleton Matching [26] in our experiments. As claimed in [34], DSE method is more stable to the small protrusions. Hence the recognition accuracy with DSE is higher than DCE for L4O CV as shown in Table I. Fig. 11(c) shows the confusion matrices of Skeleton Matching [26] with DSE [34]. From

[2] Http://www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/sc_digits.html.



(a) (b)

Fig. 13. Parameter sensitivity of α, β, and N/K, using L4O CV.

TABLE IV
MEAN ACCURACY OF SP-EMD WITH ORIENTATION NOISE

σ	5º	10º	15º	20º	25º
Mean Accuracy (L4O CV)	96.40%	96.25%	95.99%	95.72%	95.49%
Mean Accuracy (LOO CV)	98.91%	98.92%	98.88%	98.85%	98.74%

TABLE V
MEAN ACCURACY OF SP-EMD WITH SCALE NOISE

σ	5%	10%	15%	20%	25%
Mean Accuracy (L4O CV)	95.56%	93.44%	91.32%	89.26%	86.87%
Mean Accuracy (LOO CV)	98.64%	97.99%	97.24%	96.25%	95.32%

the figure, we can see that the most confusing cases are between gestures 6 and 7, and 2 and 7. Due to the characteristics of the hand gesture, the pruned skeletons have similar global structures for different gestures as shown in Fig. 12(c).

On the other two public datasets, NTU hand digit dataset [12] and ASL finger spelling dataset [32], the proposed algorithm also outperforms other existing methods including near-convex decomposition based FEMD [12], Random Forest [32], HOG [23], [24] and H3DF [23] as shown in Tables II and III. Note that the compared results are directly extracted from the corresponding reference, since the same LOO CV is applied in our experiments.

C. Sensitivity Analysis

To demonstrate the effectiveness of the proposed system, its sensitivity to the parameter, orientation, scale and view angle is further investigated in this section.

Parameter Sensitivity: There are three key parameters in the proposed SP-EMD, including the depth weigh α, fingertip coefficient β and the average size of superpixels N/K. The evaluation results on these parameters are shown in Fig. 13. It can be seen that the mean accuracy is quite stable whe α, β or N/K varies, which is another merit of the proposed SP-EMD method. For small α and β, we note that the recognition performance is slightly degraded. For example, the recognition rate is below 96% when $\alpha = 0$ and $\beta = 1$. It is because that no depth features ($\alpha = 0$) are utilized, and the moving cost of fingertips is similar to the one of the palm for $\beta = 1$. Also, from Fig. 13(b), it can be seen that the mean accuracy is not sensitive to the average size of superpixels N/K, while a small N/K should be avoided to reduce the computation complexity.

Orientation and Scale Sensitivity: Although a preprocessing step is proposed in Section III-D to ensure all the gestures have similar orientations and scales, it may be insufficient in some extreme cases. To evaluate the orientation and scale sensitivity of the system, synthetic mismatches are added to corrupt the preprocessed data of our hand gesture dataset. More specifically, they are randomly rotated with a degree θ or scaled by a factor of $(1 + \delta)$. In our experiments, θ and δ are generated using a Gaussian distribution $N(0, \sigma^2)$ with zero mean and a standard deviation of σ. Five different values of σ are tested and each test is repeated 50 times. Tables IV and V summarize the averaged accuracies with orientation and scale noise, respectively. It can be seen that the mean accuracy is robust to the orientation noise thanks to the effectiveness of the 2D ICP alignment employed. On the other hand, the performance is relatively more sensitive to the scale noise. Note that the mean accuracy for L4O CV degrades much faster than that of LOO CV, since the latter CV has 4 times as many templates as the former one, and thus it offers higher chances to observe similar level of scale mismatches.

View Angle Sensitivity: In practice, the user may not directly face the camera, which leads to severe out-of-plane rotation of the gestures. To show the effectiveness of the proposed system in this situation, a view angle sensitivity test is also conducted with samples captured from 5 different view angles (roughl $0°$, $\pm10°$ and $\pm20°$) with 5 subjects. All the parameters are set as the same as the previous experiments. The achieved mean accuracies for L4O CV and LOO CV are 94.84% and 98.07%, respectively. It can be seen that the mean accuracy does not degrade too much (2.36% drop in L4O CV and 1.03% drop in LOO CV). This test suggests that the proposed recognition system is quite robust to these view angle changes.

VI. APPLICATIONS

Hand gesture recognition finds great potential in many emerging applications such as interactive gaming and virtual reality systems over traditional input devices like keyboards and mice. Therefore, we use the hand gesture as an interface to implement two real life HCI applications, Rock-Paper-Scissors -Lizard-Spock Game and 3D Content Browser, based on the proposed hand gesture recognition system. As shown in Table I, our system runs in real-time with a considerable high accuracy, which ensures a pleasant user experience. The demo video is available in our project homepage.

A. Rock-Paper-Scissors-Lizard-Spock Game

The experiments show that our system can recognize Gesture 9, i.e. the famous "Spock" gesture, with a high accuracy. In this demo, we build a Rock-Paper-Scissors-Lizard-Spock Game (an expansion of the classic Rock-Paper-Scissors game) system played between a human and a computer. The computer will randomly choose a weapon, while user's gesture is recognized by our system. The system will show the result whether the user is the winner (smiling face), even if the user is not familiar with

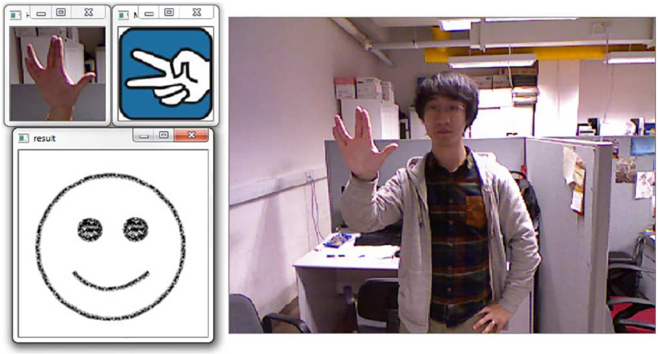


Fig. 14. Rock-Paper-Scissors-Lizard-Spock Game.

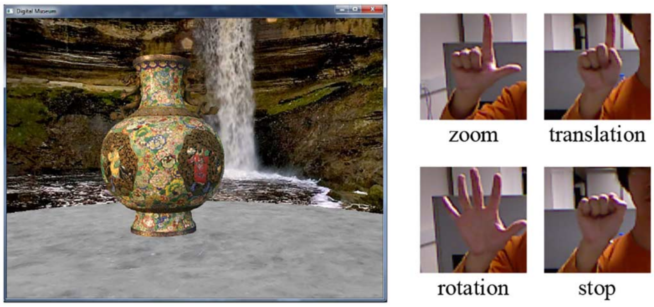


Fig. 15. Example view (left) and commands (right) of 3D content browser.

the complicated rule of the game. Fig. 14 shows an example of the demo.

B. 3D Content Browser

The greatest advantage of using hand gesture in HCI is its contactless nature. Thus we develop the second demo to introduce its potential applications in interacting virtual reality. By defining different gestures as the command to rotate, move or scale the virtual camera, users can use hand to simply interact with and navigate in the virtual 3D world. Our system detects the user's command first, and then zoom/rotate/translate the camera or the 3D object according to how far the hand moves. Since we use the body skeleton to locate the hand, it is also easy to extend to a two-hand gesture system, which will provide more interactions between the human and computer. An example of the 3D content browser is shown in Fig. 15 together with four gesture commands.

VII. Conclusion

A novel superpixel-based hand gesture recognition system using a novel SP-EMD and depth camera for contactless HCI has been proposed. It is based on a compact representation in the form of superpixels, which efficiently capture the shape, texture and depth features of the gestures. Based on this representation, a novel distance metric, superpixel earth mover's distance (SP-EMD), is proposed as the dissimilarity measurement for gesture recognition. The key partial matching issue is addressed by introducing the concept of virtual superpixels, which serves to model the folded fingers. The effectiveness of the proposed system is illustrated by extensive experiments on three challenging real-life datasets. High mean accuracies (99.1%, 99.6% and 75.8%) and fast recognition speed (average 0.067 second per gesture) for hand gesture recognition is achieved.

Comparing with previous distance measures such as FEMD, shape context distance and path similarity, the proposed SP-EMD metric achieves better performance for hand gesture recognition. Moreover, it is very computationally efficient and thus suitable for real-life HCI applications. Our future research will focus on exploring robust color features for SP-EMD and extending it to dynamic hand gesture, body posture and generic object recognition.

References

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 3, pp. 311–324, Apr. 2007.

[2] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Comput. Vis. Image Understanding*, vol. 108, no. 1–2, pp. 52–73, Oct. 2007.

[3] J. P. Wachs, M. Kolsch, H. Stern, and Y. Edan, "Vision-based hand-gesture applications," *Commun. ACM*, vol. 54, no. 2, pp. 60–71, Feb. 2011.

[4] I. Dejmal and M. Zacksenhouse, "Coordinative structure of manipulative hand-movements facilitates their recognition," *IEEE. Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2455–2463, Nov. 2006.

[5] P. G. Kry and D. K. Pai, "Interaction capture and synthesis," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 872–880, Jul. 2006.

[6] M. H. Yang, N. Ahuja, and M. Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1061–1074, Nov. 2002.

[7] Y. Wu, J. Lin, and T. S. Huang, "Analyzing and capturing articulated hand motion in image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1910–1922, Dec. 2005.

[8] M. B. Kaaniche and F. Bremond, "Recognizing gestures by learning local motion signatures of HOG descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2247–2258, Nov. 2012.

[9] M. Chen, G. AlRegib, and B.-H. Juang, "Feature processing and modeling for 6D motion gesture recognition," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 561–571, Apr. 2013.

[10] N. H. Dardas and E. M. Petriu, "Hand gesture detection and recognition using principal component analysis," in *Proc. CIMSA*, Ottawa, Canada, 2011, pp. 1–6.

[11] N. H. Dardas and N. D. Georganas, "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," *IEEE. Trans. Instrum. Meas.*, vol. 60, no. 11, pp. 3592–3607, Nov. 2011.

[12] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using Kinect sensor," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1110–1120, Aug. 2013.

[13] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas, "Real-time head and hand tracking based on 2.5D data," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 575–585, Apr. 2012.

[14] C. Wang, Z. Y. Zhu, S. C. Chan, and H. Y. Shum, "Real-time depth image acquisition and restoration for image based rendering and processing systems," *J. Signal Process. Syst.*, pp. 1–18, 2013, 10.1007/s11265-013-0819-2.

[15] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2281, Nov. 2012.

[16] H. Liang, J. Yuan, and D. Thalmann, "Parsing the hand in depth images," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1241– 1253, Aug. 2014.

[17] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *Proc. RO-MAN*, Paris, France, 2012, pp. 411–417.

[18] Y. R. Wang, W. H. Lin, and L. Yang, "A novel real time hand detection based on skin-color," in *Proc. ISCE*, Hsinchu, Taiwan, 2013, pp. 141–142.

[19] M. Kolsch and M. Turk, "Robust hand detection," in *Proc. FG*, Seoul, Korea, 2004, pp. 614–619.

[20] B. Stenger, P. R. S. Mendonça, and R. Cipolla, "Model-based 3D tracking of an articulated hand," in *Proc. CVPR*, Kauai, HI, USA, 2001, pp. 310–315.

[21] R. P. Mihail, N. Jacobs, and J. Goldsmith, "Real time gesture recognition with 2 kinect sensors," in *Proc. IPCV*, Las Vegas, NV, USA, 2012, pp. 1–7.

[22] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the Kinect," in *Proc. ICMI*, Alicante, Spain, 2011, pp. 279–286.

[23] C. Zhang, X. Yang, and Y. Tian, "Histogram of 3D facets: A characteristic descriptor for hand gesture recognition," in *Proc. FG*, Shanghai, China, 2013, pp. 1–8.

[24] N. Dalal and B. Triggs, "Histogram of orientated gradients for human detection," in *Proc. CVPR*, San Diego, CA, USA, 2005, pp. 886–893.

[25] L.-C. Wang, R. Wang, D.-H. Kong, and B.-C. Yin, "Similarity assessment model for chinese sign language videos," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 751–761, Apr. 2014.

[26] X. Bai and L. J. Latecki, "Path similarity skeleton graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1282–1292, Jul. 2008.

[27] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[28] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb. 2007.

[29] J. Heikkila, "Geometric camera calibration using circular control points," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1066–1077, Oct. 2000.

[30] P. J. Besl and N. D. McKay, "A method for registration of 3D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[31] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

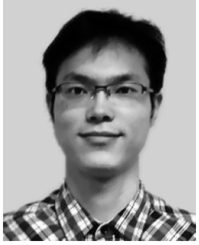
[32] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in *Proc. ICCV Workshops*, Barcelona, Spain, 2011, pp. 1114–1119.

[33] X. Bai, L. J. Latecki, and W.-Y. Liu, "Skeleton pruning by contour partitioning with discrete curve evolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 449–462, Mar. 2007.

[34] X. Bai and L. J. Latecki, "Discrete skeleton evolution," in *Proc. EMCVPR*, Ezhou, China, 2007, pp. 362–374.

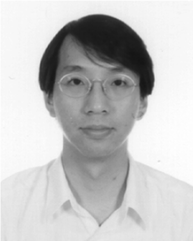
Chong Wang (S'12–M'14) received the B.Eng. degree from Zhejiang University of Technology, Hangzhou, China, in 2007, the M.Eng. degree from the University of Science and Technology of China, Hefei, China, in 2010, and the Ph.D. degree from the University of Hong Kong, Pokfulam, Hong Kong, in 2014.

His main research interests are in depth camera-assisted systems, gesture recognition, image and video restoration, image-based rendering, and parallel computing.



Zhong Liu received the B.Eng. degree from the Huaihai Institute of Technology, Lianyungang, China, in 2010, the M.Eng. degree from Zhongyuan University of Technology, Zhengzhou, China, in 2013, and is currently working towards the M.Phil. degree from the Department of Electrical and Electronic Engineering, University of Hong Kong, Pokfulam, Hong Kong.

His research interests mainly include computer vision, human body tracking and gesture recognition.



Shing-Chow Chan (S'86–M'90) received the B.Sc. (Eng) and Ph.D. degrees from the University of Hong Kong, Pokfulam, Hong Kong, in 1986 and 1992, respectively.

Since 1994, he has been with the Department of Electrical and Electronic Engineering, University of Hong Kong, Pokfulam, Hong Kong, where he is currently a Professor. His research interests include fast transform algorithms, filter design and realization, multirate and biomedical signal processing, communications and array signal processing, high-speed A/D converter architecture, bioinformatics, smart grid, and image-based rendering.

Dr. Chan is currently a member of the Digital Signal Processing Technical Committee of the IEEE Circuits and Systems Society, and Associate Editor of the *Journal of Signal Processing Systems* and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I from 2008 to 2009. He was the Chair of the IEEE Hong Kong Chapter of Signal Processing from 2000 through 2002, an Organizing Committee Member of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, and the 2010 International Conference on Image Processing.