Signal Processing: Image Communication 58 (2017) 87–98

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

A hand gesture recognition system based on canonical superpixel-graph ☆

Chong Wang [a,*], Zhong Liu [b], Minfeng Zhu [c], Jieyu Zhao [a], Shing-Chow Chan [b]

[a] Faculty of EECS, Ningbo University, Ningbo, PR China
[b] Department of EEE, The University of Hong Kong, Hong Kong
[c] College of Control Science & Engineering, Zhejiang University, Hangzhou, PR China

ARTICLE INFO

Keywords:
Hand gesture recognition
Kinect
Superpixel earth mover's distance
Canonical forms

ABSTRACT

This paper presents a new hand gesture recognition system based on a novel canonical superpixel-graph earth mover's distance (CSG-EMD) metric. It aims to improve the performance of the superpixel earth mover's distance (SP-EMD), a recently proposed distance metric designed for depth-based hand gesture recognition. In real life, people have their own habits while performing certain hand gestures, which yields a variety of hand shapes with different finger poses. Such variety may affect the accuracy of SP-EMD and hence will degrade its performance. In this paper, we propose a new distance metric CSG-EMD to alleviate the problem. Scattered superpixels are organized in the form of canonical superpixel-graph which can factor out non-standard finger poses, resulting a well-structured finger-pose-neutral shape representation for hand gestures. Moreover, a structure stress based fusion scheme is applied to formulate the proposed distance metric, i.e. CSG-EMD, for gesture recognition. Experimental results on five public gesture datasets show that the proposed CSG-EMD-based system can achieve better recognition accuracy than other state-of-the-art algorithms compared. Its superiority is further demonstrated by two real-life applications.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Vision-based hand gesture recognition has received great attention in recent years due to its importance in contactless human–computer interaction (HCI) scenarios such as virtual reality, sign language recognition [2,3]. Although a lot of related work has been done [4–8], robust vision-based hand gesture recognition remains a challenging problem. In particular, reliable hand segmentation is essential to gesture recognition and many hand detection techniques [9] have been developed for tracking and recognizing different kinds of hand gestures. With the recent development of depth cameras, such as Microsoft Kinect, Creative Senz3D or Intel RealSense, the work of hand detection and segmentation can be greatly simplified thanks to the extra depth information provided. For instance, threshold-based methods [10–12] are commonly used for efficient hand segmentation. Moreover, some researchers [12,13] take advantage of the skeleton recognized by Kinect for more convenient hand localization and tracking.

Once the hands have been isolated, a variety of hand features can be extracted and used for hand gesture recognition. The hand features can be roughly classified into three categories, i.e. depth-based, color-image-based and shape-based features. Some typical depth-based features are Histogram of 3D Facets (H3DF) [14] and 3D point distribution histogram [15], which are not sensitive to the lighting conditions. One classical but efficient color-based feature descriptor is Histogram of Oriented Gradients (HOG) [16] which is invariant to geometric and photometric transformations. Meanwhile, there are many shape-based features utilized for gesture recognition such as hand contours [2], hand skeleton [17], shape context [18], inner distance [19] and Finger-Earth Mover's Distance (FEMD) [11]. Most of them are generated based on the hand contour, which is, however, usually noisy and distorted due to the low resolution and accuracy of the current depth cameras. In other words, their performance may suffer from ambiguity due to the orientation, distortion and articulation of the hand gestures. Recently, superpixels are utilized as kinds of shape-based features with a new distance metric, namely Superpixel Earth Mover's Distance (SP-EMD) [12], which shows promising performance. However, we found in our experiments that the performance of SP-EMD will be degraded by the variety of the hand shapes resulting from the own habits of different

☆ A preliminary version of this paper appeared in the IEEE International conference on Multimedia and Expo Wang et al. [1].

* Corresponding author.
E-mail addresses: wangchong@nbu.edu.cn (C. Wang), liuzhong@eee.hku.hk (Z. Liu), zjuzmf1996@zju.edu.cn (M.F. Zhu), zhao_jieyu@nbu.edu.cn (J.Y. Zhao), scchan@eee.hku.hk (S.C. Chan).

http://dx.doi.org/10.1016/j.image.2017.06.015
Received 4 November 2016; Received in revised form 27 June 2017; Accepted 27 June 2017
Available online 6 July 2017
0923-5965/© 2017 Elsevier B.V. All rights reserved.

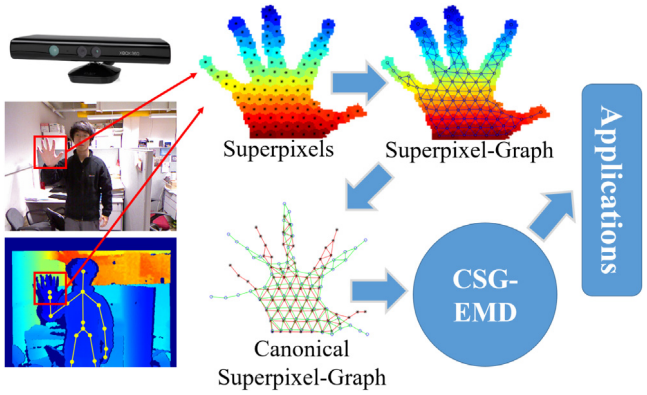

Fig. 1. The workflow of our proposed hand gesture recognition system based on CSG-EMD.

individuals. Hence, it is highly desirable to mitigate such shape variation problem.

Based on our previous work [12], a new variant of SP-EMD, namely canonical superpixel-graph earth mover's distance (CSG-EMD), is proposed in this paper for improving the performance of hand gesture recognition systems. The workflow of the proposed CSG-EMD based system is illustrated in Fig. 1. More precisely, a superpixel-graph is first formed by linking the adjacent scattered superpixels of the hand gesture. Then, the resultant superpixel-graph is further transformed into its canonical form which can factor out non-standard finger poses, yielding a well-organized finger- pose-neutral shape representation for the gesture. Moreover, a structure stress based fusion scheme is applied to formulate the proposed CSG-EMD distance metric which fuses two SP-EMDs of the original and canonical superpixel-graph for accurate hand gesture recognition.

In summary, the current work extends [12] in the following perspectives: (1) the proposed model organizes the scattered hand superpixels in a new canonical superpixel-graph to explore the structural information of hand gestures and developed a new canonical superpixel-graph earth mover's distance matrix for improving the accuracy of hand gesture recognition, (2) a new distance measure, which employs a weighted combination of the previous distance metric SP-EMD and the newly proposed CSG-EMD, is proposed for incorporating the structural information through the structural stress, and (3) the proposed hand gesture recognition algorithm was evaluated extensively on more challenging datasets including the Microsoft Kinect and Leap Motion Dataset [20] and the Creative Senz3D Dataset [21] with improved performance over conventional algorithms tested.

The rest of the paper is organized as follows: Section 2 provides a brief review of related works including the SP-EMD. We describe the main modules of the proposed CSG-EMD based hand gesture recognition system in Section 3. In Section 4, the performance of the proposed system is evaluated with extensive comparisons against the state-of-the-art algorithms. In Section 5, we present two HCI applications using the proposed system for hand gesture recognition to demonstrate its effectiveness. Finally, Section 6 concludes the paper.

2. Related works

We now discuss related works in the following three directions, namely hand segmentation, hand gesture recognition and superpixel earth mover's distance (SP-EMD).

2.1. Hand segmentation

As mentioned earlier, depth cameras simplify the process of hand segmentation, while threshold-based methods [10–12,22] are commonly used for efficient hand segmentation. In these approaches, the interest region corresponding to the hand is segmented based on the assumption that the hand is the closest object in the scene to the camera. Skeleton-based methods [12,13] take advantage of the skeleton recognized by Microsoft Kinect for more convenient hand localization and tracking. More recently, a modified expectation–maximization (EM) algorithm based on Bayesian networks is proposed in [23] for hand gesture segmentation from RGB-D images.

Besides the above-mentioned hand segmentation algorithms, algorithms for general object detection and segmentation with good performance in terms of effectiveness and efficiency have also proposed. They includes: (1) superpixel-based algorithms which include the Linear Spectral Clustering (LSC) algorithm [24] for producing compact and uniform superpixels with low computational costs, and the fast superpixel segmentation algorithm based on edge refinement and revised Wishart distance [25]; (2) machine-learning-based detection algorithm based on learning rotation-invariant convolutional neural networks [26]; and (3) semantic-annotation based method such as automatic semantic annotation of high resolution images in [27]. It is promising to apply these algorithms for hand segmentation and study the performance.

In the current work, we make use of the skeleton provided by the Kinect depth camera to track and segment the depth map of the hands for gesture recognition, which offer a simple and yet efficient method for supporting real-time applications, as demonstrated by our experimental results.

2.2. Hand gesture recognition

The depth information is widely utilized for hand gesture recognition in many recent literatures. Some algorithms are based on local fingers, such as the algorithm proposed in [28], which divided the hand into palm and finger regions and used a multi-class SVM classifier for recognition in real time. On the other hand, the k-curvature based method [22] tries to locate the fingertips over the contour extracted from RGB-D data, while dynamic time warping is used for gesture selection and recognition. A fingertip extraction method was also proposed in [29] which is based on the earth mover's distance and Lasso algorithms.

Besides the local-finger-based algorithms, [30] presents an Image-to-Class Dynamic Time Warping (I2C-DTW) approach for the recognition of both 3D static and trajectory gestures, which shows good recognition performance Some other algorithms focus on a variety of key features. For instance, in [31], a multi-layered gesture recognition method is proposed to sequentially classify the motion, location and shape components by extracting linguistic characters of gestures from both the segmented semantic units and the whole gesture sequence. Moreover, a hybrid system that fused the data obtained from depth cameras and other types of sensors was proposed to improve the recognition performance. In [32], Leap Motion [33] is employed for hand gesture recognition and three Leap Motion features with two Kinect features (3LM + 2K) are used. The numbers of features of Leap motion and Kinect are further extended to 4(4LM + 4K) in [20]. For more recent Kinect-based algorithms for hand gesture analysis, we refer interested readers to [34]. Next we shall briefly review the superpixel earth mover's distance [12] which is a recently proposed distance metric for hand gesture recognition.

2.3. Superpixel earth mover's distance (SP-EMD)

In our previous work [12], SP-EMD was proposed for hand gesture recognition. The hand is first represented in the form of superpixels as shown in Fig. 2(a) and (b), which can be viewed as piles of pixels (Fig. 2(c) and (d)). Then SP-EMD is determined by the minimum effort needed to move pixels from the superpixels of one gesture to another.

Specifically, two hand gestures are defined as $P = \left\{\left(p_1, w_{p_1}\right), \ldots, \left(p_k, w_{p_k}\right)\right\}$ and $Q = \left\{\left(q_1, w_{q_1}\right), \ldots, \left(q_l, w_{q_l}\right)\right\}$ with superpixels $p_i, i = 1, \ldots, k$ and $q_i, i = 1, \ldots, l$ respectively, while the

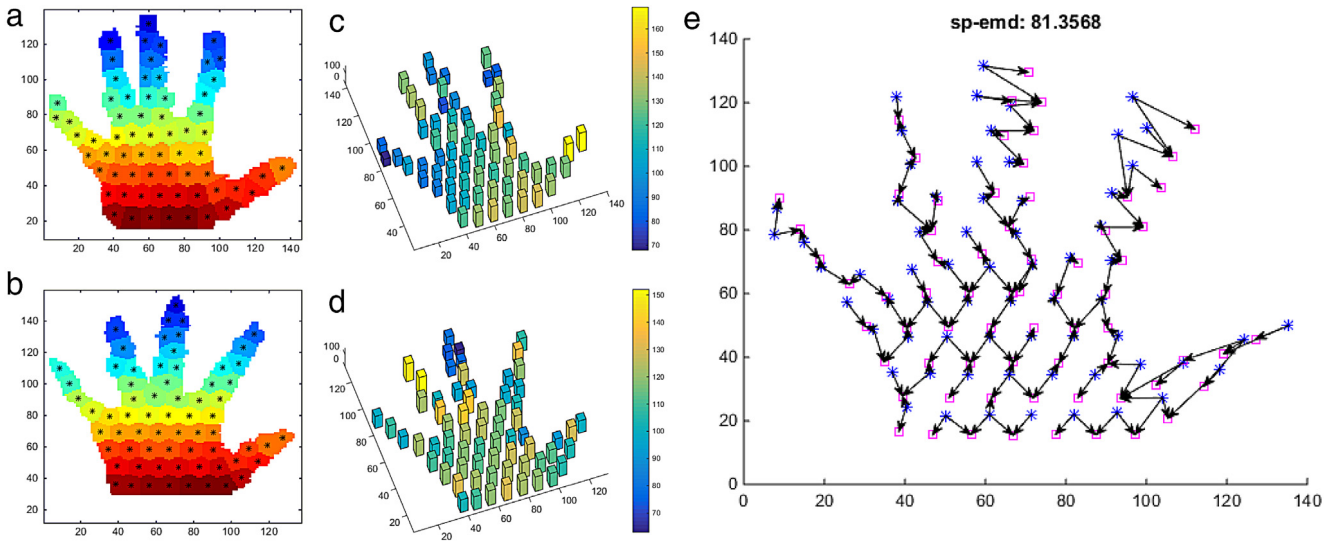

Fig. 2. An illustration of SP-EMD. (a) and (b) are the same hand gestures given by two subjects, in the superpixel representation which are color coded. Black stars are the centroids of corresponded superpixels. (c) and (d) show the number of pixels in the superpixels of gestures (a) and (b), respectively. (e) is the optimal flow from (a) to (b). Blue stars and magenta squares denote the superpixels of gesture (a) and (b), respectively. Black arrows indicate the moving flow directions.

weight w_{p_i} is defined by the number of pixels within the superpixel p_i. The centroid $\left[x_{p_i}, y_{p_i}\right]$ and the depth d_{p_i} are used to define the superpixel $p_i = \left[x_{p_i}, y_{p_i}, d_{p_i}\right]^{\mathrm{T}}$. Then the cost c_{ij} from superpixel p_i to q_j is defined as

$$c_{ij} = \left[\left(x_{p_i} - x_{q_j}\right)^2 + \left(y_{p_i} - y_{q_j}\right)^2 + \alpha\left(d_{p_i} - d_{q_j}\right)^2\right]^{\beta}, \tag{1}$$

where α is the depth weight and β is a nonlinear penalty. To address the partial matching issue, SP-EMD creates two virtual superpixels $p_0 = [0,0,0]^{\mathrm{T}}$ and $q_0 = [0,0,0]^{\mathrm{T}}$ with the weights

$$w_{p_0} = \begin{cases} 0, & w_p \geq w_q \\ w_q - w_p, & \text{otherwise,} \end{cases} \tag{2}$$

$$w_{q_0} = \begin{cases} 0, & w_p \leq w_q \\ w_p - w_q, & \text{otherwise,} \end{cases} \tag{3}$$

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 of P and Q. Then the SP-EMD is determined as,

$$SPEMD(P,Q) = \frac{\sum_{i=0}^{k}\sum_{j=0}^{l} c_{ij} f_{ij}}{\sum_{i=0}^{k}\sum_{j=0}^{l} f_{ij}}, \tag{4}$$

where f_{ij} is the flow from superpixel p_i to q_j, which is obtained by minimizing the cost function

$$\mathbf{F} = \arg\min \sum_{i=0}^{k}\sum_{j=0}^{l} c_{ij} f_{ij},$$
$$\text{s.t.} \begin{cases} f_{ij} \geq 0, & 0 \leq i \leq k,\ 0 \leq j \leq l \\ \sum_{i=0}^{k} f_{ij} = w_{q_j}, & 0 \leq j \leq l \\ \sum_{j=0}^{l} f_{ij} = w_{p_i}, & 0 \leq i \leq k, \end{cases} \tag{5}$$

where $\mathbf{F}$ is the matrix form of the flow f_{ij}. It has been shown in [12] that the SP-EMD approach performs favorably than most conventional algorithms tested. However, we found that the structural information of hand gestures, which is important to further improve the recognition performance, is not considered in this approach. As an illustration, Fig. 2(e) shows an example of the moving flows calculated between the same hand gestures from two different subjects. It can be seen that, the fingers are not fully aligned so that the flows at the finger tips (especially for the index finger) are longer than the ones at other hand parts, which leads to increased SP-EMD. That motivates us to seek canonical forms for the same gestures, in order to enhance the performance of SP-EMD based hand recognition system.

3. Canonical superpixel-graph earth mover's distance

Although SP-EMD shows promising performance for hand gesture recognition, it requires proper preprocessing of the extracted hand. In [12], hand shapes are normalized and aligned according to their depth values and palm centroids, respectively. Meanwhile, out-of-plane rotation of the hand is compensated based on the 3D point cloud. During the process of SP-EMD calculation, 2D Iterative Closest Point (ICP) [35] is employed to match the superpixels for each pair of hand gestures.

The major reason why SP-EMD requires these preprocessing steps is that the cost c_{ij} of moving pixels in (1) is directly determined by the locations of superpixels. In other words, any mismatch between the shapes of two hand gestures will contribute to an increase in SP-EMD. However, due to different user's habits, hand shapes can be quite different even for the same hand gesture. This kind of difference is mainly caused by the variety of finger poses, i.e. the different locations of the fingers' superpixels. As an illustration, Fig. 4(a) presents three samples of the same gesture which are performed by different subjects. It can be seen that the finger poses are slightly different from each other, such as the angles between the thumb and index finger. Since such kind of hand shape mismatch cannot be corrected by the preprocessing steps, the performance of the algorithm reported in [12] will be degraded. To address this issue, we propose a new algorithm equipped with canonical superpixel-graph earth mover distance (CSG-EMD) metric for better hand gesture recognition. Next we describe the key components of CSG-EMD.

3.1. Superpixel-graph

Although hand gestures can be effectively represented in the form of superpixels, scattered superpixels lack the structural information of gestures. Without this global clue, it is difficult to tell which superpixel belongs to the finger and which one belongs to the palm. In other words, it poses a challenge to relocate the superpixels without altering the main structure of the hand gesture.

To address the problem, we propose a new superpixel-graph representation, which is generated by connecting superpixels with their

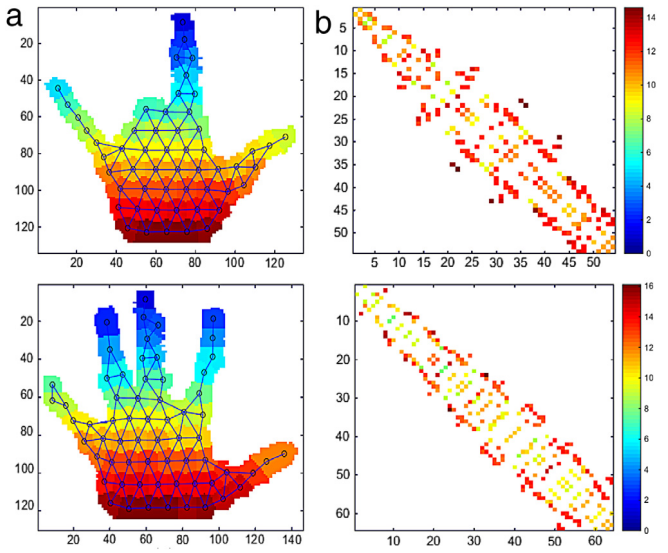

Fig. 3. Shape representation using superpixel-graph. (a) A hand shape represented by superpixel-graph. Black circles and blue lines indicate the centers of superpixels and edges between adjacent superpixels, respectively. (b) The corresponding distance matrix of the superpixel-graph in (a).

neighborhoods as shown in Fig. 3(a). This superpixel-graph is denoted as $G = (V, E)$, which consists of a set of superpixels V together with a set of edges E between adjacent superpixels. To be specific, the set V is the centroids of the superpixels, while $e = (i, j) \in E$ indicates the ith and jth superpixels are adjacent. Moreover, the Euclidean distance between adjacent superpixels can be calculated to form a distance matrix $\mathbf{A}$, whose entry a_{ij} is the distance between the ith and jth connected superpixels as shown in Fig. 3(b). The value of a_{ij} is 0 if the corresponding superpixels are not connected. For the sake of simplicity, the proposed superpixel-graph G is defined as

$$G = (\mathbf{X}, \mathbf{A}), \tag{6}$$

where $\mathbf{X} = [x_1, y_1; x_2, y_2; \ldots; x_N, y_N]$ is a matrix formed by superpixel coordinates and N is the number of superpixels.

3.2. Canonical forms of superpixel-graph

Canonical forms are widely used in non-rigid 3D model retrieval [36–38], which was first proposed by Elad and Kimmel [39]. They use least squares Multi-Dimensional Scaling (MDS) to stretch out a model's limbs so that its extremities can be distant from each other. More precisely, canonical forms are achieved by minimizing the sum of weighted squared errors between a certain dissimilarity metric and the Euclidean distance for all input points. In the case for hand gesture recognition, we aim to use the canonical forms of hand gestures to alleviate the aforementioned hand shape variation problem. Specifically, the geodesic distances between all superpixels are mapped to 2D Euclidean distances using MDS. By doing so, the canonical superpixel-graph $\hat{G} = \left(\hat{\mathbf{X}}, \mathbf{A}\right)$ can be obtained by minimizing the following structural stress $S(\mathbf{X})$,

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} S(\mathbf{X}),$$
$$S(\mathbf{X}) = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_{ij}\left(g_{ij}(\mathbf{X}) - d_{ij}(\mathbf{X})\right)^2}{\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_{ij} g_{ij}^2(\mathbf{X})}, \tag{7}$$

where $\hat{\mathbf{X}}$ is the estimated 2D superpixel coordinates of the canonical superpixel-graph, w_{ij} is a weighting coefficient assigned between ith and jth superpixel, $g_{ij}(\mathbf{X})$ and $d_{ij}(\mathbf{X})$ are the geodesic distance and Euclidean distance between the ith and jth superpixel, respectively.

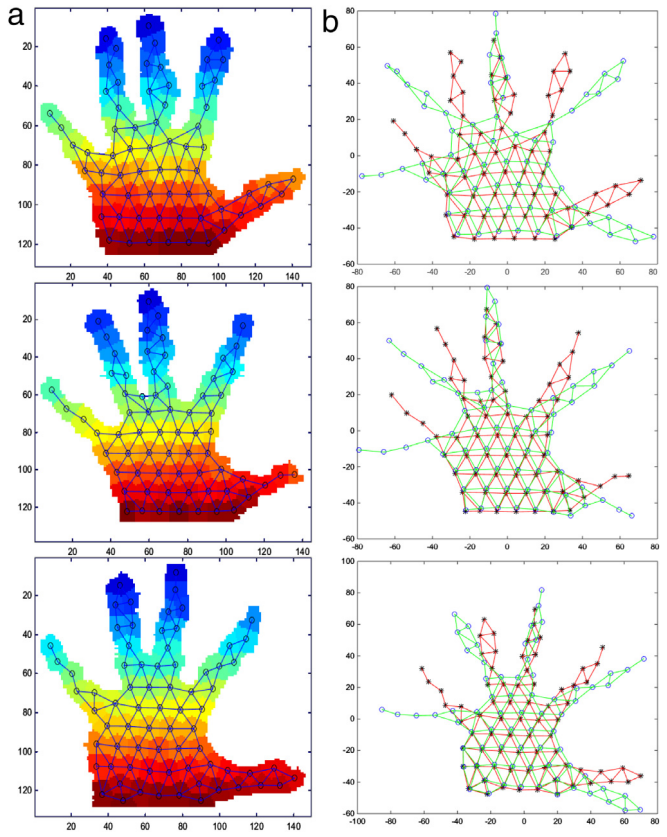

Fig. 4. Examples of the proposed canonical forms for the same hand gestures performed by different subjects. (a) The superpixel-graph representation. (b) Canonical superpixel-graphs. Black stars with red edges are the original hand gesture represented by superpixel-graphs. Blue circles with green edges are their corresponding canonical forms.

In this paper, the geodesic distance is computed using fast marching algorithm [40] over the superpixel-graph. Moreover, the weight w_{ij} is determined according to the distance matrix $\mathbf{A}$, in order to maintain the main structure of the hand gesture,

$$w_{ij} = \begin{cases} \lambda, & a_{ij} > 0 \\ 1, & a_{ij} = 0, \end{cases} \tag{8}$$

where λ is a large constant value.

By applying SMACOF (scaling by maximizing a convex function) algorithm [41], the minimization of the stress function (7) can be solved by iteratively computing the following equations,

$$\mathbf{X}_{t+1} = \mathbf{W}^{+}\mathbf{B}_{\mathbf{X}_t}\mathbf{X}_t, \tag{9}$$
$$\mathbf{W} = \sum_{i=1}^{N}\sum_{j=i+1}^{N} w_{ij}\left(e_i - e_j\right)\left(e_i - e_j\right)^{\mathrm{T}}, \tag{10}$$

where $\mathbf{X}_t$ denotes the estimation of superpixel coordinates after tth iteration, $\mathbf{W}^{+}$ is the Moore–Penrose pseudoinverse of the weighting matrix $\mathbf{W}$ and e_i is the ith column of the identity matrix. The elements of matrix $\mathbf{B}_{\mathbf{X}_t}$ are

$$b_{ij} = \begin{cases} -\dfrac{w_{ij} g_{ij}(\mathbf{X}_t)}{d_{ij}(\mathbf{X}_t)}, & i \neq j \text{ and } d_{ij}(\mathbf{X}) \neq 0 \\ 0, & i \neq j \text{ and } d_{ij}(\mathbf{X}) = 0 \\ \displaystyle\sum_{j=1, j\neq i}^{N} \frac{w_{ij} g_{ij}(\mathbf{X}_t)}{d_{ij}(\mathbf{X}_t)}, & i = j. \end{cases} \tag{11}$$

The superpixel-graph $G_t = (\mathbf{X}_t, \mathbf{A})$ is iteratively updated using (7) until the following criteria is satisfied,

$$S(\mathbf{X}_{t+1}) - S(\mathbf{X}_t) < \varepsilon, \tag{12}$$

where ε is a stopping threshold. To evaluate the efficiency of the SMACOF algorithm, the convergence curves of stress $S(\mathbf{X})$ for ten different gestures are shown in Fig. 5. It can be seen that a small number of iterations, say about 10–15, is sufficient for minimizing the stress $S(\mathbf{X})$.

The final solution will give us the desired canonical formed superpixel-graph representation $\hat{G} = (\hat{\mathbf{X}}, \mathbf{A})$ with a converged stress $\hat{S}(\hat{\mathbf{X}})$. For an illustration, Fig. 4(b) shows comparisons between the original superpixel-graphs and their canonical forms. It can be clearly noticed that the fingers are stretched out as we expect and hence more unified SP-EMD distance metric can be obtained for recognition.

3.3. Canonical superpixel-graph earth mover's distance

We see from the previous discussion that the structural information of the hand gesture is first represented by the superpixel-graph and it is implicitly coded in the canonical forms. The new superpixel locations $\hat{\mathbf{X}}$ will then indicate the finger structures (e.g. fingertips move far away from each other while the palm stays at the previous location.) It is worth noting that the stress $S(\mathbf{X})$ can be viewed as another type of global structural information introduced by the superpixel-graph. As shown in Fig. 6, the stress values of ten different gestures from the HKU hand gesture dataset [12] are distinct from each other to a certain extent, while all five subjects share very similar patterns. Therefore, $\hat{\mathbf{X}}$ and $S(\mathbf{X})$ are the key to formulate the proposed canonical superpixel-graph earth mover's distance (CSG-EMD). Although $\hat{\mathbf{X}}$ is able to represent the hand gesture structure, undesired errors may be introduced due to the articulation issue such as fingers that merged together which cannot be correctly stretched out. To better describe the characteristics of hand gestures, not only the canonical superpixel-graph but also the original scattered superpixels should be utilized to form the CSG-EMD, which is defined as the following weighted combination of two SP-EMDs:

$$\begin{aligned} CSGEMD(H, T_g) = {} & w_{raw} SPEMD(H, T_g) \\ & + w_{mds} SPEMD(\hat{H}, \hat{T}_g), \end{aligned} \tag{13}$$

where $H = (\mathbf{X}, \mathbf{A})$ and $T_g = (\mathbf{X}_g, \mathbf{A}_g)$ are respectively the original data of the input gesture and that of the template in class g, while $\hat{H} = (\hat{\mathbf{X}}, \mathbf{A})$ and $\hat{T}_g = (\hat{\mathbf{X}}_g, \mathbf{A}_g)$ are respectively the canonical forms of H and T_g.

The second structural feature, i.e. the stress $S(\mathbf{X})$, is utilized to construct the stress weights w_{raw} and w_{mds} in order to fuse the SP-EMDs calculated with the original data and its canonical form. The weights are defined as follows

$$w_{raw} = \max\left\{ \frac{S(\mathbf{X})}{S(\mathbf{X}_g)}, \frac{S(\mathbf{X}_g)}{S(\mathbf{X})} \right\}, \tag{14}$$

$$w_{mds} = \max\left\{ \frac{S(\hat{\mathbf{X}})}{S(\hat{\mathbf{X}}_g)}, \frac{S(\hat{\mathbf{X}}_g)}{S(\hat{\mathbf{X}})} \right\}, \tag{15}$$

where $S(\mathbf{X})$, $S(\mathbf{X}_g)$, $S(\hat{\mathbf{X}})$ and $S(\hat{\mathbf{X}}_g)$ are the stress values of the input gesture, the template and their canonical forms, respectively.

The merits of such new distance metric, namely CSG-EMD, are twofold. On one hand, it integrates the canonical superpixel-graphs into the framework of SP-EMD. Thus it gains most advantages of the original SP-EMD, while the gesture variation issue is handled by the proposed canonical forms. On the other hand, the proposed weights (14) and (15) are designed to penalize the inter-class CSG-EMD distance by increasing the SP-EMDs between the potentially dissimilar input gesture and template in order to improve the recognition accuracy. This is motivated by the fact that the same hand gestures shall share similar

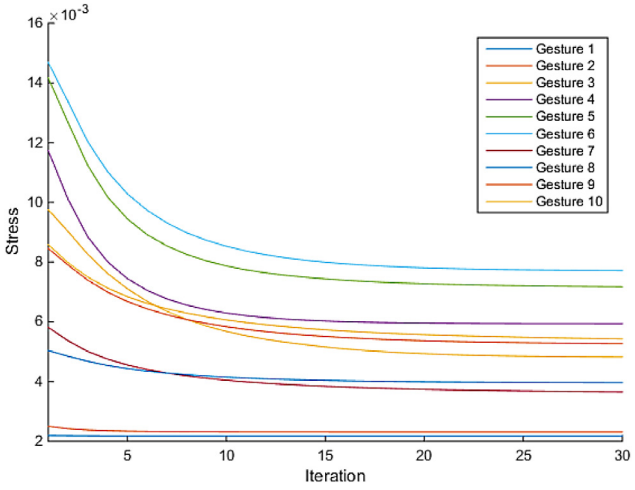

Fig. 5. Convergence curves of the stress $S(\mathbf{X})$. Different color lines indicate different gestures.

superpixel-graph structures with similar stress values. Thus the stress weights w_{raw} and w_{mds} are close to 1 for the gestures from the same class. Therefore, the CSG-EMD should be the sum of the two SP-EMDs without any scaling. In contrast, the weights are larger than 1 if two gestures are different, which increases the CSG-EMD proportionally. As a result, it helps to distinguish dissimilar hand gestures even if they have close SP-EMDs arising from gesture variation. Consequently, those gestures with more distinct stress values will be benefited more from the proposed CSG-EMD.

3.4. Recognition

The k-nearest neighbors (k-NN) classifier with the proposed CSG-EMD is utilized for hand gesture recognition, while k is chosen as 3 in all the following experiments. Since k-NN is a template based approach, the performance is closely related to the selected templates and hence the training data. In our experiments, leave-p-out (LpO) cross-validation (CV) is conducted 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. Specifically, all the hand gesture samples of the $M-p$ subjects are served as the templates for k-NN classification. This process is repeated for every combination of p subjects to compute the average accuracy.

It is worth noting that the hand gesture variation problem will be more significant when a larger value of p is chosen, since less templates are available. Therefore, in our experiments the value of p is set to the maximum value, i.e. $M-1$, to test the performance of the proposed CSG-EMD against the variation issue. Meanwhile, leave-one-out CV (LOO CV), i.e. $p=1$, is also considered in order to give a more comprehensive evaluation.

4. Experimental evaluations

We now evaluate and compare the proposed CSG-EMD based hand gesture recognition system with various state-of-the-art recognition algorithms including the original SP-EMD [12], FEMD [11], Skeleton Matching [17], HOG [15], H3DF [14], S-LBsP [42], Hybrid Features (3LM+2K [32] and 4LM+4K [20]), Distance Features [28], Convex Hull Concavities [43] and Three-Dimensional Feature Array (3DFA) [21], using five different real world datasets, namely NTU hand digit dataset [11], HKU hand gesture dataset [12], HKU multi-angle hand gesture dataset [12], Microsoft Kinect and Leap Motion Dataset [20] and Creative Senz3D Dataset [21]. Two different CV tests

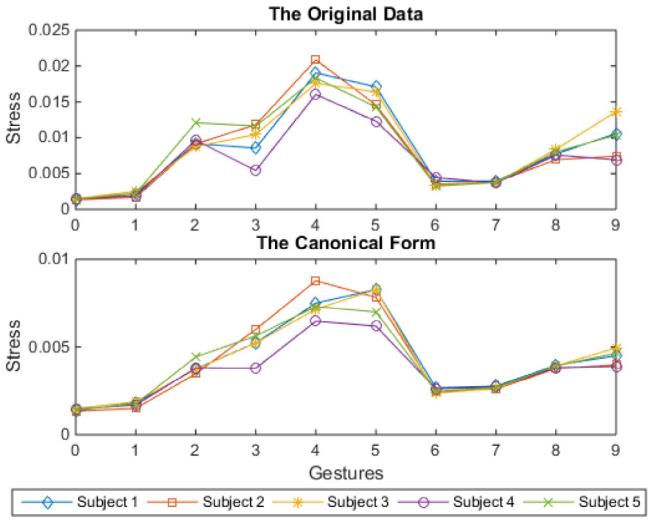

Fig. 6. Stress values of ten different hand gestures from five subjects for the original data (upper) and their canonical forms (lower).

are applied to illustrate the effectiveness of the proposed CSG-EMD distance metric. The confusion cases and parameter sensitivity will also be discussed.

4.1. Experimental setup

The datasets used in the experiments are collected with different depth sensors. They all consist of color images and depth maps of the hand gestures, while some provide additional auxiliary data. The details of each dataset are summarized below:

– *The NTU hand digit dataset* [11] is captured with Microsoft Kinect v1, which contains 1000 cases of 10 hand gestures from 10 subjects.

–*The HKU hand gesture dataset* [12] is also collected with Kinect v1, which contains 10 gestures with 20 different poses from 5 subjects with a total of 1000 cases. The body skeletons are provided for hand segmentation.

– The *HKU multi*-angle *hand gesture dataset* [12] is an extension of the HKU dataset with more challenging hand gesture samples, which are captured from 5 different view angles (roughly 0°, ±10° and ±20°) with 5 subjects. In total, there are 3000 samples for testing.

– *The Microsoft Kinect and Leap Motion Dataset* [20] contains both depth data and Leap Motion data. There are 10 different gestures from 14 different people. Each user has repeated each gesture 10 times for a total of 1400 different data samples.

– *The Creative Senz3D Dataset* [21] is acquired with the Creative Senz3D camera. The dataset contains 11 different gestures performed by 4 different people. Each gesture has been repeated by each user 30 times for a total of 1320 acquisitions. The Senz3D confidence maps are also included in the dataset.

Some gesture samples from these five datasets are shown in Fig. 7, from which we can see that the hand motion is relatively flexible including large in-plane and moderate out-of-plane rotations.

In the experiments, hand shapes are first normalized and aligned according to their depth values and palm centroids, respectively. And 2D ICP [35] is utilized to address the in-plane rotation of the hand. More details can be found in [12]. As mentioned in Section 3.3, LOO CV is conducted for all datasets. Also, leave-4-out CV (L4O CV), leave-9-out CV (L9O CV), leave-13-out CV (L13O CV) and leave-3-out CV (L3O CV) are applied for the two HKU datasets (5 subjects), the NTU dataset (10 subjects), the Kinect & Leap Motion dataset (14 subjects) and the Senz3D dataset (4 subjects), respectively.

Furthermore, we perform another set of experiments that only exploits the shape information to show the effectiveness of the proposed algorithm even if the color and depth information are not available. To be specific, the superpixels are generated solely from hand shapes and depth weight α in (1) is set to 0, while the other steps remain unchanged.

4.2. Performance evaluation

The experiments were conducted on an Intel Core™ i7-5820K 3.6 GHz CPU with 32 GB of RAM. Now we evaluate the performance of the proposed system by comparing with other state-of-the-art methods.

In all the experiments, the depth weight α, the fingertip coefficient β and the average size of superpixels are respectively fixed as 1.0, 2.0 and 81, using the same setup in [12]. For the proposed canonical superpixel graph, the constant weight λ in (8) and threshold ϵ in (12) are set to 100 and 10^{-6}, respectively. Although these two coefficients are set heuristically, we will show in Section 4.3 that the recognition performance is not quite sensitive to the parameters.

4.2.1. Mean accuracy

Depending on the availability of color and depth information, the experiments are divided into two catalogs, i.e. utilizing color textures and depth maps or not. The mean accuracies of the proposed CSG-EMD on different datasets are presented in Tables 1–5.

It can be seen that, in the tests using color and depth information, CSG-EMD achieves higher mean accuracies than the original SP-EMD on all five datasets. It is worth noting that the improvement over SP-EMD in L4O CV (or L9O CV, L13O CV, L3O CV) is noticeable, about 1.2%, 0.6%, 0.9%, 1.7% and 4.5% for five datasets respectively. In contrast, the improvement in LOO CV is slight, which is about 0.2%, 0.2%, 0.1%, 0.9% and 2.4% correspondingly.

The major reason of such big difference is that the number of templates in LOO CV is 4 (or 9, 13, 3) times larger than the ones in L4O CV (or L9O CV, L13O CV, L3O CV). It suggests that the hand gesture variety issue can be alleviated by the various gesture samples from different subjects in LOO CV. In other words, the proposed CSG-EMD will degenerate to the original SP-EMD when a large number of templates are applied. However, in real world it is not always possible to create a complete set of templates. L4O CV (or L9O CV, L13O CV, L3O CV), by contrast, only utilizes one subject's samples as templates, which is much closer to the practical situations and thus more vulnerable to the hand gesture variety issue. In fact, the more subjects involved, the more variety introduced, and thus the more serious the issue will be. This pattern can be observed by comparing the results of L9O CV and L4O CV in Tables 1 and 2, respectively. Although these two datasets have the same amount of samples and similar performance in LOO CV, the accuracy drop from LOO to L9O CV (about 3%–7%) is much bigger than the one from LOO to L4O CV (less than 2%).

The proposed CSG-EMD also shows better performance than the original SP-EMD in those experiments only using the shape information. Comparing with the results of SP-EMD, the accuracy gains in L4O CV (or L9O CV) are about 1%–2%. Again, it shows the capability of CSG-EMD in dealing with the variation problem.

From Tables 2 and 3, we can find that in general the usage of the additional color and depth information is able to enhance the recognition performance, which is also reported in [12]. However, Table 1 shows quite opposite results. For both SP-EMD and CSG-EMD, the mean accuracies using hand shape is higher than the ones with color and depth data, especially for L9O CV. This is mainly caused by the inaccurate alignment between color textures and depth maps as the NTU hand digit dataset [11] does not provide the camera registration parameters. It gives a good example of what will happen if the color and depth information is not reliable.

To further illustrate the advantage of our system, we compare it with several state-of-the-art recognition algorithms, namely FEMD [11], Skeleton Matching [17], HOG [16], H3DF [14], S-LBsP [42], Hybrid Features (3LM + 2K [32] and 4LM + 4K [20]), Distance Features [28], Convex Hull Concavities [43] and 3DFA [21]. It is noted that two

Fig. 7. Gesture samples in (a) NTU hand digit dataset [11], (b) HKU hand gesture dataset [12], (c) HKU multi-angle hand gesture dataset, (d) Microsoft Kinect and Leap Motion Dataset [20] and (e) Creative Senz3D Dataset [21].

Table 1
Performance of FEMD, HOG, H3DF, original SP-EMD and proposed CSG-EMD on the NTU hand digit dataset.

Algorithms	FEMD [11] (Near Convex)	HOG [14,15]	H3DF [14]	S-LBsP [39]	SP-EMD [12] (shape only)	CSG-EMD (shape only)	SP-EMD [12]	CSG-EMD
Mean Accuracy (LOO CV)	93.9%	93.1%	95.5%	97.9%	99.6%	99.6%	99.5%	**99.7**%
Mean Accuracy (L9O CV)	–	–	–	–	94.1%	**96.2**%	91.8%	93.0%

'–' indicates no result was reported and no code is available for implementation.

Table 2
Performance of FEMD, Skeleton Matching, original SP-EMD and proposed CSG-EMD on HKU hand gesture dataset.

Algorithms	FEMD [14] (Thresholding)	Skeleton Matching [17] (DSE [28])	SP-EMD [12] (shape only)	CSG-EMD (shape only)	SP-EMD [12]	CSG-EMD
Mean Accuracy (LOO CV)	95.0%	96.0%	98.7%	**99.4**%	99.2%	**99.4**%
Mean Accuracy (L4O CV)	91.0%	90.5%	96.1%	97.4%	97.3%	**97.9**%

Table 3
Performance of FEMD, Skeleton Matching, original SP-EMD and proposed CSG-EMD on HKU multi-angle hand gesture dataset.

Algorithms	FEMD [11] (Thresholding)	Skeleton Matching [17] (DSE [28])	SP-EMD [12] (shape only)	CSG-EMD (shape only)	SP-EMD [12]	CSG-EMD
Mean Accuracy (LOO CV)	96.2%	95.1%	95.3%	96.1%	97.8%	**97.9**%
Mean Accuracy (L4O CV)	89.7%	90.3%	92.5%	93.7%	94.7%	**95.6**%

Table 4
Performance of Hybrid Features (3LM + 2K [32], 4LM + 4K [20]), original SP-EMD and proposed CSG-EMD on Microsoft Kinect & Leap Motion Dataset.

Algorithms	3LM + 2K [32] (Kinect)	3LM + 2K [32] (Kinect + Leap Motion)	4LM + 4K [20] (Kinect)	4LM + 4K [20] (Kinect + Leap Motion)	SP-EMD [12] (Kinect)	CSG-EMD (Kinect)
Mean Accuracy (LOO CV)	89.7%	91.3%	96.3%	96.5%	95.8%	**96.6**%
Mean Accuracy (L13O CV)	–	–	–	–	82.6%	**84.3**%

Table 5
Performance of distance features, Convex Hull concavities, 3DFA, original SP-EMD and proposed CSG-EMD on Creative Senz3D Dataset.

Algorithms	Distance features [28]	Convex Hull concavities [43]	3DFA [21]	SP-EMD [12]	CSG-EMD
Mean Accuracy (LOO CV) (LOO CV)	75.0%	65.0%	90.0%	95.0%	**97.4**%
Mean Accuracy (L3O CV)	–	–	–	88.2%	**92.8**%

different finger decomposition methods, namely thresholding decomposition and near-convex decomposition, are proposed for FEMD in [11]. In our experiments, the thresholding based method is applied to two HKU datasets, while the other method is compared on the NTU dataset. For Skeleton Matching [17], discrete skeleton evolution (DSE) [44] is applied as the skeleton pruning method, which is more stable to the small protrusions and results a better recognition accuracy than discrete curve evolution (DCE) [45] as reported in [12].

Mean accuracies of the compared algorithms on different datasets are also shown in Tables 1–5. The hand shapes are segmented and preprocessed using the same method as SP-EMD and CSG-EMD adopted. It can be seen that the proposed hand recognition system achieves the highest mean accuracy. Especially, our recognition rate is at least 5% higher than other algorithms (except SP-EMD) for the L4O CV considered in Tables 2 and 3. Note that the L9O CV (or L13O CV, L3O CV) results of some compared algorithms are not available in Tables 1, 4 and 5, since no result was reported and no code is available for implementation.

4.2.2. Confusing cases

Detailed comparisons between the proposed CSG-EMD and the original SP-EMD are presented in Figs. 8 and 9 in the form of the confusion matrices of L4O CV (or L9O CV). More confusion matrices of reported results can be found at our project homepage[1] . Fig. 8 are those experimental results with color textures and depth maps, while Fig. 9 shows those without such additional information.

It can be seen that the most confused cases for SP-EMD is gestures 4 and 9 with a mismatching rate of 12.67% as shown in Fig. 8(a). In contrast, the mismatching rate between these two gestures is greatly reduced to 7.67% in CSG-EMD. Similar improvement can be spotted in Fig. 8(b) and (c) as well. The highest confusion rates are decreased from 7% to 4.5% between gestures 7 and 9 in HKU hand gesture dataset [12], and from 6.33% to 1% in HKU multi-angle hand gesture dataset [12]. Though the mismatching rates of some other gestures may increase slightly for the proposed system, the overall performance is constantly improved as compared with other algorithms. In other words, CSG-EMD is able to considerably enhance the recognition accuracies of the most confused gestures with the cost of a small performance penalty for some other gestures, which is convenient in practical applications.

Similar properties can be found in the experiments that utilize shape information solely as shown in Fig. 9. The mismatching rates between gestures 4 and 3 as well as 4 and 5 in the NTU dataset are remarkably dropped by 4%–5% as we use CSG-EMD instead of SP-EMD for hand gesture recognition. On the HKU hand gesture dataset, the recognition rate of gesture 5 is boosted to 98.25% in CSG-EMD, while 5.5% of its samples are mistakenly recognized as gesture 9 in SP-EMD. Similarly, the mismatching rate between gestures 8 and 9 is dropped from 6.08% to 5.17% in the HKU multi-angle hand gesture dataset.

4.3. Sensitivity analysis

There are three key parameters in SP-EMD, including the depth weight α, fingertip coefficient β and the average size of superpixels N/K. As claimed in [12], the mean accuracy is quite stable when α, β or N/K varies. Since the proposed CSG-EMD is based on the framework of SP-EMD, the recognition performance is not sensitive to these coefficients as well.

The remaining two tunable parameters are the constant weight λ and stopping threshold ϵ as introduced in Section 3.2. Although they are not directly involved in the calculation of CSG-EMD, the stress values and the shapes of canonical superpixel-graphs are highly related to them. To analyze their influence, mean accuracies using the proposed CSG-EMD with various setting of these two parameters are presented in Fig. 10. These tests are conducted on the HKU hand gesture dataset [12] without utilizing the color and depth information. It can be seen that the recognition rate is quite stable even for a wide range of λ (from 10 to 10^4) and ϵ (from 10^{-4} to 10^{-9}). Some examples of the canonical superpixel-graphs are also presented in Fig. 11, from which we can observe that the resultant shapes of the canonical forms generated with different λ and ϵ are quite similar.

It is worth noting that the parameter λ is the weight of original structure. Specifically, a very small value of λ may alter some local structures, while a very large value will obstruct the process of canonical form estimation. As seen in Fig. 11, when $\lambda = 10$ (a small value), the width of fingers is reduced and the recognition accuracy is slightly lower than the ones with a larger λ as shown in Fig. 10(b). On the opposite, when $\lambda = 10^4$ (a very large value), it is more difficult to get the canonical superpixel-graph representation from the original hand gesture, and hence the corresponding recognition performance is noticeably degraded as presented in Fig. 10(a) and (b). Meanwhile, the threshold ϵ determines the number of iterations in MDS as shown in Fig. 10(c). A small threshold can guarantee the convergence with the cost of time, while a large threshold may generate near-canonical forms which have not converged yet. As presented in the first column of Fig. 11, the fingers have not been totally stretched out with a large $\epsilon = 10^{-4}$.

We have also performed two tests with different template selection schemes on the HKU hand gesture dataset [12] without utilizing the color and depth information. It is noticed that the performance is not very sensitive to the selection of the templates. Interested readers are referred to our project homepage[2] for more information.

5. Applications

Hand gesture recognition finds great potential in many emerging applications such as interactive gaming, virtual reality and remote robot controlling over traditional input devices like keyboards and mice. The proposed hand gesture recognition system allows us to efficiently use the hand gesture as a contactless interface for such HCI applications. In this work, two example applications, i.e. Robotic Hand Manipulation and 3D Scene Navigation are designed. A demo video is available at https://youtu.be/QKwWzC2z-uY.

5.1. Robotic hand manipulation

Experiments in this paper show that our system can recognize various gestures with high accuracies. To further demonstrate its practical efficiency, we build a system to manipulate a robotic hand using users' hand gestures. Fig. 12 shows the system setup of this demo. The robotic hand in our demo is 3D printed and assembled based on the InMoov open source models designed by Gaël Langevin [46]. The gestures performed by users are first recognized using the proposed hand gesture recognition system running on a PC. The results are then sent to a single chip microcomputer (STM32) which controls five servo motors (MG996R) to fold or extend corresponding fingers of the robotic hand. It can be seen from the demo that the robotic hand can quickly and accurately follow users' hand gestures, which is close to the one developed by the designer using a Leap Motion [33,46].

5.2. 3D scene navigation

The greatest advantage of using hand gesture in HCI is its contactless nature. To explore its potential applications in interacting virtual reality, we develop the second demo which is referred to as 3D scene navigation. A snapshot of the system is shown in Fig. 13 where the 3D scene is rendered using the textures and depth maps from Kinect in real time. By defining different gestures as the command to rotate, move or scale the virtual camera, users can use hand to simply interact with and

[1] https://sites.google.com/site/spemdkinect/canonical-superpixel-graph.

[2] https://sites.google.com/site/spemdkinect/canonical-superpixel-graph.

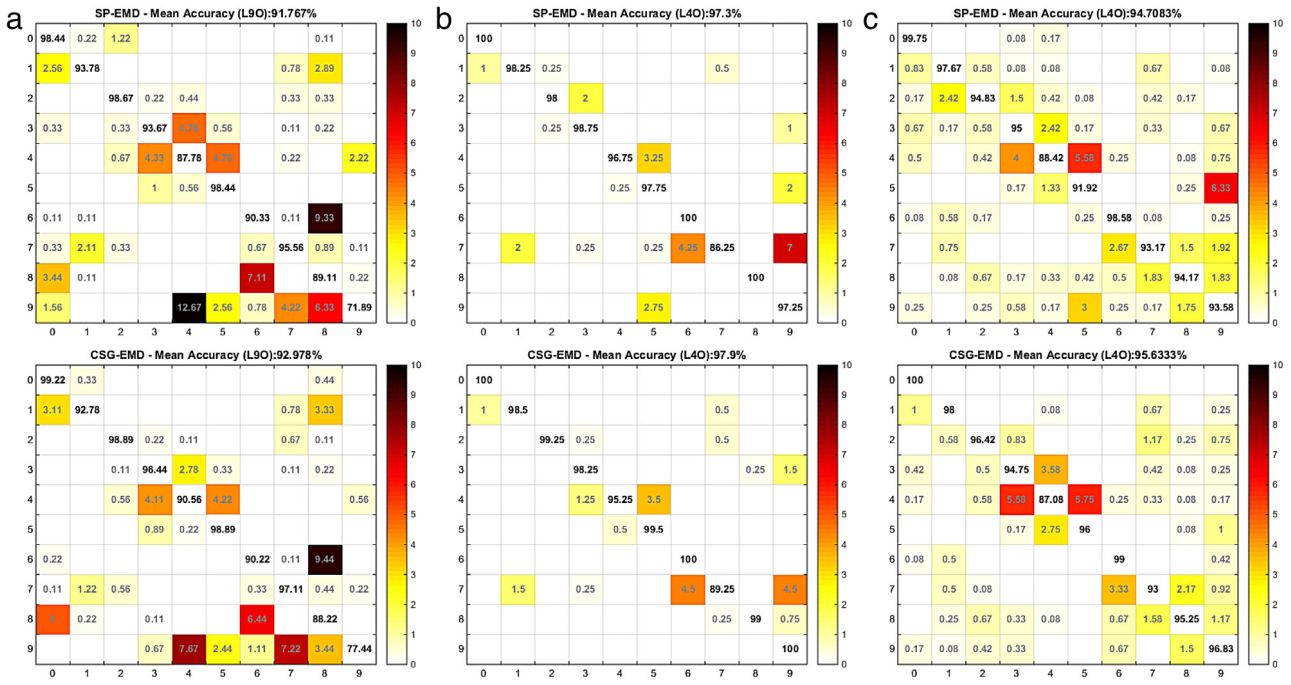

Fig. 8. Confusion matrices *with* color and depth information (unit: %). (a) NTU hand digit dataset [11] with L9O CV. (b) HKU hand gesture dataset [12] with L4O CV. (c) HKU multi-angle hand gesture dataset [12] with L4O CV. The upper and lower rows are results of the original SP-EMD and the proposed CSG-EMD, respectively.

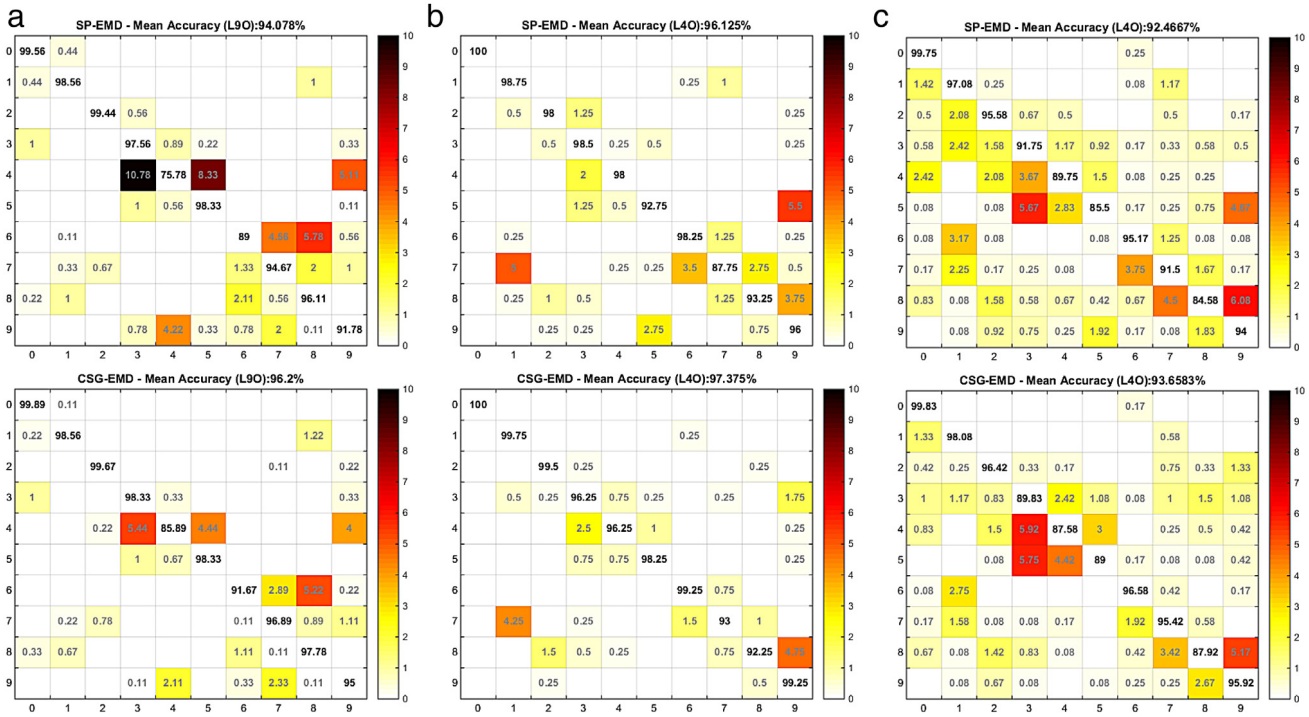

Fig. 9. Confusion matrices *without* color and depth information (unit: %). (a) NTU hand digit dataset [11] with L9O CV. (b) HKU hand gesture dataset [12] with L4O CV. (c) HKU multi-angle hand gesture dataset [12] with L4O CV. The upper and lower rows are results of the original SP-EMD and the proposed CSG-EMD, respectively.

navigate in the virtual 3D world. Our system can recognize the user's commands, i.e. zoom/rotate/reset/stop for 3D scene navigation. Since we use the body skeleton to locate the hand, it is easy to extend to a two-hand gesture system, which can provide more interactions between the human and computer.

6. Conclusion

A new superpixel-graph based system using a novel CSG-EMD for hand gesture recognition with depth camera has been proposed. The hand is represented by a set of connected superpixels, i.e. superpixel-graph, which can preserve the structural information of hand gestures.

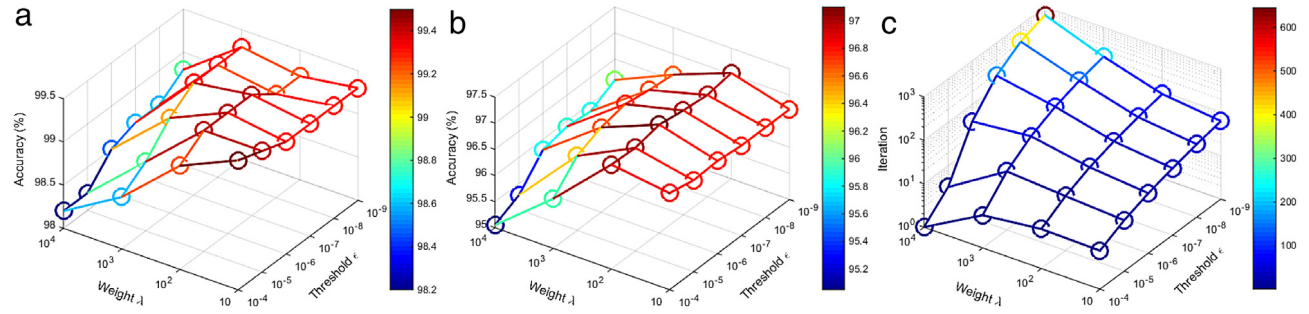

Fig. 10. Performance with different λ and ε on HKU hand gesture dataset [10]. (a) Mean accuracies (shape only) for LOO CV. (b) Mean accuracies (shape only) for L4O CV. (c) The number of iterations in MDS.

Fig. 11. Canonical superpixel-graphs with different λ and ε. From top to bottom, the values of λ are 10, 100, 1000 and 10000. From left to right, the values of ε are 10^{-4}, 10^{-5}, 10^{-6} and 10^{-7}. Black stars with red edges are the original hand gesture represented by superpixel-graphs. Blue circles with green edges are their corresponding canonical forms.

The canonical form of such represented superpixel-graph is estimated using MDS to achieve a well-structured finger-pose-neutral shape representation for hand gestures. Based on this representation, a novel distance metric, canonical superpixel-graph earth mover's distance (CSG-EMD), is further proposed as a unified dissimilarity measurement to alleviate the hand gesture variation problem. The effectiveness of the proposed system is illustrated by extensive experiments on five challenging real-life datasets. High mean accuracies (96.2%, 97.9%, 95.6%, 84.3% and 92.8%) with a small number of templates (L4O CV, L9O CV, L13O, L3O) are achieved.

Comparing with previous distance measures such as SP-EMD, FEMD and path similarity, the proposed CSG-EMD metric achieves better performance for hand gesture recognition. Moreover, it is superior in dealing with hand gesture variation problem, which is very important

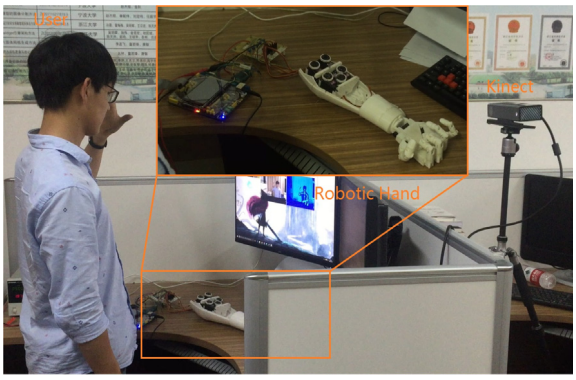

Fig. 12. Robotic hand manipulation.

Fig. 13. Example view (left) and commands (right) of 3D Scene Navigation.

in real-life HCI applications. Our future research will focus on extending the CSG-EMD framework to dynamic hand gesture, body posture and generic object recognition.

Acknowledgments

This work was supported by K.C. Wong Magna Fund in Ningbo University, National Natural Science Foundation of China (61603202, 61571247), the Open Project Program of the State Key Lab of CAD&CG in Zhejiang University (A1606), Zhejiang Provincial Natural Science Foundation of China (LQ16F030001, LZ16F030001, LY17F030002), Zhejiang Open Foundation from Information and Communication Engineering of the Most Important Subjects, China (xkxl1512, xkxl1526), Ningbo Municipal Natural Science Foundation of China (2016A610070), Research Foundation of Education Department of Zhejiang Province, China (Y201533827) and the General Research Fund (GRF) of Hong Kong Research Grant Council (RGC) (HKU710611).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.image.2017.06.015.

References

[1] C. Wang, Z. Liu, J. Zhao, Hand gesture recognition based on canonical formed superpixel earth mover's distance, in: Proc. ICME, Seattle, 2016, pp. 1–6.
[2] L.C. Wang, R. Wang, D.H. Kong, B.C. Yin, Similarity assessment model for chinese sign language videos, IEEE Trans. Multimedia 16 (3) (2014) 751–761.
[3] Y. Park, J. Kim, K. Lee, Effects of auditory feedback on menu selection in hand-gesture interfaces, IEEE Trans. Multimedia 22 (1) (2015) 32–40.
[4] G. Wu, W.X. Kang, Robust fingertip detection in a complex environment, IEEE Trans. Multimedia 18 (6) (2016) 978–987.
[5] S. Mitra, T. Acharya, Gesture recognition: a survey, IEEE Trans. Syst. Man Cybern. C, Appl. Rev. 37 (3) (2007) 311–324.
[6] P.K. Pisharady, M. Saerbeck, Recent methods and databases in vision-based hand gesture recognition, Comput. Vis. Image Underst. 141 (C) (2015) 152–165.
[7] H. Leea, S.Y. Lima, I. Leea, J. Chaa, D.-C. Chob, S. Choc, Multi-modal user interaction method based on gaze tracking and gesture recognition, Signal Process., Image Commun. 28 (2) (2015) 114–126.
[8] C. Nyirarugira, T.Y. Kim, Stratified gesture recognition using the normalized longest common subsequence with rough sets, Signal Process., Image Commun. 30 (1) (2015) 178–189.
[9] S.S. Rautaray, A. Agrawal, Vision based hand gesture recognition for human computer interaction: a survey, Artif. Intell. Rev. 43 (1) (2015) 1–54.
[10] H. Cheng, L. Yang, Z. Liu, A Survey on 3D hand gesture recognition, IEEE Trans. Circuits Syst. Video Technol. (2015) 1–14. http://dx.doi.org/10.1109/TCSVT.2015.2469551.
[11] Z. Ren, J. Yuan, J. Meng, Z. Zhang, Robust part-based hand gesture recognition using kinect sensor, IEEE Trans. Multimedia 15 (5) (2013) 1110–1120.
[12] C. Wang, Z. Liu, S.C. Chan, Superpixel-based hand gesture recognition with kinect depth camera, IEEE Trans. Multimedia 17 (1) (2015) 29–39.
[13] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, P. Presti, American sign language recognition with the kinect, in: Proc. ICMI, Alicante, 2011, pp. 279–286.
[14] C. Zhang, X. Yang, Y. Tian, Histogram of 3D facets: A characteristic descriptor for hand gesture recognition, in: Proc. FG, Shanghai, 2013, pp. 1–8.
[15] R.P. Mihail, N. Jacobs, J. Goldsmith, Real time gesture recognition with 2 kinect sensors, in: Proc. IPCV, Las Vegas, 2012, pp. 1–7.
[16] N. Dalal, B. Triggs, Histogram of orientated gradients for human detection, in: Proc. CVPR, San Diego, CA, 2005, pp. 886–893.
[17] X. Bai, L.J. Latecki, Path similarity skeleton graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 30 (7) (2008) 1282–1292.
[18] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[19] H. Ling, D.W. Jacobs, Shape classification using the inner-distance, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2) (2007) 286–299.
[20] G. Marin, F. Dominio, P. Zanuttigh, Hand gesture recognition with jointly calibrated leap motion and depth sensor, Multimedia Tools Appl. 75 (22) (2016) 14991–15015.
[21] A. Memo, P. Zanuttigh, Head-mounted gesture controlled interface for human-computer interaction, Multimedia Tools Appl. (2016) 1–27. http://dx.doi.org/10.1007/s11042-016-4223-3.
[22] G. Plouffe, A.M. Cretu, Static and dynamic hand gesture recognition in depth data using dynamic time warping, IEEE Trans. Instrum. Meas. 65 (2) (2016) 305–316.
[23] Z. Ju, X. Ji, J. Li, H. Liu, An integrative framework of human hand gesture segmentation for human–robot interaction, in: IEEE Systems Journal, Vol. PP, no. 99, pp. 1–11, http://dx.doi.org/10.1109/JSYST.2015.2468231.
[24] Z. Li, J. Chen, Superpixel segmentation using Linear Spectral Clustering, in: Proc. CVPR, Boston, MA, 2015, pp. 1356–1363.
[25] Y. Zhang, H. Zou, T. Luo, X. Qin, S. Zhou, K. Ji, A Fast superpixel segmentation algorithm for PolSAR images based on edge refinement and revised wishart distance, Sensors 16 (10) (2016) 1687 1-22.
[26] G. Cheng, P. Zhou, J. Han, Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 54 (12) (2016) 7405–7415.
[27] X. Yao, J. Han, G. Cheng, X. Qian, L. Guo, Semantic annotation of high-resolution satellite images via weakly supervised learning, IEEE Trans. Geosci. Remote Sens. 54 (6) (2016) 3660–3671.
[28] F. Dominio, M. Donadeo, P. Zanuttigh, Combining multiple depth-based descriptors for hand gesture recognition, Pattern Recognit. Lett. 50 (C) (2014) 101–111.
[29] Z. Ju, D. Gao, J. Cao, H. Liu, A novel approach to extract hand gesture feature in depth images, Multimedia Tools Appl. 75 (19) (2016) 11929–11943.
[30] H. Cheng, Z. Dai, Z. Liu, Y. Zhao, An image-to-class dynamic time warping approach for both 3D static and trajectory hand gesture recognition, Pattern Recognit. 55 (C) (2016) 137–147.
[31] F. Jiang, S. Zhang, S. Wu, Y. Gao, D. Zhao, Multi-layered gesture recognition with kinect, J. Mach. Learn. Res. 16 (1) (2015) 227–254.
[32] G. Marin, F. Dominio, P. Zanuttigh, Hand gesture recognition with Leap Motion and Kinect devices, in: Proc ICIP, 2014, Paris, France.
[33] Leap Motion, Online: https://www.leapmotion.com/.
[34] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: a review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334.
[35] P.J. Besl, N.D. McKay, A method for registration of 3D shapes, IEEE Trans. Pattern Anal. Mach. Intell. 14 (2) (1992) 239–256.
[36] D. Pickup, X. Sun, P.L. Rosin, R.R. Martin, Z. Cheng, S. Nie, L. Jin, Canonical forms for non-rigid 3D shape retrieval, in: Proc. 3DOR, Zurich, 2015, pp. 1–8.
[37] M. Ben-Chen, C. Gotsman, Characterizing shape using conformal factors, in: Proc. 3DOR, Crete, 2008, pp. 1–8.
[38] D. Pickup, X. Sun, P.L. Rosin, R.R. Martin, Euclidean-distance-based canonical forms for non-rigid 3D shape retrieval, Pattern Recognit. 48 (8) (2015) 2500–2512.
[39] A. Elad, R. Kimmel, On bending invariant signatures for surfaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (10) (2003) 1285–1295.
[40] R. Kimmel, J.A. Sethian, Computing geodesic paths on manifolds, in: Proc. Nat'l Academy of Science, 1998, Vol. 95: pp. 8431–8435.
[41] I. Borg, P. Groenen, Modern Multidimensional Scaling: Theory and Applications, in: Springer Series in Statistics, Springer, New York, 2005.

[42] A.I. Maqueda, C.R. del Blanco, F. Jaureguizar, N. García, Temporal pyramid matching of local binary subpatterns for hand-gesture recognition, IEEE Signal Process. Lett. 23 (8) (2016) 1037–1041.
[43] F. Dominio, G. Marin, M. Piazza, P. Zanuttigh, Feature descriptors for depth-based hand gesture recognition, in: Computer Vision and Machine Learning with RGB-D Sensors, Springer International Publishing, Springer, New York, 2014, pp. 215–237.
[44] X. Bai, L.J. Latecki, Discrete skeleton evolution, in: Proc. EMCVPR, Ezhou, 2007, pp. 362–374.
[45] X. Bai, L.J. Latecki, W.-Y. Liu, Skeleton pruning by contour partitioning with discrete curve evolution, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 449–462.
[46] InMoov, Online: http://www.inmoov.fr/.