SPFEMD: SUPER-PIXEL BASED FINGER EARTH MOVER'S DISTANCE FOR HAND GESTURE RECOGNITION

Yiwei Wang[†], *Cheolkon Jung*[†], *Inyong Yun*[‡] *and Joongkyu Kim*[‡]

[†]School of Electronic Engineering, Xidian University, Xian, Shaanxi 710071, China
[‡]Dept. EECE, Sungkyunkwan University, Suwon, Gyeonggi-do 16419, Korea
zhengzk@xidian.edu.cn

ABSTRACT

In this paper, we propose super-pixel based finger earth mover's distance (SPFEMD) for hand gesture recognition. For finger representation, we design SPFEMD for similarity measurement between fingers and hand gestures, and use it as the distance metric for hand gesture recognition. First, we extract hands without any user interaction using both color and depth from Kinect camera. Then, we obtain the seed for hand segmentation on the depth map and segment hand on the color image using the seed. Since fingers contain distinct features for hand gesture recognition, we decompose the hand segment into palm and fingers based on morphological operation. Finally, we perform hand gesture recognition from fingers based on SPFEMD. Experiments on publicly available and our own data sets show the superiority of the proposed method over state-of-the-arts in terms of accuracy and confusion matrices.

Index Terms— Hand gesture recognition, Kinect data, finger representation, morphological operation, superpixel segmentation.

1. INTRODUCTION

Hand gesture recognition has received much attention in recent years due to its applications to human-computer interaction (HCI), sign language translation and virtual reality [1–3]. This is a challenging task since hands have a high degree of freedom in poses and vary with viewpoints. Moreover, the hand appearance differs from person to person, and is significantly influenced by illumination. All these problems make it hard to implement hand gesture recognition for a general use. To address them, various methods for hand gesture recognition have been proposed so far [4]. Although image-based techniques have been widely studied, they are easily affected by lighting conditions and large variations of hand gestures and textures. In particular, reliable hand detection is required to perform hand gesture recognition. To track and recognize various hand gestures [5], hand color model [6] and hand shape model [7] were proposed. However, they were not robust to the dynamic environment highly depending on the models. Recently, there is a significant progress in hand gesture recognition with the advent of depth cameras such as Kinect devices. Once the hand is localized and segmented using depth maps, hand gesture recognition aims to interpret gestures into a certain sign using pattern classifiers such as k-Nearest Neighbors (kNN) [8], Hidden Markov Models [9], Principal Component Analysis (PCA) [10] and Support Vector Machine (SVM) [11]. Ren et al. [12] proposed an effective gesture recognition method based on the Earth Mover's Distance (EMD) and Template Matching Method (TMM), which shows outstanding performance.

In this paper, we propose super-pixel based finger earth mover's distance (SPFEMD) for hand gesture recognition. Hand gestures are closely related to the change of fingers to convey information. Thus, we perform finger extraction using morphological opening operation. Based on the super-pixel finger representation, we provide SPFEMD to measure the dissimilarity between fingers and hand gestures by calculating the earth mover's distance of super-pixel finger points. Fig. 1 illustrates the entire diagram of the proposed hand gesture recognition based on Kinect data. Compared with existing methods, main contributions of the proposed method are as follows: (1) We perform automatic hand segmenta-

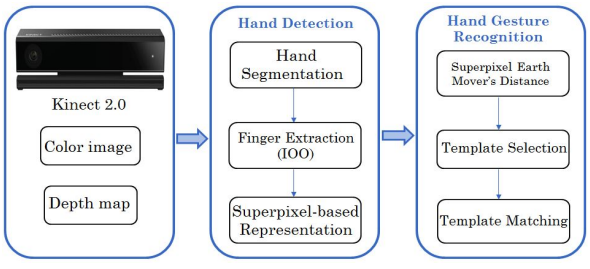

Fig. 1. Entire diagram of the proposed hand gesture recognition based on Kinect camera. IOO: Image opening operation.

This work was supported by the National Natural Science Foundation of China (No. 61872280), the International S&T Cooperation Program of China (No. 2014DFG12780) and the International Cooperation Program of Korea (No. NRF-2013K1A3A1A20047102).

tion without any user interaction using depth information; (2) We extract fingers from each hand segment by morphological opening operation; and (3) We propose SPFEMD for similarity measurement between fingers and hand gestures.

2. HAND DETECTION

Hand detection is a key process for hand gesture recognition and includes hand segmentation, finger extraction and representation. In this work, we use the depth map captured by Kinect camera for hand detection without any user interaction and extract fingers by the image opening operation (IOO). Although the color image and its depth map are captured by Kinect camera simultaneously, they are not aligned well each other. Therefore, a calibration process is required to jointly utilize color and depth information. We use Heikkilas method [13] to calibrate Kinect's color and infrared (IR) sensors. Since Kinect depth map is generated from the IR sensor, the estimated color-IR camera parameters are utilized to calibrate depth map and color image [14]. Many hand segmentation approaches [12] assumed that the hand is the frontal object from Kinect camera. Thus, this assumption allows us to quickly separate the hand from the background based on depth by Otsu's segmentation technique [15]. The threshold for binarization is determined by analyzing the standard deviation of depth values on the depth map. Thus, we get the binarized depth map which contains foreground and background, i.e. foreground segmentation. To detect the seed for automatic hand segmentation, we find the largest connected region on the binarized depth map which refers to a complete area of the hand in the foreground. Algorithm 1 describes the seed selection procedure on the depth map. From the selected seed, we obtain the hand segmentation result on the color image based on Gibbs Random Field (GRF) [16]. IOO is a simple but effective method that decomposes the hand into natural primitives such as fingers and palm. We remove the fingers by erosion operation because fingers are thinner than the palm. For erosion, we use the circle whose size is determined by the maximum distance in the seed selection procedure. Meanwhile, the palm becomes also small after erosion. Thus, we repair the small palm completely by dilation operation. In the end, we decompose the hand into fingers and palm from the segmented hand. Fig. 2 shows hand segmentation results and their decomposition into fingers and palm. To extract features for finger representation, we perform super-pixel segmentation. We use the Simple Linear Iterative Clustering (SLIC) for super-pixel segmentation [17] that effectively enforces spatial compactness of superpixels. To improve its robustness, we use the location information (x, y) and its depth value d. Assume that an image with N pixels is segmented into K superpixels. Each superpixel should have nearly equal size of N/K pixels. Let $u_i = [x_i, y_i]^{\mathrm{T}}$, and the pixel-to-pixel distance is measured as follows:

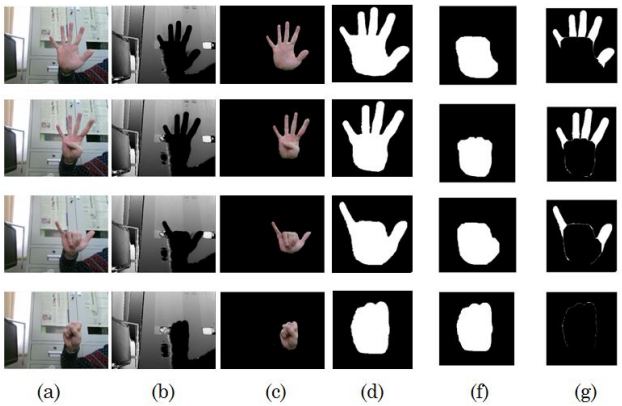

Fig. 2. Hand segmentation and decomposition results. (a) Color image. (b) Depth map. (c) Hand segmentation result. (d) Binarized map of (c). (f) Palms. (g) Fingers.

$$D_s = d + \frac{c}{N/K} d_{xy} \tag{1}$$

where c is the compactness coefficient of superpixels; $d = d_i - d_j$ and $d_{xy} = ||u_i - u_j||$ are depth and spatial distances, respectively; and $||.||$ is Euclidean norm. Thus, D_s is a weighted sum of the depth and spatial distances adjusted by the compactness coefficient c. Some examples are shown in Fig. 3.

3. HAND GESTURE RECOGNITION

The Earth Mover's Distance (EMD) is a measure of the distance between two probability distributions over a region [12]. It has been widely used in image retrieval and pattern recognition. Based on EMD, we propose a novel metric for similarity measurement, the Superpixel Finger Earth Mover's Distance (SPFEMD), to recognize the hand gesture. As shown in Fig. 4, the Finger Earth Mover's Distance (FEMD) only uses the

Algorithm 1 Automatic Seed Selection

Input: Binarized depth map;
Output: Image with the seed point.

Step 1: Mark all the connected regions;
Step 2: Calculate the size of the connected region;
  2.1: Select a point (x_0, y_0) as an initial growing point on the foreground;
  2.2: Consider its 8 neighborhood pixels (x, y) for (x_0, y_0)
    if (x, y) is the foreground points:
      yes: Accepted as a new growing point;
      else: Eliminate this point.
Step 3: Repeat Step 2 until all foreground points are detected;
Step 4: Assign the index to the largest connected region;
Step 5: Get the largest connected region;
Step 6: Detect the seed point on the the largest connected region;
  6.1: Select an initial point p in the largest connected region;
  6.2: Calculate the minimum distance D_p to the region boundary;
  6.3: Point of the maximum distance is selected as the seed P_s.
Step 7: Return the image with P_s.

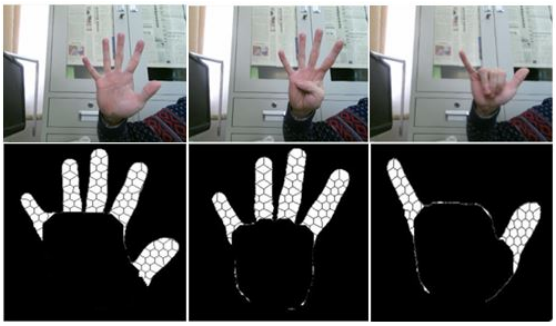

Fig. 3. Finger shape representation based on super-pixel segmentation. Top: Color images. Down: Finger shapes.

contour [12], but SPFEMD jointly measures the similarity between two hand gestures based on finger shape, location and depth information.

3.1. SPFEMD

We define the signature by fingers by a set of super-pixels $p_i, i = 1, 2, ..., k$ with the corresponding weight w_{p_i}. Formally, let $P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), ..., (p_k, w_{p_k})\}$ be the first finger signature with k superpixels, and $Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), ..., (q_l, w_{q_l})\}$ be the second finger signature with l clusters. The centroid $[x_{p_i}, y_{p_i}]^{\mathrm{T}}$ and the average depth value d_{p_i} are used to define the super-pixel $p_i = [x_{p_i}, y_{p_i}, d_{p_i}]^{\mathrm{T}}$. Compared with the texture, the depth is insensitive to illumination changes. The number of pixels, m_i, within the super-pixel is used to denote the cluster weight w_{p_i}. Then, the cost c_{ij} from the super-pixel p_i to q_j is defined as the weighted 3D distance:

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

where α is the depth weight that balances the significance between the 2D shape and depth, and β is a nonlinear fingertip coefficient.

3.2. Template Selection and Matching

Denote g as the number of finger and thus g ranges [0,5]. We classify the hand gesture into six groups. Then, we select the templates which have the same number of fingers according to the testing gesture. For template matching, we randomly select *K* samples from T_g to be the initial medoids $S^0_{C_i}$, then compare with medoids one-by-one using the next alternating two steps. In the first step, we calculate the SPFEMD distance, SPFEMD$(S_m, S^0_{C_i})$, for all samples and clusters which have the same number of fingers. A hand gesture sample S_m is assigned to the i-th cluster $S^t_{C_j}$ if the following condition is satisfied:

$$\text{SPFEMD}(S_m, S^t_{C_i}) \leq \text{SPFEMD}(S_m, S^t_{C_j}), \forall j, j = 1, ..., K, \quad (3)$$

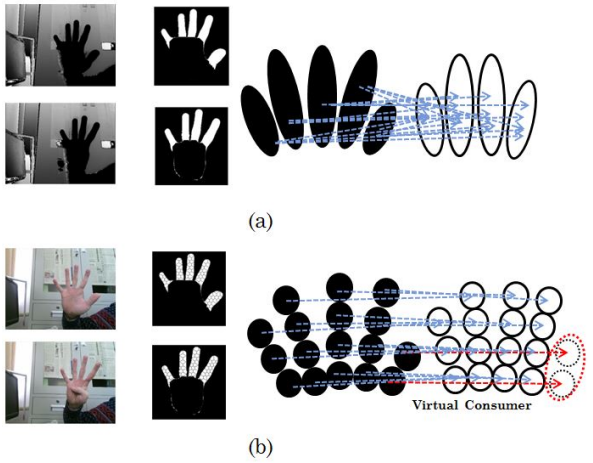

Fig. 4. Comparison between FEMD [12] and the proposed SPFEMD for similarity measurement. (a) FEMD. (b) SPFEMD.

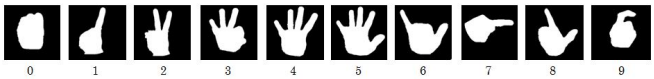

Fig. 5. Gesture samples in our experiments which are labeled from 0 to 9.

where S_m is assigned to one cluster and t is the iteration index starting from 0. The total cost, $c_{med}(\mathbf{C}^t)$, is computed as follows:

$$c_{med}(\mathbf{C}^t) = \sum_{i=1}^{K} \sum_{S_m \epsilon C_i} \text{SPFEMD}(S_m, S^t_{C_i}) \quad (4)$$

In the second step, the medoids are updated by the gesture samples to minimize the sum of distances within the corresponding clusters as follows:

$$S^{t+1}_{C_i} = arg\min_{S_m} \sum_{i=1}^{K} \sum_{S_n \epsilon C_i / S_m} \text{SPFEMD}(S_m, S_n), S_m \epsilon C_i \quad (5)$$

The iteration stops until $c_{med}(\mathbf{C}^t) \geq c_{med}(\mathbf{C}^{t-1})$, and the final medoids in $\mathbf{C}^{t-1}$ are the primitive templates.

4. EXPERIMENTAL RESULTS

We perform experiments in a PC with an Intel Core i7-6700 3.40 GHz CPU and 8 GB of RAM. In all experiments, we set the depth weight α to 1.0 so that balances the significance between the 2D shape and depth, and the fingertip coefficient β to 2.0. We set the average size of super-pixels to 81, i.e. 9×9. To evaluate the performance of the proposed method, we use two data sets – Hand Gesture Image Data sets made by the University of Padova [19] and our own data set captured

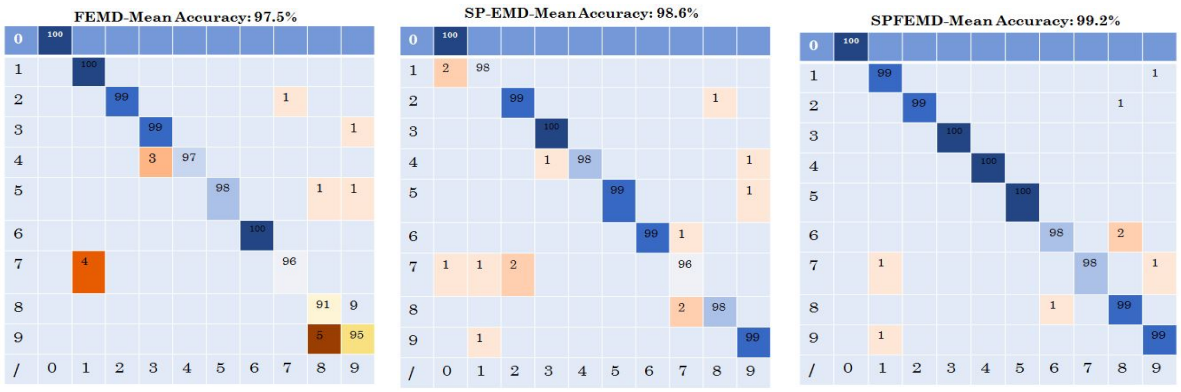

Fig. 6. Confusion matrix on our own data set: Performance comparison between different methods of FEMD [12], SP-EMD [18] and the proposed SPFEMD (unit: %).

Table 1. Mean accuracy of FEMD [12], SP-EMD [18] and the proposed SPFEMD on our own data set

Algorithms	Mean Accuracy	
	L4O CV	LOO CV
FEMD [12] (thresholding)	95.0%	97.5%
SP-EMD [18] (shape only)	97.1%	98.6%
SPFEMD (proposed)	97.7%	99.2%

Table 2. Mean accuracy of FEMD [12], SP-EMD [18] and the proposed SPFEMD on a publicly available data set [19]

Algorithms	Mean Accuracy	
	L4O CV	LOO CV
FEMD [12] (thresholding)	91.3%	93.9%
SP-EMD [18] (shape only)	96.9%	98.3%
SPFEMD (proposed)	97.3%	98.6%

by Kinect V2. It contains 10 gestures with 10 different poses from 2 subjects. Therefore, there are a total of 200 cases for testing, and each of them consists of a pair of color texture and depth map. Gesture samples are shown in Fig. 5, and are labeled from 0 to 9. The public Kinect gesture data sets contain 1,000 cases of 10 hand gestures from 10 subjects.

We compare the performance of the proposed SPFEMD with those of two state-of-the-arts: Threshold-based FEMD method [12] and the shape-based SP-EMD [18]. To illustrate the effectiveness of the proposed SPFEMD, the mean accuracy for comparison between confusion matrices of Leave-4-Out Cross Validation (L4O) and Leave-One-Out Cross Validation (LOO) on the public database and our own data sets is provided in Tables 1 and 2, respectively. It can be observed that the proposed SPFEMD achieves the best performance in both database. As listed in Table 1, the mean accuracy for the proposed method is 97.7% for L4O CV and 99.2% for LOO CV. Moreover, Table 2 shows good performance of the proposed method in mean accuracy on the publicly available data set [19]. It can be observed that LOO CV achieves better recognition rates than L4O CV in both data sets because the number of training data (or templates) in the former is 4 times larger. We provide the comparison of confusion matrices LOO on our own data set in Fig. 6. The most confusing cases are between gestures 1, 7 and 9 due to having the same number of fingers. Sometimes, two fingers are fused into one due to the distortion. That causes the error that gesture 4 is also wrongly recognized as gesture 3.

5. CONCLUSION

In this paper, we have proposed SPFEMD for hand gesture recognition based on Kinect data. First, we have realized fully automatic hand segmentation without any user interaction using both color and depth information from Kinect camera. Second, we have designed SPFEMD for similarity measurement to effectively capture finger shape, location and depth from hand gestures. Finally, we have performed hand gesture recognition based on SPFEMD and template matching. Experimental results demonstrate that the proposed SPFEMD achieves outstanding gesture recognition accuracy over state-of-the-art methods. Our future work includes investigating real-time hand gesture recognition for immersive interactive virtual reality (VR).

6. REFERENCES

[1] Sushmita Mitra and Tinku Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[2] Xinghao Chen, Hengkai Guo, Guijin Wang, and Li Zhang, "Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition," in *Proceedings of the IEEE Conference on Image Processing*, 2017, pp. 2881–2885.

[3] Guillaume Devineau, Fabien Moutarde, Wang Xi, and Jie Yang, "Deep learning for hand gesture gecognition on skeletal data," in *Proceedings of the IEEE Conference on Automatic Face & Gesture Recognition*, 2018, pp. 106–113.

[4] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[5] Juan Pablo Wachs, Mathias Kölsch, Helman Stern, and Yael Edan, "Vision-based hand-gesture applications," *Communications of the ACM*, vol. 54, no. 2, pp. 60–71, 2011.

[6] Ming-Hsuan Yang, Narendra Ahuja, and Mark Tabb, "Extraction of 2d motion trajectories and its application to hand gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1061–1074, 2002.

[7] Ying Wu, John Lin, and Thomas S Huang, "Analyzing and capturing articulated hand motion in image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1910–1922, 2005.

[8] Mohamed-Bécha Kaâniche and Francois Bremond, "Recognizing gestures by learning local motion signatures of hog descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2247–2258, 2012.

[9] Mingyu Chen, Ghassan AlRegib, and Biing-Hwang Juang, "Feature processing and modeling for 6d motion gesture recognition," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 561–571, 2013.

[10] Nasser H Dardas and Emil M Petriu, "Hand gesture detection and recognition using principal component analysis," in *Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA)*. IEEE, 2011, pp. 1–6.

[11] Nasser H Dardas and Nicolas D Georganas, "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 11, pp. 3592–3607, 2011.

[12] Zhou Ren, Junsong Yuan, Jingjing Meng, and Zhengyou Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[13] Janne Heikkila, "Geometric camera calibration using circular control points," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1066–1077, 2000.

[14] Chong Wang, Zhen-Yu Zhu, Shing-Chow Chan, and Heung-Yeung Shum, "Real-time depth image acquisition and restoration for image based rendering and processing systems," *Journal of Signal Processing Systems*, vol. 79, no. 1, pp. 1–18, 2015.

[15] Zhiwei Tang and Yixuan Wu, "One image segmentation method based on otsu and fuzzy theory seeking image segment threshold," in *Proceedings of the International Conference on Electronics, Communications and Control (ICECC)*. IEEE, 2011, pp. 2170–2173.

[16] Daniela Espinoza Molina, Dusan Gleich, and Mihai Datcu, "Gibbs random field models for model-based despeckling of sar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 73–77, 2010.

[17] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[18] Chong Wang, Zhong Liu, and Shing-Chow Chan, "Superpixel-based hand gesture recognition with kinect depth camera," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 29–39, 2015.

[19] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 1565–1569.