Integrating Color and Depth Cues for Static Hand Gesture Recognition

Jiaming Li(✉), Yulan Guo, Yanxin Ma, Min Lu, and Jun Zhang

College of Electronic Science and Engineering,
National University of Defense Technology, Changsha 410073, China
lijiaming15@nudt.edu.cn

Abstract. Recognizing static hand gesture in complex backgrounds is a challenging task. This paper presents a static hand gesture recognition system using both color and depth information. Firstly, the hand region is extracted from complex background based on depth segmentation and skin-color model. The Moore-Neighbor tracing algorithm is then used to obtain hand gesture contour. The k-curvature method is used to locate fingertips and determine the number of fingers, then the angle between fingers are generated as features. The appearance-based features are integrated to the decision tree model for hand gesture recognition. Experiments have been conducted on two gesture recognition datasets. Experimental results show that the proposed method achieves a high recognition accuracy and strong robustness.

Keywords: Hand gesture recognition · Skin-color model
Depth segmentation

1 Introduction

Hand gesture recognition is an active research problem in human-computer interaction, it provides a natural way for communication between humans and machines. Typical applications include sign language recognition, virtual reality and smart homes.

Existing hand gesture recognition methods can be divided into two categories: static and dynamic methods [20]. Although remarkable progress has been achieved in this area [7,11,20,25,27], several challenges are still faced by existing methods. A major challenge is the uncontrolled real-world environment. For static hand gesture recognition, the first problem is reliable hand segmentation. Some systems require users to wear an electronic glove to capture the key features of an hand. However, the device is usually costly and inconvenient [8]. Methods based on skin-color model [21,32,34] and hand shape model [12,13,31] have also been proposed. However, they are not robust in complex backgrounds and rely significantly on the models.

The first author (Jiaming Li) is a master student in National University of Defense Technology.

© Springer Nature Singapore Pte Ltd. 2017
J. Yang et al. (Eds.): CCCV 2017, Part I, CCIS 771, pp. 295–306, 2017.
https://doi.org/10.1007/978-981-10-7299-4_24

Recent development of depth cameras provides a new direction for static hand gesture recognition. In this work, we use a Kinect camera to capture color and depth images. We propose a static hand gesture recognition system based on depth segmentation and skin-color model (as shown in Fig. 1). We first use depth thresholding and skin-color model to segment hand from an RGB-D image, and then extract appearance-based features. Finally, the decision tree is used for hand gesture recognition. Experimental results show that the proposed system is highly accurate and efficient.

The rest of the paper is organized as follows. Related works are first reviewed in Sect. 2. The proposed method for hand segmentation, hand representation and hand gesture recognition is introduced in Sect. 3. Experimental results are discussed in Sect. 4. Finally, we conclude the paper in Sect. 5.

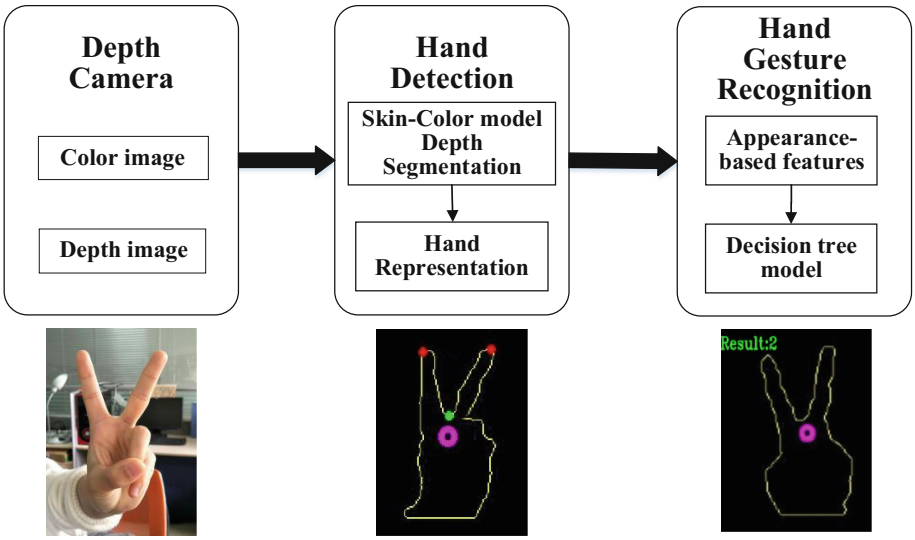

Fig. 1. The flowchart of the proposed hand gesture recognition method.

2 Related Works

A hand gesture recognition system includes data acquisition, hand localization, feature extraction, and gesture recognition [22]. Many vision-based hand gesture recognition algorithms have been proposed [7,10,20,27,29]. If depth information is available, the accuracy and robustness of hand gesture recognition can further be improved. Early hand segmentation methods mainly rely on features, such as chromatic distribution, which is sensitive to variations of skin colors. An appearance-based hand detection system is also proposed in [14], which is unreliable due to the complicated and unpredictable features. Therefore, depth

information can be used to improve hand segmentation performance. For example, the point in the depth image closest to the camera is first detected, and its neighbors within a range of depth value can be used to segment the hand. The hand segmentation result is usually inaccurate since only a small portion of an image is occupied by the hand.

After hand segmentation, various hand features can be extracted from depth image. Li *et al.* [16] proposed a bag-of-3D-points feature for activity recognition from depth image sequences. Specifically, 3D points are sampled from the silhouettes of depth images. Ren *et al.* [24] represented the hand shape by time-series curve, with each finger corresponding to a segment of the curve. The time-series curve records the relative distance between each contour vertex to a center point and reserves the topological information as well. Besides, the Finger-Earth Mover's Distance (FEMD) was used in this method. However, this distance metric is sensitive to distortion. Wang *et al.* [30] proposed Superpixel Earth Mover's Distance (SP-EMD) which shows a promising way for hand gesture recognition. In this paper, the hand is located and segmented according to the depth and skeleton information from Kinect. Chen *et al.* [6] proposed image-to-class dynamic time warping (DTW) distance to distinguish the fingerlet features of 3D hand contours. Liu *et al.* [18] explored the invariance from the hand contour and utilized Fourier descriptor, edge histogram, and boundary moment invariants for the feature extraction. Sorce *et al.* [26] proposed a neural network with backpropagation detecting the hand pose to recognize whether it is closed or not. An exponential weighted moving average noise reduction mechanism was used to suppress the noise effects of the neural network. Kuznetsova *et al.* [15] proposed an Ensemble of Shape Function (ESF) to represent the hand region. The ESF descriptor consists of concatenated histograms generated from randomly selected points in the point cloud. A Multi-Layered Random Forest (MLRF) was then trained to recognize the hand signs. Zhang *et al.* [33] defined a 3D facet as a 3D local support surface for each 3D point cloud and used the Histogram of 3D Facets (H3DF) to represent the 3D hand shape. H3DF can effectively represent the 3D shapes and structures of various hand gestures. The Support Vector Machines (SVM) with linear kernel was used for classification. SVM is the most popular classifier for the 3D static hand gesture recognition. Pugeault and Bowden [23] adopted linear SVM to predict the class label with the ASL data sets. For nonlinear kernels, Gallo *et al.* [9] trained the Radial Basis Function (RBF) classifiers to recognize four different hand postures in a 3D medical touchless interface. Bagdanov *et al.* [1] represented the hand by the concatenation of five Speeded Up Robust Feature (SURF) descriptors for a total of 640 dimensions. A SVM with a nonlinear RBF was trained for hand gesture recognition. This hand gesture recognition method is invariant to rotation, arm pose, background and distance from the sensor. The random decision forest has been experimentally compared with the SVM classification with various 3D hand features [19]. In general, the performance of RFs is dependent on the depth of the tree. The tradeoff of accuracy and speed always needs to be properly considered in its implementation.

3 Static Gesture Recognition

First, both depth and skin color are used for hand segmentation. Then, the appearance-based features are used to represent the hand. Finally, the decision tree model is trained for static hand gesture recognition.

3.1 Image Preprocessing

Preprocessing has to be performed before hand segmentation. Median filter is used to reduce salt and pepper noise. Compared to mean filter, median filter can an preserve more details around boundary structures. Besides, median filter is robust to noise.

3.2 Depth and Skin-Color Based Hand Detection

Both skin color and depth are used for hand detection. The hand is first segmented with depth threshold, histogram of the depth image is then used to distinguish the foreground from background. The depth data within a range are extracted and further combined with RGB images (using skin color segmentation) to remove unnecessary interference. The orthogonal color space $YCbCr$ is used in our paper, pixel values are considered as skin color if Cb is within the range of [77, 127] and Cr is within the range of [133, 173]. The depth range is set to [1.5 m, 1.7 m], pixels within the depth range are saved as $\boldsymbol{p}_i(x_i, y_i)$. In this paper, the K-means clustering algorithm is used to classify all pixels into two clusters. Specifically, the initial K clustering centers $\{u_j\}$ are randomly selected, then the following calculation is repeated until the sample converges. For each pixel $x^{(i)}$, the class label $c^{(i)}$ is defined as

$$c^{(i)} = \text{argmin}||x^{(i)} - u_j||^2 \tag{1}$$

where $x^{(i)}$ represents the point in sample space, the cluster center u_j is calculated as

$$u_j = \frac{\sum_{i=1}^{m} \mathbf{1}\left\{c^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\left\{c^{(i)} = j\right\}} \tag{2}$$

where is $\mathbf{1}\{c^{(i)} = j\}$=1 if $c^{(i)} = j$. Consequently, the pixels are divided into two classes: hand part and non-hand part. Figure 2 illustrates the hand segmented by our method.

3.3 Appearance Based Hand Representation

After hand segmentation, hand gesture contour is obtained using the Moore-Neighbor tracing algorithm [17]. Specifically, a pixel is compared to its 8 adjacent points, if all the adjacent points are internal points, the pixel is deleted. Otherwise, the pixel is considered as a hand contour point. This process is traversed over the entire image. Then, hand can be represented by the center of

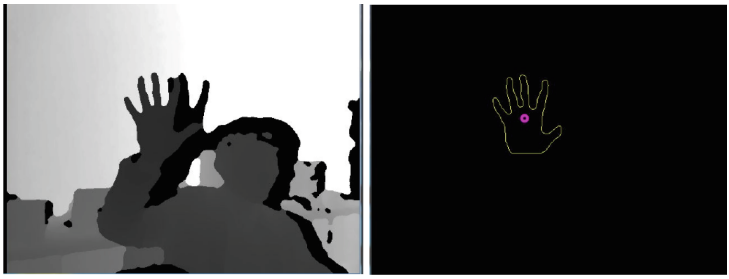

Fig. 2. The depth map and the segmentation result.

the palm, the number of fingers and the angles between fingers. The center of the palm can be determined by calculating the maximum inscribed circle of the hand, which can effectively eliminate the impact caused by hand palm motion. The number of fingertips is obtained by the k-curvature algorithm [28]. A vector is first calculated from contour points, a candidate fingertip point is then determined as the point with the largest curvature, and then the following steps are used:

(1) A set is initialized by the points on the hand.
(2) Given a candidate fingertip $\boldsymbol{p}_i$ and its neighboring points $\boldsymbol{p}_{i-k}$ and $\boldsymbol{p}_{i+k}$, two vectors are computed from these points to $\boldsymbol{p}_i$, i.e., $\overrightarrow{\boldsymbol{p}_i\boldsymbol{p}}_{i-k}$ and $\overrightarrow{\boldsymbol{p}_i\boldsymbol{p}}_{i+k}$.
(3) θ is calculated using Eq. 3. If $\theta < 45^{\circ}$, $\boldsymbol{p}_i$ can be determined as a fingertip. Otherwise, steps 2 and 3 are repeated with next unchecked point.

$$\theta = \frac{(\boldsymbol{p}_i - \boldsymbol{p}_{i-k})}{||\boldsymbol{p}_i - \boldsymbol{p}_{i-k}||} * \frac{(\boldsymbol{p}_i - \boldsymbol{p}_{i+k})}{||\boldsymbol{p}_i - \boldsymbol{p}_{i+k}||} \tag{3}$$

Consequently, the hand fingertips and the groove between two fingers can be obtained at the same time. The groove points are deleted using the cross product of the two vectors. This method can accurately find the position of fingertips for different gestures of different hands. The angle θ is relatively stable even if the hand position changes since the relative position of these three points is unchanged. Once fingertips are determined, angles between fingers are obtained by connecting the center of the palm and fingertips. Besides, the maximum angle between fingers is determined and the distances between the center of the palm and the fingertips are calculated.

3.4 Decision Tree Based Static Hand Gesture Recognition

Using the decision tree model, a complex multi-class classification problem can be transferred to a number of simple classification problems. A suitable tree structure should be carefully designed. That is, appropriate nodes and branches, the number of features and the appropriate decision rules should be determined.

The number of extracted fingers N and the maximum angle between fingers A are used as the decision tree node. The decision tree model used in our dataset is shown in Fig. 3.

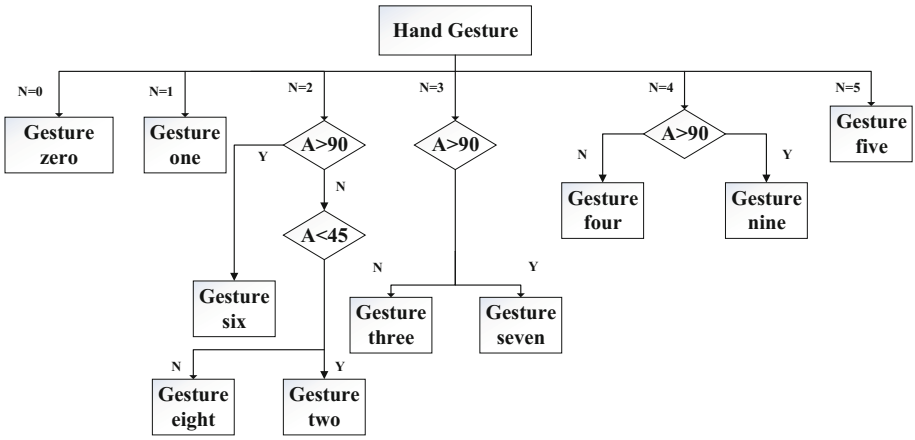

Fig. 3. The decision tree model of the proposed method.

Note that, gestures zero, one and five can be directly recognized at the root node by the number of fingers. When the number of fingers is same, a second-level node should be established to recognize gestures three, seven, four and nine using the maximum angle between fingers. Further, a third-level node should be established to distinguish gestures two, six and eight by dividing the maximum angle between fingers. An illustration of hand gesture recognition results is shown in Fig. 4.

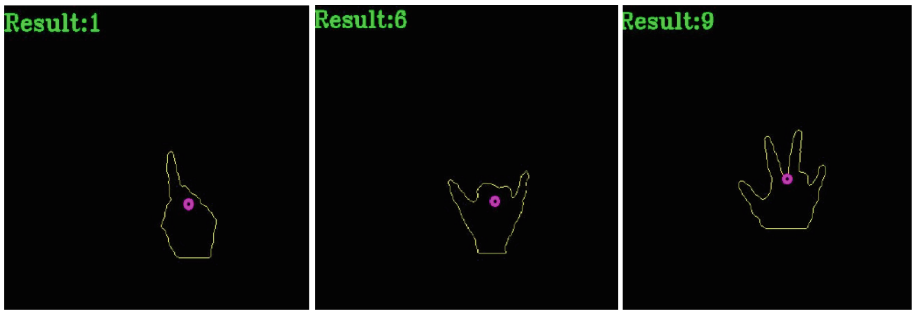

Fig. 4. An illustration of hand gesture recognition results.

4 Experiments

We evaluated the proposed gesture recognition approach on two datasets. The 10-Gesture dataset [24] contains both RGB and depth images of 10 gestures collected from 10 subjects. Each subject performs 10 different poses for each gesture. In total, the dataset has 1000 samples. To further evaluate our approach, we collected a digit dataset using a Kinect device. Our dataset contains 10 gestures of 5 subjects. For each subject, 10 different poses are performed for each gesture. Therefore, the dataset has 500 samples. Several gesture samples in our dataset are shown in Fig. 5. It should be noted that gesture 9 in the 10-Gesture dataset and our dataset is different, while the other gestures are the same. For fair comparison, we followed the experimental setup introduced in [24]. All experiments were conducted on a machine with an Intel Core i5-7200 CPU and 4 GB RAM. The proposed method was compared to existing methods in terms of mean accuracy and robustness.

Fig. 5. Gesture samples (0–9) captured by Kinect.

4.1 The 10-Gesture Dataset

The decision tree used for the 10-Gesture dataset is slightly different from that for our our dataset (Fig. 3). Specifically, the distances between the center of the palm and the fingertips rather than the angle are used to distinguish gestures 1 and 9 in the 10-Gesture dataset. As shown in Fig. 6, there are several false recognition results between gestures 1 and 9. That is because the number of fingers for gestures 1 and 9 is the same. For other gestures, a few errors are caused by individual difference for these gestures. For example, the angles between fingers for the same gesture may be different for different persons. The overall mean accuracy achieved by our method is 96.1%.

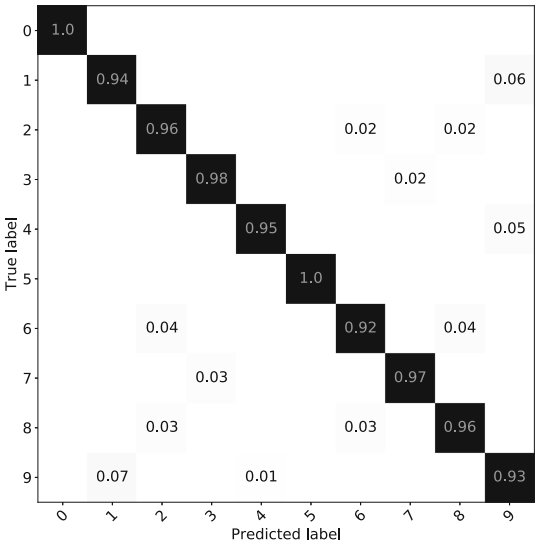

Fig. 6. The confusion matrix achieved by the proposed method on the 10-Gesture dataset.

We further compare our method to three existing methods on the 10-Gesture dataset, include Shape Context [5], Skeleton Matching [3] and FEMD [24]. Their mean accuracy results are shown in Table 1. It can be seen that the proposed method achieves the highest mean accuracy. The most confusing cases for FEMD are gestures 4 and 5, and gestures 1 and 9. That is because the fingers in those gestures are not accurately segmented due to image distortion. Moreover, different hand size can change the weight of each finger for FEMD, resulting in mismatching. For shape context based method, the most confusing cases are gestures 1 and 9, and gestures 8 and 9. That is because these gestures have similar contours. In some cases, two fingers may stick together due to distortion. That is why gesture 4 is falsely recognized as gesture 3. For skeleton matching based method, two different skeleton pruning methods can be used, including

Table 1. The mean accuracy results achieved by Shape Context, Skeleton Matching, FEMD, and our proposed method on the 10-gesture dataset.

Methods	Mean accuracy
Shape Context [5] (with bending cost)	90.9%
Shape Context [5] (without bending cost)	93.5%
Skeleton Matching [3] (DCE [4])	92.8%
Skeleton Matching [3] (DSE [2])	92.9%
FEMD [24]	94.1%
The proposed method	**96.1**%

Discrete Curve Evolution (DCE) [4] and Discrete Skeleton Evolution (DSE) [2]. DSE is more robust to small protrusions. Therefore, the recognition accuracy achieved by DSE is higher than DCE. The most confusing pairs are gestures 6 and 7, and gestures 2 and 7. That is because the pruned skeletons have similar global structures for these gestures. In summary, the proposed method achieves a higher recognition accuracy than existing methods.

4.2 The Kinect Dataset

The recognition results achieved on our Kinect dataset are shown in Fig. 7. For gestures 0, 1 and 5, the recognition rate is 100%. For other gestures, few false recognition can be found. The mean accuracy achieved by our method is 98.0%. The results show that the proposed method can achieve a high recognition accuracy.

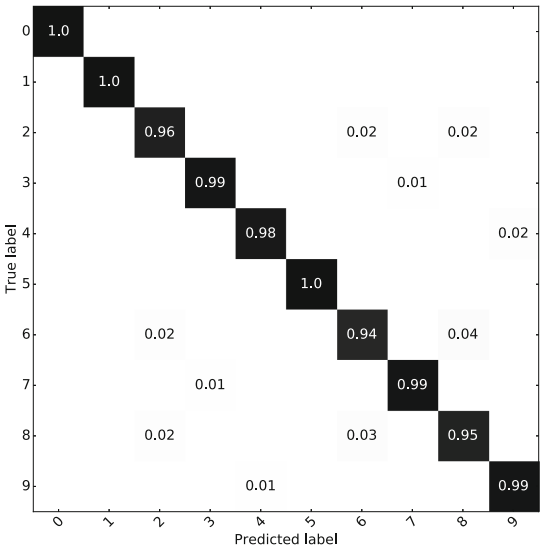

Fig. 7. The confusion matrix achieved by the proposed method on our Kinect dataset.

Hand gesture detection based on skin color can easily be affected by objects with the same color as skin. It is vulnerable to environment changes, such as illumination variation. Therefore, the combination of skin color and depth information can significantly improve the robustness of hand segmentation. To test the robustness of static gesture recognition in illumination changes and complex backgrounds, experiments were conducted under high illumination, low illumination and complex backgrounds, with each gesture being tested for 200 times. Experimental results are shown in Table 2. It can be seen that the recognition results under illumination changes are stable due to accurate hand segmentation. The accuracy drops slightly in complex background. That is mainly because

hand segmentation is sensitive to background with skin color. In summary, our hand gesture recognition method is robust to illumination changes and complex background.

Table 2. The hand gesture recognition accuracy achieved in different circumstances.

Hand gesture	High illumination		Low illuminiation		Complex background	
	True positives	Mean accuracy	True positives	Mean accuracy	True positives	Mean accuracy
0	200	100%	200	100%	198	99%
1	200	100%	200	100%	196	98%
2	199	99.5%	198	99%	194	97%
3	197	98.5%	197	98.5%	195	97.5%
4	198	99%	195	97.5%	196	98%
5	196	98%	196	98%	194	97%
6	195	97.5%	198	99%	194	97%
7	197	98.5%	197	98.5%	195	97.5%
8	199	99.5%	198	99%	195	97.5%
9	198	99%	198	99%	196	98%
Mean	197.9	98.95%	197.7	98.85%	195.3	97.65%

We also compare our method to three existing methods on this dataset, including Shape Context [5], Skeleton Matching [3] and FEMD [24]. Their mean accuracy results are shown in Table 3. It can be observed that, the proposed method achieves the best mean accuracy. The mean accuracy of each method achieved on this dataset is higher than that on the 10-Gesture dataset. That is because the 10-Gesture dataset is more challenging than our dateset, as it is collected in uncontrolled environment. The most confusing cases for FEMD are gestures 4 and 5, and gestures 1 and 7. That is because the fingers in those gestures are not accurately segmented. For shape context based method, the most confusing cases are gestures 1 and 7, and gestures 8 and 9. The main reason

Table 3. The mean accuracy results achieved by Shape Context, Skeleton Matching, FEMD and our proposed method on our Kinect dataset.

Methods	Mean accuracy
Shape Context [5] (with bending cost)	94.7%
Shape Context [5] (without bending cost)	97.1%
Skeleton Matching [3] (DCE [4])	96.5%
Skeleton Matching [3] (DSE [2])	96.6%
FEMD [24]	96.8%
The proposed method	**98.0**%

is, these gestures have similar contours. For skeleton matching based method, the most confusing pairs are gestures 6 and 7, and gestures 2 and 7. That is because the pruned skeletons have similar global structures for these gestures. In summary, the proposed method achieves higher recognition accuracy than existing methods on our Kinect dataset.

5 Conclusions

In this paper, a hand gesture recognition method has been proposed using both depth and color information. The appearance features are extracted to represent a hand gesture, and the decision tree model is then used for hand gesture recognition. The effectiveness of the proposed method has been demonstrated on two datasets collected by Kinect. High mean accuracy and strong robustness has been achieved by our hand gesture recognition method. Compared to three state-of-the-art methods, the proposed method achieves the best hand gesture recognition performance.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (Nos. 61602499 and 61471371), the National Postdoctoral Program for Innovative Talents (No. BX201600172), and China Postdoctoral Science Foundation.

References

1. Bagdanov, A.D., Del Bimbo, A., Seidenari, L., Usai, L.: Real-time hand status recognition from RGB-D imagery. In: ICPR, pp. 2456–2459. IEEE (2012)
2. Bai, X., Latecki, L.J.: Discrete skeleton evolution. In: Yuille, A.L., Zhu, S.-C., Cremers, D., Wang, Y. (eds.) EMMCVPR 2007. LNCS, vol. 4679, pp. 362–374. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74198-5_28
3. Bai, X., Latecki, L.J.: Path similarity skeleton graph matching. IEEE TPAMI **30**(7), 1282–1292 (2008)
4. Bai, X., Latecki, L.J., Liu, W.Y.: Skeleton pruning by contour partitioning with discrete curve evolution. IEEE TPAMI **29**(3), 449–462 (2007)
5. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE TPAMI **24**(4), 509–522 (2002)
6. Cheng, H., Dai, Z., Liu, Z.: Image-to-class dynamic time warping for 3D hand gesture recognition. In: IEEE ICME, pp. 1–6 (2013)
7. Cheng, H., Yang, L., Liu, Z.: Survey on 3D hand gesture recognition. IEEE TCSVT **26**(9), 1659–1673 (2016)
8. Dejmal, I., Zacksenhouse, M.: Coordinative structure of manipulative hand-movements facilitates their recognition. IEEE TBE **53**(12), 2455–2463 (2006)
9. Gallo, L.: Hand shape classification using depth data for unconstrained 3D interaction. J. Ambient Intell. Smart Environ. **6**(1), 93–105 (2014)
10. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans. Cybern. **43**(5), 1318–1334 (2013)
11. Hasan, H.S., Kareem, S.A.: Human computer interaction for vision based hand gesture recognition: a survey. AIR **43**(1), 1–54 (2015)

12. Hu, R.X., Jia, W., Zhang, D., Gui, J., Song, L.T.: Hand shape recognition based on coherent distance shape contexts. Pattern Recogn. **45**(9), 3348–3359 (2012)
13. Keskin, C., Kıraç, F., Kara, Y.E., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 852–863. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_61
14. Kölsch, M., Turk, M.: Robust hand detection. In: IEEE FGR, pp. 614–619 (2004)
15. Kuznetsova, A., Leal-Taixé, L., Rosenhahn, B.: Real-time sign language recognition using a consumer depth camera. In: IEEE ICCV Workshops, pp. 83–90 (2013)
16. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: IEEE CVPR Workshops, pp. 9–14. IEEE (2010)
17. Li, Y.: Hand gesture recognition using kinect. In: IEEE ICSESS, pp. 196–199 (2012)
18. Liu, Y., Yang, Y., Wang, L., Xu, J., Qi, H., Zhao, X., Zhou, P., Zhang, L., Wan, B., Ming, D.: Image processing and recognition of multiple static hand gestures for human-computer interaction. In: ICIG, pp. 465–470 (2013)
19. Minnenm, D., Zafrulla, Z.: Towards robust cross-user hand tracking and shape recognition. In: IEEE ICCV Workshops, pp. 1235–1241 (2011)
20. Mitra, S., Acharya, T.: Gesture recognition: a survey. IEEE TSMC **37**(3), 311–324 (2007)
21. Phung, S.L., Bouzerdoum, A., Chai, D.: Skin segmentation using color pixel classification: analysis and comparison. IEEE Trans. Pattern Anal. Mach. Intell. **27**(1), 148 (2005)
22. Plouffe, G., Cretu, A.M.: Static and dynamic hand gesture recognition in depth data using dynamic time warping. IEEE TIM **65**(2), 305–316 (2015)
23. Pugeault, N., Bowden, R.: Spelling it out: real-time asl fingerspelling recognition. In: IEEE ICCV Workshops, pp. 1114–1119 (2012)
24. Ren, Z., Yuan, J., Zhang, Z.: Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In: ACMMM, pp. 1093–1096 (2011)
25. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE CVPR, pp. 1297–1304 (2011)
26. Sorce, S., Gentile, V., Gentile, A.: Real-time hand pose recognition based on a neural network using microsoft kinect. In: BWCCA, pp. 344–350 (2013)
27. Suarez, J., Murphy, R.R.: Hand gesture recognition with depth images: a review. In: RO-MAN, pp. 411–417 (2012)
28. Trigo, T.R., Pellegrino, S.R.M.: An analysis of features for hand-gesture classification. In: IWSSIP, pp. 412–415 (2010)
29. Wachs, J.P., Kölsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. Commun. ACM **54**(2), 60–71 (2011)
30. Wang, C., Liu, Z., Chan, S.C.: Superpixel-based hand gesture recognition with kinect depth camera. IEEE TMM **17**(1), 29–39 (2015)
31. Wu, Y., Lin, J., Huang, T.S.: Analyzing and capturing articulated hand motion in image sequences. IEEE TPAMI **27**(12), 1910–1922 (2005)
32. Yang, M.H., Ahuja, N., Tabb, M.: Extraction of 2D motion trajectories and its application to hand gesture recognition. IEEE TPAMI **24**(8), 1061–1074 (2002)
33. Zhang, C., Yang, X., Tian, Y.: Histogram of 3D facets: a characteristic descriptor for hand gesture recognition. In: FG, pp. 1–8 (2013)
34. Zhu, X., Wong, K.Y.K.: Single-frame hand gesture recognition using color and depth kernel descriptors. In: ICPR, pp. 2989–2992 (2013)