ICIC Express Letters
Part B: Applications
Volume **7**, Number **8**, August **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **1755–1761**

HAND GESTURE RECOGNITION USING CONCENTRIC DEPTH DISTRIBUTION HISTOGRAM FEATURE REPRESENTATION

Min Jiang[1], Pingping Li[1], Jun Kong[1,2] and Lin Sun[1]

[1]Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education)
Jiangnan University
No. 1800, Lihu Avenue, Wuxi 214122, P. R. China
minjiang@jiangnan.edu.cn

[2]College of Electrical Engineering
Xinjiang University
No. 14, Shengli Road, Urumqi 830046, P. R. China

Received December 2015; accepted March 2016

Abstract. *The recently developed depth sensors, e.g., Kinect sensor, have made it possible to capture depth images in real time, which facilitates a variety of visual recognition tasks including hand gesture recognition. However, most existing methods use the depth images only to facilitate the segmentation, and ignore the depth information on the surface of hand gesture while recognizing. This paper focuses on hand gesture recognition task using Kinect. To explicitly represent the depth information on the surface of hand, we propose a novel feature, concentric depth distribution histogram (CDDH), which is characterized by a 2D histogram representing the distribution for depth information of several concentric rings. The fusion of the geometrical features and the CDDH feature could fully utilize the information in both 2D and 3D. Then the random forest classifier is applied for hand gesture recognition task using the normalized hand depth map. The extensive experimental results show that our method has higher precision for hand gesture recognition, and our method is invariant to translation, rotation, and scale of the hand.*

Keywords: Concentric depth distribution histogram feature, Geometrical features, Hand gesture recognition, Random forest classifier

1. **Introduction.** As one of the underlying issues in human-computer interaction (HCI) system [1], hand gesture recognition plays a significant role due to its numerous potential applications in critical tasks such as virtual reality systems, sign language recognition systems, and interactive gaming platforms. Although much progress has been made in the past decades, traditional vision-based hand gesture recognition methods have certain limitations in real-life application due to several challenges such as illumination changes and complex backgrounds. Therefore, the traditional vision-based methods are usually unable to detect and segment the hands robustly and accurately, which severely affects the results of gesture recognition. Recently, vision-based hand gesture recognition methods have been discussed and investigated thoroughly [2,3]. Plenty of interesting research work has been developed which can be mainly classified into two categories: 3D hand model based methods and 2D hand model based methods. Constructing the 3D hand surfaces [4] can provide more valuable information to eliminate the ambiguities of hand gesture, and mostly reflect the original features of gesture. However, some drawbacks of 3D hand methods including the high computational cost, complex parameters and the difficulty of 3D reconstruction, have extremely hinder the widespread adoptions in real-life applications. On the contrary, the 2D hand model based methods have less computational cost and can assure the real-time of hand recognition. However, hand gesture recognition on 2D RGB images is sensitive to background clutters and illumination variations [5], so it usually requires a clean background, which limits its application in the real world.

To overcome the drawbacks of 2D RGB images, Ren and Yuan [6] apply the depth sensor Kinect to obtain the depth images of hand gestures for the robustness to background clutters and illumination variations. In [6], a template matching based method is applied to recognize hand gestures through the distance metric of finger earth mover distance (FEMD). However, this method only considered the contour of hand but ignored the surface information of hand region which also provides important structure information of complex hand gesture.

Motivated by the above-mentioned work, we propose a novel feature concentric depth distribution histogram (CDDH) to recognize the hand gestures using Kinect, which can explicitly record the surface depth information of hand region. Extensive experiments show that the proposed method achieves more favorable performance than several competitive features based methods in terms of accuracy. The rest of the paper is organized as follows. Section 2 describes the hand gesture segmentation and preprocessing. Section 3 presents our CDDH feature and the fusion with geometrical features. The results and analysis are presented in Section 4, and conclusion and future work in Section 5.

2. Hand Gesture Preprocessing.

Kinect is a motion-sensing device developed by Microsoft [7]. As Kinect can easily produce 3D depth data, we can efficiently apply Kinect to recognize the hand gesture. Compared with the full field of camera view, hands are usually quite small. Thus it is still a vital challenge to segment the hand from the clutter-background. The technology of skeleton tracking can efficiently estimate the major joint coordinates from the depth image. Therefore, we can roughly obtain the depth value of the hand area by the location of the hand joint. Although the surface of hand should be integrated and smooth theoretically, limited by the structured light technology, it is difficult to get high-quality depth image without holes and coarse edges via Kinect. To correct the possible wrong depth value in the hand area, we evaluate the depth image by Equation (1) with the mean μ_d and variance σ_d^2 in the connected neighbor area of the hand joint. There may be several connected regions exiting in the image. Here we just pay attention to the region where hand joint located and dismiss the redundant region along the divider line which crosses the wrist joint and is perpendicular to the vector from hand joint to the wrist joint in 2D projection. Figure 1 shows the details of segmentation.

$$f(d) = \begin{cases} d & \text{when } |d - \mu_d| < \varepsilon,\ \varepsilon \leq n\sigma_d^2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The segmented hand gesture is shown in Figure 1(d). It can be obviously seen that there are some noises and coarse edges. We apply morphology to process the coarse image. The result is shown in Figure 1(e). To eliminate the harmful effect of varied distance and individual varieties of pose and hand size, all the segmented hand gestures are normalized to a fixed size 120 × 120.

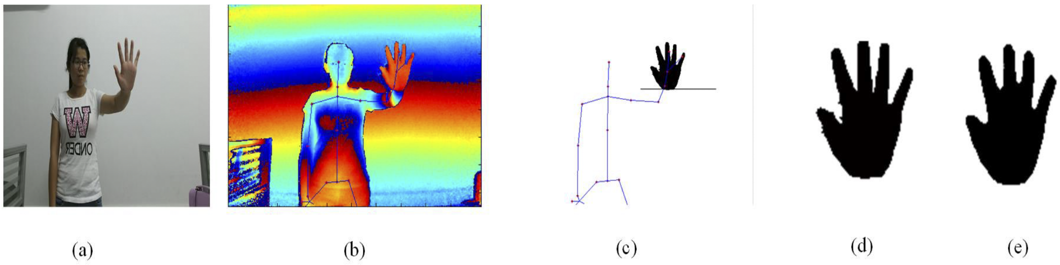


Figure 1. (a) The original color image, (b) the corresponding skeleton map over depth map, (c) the image after segmenting with the divider line, (d) the segmented hand image, (e) the smoothed hand gesture

3. Feature Extraction and Recognition.

In this paper, we propose a new feature named concentric depth distribution histogram (CDDH) feature. Combined with the features of geometric description of the hand gesture, our feature set assures the robustness and precision for hand gesture recognition. Details are explained as follows.

3.1. Concentric depth distribution histogram feature.

In this section, we describe the details of the proposed CDDH feature. CDDH feature is a depth based descriptor. It reveals the distribution of depth data for different hand gestures. It is invariant under rotations and translations. The steps of extracting the feature are explained as follows.

Step 1: Let $o(0,0,0)$ be the origin of the coordinate. As the distance between hand and the camera varies with different samples, we normalize all the samples by transforming the coordinate along the depth axis and place the new coordinate system's origin at $o(0,0,d)$, where d denotes the minimal depth value on the segmented hand sample.

Step 2: Given the segmented hand region, we calculate its centroid or center of gravity (COG), whose coordinate is $O_{centroid}(\bar{x},\bar{y})$, where $\bar{x} = \frac{\sum_{p\,(x_i,y_i)\in\Omega} x_i}{\Omega}$ and $\bar{y} = \frac{\sum_{p\,(x_i,y_i)\in\Omega} y_i}{\Omega}$. Ω denotes the hand region and $p(x_i,y_i)$ means the coordinates of i^{th} pixel in the hand region Ω.

Step 3: As shown in Figure 2, centered at $O_{centroid}(\bar{x},\bar{y})$, the minimum circumscribed circle is calculated to find the farthest distance $l_{\max} = \max\limits_{(x',y')\in\Phi} \sqrt{(x'-\bar{x})^2 + (y'-\bar{y})^2}$ from the centroid to the hand contour, where (x',y') is the point on the hand contour Φ.

Step 4: Depth maps are segmented with N concentric circles based on the centroid $O_{centroid}(\bar{x},\bar{y})$. As shown in Figure 2, the radiuses of these concentric circles are respectively denoted as $\frac{1}{N}l_{\max}$, $\frac{2}{N}l_{\max}$, $\ldots$, $\frac{N-1}{N}l_{\max}$. The effects of parameter N will be discussed in Section 4.2.

Step 5: The depth distribution histogram H_i for the ith concentric circle is generated which gives a rough sense of the density of the annular depth data. In our experiment, the number of bins $k = 8$ and the range is $\left[0, \frac{9}{8}D_{\max}\right]$, where $D_{\max}$ denotes the maximal depth value among all the training samples.

Step 6: As a result, a depth map of hand gesture is represented by the depth distribution histograms H_i $(i = 1,2,3,4,\ldots,N)$ for N concentric circles named concentric depth distribution histogram (CDDH): $F_{CDDH} = (H_1, H_2, H_3, \ldots, H_N)^T$.

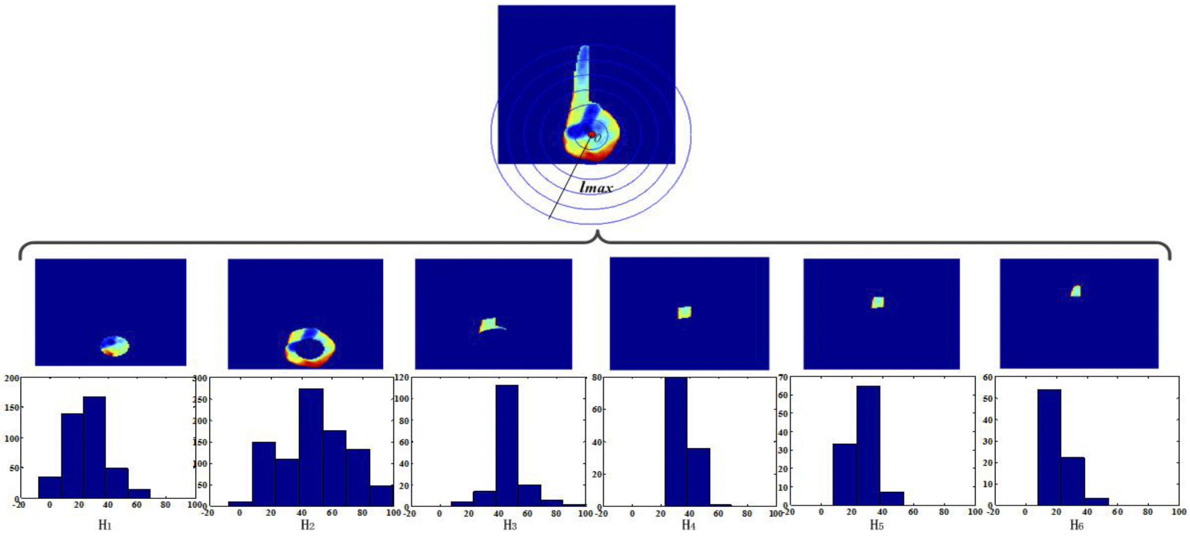


FIGURE 2. Concentric depth distribution histogram (CDDH) feature

3.2. **Geometrical features.** Two main geometrical features are adopted in our work: compactness and rectangularity. These features meet with the principles of translation and scaling invariance.
(1) Compactness (C_A): Compactness is a shape-based descriptor. It is a numerical quantity representing the degree to which a shape is compact. C_A is defined as: $C_A = \frac{Perimeter^2}{Area}$, where $Perimeter$ is the length of the contour for the hand and $Area$ is the total number of hand pixels.
(2) Rectangularity (R_A): Rectangularity defines the measure of the shape to show how much its shape is closer to the rectangle. It is calculated as: $R_A = \frac{Area}{A_{MER}}$, where $Area$ is the total pixels of the hand, and A_{MER} is the area of minimum circumscribed rectangle. The feature set of the geometrical feature is defined as: $F_{Geo} = (C_A, R_A)^T$.

Both the CDDH feature F_{CDDH} and the geometrical feature F_{Geo} are important for representing the hand gesture information. F_{Geo} describes the features in 2D and F_{CDDH} encodes the information in 3D. The combination of the two feature sets provides better discrimination than either alone. The feature set vector can be denoted as:

$$F_{total} = [F_{Geo}, F_{CDDH}] = (C_A, R_A, H_1, H_2, H_3, \ldots, H_N)^T \tag{2}$$

A linear normalization is applied to normalize the feature values to the range of [0 1].

3.3. **Fusion and classifier using the random forests.** Hand gesture recognition is a typical multi-class classification problem. In order to perform feature fusion and selection, we propose to use the random forests (RFs) method [9].

RFs are usually considered as a classifier using tree predictors by which each tree splits depending on the feature of randomly selected data. Let the feature vector be $F_{total} \in \mathbb{R}^N$, where the number of the features for each sample is N. A number $n < N$ is specified at each node of the tree, where n features are randomly selected to determine the split of that node. The randomly selected n features contain CDDH feature F_{CDDH} partially and the geometrical feature F_{Geo} partially. In this way, the feature fusion is executed randomly and naturally in the tree building process. At each node, the feature that provides the most information gain is selected to split the node. Then the information gain I can be defined as: $I_j = H(S_j) - \sum_{k \in (L,R)} \frac{|S_j^k|}{S} H\left(S_j^k\right)$, where $|.|$ is the size of the set. S_j is the set of training points at node j, $H(S_j)$ is the Shannon entropy at node j before the split, and S_j^L and S_j^R are the sets of points at the right child and left child respectively of the parent node j after the split. And H can be defined as: $H(S) = -\sum_{c \in C} p_c \log(p_c)$, where p_c is the probability of a sample being class c. In the leaf nodes, the probabilistic distribution for each class is computed. In testing, the feature set F for each test sample goes down to one of the leaf nodes in each tree t, denoted as $l_{t,F}$. Random forests classifier chooses the best class label which gets the most vote over all the trees. The class label $\hat{c}$ is determined by: $\hat{c} = \arg\max_c \frac{1}{T} \sum_{t=1}^{T} p_{l_{t,F}}^c$, where T is the number of trees, and $p_{l_{t,F}}^c$ is the posterior probabilities for class c at leaf node $l_{t,F}$, which is determined in the training procedure.

4. **Results and Discussion.**

4.1. **Hand gesture dataset.** All experiments are implemented in MATLAB and run on the Intel-Core I7 3.8 GHz CPU with 8 GB of RAM. As the depth information is a new emerging researching area and there is no public hand gesture dataset with depth information, it is necessary to build a new dataset for the experiments of our work. We build the dataset by capturing the depth image with Microsoft Kinect. We collected from 10 subjects. Each subject performed 5 kinds of hand gestures indicating numbers from 1 to 5. There are totally 250 samples in our dataset. For each kind of gesture, the subject

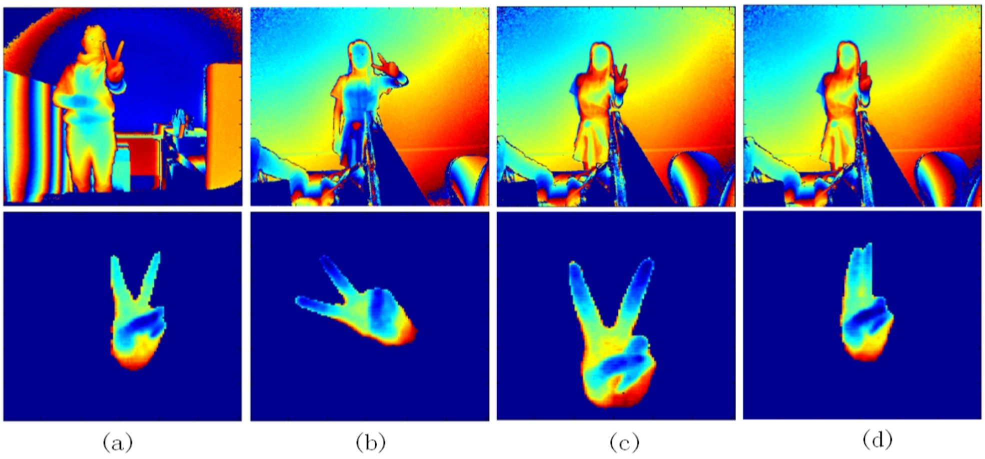
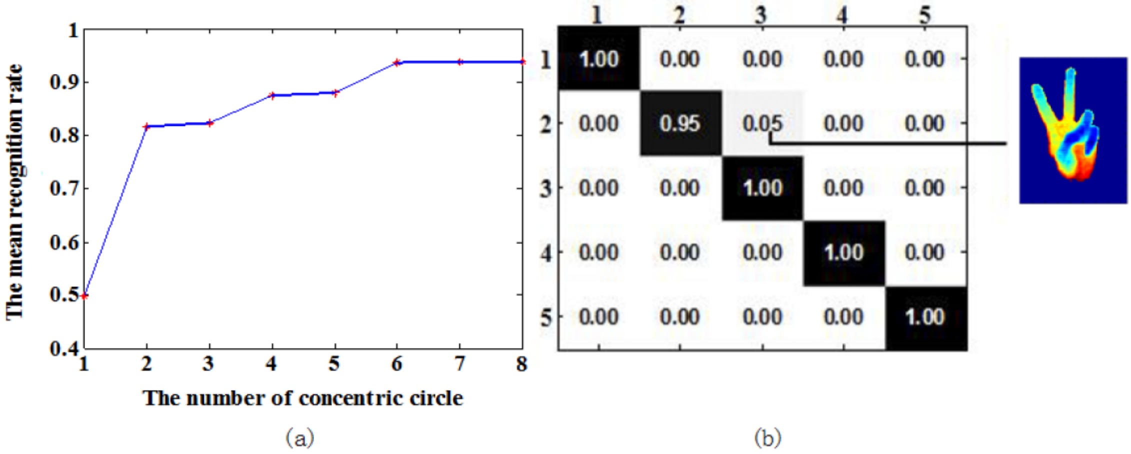


Figure 3. The different poses of Gesture 2: (a) the common gesture; (b) the rotation of gesture; (c) the different scale compared with the common gesture; (d) the different appearance of the same gesture



	1	2	3	4	5
1	1.00	0.00	0.00	0.00	0.00
2	0.00	0.95	0.05	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	1.00

Figure 4. (a) Parameter sensitivity on N; (b) confusion matrix of the proposed method

poses vary in hand orientation, scale, and articulation as shown in Figure 3. The first row illustrates the original depth images, and the second row illustrates the depth map for the segmented hand gesture. It should be noted that our dataset is collected in the natural environment. The resolution is 512 × 424. Therefore, our dataset is very challenging for the recognition of hand gestures in real life.

4.2. **Parameter sensitivity.** The number of the concentric circles N is an important parameter in our method. In order to clarify the effect of the parameter N, we only use the CDDH feature in this experiment to recognize the gesture. We performed the experiment 10 rounds to get the mean recognition precision. We randomly selected one half of the samples for training and the other half for testing. The obtained results are shown in Figure 4(a). The horizontal axis represents the number of concentric circle N, and the vertical axis gives the rate of gesture recognition. It is clear that if N is too small ($N < 6$), the mean recognition rate decreases sharply. The main reason is that the difference of the depth distribution is quite modest for different gestures. If N is too large ($N \geq 6$), the mean recognition rate converges to a limit value and the computation cost increases as well. Considering all the above mentioned factors, we make the compromise to set the parameter $N = 6$.

4.3. **Experimental results.** To prove the effectiveness of the proposed method, we compare our method with the geometrical features alone based and CDDH feature alone based methods respectively. Moreover, the method in [10] is also introduced for a more convincing proof. To fairly evaluate the ability of our method, we randomly selected one third of samples for training and two thirds of the samples for testing. The mean accuracy of the mentioned methods are shown in Table 1. We can find that our method results obtain higher accuracy than other methods. From the first three rows of Table 1, we can notice that the method fusing the geometrical features and CDDH feature can obviously increase the recognition rate. The reason of the method in [10] with slightly lower recognition rate is that the method is not robust to the adhesion of fingers, as shown in Figure 3(d). The proposed method performs much better while dealing with the adhesion problem. The main reason is that our proposed CDDH feature segments the hand gesture into N concentric circles and the adhesion of fingers does not affect the depth distribution in each circle. Figure 4(b) illustrates the confusion matrix of the experiment for our method. From the confusion matrix, it can be obviously seen that the most confusing gestures are Gesture 2, which are attached along with the confusion matrix. We argue that the main reason is the casualness of the performance and the precision of Kinect. For other gestures, our method achieves fairly higher accuracy in the experiment.

Table 1. The mean of the recognition accuracy

Different algorithms	Gesture 1	Gesture 2	Gesture 3	Gesture 4	Gesture 5
Geometrical feature + CDDH feature + RFs	99.2%	98.2%	99%	98.6%	99.2%
Geometrical feature + RFs	89.2%	86.4%	86%	87%	90%
CDDH feature + RFs	96%	95.5%	97%	95%	98%
Shape parameters [10]	97.5%	98%	98.2%	97%	97.6%

5. **Conclusion.** In this paper, we propose a new approach of hand gesture recognition based on depth spatial distribution characteristics under the complex background. We effectively take the advantage of Kinect and both of CDDH and geometrical features extracted from the depth image, which makes the extraction of the depth distribution histogram more simple and practical. Extensive experimental results show that our method performs good adaptability and stability under the complex environment and unstable light conditions compared with other methods. However, some improvements need to be done in the future work, which are shown as follows: 1) ability to recognize the hand gesture with small distinction; 2) ability to recognize dynamic hand gesture.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (61201429, 61362030), China Postdoctoral Science Foundation (2015M5817 20).

REFERENCES

[1] B. A. Myers, A brief history of human computer interaction technology, *ACM Interactions*, vol.5, no.2, pp.44-54, 1998.

[2] L. Kong and W. Yang, Saliency and similarity learning via ranking-SVM based hand gesture recognition, *Journal of Information & Computational Science*, vol.11, no.14, pp.5257-5266, 2014.

[3] M. Jiang and K. Jiang, Robust hand gesture recognition using depth image, *Journal of Computational Information System*, vol.11, no.3, pp.1093-1100, 2015.

[4] G. Dewaele and F. Devernay, Hand motion from 3D point trajectories and a smooth surface model, *Proc. of Eur. Conf. Computer Vision*, Prague, Czech Republic, pp.495-507, 2004.

[5] V. A. Ramirez and S. A. Mota-Gutierrez, A hand gesture recognition system based on geometric features and color information for human computer interaction tasks, *Proc. of Robotics Summer Meeting*, Mexico, 2011.

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

[7] J. Han, L. Shao, D. Xu and J. Shotton, Enhanced computer vision with microsoft Kinect sensor: A review, *IEEE Trans. Cybernetics*, vol.43, no.5, pp.1318-1334, 2013.

[8] O. Rashid, A. Al-Hamadi, A. Panning and B. Michaelis, Posture recognition using combined statistical and geometrical feature vectors based on SVM, *World Academy of Science, Engineering and Technology*, vol.3, no.8, pp.1570-1577, 2009.

[9] L. Breiman, *Random Forests*, Machine Learning, 2001.

[10] M. Panwar, Hand gesture recognition based on shape parameters, *International Conference on Computing, Communication and Applications*, pp.1-6, 2012.