# REAL TIME HAND GESTURE RECOGNITION VIA FINGER-EMPHASIZED MULTI-SCALE DESCRIPTION

*Jianyu Yang[1,2,*], Chen Zhu[1], Junsong Yuan[2]*

1. School of Urban Rail Transportation, Soochow University, China
2. School of EEE, Nanyang Technological University, Singapore
jyyang@suda.edu.cn, czhu@stu.suda.edu.cn, jsyuan@ntu.edu.sg

## ABSTRACT

The development of depth cameras, e.g., the Kinect sensor, provides new opportunities for human computer interaction (HCI). Although the Kinect sensor has been extensively applied for human tracking, human action recognition and hand gesture recognition, real time hand gesture recognition is still a challenging problem. In this paper, we propose a new real time hand gesture recognition method. To represent the noisy and articulated hand shape segmented from the Kinect images, a finger emphasized multi-scale descriptor is proposed. To fully utilize hand shape features, this descriptor incorporates three types of parameters of multiple scales, which emphasize the finger features. Hand gesture recognition is then achieved with both DTW algorithm and BP neural network. Extensive experimental results and the comparison with state-of-the-art methods demonstrate that our method is accurate (a 100% accuracy on a challenging hand gesture dataset), efficient (average 0.941ms per frame), and robust to noise, articulations and rigid transformations.

***Index Terms***— Multi-Scale Descriptor, Hand Gesture Recognition, Human-Computer Interaction, RGB-D

## 1. INTRODUCTION

Hand gesture recognition has always been an important topic in computer vision for its extensive applications in human-computer interaction (HCI), including virtual reality, sign language recognition and computer games [1]. Adopting hand gesture as an interface allows communications and manipulations in the non-contact environments. Traditional methods attach sensors or markers on the fingers, e.g., the data gloves [2][3], to capture hand gestures via electro-mechanical or magnetic sensing. These methods are effective to provide complete and real-time measurements of hand gestures, however, they hinder the natural motion of hand and are unapplicable in non-contact environments. Moreover, the devices are expensive for casual use and require complex calibration.

*Corresponding author. This work was supported by the National Natural Science Foundation of China (NSFC No. 61305020), and the Singapore MoE Tier-2 project MOE2015-T2-2-114.

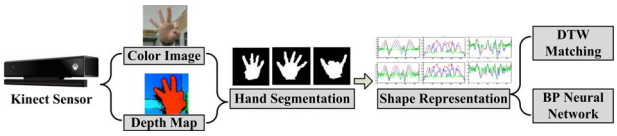


**Fig. 1**. The framework of our hand gesture recognition system.

The vision-based hand gesture recognition methods [4][5][6] give an alternative solution to the problems, which can be used naturally in non-contact environments. However, due to the limitations of the optical sensors, the captured images are sensitive to lighting conditions and cluttered backgrounds. Thus these methods usually cannot detect and track the hand robustly. Therefore, the traditional vision-based methods are far from satisfactory for real-life applications.

With the development of the depth cameras, e.g., the Kinect sensor [7], hand gesture recognition can be explored in a new form. The hand occupies a small area of the image with significant articulations, noises and distortions, which affects the recognition result. The classic shape recognition methods, e.g., the shape-context-based methods [8][9] and the skeleton-based methods [10][11], cannot recognize hand gestures robustly under severe articulations and distortions. The part-based methods [12] was proposed to solve these problems, but they cannot capture complete hand shape features for sufficient robustness and accuracy. Furthermore, these methods are not so efficient for real time applications. It is still a challenging problem to use depth sensor for real time hand gesture recognition.

In this work, a new real time hand gesture recognition method is proposed. We propose a finger-emphasized multi-scale descriptor (FMD) for hand gesture shape representation based on the IMD object descriptor [13]. Different types of parameters are defined in multiple scales for discriminative and complete representation of hand shapes, where the finger features are emphasized. Two recognition methods are built based on DTW and BP neural network for various applications.

Fig. 1 shows the framework of the proposed real time hand gesture recognition system. The Kinect sensor is used to capture both the color image and depth map of hand gestures as input. The hand is detected and segmented from the cluttered background by means of the depth map. The hand shape is represented by the proposed FMD descriptor, and recognized by the recognition methods.

Extensive experimental results validate that our method is robust to noise, articulated variation and rigid transformation. The recognition accuracy is evaluated on the latest challenging hand gesture datasets [12][14][15], and our method outperforms state-of-the-art methods a lot, achieves a $100\%$ score. The mean running time of our method based on BP neural network is less than *1ms*, which supports real time applications.

## 2. RELATED WORK

There are various vision-based hand gesture recognition methods proposed in the literature [16][17][18], and most of them are summarized in [6][19]. Generally, there are two main categories. One category is statistic model based methods, e.g., HMM models [16] and particle filtering [18]. The other category is based on a set of predefined rules [17]. The color makers are used on fingers and palms to detect the positions of joints and fingertips [4][20], which are sensitive to cluttered background. The hand region representation using skin color models [2][21] faces the similar problem that will be confused with the background.

Some researchers make use of the 3D features of hands or structured light to reconstruct the 3D hand surface [2][22], but the high computational cost restrains the real time application of these methods. Stereo camera is also used to track the trajectories of the hand surface points [23]. The multi-camera system can be used to reconstruct the 3D hand surface information. However, they also face the problem of high computational cost and the expensive devices make it far from real-life applications.

The development of depth cameras provides a robust solution of the problems. However, the hand shapes segmented from the depth maps are not accurate, which includes significant noise and articulations. These problems affect the recognition performance. Then, various hand shape representation methods are proposed. The classic shape context methods [8][9] represent the hand shape contours. The skeleton-based method [11] represents the hand shape as path topology. Bai et al. [10] proposed a skeleton pruning method to make the method robust to noise. The near-convex decomposition method [12] decomposes hand into fingers, which makes a superior performance in the relate work. The deep learning based methods [24][25] are also employed for hand gesture recognition recently.

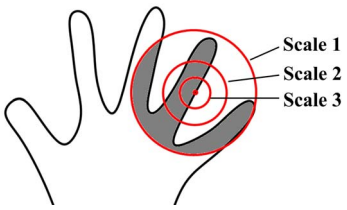


**Fig. 2**. The four circles indicate three scales 1-3. The grey area in each circle is used to calculate the parameters in respective scale.

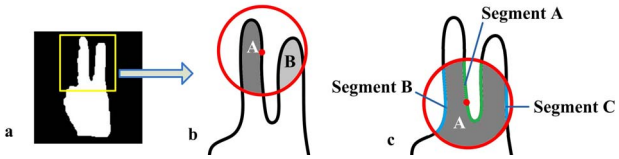


**Fig. 3**. Demonstration of the FMD parameters. Zones A in (b) and (c) are the major zones within the circle. Zone B in (b) is not a major zone. Segment A in (c) is a major segment, while Segments B and C are not.

## 3. HAND GESTURE DESCRIPTION

As shown in Fig.1, the hand shape is segmented from the RGB-D data using the similar method in [12]. To represent the hand shape, a finger-emphasized multi-scale descriptor (FMD) is proposed based on the invariant multi-scale descriptor (IMD) [13], which includes three types of parameters in multiple scales. Different from the IMD, the FMD is specially designed for describing hand gestures, where the fingers are emphasized in order to capture the salient feature.

The segmented hand shape is a closed contour consists of a sequence of contour points, denoted by $S = \{p(i)|i \in [1, n]\}$, where $n$ is the contour length and $p(i)$ is parameterized as the coordinates $p(i) = \{u(i), v(i)\}$ in the image. The FMD descriptor $I$ is defined as follows:

$$I = \{s_k(i), l_k(i), c_k(i)|k \in [1, m], i \in [1, n]\}, \qquad (1)$$

where $s_k$, $l_k$ and $c_k$ are the three types of parameters: area $s$, arc length $l$ and central distance $c$ in Scale $k$ (see Fig. 2). $k$ is the scale label, and $m$ is the total scale number. All these parameters are normalized and finger-emphasized parameters.

Consider a circle $C_k(i)$ with radius $r_k$ centered at $p(i)$, it covers zones and contour segments as shown in Fig. 3. Denote Zone A the major zone of $p(i)$ since it covers $p(i)$ inside the circle, while Zone B is not. Similarly, denote Segment A the major segment of $p(i)$ since it cross $p(i)$, while Segments B and C are not. Then, the FMD parameters are calculated from the area of Zone A, arc length of Segment A and weighted center of Zone A, as follows:

$$s_k(i) = \frac{s_k^*(i) \cdot d(i)}{(\pi r_k^2)} = \frac{\int_{C_k(i)} B(Z_k(i), x)dx \cdot d(i)}{(\pi r_k^2)}, \quad (2)$$

$$l_k(i) = \frac{l_k^*(i) \cdot d(i)}{(2\pi r_k)} = \frac{\int_{C_k(i)} B(S_k(i), x)dx \cdot d(i)}{(2\pi r_k)}, \quad (3)$$

$$c_k(i) = \frac{c_k^*(i) \cdot d(i)}{r_k} = \frac{||p(i) - w_k(i)|| \cdot d(i)}{r_k}, \quad (4)$$

$$w_k(i) = \frac{\int_{C_k(i)} B(Z_k(i), x)xdx}{\int_{C_k(i)} B(Z_k(i), x)dx}, \quad (5)$$

$$B(Z_k(i), x) = \begin{cases} 1, \text{if x belong to } Z_k(i), \\ 0, \text{for else.} \end{cases} \quad (6)$$

where $B(Z_k(i), x) : \boldsymbol{R}^2 \times \boldsymbol{R}^2 \longmapsto \{0, 1\}$ is an indicator function on Zone A ($Z_k(i)$) and Segment A ($S_k(i)$), $w_k(i)$ is the weighted center of Zone A. Then, $s_k^*(i)$, $l_k^*(i)$ and $c_k^*(i)$ are the area of Zone A, arc length of Segment A and central distance of Zone A, respectively. $s_k(i)$, $l_k(i)$ and $c_k(i)$ are normalized parameters by the area, circumference and radius of $C_k(i)$, respectively. Furthermore, there three FMD parameters are finger-emphasized by the factor $d(i)$. Since the weighted center of a hand is inside the palm, $d(i)$ is used to emphasize the features of fingers and is defined as the normalized distance from $p(i)$ to the weighted center:

$$d(i) = \frac{1}{2\sqrt{area_S}}||p(i) - p_{center}|| = \frac{1}{2\sqrt{area_S}}||p(i) - \frac{\int_{shape} xdx}{\int_{shape} dx}||, \quad (7)$$

where $area_S$ is the area of the whole hand shape, and $\frac{1}{2\sqrt{area_S}}$ is used to normalize the finger-emphasize factor $d(i)$ between 0-1. This factor is to make the FMD descriptor sensitive to fingers.

We should note that, the radius of the circle $C_k$ in each scale should be set first. In our method, the radius $r_k$ in different scales are set with respect to an initial radius $R$:

$$r_k = \frac{R}{2^k}, \quad (8)$$

where $r_k$ is half of $r_{k-1}$ in the prior scale. The setting of the initial $R$ is according to the hand size:

$$R = \sqrt{area_S}, \quad (9)$$

where $area_S$ is the area of the whole hand shape.

In Fig. 2, an example with circles from Scale 1 to Scale 3 is shown. From the figure we can see that $C_1$ covers three fingers, while the $C_3$ covers only a local area of present finger. Therefore, the FMD parameters in different scales represent full information of the hand shape.

The total scale number $m$ also needs to be carefully selected. It worth noting that, the scales are not always the more the better. A too big $m$ introduces extra noise which lower the accuracy of recognition as well as increase computational cost. In this work, $m$ is set according to a so called convergence condition: if the average difference of the FMD parameters between two neighbor scales $m$ and $m + 1$ is less than a threshold, e.g., $1 \times 10^{-2}$, the parameters in scale $m+1$ are unnecessary.

## 4. HAND GESTURE RECOGNITION

### 4.1. Recognition via DTW Alignment

Since the representation of hand gesture is a sequence of FMD parameters, the pairwise matching is an intuitive solution. The dynamic time warping algorithm (DTW) [26] algorithm calculates the best alignment (minimum matching distance) between two FMD sequences. Given two FMD sequences $I_A = \{s_k^p(i), l_k^p(i), c_k^p(i)|k \in [1, m], i \in [1, n_A]\}$ and $I_B = \{s_k^q(j), l_k^q(j), c_k^q(j)|k \in [1, m], j \in [1, n_B]\}$, the distance $d(p_i, q_j)$ between two contour points $p_i$ and $q_j$ is defined as the Euclidean distance of their FMD parameters:

$$d(p_i, q_j) = \sqrt{\sum_{k=1}^{m} ((s_k^p(i) - s_k^q(j))^2 + (l_k^p(i) - l_k^q(j))^2 + (c_k^p(i) - c_k^q(j))^2)}, \quad (10)$$

and the accumulative minimum matching distance up to the present corresponding points $p_i$ and $q_j$ is defined as follows:

$$D(i, j) = min(D(i-1, j-1), D(i-1, j), D(i, j-1)) + d(p_i, q_j), \quad (11)$$

and the total distance is $D_{A,B} = D(n_A, n_B)$. The less distance indicates the more similarity.

### 4.2. Recognition via BP Neural Network

Besides pairwise matching, a classifier is also trained based on hand gesture datasets. The BP Neural Network (BPNN) [27] is use it to model the relationship between the FMD description of hand gestures and their class labels. The network consists of input layers, concealed layers and output layers. The three-layers network with one concealed layer has been proved to be capable of approaching any multi-variable polynomial function [27]. To make the dimension of the input data consistent, the contour point sequences are regularized to the same length by average sampling a fixed number of contour points. In this work, the regularized contour length is set to 100, since a longer contour can not increase the performance of recognition in the experiment. Given a hand gesture $A$ represented by its FMD description $I_A$:

$$I_A = \{s_k(i), l_k(i), c_k(i)|k \in [1, m], i \in [1, 100]\}, \quad (12)$$

all the FMD parameters $\{s_1(1), l_1(1), c_1(1), s_1(2), \ldots, s_m(100), l_m(100), c_m(100)\}$ are used as the input data $\{x_1, x_2, \ldots, x_{n_{input}}\}$ of BP neural network, where $n_{input} = 3 \times 100 \times m$ is the total number of the FMD parameters in $I_A$. $\{y_1, y_2, \ldots, y_{n_{output}}\}$ are the output values, where $n_{output}$ is the total class number. We use each $y_i$ to indicate the recognition result of Class i. $w_{ij}$ and $w_{jk}$ are the weights of the BP neural network. The number of the hidden layer spots is determined by the empirical formula:

$$n_{hidden} = \sqrt{n_{input} + n_{output}} + c, \quad (13)$$

where $c$ is integer between 1-10. The initial weight of BP network is defined in [0, 1]. The Levenberg-Marquardt algorithm (L-M algorithm) is employed to compute and update

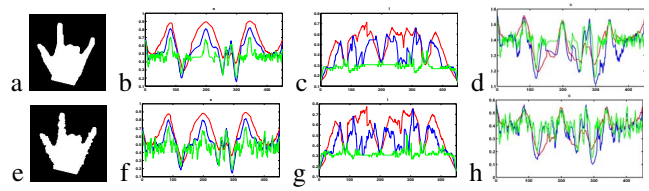


**Fig. 4**. The first column includes both the smooth and noisy hand shapes of the same gesture. Colums 2-4 are their corresponding FMD plots of $s$, $l$ and $c$. The red, blue and green plots indicate the Scales 1-3 of the FMD, respectively.

$w_{ij}$ and $w_{jk}$, since its fast convergence speed. The completely trained BP neural network is use as the prediction model of hand gestures.

It is worth noting that, the DTW has advantage that the recognition can start from an empty database without a prior training process, while the BP neural network can make use of the prior knowledge and boost high-level functions.

## 5. EXPERIMENTS

The capability of the proposed method is evaluated in two aspects: (1) demonstrate the robustness of our method to noises, articulated variations and rigid transformations; (2) evaluate the accuracy and efficiency of our method by an extensive comparative study. All the tests are implemented on a Intel Core 2 Quad 2.66 GHz CPU with 3G of RAM.

### 5.1. Robustness of Our Method

#### *5.1.1. Robustness to Noises*

The hand shapes always include significant noise as shown in Fig. 7, the bottom row. In this experiment, the FMD parameters of both the noisy shapes and smooth shapes are plotted and compared in Fig. 4. The binary image (a) is a smooth hand shape and (e) is the same gesture with significant noise. Columns 2-4 are the corresponding FMD plots: $s$, $l$ and $c$. The red, blue and green plots indicate the Scales 1-3 of the FMD, respectively. The corresponding plots in each column are very similar, which indicates that the proposed method is robust to noise. Only the third scale (green) is a little distorted, because the third scale (with smaller circle) is more disturbed by noise. Specifically, the absolute value of the bottom $l$ is higher than the top one, since the noise increases the arc length of the contour.

#### *5.1.2. Robustness to Articulated Variations*

This experiment validates the robustness to articulated variations. In Fig. 5, (a),(e),(i) are the same gesture with articulated variations, and Colums 2-4 are their corresponding FMD plots. Although the fingers of the hand gestures are heavily articulated and distorted, we can find the strong similarities among the corresponding plots in each column, which verify the robustness of our method to articulated variations.

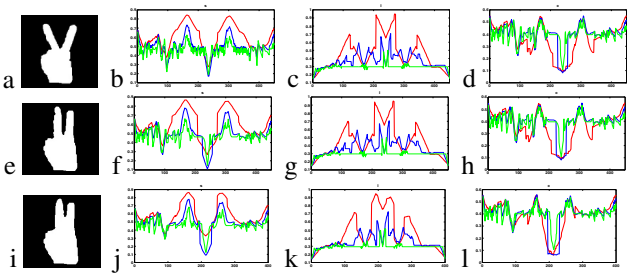


**Fig. 5**. The hand shapes with articulated variations and their FMD plots.

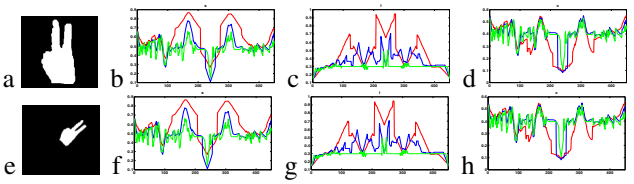


**Fig. 6**. The hand shapes with rigid transformations and their FMD plots.

#### *5.1.3. Robustness to Rigid Transformations*

This experiment validates the robustness to rigid transformations, including rotation, scale variation and translation. From Fig. 6 we can see that (a) is rotated, scaled and translated to (e), and Colums 2-4 are their corresponding FMD plots. The corresponding plots are exactly the same, which indicates that our method is robust to rigid transformations.

### 5.2. Performances of Hand Gesture Recognition

#### *5.2.1. NTU Dataset*

In this experiment, we use the challenging NTU hand gesture dataset [12] where the hand gestures are collected by a Kinect sensor. This dataset is collected from 10 subjects and includes 10 gesture classes. Each subject performs the same gesture in 10 different poses, thus the dataset has 10(people)×10(gestures)×10(poses)=1000 samples. Each of them contains a color image and a corresponding depth image. This is a very challenging real-life dataset collected in cluttered backgrounds. Moreover, the samples of the same gesture class have variations in hand orientation, scale, articulation, etc. The 10 hand gestures with corresponding shape samples are shown in Fig. 7.

We test our method with both DTW matching and BPNN classification, and half of the hand gestures are used for training and half for test. This experiment is repeated over 50 times while changing the training and testing data. The

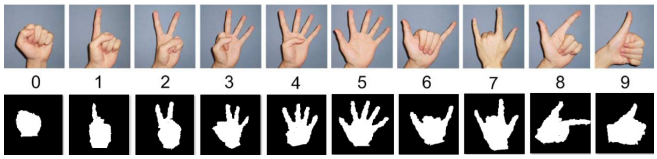


**Fig. 7**. Ten gestures of the hand gesture dataset [12] in the top row and the corresponding shape samples in the bottom row.

**Table 1**. Accuracy with Different Scale Number $m$

| Scale Number $m$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| FMD+DTW matching | 99.6 | **100** | 99.2 | 98.6 |
| FMD+BP neural network | 98.8 | **99.4** | 98.5 | 97.9 |

FMD description in this test is used with different scale numbers from 1 to 4. The mean accuracies with different scales are listed in Table 1. From the results we can find that our method performs well, especially achieves a $100\%$ accuracy when a two-scale FMD descriptor is used with DTW. Moreover, from the table we can see that the accuracy drops when the scale number is bigger than 2, which meets the convergence condition in the setting of total scale number $m$.

We also test the efficiency of the proposed methods, and compare them extensively with state-of-the-art methods as shown in Table 2. The best accuracies of our methods (when $m = 2$) are listed as well as their corresponding average efficiencies. We compare our methods with the correspondence-based algorithm, Shape Context (SC) [8][9], and the skeleton-based method, Path Similarity (PS) [10], as well as the latest part-based methods, Near-convex Decomposition (ND) and Thresholding Decomposition (TD) [12].

From Table 2 we can find that our methods achieve the best performance in both accuracy and efficiency. Both the accuracies of FMD+DTW and FMD+BPNN are very high, outperform all other methods. Especially, the FMD+DTW matching obtains a $100\%$ accuracy. Furthermore, the efficiencies of the proposed methods are significantly higher than that of other methods, which can fully support real time hand gesture recognition. Comparing the two proposed methods, the accuracy of FMD+BPNN is just a bit lower than that of FMD+DTW. However, its mean running time is less than 1ms, which is significantly faster than all other methods.

**Table 2**. Accuracy and Efficiency Comparison

| Method | Accuracy (%) | Time (s/query) |
|---|---|---|
| SC [8] | 83.2 | 12.346 |
| SC+Bending Cost [9] | 79.1 | 26.777 |
| Skeleton Match [10] | 78.6 | 2.4449 |
| ND+FEMD [12] | 93.9 | 4.0012 |
| TD+FEMD [12] | 93.2 | 0.0750 |
| FMD+DTW | **100** | 0.0157 |
| FMD+BPNN | 99.4 | **0.000941** |

**Table 3**. PadovaU Dataset I

| Method | Accuracy (%) |
|---|---|
| Marin et al. [14] | 91.3 |
| Joint Calibration [28] | 96.5 |
| FMD+DTW | **98.2** |
| FMD+BPNN | 95.8 |

**Table 4**. PadovaU Dataset II

| Method | Accuracy (%) |
|---|---|
| Memo et al. [15] | 90 |
| FMD+DTW | **99.3** |
| FMD+BPNN | 97.9 |

### *5.2.2. PadovaU Datasets*

The PadovaU datasets [14][15] are also employed in our experiments for testing and comparison. The PadovaU dataset I consists of 1400 hand gesture samples: 14 (subjects) $\times$ 10(gestures) $\times$ 10 (poses), captured from both the Kinect Sensor and Leap motion sensor. We test our methods in the same manner as the NTU dataset and compare the results with other methods in Table 3. From the result we find that the DTW accuracy outperforms other methods, while the BPNN is not so good as the result in [28] . That is because the method in [28] uses both the kinect data and the leap motion data, while our method employs only the kinect data.

The PadovaU dataset II consists of 1320 hand gesture samples: 4 (subjects) $\times$ 11(gestures) $\times$ 30 (poses), captured from the Creative Senz3D sensor. The accuracies of our methods comparing with other method are shown in Table 4. Both our methods outperform the previous method [15], and the DTW accuracy is nearly the full score. The results verify that our methods are adaptable for different sensors and applications.

## 6. CONCLUSIONS

In this work, we proposed a hand gesture recognition system using the Kinect sensor. A finger-emphasized multi-scale descriptor is proposed for hand gesture representation, which is robust to noises, hand articulations and rigid transformations. We perform two hand gesture recognition methods by DTW and BPNN, respectively. Extensive experiments on the challenging hand gesture datasets validate the robustness, accuracy and efficiency of our method. Our method is also applicable for real time applications.

## 7. REFERENCES


[1] J. P. Wachs, M. Kolsch, H. Stern, and Y. Edan, Vision-based handgesture applications, Commun. ACM, vol. 54, pp. 60-71, 2011.

[2] G. Dewaele, F. Devernay, and R. Horaud, Hand motion from 3D point trajectories and a smooth surfacemodel, in Proc. ECCV, 2004.

[3] E. Foxlin, Motion tracking requirements and technologies, Handbook of Virtual Environment Technology, pages 163-210, 2002.

[4] C. Chua, H. Guan, and Y. Ho, Model-based 3d hand posture estimation from a single 2d image, Image and Vision Computing, 20:191-202, 2002.

[5] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla, Filtering using a tree-based estimator, in Proc. of IEEE ICCV, 2003.

[6] G. R. S. Murthy and R. S. Jadon, A review of vision based hand gesture recognition, Int. J. Inf. Technol. Knowl. Manage., vol. 2, pp. 405-410, 2009.

[7] J. Shotton, A. Fitzgibbon, M. Cook, and A. Blake, Real-time human pose recognition in parts from single depth images, in Proc. IEEE CVPR, 2011.

[8] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, pp. 509-522, 2002.

[9] H. Ling and D. W. Jacobs, Shape classification using the inner-distance, IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, pp. 286-299, 2007.

[10] X. Bai and L. J. Latecki, Path similarity skeleton graph matching, IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, pp. 1-11, 2008.

[11] K. Siddiqi, S. Bouix, A. R. Tannenbaum, and S. W. Zucker, Hamilton Jacobi skeletons, Int. J. Comput. Vision, vol. 48, pp. 215-231, 2002.

[12] Z., Ren, J., Yuan, J., Meng, Z., Zhang, Robust Part-Based Hand Gesture Recognition Using Kinect Sensor, in IEEE Trans. Multimedia, vol. 15, no. 5, pp. 1110-1120, 2013.

[13] J. Yang, H. Wang, J. Yuan, Y. Li and J. Liu, Invariant multi-scale descriptor for shape representation, matching and retrieval, Computer Vision and Image Understanding, Vol. 145, pp. 43-58, 2016.

[14] G. Marin, F. Dominio, and P. Zanuttigh, Hand gesture recognition with Leap Motion and Kinect devices, IEEE Int. Conf. on Image Processing (ICIP), Paris, France, 2014.

[15] A. Memo and P. Zanuttigh, Head-mounted gesture controlled interface for human-computer interaction, Multimedia Tools and Applications, available online, 2017.

[16] A. Wilson and A. Bobick, Parametric hidden markov models for gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, pp. 884-900, 1999.

[17] M.C. Su, A fuzzy rule-based approach to spatio temporal hand gesture recognition, IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 30, pp. 276C281, 2000.

[18] C. Kwok, D. Fox, and M. Meila, Real time particle filters, Proc. IEEE, pp. 469-484, 2004.

[19] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle and X. Twombly, Vision based hand pose estimation: A review, Comput. Vision Image Understand., vol. 108, pp. 52-73, 2007.

[20] Y. Fang, K. Wang, J. Cheng and H. Lu, A real time hand gesture recognition method, in Proc. IEEE Int. Conf. Multimedia and Expo, 2007, pp. 995C998.

[21] M. H. Yang, N. Ahuja and M. Tabb, Extraction of 2d motion trajectories and its application to hand gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, pp. 1062-1074, 2002.

[22] M. Reale, S. Canavan, L. Yin, K. Hu and T. Hung, A multi-gesture interaction system using a 3D iris disk model for gaze estimation and an active appearance model for 3D hand pointing, IEEE Trans. Multimedia, vol. 13, pp. 474-486, 2011.

[23] M. Bray, E. Koller-Meier and L. V. Gool, Smart particle filtering for 3D hand tracking, in Proc. IEEE Int. Conf. Face and Gesture Recognition, Los Alamitos, CA, USA, 2004, pp. 675-680.

[24] P. Molchanov, X. Yang, S. Gupta, et al., Online Detection and Classification of Dynamic Hand Gestures With Recurrent 3D Convolutional Neural Network, in Proc. IEEE Int. Conf. CVPR, 2016.

[25] Q. Ye, S. Yuan, T. Kim, Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation, in Proc. ECCV, 2016.

[26] L. Rabiner and B.Juang, Fundamentals of Speech Recognition, Prentice-Hall, New Jersey, 1993.

[27] Robert Hecht-Nielsen, Theory of the backpropagation neural network, Int. Joint Conf. IEEE Neural Networks, IJCNN., pp. 593-605, 1989.

[28] G. Marin, F. Dominio and P. Zanuttigh, Hand Gesture Recognition with Jointly Calibrated Leap Motion and Depth Sensor, Multimedia Tools and Applications, Vol. 75, pp. 14991-15015, 2015.