Real-Time 2DHoG-2DPCA Algorithm for Hand Gesture Recognition

Omnia S. ElSaadany and Moataz M. Abdelwahab

Nile University, Egypt
omnia.samy@nileu.edu.eg,
mabdelwahab@nileuniversity.edu.eg

Abstract. Hand gesture recognition is one of the most challenging topics in computer vision. In this paper, a new hand gesture recognition algorithm presenting a 2D representation of histogram of oriented gradients is proposed, where each bin represents a range of angles dealt with in a separate layer which allows using 2DPCA. This method maintains the spatial relation between pixels which enhances the recognition accuracy. In addition, it can be applied on either hand contour or image representing hand details. Experimental results were performed on the latest existing depth camera dataset. The comparison with reported methods confirms excellent properties of our proposed method and promotes it for real time applications.

Keywords: Human-Computer Interaction, Hand Gesture Recognition, 2DHoG, 2DPCA.

1 Introduction

Hand gesture recognition is an active research topic considering the need for natural human-machine interaction, yet it is still a very challenging problem. Hand gesture recognition has many applications that vary from games, virtual reality, robotics, touchless interface in health care environment, enhancing physical daily life of handicaps, and sign language recognition.

Many research efforts were conducted on this problem[1,2]. Existing methods can be described according to environmental setup, feature descriptors and classification techniques.

1. Environmental setup is divided into two categories: wearing a glove and hand free.

Wearing a glove [3]: data gloves provide accurate fingers tracking, hand shape and position. However, these data gloves are electro-magnetic devices, expensive and sometimes require calibration. Another attempt with gloves was conducted using a colored glove, [4,5] Where a normal cloth glove with distinguishable colored patches helps in hand segmentation. Yet, wearing a glove limits user comfort and natural interface.

Hand Free: provides more natural and friendly interface than wearing a glove. The required segmentation is based on vision techniques in which a normal intensity or depth camera is used. Intensity camera is easily affected by illumination, lightening condition and background objects/colors [6]. Real time depth camera provides intensity image and the corresponding depth map, that can be used in segmentation of a cluttered background [7].

2. Feature descriptors can be Spatial features or temporal features. Spatial features describe the shape and position of the hand employing: edges, regions, silhouettes, histograms, etc..[6]. Temporal features model the dynamics of a gesture using motion parameters and trajectories.
3. Classification techniques: Template matching techniques and nearest neighbor classifier based on distance measures (*i.e.* euclidean distance, mahalanobis distance, etc..) are mostly used. Other methods involves PCA, statistical models *i.e.* Neural Network, Decision Tree, SVM , HMM and parameter estimation techniques *i.e.* Maximum Likelihood and Maximum Posteriori.

In 2011, Zhou Ren [10] proposed a new method using kinect Microsoft depth camera, by which he collected a 10 hand gestures dataset. This method is based on time curve series and proposed a Finger-Earth Mover's Distance (FEMD). Hand shape representation using time curve series records the relative distance between each contour vertex to the hand center point. Earth Mover's Distance [11] evaluates dissimilarity between two distributions, where FEMD was designed for matching the fingers histogram to the stored templates. Experimental results were conducted on the collected dataset and achieved excellent accuracy and real time performance. It worth to mention that FEMD is limited to gestures where fingers aren't internally hidden.

In this paper we propose a real time approach that depends on a 2D representation of histogram of oriented gradients (2DHoG) rather than 1DHoG [12], to represent hand contour or inner details. This new representation allows using 2DPCA [13] which maintains the spatial relationship between image pixels [15]. Our method is more general and robust as we consider the whole hand shape either contour or inner details, rather than fingers' count and position, as other methods. Experimental results on existing datasets proved that our method achieved the best accuracy and less running time compared to the reported methods.

This paper is organized as follows: Section 2 introduces the proposed method, Section 3 discusses performance evaluation and shows a comparison with the most recent work, Finally conclusion is presented in Section 4

2 Proposed Method

In this section we describe our proposed hand gesture recognition method as shown in figure 1. Hand segmentation and alignment is described in Section 2.1, feature extraction and classification are presented in Section 2.2.

Fig. 1. The overall system description for: (a) training mode, (b) testing mode

2.1 Hand Segmentation and Alignment

In this paper we used the recent depth camera dataset reported in [10] that assumes that the hand is the closest object to the Kinect sensor and the user is wearing a black belt on his wrist. Hence, detection/segmentation is done by applying depth threshold from the nearest value to a certain range. The hand detected from segmentation has two issues: It contains part of the arm and is rotated as the dataset didn't constraint the user to a certain position, as shown in figure 2 (a). Both issues will be handled by: detecting the wrist belt (hand reference), fitting a line to the belt, crop the arm part and align the hand as shown in figure 2 (b), (c).

Fig. 2. (a) segmented hand, the detected belt line is shown in red, (b) hand after cropping the arm part, (c) hand alignment to the x-axis reference

2.2 Feature Extraction and Classification

For both training and testing modes, Histogram of oriented gradients (HoG)[12] representing hand (contour or inner details) angles' magnitude is extracted and

stored in a 2D representation B layers; the 2D representation maintains the spatial relationship of hand shape, where each layer represents the orientation in a certain range of angle as illustrated in figure 3.

The 2DHoG is computed as follows:

1. Compute image gradient in both horizontal and vertical directions I_x , I_y respectively, by applying a 3×3 Sobel mask
2. Compute gradient magnitude, mag , and orientation, θ , as follows:

$$mag(x, y) = \sqrt{I_x(x, y)^2 + I_y(x, y)^2}$$

$$\theta(x, y) = \arctan \frac{I_y}{I_x}$$

3. Divide the input image into cells and store the cumulative gradient magnitude in B bins (layer), where each bin represents orientation interval from 0° to $360^\circ/B$.

Fig. 3. (a) Aligned hand, (b) Hand grid cells, (c) Calculated 2DHoG where each color represents a single bin, (d) B layers of 2DHoG bins

Training Mode. to get the best representation of the gesture features with spatial relations maintained, We apply 2DPCA for each bin, $b = 1 \dots B$, of the 2DHoG as follows:

1. Average matrix for bin b , $\bar{X}_b$, is calculated as follows

$$\bar{X}_b = \frac{\sum_{i=1}^N X_{bi}}{N}$$

where N is the number of training samples, i is the training sample index, X_{bi} is the magnitude matrix of the b^{th} bin of the i^{th} training sample.

2. Covariance matrix Cov_b is calculated:

$$Cov_b = \frac{\sum_{i=1}^N (X_{bi} - \bar{X}_b)(X_{bi} - \bar{X}_b)^T}{N - 1}$$

3. Eigen values and the corresponding vectors of Cov_b are obtained:

$$Cov_b V = \lambda_b V_b$$

where λ_b is a diagonal matrix of eigen values and V_b is a matrix of the corresponding eigen vectors.

4. Store the dominant eigen vectors $\tilde{V}_b$
5. Store the feature matrix A_{bi} obtained by projecting the training matrix X_{bi} on the dominant eigen vectors

$$A_{bi} = (X_{bi} - \bar{X}_b) \tilde{V}_b$$

6. Steps 1-5 are repeated for all bins.

Testing Mode. for all bins:

1. Calculate feature matrix At_b obtained by projecting the testing matrix Xt_b on the stored dominant eigen vectors

$$At_b = Xt_b \tilde{V}_b$$

2. Classification is obtained using minimum Euclidean distance classifier. Each bin votes by a weight obtained from it's accuracy. a weighted sum of bins' votes is used to obtain one decision/class per tested gesture. For each test image, At_b , The Euclidean distance is calculated as follows:

$$\arg \min_{i \in N} Distance(i) = \sum_b^B acc_b \cdot \sum_x \sum_y |At_b(x, y) - A_{bi}(x, y)|$$

where acc_b is the b^{th} bin accuracy and N is the number of stored training features matrices.

3 Performance Evaluation

3.1 Datasets

Our first experiment was applied on Z. Ren's Dataset. The dataset was collected using kinect depth camera[10]. The ten hand gestures dataset is very challenging as it was collected under cluttered background, uncontrolled environment and high variation gestures in terms of color, scale and orientation. The kinect sensor dataset provides both color images and the corresponding depth maps as shown in figure 4 (a),(b).

The dataset consists of 10 persons performing 10 hand gestures ten times with different hand orientations as shown in figure 5. To test our proposed algorithm on images presenting the entire details of the hand and for more different gestures like sign language, we also tested our algorithm on the international sign language dataset, a twenty four different static signs were used in this dataset[14]. It Consists of 2040 images where the number of images of each gesture/sign is listed below:A: 40, B: 40, C: 40, D: 40, E: 40, F: 40, G: 100, H: 100, I: 100, K: 100, L: 100, M: 100, N: 100, O: 100, P: 100, Q: 100, R: 100, S: 100, T: 100, U: 100, V: 100, W: 100, X: 100, Y: 100

Fig. 4. Z. Ren's dataset sample: (a) Intensity image, (b) Corresponding depth map, (c) Segmented arm, (d) Cropped hand

Fig. 5. Ten hand gestures in Z. Ren's dataset

3.2 Experimental Results and Analysis

In the first experiment, the image is scaled to 50X50, where the aspect ratio is maintained. The 2DHoG was applied using 5X5 cells with 9 signed bins. For the sake of comparison a leave-one-out strategy was employed on Z. Ren dataset, where 9 persons' gesture images were used in the training phase and the 10th person's image is used in the testing phase. Our proposed algorithm was applied on the hand contour only as shown in figure 7 (a),(b).

As shown in table 1 our proposed method achieved mean accuracy of 91.6% and overall mean running time of 0.2029 seconds, where 2DHOG-2DPCA and preprocessing (segmentation and alignment) mean running time are 0.095 and 0.1079 seconds respectively.

Compared to Z. Ren [10], thresholding decomposition + FEMD method we achieve better accuracy for less running time, which promotes our method for real-time applications. Confusion matrix is shown in figure 8.

It worth to mention that these excellent properties were obtained using hand contour information only. Our method is more general than Z. Ren as it considers the entire hand shape not matching the finger tips only. Our method is robust under unconstrained environment including illumination and cluttered background,

Fig. 6. Twenty four static sign alphabet

Fig. 7. (a) Hand contour, (b) 2DHoG applied on the hand contour in 'a', (c) hand gray scale image, (d) 2DHoG applied on entire hand details in 'c', (e) 2DHoG applied on edges with magnitude threshold = 100, (f) 2DHoG applied on edges with magnitude threshold = 200

since hand segmentation was done using the depth information. In addition, our methods tolerates hand orientation's variation. Changes in hand articulation is also tolerable as we map gradient orientation to one of the B signed Bins.

For testing with more number of gestures the inner hand details will be an important information to distinguish between different gestures. As shown in figure 9 the 'C', 'E' and 'O' signs have similar outer contour, but the inner hand details are different.

Therefore in the second experiment, we used the international sign language dataset[14], where two experiments were conducted. In the first experiment, our proposed algorithm was applied on the whole segmented hand as shown in figure 7 (c),(d). While in the second experiment, our proposed algorithm was applied only on edges of the inner hand, where the edges are obtained by magnitude thresholding as shown in figure 7 (e),(f).

In both experiments the image is scaled to 32X32, where the aspect ratio was maintained. The 2DHoG was applied using 2X2 cells with 9 signed bins.

In the first experiment, our algorithm achieved 99.17% accuracy compared to 98.7% achieved by [14]. It worth to mention that the experiment performed in [14] requires to train by images from the same user to achieve these results, while

Table 1. Accuracy and mean running time for proposed method and Z. Ren method

	Proposed Method	Thresholding Decomposition + FEMD
Mean Accuracy	91.6%	90.6%
Mean Running Time	0.2029s	0.5004s

Fig. 8. Confusion matrix of the proposed method applied on Z. Ren dataset (mean accuracy 91.6%)

Table 2. Accuracy and mean running time for the proposed method on international sign language dataset

	Experiment 1	Experiment 2
Mean Accuracy	99.17%	97.0833%
Mean Running Time	0.28321s	0.28321s

Fig. 9. C, E and O outer contour in cyan shows the similarity between those letters contour

Fig. 10. Confusion matrix of the proposed method on the international sign language dataset (mean accuracy 99.17%)

Fig. 11. Confusion matrix of the proposed method applied on Thomas dataset thresholded edges(mean accuracy 97.083%)

our method does not require to train by the same user. The confusion matrix in figure 10 shows that we achieved 100% accuracy for 23 gestures, and only 1 gesture was 80%.

In the second experiment, our algorithm achieved 97.083% accuracy. The confusion matrix in figure 11 shows that we achieved 100% accuracy for 20 gestures, 90% for 3 gesture and 60% for 1 gesture.

Overall mean running time of 0.28321 seconds was achieved, where 2DHOG-2DPCA and preprocessing (segmentation and alignment) mean running time are 0.2674 and 0.01581 seconds respectively as shown in table 2, Which confirms the excellent properties of our algorithm.

4 Conclusion

Many applications in different fields are based on hand gesture recognition, where high accuracy with real-time performance is usually required. This paper proposed a hand gesture recognition algorithm achieving an excellent accuracy with real-time performance. Our algorithm presents a 2D representation of Histogram of oriented gradients maintaining spatial relationship that can be applied either on hand contour only or the entire hand image. Efficient feature extraction was applied employing 2DPCA where spatial relations between pixels are maintained. The overall results are based on a voting criteria where each bin’s layer votes based on it’s accuracy. Experimental results confirm these excellent properties.

References

1. Wu, Y., Huang, T.S.: Vision-based gesture recognition: A review. In: Braffort, A., Gibet, S., Teil, D., Gherbi, R., Richardson, J. (eds.) GW 1999. LNCS (LNAI), vol. 1739, pp. 103–115. Springer, Heidelberg (2000)

2. Mitra, S., Acharya, T.: Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews* 37, 311–324 (2007)
3. Dipietro, A.M., Sabatini, L., Dario, P.: A survey of glove-based systems and their applications. *Systems, Man, and Cybernetics, Part C: Applications and Reviews* 38, 461–482 (2008)
4. Lamberti, L., Camastra, F.: Real-time hand gesture recognition using a color glove. In: Maino, G., Foresti, G.L. (eds.) *ICIAP 2011, Part I. LNCS*, vol. 6978, pp. 365–373. Springer, Heidelberg (2011)
5. Wang, R.Y., Popovi, J.: Real-time hand-tracking with a color glove. *ACM Transactions on Graphics (TOG) - SIGGRAPH* 28 (2009)
6. Shan Lu Imagawa, K., Igi, S.: Color-based hands tracking system for sign language recognition. In: *Proceedings Third IEEE International Conference Automatic Face and Gesture Recognition*, pp. 462–467. IEEE (1998)
7. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from a single depth image. In: *CVPR. IEEE* (2011)
8. Ming-Hsuan, N., Yang, A., Tabb, M.: Extraction of 2d motion trajectories and its application to hand gesture recognition. *Pattern Analysis and Machine Intelligence* 24 (2002)
9. Liwicki, S., Everingham, M.: Automatic recognition of fingerspelled words in british sign language. In: *2nd IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2009)*, pp. 50–57 (2009)
10. Ren, Z., Yuan, J., Zhang, Z.: Robust hand gesture recognition based on finger-earth movers distance with a commodity depth camera. In: *Proceedings of the 19th ACM International Conference on Multimedia, MM 2011*, pp. 1093–1096. ACM, New York (2011)
11. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth movers distance as a metric for image retrieval. *IJCV*, 40–99 (2000)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference Computer Vision and Pattern Recognition, CVPR*, vol. 1, pp. 886–893. IEEE (2005)
13. Jian Yang, A.F., Zhang, D., Frangi, Y.J.Y.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (2004)
14. Birk, H., Moeslund, T.B., Madsen, C.B.: Real-time recognition of hand alphabet gestures using principal component analysis. In: *10th Scandinavian Conference on Image Analysis* (1997)
15. Abdelwahab, M.M., Yousry, I., Mikhael, W.: A novel algorithm for simultaneous face detection and recognition. In: *Circuits and Systems (MWSCAS)*, pp. 670–673. IEEE (2012)