Optik 124 (2013) 6795–6798

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

A temporal hand gesture recognition system based on hog and motion trajectory

Jing Lin, Yingchun Ding*

Department of Physics, Beijing University of Chemical Technology, Beijing 100029, China

ARTICLE INFO

Article history:
Received 11 January 2013
Accepted 23 May 2013

Keywords:
Human–computer interaction
Temporal gesture recognition
Histograms of oriented gradients
Support vector machine
Motion trajectory
Mahalanobis distance

ABSTRACT

A real-time, rapid and robust gesture recognition system is usually hindered by difficulty of hand localization and complexity of hand gesture modeling, especially under complex background. For eliminating these obstacles, in this paper, we propose a method using histograms of oriented gradients features (HOG) and motion trajectory information for temporal hand gesture recognition in natural environment. We firstly localize hand in video stream based on hand detection by HOG and support vector machine algorithm (SVM). After hand localization, the motion trajectory information of consecutive hand gesture is extracted and a database of standard gestures is built. Finally, the Mahalanobis distance between input gesture and database is computed for recognition. As the experimental results shown, our method exhibits a good performance in real-time test.

1. Introduction

In recent years, with the rapid development of computers, human–computer interaction or HCI, has become an increasingly important part of our daily lives [1]. As we know, the most popular mode of HCI is based on simple mechanical device, e.g. keyboard and mice [2,3]. However, these devices inherently limit the speed and naturalness with which we can interact with the computer [1]. With dexterous functionality in communication and manipulation, human hand has been regard as the most effective, general-purpose interaction tool for HCI [4]. As a result, the gesture recognition technique has attracted more and more attentions recently [5].

The gesture recognition is a complex subject especially sign language recognition, which involves many aspects such as motion modeling, motion analysis, pattern recognition and machine learning, even psycholinguistic [2]. A real-time, rapid and robust gesture recognition system is usually hindered by difficulty of hand localization and complexity of hand gesture modeling, especially under complex background. Two types of methods are often used in hand localization: skin color cues based method and motion cues based method [1,6]. The color-based methods [7–9] are not robust because of sensitivity to illumination and restriction on other skin-colored objects. And the motion-based methods [10,11] are also require many assumptions like static background and constantly moving of hand [7]. Furthermore, hand tracking technology [12] and hand segmentation algorithm [13] are also with many drawbacks.

Histograms of oriented gradients, called HOG, were firstly used by Dalal and displayed an excellent performance in human detection [14]. Zondag et al. used then HOG features combination with two variations of AdaBoost algorithm for construction a real-time hand detector. As experimental results shown, HOG showed a good performance [15]. Yu et al. [16] used HOG features for characterizing hand "shape", and constructed an classifier by support vector machine algorithm (SVM) for static hand gesture recognition. In their experimental testing, 9 defined postures were recognized with an accuracy of 94.49%. Sha et al. [17] extended the HOG to refine the best posture region and recognition. Experimental results showed promising performance under various capture conditions. In paper [18] HOG was also chosen to represent the hand shapes for automatically learn a large number of British Sign Language (BSL) signs from TV broadcasts.

In this paper, our goal is to design a simple, rapid, and robust gesture recognition system. Given the advantage of HOG in hand detection, we combined HOG and motion trajectory information of consecutive hand gesture for temporal gesture recognition. First, we construct a hand detector using HOG features and SVM algorithm. Secondly, the trajectory features are extracted for represented consecutive hand gesture. A database composed of standard features of 6 defined varieties of gestures is then established. Finally, Mahalanobis distance is calculated to sort the input gesture.

The rest of this paper is organized as follow. The HOG detector constructing is addressed in the next section. The extraction of trajectory feature, segmentation of consecutive gestures, database

* Corresponding author. Tel.: +86 18911131198.
E-mail address: dingyc@mail.buct.edu.cn (Y. Ding).

0030-4026/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijleo.2013.05.097

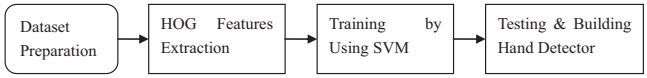

Fig. 1. The flow chart of constructing a hand detector.

establishing and final recognition are then presented in Section 3. Section 4 shows the experimental results in detail. Finally, conclusions are given in Section 5.

2. HOG detector construction

In our work, a HOG detector should be pre-constructed for hand detection. Fig. 1 shows the constructing processes.

2.1. HOG features extraction

Being different from other geometry features which considered integrality of an image, HOG feature, however, divides an image into many units which called cells. Firstly, gradient orients or edge orients and angles of each pixel are calculated all over the image using varieties of masks, e.g. central masks, Sobel masks, diagonal masks, and Prewitt masks. The histograms of gradient directions over each pixel of the cell are then accumulates. Some cells compose a region called block. Then an image can be regarded as a connection of many blocks. The concatenated histograms of the whole blocks form the vectors of HOG. Fig. 2 shows the structure of HOG.

2.2. SVM

SVM, proposed by Vapnik and his co-workers, is a supervised learning theory based on the notion of structural risk minimization which minimizes the upper bound of generalization error and provide excellent generalization ability [19]. The basic principle behind SVM is to map the input data set which contains both positive and negative examples to a high dimensional space to estimate separating by the hyperplane that has the largest distance to the neatest training data points of any class, since in general the larger the margin the lower the generalization error of the classifier.

There are four basic kernels for mapping data set as follows:

$$\text{Linear}: k(x_i, x_j) = x_i^T x_j. \tag{1}$$

$$\text{Ploynomial}: k(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \quad \gamma > 0. \tag{2}$$

$$\text{Radial basis function (RBF)}: k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \quad \gamma > 0. \tag{3}$$

$$\text{Sigmoid}: k(x_i, x_j) = \tan h(\gamma x_i^T x_j + r). \tag{4}$$

Here, γ, r, and d are kernel parameters.

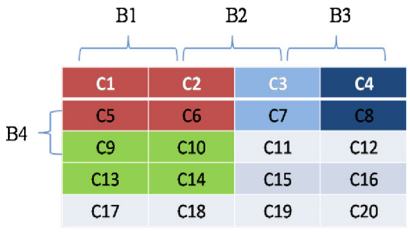

Fig. 2. Structure of HOG features.

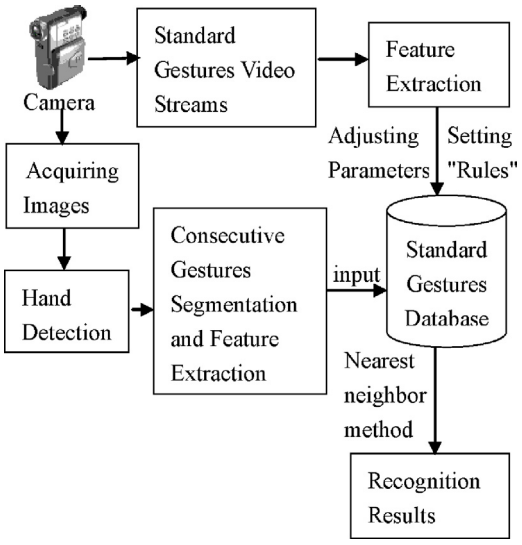

Fig. 3. Architecture of temporal gesture recognition.

2.3. Detector construction

For constructing a detector, it is crucial to prepare training samples which consist of positive samples and negative samples. In our work, the positive samples involve one class of defined hand, which acquired in different conditions including simple and complex background with varied illumination. In the acquisition process, the palm must face the camera generally. Similarly, the negative samples are collected from random views in different background. After dataset preparation, the HOG features are extracted according to the "optimal" parameters which determined referred to Jorn's empirical analysis [15]. A 324 dimensional vector composes finally HOG feature of hand in this paper.

After feature extraction, the RBF kernel is chosen for solving a nonlinear classification problem in the training step. The optimal parameters of SVM are defined using cross-validation technique [20]. An optimal trained hand model by SVM is finally chosen for hand detection after testing steps.

3. Temporal gesture recognition

Temporal gesture can be regarded as a continuous change sequence of gesture trajectory. For temporal gesture recognition, the methods based on Hidden Markov Model (HMM) are more efficient and successful, especially for sign language recognition. However, these methods need often a lot of computational time [2]. For avoiding this problem, a simple method is proposed for gesture recognition in this paper. Fig. 3 shows the architecture of our recognition system.

Firstly, we need build a standard gestures database beforehand after extracting features on standard gesture videos and setting "rules". In real-time recognition process, when a gesture inputs, the hand position are detected by hand detector. The consecutive gestures are then segmented and feature extraction is followed. Finally, nearest neighbor method is used to decide which class of gestures the input belongs to.

3.1. Trajectory features extraction

After hand detection, the hand position can be detected and the feature extraction is often followed. In this paper, hand position, velocity, and angle compose a 4 dimensional vector $V(x, y, v, \theta)$ for

Fig. 4. Positive samples and negative samples.

representing motion trajectory of gesture. The velocity and angle are computed as followed:

$$v_x = \frac{x_1 - x_2}{\Delta t} \tag{5}$$

$$v_y = \frac{y_1 - y_2}{\Delta t} \tag{6}$$

$$v = \sqrt{{v_x}^2 + {v_y}^2} \tag{7}$$

$$\tan\,\theta = \frac{y_1 - y_2}{x_1 - x_2} \tag{8}$$

$$\theta = \arctan\,\frac{y_1 - y_2}{x_1 - x_2} \tag{9}$$

where (x_1, y_1) and (x_2, y_2) are positions of hand in preceding frame and current frame respectively. v and θ represent velocity and angle of current frame.

3.2. Segmentation of consecutive gestures

A complete gesture can be usually divided into three phases: action preparation, movement and ending action. In this paper, a velocity filter is used for consecutive gestures segmenting. When the velocity v is more than a threshold V_{start}, it indicates the gesture is started. Similarly, when the velocity v is less than a threshold V_{end}, the gesture is considered to be end. Actually, it is not enough to segment consecutive gesture since the hand is not absolutely static in the end phase of gesture. Therefore, a threshold N_{end}, which indicates the continuous frames number of velocity less than V_{end}, is added for determine whether the gesture is finished or not. Furthermore, it must be guaranteed that every defined gestures with the same length. Otherwise, it cannot calculate the distance between an input gesture and standard gestures in database. For solving this problem, interpolation is used. When the length of gesture is more than the standard length L_c, some frames are removed. And when the length is more than L_c, some frames are interpolated.

3.3. Database establishing

In this paper, we need to build a standard database consisted of standard gesture feature data in advance. Firstly, some complete standard gesture, video streams are prepared by a written program. Each gesture is isolated and completed in uniform background, so it is easy to realize hand localization. For consecutive gesture segmentation, we need to set "rules" which must fixed all the defined gestures segmentation. In other words, we need to determine 4 parameters: V_{start}, V_{end}, N_{end} and L_c. In the database building process, the parameters are decided by large number of experiments. The trajectory features of each standard gesture are then extracted and consist of the database.

After establishing database, the nearest neighbor rule is applied for recognition. When a gesture inputs, its Mahalanobis distances to the standard gestures in database are calculated. The minimum distance decides which type of gestures it belongs to.

4. Experimental results

4.1. Dataset acquisition

In our work, two parts of dataset for hand detection and temporal gesture recognition respectively are almost obtained by a Unify 6100 web camera. The dataset for hand detection include positive samples and negative samples, which are shown in Fig. 4. To create positive samples, 1500 images from five different persons averagely in different conditions including simple and complex background with varied illumination. In the acquisition process, the palm must face the camera generally, since the hand shape is not changed in our defined gestures. Similarly, the negative samples consist of 5000 images from random views in different background.

The dataset for final temporal gestures recognition consist of standard samples and testing samples. The standard samples include 120 standard video streams for 6 defined gestures averagely under simple background. And the testing samples are also composed of 120 testing videos which include gestures in different conditions. Fig. 5 shows the defined gestures: "move up", "move down", "move right", "move left", "draw a circle", and "no".

4.2. Hand detection

For hand detection, 1000 positive samples and 4000 negative samples are divided to form training samples, and the rests compose the testing set.

Testing result on our testing dataset for hand detection based on HOG features shows the following result: 94.4% of detection rate and 9.4% of error rate. Therefore we can come to the conclusion that this method can meet the requirements of hand detection in our work.

4.3. Temporal gesture recognition

The parameters discussed in Section 3.2 are determined through a large number of experiments. We can finally get these optimal parameters: $V_{start} = 4$, $V_{end} = 8$, $N_{end} = 12$ and $L_c = 60$, according to the

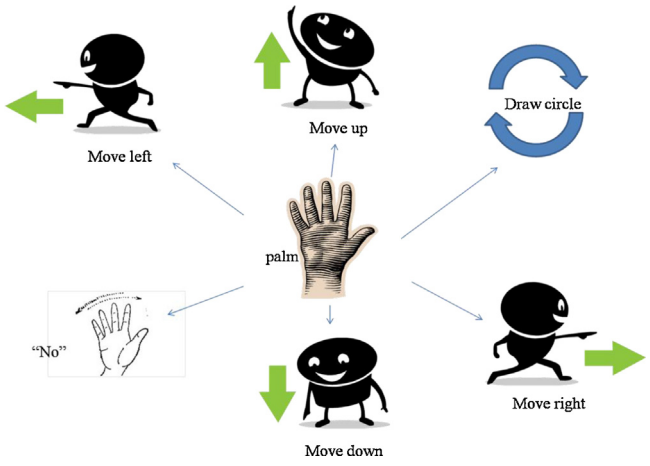

Fig. 5. The 6 defined gestures.

Table 1
Recognition rate of defined gestures.

Gesture	Move up	Move down	Move right	Move left	Draw circle	NO	Average of recognition rate
Recognition rate	100%	95%	95%	95%	80%	85%	91.7%

rule: the parameters must satisfy the segmentation and recognition of all the defined standard gestures.

Table 1 shows the recognition rates of the 6 defined gestures after the final testing. As shown in Table 1, we can find that the average recognition rate achieves 91.7%. And the experimental results verify the effectiveness of our method.

5. Conclusion

In this paper, we propose a method for 6 defined gestures recognition using HOG features and motion trajectory information. We construct firstly a hand detector based on HOG features and SVM algorithm for hand detection. The experimental results show the detector with a detection rate of 94.4% and 9.4% of error rate. Next, motion trajectory features of temporal gestures are extracted. In the feature extraction process, for consecutive gesture segmentation, 4 parameters are determined through large number of experiments. Moreover, a database consisted of standard gesture feature is built for providing criterion. Finally, the Mahalanobis distance is computed to sort the input gesture. The final testing results show our method exhibits a good performance in testing. However, in the testing process, we found that the gesture segmentation is not robust, especially for a user without pre-training. In other hand, the gestures we defined do not include complicated gestures since the palm can be only detected. As a result, this recognition system is not suitable for the complex gesture recognition. Therefore, in the future work, we would like to develop this method to adapt to complicated gestures.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under the Grant No. 60978006 and the Beijing Natural Science Foundation under the Grant No. 4122055.

References

[1] I. Vladimir, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 677–695.
[2] Y. Wu, T.S. Huang, Vision-based gesture recognition: a review, Lect. Notes Comput. Sci. 1739 (1999) 103–115.
[3] C. Shan, T. Tan, Y. Wei, Real-time hand tracking using a mean shift embedded particle filter, Pattern Recognit. 40 (2007) 1958–1970.
[4] A. Erol, G. Bebis, M. Nicolescu, et al., Vision-based hand pose estimation: a review, Comput. Vis. Image Understand. 108 (2009) 52–73.
[5] S. Mitra, T. Acharya, Gesture recognition: a review, IEEE Trans. Syst. Man Cybern. C: Appl. Rev. 37 (2007) 311–324.
[6] S.C.W. Ong, S. Ranganath, Automatic sign language analysis: a survey and the further beyond lexical meaning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 837–891.
[7] A.A. Bhuiyan, C.H. Liu, Intelligent vision system for human–robot interface, in: Proceeding of World Academy of Science, Engineering and Technology, vol. 22, 2007, pp. 57–63.
[8] D. Chai, K.N. Ngan, Face segmentation using skin-color map in videophone applications, IEEE Trans. Circuits Syst. Video Technol. 9 (4) (1999) 551–564.
[9] M. Soriano, B. Martinkauppi, S. Huovinen, et al., Skin detection in video under changing illumination conditions, in: 15th International Conference on Patter Recognition, vol. 1, 2000, pp. 839–842, 2000.
[10] N. Habili, C.C. Lim, A. Moini, Segmentation of the face and hands in sign language video sequences using color and motion cues, IEEE Trans. Circuits Syst. Video Technol. 14 (8) (2004) 1086–1097.
[11] A. Sundaresan, R. Chellappa, Multicamera tracking of articulated human motion using shape and motion cues, IEEE Trans. Image Process. 18 (2009) 2114–2126.
[12] M. Elmezain, A. Al-Hamadi, R. Niese, et al., A robust method for hand tracking using mean-shift algorithm and Kalman filter in stereo color image sequence, World Acad. Sci. Eng. Technol. 59 (2009) 283–287.
[13] V. Spruyt, A. Ledda, S. Geerts, Real-time multi-colourspace hand segmentation, in: 17th IEEE International Conference on Image Processing 2010, 2010, pp. 3117–3120.
[14] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 886–893.
[15] J.A. Zondag, T. Gritti, V. Jeanne, Practical study on real-time hand detection, in: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, pp. 1–8.
[16] R. Yu, G. Chengcheng, Hand gesture recognition based on HOG characters and SVM, Bull. Sci. Technol. 27 (2) (2011) 211–214.
[17] L. Sha, G. Wang, A. Yao, et al., Hand posture recognition in video using multiple cues, in: IEEE International Conference on Multimedia and Expo, 2009, pp. 886–889.
[18] P. Buehler, M. Everingham, A. Zisserman, Learning sign language by watching TV (using weakly aligned subtitiles), in: IEEE Conference on Computer Vision and Pattern Recognition 2009, 2009, pp. 2961–2968.
[19] Y.-T. Chen, K.-T. Tseng, Multiple-angle hand gesture recognition by fusing SVM classifiers, in: Proceeding of the 3rd Annual IEEE Conference on Automation Science and Engineering, Scottsdale, AZ, SUA, 2007, pp. 527–530.
[20] C.-W. Hsu, C. Chunng, A Practical Guide to Support Vector Classification, 2004 http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf