A mutiple-layered gesture recognition system for One-Shot Learning

Shen Wu, Wenping Pan, Feng Jiang, Yang Gao, Debin Zhao
School of Computer Science and Technology
Harbin institute of technology
Harbin, China
wu.shen.eltshan@gmail.com , wenpingpan@gmail.com fjiang@hit.edu.cn, dbzhao@hit.edu.cn

***Abstract*—Gesture recognition is the nature way of human machine interaction and so far enormous work has been done in this area, most of which are based on RGB cameras. The release of Kinect, a depth camera developed by Microsoft, injects new vitality to this well-developed field. In this paper we propose a three layered gesture recognition system. We use and improve the Principle motion, a PCA based method, as the first layer; next we present a particle based descriptor to extract dynamic information of gestures and then propose a specifically designed DTW for classification; finally, a method named edge route context, which is partly inspired by the shape context is brought by us to recognize static hand shapes. Our system performs very well on the one-shot-learning CHALEARN gesture challenge.**

I. INTRODUCTION

Gestures are the unsaid words of human which he expresses in the form of actions [1]. They are considered as the most natural expressive way for communications between human and computers in virtual system [2].Thus much recent research has been focus on gesture recognition with the purpose of interacting with machine, whose extensive applications include gaming, video surveillance, robot control, and interpreting sign language for the deaf [3]. Traditionally, researchers use RGB camera as input sensor, because it's as nature as human. Many promising system have been developed [4]. However, until now, none of them is practical available [5]. This is mainly because that even using RGB camera as input sensor, users are still facing various limitation: appearance based hand detection approaches put serious limit on users' skin color, clothes, background and lighting condition. Besides, facing view-dependence issue, users usually have to be in specific location and orientation [6], which is unacceptable for practical usage. Therefore it is fair to say that based on current feature descriptor and classifier, RGB camera based gesture recognition system has reached a bottleneck.

The release of Kinect, an affordable composite device consisting of an IR projector of a pattern and IR camera [7], gives researchers a new promising option. Through the depth image generated by Kinect, now method of segmentation and tracking, novel features and descriptors all become possible.

In this paper, we propose a multiple layered gesture recognition system. The data acquisition is though Kinect. Next, we divide recognition phrase into three layers: the first layer for fast distinguishing types of gestures; the second layer for identifying gestures though dynamic information; and finally the static hand shapes are recognized at the third layer. This division is motivated by general human recognition pattern: we tend to preferentially distinguish gestures with obvious difference, and then those similar gestures are implicitly classified as dynamic state oriented and static state oriented. That pattern is by and large simulated by our system.

At the first layer of our classifier, we implement and improve the principle motion [9], a PCA based method. It is fast, and more importantly, very robust, which makes it a reliable foundation of our system. Then, at the second layer, we propose a particle based descriptor to extract and identify dynamic information of gestures in each frame. Then, a DTW with adaptive weight was presented to classify time series. At last, a method named edge route context, which is partly inspired by the shape context is brought by us to recognize static hand shapes.

The overall framework that integrates all the above is evaluated on data from ChaLearn Gesture Dataset 2011[8]. The data are gestures recorded as RGB and depth videos with Kinect camera. Our system shows high recognition rate on various gestures. Besides, this system is robust to invariant to different environmental conditions including various backgrounds, clothing, skin colors, lighting and temperature.

We believe this system is novel in following three aspects: firstly, we design the multiple layered system based on human recognition pattern; secondly, we propose the particle-based descriptor to describe dynamic information in gestures; finally, to describe hand shapes, we present the edge route context, which we believe has the potential to be applied to similar tasks.

The remainder of the paper is organized as follows. The implementation of our system is elaborated in section II. Extensive experimental results are reported in Section, and finally, the conclusion is given in section IV.

II. MUYIPLE LAYERED SYSTEM

A. Principle Motion classifier

The first layer of our system is designed to roughly classify the gestures to be recognized into two batches: possible candidates and impossible candidates. Hence we values robust more much than accuracy. Besides, we need the method to be fast and compact so that computing resource can be left to remaining classifiers. Based on considerations above, we use the principle motion method to be the first layer of classifier.

Principle motion is the implementation of a reconstruction approach to gesture recognition based on principal components analysis (PCA). In this method, frames in test videos are projected into the PCA space and reconstructed back using each of the PCA models, one for each gesture in the vocabulary. Next the reconstruction error for each of the models was measured, we then based on the reconstruction errors classify the gestures into two categories: possible candidates and impossible candidates. The implement detail can be found in [9].

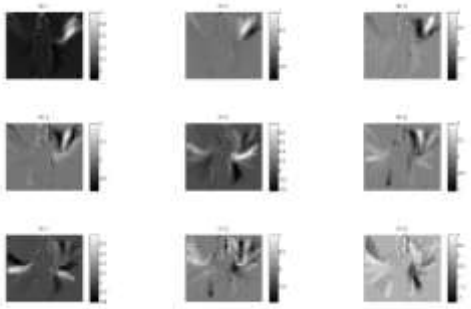

Figure.1 PCA models

In the original algorithm, only the color image or the depth image is used. However, we believe these two types of data can complement each other. Therefore we combine them to be a four-channel image (three for R, G, B and one for depth). Besides, we adjustment the number of PCA models according to the overall length of each gesture. These two improvements are both proved to be effective by our experimental results.

B. Particle based descriptor and DTW with adaptive weight

At the second layer of our classifier, we are facing much more complicated gestures, chiefly the gestures performed by moving arms and hands. Although utilizing the skeleton information provided by Kinect seems can simplify this issue, it's not a valid option because that the accuracy of that joint information cannot be fully trusted and besides when action is too fast or hands occlude each other, the positions of hands are impossible to locate. Hence, we reverse the conventional thought and comprehend this kind of gestures as probability distribution of motion in the three-dimensional space.

Firstly, we embody motion information to be the spatial location of users' arm and hand. Apparently, we can't use all the pixels in one frame due to the consideration of computational complexity; we thus propose a particle description based on motion information. This method is theoretically motivated by two simple observations: 1) at certain extent, we can rarefy the images in a gesture until each frame is represented only by dozens pixels, the gesture can still be recognizable to human. 2) The motion information in each gesture is not uniform distributed; there are always certain particles that are decisive to recognition. There we define these particles as "key particles". Accordingly, our goal in this method is twofold： finding the proper particles to represent the motion information, and distribute weights to these particles to maximize the importance of those key particles. In other words, at first step, we focus on excluding futile information, and then at the second step, we extract and then emphasize the crucial information. Specific algorithms are introduced in the following presentation.

One reason we choose particle based description is that we want to avoid detecting human hands, which would put serious limit on a gesture recognition system. However, it's still necessary to roughly segment the body part that contains the most motion information, the arms. This can be done by simply subtracting the initial position, which is defined as arms hanging down, from current frame. The result we get, which is the difference between current position posture and initial position, is defined as dynamic information.

After deciding specific region that contains the dynamic information, we further implement k-means cluster algorithm. The number of cluster, the k, is decided by the total number of pixels in that region and clustering distance is calculated by the relative spatial distance. We then extract clustering centers to represent particles that describe current dynamic information. And these particles arranged as time series are the features we extract to represent a gesture. The distance between two frames is defined by the minimal distance between the particles that in each frame. This can be calculated by implementing the Hungarian algorithm.

To classify time series composed of frames of particles, Hidden Markov model (HMM) and Dynamic time warping (DTW) both seem to be considerable choices. The former, HMM gains great success in speech recognition field and now are the popular classifier at gesture recognition related areas. The DTW seems to be less popular however it runs relatively fast and has great potential for further improvements and more characteristic due to its compactness. In our work, we propose a DTW with adaptive weight (DTWAW). DTWAW runs much faster than regular DTW, besides, the weight of each frame can be automatically distributed to maximize the differences of similar gestures. The algorithmic detail is described as follows:

1). the traditional DTW process the time series matching and distance calculation at the same time. For complex features, it will cost a lot of time. Therefore, we firstly use simple feature vector combined by change amount of motion, change rate of motion, and motion direction. DTW is implemented on these features to decide the optimal match. By this processing we decrease the time complexity from n^2 to n .

2). usually the dissimilarity between two time series is the sum of the distances between frames in those two sequences. However, this measurement implicitly gives all the frames the same weight while at many cases gestures are distinguished by only few frames. Therefore, at training stage we design the automatic weight distribution algorithm to emphasize the difference frames between gestures. The algorithms are described in brief as follows.

At training state, firstly we implement regular DTW to calculate the distances between current gesture and all the other gestures in the example poor. Then we record the cost of each frame. Next, we find the gesture with the minimum distance, which is the most similar one. Based on the cost of each frame between current gesture and the most similar one, we adjust the weight of each frame to maximize the total distance of these two sequences. If the most similar gesture changed after adjustment, we repeat the second step until it no longer changes.

It is noticed that for one-shot-learning, the DTWAW could result in over fitting issue. However our experiments prove that the issue is outweighed by the advantage it brings.

C. Edge Route Context for classifing hand shapes

The last layer of our recognizer is hand shape recognizer. Hand shape recognizing has always been the core problem of gesture recognition, because in most cases, hand shape expresses much more information than any other body part. Besides, hand shape is the hardest problem in gesture recognition and sign language system because human hand is a complex articulated object with many connected joints and links, which forms the 27 degrees of freedom [10] for the hand.

In this paper we propose a method named Edge Route Context (ERC) which is inspired by the traditional shape context [11][12].

We comprehend the hand shape classification problem as an appearance- based shape recognition process. Currently, the most popular image feature descriptors are SIFT and derived algorithms and HOG and derived algorithms. They are both suitable for textured and structured, rather than object recognition. However, although we mentioned above that hand shapes can be complex, they do not have robust texture and structured appearance. Thus, a descriptor suitable for contour based object, the shape text, is a more suitable choice.

However, shape context has its own issues:

1) It cannot properly deal with the variant of hand.

2) Traditional shape context build histograms on all contour points of shape, which implicitly endow them the same weight. However, at most cases, certain points in a shape are more representative and thus more important than others.

3) It can only contain information of planimetric position, while hand shapes sometimes only distinguishable in 3d space.

To solve these issues, we brought the ERC.

1) We divide 2d plane into bins like shape context.

2) To offset the error brought by rotation, we combine continues adjacent bins. This act is to make sure those bins overlap each other, which makes the descriptor rotation invariant.

3) Rather than calculating number of points in each bin as shape context, we estimate the average depth value.

Traditional shape context treat contour points as randomly distributed, which makes the recognition a problem of finding maximum-weight matching in bipartite graphs. The Hungarian is chose to solve this problem. However, outer edge points of hand shape are apparently sequential. We thus propose a Spatial Path Warping algorithm (SPW), which is the space version of DTW, to match edge points. In SPW, we handle the start/end points of an edge points as the Initial/termination of a time series in DTW. An adaptive window was set to limit maximal length one point can match or be matched to.

We also add the feature of automatic distribution of weight. Through training, different weights are set to edge points. The specific method is described below:

1) Firstly, we extract all points of static information in the key frame, basically the arm part.

2) Secondly, canny edge detector is used to extract the contour points

3) We clockwise calculate and save edge context of each contour point

4) We implement SPW to estimate the distance between current hand shape and all training examples. The one with the minimal distance is selected to be recognition result.

III. EXPERIMENT AND ANALYSIS

In this section, extensive experimental results are presented to evaluate the performance of the proposed system. All the experiments are performed in Matlab 7.12.0 on a Dell OPTIPLEX computer with Intel(R) Core(TM) 2 Duo CPU E8400 processor (3.00GHz), 3.25G memory, and Windows 7 operating system. Due to the large amount of data, we use the parallel processing to accurate the recognition process.

The gesture data we use for experiment is ChaLearn Gesture Dataset (CGD2011).The dataset is recorded for the one-shot-learning CHALEARN. Example is shown in Fig.2.

Fig.2 example gestures

It's the largest dataset of gestures recorded as RGB and depth videos with a Kinect(TM) camera. The data is described as follows

The ChaLearn Gesture Dataset has 50,000 gestures, with image sizes 240 x 320 pixels, at 10 frames per second, recorded by 20 different users, grouped in 500 batches of 100 gestures, each batch including 47 sequences of 1 to 5 gestures drawn from various small gesture vocabularies of 8 to 15 gestures, from over 30 different gestures.

In the experiment, we combine all the three layer classifiers. This recognition is implemented on all 480 batches in CGD2011. The results are shown in Table.1.

Table.1

Batch number	Error rate (%)	Recognize time(s)
devel 01	1.11	359.20
devel 02	24.35	1000.51
devel 03	39.95	561.71
devel 04	6.93	230.41
devel 05	4.77	305.29
devel 06	18.51	342.34
devel 07	8.51	179.61
devel 08	5.71	294.51
devel 09	6.44	301.70
devel 10	16.52	395.01
devel 11	18.93	631.31
devel 12	7.06	234.16
devel 13	12.93	199.96
devel 14	27.98	235.03
devel 15	6.21	219.59
devel 16	19.41	694.89
devel 17	16.32	539.32
devel 18	53.55	160.05
devel 19	27.61	284.71
devel 20	10.61	201.55
21~480	19.624	349.808
average	19.500	355.291

It can be seen that our recognition system have achieved the accuracy rate of eight percent, which is relatively impressive considering the complexity of the experimental data. Besides, the recognition process is very fast, which makes the system to meet the requirement of real-time recognition.

IV. CONCLUSION

In this paper, we introduced a multiple-layered gesture recognition system based on Kinect, which is motived by the general human recognition pattern. Specifically, we use the principle motion, a PCA based method for the first layer of classifier due to its robust and compactness; A particle based descriptor is proposed to describe the dynamic information and a DTW with adaptive weight is presented to recognize the time series. Finally, at the third layer we propose the edge route context descriptor to identify static hand shapes. The performance of the proposed system is evaluated on the ChaLearn Gesture Dataset, and finally the accuracy rate of eighty percent is achieved. The results prove that our work can efficiently recognize complex gestures including body posture, motion based gesture and even hand shape based gesture. We believe that our work, although intuitive at some extent, reflect a novel idea of gesture recognition which has the potential to be further explored.

REFERENCES

[1] Gesture, K.A.: Visible Action as Utterance. Cambridge University Press, UK (2004)

[2] S.Mitra, T.Acharya. Gesture Recognition: A Survey. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 37, NO. 3, MAY 2007.

[3] C. L. Lisetti and D. J. Schiano, "Automatic classification of single facial images," Pragmatics Cogn., vol. 8, pp. 185–235, 2000.

[4] Sylvie C.W. Ong and Surendra Ranganath. Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 6, JUNE 2005

[5] S. Mitra, and T. Acharya, 2007. Gesture Recognition: A Survey. IEEE Transactions on systems, Man and Cybernetics, Part C: Applications and reviews, vol. 37 (3), pp. 311-324, doi: 10.1109/TSMCC.2007.893280.

[6] H. Cooper, B. Holt, and R. Bowden. Sign Language Recognition. Chapter 27. Visual Analysis of Humans..

[7] ChaLearn Gesture Dataset (CGD2011), ChaLearn, California, 2011.

[8] H. Jair Escalante and I. Guyon. "Principal motion: PCA-based reconstruction of motion histograms". Technical Memorandum, June 2012. http://www.causality.inf.ethz.ch/Gesture/principal_motion.pdf

[9] V. I. Pavlovic, R. Sharma, and T. S. Huang, 1997. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 19(7), pp. 677- 695.3

[10] S. Belongie and J. Malik (2000). "Matching with Shape Contexts". IEEE Workshop on Contentbased Access of Image and Video Libraries (CBAIVL-2000).

[11] H. Chui and A. Rangarajan (June 2000). "A new algorithm for non-rigid point matching". CVPR. 2. pp. 44–51.