HUMAN GESTURE RECOGNITION USING ORIENTATION SEGMENTATION FEATURE ON RANDOM ROREST

Weihua Liu[1] *Yangyu Fan*[2] *TaoLei*[2] *ZhongZhang*[3]
[1,2]Northwestern Polytechnical University Xi'an China
[3]The University of Texas at Arlington Texas USA

ABSTRACT

In the field of gesture recognition, one of the major challenges lies in that different user may sign different style of gesture. Traditional exemplar-based methods are vulnerable to gesture scaling and hand location translating. To overcome such disadvantage, we propose an efficient and inexpensive solution for classifying hand gestures by defining an invariant feature and applying it on random forest. One of the prominent characteristics of gesture is the underlying sequence structure, which can be greatly distinguished from other gestures. Hence, direction of gesture sequence segments has been established as simple comparison features for training random forest classifier, and then predicting gestures at sign piece level. The property of this feature determines that our recognition method can invariant to gesture scaling and hand location translating. It is free to act gesture at any angular field of view and not subject to different acting style of signer. The results show that the performance of proposed method outweighs other state-of-art methods for gesture recognition.

Index Terms—Gesture recognition, act style, gesture sequence, invariant feature, random forest

1. INTRODUCTION

To build a reliable and fast system for human gesture recognition is one of important issues in computer vision and pattern recognition community. The user should ideally interact with computer as nature as possible and without spatiotemporal restriction. Hand and arm gesture are often subtle and vulnerable to various change, such as position, orientation, and distance of the people performing gesture with respect to camera. Most of common approaches for gesture classification and recognition can be generally involved in indirect and direct way. The main trend methods for isolated gestures are related to indirect approach, such as Dynamic Programming(DP) methods, e.g., Dynamic Time warping(DTW), Dynamic Spatiotemporal warping(DSTW), Continuous Dynamic Programming, various form of Hidden Markov Models and Conditional Random Fields. An exemplar-based approach like DTW [1,2,3,4], DSTW[5], relies on aligning temporal trajectory between query and model sequence for similarity comparison. In such approaches, trajectory locations are usually adopted as an important match feature for gesture classification, however, they cannot invariant to scale and translation with respect to performing position, orientation and distant to camera. This restriction makes it difficult to accommodate various ways of acting behaves.

Graphical model is another successful framework in modeling temporal sequences. Hidden Markov Models and Conditional Random Fields are two representative methods exist in directed and undirected graphical model domain. A generative HMM [6] model and many extensions has been successfully used for gesture and sign language recognition. S.Bor [7] apply Hidden CRF model spatial dependencies for gesture spotting, which has the ability of CRFs to use long range dependencies, and the ability of HMMs to model latent structure. However, training both models might represent a cumbersome task and a high time-complexity, hence, various approaches were proposed to facilitate the training process.

Random forest [8,9] is a machine learning model which is built on an ensemble of decision trees to training samples of the training sets. It has been widely applied in human action and poses recognition [10] since the advent of Kinect and other high resolution camera. As for hand gesture classification, most of articles investigated in static hand gesture [11] or pose [12], which mean they concern about the hand appearance difference in spatial domain however neglect the entire sign of hand and arm in time domain.

Inspired by recent object recognition work [13, 14] that divided objects into parts and considered the common traits of gesture sequence, we treat the gesture as multi-small segmentation and build decision trees by evaluating those directions of segmentations. Those direction feature avoid using hand locations at each frame, thus can greatly invariant to scale and translation. For proposed feature, it only provide a fraction of gesture orientation signal which is weak and meaningless for the whole gesture recognition, however, in combination in a decision forest they become a strongly classifier and are sufficient to accurately distinguish the long-term gesture. The most saliency feature of gesture sequence is hand location at each frame, which can be easily achieved by utilizing hand detector. Motivated by gesture invariance issue, we design the Orientation Segmentation Feature (OSF) based on hand location. It transfers the feature into temporal domain only, which can also reduce the computational complexity.

In this paper, we proposed a method for quickly and accurately recognizing hand gesture by using Orientation Segmentation Feature on random forest. To establish those features, hand location should be well detected and gesture trajectory should be normalized ahead. For hand detection, we adopt multi-vision-based feature [15] to build an efficient and reliable hand detector. Those are hand color, temporal motion, and motion residue information which can yield promising results than other combination features. Normalized trajectories are created from detected hand locations at each frame, then fitting original trajectory to curve by using interpolation, finally, re-sampling each trajectory to same length to ensure subsequent training and testing. It is important to note that the gesture is a directional video sequences which distinguish from hand pose. For example, we evaluate our system on a vision-based digit gesture task which ask user to sign a digit ranging from 0 to 9 and classify that given digit. Although the complete digital trajectory can be extracted from hand detection, however, we should not simply treat it as a hand writing character but a directional sequence, as depicted in Figure 1(a).

2. GESTURE ACQUISITION

2.1. Hand Detection

Building a reliable and efficient hand detector is an essential step for recognizing signs and gesture. In our system, we adopt multi-vision-based methods which linearly combine three different features to achieve a strong detector. According to pervious work, we directly use hand color features such as hand skin color, temporal motion, and motion residue, which are robust enough for hand detection in clean background.

2.2. Gesture Sequences Normalization

Different gesture has different length. It may also have different length interval in every two adjacent gesture frame as showing the different signers. For the proposed features of random forest which describe in Section 3, we need to normalize each sequence to same number of hand spots. For specifically, gestures trajectory are linearly interpolated between every two adjacent detected hand location, then M locations are equidistant extracted from all gesture spots, so that each sequence has same length, as shown in Figure 1(b)(c). In our method, this sequences normalization is significantly important because we need to treat entire gesture sequence as a combination of multi-piecewise gesture and compare the direction of corresponding piecewise in different gestures for classifying. Normally, the direction of the gesture segment with same sequence order from same gestures will not have too much different; however, it varied greatly from different gestures. Hence, the main benefit of normalization lies in that it helps gesture recognition can invariant to scale and translation.

3. GESTURE RECOGNITION

3.1. Gesture Orientation Segmentation Feature

Consider that the trajectory of gesture can be decomposed to small segments, we employ each gesture segment as Orientation Segmentation Feature (OSF). At kth segment, the feature can be express as

$$\theta_n(v_n) = \begin{cases} \dfrac{\arctan(v_n)\times 180}{\pi} & ; x_{n+k} - x_n > 0 \\ 180 + \dfrac{\arctan(v_n)\times 180}{\pi} & ; y_{n+k} - y_n > 0; x_{n+k} - x_n < 0 \\ -180 + \dfrac{\arctan(v_n)\times 180}{\pi} & ; y_{n+k} - y_n < 0; x_{n+k} - x_n < 0 \end{cases} \tag{1}$$

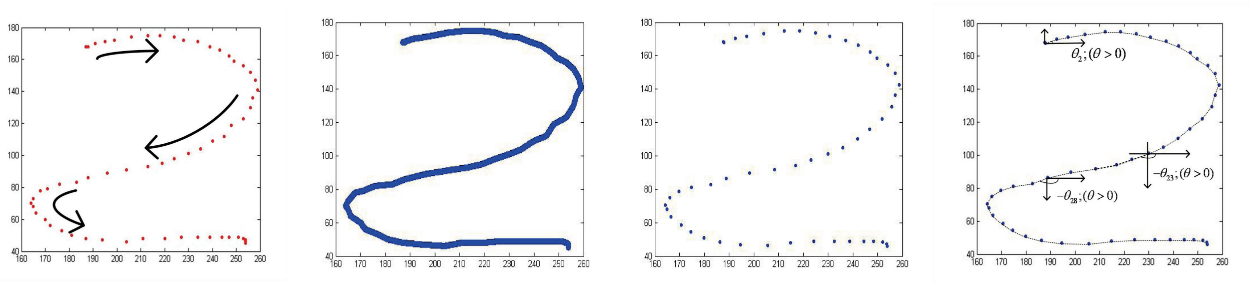
where $v_n = \dfrac{y_{n+k} - y_n}{x_{n+k} - x_n}$ represented the relationship between two vicinity hand location vector $p_n = (x_n, y_n)$ and $p_{n+k} = (x_{n+k}, y_{n+k})$ at spot n , $n+k$. Here we set the total number of gesture spots as M . Feature $\theta_n(v_n)$ essentially finds the angle of vector $(p_n - p_{n+k})$ in image coordinate, the angle range from $[-180^\circ, 180^\circ]$, the schematic of computing segments angle as shown in Figure 1(d). Also, the feature vector of related gesture can be written as:

$$\boldsymbol{\theta}(\boldsymbol{v}) = (\theta_1(v_1), \cdots, \theta_N(v_N)) \tag{2}$$

where $n \in [1, N]$, it also counts for the element number of vector. For computing each next $\theta_n(v_n)$, its subscript n should be updated from previous: $n + k \rightarrow n$. The vector length N of $\boldsymbol{\theta}(\boldsymbol{v})$ determined by the amount of gesture spots M and the frame span value k . The relationship between (N, M, k) can be written as:

$$N = \left\lceil \frac{(M-1)}{k} \right\rceil \tag{3}$$

Individually these features provide only a weak classifier for which part of gesture belongs to, however, in combination of multi-decision forests, they are sufficient to accurately evaluate entire gesture. The design of these features maintains a primary principal that it should make the classification invariant to gesture scale and translation. Hence, consider the characteristic of the gesture trajectory that the same meaning gestures has more or less similar angle on the corresponding Orientation Segmentation Feature. Inversely, the corresponding angle has greatly disparity in different meaning gestures. To this view, we take advantage of the angle of gesture segment as the discriminate feature.



(a) original gesture trajectory with direction arrow (b) gesture trajectory fitting (c) gesture trajectory resampling (d) gesture segment angle

Fig1. Preprocessing on digital gesture "2" model (a) original gesture trajectory with direction arrow indicated (b) gesture trajectory fitting with 1000 interpolated between every two hand location(c) gesture trajectory with 50 spots resampling (d) example of computing gesture segment angle

3.2 Random Decision Forests

The idea of random forest is integrate multi-weak class classifiers to a strong class classifier for classification problem. It has been prove a widely use in many computer vision field. For human gesture classifying, a forest is ensemble by T randomized decision trees, each tree consist by split leaf node which be treated as weak classifier. To classifier one of gesture segments, feature θ_n and two thresholds $\phi=(\sigma,\tau)$ are given to each split node. It starts from the root node and traverse a path according to repeatedly comparison to node threshold that branching the path left or right. This path discrimination following weak classification rule:

$$h(v_\sigma;\tau)=[\theta_\sigma(v_\sigma)\le\tau] \tag{4}$$

where $[.]$ is the indictor function, if inequality satisfied, $h(v_\sigma;\tau)=1$;otherwise it equal to 0 ;

Here, $\sigma\in[1,N]$ denote the order number of elements θ_n in feature vector $\boldsymbol{\theta}(\boldsymbol{v})$; τ denote the threshold parameter which range from $[-180^\circ,180^\circ]$ and with unit of 1° .

At each leaf node of tree t , a learned probability distribution $P_t(c\,|\,\boldsymbol{\theta}(\boldsymbol{v}))$ over gesture labels c is obtained and the final distribution of which class is acquired by averaging the probability of that label for all trees in the forest.

$$P(c\,|\,\boldsymbol{\theta}(\boldsymbol{v}))=\frac{1}{T}\sum_{t=1}^{T}P_t(c\,|\,\boldsymbol{\theta}(\boldsymbol{v})) \tag{5}$$

To train the random forest for classification, each training gesture is preprocessed as the form of angle vector $\boldsymbol{\theta}(\boldsymbol{v})$ as noted above. All these training features are used to grow each tree by implementing the following algorithm:

1. Randomly generate two splitting parameter sets $(\boldsymbol{\sigma},\boldsymbol{\tau})$ separately. Vector $\boldsymbol{\sigma}$ contains a set of order number parameters which can stochastic pick one of angle elements $\theta_\sigma(v_\sigma)$ in vector $\boldsymbol{\theta}(\boldsymbol{v})$. Vector $\boldsymbol{\tau}$ contains 180 threshold parameters for comparing with that chosen angle element.

2. Divide the set of training examples $Q=\{\boldsymbol{\theta}(\boldsymbol{v})\}$ into left and right subsets by using selected $\phi=(\sigma,\tau)$:

$$\begin{aligned}Q_l(\phi)&=\{(\boldsymbol{\theta}(\boldsymbol{v}))\,|\,h(v_\sigma;\tau)=0\}\\Q_r(\phi)&=\{(\boldsymbol{\theta}(\boldsymbol{v}))\,|\,h(v_\sigma;\tau)=1\}\end{aligned} \tag{6}$$

3. Compute the information gain for each split node:

$$G(\phi)=E(Q)-(\frac{|Q_l(\phi)|}{|Q|}E(Q_l(\phi))+\frac{|Q_r(\phi)|}{|Q|}E(Q_r(\phi))) \tag{7}$$

where $E(Q)=-\sum_{c\in C}P_Q(c)\ln P_Q(c)$ represent the Shannon entropy for normalized histogram of gesture labels c for corresponding Q .

4. Select the maximum information gain as the optimal node split parameter.

$$\phi^*=\arg\max_{\phi}G(\phi) \tag{8}$$

5. The computation for left and right subsets $Q_l(\phi^*)$ and $Q_r(\phi^*)$ will not stop until the optimal gain $G(\phi^*)$ fall into threshold or the depth of current node reaches the maximum depth of that tree; otherwise, the algorithm recursively from step1 through 4 for acquiring left and right node and its corresponding optimal gain.

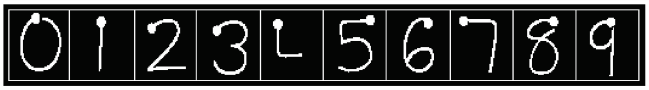


Fig.2. hand sign digital model

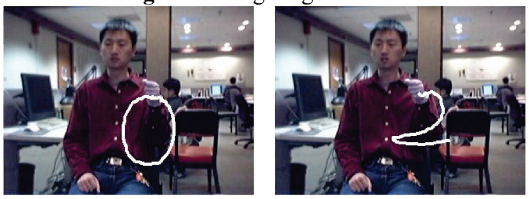


Fig.3. Example model digitals extracted

4 EXPERIMENTS

4.1 Data Description

In this section we evaluate our method in hand-signed digit dataset. In these dataset, 74 video clips with 10 gesture digits each are repeatedly signed from different participants by using a Unibrain Graffiti camera, as shown in Figure 2 and 3. We divide the dataset into training and testing parts separately. In training examples, 440 digit exemplars are stored with 44 per class. In testing examples, 300 digit exemplars are used as queries data with 30 per class. For more exactly validating the robustness of proposed method, we also flip the training and testing datasets as 2-fold cross-validation. It is important to note that neither part of datasets is used for training and testing simultaneously, which mean each part of data is performed independently during training or testing process.

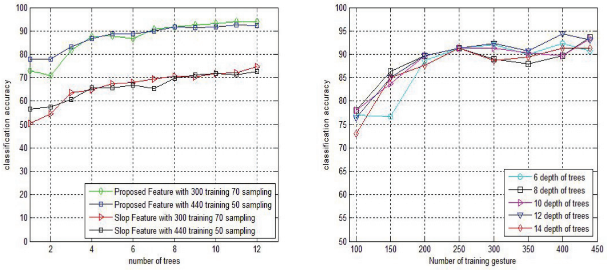
4.2 Classification accuracy

The experiment includes testing isolated gesture classification accuracy on several training parameters and comparing it with related methods. Given the limited data set, we carried out 2-fold cross-validation in which two group of data set were in turn used as training and testing.

In Fig.4(a) we show how classifying accuracy related to training feature and number of trees. As shown below, although the number of trees greatly impact on the capacity of both feature classifiers, the classification accuracy with proposed feature was overall higher than with the slop feature. Also, it approaches to stable after ensemble 10 trees with regard to both training sets and reaches the highest 94.33% accuracy for proposed feature (12 trees, 12 depth, 300 training set, 70 sampling) to compare with 75.4% accuracy for slop feature with same parameters. It is important to note that overmuch trees can result in over fitting problem, hence, we choose 12 trees as the maximum number.

In Fig.4 (b), we adopt 12 ensemble trees to evaluate the algorithm performance by changing number of training gestures and depth of trees in 440 exemplars training set (12 trees, 12 depth, 70 sampling). As shown below, with varying depth of trees, classification accuracy can be greatly impacted by increasing training data. The result shows that deeper layer may not bring higher accuracy rate, however, it also cost extra run-time computation and large memory. Hence, an efficient training strategy has been achieved at depth 12 with relatively high accuracy.

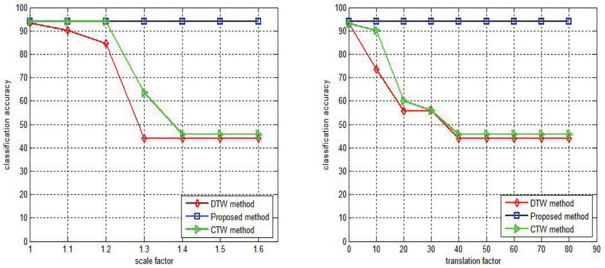
We also compare our approach to other state-of-art methods with parameters same as pervious experiment. Fig.5 (a), (b) represent a considerably recognition precision improvement on gesture scale and translation respectively. With gradually increasing scale and translating factors, the classification accuracy of dynamic time warping (DTW), canonical time warping (CTW) methods are rapidly dropped below 50%, however, the proposed methods still remain high accuracy of 94.33% and without change. It means our methods can invariant to gesture scale and translation. This advantage owns to the Orientation Segmentation Feature can convert gesture location information to gesture direction information, which preserve the main aspect ratio of a gesture and make it possible to recognize similarity gesture even with slightly variation.



(a) Num. training trees and training feature

(b) Num. training gesture and depth of trees

Fig.4. Classification accuracy vs. training parameters (a) Number of training trees and training features on both sets (b) Number of training gesture and depth of trees on 440 exemplars training sets.



(a) Scale factor

(b) Translation factor

Fig.5. Classification accuracy vs. gesture scale and translation (a) scale factor (b) translation factor

5 CONCLUSION

A novel human gesture recognition algorithm based on random forest was proposed in this paper. This algorithm firstly fitting original gesture trajectory to gesture curves, then break these curve into small segments with unified hand spots number in spatiotemporal domain. By assigning orientation parameter to each small part, learning features has been generated for training a random forest model. The experiment results show that apply these multi-directional segmentation feature in random forest can resist to gesture translation and scale and achieve comparative or even better classification rate than other state of the art algorithm. Our future work will focus on expand the proposed work to 3D space based on more advanced depth camera and improve its accuracy to more complex gesture.

6 AKNOWLEGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 61202314 and by the National Science Foundation for Post-doctoral Scientists of China under Grant 2012M521801.

7 REFERENCES

[1] P. Doliotis, A. Stefan, and McMurrough, C., "Comparing gesture recognition accuracy using color and depth information," in *Proc. of the 4th International Conference on PErvasive Technologies Related to Assistive Environments,* (PETRA 2011), Crete, Greece, 2011.

[2] H.J. Wang, A. Stefan, and Moradi, S., "A system for large vocabulary sign search," *in Proc. of the 11th European conference on Trends and Topics in Computer Vision,* Crete, Greece, 2010,pp.342-353.

[3] A. Stefan, V. Athitsos, and J. Alon, "Translation and scale-invariant gesture recognition in complex scenes," in *Proc. of the 1st International Conference on PErvasive Technologies Related to Assistive Environments,* (PETRA 2008) Athens, Greece, 2008.

[4] A. Jonathan, V. Athitsos, and S. Sclaroff, "Accurate and efficient gesture spotting via pruning and subgesture reasoning," in *Proc. of the 2005 International Conference on Computer Vision in Human-Computer Interaction.* (ICCV 2005), Beijing China, 2005, pp.189-198.

[5] A. Jonathan, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *Pattern Analysis and Machine Intelligence.* (PAMI 2009) IEEE Computer Society, vol.31, No.9, pp.1685-1699, Sept, 2009.

[6] F.S. Chen, C.M. Fu, and C.L. Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models," *Image and Vision Computing*, Elsevier, vol.21, No.8, pp.745-758, Aug, 2003.

[7] S. Bor, A. Quattoni, L.P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," *in Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition.*(CVPR 2006), New York, USA, 2006, pp.1521-1527.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning,* Springer, chapter.15, 2009.

[9] D. Demirdjian, C. Varri, "Recognizing events with temporal random forests," *In Proc of the 2009 International Conference on Multimodal Interfaces,* (ICMI 2009), Cambridge, USA, pp. 293-296. 2009.

[10] T. Deselaers, A. Criminisi, J. Winn, and A. Agarwal, "Incorporating On-demand Stereo for Real Time Recognition," *in Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition.*(CVPR 2007), Minneapolis, USA, 2007, pp:1-8.

[11] A. Kuznetsova, Laura, L.T., and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," *in Proc. of the 3rd IEEE Workshop on Consumer Depth Cameras for Computer Vision,* Sydney, Australia, 2013.

[12] X. Zhao, S.J. Guo, "Real-Time Hand Gesture Detection and Recognition by Random Forest," *Communications in Computer and Information Science,* Springer Berlin Heidelberg, Part II, vol.289, pp. 747-755, 2012.

[13] J. Shotton, A. Fitzgibbon, and M. Cook, et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, ACM, vol.56, No. 1, pp.116-124, Jan, 2013.

[14] C.J. Yong, S.W. Nam, "Fast random forest based human pose estimation using a multi-scale and cascade approach," *ETRI Journal,* Electronics and Telecommunications Research Institute, vol.35, No. 6, pp.2013.

[15] Z. Zhang, R. Alonzo, and V. Athitsos, "Experiments with computer vision methods for hand detection," in *Proc. of the 4th International Conference on PErvasive Technologies Related to Assistive Environments.*(PETRA 2011), Crete, Greece, 2011.