Robust Hand Gesture Recognition with Feature Selection and Hierarchical Temporal Self-Similarities

Yuhang Wu, Liangke Zhao, and Hui Ding

Abstract—**There are increasing numbers of hand interaction system based on Microsoft Kinect for its providing the extra depth information which facilitates the hand segmentation. However, most of the recent papers focus on the stable hand recognition whose assumption can hardly be fulfilled in highly interactive HCI environment. We proposed a robust method to recognize both the moving and stable hand. We explored several features in both RGB and depth images based on a novel descriptor, and then built a feature selection model to find the robust features for recognition. Then, we exploited the hierarchical self-similarity matrix (H-SSM) to detect the gesture transformation to further enhancing the system robustness. Our method can be widely used in gesture recognition and the effectiveness has been verified through experiment.**

Index Terms—**Self-similarity matrix, feature selection, HOG feature, kinect**

I. Introduction

Hand plays a crucial rule in human daily communication through moving and gesture transforming. Unfortunately, the huge varieties of non-rigid transformations bring much intractability in recognition.

Recently, along with the off-the-shelf Microsoft Kinect becoming popular, complicated background is no longer a insurmountable barrier and a real-time hand segmentation can be attained based on depth information given by the sensor [1,2,3]. In the depth image, each value of the pixel refers to the distance from the person to Kinect and the hand can be segmented as the nearest region from sensor. Even though Kinect successfully solved the intractable segmentation problem, most of the depth image based approaches are built on the strong assumption that the hand remain unmoved [2], [4], [5]. Unfortunately, when the hand is moving or has small but sudden rotation, the boundary of the hand becomes so unstable that it is hard to recognize the hand as ideal circumstances. In order to build a robust recognition system, we developed a novel descriptor on depth image, several spatial, texture features were extracted and then selected according to our proposed feature selection approach. Later, we divided the gestures into two observation training sets according to the hand's moving speed and learned the decision model respectively. To further improve the robustness of system, we developed a hierarchical self-similarity matrix (H-SSM) based classifier to detect the gesture transformation. As the best of our knowledge, this is the first paper using self-similarity matrix to enhance the classifier's recognition power. Experiments showed that our method efficiently improved the accuracy.

Manuscript received November 9, 2012; revised March 2, 2013. This work was supported in part by the Foundation of Laboratory of Capital Normal University, Beijing, PRC.

Yuhang Wu and Hui Ding are with the Information and Engineering College, Capital Normal University, Beijing, PRC (e-mail: sscommanderh@126.com, dinghui@ie.cnu.edu.cn)

Liangke Zhao is with Capital Normal University, PRC. He is now an internship student with the NLPR, Chinese Academy of Science (e-mail: bensvage1989@yahoo.com.cn)

II. Related Work

In order to attain good recognition performance, an appropriate hand model is needed. The demand of application plays an important role in model selecting [6,7].

Our method closely relates to the appearance-based approaches, which employs an offline training process to establish a corresponding relationship from image features to a finite predetermined status [3]. Most of the real-time recognition methods are based on this model for its stability and efficiency [7]. This kind of method highly relies on the invariance and robustness of the extracted features.

Recently, there is a new trend to combine the merits of RGB and depth features [8], [9]. Simply concatenating multiple feature vectors might be a choice but the result is quite intractable because the ensemble vector including many unbalanced elements from different sources [10], [11]. Hence, a feature selection method is highly needed. We built a model based on the traditional K-Means [12] to evaluate the discriminative capability of different kinds of features and selected the most robust one for gesture recognition.

The self-similarity matrix (SSM), which we used to enhance the recognition accuracy through detecting the action inconsistency, had been invented to detect the periodic motion[13]. Imran Junejo developed the method to recognize human actions under view changes [14, 15]. Our method was inspired by the work of [16] which aim at finding the start and end point of facial expression. In this paper, we embedded the output of H-SSM into SVM to stabilize the classifier's outcome. We believe there will be more researches to broaden the application of SSM in following years.

III. Feature Extraction

Inspired by the work of [1], our descriptor is composed by N concentric circles as Fig. 1. Shows. The geometric center of hand lies in the center of the N rings. After get a rectangular boundary of hand on depth image through

DOI: 10.7763/IJIEE.2013.V3.368

segmentation, each circle's radius (r_i) is determined by (1):

$$r_i = (r_0 - q)(1 - \frac{1}{(i-1)N}) \quad (1)$$

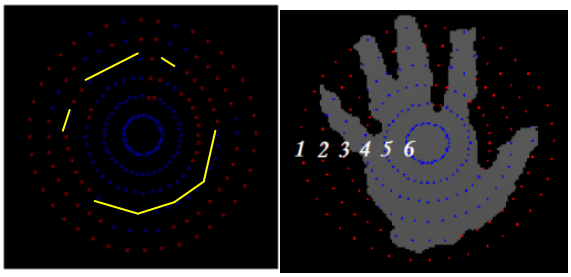


Fig. 1. (a,b). Structure of the novel descriptor

where r_0 represents 1/2 length of the diagonal line in the rectangular boundary. *q* is the distance assigned between the rectangular boundary and the outer ring r_N. We uniformly sampled M positions in each concentric circle. In experiment, we set $_N$=6 and $_M$=40 as Fig. 4 shows. We used this spatial structure as as a platform and extracted the following features on each frame.

1) *Continuous measurement:* this spatial feature is mainly designed to represent the layout of fingers. We encoded the feature as *N*×1 histogram, each bin is corresponding to the number of connected regions in each ring as Fig. 1(b) shows.
2) *Diameter Difference:* this depth feature is designed to detect the relative location of hand's parts in 3D space. Each bin lies in the histogram corresponding to the difference of intensity of any two sample points located on the end of any one diameter in r_i.
3) *Difference among circles:* this depth histogram is also designed to detect the relative position of hand's parts. After we got *N* average intensity value of the all *M* sample points in each ring, we calculated the difference of the value between two adjoined rings. We considered the (*N-1*)×1 histogram might help to discriminate some flat and protrude postures, e.g. palm and fist.
4) *SURF and FAST:* We projected the concentric structure into RGB image and extracted SURF[17] and FAST corner[18] feature on hand. We counted the number of features located between any adjoined circles and formed (*N*-1) ×1 vectors respectively.
5) *HOG:* The histogram of gradient[19] was computed on the whole rectangular region of RGB hand image on the second level of standard image pyramid.

IV. Trainning-Set Dividing

In experiment, we found that the same gesture's appearance could be totally different considering the change of moving speed. Considering the fact that the intra-class samples (belong to the same gesture) with lower similarity are more easily misclassified than those with higher similarity, we manually divide the training set into 2 observation subsets according to the hand's moving speed and built two decision model respectively. So the classifier would no longer be trained to put the instances with (mathematical) semi-similarity into one class according to human prior semantic knowledge. By this means, almost all gestures got two observation states in both sets as Fig. 2 and Fig. 6 shows. Hence, all we need to do in recognition was got the observation states and found its actual corresponding status through mapping.

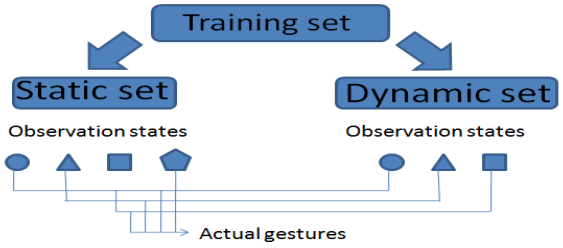


Fig. 2.Training set is divided into two observation states according to moving speed of hand

V. Feature Selection

Feature selection is important when there are multiple feature candidates. Many features are designed to handling a specific task but may prove to be inappropriate in another application. We tried to propose an approach to evaluate the feature's discriminative capability under certain environment.

Our method would be based on a traditional clustering method called K-means. To generalize the problem, let x and y be a feature vector and its label. Based on the clustering assumption [20] that $p(y|x)$ is not tend to change much in a dense area, which means the decision boundary more likely to occurs at the area where $p(x)$ is lower. If the clustering boundary for the region where *x* is located highly corresponding to the decision boundary of *y* provided by supervisor, we may assume they have comparable distribution. Hence, we can infer that the feature vector x has more discriminative capability than the any other *x*' whose distribution incompatible with y.

Specifically, let E_g be the feature's discriminative capability.

$$E_g = 100 \times \frac{D_p}{n^2} \sum_{j=1}^{n} \delta_j \quad (2)$$

n is the number of class in training data. D_P Reflects the diversity of clusters after K-Means. If we using *Pj* to notes the

label of the largest cluster among the clusters composed by the elements belong to class *j*, D_P would be the number of different elements in set *P*. δ_j is the weight can be updated according to (3):

$$\delta_j = \frac{1}{S_j} \sum_{k=1}^{m} \varepsilon^{1-k} S_k \quad (3)$$

where *m* is the number of clusters after K-means composed by the instances in class *j*. S_k refers to the number of elements in cluster *k*. ε is the decay weight. In experiment, we set $\varepsilon = e$.

VI. H-Ssm and Motion Inconsistancy Detection

A. Definition of Hierarchical Self-similarity Matrix

For a temporal sequence $I = I_1, I_2, \cdots, I_T$ where *T* is the number of frame, SSM is a symmetric matrix can be

written as equation (4):

$$[d_{i,i}]_{i,j=1,2,\cdots,T} = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1T} \\ d_{21} & 0 & d_{23} & \cdots & d_{2T} \\ \vdots & \vdots & \vdots & & \vdots \\ d_{T1} & d_{T2} & d_{T3} & \cdots & 0 \end{bmatrix} \quad (4)$$

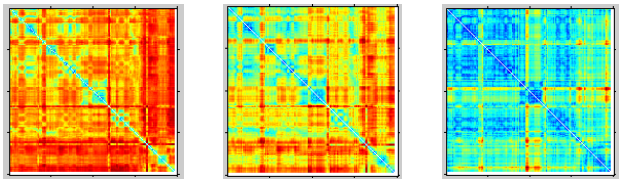


Fig. 3. The cool hue in SSM represents the area which has comparatively higher similar value than the warm hue.

where $d_{i,j}$ represents the distance between the same kind feature extracted in frames I_i and I_j respectively, when *i=j,* the distance is 0. We recommend readers to look up [14] to get more information.

We used the Euclidean distance in our experiment and built 3-layer hierarchical SSM on the 3-level Gaussian pyramid of the input image. From bottom to top layer, the SSM is denoted by M_i, *i* could be 1,2 and 3 here.

B. Definition of Sub-SSM

Due to the complexity of computing SSM is linear increasing with time elapse, we only calculated on the last N_{M_i} frames to build sub-SSM. We need to highlight that N_{M_i} is also the length of sub-SSM of level M_i. N_{M_i} is determined by the sensitivity of SSM in different level. In our experiment we select N_1=6, N_2 =8, N_3=10. Considering the sub-SSM is updated from right to left, the frame which is far from the current one (right-bottom block) contribute less than the frames just past, so we rendered the elements on the right column of M_i with decreasing weights from bottom to up when updating sub-SSMs.

C. Boundary and Motion Change Detection

In Fig. 3, we can see many colorful square-like pattern lying around the diagonal of each matrix. Each of the square represents a sequence of frames with similar features. And the cross patterns lying on the boundaries of squares represent the motion inconsistencies corresponding to the gesture change. Many works have concentrated on the pattern of the squares [14], [15], however, as the best of our knowledge, we are the first explored the cross pattern (the boundaries of squares).

When a real gesture transformation detected, we found all levels of the sub-SSM appeared to have clear boundaries covered larger than 2 columns with gradual color transition. However, when a noise took place, the boundaries would usually be sharpened. So when we detected a pattern with sharp contrast, we neutralized the error through replacing the new candidate column by the last past one in sub-SSMs.

In order to describe the change of motion, we first divided every sub-SSM into two parts as Fig. 4 shows. The area *A* is a $(N_{M_i} - m) * m$ rectangular region (in recognition we assigned *m*=3). Then we calculated the average value in *A* and denoted as A_i^*. Correspondingly, we calculated the average value in all the left blocks in SSM and represented as B_i^*. We then got:

$$C_i = (\frac{A_i^*}{B_i^*}) \quad (5)$$

C_I Had the highest responding when there was a boundary with size *m* in last *m* frames. Let W_i denotes the weight we assigned to the different levels of SSM. We got C_f which described the overall performance of sub-SSMs:

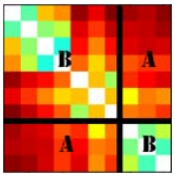


Fig. 4. The sub-SSM is divided into two rectangular regions: A and B.

$$. C_f = \prod C_i * W_i \quad (6)$$

The relationship between C_i and C_f is shown in Fig. 5.

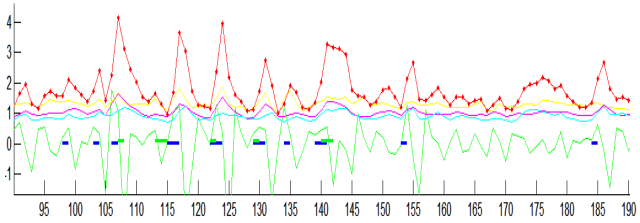


Fig. 5. C_1, C_2, C_3 are shown as yellow, pink, blue folded line respectively . Red line represents the C_f and the green folded line represents the *P*. The green horizontal line represents the action change points tagged by people. And the blue horizontal line shows the frame which met the condition: $\varepsilon_1 < D < \varepsilon_2$ in Algorithm 1. We adopted a high recall strategy here to make sure every gesture change could be detected.

At last, we used Δ to measure the stability of the current gesture. Let k_f, k_{f-1} denote the slope of C_f in continuous 2 frames and we got D=k_f-k_{f-1} and Δ can be calculated through the following algorithm 1.

Algorithm 1:

Input: D, Thresholds: $\varepsilon 1, \varepsilon 2, \varepsilon 3, \ell$
Output: Δ (change_weight), *action*
Algorithm:
If $\varepsilon_1 < D < \varepsilon_2$
  $\Delta = exp(-\ell * (D - \varepsilon 1));$
  action=1;
elseif $D < \varepsilon 1$
  $\Delta = (1 + exp(decay)) * \varepsilon 3;$
  action=0;
    if $D < 0$
      $\Delta = exp(-\ell * D)$
    end
else $\Delta = (1 + exp(decay)) * \varepsilon 3;$
end

ℓ would be gained through configuring an initiate stability rate S_0 and threshold $\varepsilon 1$: $\ell = \frac{S_0}{\varepsilon 1}$; The decay would be reset to 2 when the classifier's output change to another class mentioned below.

D. A Modified Robust Classifier

In traditional classifier e.g SVM[21], the feature vector is treated as isolated input so the sudden blur or the intermediate status of gesture transforming is always be misclassified into finite class without considering the continuity of gesture in temporal sequence. We using the following algorithm to restricted the classifier's output to its former label when there were some small interruptions detected, and encouraged the classifier to change its output when there was a real gesture change detected. In ideal circumstances, new decision would be made correspondingly to the gesture change.

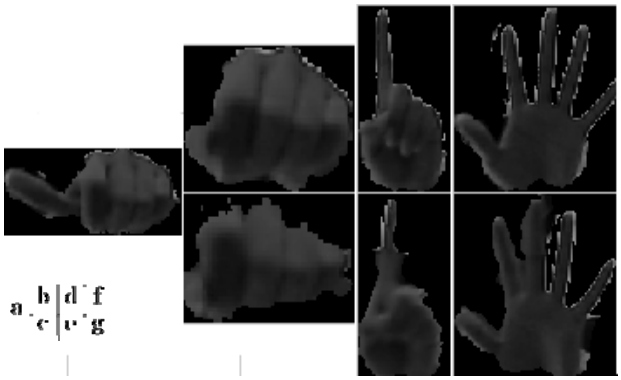


Fig. 6. The seven observation states of the four predetermined gestures.

In following algorithm 2, $P(c_j|x_m)$ denotes the probability of classifying the current sample x_m to class c_j by SVM.

Algorithm 2:

Input: The probability decision value: P_j; Stability: S_j, last predicted class: C_{m-1}; $Decay, action$

Output: Predicted class: C_m, $Decay$

Algorithm:

1.if $action$==0

$S_j = exp(\Delta);$

$S_{j'} = exp(\Delta);$

Else

$S_j = \Delta;$

$S_{j'} = \frac{1}{\Delta};$

2. $P(C_{m-1}|x_m) = P(C_{m-1}|x_m) * S_j;$

3. $P(C_n|x_m) = P(C_n|x_m) * S_{j'}$; n belongs to the class in C_m apart from C_{m-1};

4. $C_m = max(P(C_i|x_m))$

5. if C_{m-1} not equal to C_m, $Decay$=2;

6. $Decay$=$Decay$-1;

VII. Experiments and Results

A. The Dataset and Toolbox

There are 4 predetermined gestures and 7 observation states as Fig. 6 shows. 4,000 frames performed by 6 persons have been collected. 2,000 frames performed by one person as training set; another 2000 frames captured from 5 other volunteers as testing set. People were asked to stand 1-2 meters in front of KINECT and feel free to moving their hand and changing gestures. We used the OpenNI platform[22] and static threshold on depth image to segment the hand from background. We adopted Libsvm [23] for gesture classification through *one-against-one* strategy. The H-SSM embeded classifier's recognition speed attained 175fps in MATLAB 2012a on Intel i5 2500 platform .

B. Feature Selection

For each feature, we set the number of cluster as 2,4,6,8 10 in K-Means and each clustering has been done 1,000 times to neutralize the unbalanced random initiation. We used the maximum E_g of each feature to show the feature's discriminative capability. In Fig. 9, we can see the *HOG feature* outperformed others. And the *continue management* ranked the second. It is important to notice that both of the features reflect the spatial layout of the hand. The famous

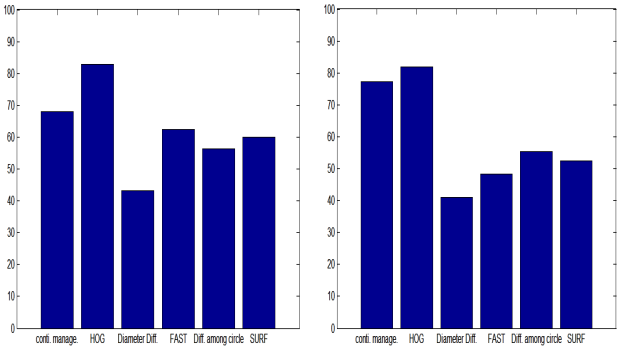


Fig. 7. The E_g calculated on the static and dynamic training sets.

SURF and FAST feature got relatively low scores which might due to the lack of key points in low resolution images. Other features ranked the last may due to the strong blur when the hand was not toward the Kinect directly

C. Gesture Estimation

We used the training set without dividing to learn the decision model of SVM as the baseline. Then we learned a decision model based on our special training strategy mentioned in Section IV. Finally we added the HOG feature based H-SSM in on-line recognition. We can see that the average accuracy has been improved by our method through Fig. 8.

In experiment, our method reflects strong robustness when handling the frame with noise and sudden change as Fig. 9 shows. However, few symbols that run against this tendency have also been detected. We found that when there was some smooth transformation of gesture, our SSM based method might fail to detect them for adding extra inertia to the original gesture and hereby undermine the sensitivity of the classifier. Hence, further explorations are highly needed.

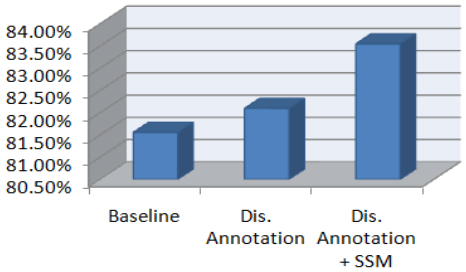


Fig. 8. Average accuracy of recognition

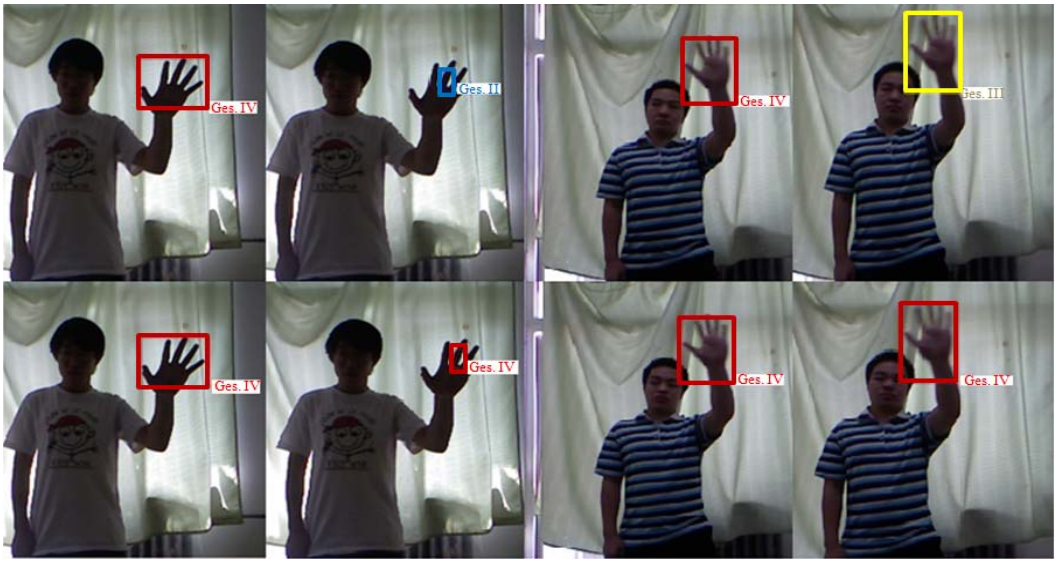


Fig. 9. The first row corresponding with the original output of classifier when recognize continuous two frames with noise, which reflects it is susceptible to the sudden change in temporal sequence The second row is the output of our modified classifier which reflects its robustness.

VIII. Conclusion

A new method for improving the robustness of hand gesture recognition is presented. Our method is simple and practical to be applied in real-time interaction system. A model is built to find the most robust feature for recognition. A modified robust classifier has been employed by integrating SVM with the motion consistency information through hierarchical self-similarity matrix. The proposed framework has been proved efficient through a real-world HCI-oriented testing set. We believe that our approach can be applied to modifying most of the traditional gesture recognition approach.

Acknowledgment

The author would like to gratitude Dr.Xiuzhuang Zhou, Dr.Xiangyang Huang and Dr.Guang Jiang for their valuable suggestions.

References

[1] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos, "Comparing gesture recognition accuracy using color and depth information," in *Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments* - PETRA, 2011.

[2] J. L. Raheja, A. Chaudhary, and K. Singal, "Tracking of Fingertips and Centers of Palm Using KINECT" *Third International Conference on Computational Intelligence, Modelling & Simulation*, pp. 248–252, 2011.

[3] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Tracking the articulated motion of two strongly interacting hands," in *CVPR*, pp. 1862–1869, 2012.

[4] Z. Ren, "Robust Hand Gesture Recognition Based on Finger- Earth Mover ' s Distance with a Commodity Depth Camera," in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1–4, 2011.

[5] D. L. Gorce, D. J. M. Fleet, and N. Paragios, "Model-Based 3D Hand Pose Estimation from Monocular Video," *IEEE transactions on pattern analysis and machine intelligence*, pp. 1–14, 2011.

[6] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review. Computer Vision and Image Understanding," pp. 52–73, 2007.

[7] M. Hasan and P. Mishra, "Hand Gesture Modeling and Recognition using Geometric Features," *A Review. Canadian Journal on Image Processing and Computer Vision*, vol. 3, no. 1, March, 2012.

[8] A. Bleiweiss and M. Werman , "Fusing time-of-flight depth and color for real-time segmentation and tracking," *Dynamic 3D Imaging* , pp. 1–13, 2009.

[9] M. V. D. Bergh, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *Applications of Computer*, pp. 66–72, 2011.

[10] Y. Fu, L. Cao, G. Guo, and T. S. Huang, "Multiple feature fusion by subspace learning," in *Proceedings of the 2008 international conference on Content-based image and video retrieval* - CIVR , vol. 127, 2008.

[11] J. Dargham, A. Chekima, and E. Moung, "Fusing Facial Features for Face Recognition," *Advances in Intelligent and Soft Computing*, vol. 151/2012, pp. 565-572, D, 2012.

[12] J. B. M. Queen, "Some methods for classification and analysis of multivariate observations," *5th Berkeley Symp. Math. Stat. Proba,* 1967.

[13] R.. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis, and appli- cations," *PAMI* 22, pp. 781–796, 2000.

[14] I. N. Junejo, E. Dexter, I. Laptev, and P. Patrick, "View-Independent Action Recognition from Temporal Self-Similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2011.

[15] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "Cross-view action recognition from temporal self-similarities," *ECCV*, 2008.

[16] F. D. L. Torre, T. Simon, Z. Ambadar, and J. F. Cohn, "Fast-FACS: A Computer-Assisted System to Increase Speed and Reliability of Manual FACS Coding," *Affective Computing and Intelligent Interaction*, vol. 6974/2011, pp. 57-66, 2011.

[17] H. Bay, "Tinne Tuytelaars and Luc Van Gool, SURF: Speeded Up Robust Features," *ECCV*, 2006.

[18] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection, " *ECCV*, 2006.

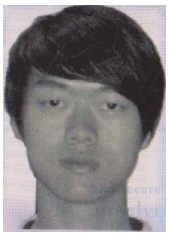
[19] Dalal and B. Triggs, "Histograms of oriented gradients for human detection" in *Computer Vision and Pattern Recognition*, 2005.

[20] B. S. O. Chapelle and A. Zien., "Semi-supervised learning," *The MIT Press*, 2006.

[21] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 273-297,1995.

[22] OpenNI. [Online]. Available: http://www.openni.org/

[23] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACMTransactions on Intelligent Systems and Technology*, 2011.



Yuhang Wu was borned in Beijing, China, 1991. He is now an senoir undergraduate student in Intelligent Science and Technology Department, College of information and engineering, Capital Normal University. He is expecting to receive a B.E degree in 2013. His research interest focuses on the computer vision and artificial intelligence.

He has intership experience in ICT, Chinese Academy of Science and has conducted several experiments covered facial expression recognition, scence image recognition and natural language processing.

Mr. Wu is a student member of IEEE, ACM, AAAI and the student leader in his department.

Liangke Zhao was borned in Beijing, China, 1989. He graduated from College of Information and Engineering, Capital Normal University. He received a B.E degree in computer science in 2012,. Currently, he is an intermship in Institution of Automation, Chinese Academy of Sciences with the concentration in computer vision.

Hui Ding received the PhD degrees in navigation, guidance and control from school of information science & technique of Beijing Institute of Technology, Beijing, China, in 2006. Then she has been a postdoctoral fellow at department of electrical engineering, Tsinghua University in 2006-2008. She is currently an associate professor in the Information Engineering College at the Capital Normal University, Beijing, China. Her research interests include computer vision, image analysis and feature detection, video surveillance, visual tracking and pattern recognition