MVA2013 IAPR International Conference on Machine Vision Applications, May 20-23, 2013, Kyoto, JAPAN

Vision-based Raising Hand Detection in Classroom

Cheng-Chieh Chiang
Department of Information Technology
Takming University of Science & Technology
No. 56, Sec. 1, Huan-Shan Rd.
Taipei 114, Taiwan, R.O.C.
kevin@csie.ntnu.edu.tw

Cheng-Chuan Tsai, Greg C. Lee
Dept. of Computer Science and Info. Eng.,
National Taiwan Normal University
No. 88, Sec. 4, Ting-Chou Rd.
Taipei 114, Taiwan, R.O.C.
leeg@csie.ntnu.edu.tw

Abstract

Raising hand is one of the most important types of interaction between students and lecturers in classroom. When an automatic system can be installed in classroom to figure out which students raise their hands, it is possible to design more advanced applications for education goals. This paper proposes a system that employs computer vision technologies to automatically detect the student action of raising hand. We first design a foreground extraction method to segment student bodies in consecutive video frames. Next, a shape-like appearance signature that represents human gestures is designed based on the scale-invariant feature transform (SIFT) descriptor. A gesture classifier for raising hands is also designed using the support vector machine (SVM) approach. This paper designs several experiments to demonstrate the performance of our proposed system in a real classroom.

1. Introduction

Machine vision has been widely applied to many goals of applications in order to improve human life in the modern society. During the past years, our team, VIPLab in National Taiwan Normal University, has paid more attention to building an advanced environment for future classrooms to help students improve the learning performance in school/university. This paper focuses on automatically detecting raising hands in a real classroom.

Raising hand is one of the most fundamental actions that can be considered an interaction between students and lecturers in classroom. When an automatic system can be installed in classroom to figure out which students raise their hands, it is possible to design more advanced applications to analyze interactions between students and teachers.

The raising hand detection in classroom is not a trivial task due to the following reasons. First, there are often a lot of students stayed in a classroom. The designed detection system has to deal multiple persons with action classification. Second, many other objects, besides students, may also appear in a classroom such as book, cup, laptop, and bag. All of foreground subjects including both students and other objects may change their positions, and hence we cannot segment them by using a simple background subtraction method. Next, different students should attend different classes. Thus, we cannot expect that students appearing in class are fixed. Moreover, various environment conditions of class such as lighting, seat position and subject occlusion also make the system design more difficult.

Our intuitive idea of the student gesture detection is to employ Kinect [18] which is developed by Microsoft to sense subject moments without any touch. However Kinect is not appropriate for a real classroom because its effective sensing distance is limited. In past, many researchers have published their works related to detection and recognition for human gestures. Hence, we employed several computer vision technologies to treat the raising hand detection in classroom.

S. Maitra and T. Acharya classified issues of gesture recognition into three categories: (i) hand and arm gestures, (ii) head and face gestures, and (iii) body gestures [10]. R. Poppe also published a literature survey of the human action recognition using vision-based approaches [12]. S. S. Rautaray and A. Agrawal provided a comprehensive review of human and computer interaction focusing on using vision-based hand gesture recognition [11]. Our work to detect raising hands mainly covers the recognition of both the arm and the body parts. A number of researchers have paid more attention to these issues, such as [1][3][7] on body and arm gestures, and [2][4][5] on hand gestures.

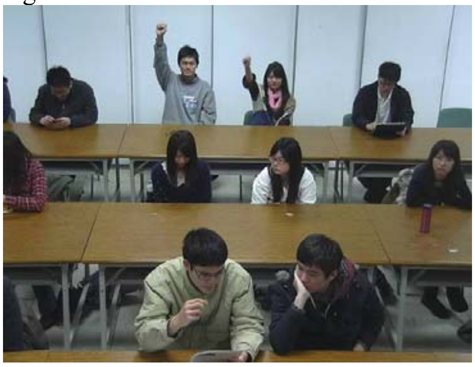


Figure 1. A snapshot of classroom in our experiment.

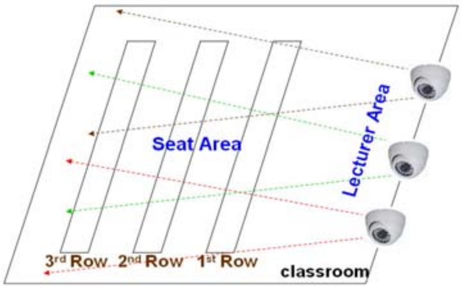


Figure 2. Our classroom contains three rows of seats, covered by three cameras.

Figure 1 illustrates the real scene of our experiment in a university classroom, also drawing the camera setup in

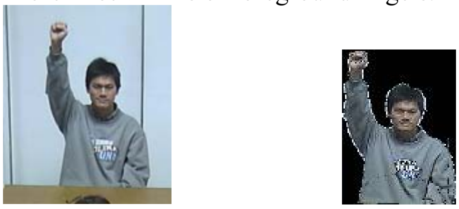
Figure 2. This classroom contains three rows of seat, and three cameras are installed under the ceiling to cover the whole seat area. Students can arbitrarily select their own preferred seats. Cameras are installed and fixed near the ceiling, and then the camera views are also fixed. Three gestures of student raising hands are detected in this work: right hand, left hand, and normal, where normal means that this student may appear any kind of gestures except raising left or right hand.

Given consecutive video frames from a camera, this work first segments the foreground areas to locate the student positions. Then, a set of shape-like appearance signatures that is based on the scale-invariant feature transform (SIFT) descriptor [8][9] is extracted from each of located student bodies. We therefore employ the support vector machine (SVM) [15] to build a classifier that can determine what the student gestures appear in video. We also perform several experiments in a real classroom to show the performance of our proposed system.

The remainder of this paper is organized as the follows. Section 2 to 4 present the methods used in this work, containing foreground subject detection, feature extraction, and classification, respectively. Moreover, several experiments have been performed to demonstrate the performance of our system in Section 5. Finally, we draw conclusions and future works in the last section.

2. Foreground detection

To detect foreground targets in video sequences is the first task to automatically recognize student actions in our work. In general, the definition of the foreground target contains all objects that are not included in the background. For example, the foreground targets may contain human, bag, book, and cup since they are not in the classroom in original. Sometime these foreground targets can change their positions, but it is also possible that they have less motions in class. The most trivial approach is to construct a background modeling for the camera view. In order to detect student actions as soon as possible, a background modeling needs to be sensitive to foreground motion. However a sensitive background modeling could also generate a lot of noises that will make more false alarms of foreground targets.



(a). the original frame (b). the foreground segment

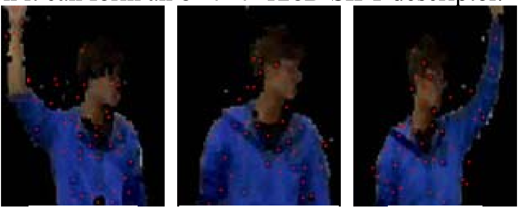
Figure 3. Student body detection

This paper employs two well-known approaches to deal with the foreground extraction mentioned above: GMM [15] and temporal differencing [13]. The GMM method is widely used for constructing a dynamic background modeling in a video sequence. While a foreground subject does not have any motion in a time period, this subject will be involved in the background modeling gradually if using the GMM approach. The temporal differencing approach can be very sensitive to detect tiny motions of moving parts in a video sequence. Our system incorporates with these two methods to effectively detect student regions of raising hands in video frames. Given extracted foreground regions in video frames, we employ the adaboosting approach [16] of face detection and the skin detection approach [14] to locate student bodies in classroom. Note that only the upper-body areas are captured due to students may be occluded by tables by part. Figure 3 presents an example of an original frame and the corresponding student body.

3. Shape-like Feature Based on SIFT Descriptor

When student bodies have been localized by the methods mentioned in the above, we define a shape-like feature to describe the appearances of the student gestures based on the SIFT descriptor [8][9]. SIFT descriptor is first proposed by D. Loew in 1999 to extract distinctive invariant features from images. Many researches have shown that SIFT descriptor can be used to perform reliable matching between different views of an object or scene. The details of extracting SIFT descriptor is referred to Dr. D. Loew's publications in [8][9]. Here we only present a brief of the extraction for short.

The procedure of extracting SIFT descriptors contains the two stages: detector for keypoint localization and descriptor for keypoint description. First, in order to determine location candidates of keypoints in a scale space, the difference-of-Gaussian function is performed to detect the scale-space extrema by computing the difference of two nearby scales separated. Then, two types of keypoint candidates are rejected: one with low contrast and the other localized along an edge. When robust keypoints have been localized, the second task is to extract their descriptors in image. In order to achieve orientation invariant, a consistent orientation based on local properties of image is assigned and an orientation histogram of gradient is built. The orientation histogram contains 8 directions and accumulates over a 4x4 subregions, and then it can form an 8×4×4=128D SIFT descriptor.



right hand normal gesture left hand

Figure 4. Sampling of feature points by SIFT descriptors for raising hands.

The SIFT descriptor can represent significant contents in an object illustrated as Figure 4, but our goal is to verify the shape-like information of raising hands. Hence we extract a fixed number K of SIFT descriptors, K=100 in our implementation, from a segmented human body with the most strong magnitudes to sample the region of the gesture. Assume these K descriptors locate at (w_1, h_1), ..., (w_K, h_K) with the row-major order in the image coordinate and with magnitude m_1, ..., m_K, respectively. The corresponding feature vector can be defined as (w_1, h_1, m_1, ..., w_K, h_K, m_K) with 3×K dimensions.

4. Classification

When collecting enough training data for student gestures and extracting SIFT descriptors of foreground segments, a classifier is then learned to automatically determine whether a student raises hands or not. In this work, the SVM classifier [15] is adopted to perform our classification task. SVM is a supervised learning model and is well-known to achieve a good performance in classification. In implementation, we adopted LIBSVM library [17] with a radial basis kernel. LIBSVM library was developed by Machine Learning and Data Mining Group, National Taiwan University, and can support the main functions of SVM classifiers.

Assume a set of training data $\{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, y_n)\}$ given where $\mathbf{x}_i$ means the feature vector of training data, and $y_i \in \{1, -1\}$ is the labeled of $\mathbf{x}_i$. In order to define a linear classifier $y_i = \mathbf{w}\mathbf{x}_i + b$, the SVM method solves:

$$\min_{\mathbf{w},b}\{\frac{1}{2}\|\mathbf{w}\|^2\} \tag{1}$$

subject to $y_i(\mathbf{w}'\mathbf{x}_i + b) \geq \mathbf{1}$

Since it is in general a non-linear classification problem, a nonlinear function ϕ is necessary to map data to a higher dimensional feature space due to Cover's theorem [6], which guarantees that the mapped data are linearly separable in the transformed feature space. This has been proven a well-known quadratic optimization problem and can be solved by a dual form

$$\max_{\lambda}\left(\sum_i \lambda_i - \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j \phi(\mathbf{x}_i)'\phi(\mathbf{x}_j)\right) \tag{2}$$

subject to $\sum_i \lambda_i y_i = 0$ and $\boldsymbol{\lambda} \geq 0$

where $\boldsymbol{\lambda}$ are Lagrange multipliers corresponding to the constraints of equation (1). The nonlinear mapping ϕ in equation (2) forms an inner product, hence it is possible to define a kernel function,

$$K(\mathbf{x}_i, \mathbf{x}_i) =< \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >= \phi(\mathbf{x}_i)'\phi(\mathbf{x}_j), \tag{3}$$

for solving this equation without having to compute the mapping ϕ explicitly. Finally, we can have the solution of the dual problem,

$$\mathbf{w} = \sum_i y_i \lambda_i \phi(\mathbf{x}_i), \tag{4}$$

and the classifier can be defined as

$$f(x) = sign\left\langle \sum_i y_i \lambda_i K(\mathbf{x}_i, \mathbf{x}) + b \right\rangle \tag{5}$$

where b can be easily computed from $\boldsymbol{\lambda}$.

5. Experiments

5.1. Data set

The proposed system has been installed in a classroom of National Taiwan Normal University to collect our training and test videos. Three cameras are involved in our system to cover the whole classroom, and these camera views are fixed for simplification in the system implementation. Our experimental data set contains two subsets, which are captured with different time and with different students. These two sets are called D_1 and D_2 that are partitioned into two parts with the equal size roughly: one for training and the other for test. Thus, we can denote data set $D_i = M_i \cup V_i$ for i=1 and 2, where M_i means the training part and V_i the test part, respectively. The two data sets can be summarized as Table 1.

Table 1. The sizes of our training and test data set, where the digits in this table means the numbers of instances. Two sets D_1 and D_2 are captured from different classes, and they are divided into two sets with rough equal-sizes.

Data		Left Hand	Normal	Right Hand	Total No.
D_1	M_1	738	2355	757	3850
	V_1	734	2363	789	3886
D_2	M_2	1380	6939	1759	10078
	V_2	1370	6961	1727	10058

5.2. Results

Table 2. Classification rates that the training and test data are from the same source, with the whole average 0.892.

	Left	Normal	Right
Left hand	**0.954**	0.025	0.021
Normal	0.045	**0.878**	0.077
Right hand	0.019	0.089	**0.891**

Table 3. Classification rates that the training and test data are from different sources: M_1 for training and V_2 for test. The whole average rate is 0.757.

	Left	Normal	Right
Left hand	**0.768**	0.174	0.058
Normal	0.13	**0.782**	0.088
Right hand	0.138	0.218	**0.644**

Table 4. Classification rates that the training and test data are from different sources: M_2 for training and V_1 for test. The whole average rate is 0.62

	Left	Normal	Right
Left hand	**0.837**	0.035	0.128
Normal	0.168	**0.47**	0.361
Right hand	0.034	0.09	**0.876**

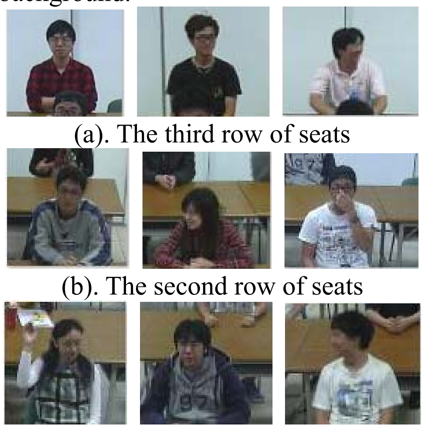
Table 5. The detailed classification rates of the three seat rows corresponding to the experiments in Table 3 and 4.

	Training: M1 Test: V2 corr. to Table 3	Training: M2 Test: V1 corr. to Table 4
3rd row	0.93	0.918
2nd row	0.663	0.521
1st row	0.677	0.416

Our experiments are mainly divided into two classes. First, the training and test data of the classification are performed on the same data source, i.e., V_1 based on M_1 and V_2 based on M_2. Since the training and test sets are captured in the same class, lighting and other conditions can be assumed consistent. Table 2 shows the confusion matrix of the classification rates of the three student gestures. Note that the result of the normal gesture is smaller than that of the other two gestures due to the normal gesture may contain different kinds of sitting postures.

The results shown in Table 2 seem not bad, but, unfortunately, the setup of the experiment in Table 2 is not reasonable. In practice, it is impossible to classify student gestures using the training data that are captured from the same class of the test data. Hence, we designed a further experiment that apply different data sources to training and test data. Two classifications are performed based on two pairs of training and test set: (M_1, V_2) and (M_2, V_1), showing their results in Table 3 and 4, respectively. Note that the whole average rates of these two tables are 0.757 and 0.62, respectively.

In order to realize what factors affect the performance of the two experiments, we individually analyze the classification rate for each row of seats in classroom, shown in Table 5. Note that our approach can achieve high rates in the third row but not very successful in the other two rows. Figure 5 illustrates several cases of our test data that are located at different rows of seats in classroom. Human bodies in the first two rows may be mixed with other foreground parts. It is more difficult to achieve a correct classification when the foreground bodies are not correct. However, the tests in the third row are very successful in Table 5; that means our approach basically can work well in a real environment (training and test data are from different data sources, and the experiment was performed in a real classroom). Hence, the most important issue to improve this work is to design a good method to well segment the foreground subjects in the cluttered background.



(a). The third row of seats

(b). The second row of seats

(c). The first row of seats

Figure 5. Examples of test gestures at different rows of seat. Human bodies in the first and the second rows may be strongly affected by people in the other rows.

6. Conclusion

This paper aims to design a vision-based system to automatically detect the student gesture of raising hands in classroom. We introduce the details of our approaches and show the experimental results to discuss the efficacy of our proposed system. This paper presents the first results of our effort for detecting raising hands in a real classroom. Even though the shown performances are not excellent, the experiments indicate that our approaches can work well if student bodies can be well segmented. Hence, our most important task in the future is to design an advanced approach to carefully tracking students and capturing their body areas in video frames. Moreover, we are implementing an integrated system for the future smart classroom by involving our proposed approach.

Acknowledgement

This paper was supported by National Science Council, Taiwan, under Grant No. 100-2511-S-003-020-MY2.

Reference

[1] J. Alon, V. Athitsos, Quan Yuan, and S. Sclaroff: “A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 9, pp. 1685-1699, Sep. 2009.

[2] V. Athitsos, H.-J. Wang, and A. Stefan: “A Database-based Framework for Gesture Recognition,” *Personal and Ubiquitous Computing*, Vol. 14, No. 6, pp. 511-526, Sep. 2010.

[3] M. Bayazit, A. Couture-Beil, and G. Mori: “Real-time Motion-based Gesture Recognition using the GPU,” in Proceedings of MVA, Japan, 2009.

[4] W.-Y. Chang, C.-S. Chen, and Y.-D. Jian: “Visual Tracking in High-Dimensional State Space by Appearance-Guided Particle Filtering,” *IEEE Trans. on Image Processing*, Vol. 17, No. 7, pp. 1154-1167, July 2008.

[5] Q. Chen, N. D. Georganas, and E. M. Petriu: “Hand Gesture Recognition Using Haar-Like Features and a Stochastic Context-Free Grammar,” *IEEE Trans on Instrumentation and Measurement*, Vol. 57, No. 8, pp. 1562-1571, Aug. 2008

[6] T. M. Cover: “Geometrical and Statistical Properties of Systems of Linear Inequalities with Application in Pattern Recognition,” *IEEE Trans. on Electronic Computers*, Vol. 14, pp. 326-334, 1965.

[7] A. Justa and S. Marce: “A Comparative Study of Two State-of-the-art Sequence Processing Techniques for Hand Gesture Recognition,” *Computer Vision and Image Understanding*, Vol. 113, No. 4, pp. 532-543, Apr. 2009.

[8] D. G. Lowe: “Object Recognition from Local Scale-invariant Features,” in Proceedings of 7th ICCV, pp. 1150-1157, 1999.

[9] D. G. Lowe: “Distinctive Image Features from Scale-invariant Key Points,” *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004.

[10] S. Mitra and T. Acharya: “Gesture Recognition: A Survey,” *IEEE Trans. on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, Vol. 37, No. 3, pp. 311-324, May 2007.

[11] S. S. Rautaray and A. Agrawal: “Vision Based Hand Gesture Recognition for Human Computer Interaction: a Survey,” *Artificial Intelligence Review*, Nov. 2012.

[12] R. Poppe: “A Survey on Vision-based Human Action Recognition,” *Image and Vision Computing*, Vol. 28, pp. 976-990, 2010.

[13] L. G. Shapiro and G. C. Stockman: “Computer Vision,” 1st Edition, Prentice Hall, 2001.

[14] L.-P. Son, A. Bouzerdoum, and D. Chai: “A Novel Skin Color Model in YCbCr Color Space and its Application to Human Face Detection,” in Proceedings of ICIP, Vol. 1, pp. 289-292, 2002.

[15] S. Theodoridis and K. Koutroumbas: “Pattern Recognition,” 4th Edition, Academic Press, 2008.

[16] P. Viola and M. Jones: “Robust Real-Time Face Detection,” *International Journal of Computer Vision*, Vol. 57, No, 2, pp. 137–154, 2004.

[17] LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/.

[18] Microsoft Kinect: http://www.microsoft.com/en-us/kinectforwindows/