International Journal of Engineering Innovation & Research
Volume 3, Issue 2, ISSN: 2277 – 5668

# Automatic Raising Hand Detection in an Intelligent 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.
Email: kevin@csie.ntnu.edu.tw

***Abstract* – Future classroom should cover many intelligent features to improve the learning performance of students. Raising hand is one of the most fundamental interactions between students and lectures in classroom. When an intelligent system can automatically understand which students raise their hands, it becomes possible to design more advanced features that can help the teaching in the future classroom. This paper aims to design a practical system for detecting the gesture of raising hand in a real classroom. Getting a video sequence in a classroom, the foreground parts that only contain human bodies are first segmented. We design, therefore, a shape-like descriptor to represent the human bodies and then employ the support vector machine to recognize the student gestures. In order to more precisely understand the student gesture, the student gestures in this work are classified into three classes: left hand, right hand, and normal which means other gestures except raising left or right hand. This work has been implemented and installed in a university classroom. We performed several experiments for this system and demonstrate the results in this paper to present the efficiency.**

***Keywords* – Future Classroom, Gesture Recognition, Foreground Detection, SIFT, SVM.**

## I. INTRODUCTION

Many researchers have paid attention to applying existed technologies of computer vision to solve different kinds of problems in our life. For instance, gesture recognition [13][15] is a key technology to understand human actions appeared in images or videos. Many state-of-the-art methods for gesture recognition have been published. However, the application domain deeply affects the practical design of the gesture recognition.

Our researches focus on building a smart environment for future classrooms in order to help students improve the learning performance in school and university. One of the most fundamental actions for students is raising hand 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 becomes possible to design more intelligent applications to analyse interactions between students and lecturers.

Raising hand is a simple action in classroom, but it is not trivial to design an automatic detection system for raising hand. The main reasons are summarized as the follows. First, there are often a lot of students stayed in a classroom. A practical detection system has to deal multiple persons with action classification. Next, 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 it is hard to segment them by using a simple background subtraction method. Third, different students should attend different classes. Thus, we cannot expect that students appearing in a classroom can keep fixed. Moreover, various conditions in a classroom such as lighting, seat position and subject occlusion also make the system design more difficult.

We have implemented a practical system to automatically detect raising hands of students in classroom. The proposedsystem was install in a real classroom shown as Fig. 1. Our intuitive idea of the student gesture detection is to employ Kinect[21][23] which wasfirst launched by Microsoft in November 2010 to sense subject moments without any touch. However Kinect is not appropriate for a real classroom because its effective sensing distance is limited. In this work, we adopt a vision-based approach that employ computer vision technologies to treat the raising hand detection in classroom.

This classroom shown in Fig. 1 contains three rows of seat. Students can arbitrarily select their own preferred seats. In order to simplify our problem, three cameras were installed and fixed under the ceiling to cover the whole seat area shown as Fig. 2. That is to say, the camera views are also fixed in our problem. 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.

Fig. 3 presents the flowchart of our proposed system for detecting raising hands in classroom. Given consecutive video frames from a camera, the foreground areas are first segmented to locate the student positions. Then, a set of shape appearance signatures that is based on the scale-invariant feature transform (SIFT) descriptor [9][10] is extracted from each of located student bodies. We therefore employ the support vector machine (SVM) [19] 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 provides a literature review related to this work. Section 3presents the extraction process of our proposed shape-like descriptor of a human body to represent the student gestures. Then, the classification method using

SVM is stated in Section 4. 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.

Fig.1.The scene of the classroom in our experiment.

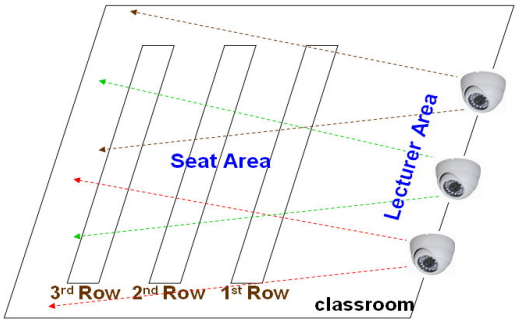



Fig.2.The classroom contains three rows of seats, covered by three cameras.

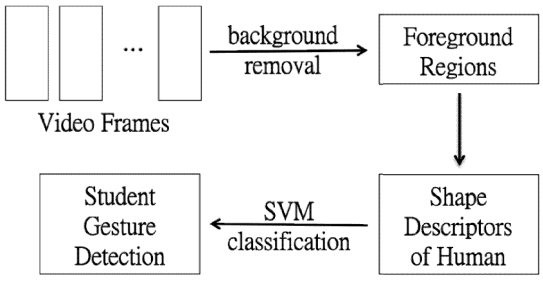



Fig.3.Theoverall flowchart of our proposed system.

## II. Related Works

Gesture recognition is one of important applications for computer vision. 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 [12]. R. Poppe also published a literature survey of the human action recognition using vision-based approaches [15]. S. S. Rautaray and A. Agrawal provided a comprehensive review of human and computer interaction focusing on using vision-based hand gesture recognition [13]. 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][8] on body and arm gestures, and [2][4][5] on hand gestures.

Microsoft Kinect provides a potential solution to estimate human gestures more accurately. There have been many researches published to treat gesture recognition using Kinect [14][18]. However, the sensing distance of the Kinect is very short, only about 3 meters. Hence, the Kinect is not appropriate for a common classroom.

The appearance representation of human body is also an important issue for gesture recognition. Many local descriptors have been proposed for image representation [11]. SIFT descriptor [9][10] 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.N.Dalal and B. Triggs proposed the Histogram of Oriented Gradients (HOG) [7]in 2005 to describe the human appearance for pedestrian detection in a still image. The idea of HOG is to count occurrences of gradient orientation in dense grids of uniformly spaced cells.In this work, we design a shape-like feature that is based on the SIFT descriptor to describe the shape appearance of raising hand gestures.

## III. Shape-Like Descriptor of Human Body

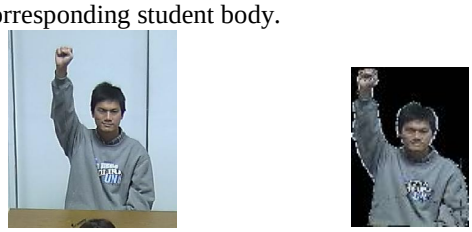
When cameras capture video sequences in classroom, the appearance descriptors of foreground areas in frames need to be extracted for representing student gestures.

### *A. Background Removal*

In order to detect foreground targets in video sequences, the first task is to automatically recognize student gestures 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 originally in the classroom. 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 modelling for the camera view. In order to detect student actions as soon as possible, a background modelling needs to be sensitive to foreground motion. However a sensitive background modelling could also generate a lot of noises that will make more false alarms of foreground targets.

This paper employs two well-known approaches: Gaussian Mixture Model (GMM)[19] and temporal differencing [16], to perform the background extraction for extracting the foreground regions in classroom. The GMM method is widely used for constructing a dynamic background modelling 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 modelling 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 eliminate the background parts in video frames.

The background areas, however, may cover non-human objects that make student gesture detection more difficult. Given extracted foreground regions in video frames, we employ the adaboosting approach [20] of face detection and the skin detection approach [17] 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. Fig.4 presents an example of an original frame and the corresponding student body.



(a) The original frame (b) The foreground area

Fig.4. Student body in the foreground area

### B. *Descriptor Extraction*

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 [9][10]. The details of extracting SIFT descriptor is referred to Dr. D. Loew's publications in [9][10]. 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

Fig.5. Sampling of feature points by SIFT descriptors for raising hands.

The SIFT descriptor can represent significant contents in an object illustrated as Fig. 5, 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 $3\times K$ dimensions.

## IV. 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 [19] 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 [22] 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 $\{(x_1, y_1), ..., (x_n, y_n)\}$ given where $x_i$ means the feature vector of training data, and $y_i \in \{1, -1\}$ is the labeled of $x_i$. In order to define a linear classifier $y_i = wx_i + b$, the SVM method solves:

$$\min_{\mathbf{w},b}\{\frac{1}{2}\|\mathbf{w}\|^2\} \quad \text{subject to} \quad y_i(\mathbf{w}^t\mathbf{x}_i + b) \geq \mathbf{1} \tag{1}$$

Since it is in general a non-linear classification problem, a nonlinear function $\phi$ 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

$$\max_{\lambda}\left(\sum_i \lambda_i - \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j \phi(\mathbf{x}_i)^t\phi(\mathbf{x}_j)\right) \quad \text{subject to} \quad \sum_i \lambda_i y_i = 0 \ \text{ and } \boldsymbol{\lambda} \geq 0 \tag{2}$$

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

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

for solving this equation without having to compute the mapping $\phi$ 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 $\lambda$.

International Journal of Engineering Innovation & Research
Volume 3, Issue 2, ISSN: 2277 – 5668

## V. EXPERIMENTAL RESULTS

The proposed system has been installed in a university classroom to collect our training and test videos. The camera setup in the classroom for our experiments is drawn in Fig. 2. Three cameras are involved in our system to cover the whole classroom, and these camera views are fixed for simplifying the system implementation.

The experimental data contains two sets, which are captured with different time and with different students. These two sets are called $D1$ and $D2$ that arepartitioned into two parts with the equal size roughly: one for training and the other for test. That is to say, we can denote data set $D_i=M_i \cup V_i$ for $i$=1 and 2, where $M_i$indicates 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 indicate 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 |

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, 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 applies 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.

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: $M_1$ Test: $V_2$ corr. to Table 3 | Training: $M_2$ Test: $V_1$ corr. to Table 4 |
|---|---|---|
| 3rd row | 0.93 | 0.918 |
| 2nd row | 0.663 | 0.521 |
| 1st row | 0.677 | 0.416 |

(a). The third row of seats

(b). The second row of seats

(c). The first row of seats

Fig.6. 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.

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. Our approach can achieve high rates in the third row but not very successful in the other two rows. Fig.6 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.

## VI. Conclusion and Future Works

This paper presents our effort to design a raising hand detection system using a vision-based method in a classroom. We have implemented the proposed approach and installed this system in a university classroom. We introduce the details of our approaches and show the experimental results to discuss the efficacy of our proposed system. In the future, we plan to design an advanced approach to carefully tracking students and capturing their body areas in video sequences. A good result of student tracking can help the system improve the analysis of the student behavior. Another possible way to extend this work is to involve more gestures/actions in classroom such as eating, drowse, and chat.

## References

[1] J. Alon, V. Athitsos, Quan Yuan, and S. Sclaroff, "A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation," IEEE Transactions 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 Machine Vision Applications, 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 Transactions 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 Transactions on Electronic Computers, Vol. 14, pp. 326-334, 1965.

[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Vol. 1, pp. 886-893, 2005.

[8] 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.

[9] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," in Proceedings of 7th International Conference on Computer Vision, ICCV, pp. 1150-1157, 1999.

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

[11] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 10, Oct. 2005.

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

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

[14] Z. Ren, J. Meng, J. Yuan, and Z. Zhang, "Robust Hand Gesture Recognition with Kinect Sensor", in Proceedings of the 19th ACM international conference on Multimedia, ACMMM, pp. 759-760, 2011.

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

[16] L. G. Shapiro and G. C. Stockman, "Computer Vision," 1st Edition, Prentice Hall, 2001.

[17] 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 International Conference on Image Processing, ICIP, Vol. 1, pp. 289-292, 2002.

[18] J. Suarez and R. R. Murphy, "Hand Gesture Recognition with Depth Images: A Review", in Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 411-417, 2012.

[19] S. Theodoridis and K. Koutroumbas, "Pattern Recognition," 4th Edition, Academic Press, 2008.

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

[21] Z. Zhang, "Microsoft Kinect Sensor and Its Effect", IEEE MultiMedia, Vol. 19, No. 2, pp. 4-10, Feb. 2012.

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

[23] Microsoft Kinect:http://www.microsoft.com/ en-us/kinectfor windows/

## Author's Profile

**Cheng-Chieh Chiang**
received a Ph.D. degree in Computer Science from National Taiwan Normal University, Taipei, Taiwan, in 2007. He is currently an Assistant Professor at the Department of Information Technology, Takming University of Science and Technology, Taipei, Taiwan. His research interests include multimedia system, pattern recognition, and computer vision.