Haar-Feature Based Gesture Detection of Hand-Raising for Mobile Robot in HRI Environments

Hong Liu, Dengke Gao

Key Laboratory of Machine Perception and Intelligence,
Shenzhen Graduate School, Peking University,
Shenzhen, China

E-mail**:** hongliu@pku.edu.cn, dengke.g@gmail.com

Abstract

This paper proposes a method for hand-raising gesture detection that can be used for mobile robots in indoor environments. Different from traditional methods which are capable of the detection of hand-raising gesture with a static camera, our method can process the detection on a non-stationary platform. At first, haar-like features of raised arms are extracted and a cascade adaboost classifier is trained. Then the classifier is used to detect whether there are hands raised in specific regions which are established by results of face detection. The detector completes the detection in different scales at all locations of an input image. Experiments on a mobile robot are implemented in indoor environments where several persons are walking or standing randomly. Experimental results show that our method is suitable for real-time hand-raising gesture detection in Human-Robot Interaction in indoor environments.

Key words: hand-raising gesture, Human-Robot Interaction, haar-like features, cascade adaboost classifier.

1 Introduction

There have been many researches on computer vision based human motion detection and analysis because of their wide applications such as Human-Robot Interaction (HRI), video surveillance, Human-Computer Interaction (HCI), distance learning and clinical studies motion analysis [1,2]. Hand-raising gesture is common in people-people interaction, which is frequently used to attract one's attention. If a computer vision system has the ability to detect hand-raising gesture, it will be very useful in HCI, HRI, distance learning and so on. Here we take HRI for example. If there is a mobile service robot in a hall and someone wants to ask the robot for some interaction, speaking seems to be a good choice. However, when the environment is noisy and the robot is a bit far, a better and humanized way is to raise hands and wave to the robot just like waving to a person. The robot captures hand-raising motion or gesture, and then moves closely for the next interaction.

It is a challenging task to detect hand-raising motion or gesture. Firstly, non-rigidity of human body brings the variety of whole-body appearances when a person is raising his hands. The variety complicates the problem. Secondly, if persons are raising their hands while walking, the trajectories of the raising hands are various. It makes analyses from trajectories more difficult. Thirdly, if the camera is moving, it will be difficult to detect human bodies from the scene. Fourthly, occlusion often occurs in the circumstances. A human body may be partially occluded by others. In HRI, how to detect hand-raised gesture in real time is also very important. Because the less time spent, the more friendly the system is. Several methods have been brought forward from the viewpoint of gesture or motion.

It is common to detect hand-raising by analyzing motion cues. HMM based method was utilized by Hossain [3] and Kapralos [4]. Both of them used a sequence of images and a set of HMMs to complete the detection. Kapralos et.al detected hand-raising gesture for a remote learning application with an omni-directional camera. Jie et al. implemented a system of arm gesture detection used in classroom [5]. It made use of temporal and spatial segmentation, skin color identification, shape and feature analysis. In the scenario using these methods, the person to be detected is standing still in the room. If a person walks with a hand raised, or the camera is moving, the trajectory of the hand-raising motion will be very different, and methods above will be invalidated.

Liu and Duan introduced a method to detect hand-raising gesture from the view of gesture, detecting with the body silhouette analysis [6]. Instead of the use of appearances and movement of bodies, they searched raised arms and hands through body silhouette analysis. Their method could be used in scenarios that have several persons moving and solves problems brought by body movement. Because of the using of background subtraction, the result of detection with body silhouette analysis will be worse if the camera is moving.

In HRI, robots and persons are not always stock-still. Persons may walk into the visual field of a robot with hands raised and the robot is non-stationary, which makes those methods above to be unavailable. We present a novel vision-method to detect hand-raising for mobile robot from the viewpoint of gesture. Haar-like features are utilized to train a classifier of raised arms using cascaded adaboost at first. Then the whole of an input image is scanned with the classifier in many scales and at all locations to test whether the sub-window scanned is a raised arm or not. The method doesn't use any background information and motion cues, and it just operates on an image, thus it solves the problems brought up by the movement of a camera or human body. It can be used in mobile service robots in indoor environments.

The rest of this paper is organized as follows. Section 2 introduces the extraction of haar-like features. Section 3 illustrates the architecture of our method. Online and offline experiments are detailed at Section 4. Section 5 draws the conclusion and some ideas for future work.

2 Features Extraction

Haar-like features are simple and easy to compute. The set of haar-like features we utilize is introduced by Rainer et al.[7]. which was inspired by the over-complete haar-like features used by Papageorgiou et al. in [8, 9] and their fast computation method proposed by Viola et al. in [8]. The features we use are presented in Figure 1. They are edge features, line features and center-surround features. The value of these features, which is simple to compute, can be calculated as:

$$V_I = \sum_{i \in I = \{1,2,\ldots,N\}} \omega_i \cdot Srec(r_i) \qquad (1)$$

where the weight ω_i is real, $Srec(r_i)$ is the sum of pixels within rectangle r_i. The value of a feature is just the difference of pixels within white rectangles and black rectangles. Take line feature of Figure 2(b) for example. Given the top left corner is at (10, 10) with the total height of 1 pixel and total width of 4 pixels, the feature can be written as: V_I= -1·*Srec*(10,10,4,1,0) +2·*Srec*(11,10,2,1,0).

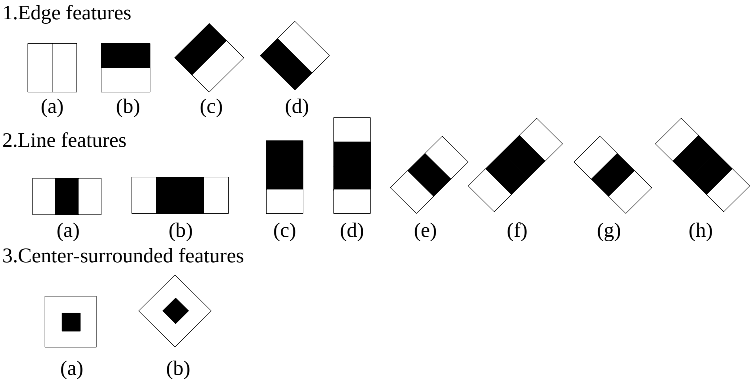
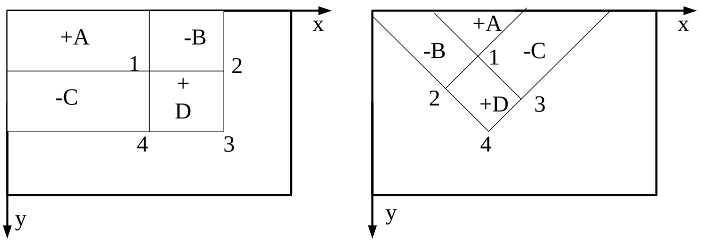


Figure 1 Feature prototypes

In [8] the proposers come up with a fast computation method "integral image" for the features, and [7] enrich the features pool by adding rotated features with a computation method for the corresponding features.

Given that the basic window to be detected is of $W\times H$ pixels, a rectangle in the window can be denoted by the tuple (x, y, w, h, α) , where (x , y) is the coordinate of the top left corner pixel with $0 \le x$, $x + w \le W$, $0 \le y, y+h \le H$, $\alpha \in \{0,45\}$is the rotation angle, *w* and *h* are respectively the width and the height of the rectangle. The rectangle features can be computed rapidly with the integral image.

For the upright rectangle, integral image at location (*x, y*) contains the sum of the pixels above and to the left of (*x, y*).



(a) (b)

Figure 2 (a) Calculation scheme of sum of upright rectangle; (b) Calculation scheme of sum of rotated rectangle.

$$Srec(x,y) = \sum_{x'\le x,y'\le y} I(x',y') \tag{2}$$

It can be calculated as follows:

$$Srec(x,y) = Srec(x, y-1) + Srec(x-1, y) + I(x, y) - Srec(x- 1, y - 1), \tag{3}$$

where *Srec (-1,-1) = Srec (-1,y) = Srec (x,-1)* = 0. From this any upright rectangle can be calculated by four table lookups (see Figure 2(a)).

For 45° rotated features, the integral image is *Srot(x,y),* which is defined as the sum of the pixels of a 45° rotated rectangle with the bottom most corner at (*x, y*) and extending up to the boundaries of the image.

$$Srot(x,y) = \sum_{y'\le y,y'\le y-|x-x'|} I(x',y'), \tag{4}$$

It can be calculated by:

$$Srot(x,y) = Srot(x-1,y-1) + Srot(x+1,y-1) + I(x,y)-Srot(x,y-2) + I(x,y-1), \tag{5}$$

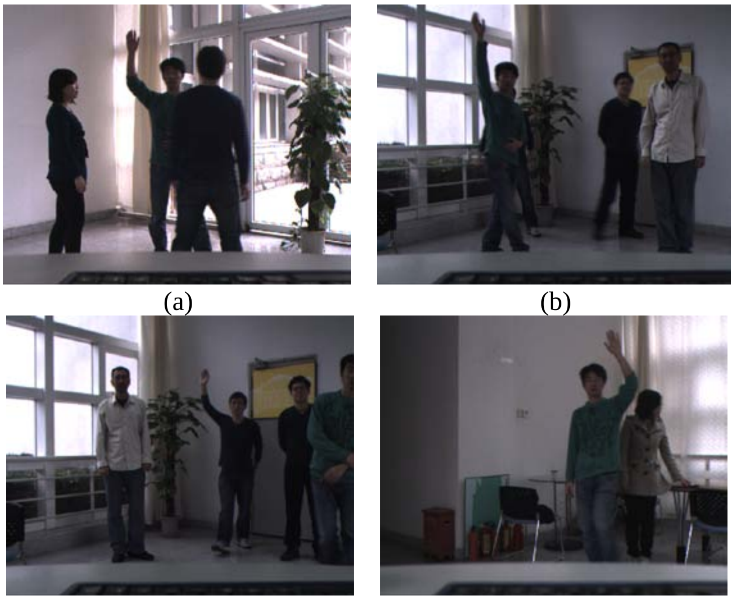
where *Srot(-1,y) = Srot(x,-2) = Srot(x,-1)* =0, and *Srot (-1,-1) = Srot (-1,-2)* = 0. From this any 45° rotated rectangle can be calculated by four table lookups (see Figure 2(b)).

3 Cascade Detection and Classification

The main motivation of this paper is to detect the hand-raising gesture real-time under a non-stationary background on a mobile robot in indoor environments. Background subtraction and analysis of motion cues are ineffective for the scenario. Therefore, we accomplish the detection by scan every frame of the video stream with detectors. Firstly, sub-regions where objects of interest may be in are obtained to reduce the time. Secondly, the sub-regions are scanned by the detector trained with cascaded adaboost.

3.1 Sub-Regions of Interest

From images in Figure 3 and Figure 4, it can be seen that the gestures of hand-raising of the whole-bodies are various because of the non-rigidity of human body. From the whole body we cannot get useful information, but there are some similarities between the parts encompassed by the rectangles in Figure 4. This is a key region which can show that an arm has been raised. The parts in the rectangles have little changes. Therefore, it is effective to detect this part instead of the whole-body or others. The next step is to choose a suitable way in which the key feature of the region can be expressed well.

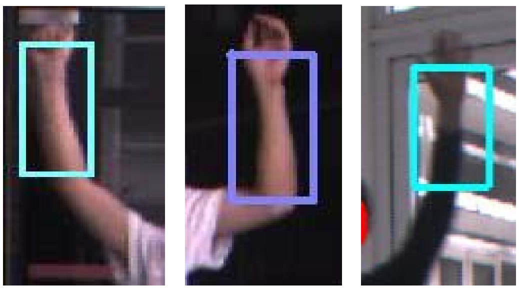


(a) (b)

(c) (d)

Figure 3 Different kinds of hand-raising gesture of human body

There are many methods that can extract main features well and get a high detection rate in detection. However what we need is not only a high detection rate, but also a rapid speed for feature computation. Because little time spent on detection will make the interaction more friendly. The extended set of haar-like features is chosen after comparing several methods for feature extraction, and it will be stated in next section.



(a) (b) (c)

Figure 4 Different kinds of hand-raising gesture of human body

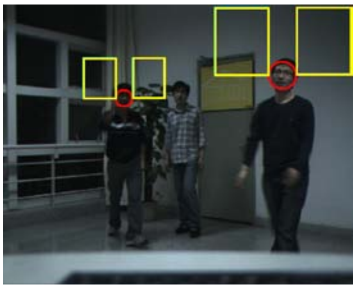


Figure 5 The radius of face decide the regions to detect

If locations of persons are known, the area of arm raised is easy to obtain. Figure 5 shows sub-regions that may have objects of interest. Due to the constraints of human body, a hand raised or an arm raised will be just in the regions beside the head. The location and the size of the sub region can be computed from the location and diameter of the circles that the face detected. The width W_{SR} and height H_{SR} of the sub region can be computed as follows:

$$W_{SR} = \alpha_1 \cdot D, \tag{6}$$

$$H_{SR} = \alpha_2 \cdot D, \tag{7}$$

here, D denotes the diameter of the circle, α_1 and α_2 are the factors. The constraint of human body makes the value of α_1 and α_2 constant. The size of the sub region is relevant to the distance between the person and the robot. When a person is far away from a robot, the sub regions is small, and when the person is near the robot, the sub regions are bigger (see Figure 5).

The location is represented by the coordinate of the top left corner of the sub region.

$$P_{xL}= C_x -\beta_{11}\cdot D, \tag{8}$$

$$P_{yL} = C_y -\beta_2\cdot D, \tag{9}$$

$$P_{x,R}= C_x +\beta_{12}\cdot D, \tag{10}$$

$$P_{yR} = C_y -\beta_2\cdot D, \tag{11}$$

here, β_{11},β_{12} andβ_2 are the constant factors like α_1 and α_2. (P_{xL}, P_{yL}) and (P_{xR}, P_{yR}) are the coordinates of the top left corners of the left and right sub region respectively. (C_x, C_y) is the coordinates of the center of the circle. The size and location of the region will change according to the person's location so that the hands raised are in the regions.

3.2 Cascaded Classifier

After regions for detection are obtained, what needs to be done is to detect the regions with a detector. Cascaded classifier from [10] is applied in this paper to handle the detection. This structure rejects majority of the non-object regions with the utilization of adaboost [14] in each stage, which uses weak classifiers to make up of a strong classifier.

If the detector scans the full image at all locations with several scales, there will be a large number of features to be computed. Even though each feature is easy to compute, it will take much time. Adaboost is introduced to select main features in each stage of the cascaded classifier.

This method uses the concept of cascade. It can not only improve the detection rate, but also greatly reduce the time for detection. For example, at the first stage the classifier chooses some features with high direction rate and low false negative rate with rejecting the majority of negative sub-windows. Then it triggers the second stage classifier that will reject some negative sub-windows in the image like the first one. And the third is triggered by the second until the expected detection rate and false positive rate are reached. Figure 6 illustrates the process.

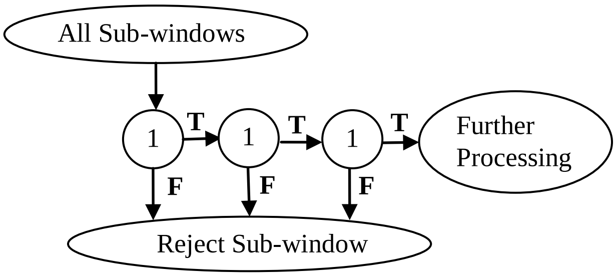


Figure 6 Illustration of detection with cascade

3.3 Architeture

The detection is accomplished by scanning a frame of the video stream with detectors to judge whether the sub-window is the object of interest. The architecture is shown in Figure 7.

In the architecture, the concept of cascade is used in the detection to reduce the areas for scanning. Two classifiers have been trained respectively at first. The method scans all locations of the input image with the face detector to find person, and then scans the specific regions, which are obtained by the result of face detection, with an arm or hand-raised detector. A positive output from the hand-raised detector means a person with the gesture of hand-raising is found.

When the detector scans a sub region, it will scan all the locations of the region with initial scale 1 at first. The detector will scan the region many times in other scales next. If scaling the image every time, the method will need much more time on the computation. It saves the time for detection to scale the detector. The method we adopt can avoid the detection on the regions unlikely to be the object of interest, and lower the probability of false positive.

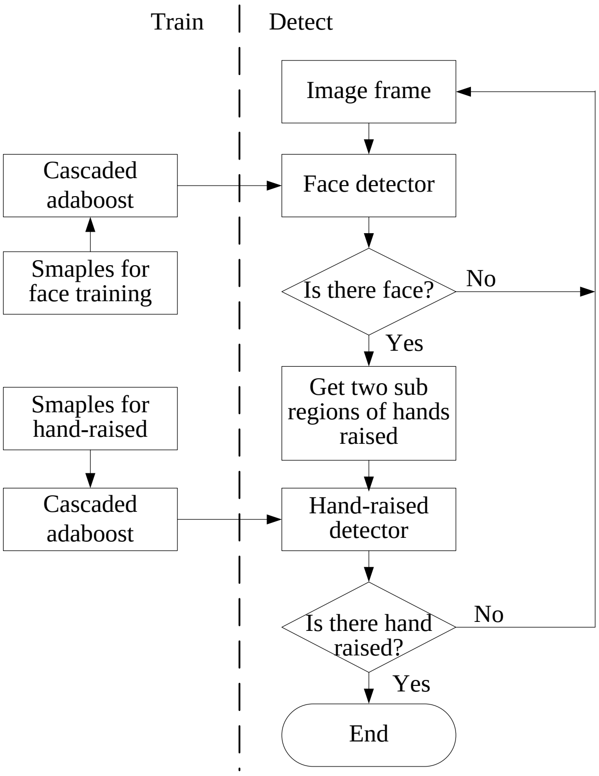
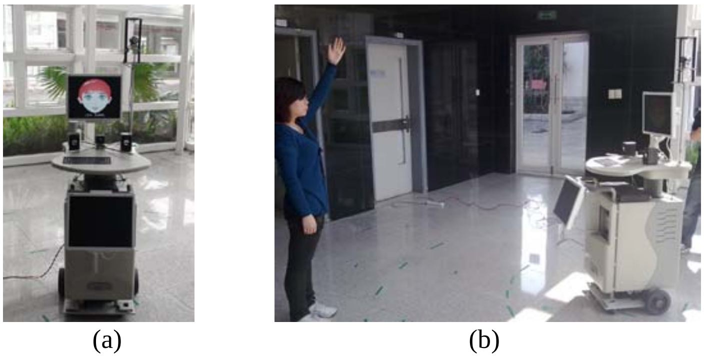


Figure 7 Flow chart of our method

4 Experiments and Discussions

The method of this paper is implemented on a mobile robot "PengPeng II", and lots of experiments have been carried out in indoor environments. Online experiments and offline experiments are completed with the same hardware. The CPU is Intel Core2 Quad 2.4GHz. The resolution of the video stream is 640x512 pixels.



(a) (b)

Figure 8 (a) Mobile Robot "PengPeng II"; (b) Interaction between "PengPeng II" and a person

In experiments, persons can walk with hands raised, or raising a hand while walking. And there is only one person whose hand is raised at a time in the scenario. The robot can turn right, turn left, move forward and move backward to change the background. Many situations for the detection were tested, such as a person who raises a hand standing far away from the robot, near the robot, or walking into the view field of the robot and so on. Some of the experiments are done at day and some are at night.

A hand-raised detector of 20 stages was trained with 1394 positive samples and 3200 negative samples. The detector has the width of 30 pixels and the height of 48 pixels. For the size of sub-regions of raised arm, parameters in (6) and (7) are set as α_1=2.0 and α_2=2.25. For the locations of the sub-regions, we set β_{11}=2.5,β_{12}=0.5 and β_2 = 2.5. The scale factor is set to 1.1. The final number of stages is 13, which is the result of a trading off. For example, when the number of stages is 17, the false positive is very low, but the false negative is very high so that the number of hand-raising missed is large. If the number of image frames missed is large, the detection time will be long, even the hand-raising can't be captured. When the number of stages is too small just as 6, the false positive is high so that the detection rate is high in the experiments offline. After the comparing, 13 stages has a good balance between the detection rate and time.

In experiments, time and detection rate are the key factors. Experiments about time are completed and prove that our method is a better choice for HRI. There have been many methods for feature description, and they perform well in detection rate for human detection, such as HOG [11,12], Glac [13] and Edgelet [15]. We take HOG and Glac to compare with haar-feature. It is obvious that the computation of haar-feature is simpler than that of HOG and Glac. Experiments show that the method with haar-feature is much faster than methods with HOG and Glac.

The online experiments are different from offline experiments. Because of the computation time of the current frame, not all the frames are detected. For the experiments online, it is difficult to tell the robot when to count time and when to stop the counter, so we only counts the number of actual hand-raising (NOA), the number of hand-raising captured (NOC), the number of hand-raising missed (NOM) and the number of false detection (NOF). The direction rate means NOC/NOA.

Table 1 Results of detection online

		NOC	NOM	NOF	NOA	Detection rate
Day	Robot moving	165	10	25	200	82.5%
	Robot still	224	6	20	250	89.6%
Night	Robot moving	161	12	27	200	80.5%
	Robot still	214	7	29	250	85.6%
Sum		764	35	101	900	84.9%

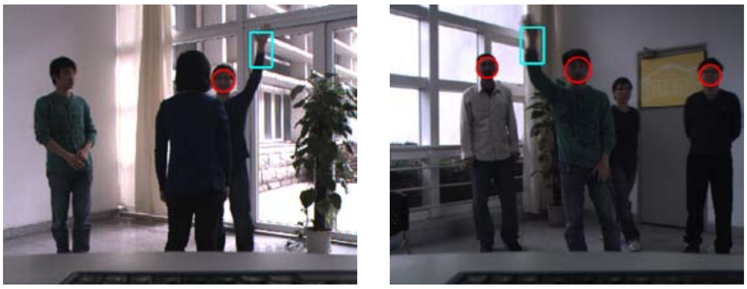
Scenarios for the offline experiments are same to the online one. For example, experiments are implemented both at day and night with the movements of the robot and persons.

Table 2 Results of detection offline

		AVT(ms)	NoM	NoF	NoA	Detection rate
Day	Robot moving	94	145	10	722	84.2%
	Robot still	97	152	15	810	92.0%
Night	Robot moving	96	123	27	693	82.1%
	Robot still	98	133	29	755	84.7%
Sum			553	81	2350	85.6%

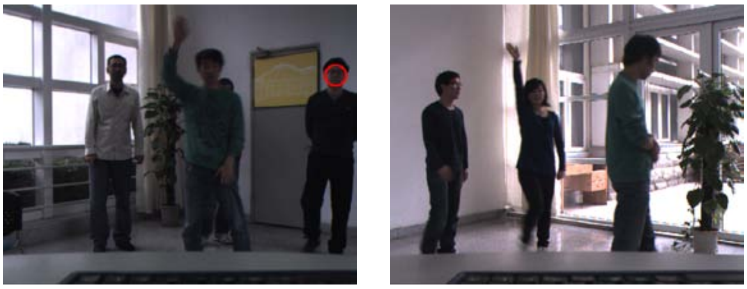
In Table 2, AVT is the average time spent on detection of hand-raising gesture per frame in the video sequences. NoM, NoF and NoA are the number of missed frames, the number of frames that are false positive, and the total number of the frames that have hand raised respectively.

The tables above and figures show the results of our method. As shown in Figure 9, our method performs well at different background images. In the images there are many cylinders, which are easy to be mistaken as hands raised. When the number of stages is 13, the mistake disappears. Though some frames in which there are hands raised are mistaken as negative, the hand-raising are still detected by the frames next. In Figure 9(a), the person is occluded by another person, but the method can still detect the hand-raising gesture.



(a) (b)

Figure 9 Correct Results



(a) (b)

Figure 10 Frames Missed. In (a) and (b) the hands raised are not captured

5 Conclusions

In this paper, we utilize haar-like features and cascaded classifier on mobile robots to detect hand-raising gesture. In the method, a person is found by face detection, and then the constraint of human body structure is used to establish the candidate regions for raised hands. It can reduce the computation time and probability of false positive to detect hands raised in the candidate regions. Experiments offline and online have shown that the method can detect the gesture of hand-raising in a humanized way with rapid detection and high detection rate on the mobile robot in indoor environment.

There are still some problems need to be solved. When two persons stand closely the method may mistake one's arm as the other's. It could be solved by the other method like research the connection of hand and the face with line, gradient or color features.

6 Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Shenzhen Scientific and Technological Plan and Basic Research program (No.JC200903160369A), Natural Science Foundation of Guangdong(No.9151806001000025).

References

[1] T.B.Moeslund and E. Granum, A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding*, vol.81, pp.231-268,2003

[2] T.B.Moeslund, A.Hilton, and V.Krüger, A survey of advances in vision-based human motion capture and analysis, *Computer Vision and Image Understanding*, vol.104, pp.90-126, 2006.

[3] M.Hossain and M. Jenkin, Recognizing ahnd-raising gestures using hmm, in *Canadian Conference on Computer and Robot Vision*, pp. 405-412, 2005.

[4] B.Kapralos, A.Hogue, and H. Sabri, Recognition of hand raising gestures for a remote learning application, in *International Workshop on Image Analysis for Multimedia Interactive Services*. pp. 38-41, 2007.

[5] J.Yao and J.R.Cooperstock, Arm gesture detection in a classroom environment, in *Workshop on Applications of Computer Vision*. pp. pp.153-157, 2002,.

[6] X.Duan and H.Liu, Detectioin of hand-raising gestures based on body silhouette analysis, in *International Conference on Robotics and Biomimetrics*. pp. 1756-1761, 2008.

[7] R.Lienhart and J.Maydt, An extended set of Haar-like features for rapid object detection, IEEE ICIP. 2002, Vol.1, pp. 900-903, 2002.

[8] A.Mohan, C.Papageorgiou, T.Poggio, Example-based object detection in images by components, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.23, No.4, pp.349-361, 2001.

[9] C.Papageorgiou, M.Oren, and T.Poggio, A general framework for Object Detection, *International Conference on Computer Vison*, pp.555-562,1998.

[10] Paul Viola and Michael J.Jones, Rapid Object Detection using a Boosted Cascade of Simple Features, *IEEE CVPR*, pp.511-518, 2001.

[11] N.Dalal and B.Triggs, Histogram of oriented gradients for human detection, *IEEE, CVPR*, Vol.1,pp.886-893, 2005.

[12] Q.Zhu, S.Avidan, M.C.Yeh, and K.T.Cheng, Fast human detection using a cascade of histograms of oriented gradients, In Proc. *IEEE Conf. on Computer Vision and Pattern Recognition*, New York, NY, Vol.2, pp. 1491-1498, 2006.

[13] T.Kobayashi and N.Otsu, Image Feature Extraction Using Gradient Local Auto-Coorelations, *Proceedings of the 10th European Conference on Computer Vision*:Part 1, pp. 346-358, 2008.

[14] Freund Y and Schapire RE, A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence* Vol.14, pp.771-780, 1999.

[15] B.Wu, R.Nevatia, Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detection, *International Conference on Computer Vision*, Vol.1,pp.90-97, 2005.