Detection of a Hand-Raising Gesture by Locating the Arm

Nyan Bo Bo, Peter Van Hese, Dimitri Van Cauwelaert, Peter Veelaert, Wilfried Philips

Abstract— **This paper proposes a novel method for detecting hand-raising gestures in meeting room and classroom environments. The proposed method first detects faces in each frame of the video sequence in order to define the region of interest (ROI). Then the system locates arms in the region of interest by analyzing the geometric structure of edges on the arm instead of directly detecting the hand. The location and the orientation of a detected arm respect to the location of the face is used to make a decision on whether or not a person is raising hand. Finally, the frequency of a raised hand detected in previous frames is used to eliminate false positive detections and robustly detects persons who are raising a hand. Unlike major visual gesture recognition systems, our method does not rely on skin color or complex tracking algorithms, while achieving 92% sensitivity and 92% selectivity.**

I. INTRODUCTION

People raise their hand to draw the attention of others, especially in meeting room and classroom environments. The system that is able to robustly detect people who are raising a hand can have many applications in smart meeting rooms, intelligent surveillance, remote classrooms and human-robot interaction. In smart meeting room or remote classroom applications, the camera can focus on a person once the system detects that person is raising a hand. When the system knows which person is raising a hand, facial recognition can be used to gather further information such as identification of the person seeking attention. Moreover, a person can get attention from a robot so that the person can request services from the robot.

Much work has been done in the area of visual human gesture and action recognition [1], [2] but very few research papers focus on the recognition of a hand-raising gesture. The system proposed by Kapralos *et al.* [3] extracts the motion cues from a sequence of omni-directional images and then recognizes the hand-raising gestures from those motion cues using pre-defined Hidden Markov Models (HMM). However it requires to manually mark the location of hands in each image in the sequence for both training and recognition, limiting its usability potential for many applications. Although it is possible to automatically locate a hand, the detection of an arbitrary posed hand itself is very a challenging problem.

The robust hand detection system in [4] uses an object detection method proposed by Viola *et al.* [5] to robustly detect the view-specific hand posture in images. As mentioned in their paper, the system only performs well on the detection of view-specific hand postures, which are used to train the system, it is not suitable for detecting the arbitrary posture of a raised hand. For example, persons in the meeting room environment may hold a pen while raising their hand. Recently, Spruyt *et al.* [6] proposed a hand detection system based on color and motion cues together with a particle filter tracking algorithm. Although the testing results prove that the system works well in an unconstrained environment, it relies on skin color, which makes the detection difficult when the color of the hand is altered somehow, for example, by wearing gloves or by a shadow falling on the hand.

Some systems directly detect a raised hand without a need for detecting the hand first. A system proposed by Yao *et al.* [7] detects a raised hand in a classroom environment with the assumption that the students' heads are randomly distributed in approximately a horizontal plane. Their system segments foreground from background by using temporal and spatial segmentation to define an object of interest map. Then, edge detection, skin detection and shape and feature analysis are used to define if there is a raised hand or not. Although their experimental results show adequate performance, the system tends to perform poorly when the arm is largely covered by clothing of a similar color to the background.

Recently, Duan *et al.* [8] proposed a hand-raising gesture detection system based on body silhouette analysis. Their system first subtracts foreground from background in the image and then candidate regions are searched on the upper quarter of the foreground silhouette according to topology of the human body. From each candidate region, features are extracted using the $\mathcal{R}$-transform and PCA is employed to reduce dimensionality. Finally, these features are fed to SVN classifiers to make a decision if the candidate region contains a raised hand or not. The system is further improved by Liu *et al.* [9] using additional edge features and SVN classifiers. Their system is highly dependent on foreground subtraction which is sensitive to environmental changes such as lighting. In the case where foreground subtraction fails, there may be some false positives and missed detections.

Our proposed method detects the hand-raising gesture in the images based on the fact that the edges on each side of a raised arm are nearly parallel and are located at the same height as the location of the face. Therefore, it finds nearly straight edge segments in a region of interest which is defined based on the location and size of the face. Then it selects edge pairs, which are nearly parallel, and further selects among these pairs using prior knowledge of the geometric properties between the edges of the arm. The robustness of

Nyan Bo Bo is with Faculty of Technology and Environment, Prince of Songkla University, 80 Moo 1, Vichit-Songkram Road, Kathu, Phuket 83120, Thailand nyan.b@phuket.psu.ac.th

Peter Van Hese, Dimitri Van Cauwelaert, Peter Veelaert and Wilfried Philips are with Image Processing and Interpretation, Ghent University/IBBT, St-Pietersnieuwstraat 41, B9000 Ghent, Belgium

Peter Veelaert and Dimitri Van Cauwelaert are with Vision Systems, University College Ghent, Schoonmeersstraat 52, B9000 Ghent, Belgium

the method is further improved by making a decision whether or not a person is raising hand based on the frequency of detected raised hands of that particular person in previous frames.

Our method can work on grayscale video if the face detection system to be used does not require color information. This is desirable when the quality of color information in the fame is not reliable. Our method requires lower computation cost than the majority of the existing systems while maintaining a 92% sensitivity and 92% selectivity. The rest of this paper is organized as follows. Section II describes how video sequences are captured. Section III explains how the proposed method detects a person who is raising a hand in the video sequences. Section IV evaluates the system describes the results. Finally, this paper is concluded in Section V together with a plan for future work.

II. Data

The video sequences for parameter tuning and evaluation of the proposed method are captured in a meeting room set up, where two participants are facing the camera and one participant is sitting in such a way that the left side of the participant is seen in video sequences most of the time. The video sequences are captured at 50 frames per second with a resolution of 960×720 pixels.

Since the main focus of this paper is the detection of a raised hand based on the known location of the face, the evaluation of the method must be isolated from the performance of face detector as much as possible. In order to accomplish this, faces were detected using face detector, which is described in Subsection A of Section III, in 18 video sequences containing 29 cases of people raising a hand. A hand raising case is selected for evaluation if face of a person is detected in 50% of 20 consecutive frames while raising a hand.

The total of 17 video sequences containing 24 cases of people raising a hand were obtained for the evaluation of our proposed method. These 17 video sequences contain a total of 2,940 frames. During the capture of video sequence number 11 and 12, the illumination intensity of lighting in the meeting room is varied manually for the evaluation of the system performance in the environment with varying lighting condition.

III. Hand-Raising Gesture Detection

Our hand-raising gesture detector consists of four parts: region of interest (ROI) selection, straight line fitting on edge segment, upright arm localization and statistical decision making.

A. Region of Interest Selection

First, each frame of the input video sequence is converted into a grayscale image. Next faces are detected using the object detector that was initially proposed by Viola *et al.* [5] and later improved by Lienhart *et al.* [10]. The threshold for the minimum number of neighbors in each detected target is experimentally set to have a very low false positive rate. Since the focus of this paper is not on face detection, the details of the face detector are not provided. However, it is possible to use a more robust face detector to ensure better face detection rates. The width of each detected face w_{face} is expanded by factor of five and the height h_{face} is expanded by factor of three to define width w_{ROI} and height h_{ROI} of the region of interest as shown in Fig. 1.

$$w_{ROI} = 5w_{face} \tag{1}$$

$$h_{ROI} = 3h_{face} \tag{2}$$

Because of the anatomical constraints between the human face and arm, the raised hand of a person can only appear within the ROI. Only the selected ROIs are used for further processing.

B. Straight Line Fitting on Edge Segment

In each selected ROI, edges are detected with a Canny edge detector where $\sigma = 1$, $low\ threshold = 0.1$ and $high\ threshold = 0.2$. The output of the edge detector is a binary image in which 1s represent edge pixels and 0s represent non-edge pixels. Then a morphological cleaning operation is performed to remove isolated edge pixels that are surrounded by non-edge pixels. The remaining connected edge pixels are skeletonized by classical morphological skeletonization operation to make sure that the width of the edge segment is only one pixel as shown in Fig. 2 (a).

The resulting binary image contains edge segments which have a thickness of one pixel. The edge segments that have branches are further split at the branch junctions, so that each branch becomes an individual edge segment. The edge segments that contain less pixels than the threshold $t_{\#pixel} = 0.75w_{face}$ are removed because edges on the arm of a human are always longer than the width of the face. If the size of a detected face is large, the threshold $t_{\#pixel}$ is large also. At this point, most edge pixels from noise and background are eliminated as shown in Fig. 2 (b).

The next step is to fit a straight line with maximum possible length to each edge segment. While doing so, line

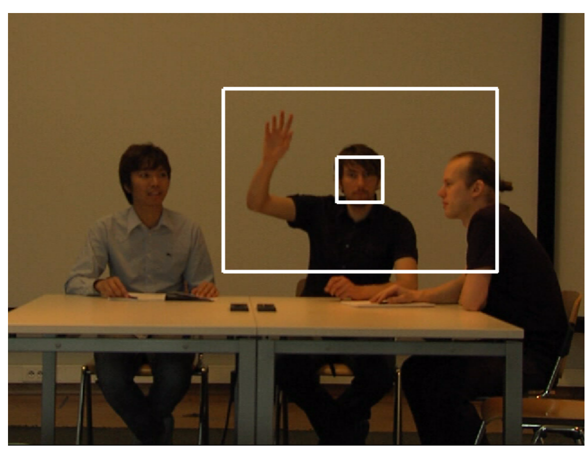


Fig. 1. The region of interest (ROI) is defined based on the size of the detected face. The inner rectangle shows the bounding box of the detected face and the outer rectangle shows the boundary of the ROI.

requires the maximum deviation between the edge pixels and the line do not exceed the threshold t_{dev}. To fit a line, one end of the straight line $p_1(x_1, y_1)$ is fixed to one end of the edge segment and the other endpoint of the line $p_2(x_2, y_2)$ moves one pixel at a time along the edge segment towards the other end of the edge segment until the maximum deviation reaches the threshold t_{dev} or reaches the end of the edge segment. The deviation dev of an edge pixel $p_{edge}(x_{edge}, y_{edge})$ is (3).

$$dev = \frac{|x_{edge}(y_1 - y_2) + y_{edge}(x_2 - x_1) + c|}{\sqrt{(x_1 - x - 2)^2 + (y_1 - y_2)^2}} \tag{3}$$

where $c = x_1 y_2 - x_2 y_1$. The maximum deviation of n edge pixels from a line is simply:

$$dev_{max} = \max(dev_1, dev_2, dev_3, ..., dev_n) \tag{4}$$

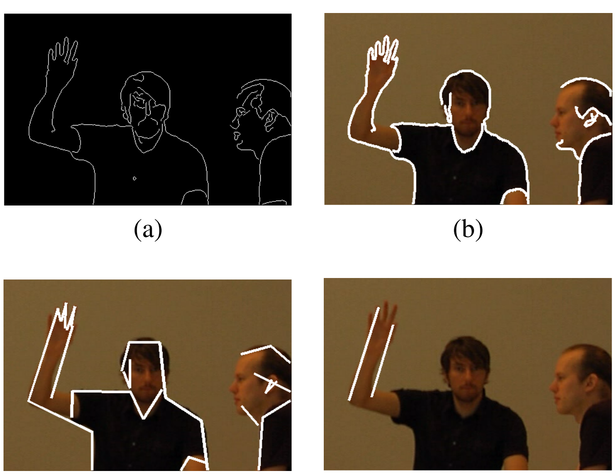
We experimentally selected the value of t_{dev} to be 20% of the width of the detected face w_{face}:

$$t_{dev} = 0.2 w_{face}. \tag{5}$$

When the maximum deviation reaches the threshold t_{dev}, fitting process for the current line stops and the fitting of a new line starts from the point where the current line ends. So, it is possible that one edge segment may result in more than one straight line for a given t_{dev}. The example output image of this process is illustrated in Fig. 2 (c).

C. Upright Arm Localization

Now, each ROI contains a set of n straight lines L_i where $i = 1, 2, 3, ..., n$. We compute the angle θ between each line and the x-axis. The absolute difference in θ between all of the lines in the ROI is computed. Only the line pairs that have a θ difference of less than 15 degrees are passed on to the next step. This rejecting of line pairs with a θ difference more than 15 degrees account for edges on each side of the arm being nearly parallel.



(a) (b) (c) (d)

Fig. 2. (a) ROI after edge detection, (b) ROI after edge linking, (c) straight line is fitted on linked edge in ROI, (d) only line segments on the edges of arm is left after filtering.

Moreover, location and orientation of two nearly parallel line segments formed by the edges on each side of the arms can be as shown in 3 (a) or (b). In the case of Fig. 3 (a), perpendicular projection from both end points of line segment L_1 falls on line segment L_2. But perpendicular projection of only one end point of L_1 falls on L_2 and that of L_2 falls on L_2 in the case of Fig. 3 (b). The value of d_proj is an average value of d_proj1 and d_proj2.

The perpendicular projection point p_{proj} on a line connecting between two end points p_1 and p_2 from a given point p_3 in x-y coordinate system can be found as follows:

$$\begin{aligned} p_1 &= [x_1, y_1] \\ p_2 &= [x_2, y_2] \\ p_3 &= [x_3, y_3] \end{aligned} \tag{6}$$

All three points p_1, p_2 and p_3 are matrices containing x and y coordinates.

$$t = \frac{v_{12}(p_3 - p_1)}{v_{12} v_{12}} \tag{7}$$

where $v_{12} = p_2 - p_1$. If the value of t is less than zero or greater than one, then no perpendicular projection point exists on the line. So, the value of x_{proj} and y_{proj} is set to infinity. Otherwise, the x_{proj} and y_{proj} can be computed.

$$p_{proj} = \begin{cases}]-\infty, +\infty[& \text{if } t < 0 \text{ or } t > 1 \\ p_1 + v_{12} t & \text{otherwise} \end{cases} \tag{8}$$

Once the perpendicular projection point p_{proj} is known, the euclidean distance between the point and one of the end points of the line can be computed. Then, we compute the

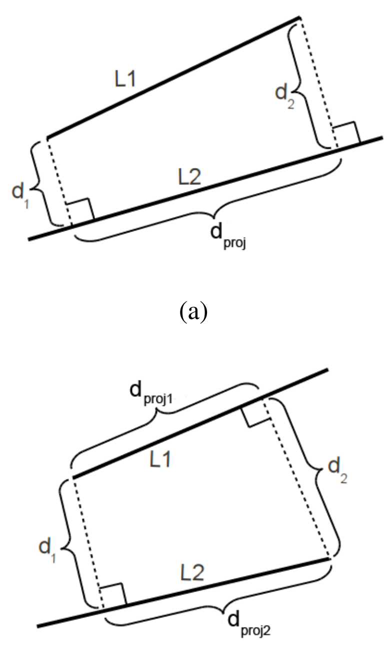


Fig. 3. Two possible location and orientation of line segments L1 and L2: (a) both end point of L1 can have perpendicular projection on L2, (b) one end point of L1 can have perpendicular projection on L2 and one end point of L2 can have perpendicular projection on L1.

d_proj for both cases: Fig. 3 (a) and (b). We experimentally found that the d_p computed from two lines on each side of the arm is usually larger than the width of the detected face w_{face}. Therefore, line pairs that have d_proj less than w_{face} are discarded.

Moreover, the distance between two lines $d_{between}$ also determines the possibility of these lines belonging to each side of the arm. The value of $d_{between}$ can be computed from d_1 and d_2 as shown in Fig. 3 using (9).

$$d_{between} = \frac{d_1 + d_2}{2} \tag{9}$$

The upper threshold $t^{upper}_{between}$ and the lower threshold $t^{lower}_{between}$ for the distance between two lines are experimentally determined as 75% of w_{face} and 25% of w_{face} respectively.

$$\begin{aligned} t^{upper}_{between} &= 0.75 w_{face} \\ t^{lower}_{between} &= 0.25 w_{face} \end{aligned} \tag{10}$$

The line pairs with $d_{between}$ not between $t^{upper}_{between}$ and $t^{lower}_{between}$ are rejected. The remaining line pairs are regarded as positive detections in which two lines in each pair represent the approximate location of the two sides of the arm. An example of a localized arm is shown in Fig. 2 (d).

D. Statistical Decision Making

This is the final stage of the system that makes a decision if a person is raising a hand in the video sequence based on the frequency of a raised hand detected within a particular ROI in the previous frames. In our method, a person is marked as a person raising a hand if there are at least seven raised hand detections in ten frames containing a face detected before the current frame. This step reduces the number of false positive detections. The output of this stage is the location a person's face who is detected as a person raising a hand. The higher-level applications such as a vision system of a robot may use this location information to instruct the robot which direction it must go to reach a person with a raised hand.

IV. Results and Discussions

The performance of our proposed method was tested on 17 video sequences containing people raising a hand 24 times. As a result, out of 24 hand raising cases, 22 cases were correctly detected as a person raising a hand, i.e. 92% sensitivity. There were two false positive detections, i.e. 92% selectivity, due to a background object which has an arm-like appearance. The hands-raising gesture detector in [9] achieves only 81% sensitivity and 84% selectivity as the evaluation was done on individual frame in seven video sequences. However, in this paper, we evaluate whether or not our method detects a hand-raising event based on the statistics over multiple frames.

The detail detection result for each video sequence is shown in Table I, NoRHE, NoTP, NoFN and NoFP stands for number of hand-raising events, true positive, false negative and false positive respectively. In video sequence number 11 and 12, four out of six of the hand raising cases are detected as true positives, indicating that the system performs quite well in the environment with varying lighting condition.

TABLE I

Detection results for each video sequence

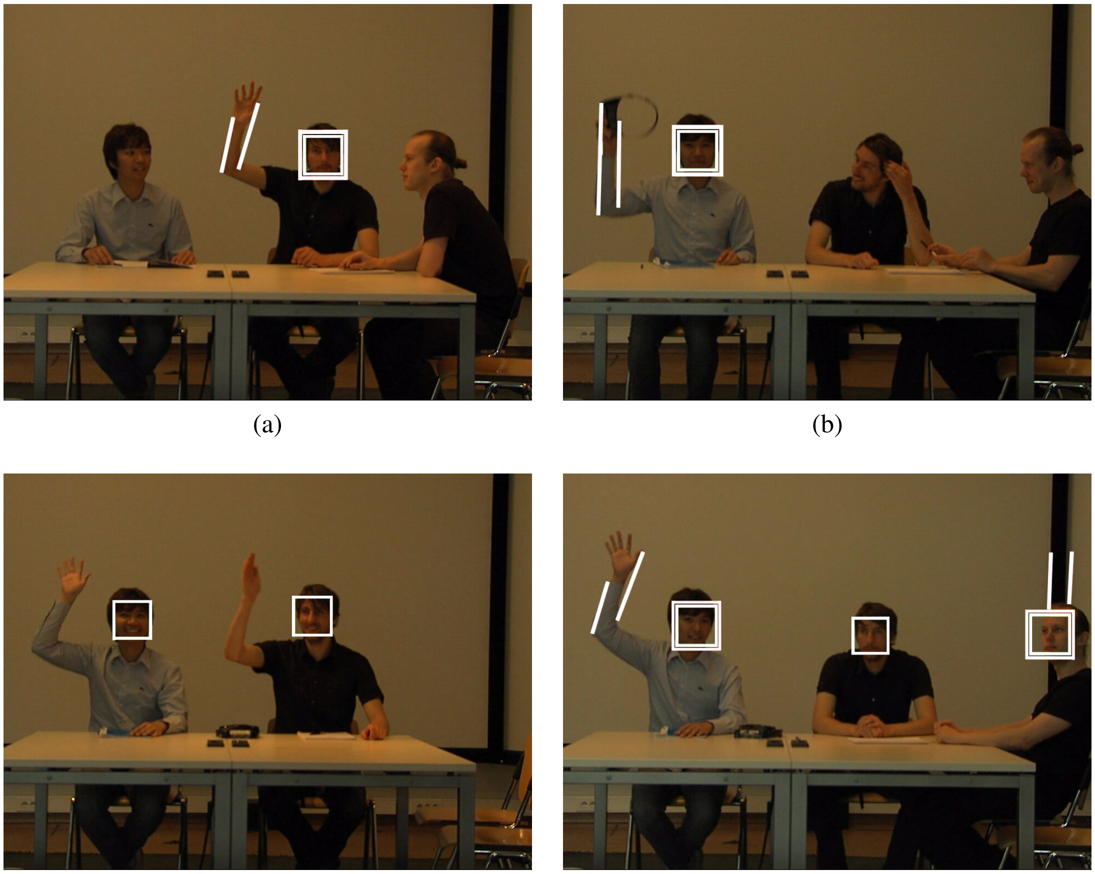
Sequence	# of Frames	NoRHE	NoTP	NoFN	NoFP
1	97	1	1	0	0
2	159	1	1	0	0
3	265	1	1	0	0
4	204	2	2	0	0
5	138	1	1	0	0
6	218	1	1	0	0
7	183	1	1	0	1
8	103	1	1	0	0
9	183	1	1	0	0
10	332	3	3	0	0
11	276	4	2	2	0
12	172	2	2	0	0
13	106	1	1	0	0
14	123	1	1	0	0
15	193	1	1	0	0
16	90	1	1	0	0
17	98	1	1	0	1
Total	2,940	24	22	2	2

The example detection results are illustrated in Fig. 4. The two white line segments on each side of an arm show the arm is detected. Single white bounding shows that the face is detected and double white bounding boxes show that the person is detected as a person who is raising a hand. The detection result in Fig. 4 (a) and (b) show an example of true positive detections. Moreover, Fig. 4 (b) shows that the system is able to locate the arm and detect a hand-raising gesture even though the person is holding an object. The other methods in which attempt to detect the shape of the hand may fail in this situation. The arm in Fig. 4 (c) is not detected due to the low contrast between the arm and the background. An example of false positive detection caused by background objects with arm-like appearance is shown in Fig. 4 (d).

From the analysis on the case in which hand-raising gesture are not detected, we found that the main reason for missing a hand-raising event is the failure of edge detection. Since the contrast between arm and background is quite low, the edges of the arm are not properly detected, resulting in short fitted lines that lead to arm localization failure. Using other more robust edge detectors may help improve the robustness.

V. Conclusion

In this paper, we propose a novel method for detecting a person who is raising a hand in video sequences with a 92% sensitivity and 92% selectivity. The method is relatively simple and computationally inexpensive so that it is possible to be used for real-time applications. Since the arm is located instead of locating a highly articulated hand, it is able to locate the arm even if the raised hand is holding an object. Moreover, our method does not require background modeling for foreground subtraction so that it can detect a raised hand without requiring background re-modeling.



(a) (b) (c) (d)

Fig. 4. Detection result examples: (a) a true positive detection result, (b) a true detection result in which a person is holding an object, (c) arms are not located due to camera noise and low contrast between arms and the background, (d) a true positive detection (leftmost) and a false positive detection (rightmost) on the background object with arm-like appearance.

The performance of the system was only evaluated on the person who is facing toward the camera so far. However, applications which make use of a multiple camera network, such as a smart meeting room, can overcome this view-independence problem by using an optimal camera selection algorithm [11] to find the camera with the best frontal view of each person. As future work, the performance of this method will be evaluated on video sequences with more people and a highly cluttered background. We will also find the way to integrate motion information between frames to discriminate edges of moving arms from edges of stationary background objects or body parts.

VI. Acknowledgment

This research project is supported by the Faculty of Technology and Environment, Prince of Songkla University (Phuket Campus), Phuket, Thailand jointly with Image Processing and Interpretation Research Group (IPI), Ghent University, Ghent, Belgium.

References

[1] J. Liu, J. Luo, and M. Shah, "Action recognition in unconstrained amateur videos," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1214–1220.

[2] L. Wang and L. Cheng, "Human action recognition from boosted pose estimation," in *Proceedings of the International Conference on Digital Image Computing and Applications*, 2010, pp. 308–313.

[3] B. Kapralos, A. Hogue, and H. Sabri, "Recognition of hand raising gestures for a remote learning application," in *Proceedings of the Eight International Workshop on Image Analysis for Multimedia Interactive Services*, 2007.

[4] M. Kolsch and M. Turk, "Robust hand detection," in *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 614–619.

[5] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2002.

[6] V. Spruyt, A. Ledda, and S. Geerts, "Real-time multi-colourspace hand segmentation," in *Proceedings of the 17th IEEE International Conference on Image Processing*, 2010, pp. 3117–3120.

[7] J. Yao and J. R. Cooperstock, "Arm gesture detection in a classroom environment," in *Proceedings of the 6th IEEE Workshop on Applications of Computer Vision*, 2002, pp. 153–157.

[8] X. Duan and H. Liu, "Detection of hands-raising gestures based on body silhouette analysis," in *Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics*, 2008, pp. 1756–1761.

[9] H. Liu, X. Duan, Y. Zou, and D. Gao, "Detection of hands-raising gestures using shape and edge features," in *Proceedings of the 2009 IEEE International Conference on Robotics and Biomimetics*, 2009, pp. 1480–1483.

[10] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proceedings of the 2002 International Conference on Image Processing*, 2002, pp. 900–903.

[11] M. Morbee, L. Tessens, H. Lee, W. Philips, and H. Aghajan, "Optimal camera selection in vision networks for shape approximation," in *Proceedings of the IEEE 10th Workshop on Multimedia Signal Processing*, 2008, pp. 46–51.