Modeling of Human Upper Body for Sign Language Recognition

Sara Bilal[1], Rini Akmeliawati[2], Amir A. Shafie, Momoh Jimoh El Salami
Department of Mechatronics Engineering
International Islamic University Malaysia (IIUM)
Jl Gombak 53100, Kuala Lumpur, Malaysia
smosb@hotmail.com [1] rakmelia@iium.edu.my[2]

***Abstract*- Sign Language Recognition systems require not only the hand motion trajectory to be classified but also facial features, Human Upper Body (HUB) and hand position with respect to other HUB parts. Head, face, forehead, shoulders and chest are very crucial parts that can carry a lot of positioning information of hand gestures in gesture classification. In this paper as the main contribution, a fast and robust search algorithm for HUB parts based on head size has been introduced for real time implementations. Scaling the extracted parts during body orientation was attained using partial estimation of face size. Tracking the extracted parts for front and side view was achieved using CAMSHIFT [24]. The outcome of the system makes it applicable for real-time applications such as Sign Languages Recognition (SLR) systems.**

Keywords: Human upper body detection; Scaling; Tracking using CAMSHIFT; Sign Language Recognition.

I. Introduction

Many Human to Computer Interaction (HCI) applications require the detection and tracking of HUB parts such as human activity analysis, in which human motion and gestures are detected, tracked and recognized from cameras. Many systems have been developed to detect HUB parts but the challenge is to find effective methods for detecting, labeling and tracking the HUB parts for a particular application. The developed system for HUB detection should be robust to avoid occlusion and to recover easily from tracking errors when the object is not in the scene. Therefore, for applications such as Sign Language (SL), it is very important to understand the working environment. Lighting condition in indoor and outdoor environment is very crucial in determining the recognition system specifications. Also for SL, hand location with respect to head, shoulder and chest convey a lot of meaning.
These parts can be used as reference for static and dynamic gestures see Figure 1.

Figure 1. A Sign stands for Islam race from MSL database

In this work, a new approach has been introduced to detect HUB parts. Features have been extracted to segment HUB and ellipses were fitted to each segment.
The paper is organized as follows; Section II presents the research background. The proposed method for modeling the HUB parts will be described in Section III. The experimental results are presented in Section IV. Section V states the conclusion.

II. Research Background

Many approaches have been developed for human pose recognition using computer vision and pattern recognition approaches. Hyeon [1] built a model of HUB by segmenting the upper body regions of humans in images and then compared the edges in the image with a predefined curvature model. Another HUB modeling approach using Ω model was developed by [2, 3]. Their system makes use of Ω model which has different scales to describe the HUB. Nicolas Burrus and Justus Piater [4] have developed top-down approaches based on pictorial models for Dutch Sign Language. The approach simultaneously models the geometry of human parts, the appearance of each part and the temporal continuity in a unified statistical framework. The modeling methods include the use of template-based distance measures, such as Chamfer Distance as in [1] and parameterized approaches, such as Mixture Density Models. David et al. [5] continuously segment the image using stereo range. They build a model for human torso and Kalman

Research Matching Grant Scheme RMGS 09-03,
International Islamic University Malaysia (IIUM).

filter was used for tracking. Build a model as presented in the previous approaches to represent the HUB can have high detection rate but slow the overall computational process time. Another work using stereo vision was presented by Darrell et al. [6]. Their system has been developed using skin colour and face position as a starting point to identify human pose.

Far from modeling the HUB, Govindaraju et al [7-10] has detected head-and-shoulder based on the geometry properties of face profile. Prior information was required and their system has a high error rate. In reference [11] an Adaptive Combination of Classifiers (ACC) has been built to detect head-and-shoulder and partially occluded people in a static image. Yi Sun et al. [12] have proposed a head-and-shoulder detection algorithm based on wavelet decomposition technique and support vector machine (SVM). Using such techniques in low level feature extraction may slow the overall process [11-12].
The goal of this work is to detect HUB parts in a sequence of video frames for real-time applications. The HUB detection system must be fast and robust to fulfil the requirements of real-time applications.
Thus, this work presents a system that is able to detect the HUB parts for real-time applications such as SLR system using a fast search algorithm as shown below.

III. Modeling the Human Upper Body

Many existing methods which use geometric modeling, boosted classifiers and SVM have been introduced [7-12]. But real-time applications, such as SL recognition systems, require a fast, compatible and synchronized method for HUB parts detection and tracking. Therefore, a fast and robust search algorithm for HUB parts based on the study in [13] has been introduced as main contribution in this work. It assumes that all body parts can be measured with head size. Face location can be used as an initialization for the system. Head size differs from face size with only few measurements as shown in Figure 2. The head measurement is found first by dividing the head into half, then, defining that 1/3 below the nose is the lip. By using that ratio, one can estimate that forehead is 2/3 the upper pupil half. Then 1/3 above the forehead is the hair end. The algorithm for finding head measurements is as follows

- The pupils are in the middle of the head, top to bottom equal to *1/2*x*h*.
- The bottom of the nose is between the pupils and bottom of the chin.
- Below the nose and between the lips *(NL) = 1/3*x*1/2*x*h*.
- 2/3's below the nose is the chin crease.
- Forehead size *(Fs) =1/3*x*1/2*x*h*.

where *h*=head height and *w*=head width.

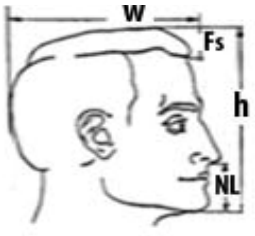


Figure 2. Head and face dimensions.

The structure of our system in which a face and HUB parts are detected in an image is shown in Figure 3.

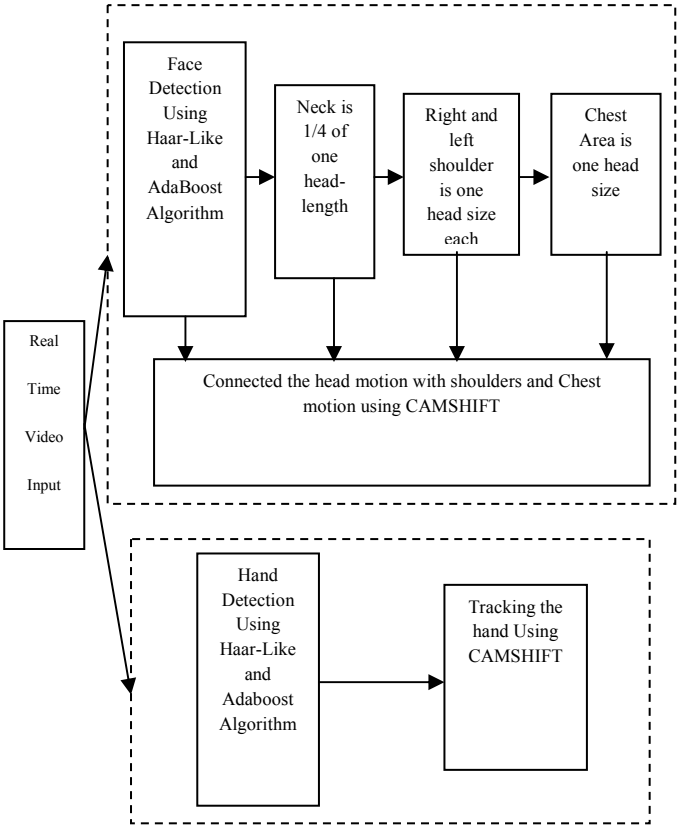


Figure 3. Human Upper Body (HUB) parts detection

A. Head and Face Detection with Haar-like Features

Initialising the system with a face will provide an adequate search region for other HUB parts. Neural network methods, support vector machines and other kernel methods have been used for face posture detection [14, 15, 16, 17]. However, most of these algorithms use raw pixel values as features which make the system sensitive to noise and changes in illumination. Instead, other approaches such as Haar-like features, which are similar to Haar basis functions has been proposed by [18]. The features encode differences in average intensities between two rectangular regions, and they can extract textures independent on absolute intensities. Our proposed method uses Haar-like features and AdaBoost algorithm for face or hand detection.

-Haar-Like Feature Family and AdaBoost Learning Algorithm

The simple Haar-like features (so called because they are computed similarly to the coefficients in the Haar wavelet transform) are introduced by [18, 19]. There are two motivations for the employment of the Haar-like features

rather than raw pixel values. The first motivation is that the Haar-like features can encode ad-hoc domain knowledge, which is difficult to describe using a finite quantity of training data [20]. Compared with raw pixels, the Haar-like features can efficiently reduce (increase) the in-class (out-of class) variability and thus, make classification easier [21].

The second motivation is that a Haar-like feature-based system can operate much faster than a pixel-based system. AdaBoost learning algorithm is a method to improve the accuracy based on a series of weak classifiers stage–by-stage [21]. Initially, it maintains a uniform distribution of weights over each training samples. In the first iteration, the algorithm trains a weak classifier using one Haar-like feature that achieves the best recognition performance for the training samples. In the second iteration, the training samples, which were misclassified by the first weak classifier receive higher weights so that the newly one is selected. The iteration goes on and the final result is a cascade of linear combinations of the elected weak classifiers, i.e. a strong classifier, which achieves the required accuracy. Therefore, the above technique is used in the proposed method to detect the face region which has reduced the image processing time. After the face region has been detected in the image, other HUB regions were found as described next in Section III B.

B. Shoulders and Chest Detection

HUB parts such as neck, shoulders chest and hands are very crucial to understand human gestures. The detected face using Haar-like and AdaBoost algorithm has been used as an initial position to find these HUB parts. A study in [13] shows that the human body can be measured using head dimension. A human figure adjusted for artists based on accurate 8 head size is shown in Figure 4. It assumes that:

1. The neck space is 1/4 of one head-length, and it starts under the chin of that top first head.
2. The second head is the shoulders head. It is the top of three trunk heads and is drawn under this neck space.
3. One quarter of one head down in this second head is the shoulder line. This allows space for the neck-support muscles above the clavicle.
4. This shoulder line is two head-lengths (two widths on a female) wide and is the top line of the *torso triangle* that extends down to the space between the legs, or the *chest triangle* that only extends down to the hip line.

Hands are not included in the assumptions above. A hand has been detected using the Haar-like and AdaBoost approach which has been used for face detection. The system has been trained for hand detection using specific hand shape. After all the HUB parts have been detected in the first subsequent of frames, tracking was done using Continuously Adaptive Mean Shift (CAMSHIFT). This will be described in Section III C.

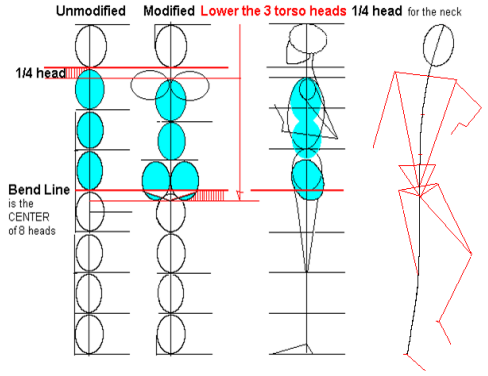


Figure 4. Accurate 8-head-high adult male human figure, adjusted for artists. [13]

C. Tracking the whole HUB Parts

Pose detection (or initialisation) is typically performed on an initial video frame followed by pose tracking where the pose parameters obtained from the current frame is used as a starting value for the subsequent video frame. Many methods exist to track pose using motion histories [22] and optical flow [23] technique. CAMSHIFT [24] is designed for face and coloured object tracking. It is adapted from Mean Shift Algorithm (MSA) which uses probability distribution. The colour histogram is used to build the probability distribution of the object in the video frame sequences. When the colour distribution of the object changes over time, the MSA has to be modified. Then, CAMSHIFT can adapt the system dynamical to achieve the tracking of the object over the video frame sequences.

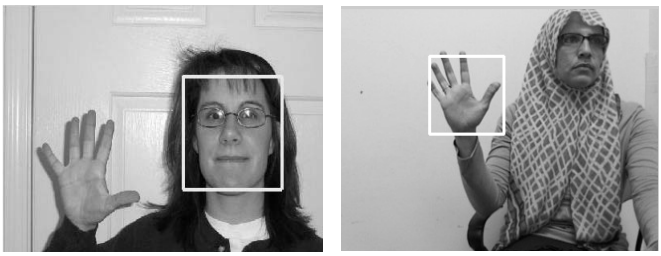
IV. Experimental Results for Detecting the Human Upper Body

Detecting human upper body for SL was achieved using VC++, OpenCV and the assumptions stated in Section III B.

A. Head and Hand Detection

First, the system has used Haar-like feature for extracting information from the face and hand using two line features, four edge features, one centroid feature and one diagonal feature, followed by AdaBoost algorithm for learning. After the face and hand have been detected, see Figure 5 (a, b), we have obtained the head size from the face by finding Equation (1):

$$\text{Head size} = \frac{1}{3} \times \left(\frac{facesize}{2} \right) + facesize \qquad [1]$$



a b

Figure 5. Haar-like and AdaBoost Algorithm used for detecting a. Face b. Hand

B. Neck, Shoulder and Chest Detection

The HUB parts were located using the head size algorithm as stated in Section III and shown in Figure 4. When the human stands in a side view, see Figure 6, the dimensions of the oriented head and face are not following the other HUB part orientation. In such case, first, the system calculates the size of the head in a front view. Then, if the size is reduced to 1/3 of the head size, the other detected HUB parts will keep its preliminarily size. This is done because for SL recognition systems, the whole HUB parts must be completely covered. Signs could be started or end from/at head, shoulder or chest.

Figure 6. A Sign stands for Hindu race from MSL database

C. Tracking the HUB Parts

Since all the HUB parts have been detected, as shown in Figure 7, tracking was achieved using CAMSHIFT. Hand colour histogram will be initiated separately from head and other HUB parts. This is due to the fast movement of hand gesture which will have rapid colour updating overtime. So two CAMSHIFT trackers have been used, one for the hand and another tracker for the face motion connected to the other HUB parts.

Figure 7. Shows the detected and tracked HUB parts.

The overall process of the hand, head, neck, shoulders, chest detection and tracking algorithm is shown below.

1. Detect face and hand.
2. Getting the head size following equation [1].
3. Neck space is 1/4 of the head.
4. Down to neck end, one head is the chest.
5. The right shoulder is one head right.
6. The left shoulder is one head left.
7. Initiate tracker for hand using CAMSHIFT.
8. Initiate tracker for face and other HUB using CAMSHIFT.

V. Conclusions & Future Work

The location of hand with respect to the head, and other HUB parts conveys a lot of meanings to understand SL. Therefore, tracking HUB parts as well as scaling the size while rotating is an important issue while classifying hand gestures. Many methods have been developed for HUB detection and tracking but real-time application requires fast computational time as well as the storage requirements is extremely important. In this paper, a fast and robust HUB parts detection algorithm has been introduced to detect and track the HUB parts. These performances make it applicable for applications such as SL recognition systems.
The assigned locations of the hand with respect to HUB will be used to classify gestures for developing an Automatic Malaysian Sign language Translator (AMSLT) system.

References

[1] D. H. Hyeon et al. “Human Detection in Images Using Curvature Model”. International Conference on Circuits/Systems Computers and Communications (ITC-CSCC), 2001.
[2] A. Broggi et al. “Shape-based Pedestrian Detection”. Proceedings of the IEEE Intelligent Vehicles Symposium, pp. 215-220, 2000.

[3]. D. Beymer, K. Konolige. “Real-time tracking of multiple people using continuous detection”. Proceedings of IEEE International Conference on Computer Vision, 1999.
[4] Nicolas Burrus and Justus Piater, “Monocular human upper body pose estimation for sign language analysis”, Talk at the 4th Multitel Spring Workshop, Mons, Belgium, 2009.
[5] David Beymer and Kurt Konolige. “Real-Time Tracking of Multiple People Using Continuous Detection”. International Conference on Computer Vision (ICCV) Frame-rate Workshop, 1999.
[6] T. Darrell, G. Gordon, M. Harville, and J. Wood_ll. “Integrated person tracking using stereo, color, and pattern detection”. In Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, pp. 601-608, 1998.
[7]. Venu Govindaraju, Sargur. N. Srihari, David B. Sher. “ A computational model for face location”.Proceedings of the IEEE Third International Conference on Computer Vision, pp. 718-721, 1991.
[8] Venu Govindaraju, David B. Sher, Rohini K. Srihari. “Locating human faces in newspaper photographs”. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 549-554, 1989.
[9] Venu Govindaraju, Sargur. N. Srihari, DavidS her. “A computational model for face location based on cognitive principles”. Proceedings of the American Association for Artificial Intelligence (AAAI), pp. 350-355, 1992.
[10] Venu Govindaraju. “Locating human faces in photographs”. International Journal of Computer Vision, Vol.19. pp. 129-146, 1996.
[11] A. Mohan, C. Papageorgiou, T. Poggio. “Example-based object detection in images by Components”. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23. , pp. 349-361, 2001.
[12] Yi Sun, Yan Wang, Yinghao He and Yong Hua. “Head-and-Shoulder Detection in Varying Pose”, Advances in Natural Computation, First International Conference, ICNC 2005, Changsha, China, Proceedings, Part II, 2005.
[13] Full Real Color Wheel Course, (July, 2011). Retrieved from http://www.realcolorwheel.com/human.htm.
[14] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. IEEE Trans. on PAMI, 20(1):39–51, 1998.
[15] H. A. Rowley, S. Baluja, and T. Kanade. Neural network based face detection. IEEE Trans. on PAMI, 20(1):23–38, 1998.
16] E. Osuna, R. Freund, and F. Girosi. Training support vectormachines: an application to face detection. Proc. of CVPR, pages 130–136, 1997.
[17] B. Heisele, T. Poggio, and M. Pontil. Face detection in stillgray images. A.I. Memo, (1687), 2000.
[18] P. Viola, M. Jones. “Robust Real-time Object Detection”. Cambridge Research Laboratory Technical Report Series CRL2001/01, pp. 1-24, 2001.
[19] Qing Chen, Nicolas D. Georganas, Emil M. Petriu: “Real-time Vision-based Hand Gesture Recognition Using Haar-like Features”. Instrumentation and Measurement Technology Conference – IMTC , 2007.
[20] R. Lienhart, J. Maydt. “An extended Set of Haar-like Features for Rapid Object Detection”. Proc. IEEE International Conference on Image Processing ICIP, Vol. 1, pp. 900-903, 2002.
[21] Y. Freund, R. E. Schapire. “A Short Introduction to Boosting”. Journal of Japanese Society for Artificial Intelligence, Vol. 14(5), pp. 771- 780, 1999.
[22] Gary R. Bradski1 and James W. Davis. “Motion segmentation and pose recognition with motion history gradients”, Machine Vision and Applications 13: 174–184, 2002.
[23] Vincent van Megen . “3D pose tracking using optical Flow”, thesis, 2010.
[24]Gary R. Bradski, “Computer Vision Face Tracking For Use in a Perceptual User Interface”. Intel Technology Journal, Vol. 2 (2), pp. 12–21, 1998.