2D Upper Human Body Pose Modelling Using Windowing and Template Matching Techniques

R. R. Porle[1], A. Chekima[1], F. Wong[1], G. Sainarayanan[2]
[1]School of Engineering and Information Technology
Universiti Malaysia Sabah, Locked Bag No. 2073, 88999 Kota Kinabalu, Sabah, Malaysia
[2]ICT Academy of Tamil Nadu, ELCOT Complex, 2-7 Developed Plots,
Industrial Estate, Perungudi, Chennai – 600 096, India.
rosalynrporle@yahoo.com.my

***Abstract*- This paper presents the upper human body pose modelling using windowing and template matching techniques. Indoor image sequences are used as input. Silhouette and colour features are extracted in the human body parts detection stage. The head and torso pose are estimated using windowing technique. Meanwhile the upper and lower arms are estimated using template matching technique. Each body part is represented by rectangular shape where its sizes depend on the predetermined approximation. The proposed techniques are tested using 30 image sequences from different users. These techniques can overcome self-occlusion, hand crossing and it is adaptive to illumination changes.**

I. INTRODUCTION

2D human body pose modelling is a system that detects the human body parts, estimates their posture and then models them in an image plane using specified shape. It is an active and growing research area in the last few decades; motivated by the need of man-machine systems to detect, recognize, track, model and analyze the human as automatically as possible in order to interact effortlessly with human or user. The human body pose modelling is applicable in various areas such as gesture identification for user interface systems, person tracking for intelligent visual surveillance systems and motion analysis in the sport and medical fields.

The technique of modelling the human body pose is directly influenced by the type of image feature to be used, its model representation and also the application of the modelling system. The human body detection usually involves background modelling, image segmentation and object classification. Description about these stages can be found in [1]. Image features such as appearance, silhouette, edge, colour, depth, and also optical flow can be used either independently or combined for detecting the targeted body part. For 2D models, the rectangle shape [2] and elliptical shape [3] are commonly employed for representing the body parts. The pose estimation can be performed by fitting the constructed model to the image feature. Statistical approach [4-6] and filtering approach [7, 8] are two general approaches that commonly used for the estimation procedure.

In this paper, the windowing and template matching techniques are proposed for the pose estimation of the upper human body in indoor environment. These techniques are part of the filtering approach. The body parts to be estimated include the head, torso, upper arms and lower arms. Each body part is represented by a rectangular shape, where its sizes depend on the predetermined approximation. Fig. 1 shows the diagram of human body and the 2D model of the upper human body parts. From this figure, the head model covers the head and neck of the human. The torso length is between the shoulder and lower hip of the human.

This paper is divided into five sections. Section I presented the introduction of the 2D human body pose modelling. Section II describes the human body parts detection where silhouette and colour features are extracted. The pose estimation techniques for the head, torso and arms are presented in section III. The windowing technique is proposed for the head and torso pose estimation. Meanwhile the template matching technique is employed for arms pose estimation. The results and discussion are presented in section IV and lastly the conclusion is presented in section V.

II. HUMAN BODY PARTS DETECTION

Human body parts detection is the fundamental stage in the human body pose modelling system. The main idea in this stage is that; given a raw image sequences acquired from a digital video camera, process these images so that only the human features are highlighted in a binary form. The input

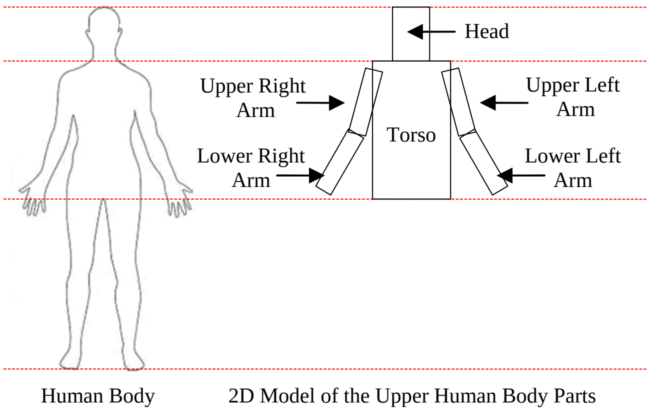

Fig. 1. The diagram of human body and 2D model of the upper human body parts

images are in RGB colour format. Approximation images are extracted from these input images using Haar wavelet transform at level one [9]. This transformation shrinks the input images and at the same time smoothes their texture. The approximation effect also reduces the lightness variations in the images. Two features namely silhouette and colour features are extracted from the transformed images. Silhouette feature is used to detect the human silhouette in the images and colour feature is used to extract the occluded arms within the torso region. The description for the human silhouette and colour extraction is as follows:

A. Human Silhouette Extraction

The human silhouette can be extracted using background subtraction. Colour information is not essential in this stage; therefore, the transformed images are grey scaled using (1).

$$I_G = 0.229R_W + 0.587G_W + 0.114B_W \tag{1}$$

The grey scaled image is denoted as I_G. Meanwhile the R, G and B colour components of the approximation images are denoted as R_W, G_W and B_W respectively.

In the background subtraction, two images namely background image, I_{GBI}, and foreground image, I_{GFI}, are subtracted. The absolute differences of these images are thresholded to obtain a binary image. The background image is the reference image that contains only background scene. The foreground image, on the other hand, is the image that contain human in the similar scene. The background subtraction equation is given as

$$I_{BS} = \left| I_{GFI}(f+1) - I_{GBI} \right|; f = 1,2,3,...,F \tag{2}$$

where F is the maximum number of frame in the image sequence. The threshold value for the background subtracted image, I_{BS}, is obtained using (3), where R_{BS} is the ratio between 0 and 1.

$$T_S = R_{BS} \times \max(I_{BS}) \tag{3}$$

The image is then thresholded based on the optimal threshold value, which is evaluated from the experimental stage. This is expressed as

$$I_T = \begin{cases} 1, & I_{BS} \geq T_S \\ 0, & I_{BS} < T_S \end{cases} . \tag{4}$$

In order to remove the scattered ON pixels in the thresholded image, I_T, morphological opening is performed [10].

B. Occluded Arms Extraction

Estimating the arms pose is a challenging task as the arms tends to move farther, faster and more often than the other body parts. Silhouette feature may not be sufficient and effective for arms detection especially when the arms were occluded in the head or torso regions. In this system, the colour extraction is performed if the arms are occluded within the torso region. Colour is a useful feature for arms detection. It provides computationally effective yet, robust information against rotation, scaling and partial occlusions [11]. Performing skin colour segmentation for arms detection in account of all variation such as race factors, illumination, and blurriness due to fast movement is difficult and time consuming. Several issues have to be considered. This includes the choices of colour space to be implemented and how the skin colour distribution should be modelled. In the human body pose modelling, this task is relatively easy since the skin colour template can be constructed specific to the human subject considered in the system [12]. The skin colour distribution in the face region can be referred and analyzed to detect the arms regions. The colour space of the approximation images are transformed into CIELAB colour space. The new colour space image is then separated into colour components. For example, the RGB colour image that is transformed into CIELAB colour space can be separated into L_{CIELAB}, A_{CIELAB} and B_{CIELAB} component respectively. In this work only the $\mathrm{A}_{\mathrm{CIELAB}}$ and $\mathrm{B}_{\mathrm{CIELAB}}$ component are utilized. The head region is extracted in the head and torso pose estimation stage. The colour information in the head region is extracted and its colour distribution is analyzed using histogram. From the histogram distribution, it is easier to determine the suitable range of the skin colour for that human. The histogram is obtained by counting the occurrence of the colour pixel value in each of the colour bin. Each of the colour components provides different range of colour histogram distribution. The occurrence of the colour pixel is referred as the frequency. The maximum frequency, f_{max}, in the histogram is obtained. Then using this value, the required frequency in the head region that can be assumed as minimum skin colour occurrence, f_{creq}, in the head region is denoted as

$$f_{creq} = f_{\max} \times R_f \tag{5}$$

The R_f is the predefined ratio for the colour component that is set between the range of 0 and 1. The range where the skin colour value is above the f_{creq} is referred as the skin colour range. From this range, two thresholds can be extracted. The low threshold, denoted as T_{CL}, is the minimum colour value in the skin colour range. Meanwhile, the high threshold, denoted as T_{CH}, is the maximum colour value in the skin colour range. Once the thresholds are obtained, the skin region of the human can be extracted. The colour components is thresholded using the following equation

$$I_C = \begin{cases} 1 & if\ T_{CL} \leq I_C \leq T_{CH} \\ 0 & else \end{cases} \tag{6}$$

where I_C is the colour component used in the work. To obtain the finalized skin region, the colour components are combined using logical AND operator. The resultant image is then superimposed with the silhouette image to exclude the ON pixels in the background. After that, the ON pixels in the head region is set to OFF pixels so that only the occluded arms regions is remained in the image. Finally morphological opening is performed to smooth the contour of the occluded arms regions.

III. POSE ESTIMATION

A. Head and Torso Pose Estimation

The head and torso pose can be estimated from the silhouette image. Each body part is modelled by rectangle shape, whereby its width and height are calculated using distance transform. The models are parameterized with a vector of *rXi* and *cXi*, where *r* and *c* are the row and column positions of the body part corners respectively, *X* is the targeted body part and *i* is a reference number. Fig. 2 shows the head and torso models along with their corner's variables.

Let DT_{max} denotes the maximum pixel value of the complement distance transform image. The torso width, $T_{width,}$ is computed as

$$T_{width} = R_{TW} \times DT_{max} \tag{7}$$

where R_{TW} is the ratio for the torso width. Using T_{width} value, the respective head width, H_{width}, is obtained using (8).

$$H_{width} = R_{HW} \times T_{width} \tag{8}$$

R_{HW} is the ratio for the head width. The values of R_{TW} and R_{HW} above are computed experimentally. After that, the head height, H_{height}, is computed as

$$H_{height} = 1.7 \times H_{width} \tag{9}$$

From (8) and (9), the head area, H_{area}, is calculated as

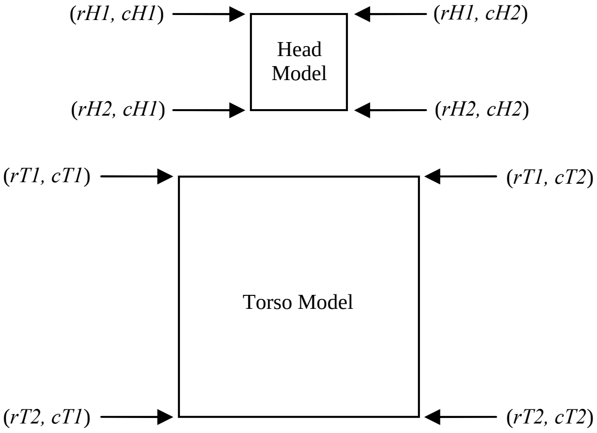

Fig.2. Diagram of the head and the torso models along with their corner's variables

$$H_{area} = H_{height} \times H_{width} \tag{10}$$

To locate the head, the half top of the human silhouette region in the silhouette image is divided into small windows and then the total sum of ON pixels in each window is computed. This is given as

$$H_{TP}(r,c,:) = \sum_{r}\sum_{c} I_s\left(r{:}r + H_{height}, c{:}c + H_{width}\right) \tag{11}$$

where *r* and *c* are the row and column positions of the half top of the human silhouette in the image. The reference point, (*rH1*, *cH1*) of the head is determined from the first element in (11) that is more than 80% of the H_{area}. Then *rH2* and *cH2* can be computed using (12) and (13).

$$rH2 = rH1 + H_{height} \tag{12}$$

$$cH2 = cH1 + H_{width} \tag{13}$$

For locating the torso, the initial torso height is set as

$$T_{height} = rH2 + r_{max} \tag{14}$$

where r_{max} is the bottom position of the human silhouette. The approximated torso area is then given as

$$T_{area} = T_{height} \times T_{width}\,. \tag{15}$$

The same windowing concept used in head pose estimation can be applied for the torso part. Total sum of ON pixels in the specified window for the torso is given as

$$T_{TP}(r,c,:) = \sum_{c} I_s\left(rH2{:}r_{max}, c{:}c + T_{width}\right) \tag{16}$$

The reference point (*rT1*, *cT1*) of the head is determined from the first element in (16) that is more than 80% of the T_{area}. Then *rT2* and *cT2* are calculated as

$$rT2 = rT1 + 0.5 \times T_{height} \tag{17}$$

$$cT2 = cT1 + T_{width}\,. \tag{18}$$

B. Arms Pose Estimation

The arms are divided into right and left arms. Each side is further subdivided into upper and lower arms. The arms pose estimation is performed on the upper right arm first. This is followed by the lower right arm, the upper left arm and lastly the lower left arm. Each part is represented by a rectangle and this rectangle is parameterized by five points, P_{Xi}, and a rotational angle, θ_X. *X* and *i* are the targeted body part and the reference number respectively. Four points indicates the

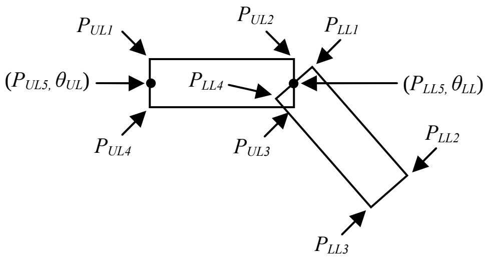

Fig.3. Diagram of the left arm models along with their parameters

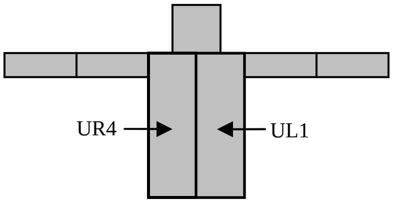

Fig.5. Diagram of the occluded upper arms regions

corners of the rectangle and the other point is the origin axis for the arm rotation. Fig. 3 shows the model of the left arm. The arms pose estimation is based on the template matching technique. It can be divided into six steps as follows.

1) Classify the Arms Pose: The arms pose can be classified either as non-occluded or occluded. The non-occluded arm is located outside the torso region. Meanwhile, the occluded arm is located within the torso region. The classification is performed by evaluating the length of the ON pixels, which are located outside the head and torso region. If the length is more than a certain ratio, the arm is classified as non-occluded pose. Otherwise, the arm is classified as occluded pose. If the arm pose is classified as occluded, the occluded arm extraction, as explained in section II, is performed before the next step.

2) Determine the Arms Size: The arms size is approximated based on the torso size. In this case, the height, A_{height}, is approximately 25 percent of the torso height and the width, A_{width}, is approximately 35 percent of the torso width.

3) Targeted Arm Extraction: For the right arm pose estimation, the right arm has to be extracted from the feature image. The pixels in the head, torso, and left arm regions are set to OFF pixels. For the left arm pose estimation, the left arm is extracted and the pixels in other regions are set to OFF pixels.

4) Arms Regions Classification: The arms regions classification is developed to minimize the search area for the targeted arm and each classified regions will influenced the range of rotational angle to be used in the pose estimation stage. The regions for the non-occluded upper right arm detection are denoted as UR1, UR2 and UR3. Meanwhile, the regions for the upper left arm detection are denoted as UL2, UL3 and UL4. The top torso corners are the reference points for dividing the regions. The size for UR1, UR2, UL3 and UL4 is given as A_{height} × A_{height}. As for the UR3 and UL2 regions, the height is given as A_{height} and its width is the range between the top torso corners to the centre of the head. The location of all the regions is illustrated in Fig. 4.

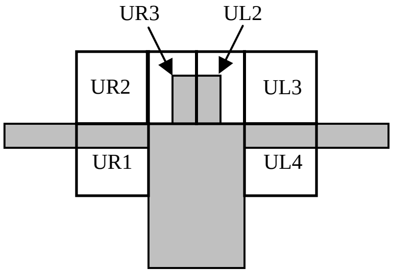

Fig.4. Diagram of the non-occluded upper arms regions

In the case of occluded arms, regions UR4 and UL1 are assigned for the upper right arm and upper left arm. These regions are illustrated as in Fig. 5.

For the lower right arm, the regions are denoted as LR1, LR2, LR3 and LR4. Meanwhile, for the lower left arm, the regions are denoted as LL1, LL2, LL3 and LL4. The size for all regions is given as A_{height} × A_{height}. The location of all the regions is illustrated in Fig. 6. The origin axis of the lower arm is set as the reference point for the region division. Therefore, depending on the poses of the targeted lower right arm, the location of the region division is also varied.

5) Determine the Rotational Angle: The range of rotational angle that will be used in the pose estimation stage is influenced by the location of the specified regions. For regions UR1, UL1, LR1 and LL1, the rotational angle is set between 0° to 90°. For regions UR2, UL2, LR2 and LL2, the rotational angle is set between 90° to 180°. For regions UR3, UL3, LR3 and LL3, the rotational angle is set between 180° to 270°. Finally for regions UR4, UL4, LR4 and LL4, the rotational angle is set between 270° to 360°. For each arm part, the total ON pixels in the defined regions is compared. The region that has the maximum total ON pixels is selected and its range of rotational angle is used for the pose estimation.

6) Pose Estimation: The upper right arm is first estimated by rotating the constructed template along the origin axis. For every rotational angle, the template image is superimposed with the feature image and these two images are then multiplied. The multiplication result indicates how well the template correlates with the feature image. High value of the multiplication result shows that the template is highly matched with the feature image in terms of position. On the other hand, low value of the multiplication result shows that the template is poorly matches with the feature image. To determine the

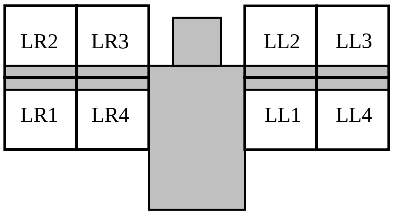

Fig.6. Diagram of the lower arms regions

best angle that gives the highest matched position, the highest total pixel value among the computed angles is selected. Once the desirable angle is selected, the four corners of the finalized template can be extracted for creating the finalized model. After the position of upper right arm is computed, the pose of the lower right arm is estimated. Before the pose can be estimated, the ON pixels of the upper arm in the feature image are set to OFF pixel to avoid it to be overlapping with the template for the lower arm. To rotate the template of the lower arm, the origin axis for the lower arm can be computed using the corner points of the upper arm. The pose estimation procedure is similar as the pose estimation of the upper right arm. The only difference is the region to be evaluated; in this case four regions are involved. After the pose of right arm is estimated, the next step is to estimate the pose of the left arm. Similar procedure described earlier is applied with the difference of the location of the region to be evaluated.

IV. EXPERIMENT RESULTS

The human body pose modelling described previously is applied to image sequences captured in indoor area. The video camera model is SONY DCR-PC115E PAL and the size of the input images is 480 × 640 pixels. Few conditions are set in the experimentation. The first condition is that the user must be standing upright and facing the video camera for the pose to be correctly estimated. The second condition is that the user wears short sleeve shirt to enable occluded arms to be detected. The proposed techniques were tested using 30 image sequences from different users. Fig. 7 shows the results of human body pose estimation in an image sequence. The rectangle model is superimposed with the grey scaled image to demonstrate the accuracy of the pose estimation. From the results, it is shown that the proposed techniques are able to estimate the pose if most of the targeted body part inside the specified rectangle shape.

V. CONCLUSION

This paper presented an approach of estimating the upper human body part using windowing and template matching techniques. A rectangle shape is used to represent the head, torso, upper and lower arms of the human. Silhouette feature is used to estimate the head, torso and non-occluded arms. Meanwhile, colour feature is used to estimate the occluded arms. The proposed techniques are adaptive to illumination changes and overcome self-occlusion and also hand crossing. Future work includes eliminating the two conditions that has been applied in the current system.

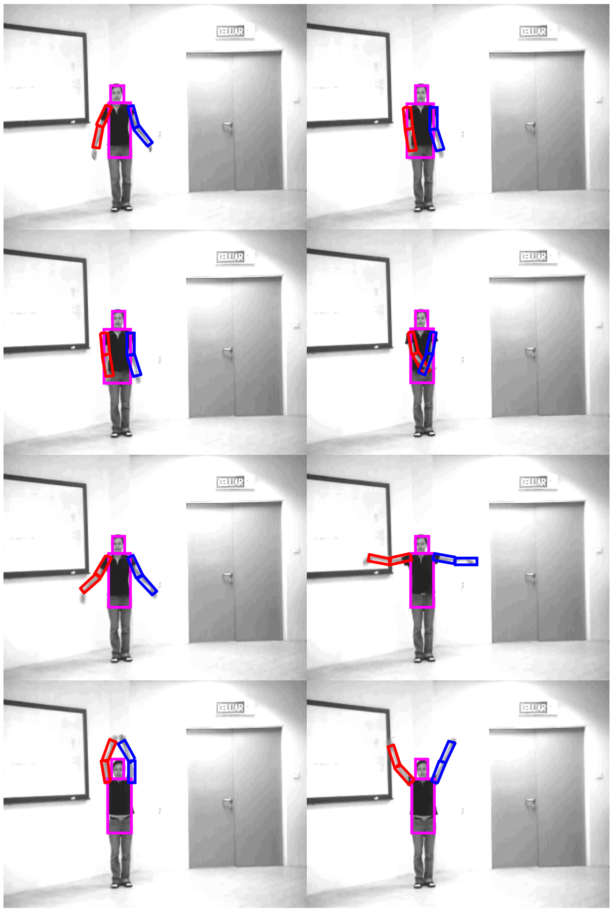

Fig.7. Results of human body pose estimation in an image sequence

REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, “A Survey on Visual Surveillance of Object Motion and Behaviors” IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, vol. 34(3), pp. 334–352, 2004.

[2] S. X. Ju, M. J. Black, and Y. Yacoob, “Cardboard People: A Parameterized Model of Articulated Image Motion” in Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition, Killington, Vermont, 1996, pp. 38–44.

[3] I. Haritaoglu, D. Harwood, and L. S. Davis, “W: Real-Time Surveillance of People and Their Activities” IEEE Transaction on Pattern Analysis Machine Intelligent, vol. 22, pp. 809–830, 2000.

[4] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-Time Tracking of the Human Body” IEEE Transaction on Pattern Analysis and Matching Intelligence, vol. 9(7), pp. 780–785, 1997.

[5] L. Zhao, and C. Thorpe, “Recursive Context Reasoning for Human Detection and Parts Identification” in Proc. Of IEEE Workshop on Human Modeling, Analysis, and Synthesis, 2000.

[6] G. Hua, M. H. Yang, and Y. Wu, “Learning to Estimate Human Pose with Data Driven Belief Propagation” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 vol. 2, pp. 747–754.

[7] J. Vignola, J. –F. Lalonde, and R. Bergevin, “Progressive Human Skeleton Fitting” in Proc. of the 16th Vision Interface Conference, Halifax, Canada, 2003, pp. 35–42.

[8] A. S. Micilotta and R. Bowden, “View-based Location and Tracking of Body Parts for Visual Interaction” in Proc. of British Machine Vision Conference, 2004, vol. 2, pp. 849–858.

[9] R. M. Rao, and A. S. Bopardikar, Wavelet Transforms Introduction to Theory and Applications, India: Pearson Education, 2000.

[10] J. C. Russ, The Image Processing Handbook, 2nd ed. Boca Raton, Florida: CRC Press, 1995.

[11] P. Kakumanu, S. Makrogiannis and N. Bourbakis, “A Survey of Skin-Color Modelling and Detection Methods”, Pattern Recognition, vol. 40(3), pp. 1106–1122, March, 2007.

[12] J. Mulligan, “Upper Body Pose Estimation from Stereo and Hand-Face Tracking”, in Proc. of the Second Canadian Conference on Computer and Robot Vision, 2005, pp.413–420.