An Automatic System for Multiple Human Tracking & Actions Recognition in Office Environment

Tan C.C. Henry, E. G. Ruwan Janapriya & Liyanage C. de Silva

Department of Electrical & Computer Engineering
The National University of Singapore
4 Engineering Drive 3, Singapore 117576
{eletanh, elev7, elelcds}@nus.edu.sg

Abstract - This paper presents an automatic system that detects humans, tracks to maintain knowledge of up to 2 human objects' whereabouts in an office environment and recognizes their simple actions (such as walking, sitting down, getting up, squatting down and standing up, in both lateral and frontal views) in color video for activities logging. The detection, tracking and human motion classifications are automatic and suitable for use in smart room/office environment.

Keywords: Skin color detection, tracking in multiple cameras, multiple human objects, monocular, activities/action recognition.

1. Introduction

1.1 Motivation

Any project on multiple human objects tracking and activities recognition in smart room/office environment definitely entails many major research areas like motion detection, presence of human verification, tracking of human objects, people identification through face recognition, advanced user-interface via facial expression recognition and/or voice recognition, activities recognition and logging through motion and postures classification, etc. The attention given to any one or a combination of these areas from the research community has been on the rise. Examples of some recent work include [1-3]. The main motivations are the high number of promising applications, and the common desire to come up with more intelligent machines and discerning vision systems that are able to understand the humans surrounding them and react to the humans so as to suit their needs, e.g. by making a certain environment more conducive for the humans based on the facial expression detected; responding to the humans' requests upon sensing a certain action or a series of motions, etc.

One of our aims for the smart room research project held at our laboratory is to explore the tracking of the human objects, recognition and logging of activities within the enclosed office area. Knowing the location and identity of people in the room is then the most vital prerequisite for many of the services that the smart room can provide. Services like displaying an appropriate message on a LCD panel for a particular user when he enters the room, or zooming in for a close-up video capture when someone has been loitering around the cabinet containing 'Confidential/Secret' documents for too long, etc. All these require the knowledge of the whereabouts and identities of the users in the smart room.

In this paper, we present our current work on the detection, tracking, location estimation and recognition of 10 common actions found in office environment, i.e. walking, sitting down, getting up, squatting down and standing up, in both lateral and frontal views with respect to our Camera 2. In this sense, our work here most closely resembles the work done in [4].

1.2 System overview

Our system setup includes 3 fixed Sony CCD camera installations, one Sony camera adaptor, 3 Euresys Picolo frame grabbers of rate 25fps and one Pentium-4 PC server running on Win 2000, installed with digital video recording software, Video Savant v.3.0. At present, all cameras are used monocularly. Cameras are installed as shown in Figure 1 – Camera 1 is fixed on the wall directly facing the door, the other two on one side of the room (Camera 2 is close to the door and Camera 3 is farther away.) They are all located at about the same height, approximately 1.85m from the floor.

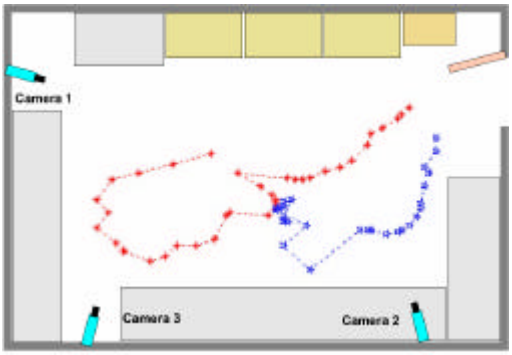

Figure 1. Plan-view of room and cameras locations.

Essentially, our algorithm consists of 3 modules, the Person Detector (PD), the Location Estimator (LE) and the Action Classifier (AC).

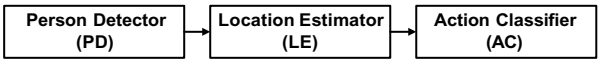

Figure 2. The 3 main modules of our system.

By using skin-color detection and image enhancement procedures, module PD verifies if a human is present. As soon as it confirms that a person is present at the door of the room, the LE starts tracking the person, or up to 2 persons, in the room. By means of background subtraction and a linear transformation of the xy coordinates of the centroid of the heads, these persons' locations are estimated so that a trajectory for each person can be plotted on a 2D plan-view of the room, as shown Figure 1. At the same time, these xy coordinates are recorded and passed to the AC for recognition of the action that has just taken place. Adapting the assumption made in [5], we realized that by setting the constraint that our heads are always

"above" our torsos, we can simplify things and just use the movement of the heads to identify the actions performed. A Linear Vector Quantization (LVQ) neural network is then implemented to classify the recorded action into one of the 10 common actions found in office environment.

The following section of this paper describes the algorithms and implementations of the 3 modules. Section 3 presents the experimental results obtained and a short discussion. Section 4 summarizes the main conclusions and sketches our future directions of research.

2. Methodology

2.1 The Person Detector

In this first module of the system, we seek to verify that a human is indeed present at the door. This is achieved by comparing the size of the face to a preset threshold, S, and matching its shape to an elliptical model. In addition to this testing on the images obtained from Camera 1, we also take the images available at Camera 2 into consideration. The background-subtracted and morphologically-enhanced blob of the object at Camera 2 is compared to another threshold, T. If all the above conditions are true, the system proceeds to the next module, as depicted in Figure 3 below. Immediately following it are the detailed descriptions of the various steps.

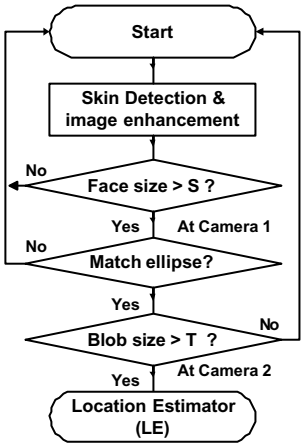

Figure 3. Flow chart of the Person Detector.

i. Skin color detection

Although skin color varies in people from different ethnic groups, previous studies have proved that all skin colors can be approximated to a map in YCbCr space [6,7]. Also, one of our studies has found that the illuminance value (Y) of the skin color is heavily dependent on the camera and the environment [8]. Thus in the present approach, we will discard Y value from skin colour detection algorithm. For the development of our hypothesis, we used a database consisted of skin samples gathered from internet belong to different ethnic groups such as Caucasian, Negroid, Asian, etc. Hence, the skin region we use could be approximated to the area surrounded by the four lines, as shown in Figure 4.

Thus, a pixel is identified as to have skin color if its Cr value fulfills the following constraints:

$$317 - 1.67Cb < Cr < 358 - 1.67Cb$$
$$\textbf{AND} \qquad 138 < Cr < 196 - 0.22Cb \qquad (1)$$

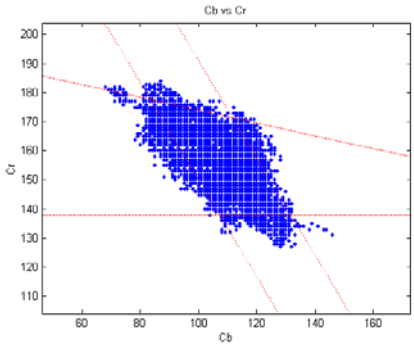

Figure 4. Human skin color model – area bounded by the 4 lines

In this manner, all the skin color regions of the input image taken from Camera 1 are extracted. And since this does not use complex computation, it is efficient as a real-time algorithm.

Following this extraction, Gaussian filtering and morphological operations are done to enhance the image for further processing.

ii. Face size check

Here, we use empirical values of about 700 pixels for the threshold S to differentiate potential facial region from skin regions that are too small to be considered as face.

iii. Elliptical face model

By considering the contours of merged skin color regions and approximating them to ellipses, faces can be easily singled out from other skin color parts, eg. hands, shoulders, etc. As established in [9], if the approximated ellipses have the following properties, it is a candidate to a face.

$$(\text{Minor Axis} \div \text{Major Axis}) > 0.58 \qquad (2)$$

iv. Blob size check

If the requirements in both parts (ii) and (iii) are fulfilled, we conclude that a face is present, which may mean the presence of a person. To confirm it is indeed a human object that is "entering to the room", data from Camera 2 is used. If the background subtracted blob exceeds a certain threshold, T, a person is confirmed to be present at the entrance of the room.

2.2 The Location Estimator

After module 1 has detected the human object, it passes the control to the second module of the system. The objectives of this module are to determine the whereabouts of the persons in the room and approximate each person's location on a plan-view drawing of the room. Putting together the time-varying positions of these objects, a plot of the paths taken by each person in the room can be obtained, shown in Figure 1. This plot not only records the trajectories but also gives hints to the activities that have been taken place in the room.

Figure 5 summarizes the process flow of this module, whose main operations can be described as follows:

i. Motion detection

This is achieved by background subtraction of the current frame. The background is modeled by computing the mean for each pixel in the colour images over a sequence of about 40 frames, which were taken when there is no motion in the smart room. And by means of thresholding, the silhouettes of the persons in the room are then obtained.

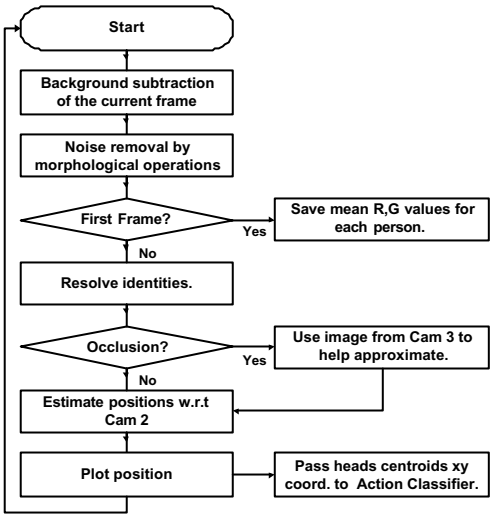

Figure 5. Flow chart of the Location Estimator.

ii. Noise removal

A number of morphological operations are then performed to isolate the actual human motion from the background noise, e.g. flickering of florescent lamps, shadows, etc. This left only the human objects of our interest in the image for further processing.

iii. Maintaining identities

The mean Red and Green values of the different subjects in the first frame of the input sequence are calculated and recorded for comparison and differentiation of the subjects appeared in the later frames, so as to preserve their identities throughout the sequence. Although the HSV and YCbCr colour schemes had been tried and found to give slightly better results, the RGB scheme is still chosen because the images are inherently RGB. Thus, color conversion is not necessary.

iv. Handling occlusion

Under non-occlusion circumstances, only images from Camera 2 are being processed. In the event when one person is occluded by the other from the view of this camera, the images taken from the other two cameras, as in Figure 6, are used to estimate the spatial proximity of the two subjects. These spatial measurements are in turn weighted and used as cross-reference in estimating the locations of the subjects in the scene. The weights have been obtained heuristically.

v. Estimating objects' locations

In order to get a more accurate reference to the objects' locations, the regions corresponding to the persons' heads are segmented from the silhouettes in every frame, and the centroids of all heads in a frame are computed. The xy coordinates of these centroids are then taken as the location reference points of different subjects in that frame.

We use a 100x100 grid to model the floor space of the image. The grid nodes nearest to the reference locations of the subjects are extracted and linearly transformed into points that are superimposed onto a plan-view of the room. The transformation function has been obtained empirically. The transformed points serve as the estimation of the subjects' locations in the images. The 100x100 grid was found to be sufficiently accurate for estimation of the trajectories when projected onto a 407x715-pixels hand-drawn map of the room, depicted in Figure 1. Two means of viewing the trajectories have been incorporated in the current prototype – i.e. via static plot of location points joined by lines; and the avi movie for viewing the path taken by subjects in sequence.

Once the xy coordinates have been plotted, these heads' centroids references are passed to the next module for classification of actions.

2.3 The Action Classifier

Neural networks are normally used in applications where it is required to learn hidden patterns and store them in an associative memory. Among the various networks, we find that LVQ neural network is simple and effective enough to serve our purpose of recognizing the common actions that we find in the office environment.

LVQ is a supervised learning technique that uses the known class information to move the Voronoi vectors slightly so as to improve the quality of the classifier decision regions [10]. The 2-layered network classifies input vectors into target classes by using a competitive layer to find subclasses of input vectors, and then combining them into the target classes.

In our case, the network has been trained with labeled input vectors of 8 subjects, of various builds and sizes. Each subject was requested to perform 10 different actions, namely walking across the room, sitting down, getting up from chair, squatting down, standing up, in both the lateral and frontal views with respect to Camera 2, as shown in Figure 7. As it was observed that most of these actions can be completed in roughly 1.1 sec by all the subjects, we use the head centroid's coordinates in only 28 frames for each action for each subject.

i. Feature Extraction & Feature Vectors

The xy coordinates for the centroids of the heads from the Location Estimator module are used to form an action matrix for each action *j* performed by individual *i*, and this can be written as:

$$A_{i,j} = [\mathbf{y}\ \mathbf{x}]_{i,j}, \qquad (3)$$

where i=1,2,..,8; j=1,2,..,10 and **y** and **x** are the 28-element-column-vectors of the y and x coordinates, respectively, of the centroid of the head in all 28 frames. Hence, for each individual, we can consolidate all actions matrices and simplify as:

$$A_i = [A_{i,1}, A_{i,2}, \ldots, A_{i,10}]. \qquad (4)$$

Also, for each individual, a difference matrix D_i can be formed by computing the difference in coordinates over successive frames in an action sequence, for all 10 actions, such that:

$$D_i = [D_{i,1}, D_{i,2}, \ldots, D_{i,10}], \qquad (5)$$

where $D_{i,j} = [y_{k+1} - y_k \quad x_{k+1} - x_k] = [d_Y \quad d_X]$ for all i and j. Each $D_{i,j}$ is a 27x2 matrix.

By concatenating d_Y & d_X, every $D_{i,j}$ is now a 54-element-column-vector.

As 4 subjects had volunteered for a second take on another occasion, we have, in all, 12 of these 54x10 feature matrices for training as well as testing purposes.

ii. Training with LVQ

In our implementation, the input layer is made up of 140 hidden neurons and the output layer has 10 nodes, each representing a

class of action. This offline action classifier has been implemented in Matlab v.6.1.

Out of the 12 feature matrices, 7 of them were randomly chosen as the training samples, and the remaining 5 were used as the test sets for the networks.

3. Results & Discussions

Table 1 gives the classification results of the individual action sequences.

Action	Total Tested	Correctly Identified	% Accuracy
Lateral view:			
Walk	5	5	100
Sit Down	5	5	100
Get Up	5	3	60
Squat Down	5	4	80
Stand Up	5	3	60
Frontal view:			
Walk	5	5	100
Sit Down	5	4	80
Get Up	5	3	60
Squat Down	5	4	80
Stand Up	5	3	60

Table 1. Classification results by LVQ.

Of all the 50 action sequences tested, we got 39 of them correctly identified. This gives us a classification rate of 78%. It is satisfactory given that the system is simple, only a xy coordinate pair is needed for every head in every frame. Also, all the subjects tested have quite different builds and sizes. The misclassifications mostly occur when the actions are very similar, e.g. getting up from chairs and standing up from squatting position.

Unlike [4], our approach, extracts the features (i.e. xy coordinates of the centroid of the heads) automatically from the images, despite having to specify which 28 frames to use for each action for each subject.

4. Conclusion

In this paper, we have presented our current work on the detection, tracking, location estimation and recognition of 10 common actions found in office environment, i.e. walking, sitting down, getting up, squatting down and standing up, in both lateral and frontal views with respect to our Camera 2.

The human object detection and its presence verification are based on the CbCr skin color model, which has demonstrated to be feature invariant and efficient.

The method of maintaining identities using the mean R and G values of the persons' images has proved to work well, judging from the trajectories plot that we obtained. Even though we are considering only up to two persons in the room at this stage, tracking more human subjects should be just an extension of the methods used.

Our LVQ action classifier is somewhat robust enough to classify the 10 action sequences executed by subjects of different builds and sizes, attaining a classification rate of 78%. But there is still room for improvement. In the next phase, we intend to try out the classical Fisher's Linear Discriminant (FLD) method for dimensionality reduction as well as classification.

In this work, only the distinctions between individuals are made. The exact identity verification can be achieved by incorporating the face or biometric recognitions algorithms.

Acknowledgements

The authors wish to thank everyone who has helped and taken part in the recording of motion sequences and testing of the system.

References

[1] Haibing Ren, Guangyou Xu, "Human action recognition in smart classroom", *Automatic Face and Gesture Recognition, 2002. Proceedings of the Fifth IEEE International Conference,* pp. 417-422, 2002.

[2] Sato, K., Aggarwal, J.K., "Tracking and recognizing two-person interactions in outdoor image sequences", *Multi-Object Tracking, 2001. Proceedings of the 2001 IEEE Workshop,* pp. 87 -94, 2001.

[3] Krumm J., Harris S., Meyers B., Brumitt B., Hale M., Shafer S.,"Multi-camera multi-person tracking for EasyLiving", *Visual Surveillance, 2000. Proceedings of the Third IEEE International Workshop*, pp. 3-10, 2000

[4] Madabhushi A., Aggarwal J.K., "A Bayesian approach to human activity recognition*", Second IEEE Workshop on Visual Surveillance*, pp. 25 -32, 1999.

[5] Haritaoglu I., Harwood D., Davis L.S., "Hydra: multiple people detection and tracking using silhouettes", *Second IEEE Workshop on Visual Surveillance*, pp. 6 -13, 1999.

[6] Rein-Lien Hsu, Abdel-Mottaleb M, Jain AK., "Face detection in color images", *Pattern Analysis and Machine Intelligence, IEEE Transactions on , Volume: 24 Issue: 5*, pp. 696 -706, May 2002.

[7] Terrillon J.C., Shiraz M.N., Fukamacih H., Akamatsu S., "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images", *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference*, pp. 54 -61, 2000

[8] Hua R.C.K, De Silva L.C., Vadakkepat P., "Detection and Tracking of Faces in Real-Time Environments", *Proceedings of 2002 International Conference on Imaging Science, Systems, and Technology (CISST'02), Las Vegas, USA,* June 24-27, 2002.

[9] Yang Ming-hsuan, Narendra Ahuja, "Detection of human faces in color images", *Proceedings of the International Conference on Image Processing,* pp. 127-130, 1998.

[10] Haykin S., "Linear Vector Quantization", in "Neural Networks: A Comprehensive Foundation", *2nd Ed., Prentice Hall*, pp. 466-470, 1994.

Figure 6. Images taken from all 3 cameras.

Figure 7. Snapshots of some action sequences.