Using Head Movement to Recognize Activity *

Anant Madabhushi and J. K. Aggarwal
Computer and Vision Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712, USA
aggarwaljk@mail.utexas.edu

Abstract

This paper presents a methodology for automatically identifying human actions in either the frontal or the lateral view. By tracking the movement of the head of the subject over successive frames of a monocular grayscale image sequence, we recognize 12 different actions. The head is segmented automatically in each frame, and the feature vectors extracted. Input sequences captured from a fixed CCD camera are matched against stored models of actions. The system uses the nearest neighbor classifier to identify the test action.

1 Introduction

Human actions are extremely diverse, and to build a system that can be used to successfully identify any type of action is a challenging problem. Aggarwal and Cai [1], in their work on human motion, discuss the different approaches used in the recognition of human activities. Ayers and Shah [2] have developed a system that makes context-based decisions about the actions of people in a room. In [3], Davis and Bobick computed motion energy and motion history images, which were employed for recognition using template matching. Some researchers have attempted the full three-dimensional reconstruction of the human form from image sequences, presuming that such information is necessary to understand the action taking place [10]. Polana and Nelson [9] used low-level non-parametric representation to represent repetitive activity.

An interesting fact about human activity is the inherent similarity in the way actions are carried out. That is, people sit, stand, walk, bend down and get up in a more or less similar fashion. An important part of human activity recognition has to do with the tracking of body parts [6]. We have found that the head of the subject is more distinctive in defining human action than any other body part. For instance, in the standing up action, the head moves forward initially and then backward, while moving upward continuously as well. In our system, we recognize 12 different types of actions by constructing a feature vector, which is the difference in the centroid of the head over successive frames. Based on the best match between the test sequence and the training models, the system classifies the test sequence as one of the training sequences. In a preliminary implementation [8], a Bayesian framework was used to classify actions based on manual segmentation. Our present implementation uses a nearest neighbor framework with automatic segmentation for recognition.

*This research was supported in part by the Army Research Office under contracts DAAH04-95-I-0494 and DAAG55-98-1-0230, and by the Texas Higher Education Coordinating Board, Advanced Research Project 97-ARP-275.

2 Modeling and Classification

In this section we describe the various steps in modeling our system and our procedure for identifying the test sequences.

2.1 Extracting feature vectors

By modeling the movement of the head for each action, we have means of recognizing that action. To do this, we estimate the centroid of the head in each frame. These are given as $[x_1, y_1] \ldots [x_{n+1}, y_{n+1}]$. The difference in coordinates over successive frames is given by $[Dx_k, Dy_k]$.

$$Dx_k = x_{k+1} - x_k \quad (1)$$

$$Dy_k = y_{k+1} - y_k \quad (2)$$

The feature vectors in our case are the difference in the centroid of the head over successive frames.

$$X = [Dx_1, Dx_2, \ldots, Dx_n] \quad (3)$$

$$Y = [Dy_1, Dy_2, \ldots, Dy_n] \quad (4)$$

where X and Y are the feature vectors for the difference in x and y coordinates of the centroid of the head respectively. Dx_k, Dy_k are not absolute differences, since we would not be able to distinguish between similar pairs of actions like getting up-bending down, sitting down-standing up and rising-squatting. Each of these pairs of actions are almost identical as far as the movement of the head is concerned, except that they proceed in opposite directions. Standing up, for instance, is the reverse of the sitting down action. Using absolute differences would have confused the system as to whether the action was standing up or sitting down.

0-7695-0750-6/00 $10.00 © 2000 IEEE

2.2 Nearest neighbor formulation

The nearest neighbor classifier is a useful classification system when the number of training samples is small [4]. The nearest neighbor classifier assigns the feature vectors $\{X, Y\}$ to the same class Ω_ω (where $\omega \in \{1, 2, \ldots, 12\}$) as the training feature vectors nearest to it in the feature space. The test sequence is assigned to the training class that satisfies equation 5.

$$\min_{\omega}\{\sum_{u=1}^{n} |Dx_u - Dx_{\omega u}| + \sum_{u=1}^{n} |Dy_u - Dy_{\omega u}|\} \quad (5)$$

Thus we compute the absolute difference of the elements of the input test feature vectors $\{Dx_u, Dy_u\}$ and the elements of the training feature vectors $\{Dx_{\omega u}, Dy_{\omega u}\}$ and sum these differences over all the frames. The test sequence is assigned the label of the training class for which this sum is the least.

3 Detection and Segmentation

The detection and segmentation of the head is central to the recognition algorithm. We model our system by estimating the centroid of the head in each frame. Many human activity recognition algorithms depend on efficient tracking of a moving body part [5, 6]. Similarly, in our case the entire recognition algorithm is based on reliably tracking the centroid of the head. In [11] Sirohey used an upright ellipse to find the head in grayscale images. In [5] the head was found using the assumption that it is the highest point on the silhouette of the body. Our system recognizes action sequences in which the head is in various positions. Further, during some actions like the bending down action, the head is lower than the back. To overcome the above constraints we combine two motion-based segmentation techniques to find the head in each frame.

3.1 Frame Differencing

Motion is a strong cue that can aid segmentation [7]. We apply successive frame differencing to segment the head from the rest of the scene. By assuming that the background does not change over successive frames, we isolate only those objects that are moving.

$$\Delta I = |I_t - I_{t+1}| \quad (6)$$

where ΔI is the absolute difference of the t and the $t+1$ frames. Successive frame differencing allows us to exploit the fact that, for most actions, the head is the most mobile part of the body. In several action sequences, the head appears as the largest blob in the binarized difference image. After differencing, we perform connected component labelling and region removal to get rid of the smaller blobs. To identify the blobs belonging to the head, we use the important observation that the head is always at the extreme end of the body, if not at the highest point. Since there are no moving objects in the background, we can assume that the blob corresponding to the head contains pixels that are either at the top of the frame or at an extreme end of the frame. By retaining the blobs that contain pixels belonging to the leftmost column and the top row in the frame, we reduce the number of blobs that could be the head to two. The centroid of the blob is computed as the average of the position of all the pixels in it.

$$i_{fd} = \begin{bmatrix} i_{fd11} & i_{fd21} \\ \vdots & \vdots \\ i_{fd1n} & i_{fd2n} \end{bmatrix}; j_{fd} = \begin{bmatrix} j_{fd11} & j_{fd21} \\ \vdots & \vdots \\ j_{fd1n} & j_{fd2n} \end{bmatrix} \quad (7)$$

where i_{fd} and j_{fd} are the x and y coordinates of the centroids of the two blobs for all frames of an action sequence. The symbol fd refers to frame differencing. The first digit in the subscript refers to the blob number and the second refers to the frame number. Hence i_{fd13} refers to the x coordinate of the centroid of blob 1 in the third frame of the sequence. If there is only one blob in the third frame, then both i_{fd23} and j_{fd23} will be zero. Since we are differencing successive frames, both blobs may represent the head. In figure 1, (a) and (b) represent consecutive frames of the same sitting down sequence. The differenced image, after connected component labeling and region removal, contains two blobs. We see from 1(c) that both blobs, albeit unconnected, belong to the head. Sometimes, the two blobs may correspond to different body parts. We need to identify the blob belonging to the head. For frames that have two blobs, we compute the distance between their centroids and determine whether it is less than a threshold. This threshold is based on the average size of the blobs over all frames of the sequence. If the distance between the centroids of the two blobs is less than the threshold, the two blobs are regarded as being one, otherwise they are considered to be two separate blobs. However, if the two blobs are separate, we do not know which blob is the head. We use a background subtraction technique to identify the blob belonging to the head.

Figure 1: (a) First frame and (b) Second frame of action sequence (c) Filtered difference image.

3.2 Background subtraction

We use background subtraction to remove the slowly varying background gray levels in each frame, and median filtering to reconstruct the background [5]. We

then subtract the background image from every frame in the sequence.

$$I_{sub} = |I_{bs} - I_{or}| \tag{8}$$

where I_{sub} is the background subtracted image, I_{bs} is the reconstructed background image and I_{or} is the original image from the action sequence. To the subtracted image, we apply the preprocessing techniques mentioned in section 3.1 to retain only one or two blobs. In figure 2 we see (a) the reconstructed background, (b) an individual frame in the sequence, and (c) the result of background subtraction. The centriods of the blobs are computed as in section 3.1 and given as:

$$i_{bs} = \begin{bmatrix} i_{bs11} & i_{bs21} \\ \vdots & \vdots \\ i_{bs1n} & i_{bs2n} \end{bmatrix}; j_{bs} = \begin{bmatrix} j_{bs11} & j_{bs21} \\ \vdots & \vdots \\ j_{bs1n} & j_{bs2n} \end{bmatrix} \tag{9}$$

where i_{bs} and j_{bs} are the x and y coordinates of the centroids of the two blobs for all frames of an action sequence. The symbol bs refers to the background subtraction technique. The numbers in the subscripts have the same meaning as in section 3.1.

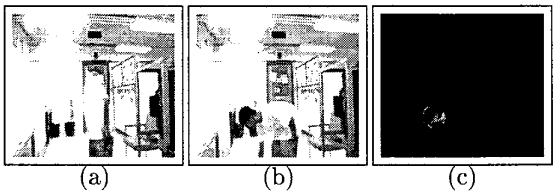

Figure 2: (a) Reconstructed background using median filtering (b) A frame of the action sequence (c) Segmented result of background subtraction.

3.3 Integrating the two paradigms

In this section we discuss how the results obtained from the two paradigms are combined to yield the centroid of the head. Ideally, the centroid of the head should be equal to the centroid of blob 1, obtained from the two techniques. For those frames in which the head was successfully segmented by both paradigms, the centroids of the head obtained from the two techniques should be almost the same. For every frame, we compute the difference of the x coordinate of the centroid of blob 1 for both paradigms. If this difference is less than a threshold (λ), the centroid of the head is the centroid of blob 1 of the background subtraction technique. If this is not so, we check to see if the second blob from the background differencing technique is present. If it is, we compute the difference between the x coordinate of the centroid of blob 1 obtained by frame differencing and the x coordinate of the centroid of blob 2 obtained by background subtraction. If this difference is greater than a threshold (α) or if the second blob in the background

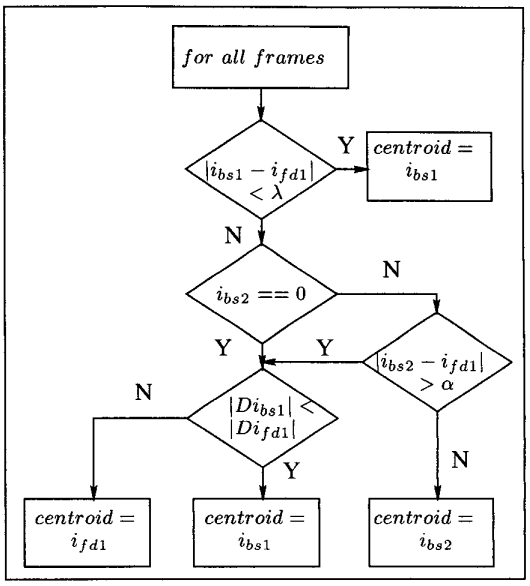

Figure 3: Algorithm used to resolve centroid of head.

subtraction technique is absent, we compare the x coordinate of the centroid of blob 1 in the current frame with it's counterpart in the next frame for both paradigms. The centroid of the head is determined to be the centroid of blob 1 obtained from the technique that has a smaller difference. We observed that a variation in the x coordinate of the centroid of the blobs was accompanied by a similar variation in the y coordinate of the centroid. Hence we only considered i_{bs}, i_{fd} in our algorithm. The flowchart in figure 3 describes the algorithm we use to identify the blob that belongs to the head. Di_{bs1}, Di_{fd1} are the differences in x coordinates of the centroid of blob 1 over two successive frames, for the two paradigms. The subscripts 1 and 2 refer to the two blobs.

4 System Implementation

A fixed CCD camera with a wide field of view working at 2 frames/sec was used to obtain the sequences. We used 36 training sequences, i.e. 3 training sequences for each class. People of diverse physical sizes were used to model the actions. The system detects the head of the subject in each frame and extracts the feature vectors. The test sequence is matched against stored models of different actions. The test sequence is identified as the training sequence to which it is closest. The subjects performed the actions at a comfortable pace. We assumed that each action was performed in roughly five seconds. We found that on average, ten frames were required to completely describe an action, i.e. n had the value 9. Most of the test sequences were performed at

this frequency, although we also tested our model successfully on action sequences having 5, 6 and 7 frames. For an input sequence that has only five frames, we select alternate elements of the training feature vector.

5 Conclusion

In this paper, we have presented a system that can accurately recognize 12 actions using only the movement of the head. The system is robust for subjects of varying heights and weights. Unlike other head segmentation techniques, our approach does not make any assumptions about the position of the head in the frame [5, 11]. Further, the segmentation is robust to background clutter. Figure 4 shows five frames of the getting up sequence. Figures 5 and 6 show the results of background subtraction and frame differencing respectively. Of the 41 sequences that the system was tested on, 34 were correctly identified, giving us a classification rate of 83%. Table 1 gives the classification results of the individual action sequences. FV refers to the frontal view.

At this point, we have not been able to model and test all of the actions in [8] due to the problem of self occlusion for some actions in the frontal view. Further, the system is sensitive to the duration of the action and can only recognize sequences for frame rates in the vicinity of 2 frames/sec. This problem could be handled by increasing the dimensionality of the feature vectors. We also intend to normalize the feature vectors to make the system independent of the distance between the subject and the camera. Finally, the system can recognize only actions that involve head movement.

Type of Sequence	Total Number	Correctly Classified	Success Percent
Standing up	3	2	67
Sitting down	3	3	100
Bending down	4	3	75
Getting up	4	4	100
Walking	3	3	100
Rising (FV)	3	3	100
Rising	5	5	100
Squatting	3	2	67
Squatting (FV)	3	2	67
Side bend	3	2	67
Side bend (FV)	4	3	75
Hugging	3	2	67

Table 1: Classification of individual sequences

References

[1] J. K. Aggarwal and Qin Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, pages 428–440, 1998.

[2] Douglas Ayers and Mubarak Shah. Recognizing human action in a static room. *In Proceedings Computer Vision and Pattern Recognition*, pages 42–46, 1998.

[3] James Davis and Aaron Bobick. The representation and recognition of action using temporal plates. *In Proceedings Computer Vision and Pattern Recognition*, pages 928–934, 1997.

[4] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* New York Wiley, 1973.

[5] I. Haritaoglu, D. Harwood, and L.S Davis. Hydra: Multiple people detection and tracking using silhouettes. *IEEE Workshop on Visual Surveillance*, pages 6–13, 1999.

[6] Stephen S. Intille, James Davis, and Aaron Bobick. Real time closed world tracking. *In Proceedings IEEE International Conference on Computer Vision and Pattern Recognition*, pages 697–703, 1997.

[7] R. Jain, W.N. Martin, and J.K. Aggarwal. Segmentation through the detection of changes due to motion. *In Proceedings of Computer Graphics and Image Processing*, 11:13–34, 1979.

[8] Anant Madabhushi and J.K. Aggarwal. A Bayesian approach to human activity recognition. *IEEE Workshop on Visual Surveillance*, pages 25–32, 1999.

[9] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision*, pages 261–282, 1997.

[10] J. Rehg and T. Kanade. Model based tracking of self-occluding articulated objects. *Proceedings of the 5th International Conference on Computer Vision*, pages 612–617, 1995.

[11] Saad Ahmed Sirohey. Human face segmentation and identification. Master's thesis, University of Maryland, 1993.

Figure 4: Frames of the getting up sequence

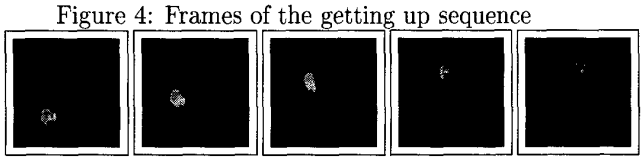

Figure 5: Frames after background subtraction

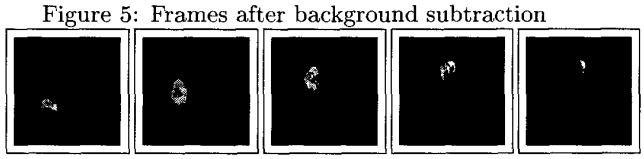

Figure 6: Frames after frame differencing