Segmentation and Recognition of Continuous Human Activity

Anjum Ali
Computer and Vision Research Center
Department of Electrical and Computer Engineering
The University of Texas, Austin
ali@ece.utexas.edu

Abstract

This paper presents a methodology for the automatic segmentation and recognition of continuous human activity. The subject executes a series of actions in the lateral view. There are no distinct breaks or pauses between the execution of different actions. We have no prior knowledge about the commencement or termination of each action. The aim is to segment the sequence of human activity into separate actions and correctly identify each one. We compute the angles subtended by three major segments of the body with the vertical axis namely the torso, the upper segment of the leg and the lower segment of the leg. The angle of inclination of the torso is used to detect *'maxima frames'*. **These are frames in which the angle traverses a local maxima. Once these are determined a forward and backward search is applied to recognize** *'breakpoint'* **and** *'non break point'* **frames. Breakpoints indicate an action's commencement or termination. Frames are classified into breakpoint and non breakpoint frames using the angles subtended by the torso, the upper leg and the lower leg of the human body with the vertical axis. We use discrete action sequences for the training data set. The test sequences are continuous sequences of human activity that consist of three or more actions in succession. The system detects the breakpoints between actions of walking, sitting down, standing up, bending, getting up, squatting and rising, and classifies each action.**

1 Introduction

Human activity is a continuous flow of single or discrete human action primitives in succession. An example of a human activity is a sequence of actions in which a subject enters a room, sits down, then stands up, walks ahead, bends down to pick up something and then gets up and walks away. Each of the components of the human activity such as walking, sitting down, standing up, bending down and getting up are discrete action primitives. Methodology for automatic interpretation of such continuous activity is presented in this paper. When humans move from one action to another, they do so smoothly and transitions between actions are not clearly defined. There is no clear beginning or end of an action. Therefore for the recognition of a continuous activity sequence such as the one described earlier, the detection of transitions between actions is crucial. Most human activity recognition systems can only recognize only single or individual action sequences. In other cases, systems recognize poses associated with an action and not a complete action. This enables them to recognize actions but not necessarily to give an accurate temporal description of each action. In this paper, we analyze a continuous human activity by first automatically segmenting it into discrete actions.

2 Review of Previous Work

Most work in the area of human activity recognition has dealt with the recognition of discrete action primitives. Segmentation and classification of continuous actions is virtually unexplored, and we believe that our system may be the first. The system presented by Madabhushi and Aggarwal [8] classifies twelve different classes of actions. These actions are walking, sitting, standing up, bending, getting up, bending sideways, falling, squatting, rising and hugging in the frontal or lateral view. However, each test sequence was a discrete action primitive. Bobick and Davis [4] use temporal templates for the representation and recognition of human actions. They derived their feature vector from 'motion energy' and 'motion history' images. Using statistical measures extracted from these images, they classified 18 aerobic exercises in 7 different orientations. Once again the approach was to recognize discrete actions, or in their case aerobic exercises. An approach to recognize postures has been presented by Davis and Haritaoglu [5]. A static matching is done between the stored models and the test frames. Their method works well with static frames of an action that consist of postures but might give erroneous results when applied to frames that lie in between their defined four postures of sitting, lying down, standing and crawling. As a result the action recognition process is not automatic and requires manual selection of frames that consist of valid postures. Ayers and Shah [2] present a context based action recognition system capable of determining the actions taking place in a room. The actions recognized by them include walking into a room, opening a cabinet, picking up a telephone, using a computer terminal and so on. A system that talks about recognition of continuous actions in a limited aspect is the one that is presented by Campbell and Bobick [3]. Their system recognizes nine fundamental steps of ballet in an 'unsegmented stream of motion'. They used a set of anatomical constraints to model each step. They were not concerned with the transitions between steps. As long as a certain set of constraints were observed somewhere during the course of a sequence the system announced the corresponding step.

Human body motion is the coordinated movement of its different body parts. These body parts are connected by joints and thus the joints also respond according to the movement or action taking place. We believe that knowledge regarding the limb and joint angles is useful in detecting the termination and commencement of different actions. A number of studies have used information from the movement of the body parts such as the trunk, the arms and the lower limbs to analyze human motion. Rohr [10] described human walking with joint angles of the hip, knee, shoulder and elbow. Along the same vein Bharatkumar *et al* [1] used kinesiology data as the basis for their human walking model. Fujiyoshi and Lipton [6] used the angle of inclination of the torso as a cue to the recognition of walking and running. Niyogi and Adelson [9] exploited the repetitive information of the lower limb trajectory for recognition of human walking.

In our work we present a system that automatically detects the breakpoints or transitions between actions by using the angles subtended by the three major segments of the body, the torso, the upper segment of the legs and the lower segment of the legs. It then classifies each action into one of the action types present in the database. The organization of this paper is as follows. In section 3 we list the different steps of the algorithm which is followed by their detailed description and examples. Section 4 describes the module for discrete action recognition enumerating the features used for the classification of the individual actions thereafter. Section 5 explains the system implementation. Results are presented in section 6. Conclusions and future directions are outlined in section 7.

3 Algorithm for detection of 'breakpoints' of actions

The algorithm for the detection of breakpoints uses the angle of inclination of the torso denoted by θ_t, angle of inclination of the upper segment of the legs, θ_u and angle of inclination of the lower segment of the legs, θ_l. The subscripts t,u and l stand for torso, upper segment and lower segment respectively. These three angles form a feature vector $[\theta_t, \theta_u, \theta_l]$. The steps of the algorithm are given below and also are illustrated in Fig.1.

1: Detection of 'Maxima frames'

Maxima frames are detected by considering frames in which the angle of inclination of the torso θ_t crosses a predefined threshold and then picking the local maxima within these frames. This

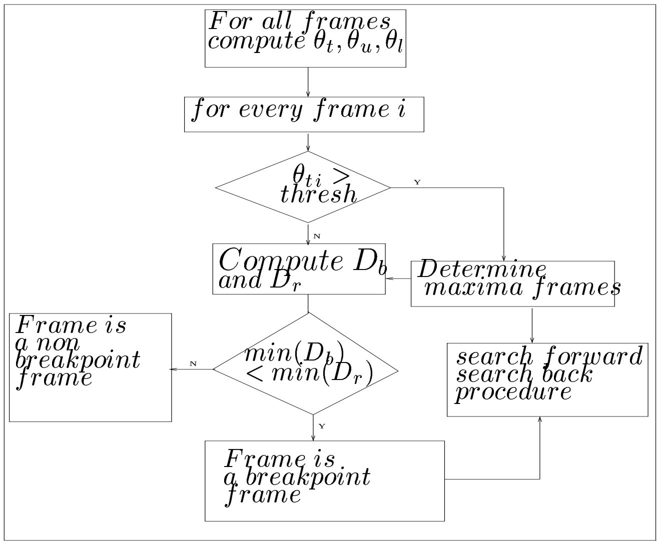
For all frames compute $\theta_t, \theta_u, \theta_l$
for every frame i
$\theta_{ti} >$ thresh
Y
N
Compute D_b and D_r
Determine maxima frames
Frame is a non breakpoint frame
N
$min(D_b) < min(D_r)$
Y
search forward search back procedure
Frame is a breakpoint frame

Figure 1: Algorithm flowchart

threshold is defined by *thresh*.

2: Classification of frames into breakpoint frames

There are two training classes for this classification, a breakpoint training class $[\theta_{tb}, \theta_{ub}, \theta_{lb}]$ and a non breakpoint training class $[\theta_{tr}, \theta_{ur}, \theta_{lr}]$. We compute Euclidean distance measures D_b and D_r. They are the set of distance measures between the test feature vector of each frame and the training feature vectors of the breakpoint and non breakpoint class. Sample training frames for breakpoint and non breakpoint classes are chosen from continuous sequences. If the minima of the set of D_b distance measures is less than the minima of the set of D_r distance measures, the frame is classified as a breakpoint frame.

3: Segmenting individual actions from the continuous sequence.

Once we have classified frames into breakpoint and non breakpoint frames we integrate this information with the previously detected maxima frames, to delineate the discrete action components. Starting from every maxima frame, we do a search forward and search backward procedure that looks for the two breakpoint frames that occur prior to and after a maxima frame. The frames thus enclosed within the two breakpoint frames constitute an individual action component.

3.1 Detection of Maxima frames

The angle of inclination of the torso for a continuous activity sequence traverses a series of maximas. Each discrete action is characterized by one or more local maxima occurring during the execution of the action. If we consider the case of a person bending down, the angle of inclination of the torso increases, and when the subject reaches the maximum position of bending, the angle is maximum. As a result if we consider a bending down action the maxima frame is the last frame. Similarly in the case of a sitting down action (Fig.2), the maximum angle of inclination of the torso is observed midway in the execution of the action. The maxima frame for this sequence is frame 6. We detect maxima frames that are indicative of that frame in an activity sequence in which the subject's torso has the maximum angle of inclination for a local set of frames. However in order to extract the angles subtended by the segments of the body the test sequence frames have to be segmented and skeletonized.

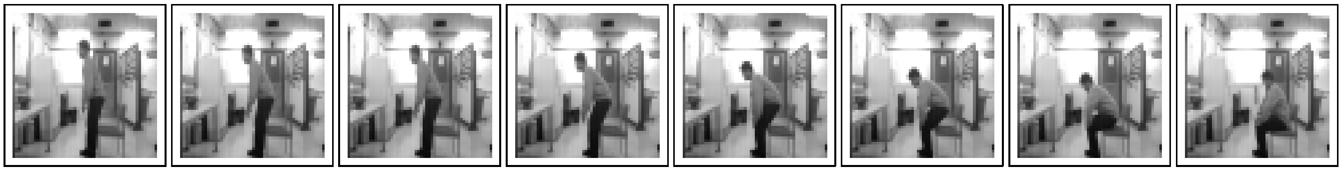

Figure 2: Successive Frames of a sitting down action

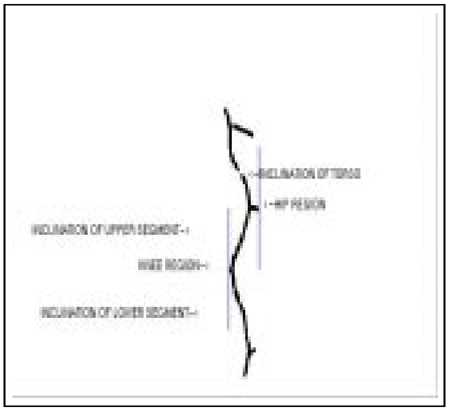

Figure 3: Skeletonization result of one frame of a sequence showing the segment angles.

Skeletonization is used to represent the shape of a region. It has been used extensively in the field of human motion analysis to extract a skeletonized image of the human subject or to generate stick figure models. Bharatkumar *et al.* [1] used the medial axial transformation to extract stick figures and compared the two dimensional information obtained from stick figures with that obtained from anthropometric data. Guo *et al.* [11] also used skeletonization on the extracted human silhouette to yield stick figure models. They then classified human motion based on the positions and angles of different segments and joints in the stick figure. Since we are working with actions in the lateral view, skeletonization yields good results and can be used to efficiently obtain the three main segments of the body used in our algorithm namely, the torso, the upper segment of the legs and the lower segment of the legs. Preprocessing steps include background subtraction, thresholding, morphological operations and finally skeletonization. Subsequent estimation of the hip and the knee regions is a difficult problem. A number of prior works have used apriori information about the position of the hip and the knee joints. Some have used markers on the body for this purpose. Work by Kakadiaris *et al.* [7] used body part decomposition and joint location from image sequences of the human subject using a physics based framework. In our case we use the high points of curvature of the skeleton as potential locations for the hip and the knee.

For the estimation of the hip and knee regions we use an algorithm that describes a point of high curvature as a location where a triangle of specified angle can be inscribed in the curve of the skeleton. For lack of space we cannot go into the details of the algorithm but it is sufficient to say that the algorithm detects the highest points of curvature on the skeleton.

The hip is the first high point of curvature on the skeleton curve. The knee is then defined as the highest point of curvature below the hip. Once the hip and the knee regions are detected the different angles are computed using the vertical axis passing through the hip and the knee. Fig.3 shows the three segments on the skeleton for one frame of a sequence.

3.2 Classification of frames into Breakpoint frames

Each frame of a continuous sequence is represented by a feature vector having three components, the angle of inclination of the torso θ_t, the angle of inclination of the upper segment θ_u and the angle of inclination of the lower segment θ_l. We define two classes, a breakpoint class and a non breakpoint class. Training vectors for both these classes are chosen from continuous sequences. Frames of a continuous sequence in which a person is at the commencement or the termination of an action are chosen to form the breakpoint class. This class is represented by a three element feature vector for every frame represented by

$[\theta_{tb}, \theta_{ub}, \theta_{lb}]$

Fig.4 shows some of the training frames that have been used for the breakpoint class. 16 sample frames have been used for training the breakpoint class. Conversely frames in which the subject is midway in the execution of an action are chosen to form the non breakpoint class. This class is also represented by three element feature vector for every frame represented by

$[\theta_{tr}, \theta_{ur}, \theta_{lr}]$.

28 sample frames have been used for training the non breakpoint class. The values of the angles of inclination of the body segments at breakpoints between actions are distinctly different from that observed during the course of the actions. If we consider a sitting down action, when the subject has completely sat down that is at a breakpoint the angle subtended by the torso with the vertical axis is minimal, that subtended by the upper segment of the leg (thigh) is almost a right angle and that by the lower segment is again minimum. These values differ appreciatively from a

frame in which the person is standing straight or has not completely sat down, more so it is the combination of the values of all three angles of inclination that is distinctly different. Similarly when the subject ends a bending action the angle of inclination of the torso is high and that of the other two segments is low. When a person has completely squatted down the angles subtended by the upper and lower segments of the leg are at a maximum due to maximum flexion of these segments. Though more than one frame at the point of transition maybe classified as a breakpoint frame, the frame whose feature vector falls closest to the breakpoint class in terms of the Euclidean distance is chosen. For each test frame we compute the angles subtended by the three segments with the vertical axis. This three element test feature vector $[\theta_t, \theta_u, \theta_l]$ is then compared with the training feature vectors from each class. The algorithm computes two Euclidean distance measures D_b and D_r between the feature vector for each frame i and the training feature vectors of the breakpoint and non breakpoint classes. They are computed as

$$\mathrm{D}_{bk} = \sum_{k=1}^{12} (\theta_{ti} - \theta_{tbk})^2 + \sum_{k=1}^{12} (\theta_{ui} - \theta_{ubk})^2 + \sum_{k=1}^{12} (\theta_{li} - \theta_{lbk})^2 (1)$$

$$\mathrm{D}_{rk} = \sum_{k=1}^{28} (\theta_{ti} - \theta_{trk})^2 + \sum_{k=1}^{28} (\theta_{ui} - \theta_{urk})^2 + \sum_{k=1}^{28} (\theta_{li} - \theta_{lrk})^2 (2)$$

A frame is classified as a breakpoint frame if the minima of the distance measure D_b is less than the minima of measure D_r.

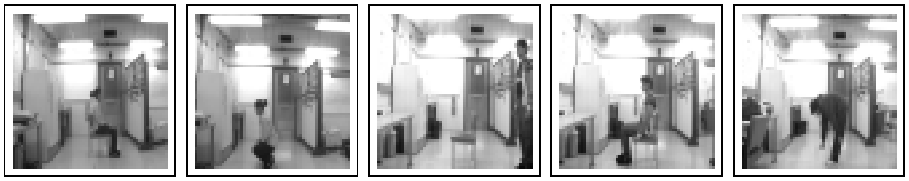

Figure 4: Samples of breakpoint training frames

3.3 Forward and Backward search

After classifying all the frames of a test sequence into breakpoint or non breakpoint frames, a backward and forward search is done. Starting from each maxima frame we look for the two breakpoint frames that are proximal to the maxima frame under consideration, from either direction. The contiguous set of frames enclosed by the two breakpoint frames are then marked as a single discrete action. Once all the maxima frames have been analyzed we check to see if there

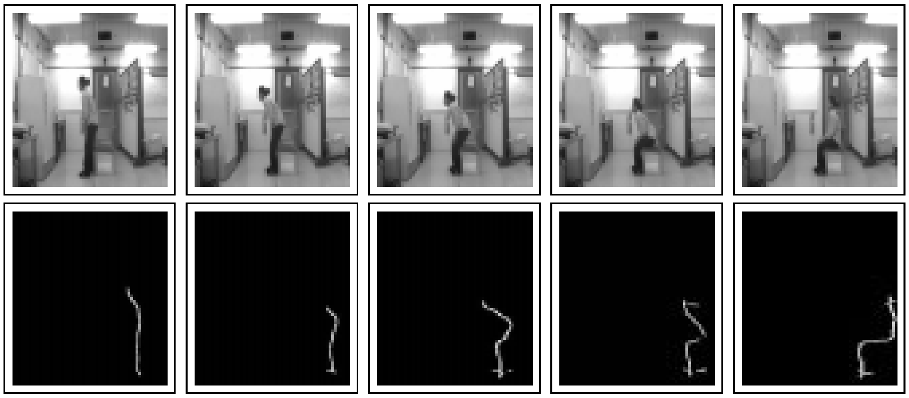

Figure 5: Frames of a sitting down action and it's skeleton showing the segments

any action frames that have not as yet been classified. If so then these frames are taken now and classified. These frames usually belong to the walking action.

4 Discrete action recognition

Similar to [8], our system models the movement of the head to classify the different action primitives. The x and y coordinates of the head are first normalized. This is done by dividing the x and y coordinates of the head, for each segmented action by the maximum x and maximum y coordinate respectively for that particular action. We then compute successive differences of these coordinates. However along with the difference in the coordinates of the head we also use the angle of inclination of the three major body segments for classifying each discrete action.

We observe that the angle of inclination of the torso, the upper segment and the lower segment of the legs follows a characteristic path during the execution of each action. During sitting down (Fig.5) the angle of inclination of the torso first increases, passes a maxima and then decreases as the subject assumes the sitting posture. The angle of inclination of the upper segment of the leg initially is minimum, but then increases and in the sitting posture is maximum. A similar observation can be made while standing up except that the inclination of the upper segment of the legs is initially maximum and then decreases. Similarly during bending the angle of inclination of torso increases till the subject ends the bending action and then decreases as he starts the getting up action. In the case of bending and getting up, the angle of inclination of the lower and upper segment of the leg remains minimum throughout the action. If we consider squatting angle of inclination of the torso also shows a midway maxima like sitting down but the difference is that the angle of inclination of the upper and the lower segments of the leg, both increase to a maximum

Thus by modeling the movement of the head as well as that of the three segments of the body we can classify each action. The feature vectors are the difference in the centroid of the head over successive frames and the angles subtended by the torso, the upper segment and the lower segment of the leg with the vertical axis.

$$X = [Dx_1, Dx_2, \ldots, Dx_n] \tag{3}$$

$$Y = [Dy_1, Dy_2, \ldots, Dy_n] \tag{4}$$

$$A1 = [\theta_{t1}, \theta_{t2}, \ldots, \theta_{tn}] \tag{5}$$

$$A2 = [\theta_{u1}, \theta_{u2}, \ldots, \theta_{un}] \tag{6}$$

$$A3 = [\theta_{l1}, \theta_{l2}, \ldots, \theta_{ln}] \tag{7}$$

where X and Y are the vectors for the difference in x and y coordinates of the centroid of the head respectively. $[Dx_k, Dy_k]$ are the difference in coordinates over successive frames. $A1$ is the vector for the angle of angle of inclination of the torso for frames of an action sequence. $A2$ is the vector for the angle of inclination of the upper segment of the leg and $A3$ is the vector for the angle of inclination of the lower segment of the leg for successive frames of an action sequence.

The nearest neighbor classifier assigns the feature vector $\{X, Y, A1, A2, A3\}$ to the same class Ω_ω (where $\omega \in \{1, 2, \ldots, 7\}$) as the training feature vectors nearest to it in the feature space.

For the segmentation of the head we use the algorithm presented by Madabhushi and Aggarwal in [8]. They combine background subtraction and frame differencing to segment the head of the subject in every frame of a sequence.

5 System Implementation

Image sequences are obtained using a fixed CCD camera. The initial rate of capture of the camera was 2 frames per second; later the system was tested on frame speeds of 15 frames per second as well. The test sequences range in size from 30 to 80 frames. The training frames of the breakpoint and non breakpoint class are breakpoint frames chosen from continuous sequences. A total of 40 such frames have been used to train the system for breakpoint and non breakpoint frame classification, 12 belong to the breakpoint class and 28 to the non breakpoint class. The system segments the head and the body of the subject. The resultant image after the body segmentation is given to the skeletonization routine. Once the skeletonization is complete, the system detects the highest points of curvature on the skeleton. Using the head as reference these are identified as the hip and the knee region. The angles subtended by the torso and the upper and lower segments with the vertical axis are computed. This information is then processed to detect the breakpoints. Once the break points are detected, each segmented action primitive is classified separately using the difference in the coordinates of the head for successive frames as well as the angles subtended

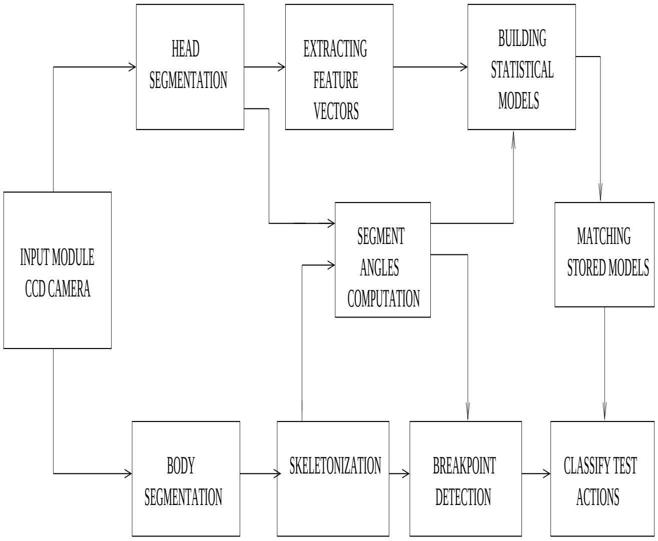

Figure 6: System Implementation

by the three segments of the body. Each action is compared with stored models of actions and classified to the test action to which it is closest. A flow chart explaining these steps is shown in Fig.6.

6 Results

The algorithm has presently been tested on 30 continuous activity sequences. Each sequence consists of 60 or more frames of actions being performed in a continuous manner with no distinct breaks or pauses. For lack of space the results have been tabulated for 8 of the test sequences in Table.1. The table gives the actual duration and type of each action along with the the action and duration that were detected using our algorithm for breakpoint detection and further classification. In 80 of the cases the breakpoints have been correctly identified.

7 Conclusion and Future Work

We have presented an approach for the automatic segmentation and recognition of continuous human activity. The input activity sequence consists of actions that are performed in succession without any forced breaks or pauses. The system uses the information from the angles subtended by the three major segments of the body to classify frames of the sequence into breakpoint and non breakpoint frames. These segments are the torso, the upper leg and the lower leg. After identifying the local maxima in each action, a search is done to separate each action component from the sequence. Since we use the angles subtended by the segments of the leg, an obvious step ahead is to increase the number of actions. Actions that involve considerable movement of the legs like kicking and crawling are a few under consideration. Also, we would like to find more efficient means of detecting the hip and knee regions. Currently our system is limited to actions performed in the lateral view. We intend to extend our system to segment and recognize continuous actions in front view as well. This would help in recognizing a larger number of actions and also longer sequences that involve both, the lateral and frontal views. A state space approach would help in predicting successive actions in a sequence and thus only those actions can be checked for by the classifier.

References

[1] M.G. Pandy Qin Cai A. G. Bharatkumar, K.E. Daigle and J.K. Aggarwal. Lower limb kinetics of human walking with the medial axis transformation. *In Proceedings IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 70–76, 1994.

[2] Douglas Ayers and Mubarak Shah. Recognizing human actions in a static room. *In Proceedings Applications of Computer Vision, WACV'98, Fourth IEEE Workshop*, pages 42–47, 1998.

[3] Lee Campbell and Aaron Bobick. The recognition of human body motion using phase space constraints. *In Proceedings Fifth International Conference on Computer Vision*, pages 624–630, 1995.

[4] James W. Davis and Aaron F. Bobick. The representation and recognition of human movement using temporal templates. *In Proceedings Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 928–934, 1997.

[5] Larry Davis and Ismail Haritaoglu. Ghost: A human body part labeling system using silhouettes. *In Proceedings Fourteenth International Conference on Pattern Recognition,Brisbane, Australia*, pages 77–82, 1998.

[6] H. Fujiyoshi and Alan J. Lipton. Real time human motion analysis by image skeletonization. *Fourth IEEE Workshop on Applications of Computer Vision, WACV 1998*, pages 15–21, 1988.

[7] Dimitris Metaxas Ioannis Kakadiaris and Ruzena Bajscy. Active part-decomposition, shape and motion estimation of articulated objects: A physics-based approach. *IEEE Computer Society's Conference on Computer Vision and Pattern Recognition*, pages 980–984, 1994.

[8] Anant Madabhushi and J. K. Aggarwal. Using head movement to recognize human activity. *International Conference on Pattern Recogntion, Baracelona, Spain*, 2000.

[9] S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in xyt. *IEEE Computer Society's Conference on Computer Vision and Pattern Recognition*, pages 469–474, 1994.

[10] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP Image Understanding*, pages 94–115, 1994.

[11] Gang Xu Yan Guo and Saburo Tsuji. Understanding human motion patterns. *Proceedings of the 12th IAPR Conference on Computer Vision and Image Processing*, pages 325–329, 1994.

Test Sequence 1								
Actual frames	1-9	9-17	17-25	25-30	30-37	37-43	44-53	53-60
action	walk	sit	standup	walk	squat	rise	bend	getup
Observed frames	1-7	7-15	15-21	21-29	29-35	35-43	44-49	49-58
action	walk	sit	standup	walk	walk	rise	bend	getup
Test Sequence 2								
Actual frames	1-10	10-19	20-27	28-34	34-41	41-51	51-60	–
action	sit	standup	walk	bend	getup	squat	rise	–
Observed frames	3-12	12-19	20-26	26-32	32-39	39-49	49-56	
action	sit	standup	walk	bend	getup	bend	rise	
Test Sequence 3								
Actual frames	1-10	10-20	20-30	30-37	39-47	47-55	55-63	–
action	sit	standup	walk	squat	rise	bend	getup	–
Observed frames	sit	standup	walk	squat	getup	bend	getup	–
action	2-9	9-18	21-29	30-38	39-46	47-56	57-64	–
Test Sequence 4								
Actual frames	1-9	10-19	19-28	28-36	36-43	44-55	55-64	–
action	walk	sit	standup	bend	getup	squat	rise	–
Observed frames	1-9	11-20	20-27	27-36	37-44	45-55	55-66	–
action	walk	sit	standup	bend	getup	squat	rise	–
Test Sequence 5								
Actual frames	1-8	8-16	16-27	27-33	33-39	39-46	46-52	52-60
action	walk	sit	standup	walk	squat	rise	bend	getup
Observed frames	1-9	10-17	17-28	28-34	34-40	40-47	47-50	50-54
action	walk	sit	rise	walk	squat	rise	bend	rise
Test Sequence 6								
Actual frames	1-10	10-19	19-27	27-32	32-40	40-48	48-56	–
action	sit	standup	walk	bend	getup	squat	rise	–
Observed frames	3-11	11-19	19-28	28-33	33-41	42-49	49-58	–
action	sit	standup	walk	bend	getup	squat	rise	–
Test Sequence 7								
Actual frames	1-13	13-21	22-29	29-36	36-44	44-54	–	–
action	walk	bend	getup	squat	rise	walk	–	–
Observed frames	3-14	14-22	22-30	31-38	39-45	45-55	–	–
action	walk	bend	getup	squat	rise	walk	–	–
Test Sequence 8								
Actual frames	1-10	11-19	21-29	30-38	39-44	45-50	50-58	–
action	walk	sit	standup	bend	getup	walk	squat	–
Observed frames	2-9	11-16	16-21	21-29	30-39	39-46	46-50	50-59
action	walk	squat	walk	standup	bend	getup	walk	squat

Table 1: Results for Test sequences