A Bayesian Approach to Human Activity Recognition *

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

Abstract

This paper presents a methodology for automatically identifying human action. We use a new approach to human activity recognition that incorporates a Bayesian framework. By tracking the movement of the head of the subject over consecutive frames of monocular grayscale image sequences, we recognize actions in the frontal or lateral view. Input sequences captured from a CCD camera are matched against stored models of actions. The action that is found to be closest to the input sequence is identified. In the present implementation, these actions include sitting down, standing up, bending down, getting up, hugging, squatting, rising from a squatting position, bending sideways, falling backward and walking. This methodology finds application in environments where constant monitoring of human activity is required, such as in department stores and airports.

1 Introduction

Human action recognition is an important topic in computer vision. The task of recognizing human actions poses several challenges. Human action is extremely diverse, and to build a system that can be used to successfully identify any type of action is a serious problem indeed. 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, assuming, of course, there is no impediment in the performance of these actions. An important part of human activity recognition has to do with tracking the body parts. Among the various body parts, it is the head of the subject that is most distinctive in its movement. The head of the subject moves in a characteristic fashion during these actions. For instance, in the standing action, the head moves forward and then backward while moving upward continuously as well. Similarly, in the sitting action, the head moves slightly forward and then backward and downward. Likewise, when a person falls down backward, the head moves down, and at the same time traces the arc of a circle, with the feet of the person serving as the center of this arc. The curve traced out by a backward falling action is different from the curve traced out by the head during a sideways bending action, where the center of the curve is restricted to the center of the torso. Thus each action can be distinguished by the characteristic movement of the head, which differentiates it from other actions.

In our system, we track the movement of the head over consecutive frames and model our system as the difference in the co-ordinates of the head over successive frames. The system is able to recognize sequences where the gait of the subject in the input sequence differs considerably from the training sequences on which it has been modeled. Since the system uses the difference in co-ordinates of the head as its feature vectors, it is able to recognize actions for people of varying physical stature, i.e., tall, short, thin, fat, etc. Hence the system can recognize the bending down action of both a short as well as tall person. For instance, in a sideways bending action where the head traces a curve whose radius is roughly equal to half the length of the body, the size of the radius itself may differ depending on the height of the person; however, the shape of the curve traced out in each case is the same. Thus, our system is not sensitive to the physical stature of the subject.

Much work has been done in the area of human activity recognition. Cai and Aggarwal [1] discuss the different approaches used in the recognition of human activities. They classify the approaches towards human activity recognition into state-space and template matching techniques. Liao et al [2] discuss methodologies which use motion in the recognition of human activity. Ayers and Shah [3] have developed a system that makes context-based decisions about the actions of people in a room. These actions include entering a room, using a computer terminal, opening a cabi-

*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.

net, picking up the phone, etc. Their system is able to recognize actions based on prior knowledge about the layout of the room. Davis, Intille and Bobick [9] have developed an algorithm that uses contextual information to simultaneously track multiple, non-rigid objects when erratic movements and object collisions are common. However, both of these algorithms require prior knowledge of the precise location of certain objects in the environment. In [3], the system is limited to actions like sitting and standing. Also, it is only able to recognize a picking action by knowledge of where the object is and tracking it after the person has come within a certain distance of it. In [7], Davis uses temporal plates for matching and recognition. The system computes history images (MHI's) of the persons in the scene. Davis [7] computes MHI's for 18 different images in 7 different orientations. These motion images are accumulated in time and form motion energy images (MEI's). Moment-based features are extracted from MEI's and MHI's and employed for recognition using template matching. Although template matching procedures have a lower computational cost, they are usually more sensitive to the variance in the duration of the movement.

A number of 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, 6, 14]. Others have proposed methods for recognizing action from the motion itself, as opposed to constructing a three-dimensional model of the person and then recognizing the action of the model [11, 4]. We provide an alternative to both of these approaches by proposing that our method of two-dimensional successive differencing of the centroids of the head eliminates the need to construct three-dimensional models as a prerequisite for recognition.

Our methodology, like Rosario and Pentland [12], uses the Bayesian framework for modeling human actions. Given the correct probability density functions, Bayes theory is optimal in the sense of producing minimal classification errors. State space models have been widely used to detect, predict and estimate time series over a long period of time. Many state space systems use the hidden Markov model (HMM), a probabilistic model for the study of discrete time series. In [12, 15], HMMs have been applied to human activity recognition. However, our approach, unlike [12, 15], computes statistical data about the human subject and models the actions based on the mean and covariance matrix of the difference in co-ordinates of the centroid of the head obtained from different frames in each monocular grayscale sequence. Thus we are able to design a system that is simple in design, but robust in recognition.

Human action recognition finds application in security and surveillance. A great deal of work has centered on developing systems that can be trained to alert authorities about individuals whose actions appear questionable. For instance, in an airport a system could be trained to recognize a person bending down to leave some baggage and then walking off, leaving it unattended, as a cause for concern and requiring investigation. Similarly, in a department store, a person picking up an article and leaving without paying could be interpreted as a suspicious activity. Thus, an intelligent, efficient recognition system could make manual surveillance redundant or, at any rate, reduce the need for human monitoring.

This paper is structured as follows: Section 2 presents our modeling and classification algorithm, section 3 describes the techniques for segmentation and tracking of the head of the subject, and section 4 describes the system implementation. Section 5 presents the experimental results obtained, while section 6 summarizes the main conclusions and sketches our future directions of research.

2 Modeling & 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

The motion of the head forms the basis of our detection and matching algorithm. The head of the person moves in a characteristic manner while walking, sitting, standing, hugging, falling down, etc. Thus each action is distinguished by the distinctive movement of the head in the execution of that particular action. By modeling the movement of the head for each of the individual actions, we have means of recognizing the type of action. To do this, we proceed by estimating the centroid of the head in each frame. The centroids of the head for the different frames of each sequence are given as $[x_1, y_1] \ldots [x_{n+1}, y_{n+1}]$. After computing the centroids of the head in each frame, the difference in the absolute co-ordinates in successive frames was found. $[dx_i, dy_i]$ are the difference in centroids of the head over successive frames.

$$dx_i = x_{i+1} - x_i \tag{1}$$

$$dy_i = y_{i+1} - y_i \tag{2}$$

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

$$X = [dx_1, dx_2, \ldots, dx_n] \tag{3}$$

$$Y = [dy_1, dy_2, \ldots, dy_n] \tag{4}$$

where X and Y are the feature vectors for the difference in x and y coordinates of the head respectively. Since

there are $n+1$ frames in each sequence, each feature vector is n elements long. Thus each feature vector is an n dimensional vector. Next, the mean and covariance matrix for the feature vector was found. This was repeated for all the monocular grayscale sequences.

2.2 Computing probability density functions

We assume independence of the feature vectors X and Y and a multi-variate normal distribution for all sequences. From the independence assumption we have:

$$p(X,Y) = p(X)p(Y) \tag{5}$$

where

$$p(X) = \frac{1}{(2\pi)^{n/2}|\Sigma_X|^{1/2}} e^{[\frac{-1}{2}(X-\mu_X)^t \Sigma_X{}^{-1}(X-\mu_X)]} \tag{6}$$

$$p(Y) = \frac{1}{(2\pi)^{n/2}|\Sigma_Y|^{1/2}} e^{[\frac{-1}{2}(Y-\mu_Y)^t \Sigma_Y{}^{-1}(Y-\mu_Y)]} \tag{7}$$

where X is the n-component feature vector in the x direction, Y is the n-component feature vector in the y direction, μ_X and μ_Y are the *mean vectors* of the normal distribution and Σ_X and Σ_Y are the $n-by-n$ *covariance matrices*. Unbiased estimates for Σ_X and Σ_Y are supplied by the sample covariance matrices [8].

$$C_X = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu_X)(X_i-\mu_X)^t \tag{8}$$

$$C_Y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\mu_Y)(Y_i-\mu_Y)^t \tag{9}$$

2.3 Bayesian formulation of the approach

Using the feature vector obtained from the test sequence, *a posteriori* probabilities are calculated using each of the training sequences. This is done using Bayes rule, which is a fundamental formula in decision theory. In the mathematical form it is given as [8]

$$P(\omega_i/X,Y) = \frac{P(\omega_i)p(X,Y/\omega_i)}{p(X,Y)} \tag{10}$$

where X, Y are the extracted feature vectors, and, $\rho(X,Y) = \sum_{i=1}^{m} p(X,Y/\omega_i)P(\omega_i)$. $P(\omega_i/X,Y)$ is the *a posteriori* probability of observing the class ω_i given the feature vectors X and Y. $P(\omega_i)$ is the *a priori* probability of observing the class ω_i, $p(X_i,Y_i/\omega)$ is the conditional density and m refers to the number of classes.

2.4 Recognition of input sequence

We assume in the recognition of our input sequence that each sequence is uniquely described by the value of its *a posteriori* probability. For our problem, we assume all *a priori* probabilities (the probability of any of the actions occurring) to be equal and, thus, find density functions for each of the classes where each class is an action. Thus, twenty such densities were found, corresponding to the ten different actions in the two orientations. Having obtained these twenty values for each of the classes, the most likely action is the class with the highest value.

$$P = max[P_1, P_2, P_3 \ldots P_m] \tag{11}$$

where P is the probability of the most likely class and $P_1, P_2, P_3 \ldots P_m$ are the probabilities of m different actions.

The frontal and lateral views of each action are modeled as individual action sequences. Hence, we are able to recognize each view by treating it as a distinct action sequence and without having to incorporate information from the other view.

2.5 Discriminating similar actions

For certain actions the head moves in a similar fashion. For instance, when viewed from the front, during squatting, sitting down and bending down, the head moves downward without much sideward deviation. Similarly, during standing up, rising and getting up actions, the head moves upward without much sideward deviation. In order to distinguish these actions from one another, we consider a discriminant number, whose value depends on how low the head goes in the performing of these actions. During bending down, the head goes much lower than in sitting down, and in sitting the head goes lower than in squatting. Let

$$g = max(y_{input})/max(y_{training}) \tag{12}$$

In general,

$$max(y_{gettingup}) > max(y_{sitting}) > max(y_{squatting}) \tag{13}$$

where g is the discriminant number obtained as a ratio of the maximum y co-ordinate in the input sequnce to the maximum y co-ordinate in the training sequences, $max(y_{gettingup})$, $max(y_{sitting})$, $max(y_{squatting})$ are the maximum values of the y co-ordinate of the head in the getting up, sitting and squatting actions in the front view. We compute $g_{gettingup}$, $g_{sitting}$, $g_{squatting}$, as the discriminant numbers corresponding to the three classes, namely getting up, sitting and squatting in the front views, which are obtained using equation 12. Thus whenever the system finds that the input action

is one of the above three, it decides the most likely action by choosing that action which has the maximum discriminant number. A similar process is invoked for the rising from the squatting position, standing and getting up actions. Other actions that are similar with respect to the motion of the head can be distinguished by considering the size of the head in successive frames. Thus, a walking action in the frontal view, which is similiar to the backwards bending action, can be distinguished by making use of the fact that the size of the head increases over successive frames as the subject approaches the camera.

$$\alpha = max(\lambda)/min(\lambda) \tag{14}$$

where λ is the size of the head in one frame of the action sequence and α is the ratio of the maximum and minimum sizes of the head taken over all frames of that action sequence. If $\alpha > \delta$, where δ is a predefined threshold, then the computed probability for the walking action in the front view is multiplied by a weighting factor W_i.

3 Detection & 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, 9]. Similarly, in our case, the entire recognition algorithm is based on reliably tracking the centroids of the head. At this stage of the project we do the segmentation by hand, isolating the head from the rest of the scene by first constructing a bounding box around the head of the subject in each frame. This bounding box is used to keep track of the head over successive frames of each sequence. We fill this bounding box with one color and assign a different color to the rest of the background. Hence we segment the entire scene into two regions, namely, the head of the person(black) and the background(white). This was done using the COREL PHOTOHOUSE program. We compute the centroid of the head in each frame as the average of the positions of all the black pixels. Figure 1 shows the steps in the detection and segmentation of the head. In figure 1(a) we have a grayscale image of the subject. In figure 1(b), a bounding box is placed over the head, and in figure 1(c), the head is segmented from the rest of the background by assigning it a different color. Obviously we would like to incorporate an approach that can automatically detect the head and segment it from the rest of the scene.

We are currently exploring the possibility of generalizing an algorithm based on Saad Ahmed Shiroey's thesis on human face segmentation and identification [13]. In this approach, pre-processing is done on edge

Figure 1: (a) grayscale intensity image and (b) bounding box placed over subjects head (c) segmenting the head from the rest of the background.

detected images of the scene to find the labeled edges that, when combined, are fitted to an ellipse in a least squares sense. The head is modeled as the largest ellipse in the scene. However, this approach is geared towards human face identification. In images on which this algorithm has been tested, the face, occupies the largest portion of the scene. This is not true in our case, since our frames include the entire body of the person. Edge detection of our scene produces far more labeled segments than the algorithm was originally intended for, making the ellipse fitting computationally very expensive. We are working to develop an algorithm that can robustly detect the head for our system as well.

4 System Implementation

A CCD static camera with a wide field of view working at 2 frames per second was used to obtain sequences of monocular grayscale images of people performing the different actions. The frames were taken in the front view and the lateral view. In order to train the system, 38 sequences were taken of a person walking, standing, sitting, bending down, getting up, falling, squatting, rising and bending sideways, in both the frontal and lateral views. People with diverse physical appearances were used to model the actions.

Figure 2 describes the processing loop and the main functional units of our system. The system detects and tracks the subject in the scene and extracts a feature vector describing the motion and direction of the subject's head. The feature vector constitutes the input module, which is used for building a statistical model. Based on the input sequence, the model is then matched against stored models of different actions. Lastly, the action is classified as the one whose probability is the highest.

The subjects were asked to perform the actions at a comfortable pace. This was done for all action sequences and for different subjects. Human motion is periodic; hence, we can divide the entire sequence into a number of cycles of the activity. After observing

the various sequences and different subjects executing these actions, it was found that on average ten frames were required to completely describe an action. We found this to hold true for all twenty actions that were modeled and tested. Hence, we designed our system to consider only the first ten frames of each sequence, ignoring the rest. The rate of capture of the images was 2 frames/second. Thus we assumed that each action was performed in roughly five seconds. We also tested our model on action sequences done at a faster rate, for instance, actions that required only five, six or seven frames. Hence, for an input sequence that has only five frames, we select only four of the 9 elements of the X and Y feature vectors obtained from the training samples and use them to compute a 4 by 4 covariance matrix. The model was able to recognize the action correctly in most cases. However, for actions that required fewer frames than this, the model was not that successful.

For the threshold and the weighting factor we used $\delta = 5.2$ and $W_i = 2.15$.

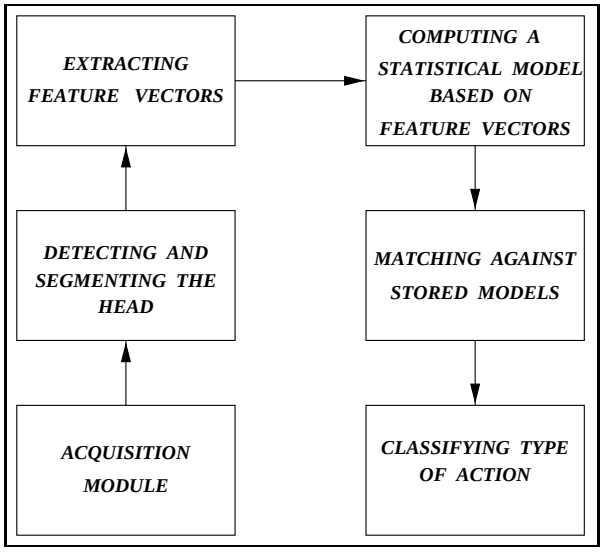

Figure 2: System Overview.

5 Results

This section describes the results obtained from experiments performed on a database of 77 action sequences. Of these, 38 were used for training and 39 were used for testing. Of the 39 test sequences, the system was able to correctly recognize 31, giving a success rate of 79.74. In 6 of the 8 test sequences that were incorrectly classified, the system classified the action correctly, but as belonging to the wrong field of view. The system was able to recognize actions for people of varying physical appearances, from tall to short and from slender to fat. Figures 3-6 show sequences of a subject executing different actions. Owing to a paucity of space, only five key frames in each sequence have been shown. In Figure 3(a), a person is standing up in the lateral view. In the segmented sequence of the standing up action, Figure 3(b), we can see the distinct movement of the head as it moves forward initially, then slightly downward and progressively upward and backward. In Figure 4(a), the subject is seen executing a bending over action in the front view. Figure 4(b) reveals the characteristic downward motion of the head in the front plane. Figure 5(a) shows the subject executing a sideways bending action in the front view. In Figure 5(b), the segmented version of the same, we see the head of the body trace the arc of a circle that has a radius equal to the length of the upper body torso. The center of this arc lies roughly at the center of the body. Finally, in Figure 6(a) we see the subject hugging another person. Notice, in Figure 6(b), the manner in which the head moves forwards horizontally and then dips slightly in the last frame. Table 1 shows the results of classification for 39 test sequences. There were 16 action sequences in the front view (FV) and 23 sequences in the lateral view (LV). Table 2 shows the results of classification for the individual action sequences.

6 Conclusion

In this paper, we have presented a system that can accurately recognize ten different human actions in the frontal or the lateral views. The ten actions are sitting down, standing up, bending over, getting up, walking, hugging, bending sideways, squatting, rising from a squatting position and falling down. The system is not sensitive to variations in the gait of the subject or the height or physical characteristics of the person. Our system was able to correctly recognize subjects of varying height and weight. Thus, it has an advantage over systems that use template matching, in which variations in physical dimensions can produce erroneous results. Further, by modeling the system on the difference in co-ordinates of the head, we do not need to construct three-dimensional models of the subject as a prerequisite to recognition, which is a separate problem in itself.

Our system does, however, have its limitations. So far we have used hand segmentation to isolate the head. Before we can consider recognition in real time, we need to be able to automatically detect and segment the head. Further, our system has had only a limited number of trials. We need to test it on a larger number of sequences to ensure its robustness. Also, thus far it is able to recognize only one action in a sequence. If a person enters a scene and then sits down, it is unable to identify both the walking and the sitting actions. We would like to be able to recognize sequences in which

Test sequences			Correct Classification			Incorrect Classification			% success		
FV	LV	Total	FV	LV	Total	FV	LV	Total	FV	LV	Total
16	23	39	14	17	31	2	6	8	87.5	73.91	79.76

Table 1: Results of Classification

Type of Sequence	Total Number	Correctly Classified	% Success	Type of Sequence	Total Number	Correctly Classified	% Success
Standing	4	3	75	Squatting	4	4	100
Sitting	5	4	80	Rising	4	4	100
Bending down	4	4	100	Hugging	2	1	50
Getting up	5	3	60	Falling	4	3	75
Walking	3	1	33	Bending sideways	4	4	100

Table 2: Classification of the individual action sequences

several actions are concatenated. We intend to work on these problems in the next phase of our implementation and expand our system to be able to identify more complex actions and recognize sequences involving combinations of actions. However, we believe that our system provides a starting point for more complex action recognition. We have, towards this end, also experimented with trying to recognize two actions in a single sequence. The results seem to be promising, however more work is called for before we can present any results.

Acknowledgements

We would like to thank Ms. Debi Paxton for her generous help in editing the paper and also special thanks to Ms. Hua Bin Zhao, Mr. Qasim Iqbal, Mr. Alexander Strehl and Mr. Shishir Shah for their invaluable advice and comments.

References

[1] J. K. Aggarwal and Qin Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, pages 428–440, 1999.

[2] J.K. Aggarwal, Qin Cai, W. Liao, and B. Sabata. Nonrigid motion analysis: Articulated and elastic motion. *Computer Vision and Image Understanding*, pages 142–156, 1997.

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

[4] A. Bobick and J. Davis. Appearance-based motion recognition of human actions. Master's thesis, Massachusetts Institute of Technology, 1996.

[5] Qin Cai and J.K. Aggarwal. Automatic tracking of human motion in indoor scene across multiple synchronized video streams. *International Conference on Computer Vision*, 1998.

[6] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. *IEEE International Conference on Computer Vision Proceedings of the 5th International Conference on Computer Vision*, pages 624–630, 1995.

[7] 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.

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

[9] 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.

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

[11] K Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP, Image Understanding*, 59 n:94–115, 1994.

[12] Nuria Oliver Barbara Rosario and Alex Pentland. A bayesian computer vision system for modeling human interactions. *Proceedings of ICVS99*, 1999.

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

[14] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time sequential images using hidden markov models. *In Proceedings of IEEE Interantional Conference on Computer Vision and Pattern Recognition*, pages 379–385, 1992.

[15] Jie Yang, Yangsheng Xu, and Chiou S. Chen. Human action learning via hidden markov model. *IEEE Transactions on Systems, Man and Cybernetics*, A:34–44, 1997.

Figure 3(a):Sequence of a person standing up in the lateral view.

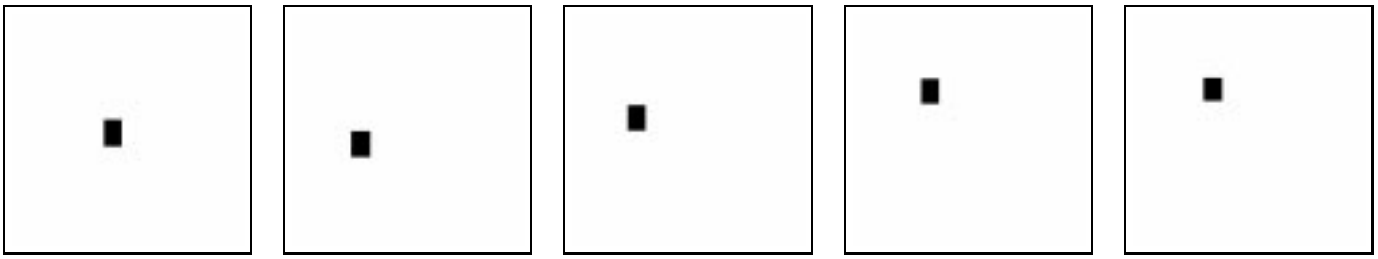

Figure 3(b):Segmented sequence of a person standing up in the lateral view.

Figure 4(a):Sequence of a person bending over in the front view

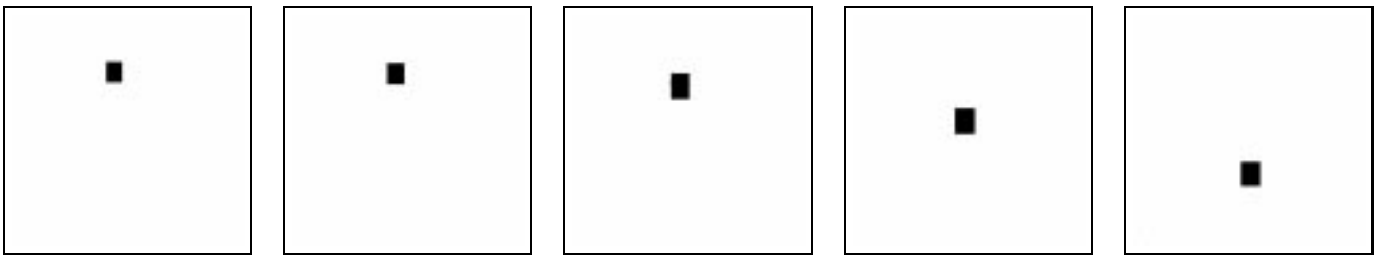

Figure 4(b):Segmented sequence of a person bending over in the front view.

Figure 5(a):Sequence of a person bending sideways in the front view.

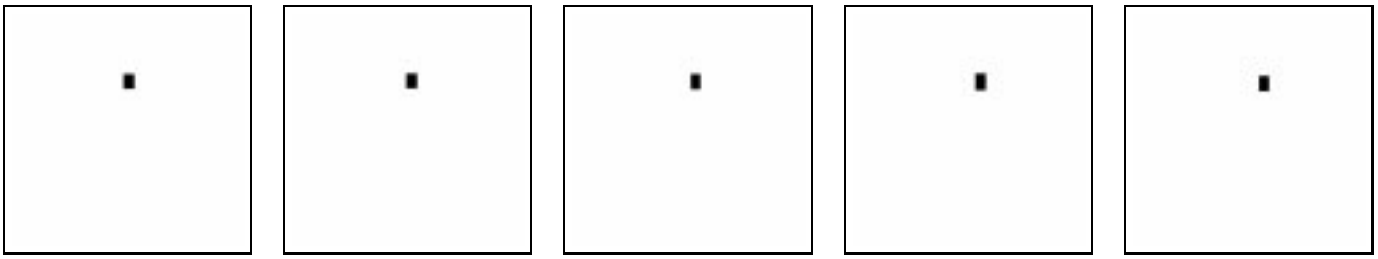

Figure 5(b):Segmented sequence of a person bending sideways in the front view.

Figure 6(a):Sequence of a person hugging another in the lateral view.

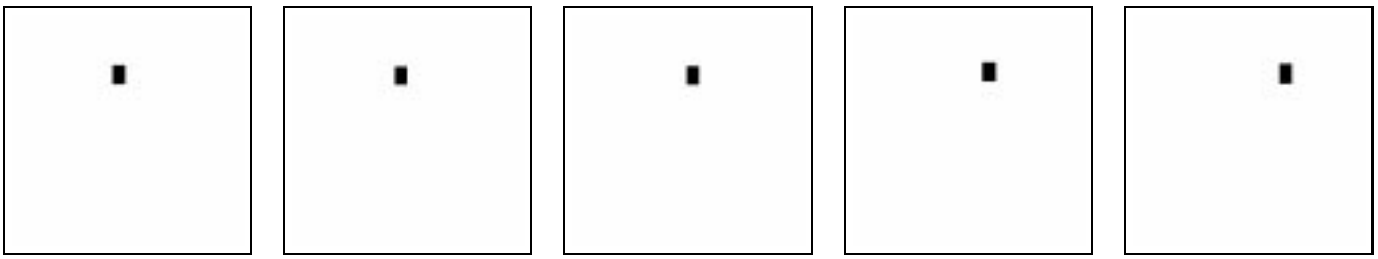

Figure 6(b):Segmented sequence of a person hugging another in the lateral view.