Activity Representation Using 3D Shape Models

Amit K. Roy-Chowdhury
University of California
Riverside, CA 92521
amitrc@ee.ucr.edu

Rama Chellappa
University of Maryland
College Park, MD 20740
rama@cfar.umd.edu

Umut Akdemir
University of Maryland
College Park, MD 20740
uakdemir@cfar.umd.edu

Abstract

Understanding human activities from video sequences is an extremely challenging problem because of the large number of possible events, intra and inter person variations, occlusion of different body parts, etc. Analysis of the trajectories of the various points involved in the activity has been a standard technique for solving such problems. Though trajectories do contain a lot of information regarding the events, merely tracking a set of points is often not enough to characterize an activity. For example, trajectories are not view invariant. However, motion trajectories can be used to build models corresponding to each event/activity, and these models can then be used for classification. In this paper, we propose that each activity can be modeled by a non-rigid 3D shape model. The models are learned from the trajectories of the various points using the factorization theorem for structure and motion estimation. The distance between various models is used for activity classification. An estimate of the number of activities is also obtained by performing a spectral analysis on the trajectory observation vector. Because of the intermediate step of creating the 3D model, the method provides view-invariant interpretation of the events and can be scaled to work with a video sensor network. We present results of our method to understand the activities of a group of moving people in an airport surveillance example and different activities of a single individual.

1. Introduction

Activity modeling and recognition from video sequences has become one of the central problems in computer vision. Traditionally, there has been a keen interest in studying human motion in various disciplines. In psychology, Johansson conducted classic experiments by attaching light displays to various body parts and showed that humans can identify motion when presented with only a small set of these moving dots [5]. Muybridge captured the first photographic recordings of humans and animals in motion in his famous publication on animal locomotion towards the end of the 19-th century[6]. In kinesology the goal has been to develop models of the human body that explain how it functions mechanically [3]. Possible areas of application of computer vision techniques to human motion analysis are video surveillance and monitoring, human computer interaction, video transmission and analysis, medicine, computer graphics and virtual reality. Various techniques have been used for the study of actions from sequences of images (e.g. [2], [4], [7], [1], [11]).

In order to recognize different activities, it is necessary to construct an ontology of various normal (both frequent and rare) events. Deviations from a pre-constructed dictionary can then be classified as abnormal events. It is also necessary that the representation be invariant to the viewing direction of the camera, and independent of the number of cameras (i.e. should be scalable to a video sensor network). Trajectories, usually computed from 2D video data, are a natural starting point for activity recognition systems. Trajectories contain a lot of information about the underlying event that they represent. However, most prevalent systems do little more than tracking a set of points over a sequence of images, and try to infer about the event from the set of tracks. Trajectories are ambiguous (different events can have the same trajectory) and depend on the viewing direction. Also, identifying events from trajectories requires the enunciation of a set of rules (often ad-hoc), which can vary from one instance to another of the same event. Hence, it is important to have a *proper* intermediate step in the leap from trajectories to event models (see Figure 1). In a recent paper, Rao, Yilmaz and Shah [8] proposed a method of representing a trajectory in terms of dramatic changes in its speed and direction (dynamic instants). In [11], the authors propose a shape model (along the lines of Kendall's shape theory) on the set of points in each image frame and describe an activity by the dynamics of the shape.

In this paper, we propose a different approach to bridge the set of trajectories with the class of activity models. The intermediate processing step of Figure 1 is a 3D non-rigid representation of the activity. We propose that each activity can be represented by a non-rigid shape model. The 3D representation captures the 3D configuration and dynamics of the set of points taking part in the activity and is independent of the viewing direction of the camera. Also, the method works whether we have a single camera or a net-

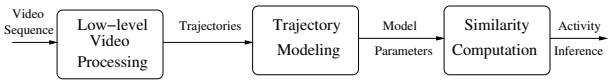


Figure 1: The framework for activity inference.

work of cameras looking at the scene. The 3D shape estimation is done using the factorization theorem [9], modified for non-rigid shapes [10]. A similarity measure between different 3D models is used to classify between the various activities. We demonstrate the applicability of our approach in an airport surveillance problem. We also show how this approach can be used in the problem of video summarization. We would like to clarify that we do not address the issue of obtaining reliable trajectories in this paper, since we consider this to be a separate research problem. A set of heuristics, along with low-level image processing tools, was used to generate reliable tracks in our test video data.

2. Shape Based Activity Models

2.1 Motivation

The model proposed in this paper is a non-rigid 3D structure for each activity, which can be estimated from the trajectories. It is based on the empirical observation that many activities have an associated structure and a dynamical model . Consider, as an example, a dancer or figure skater, who is free to move her hands and feet any way she likes. However, this random movement does not constitute the activity of dancing. For humans to perceive and appreciate the dance, the different parts of the body have to move in a certain synchronized manner. In mathematical terms, this is equivalent to modeling the dance by the structure of the body of the dancer and its dynamics. Similar comments can be made for other activities performed by a single human, e.g. walking, jogging, sitting, etc. An analogous example exists in the domain of video surveillance. Consider people getting off a plane and walking to the terminal, where there is no jet-bridge to constrain the path of the passengers (see Figure 2). Every person after disembarking, is free to move as he/she likes. However, this does not constitute the activity of people getting off a plane and heading to the terminal. The activity here is comprised of people walking along a path that leads to the terminal. Again, we see that the activity is defined by a structure and the dynamics associated with the structure. Using a shape-dynamical model is a higher level abstraction of the individual trajectories and provides a method of analyzing all the points of interest together, thus modeling their interactions in a very elegant way.

2.2 Computing the 3D Model

We hypothesize that each activity can be represented by a linear combination of 3D basis shapes. The difference between the basis shapes can be used to compute the similarity

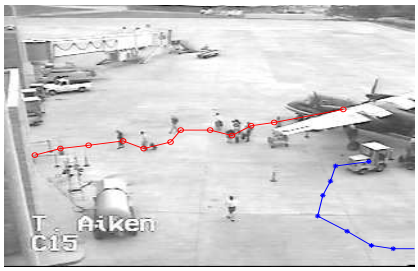


Figure 2: An example of people disembarking from an airplane.

between two activities. Mathematically, if we consider the trajectories of P points taking part in the activity, then the overall configuration of the P points is represented as a linear combination of the basis shapes as

$$S = \sum_{i=1}^{K} l_i S_i, \qquad S, S_i \in \Re^{3\times P}, l \in \Re. \tag{1}$$

The choice of K will depend on the particular application and we will explain the details of it when we describe the experiments. We will assume that we have methods to obtain the trajectories accurately. Also we will assume a weak perspective projection model for the camera.

A number of methods exist in the computer vision literature for estimating the basis shapes. In [9], the authors considered P points tracked across F frames in order to obtain two $F \times P$ matrices $\mathbf{U}$ and $\mathbf{V}$. Each row of $\mathbf{U}$ contains the x-displacements of all the P points for a specific time frame, and each row of $\mathbf{V}$ contains the corresponding y-displacements. It was shown in [9], that for 3D rigid motion under orthographic camera model, the rank, r, of $\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ has an upper bound of 3. In [10], it was shown that for non-rigid motion, the above method could be extended to obtain a similar rank constraint, but one that is higher than the bound for the rigid case. We will use the last mentioned method for computing the basis shapes. We will outline the basic steps of their approach in order to clarify the notation for the remainder of the paper.

Given F frames of a video sequence with P moving points, we can obtain the trajectories of all these points over the entire video sequence. These P points can be represented in a measurement matrix as

$$\mathbf{W}_{2F\times P} = \begin{bmatrix} u_{i,1} & \cdots & u_{i,P} \\ v_{i,1} & \cdots & v_{i,P} \end{bmatrix}_{i=1,\ldots,F}, \tag{2}$$

where $u_{f,p}$ represents the x-position of the p^{th} point in the f^{th} frame and $v_{m,p}$ represents the y-position of the same point.

Under weak perspective projection, the P points of a configuration in a frame f, are projected onto 2D image

points $(u_{f,i}, v_{f,i})$ as

$$\begin{bmatrix} u_{f,1} & \cdots & u_{f,P} \\ v_{f,1} & \cdots & v_{f,P} \end{bmatrix} = \mathbf{R}_f \left(\sum_{i=1}^{K} l_{f,i} S_i \right) + \mathbf{T}_f, \quad (3)$$

where, $\mathbf{R}_f$ represents the first two rows of the full 3D camera rotation matrix and $\mathbf{T}_f$ is the camera translation. The translation can be eliminated by subtracting out the mean of all the 2D points, as in [9]. Henceforth, $\mathbf{W}$ will represent the measurement matrix with the means of each of the rows subtracted out. Using (2) and (3), it is now easy to show that [10]

$$\mathbf{W} = \mathbf{Q}_{2F \times 3K} . \mathbf{B}_{3K \times P}. \quad (4)$$

The matrix $\mathbf{Q}$ contains the pose for each frame of the video sequence and the weights $l_1, ..., l_K$. The matrix $\mathbf{B}$ contains the basis shapes corresponding to each of the activities. In [10], it was shown that $\mathbf{Q}$ and $\mathbf{B}$ can be obtained using singular value decomposition (SVD) as $\mathbf{W}_{2M \times P} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and $\mathbf{Q} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$ and $\mathbf{B} = \mathbf{D}^{\frac{1}{2}}\mathbf{V}^T$.

2.3 Activity Inference

Having obtained the 3D models, the next step is to classify between various activities. Our approach for activity inference consists of a learning/training phase and a testing one. During the training phase, the 3D models for various activities are computed. Given a test sequence, the 3D model estimated from this sequence is compared with that learned before and a similarity score is computed based on a measure of the difference of the two 3D models. The exact method for computing this difference is based on the particular application. We consider the activity of a group of objects (people and vehicles in an airport), each object being represented as a point. The paths followed by the passengers and the vehicles are very different, and the 3D model of these paths are estimated and compared. The values of the weighting coefficients, l_i, are used to differentiate between the activities. Details of the processes are available in Section 3.

3. Experimental Results

In this section, we consider a video surveillance application in an airport scenario. A group of people get off an airplane and walk to the terminal. Also, there are other moving objects like vehicles, airport personnel, etc. The goal is to classify between the activities of the different groups of objects (people vs. vehicles) and to identify an abnormal behavior (e.g. a passenger straying from the normal path), using the information available in the trajectories.

Given a video sequence with each moving point representing the motion of a different activities/objects, we can obtain the trajectories of all these points over the entire video sequence. The trajectory defines the particular activity. For the case of people getting off an airplane, each person is represented by a point. An average trajectory over all the people represents the activity of people getting of the plane. If we have M different **training** video sequences with different instances of the same activity, we can obtain many such example trajectories. Each of the example trajectories can be sampled uniformly to produce a set of P points, each represented as a pair of x and y co-ordinates, for each video sequence. Note that the number of rows in the matrix $\mathbf{W}$ in (2) depends on the number of training sequences, i.e. $F = M$.

During training, we compute the rotation matrix and the average shapes as explained above. For the m^{th} video sequence, consider the rows $(2m-1)$ and $2m$ of the matrix $\mathbf{W}$, and represent it by W_m. It represents the average trajectory of the activities in the m^{th} training sequence. From (3), we see that $l_{m,i}$ can be computed by taking the inner product of W_m with $\mathbf{R}_m S_i$, i.e.

$$l_{m,i} = < W_m, \mathbf{R}_m S_i > \quad (5)$$

for each activity $i = 1, ..., N$ and for each training video sequence $m = 1, ..., M$. Thus for each activity i, we have M values of l_i. These multiple values of l_i represent a significant part of the range of values that can be taken by different instances of these activities. Since a fixed camera is looking at the same set of activities, the rotation matrices will not be very different between the different instances of the same activity (see Fig. 3(a)). Hence, all the l_i for each activity cluster together and can be used for recognition (see Fig. 4(b)).

During **testing**, we consider the trajectory of each object in the video sequence. The procedure described above can be re-applied to the set of tracked points in the sequence in order to obtain the configuration weights by projecting onto the rotated basis shapes, as in (5). The cluster to which the computed l_i belong can be used to identify the activity. The intuitive idea is that the set of weights learned from the training examples cover most of the possible ones for normal activities. Thus, if projections for the test activity lie within a cluster for one of the activities, then we can claim to have recognized that particular activity. In practice, we can set a threshold, $T < M$, for the number of projections that need to lie within a cluster for the activity to be recognized as such. By this method, the activity of each object is individually detected and verified in this 3D shape space. One of the advantages of our method is that it is computationally very inexpensive, since all that it does for classification and verification is to compute projections of tracked features onto basis shapes learned a-priori.

In Figure (4)(a), we plot the average centered shapes (i.e. after the mean of every row of $\mathbf{W}$ is subtracted out) for the two major activities, path of passengers disembarking (represented by S_1) and the path of the luggage cart or fuel tank (represented by S_2). The airport personnel are identified a-

priori and their motion is neglected for the purposes of this analysis. The plot of the various values of $l_{m,1}$ and $l_{m,2}$ for all m, learned from the training sequences, is shown in Figure 4(b), thus showing the clear demarcation between the two activities. In Figure (5)(a), we show the plots of the projections of the activity of passengers deplaning on the two sets of rotated basis shapes, learned during the training phase, i.e. $\mathbf{R}_m S_1$ and $\mathbf{R}_m S_2$, for $m = 1, ..., 150$. The projections of the path taken by the luggage cart on the two sets of rotated basis shapes is shown in Figure (5)(b). The plots in Figure (5) can be used to distinguish between the two activities, given just their motion trajectories by setting an appropriate threshold and declaring an activity to be either one or two, depending on the number of points on either side of the threshold.

The next task is to determine any abnormalities. By this we mean the detection of the case shown in Figure (3)(b). Since the testing is done for each object at a time, the process can identify the concerned individual or object. Since we did not have real video sequences of such behavior, we simulated it by pulling a passenger away from the normal path. Figures (3)(c) plots the projections for the abnormal activity and a normal one on the set of rotated basis shapes. The clear difference in the projections shows the difference in the two activities, which can help to identify the abnormal one.

The Receiver Operating Characteristic (ROC) of the activity detection algorithm, is shown in Figure 4(c). The plots are obtained through simulations by varying the threshold of detection for the two normal activities, as well as the abnormal one. For classification between the two activities, a detection occurs when a test activity, say A, is recognized correctly from the projections onto the set of rotated basis shapes of A, while a false alarm is defined as the case when the projections onto the rotated basis shapes of A of the trajectory obtained from some other activity exceeds the detection threshold. For an abnormal activity, a detection occurs when it is correctly identified as abnormal, while a false alarm occurs when a normal activity is flagged as abnormal.

3.1 Video Summarization

We performed an experiment to summarize a three minute segment of video obtained for the airport surveillance example in the activity shape space using the subspace analysis method. The motion trajectories of all moving objects were considered. They included the passengers, a luggage cart and an airport personnel (whose motion has not been modeled as part of the training procedure, but who can be seen at the bottom of Figure 2). The motion trajectory of each individual object was projected onto the set of rotated basis shapes $\mathbf{R}_m S_i$, for $m = 1, ..., 150$, $i = 1, 2$ learned from the training examples, as explained before. Figure 5(c) shows the projections form three clusters, corresponding to the motion trajectories of 10 passengers, the luggage cart and an airport personnel. These three clusters contain information about all the moving objects in the three-minute segment of the video. Hence we see that it is possible to summarize the motion of all objects in the scene in the shape space.

3.2 Human Motion Analysis

The approach outlined in this paper was also applied to the problem of recognizing the activities of a single individual. For want of space, we cannot describe the details of the experiment here. However, we present the final result in the form of a similarity matrix (Figure 6(b)). Examples of the first basis shape for some of the activities is shown in Figure 7 and the criterion for computing the similarity between the basis shapes is shown in Figure 6(a). It is apparent from the matrix that similar activities are grouped together, thus validating this approach. Moreover, this experiment involved significant changes in viewing direction and the use of multiple cameras.

4. Conclusion

In this paper we have proposed a method for activity modeling and inference using the non-rigid 3D structure of the configuration of points taking part in the activity. The 3D shape is estimated from the motion trajectories of the points under the assumption of scaled orthographic projection. The approach fits into the general framework of inferring high level information about different activities starting from the trajectories. Our approach is independent of the viewing direction of the camera and can be extended to the situation of a video sensor network looking at the scene. Experimental results are shown for classifying between various activities of a group of people and vehicles in an airport surveillance scenario.

References

[1] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *CVPR*, pages 928–934, 1997.

[2] W. Grimson, L. Lee, R. Romano, and C. Stauffer. Using adaptive tracking to classify and monitor activities in a site. In *CVPR*, pages 22–31, 1998.

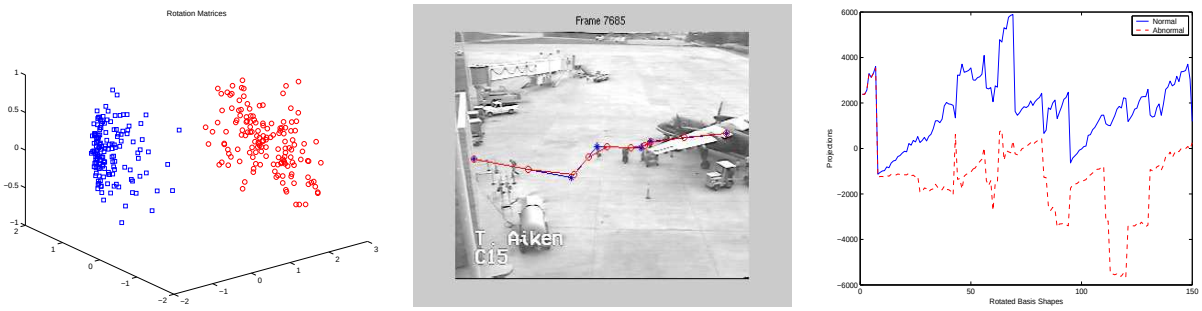
[3] G. Harris and P. Smith (Editors). *Human Motion Analysis: Current Applications and Future Directions*. IEEE Press, 1996.

[4] S. Hongeng and R. Nevatia. Multi-agent event recognition. In *ICCV*, pages II: 84–91, 2001.

[5] G. Johansson. Visual perception of biological motion and a model for its analysis. *PandP*, 14(2 1973):201–211, 1973.

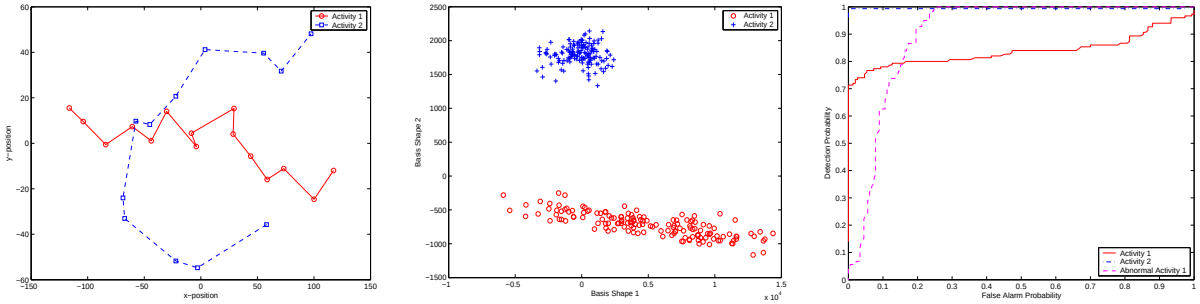
[6] E. Muybridge. *The Human Figure in Motion*. Dover Publications, 1901.

[7] V. Parmeswaran and R. Chellappa. View invariants for human action recognition. In *CVPR*, 2003.



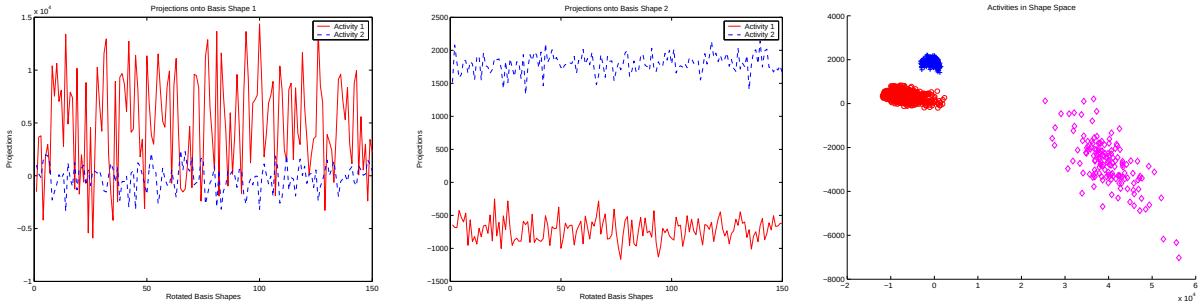
(a) (b) (c)

Figure 3: (a): Plot of the first and second rows of the rotation matrices. (b): An example of an abnormal activity where the average trajectory is distorted to simulate an abnormal behavior. (c): Projections of the abnormal activity and a normal one on the rotated basis shapes for the first activity.



(a) (b) (c)

Figure 4: (a) Plot of the centered shapes formed from the average trajectories of the two activities. (b): Plot of the projections of the various instances of the two activities, as available in the training data, onto the rotated basis shapes. (c): ROC plots for classification of the two normal activities and the abnormal one.



(a) (b) (c)

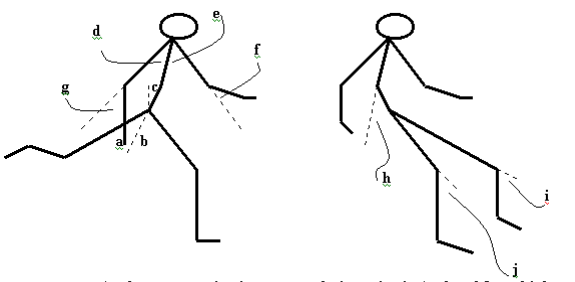
Figure 5: Projections of the two activities on the rotated basis shapes for the first one are shown in (a), while the projections on the rotated basis shapes for the second one are shown in (b). (c): A video summarization example: projections of all the motion trajectories in a three minute segment of the video sequence onto the basis shapes. The red cluster contains the projections of the passengers, the blue of the luggage cart and magenta of the airport personnel whose motion was not modeled as part of the training examples.

[8] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

[9] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9:137–154, November 1992.

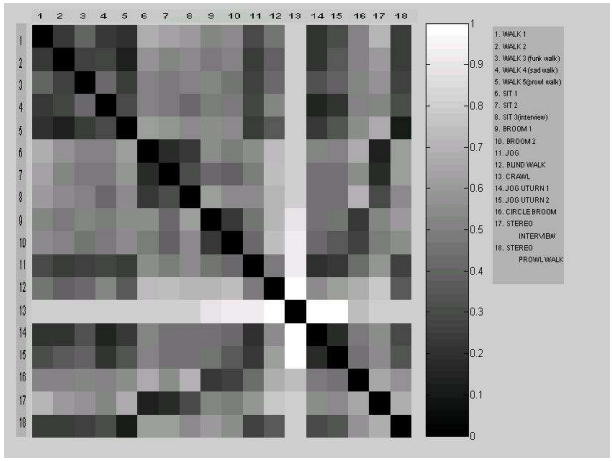
[10] L. Torresani and C. Bregler. Space-time tracking. In *ECCV*, 2002.

[11] N. Vaswani, A. RoyChowdhury, and R. Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. In *CVPR*, 2003.



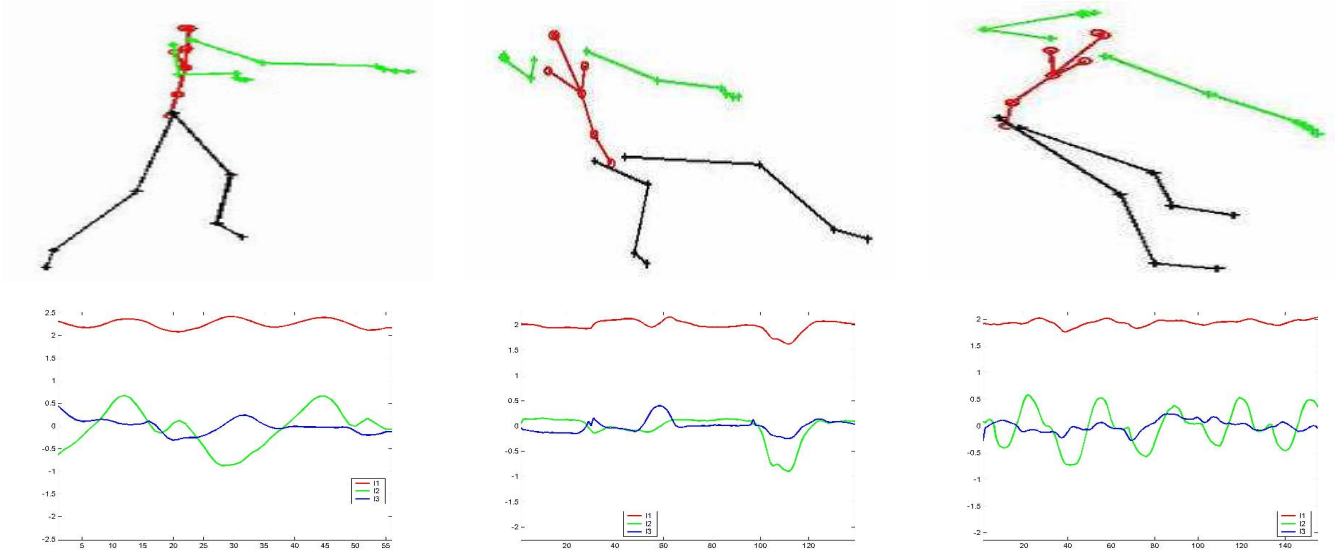
Angles we are using in our correlation criteria (ordered from highest weighted to lowest):

1. c -> angle btw. Hip-abdomen and vertical y axis
2. h -> angle btw. Hip-abdomen-chest
3. (a+b)/2 -> average angle btw. two legs and abdomen-hip axis
4. (b-a)/2 -> the angle difference between two upper legs
5. (i+j)/2 -> average angle btw. upper legs and lower legs
6. (d+e)/2 -> average angle btw. upper arms and abdomen-chest axis
7. (f+g)/2 -> average angle btw. upper arms and lower arms



(a) (b)

Figure 6: (a): The various angles used for computing the similarity of two models is shown in the Figure. The text below describes the seven dimensional vector computed from each model and whose correlation determines the similarity scores. (b): The similarity matrix for the various activities, including ones with different viewing directions and multiple cameras.



(a) (b) (c)

Figure 7: (a) - (c): Plots of the first basis shape, S_1 and combination coefficients l_i (against time) for walk, sit and broom sequences, respectively.