A Factorization Approach for Activity Recognition

Amit Roy Chowdhury, Rama Chellappa
Center for Automation Research
University of Maryland
College Park, MD 20742.

Abstract

Understanding activities arising out of the interactions of a configuration of moving objects is an important problem in video understanding, with applications in surveillance and monitoring. A special situation is when the objects are small enough to be represented as points on a 2D plane. In this paper, we introduce a novel method of representing the activity by the deformations of the point configuration in a properly defined shape space. Instead of inferring about the activity directly from the motion tracks of the individual points, we propose to model an activity by the polygonal shape formed by joining the locations of these point masses at any time t, and its deformation over time. Given the locations of the 2D points over a sequence of frames in the video, the factorization theorem for matrices is used to obtain a set of basis shapes for each activity. An unknown activity can now be recognized by projecting onto these basis shapes. Also, once a specific activity is recognized, the deviations from it can be modeled by the deformations from the basis shape. This is used to identify an abnormal activity. We demonstrate the applicability of our algorithm using real-life video sequences in an airport surveillance environment. We are able to identify the major activities that take place in that setting and detect abnormal ones.

1 Introduction

Monitoring activities from video data is an important surveillance problem, requiring the understanding of the interactions of various objects, as well as their evolution over time. Surveillance problems, therefore, provide rich application areas for event recognition algorithms. In [1], the authors proposed building a tracking and monitoring system using a "forest of sensors" distributed around the site of interest. Their approach involved tracking objects in the site, learning typical motion and object representation parameters (e.g. size and shape) from extended observation periods and detecting unusual events in the site. In [2], the authors proposed a method for recognizing events involving multiple objects using Bayesian inference. The above approaches use the motion tracks of individual objects and their interaction with other objects in the scene for event analysis. A special scenario arises in the case of very low resolution surveillance video where the moving objects are small enough to be modeled as point objects in a 2D plane. Instead of inferring about the activity directly from the motion tracks of the individual objects, we propose a different approach to the problem using the 3D non-rigid shape of the configuration of moving points.

The term shape has been widely used in image understanding for a variety of applications and there exists a huge body of work on shape tracking, analysis and similarity. The problem of studying the similarity of two shapes was posed in a theoretical framework in [3]. Many researchers have adopted different frameworks for analyzing different applications using shape theory. Evaluating the similarity of two shapes by the transformations of the local deformations needed to change one into the other was presented in [4]. A method of creating "active shape models" by learning patterns of variability which are characteristic of the class of objects that it represents was proposed in [5]. Recently, a method for activity analysis using Kendall's statistical shape theory [6] has been proposed [7]. In this paper, we refer to the 3D shape that can be recovered from the motion tracks of points in a 2D image sequence. Towards this end, we extend the idea in [8] for recovering 3D non-rigid shape from 2D image sequences, to the domain of activity analysis and monitoring.

Recognizing activities is an extremely complicated task at which even humans are often less than perfect. It is improbable that there is one single algorithm that would be able to recognize all kinds of activities. Also, pure vision techniques will probably not lead to very robust recognition algorithms. Nevertheless, vision tools can provide very accurate solutions in some scenarios, and very good inputs to higher level logical modules in others. In this paper, we look at one such application where the physical setup offers certain constraints which enables us to apply known techniques in computer vision to recognize certain kinds of activities. Specifically, we consider a rigidly mounted video surveillance camera observing the same kinds of activities happen over and over again. As an example, consider an airport

scenario where after the plane stops, passengers get off and on, luggage carts are loaded and unloaded and the plane is refueled. The activities are very regular and repetitive; however, identifying any anomalies in this regular pattern is an extremely important security issue. Our aim here is to exploit the repetitiveness of the pattern of activities to automatically recognize them and, subsequently, any significant deviations from them.

Our input is a video sequence of the concerned activities that need to be monitored. The different objects that comprise the activity (e.g. people, vehicles, etc.) are represented as points on the 2D ground plane. The locations of these points over the entire video sequence is obtained as a measurement matrix. In [8] (and in their earlier papers on related work), the authors showed that the 2D measurement matrix could be factorized into 3D pose, object configuration and 3D basis shapes using singular value decomposition (SVD). Their work was built on the well-known factorization theorem of linear algebra for the product of two matrices, which was originally used in [9] for solving the structure from motion problem under orthographic projection assumptions. Many extensions of this method have been proposed; the one most relevant for this work is that of Costeira and Kanade that relaxes the single-body constraint [10]. We extend the method of [8] for our problem. We hypothesize that each activity can be modeled by a basis shape corresponding to it. From training videos of the various activities, these basis shapes can be learned. Given a test video sequence containing the unknown activities, the measurement matrix can be formed. The various activities can be identified by computing the projections onto each of the basis shapes. Once an activity is identified, any large deviation from the normal learned nature of that activity can also be inferred in this shape space. Thus we are able to identify the activity, as well as detect any appreciable deviation from normalcy, for each one of them. We work under the scaled orthographic projection approach, which is valid since the objects are far enough from the camera. In this work we assume that the total number of activities occurring in the scene is known. This assumption will be relaxed in future work.

In the next section, we present our theory for modeling activity in 3D shape-space using the factorization theorem. A synopsis of the algorithm is given in Section 3. Section 4 provides an experimental justification for our shape-based activity model. Section 5 presents the experimental results on a real-life video surveillance problem, and Section 6 concludes the paper.

2 Shape-Space Theory for Activity Modeling

We first explain how the problem of non-rigid shape and structure estimation can be recast as an activity modeling and inference problem. We show that it is possible to infer about the nature of different activities, using the fact that 3D non-rigid motion puts rank constraints on 2D image motion. Next, we show how this idea can be used to detect abnormal events within the class of each of those activities.

2.1 Factorization Algorithm for Multiple Activities

Given F frames of a video sequence with moving points representing N different activities, we can obtain the trajectories of all these points over the entire video sequence. An average trajectory for each of the activities can then be obtained. The trajectory defines the particular activity. As an example, consider 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 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 for each video sequence. These M pairs of P points can be represented in a measurement matrix as

$$\mathbf{W}_{2M\times P} = \begin{bmatrix} u_{1,1} & \cdots & u_{1,P} \\ v_{1,1} & \cdots & v_{1,P} \\ \vdots & \vdots & \vdots \\ u_{M,1} & \cdots & u_{M,P} \\ v_{M,1} & \cdots & v_{M,P} \end{bmatrix}, \qquad (1)$$

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

A comparison with the factorization theorem for structure and motion estimation is in order here. 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 has been shown previously in [10], 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 4. The rank constraint is derived from the fact that $\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ can be factored into two matrices $\mathbf{M}_{2F\times r}$ and $\mathbf{S}_{r\times P}$, corresponding to the pose and 3D structure of the scene, respectively. In [8], 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.

Assume that the set of P points in (1) can be represented in the 3D world in terms of a set of basis shapes $S_1, ..., S_K$, where each S_i is a $3 \times P$ matrix describing P points. The overall configuration of the P points is represented as a linear combination of the basis shapes, i.e.,

$$S = \sum_{i=1}^{K} l_i S_i, \qquad S, S_i \in \Re^{3\times P}, l \in \Re. \tag{2}$$

The method described in this paper consists of two parts: a training part and a testing part. During training, given different video sequences of the same activities, these basis shapes are learned. It is assumed that each activity can be represented by a single average shape, with variations modeled as deformations about that shape. We later present experimental justification for this assumption. Thus for N training activities, there are N basis shapes, $S_1, ..., S_N$, i.e. $K = N$. The representation can take into account the fact that deformations from the basis shape will also take place and these are represented by projections onto the basis shapes for the other activities. During testing, given a video sequence with an unknown activity, the activity is first identified using the learned basis shapes for known activities, followed by detection of an abnormality by analyzing the projections onto the basis shapes.

We next consider the measurement matrix $\mathbf{W}$ of size $2M \times P$ in (1). For the moment, assume that we can somehow arrange $\mathbf{W}$ such that the columns representing the points belonging to the same activity are organized together. As an example, assume that the scene consists of only two activities. Also, assume that the set of P points consists of P_1 points from the first activity and P_2 points from the second one. Imagine for the moment that we know which point belongs to which activity and we can arrange $\mathbf{W}$ such that the first P_1 columns represent the first activity and the next P_2 columns the second activity. Under weak perspective projection, the P points of a configuration in a training example m, are projected onto 2D image points $(u_{m,i}, v_{m,i})$ as

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

where,

$$\mathbf{R}_m = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \end{bmatrix}. \tag{4}$$

$\mathbf{R}_m$ represents the first two rows of the full 3D camera rotation matrix and $\mathbf{T}_m$ is the camera translation. The translation can be eliminated by subtracting out the mean of all the 2D points, as in [9]. We now form the measurement matrix $\mathbf{W}$, which was represented in (1), with the means of each of the rows subtracted out.

The weak perspective projection assumed in (3) is usually valid if the range of depths of the object is small compared to its distance from the camera. In that case, it is reasonable to replace the scaling factor $\frac{f}{Z_0+\Delta Z}$ for perspective projection by an average scaling factor $\frac{f}{Z_0}$. Since we assume that the objects are far enough from the camera to be treated as points on a 2D plane, this is a valid assumption. The weak perspective scaling factor is implicitly coded in the configuration weights, $\{l_{m,i}\}$.

Using (1) and (3), it is now easy to show that

$$\mathbf{W} = \begin{bmatrix} l_{1,1}\mathbf{R}_1 & \cdots & l_{1,K}\mathbf{R}_1 \\ l_{2,1}\mathbf{R}_2 & \cdots & l_{2,K}\mathbf{R}_2 \\ \vdots & \vdots & \vdots \\ l_{M,1}\mathbf{R}_M & \cdots & l_{M,K}\mathbf{R}_M \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_K \end{bmatrix} \tag{5}$$

$$= \mathbf{Q}_{2N\times 3K}.\mathbf{B}_{3K\times P}. \tag{6}$$

Thus the measurement matrix has a maximum rank of $3K = 3N$, where N is the total number of activities in the sequence. The matrix $\mathbf{Q}$ contains the pose for each representation of the average trajectory and the weights $l_1, ..., l_K$. The matrix $\mathbf{B}$ contains the basis shapes corresponding to each of the activities.

In [8], it was shown that $\mathbf{Q}$ and $\mathbf{B}$ can be obtained using singular value decomposition (SVD) as

$$\mathbf{W}_{2F\times P} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{7}$$

and $\mathbf{Q} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$ and $\mathbf{B} = \mathbf{D}^{\frac{1}{2}}\mathbf{V}^T$. Obtaining the basis shapes for each of the activities in this way assumes that the columns of $\mathbf{W}$ are arranged according to the different activities, as explained earlier. We will get back to this point a little later.

We still need to obtain the weights $l_{m,1}, ..., l_{m,K}$ and the rotation matrices $\mathbf{R}_m$ for each m. In [8], it was shown that by considering a reordering of a row of the matrix $\mathbf{Q}$, it was possible to obtain these factors, again using SVD. A modification of the SVD approach was also proposed in [8]. We follow a similar approach in order to obtain the configuration weights and rotation matrices. The steps in the mathematical calculations remain the same and we do not feel that it is necessary to repeat them again here. Suffice it is to say that for each activity i we can obtain the rotation matrices $\mathbf{R}_m, m = 1, ..., M$ and the basis shape S_i corresponding to that activity.

We have now laid the groundwork for modeling activity using a set of basis shapes. We have shown that it is possible to recover non-rigid 3D structure of the model representing the activity. Next, we show how this can be used to train and classify various activities and deviations from them.

2.2 Activity Recognition

2.2.1 Training

The first step in the activity recognition algorithm is to obtain the set of basis shapes representing each of the activities. We consider a video surveillance scenario where the camera is fixed and rigidly mounted. The activities that we wish to monitor follow certain nominal trajectories, a mathematical representation of which we seek to learn. Using the method described above, we can obtain the basis shapes $S_1, ..., S_N$ and the rotation matrices $\mathbf{R}_1, ..., \mathbf{R}_M$. For each such video sequence, we consider the rotation matrix and the average shapes. For the m^{th} video sequence, the weights $l_{m,i}, i = 1, ..., N$, $m = 1, ..., M$, can be computed as follows. Consider the rows $(2m-1)$ and $2m$ of the matrix $\mathbf{W}$, and represent it by W_m. It represents the average trajectory 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.

$$l_{m,i} = < W_m, \mathbf{R}_m S_i > \tag{8}$$

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. Hence, all the l_i for each activity cluster together and can be used for recognition.

2.2.2 Identifying Multiple Activities and Detecting Abnormalities

Given a video sequence with unknown activities, the procedure described above can be re-applied to the set of tracked points in the sequence in order to obtain the rotation matrix, basis shapes and configuration weights. The cluster to which the computed l_i belong are can be used to identify the activity. However, detecting an abnormality using this method can lead to errors. This is because, if we follow the method described above, the rotation matrices and S_i may be different because of the large deviation as a result of the abnormal behavior, yet the l_i may lie in one of the different clusters and be detected as normal. Thus the procedure needs to be modified slightly for the detection and verification of the activities.

The points corresponding to the different objects are tracked across the video sequence. The average trajectory of each such object is considered and sampled appropriately so as to obtain a matrix similar to the left hand side of (3). The matrix has two rows consisting of the x and y positions of the sampled points on the trajectory of the object. Let us denote it by W_{test}. For each possible activity i, the inner product of W_{test} with $\mathbf{R}_m S_i$, for every rotation matrix $\mathbf{R}_m$, is computed according to (8). The intuitive idea is that the set of rotation matrices learned from the training examples cover most of the possible ones for normal activities, because of the constrained nature of the problem explained before. Thus if we test the unknown activity with all possible rotation matrices and if each of the projections lies 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 be recognized as such. By this method, the activity of each object is individually detected and verified in this 3D shape space.

We consider a simple example to clarify the above procedure. Consider the activity corresponding to passengers getting off the airplane. Each passenger is tracked and the tracks are given as an input to our recognition algorithm. If the passenger follows a normal path, his/her trajectory would be similar to one of the learned ones and can be modeled by the learned average shape of the path taken by deplaning passengers and the learned rotation matrices. Hence the projections would also lie close to the learned ones. If there is a substantial deviation from the normal trajectory, the learned values of the projections can no longer model it, and it will be detected as an abnormality. The procedure can be repeated for the tracks of all objects, passengers and vehicles alike. In fact, it can be used to identify the object as a passenger or a vehicle, under the assumption that their motion tracks would be very different.

One issue still remains, that of arranging $\mathbf{W}$ such that the points corresponding to the same activity occur in adjacent columns. This is required in the training phase when the different basis shapes and rotation matrices need to be learned. Since this is done off-line during training, any method, including a manual one, can be used. However, we would like to point out that this problem was considered in the multi-body factorization problem in [10] and a method was proposed by considering the effects of swapping the rows and columns of a matrix $\mathbf{Q} = \mathbf{V}\mathbf{V}^T$ (from (7)). However, this is a computationally complex process and we did not use it in our experiments. During the testing phase, the algorithm works by considering the trajectory of each object and verifying its motion. Hence, the question of arranging the columns of $\mathbf{W}$ according to each activity does not arise. However, if the trajectories corresponding to all the activities are considered together in a single matrix $\mathbf{W}$, the method in [10] can be used to group the points for each separate activity together into adjacent columns of $\mathbf{W}$.

3 Algorithm

Training:

1. Given M training video sequences consisting of N possible activities, form the matrix $\mathbf{W}$ in (1) using the average trajectories defining each activity.
2. Obtain $S_1, ..., S_N$ and $\mathbf{R}_1, ..., \mathbf{R}_M$ as explained in Section 2.
3. Obtain the set of weights $l_{m,i}$ according to (8) for all $m = 1, ..., M$ and $i = 1, ..., N$. For each activity i, all the values of l_i form a cluster.
Testing:
1. Given the motion track of an unknown activity, obtain the projections $\mathbf{R}_m S_i$, for all $m = 1, ..., M$ and $i = 1, ..., N$.
2. If $T < M$ projections lie in one of the clusters, the activity belongs to that cluster,
3. If an activity does not lie in any of the clusters, it is inferred to be an abnormal one.

4 Experimental Justification for Shape-Based Activity Model

Figure (1) shows two example frames from two different video sequences representing the different activities around an aircraft standing near the airport terminal. After a plane arrives at the gate, a number of activities take place and are viewed by a camera mounted at a high elevation. Figure (1)(a) shows the average motion trajectory of the two main activities, the path followed by the passengers and the path followed by the luggage cart or the fuel tank. These trajectories are an example of one single instance of these activities occurring. In (1)(b), we show an example of what would be treated as an abnormal event and get flagged as an abnormal activity. We would like to clarify that this kind of activity is simulated by pulling one point away from the normal trajectory. This is because we do not have a video sequence where such an abnormal behavior actually occurs. In this case a passenger moves away significantly from the normal path.[1]

Event analysis for the type of problem that we deal with in this paper typically involves analyzing the trajectories of the various objects and drawing conclusions based on these, as in [1]. Using a shape 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. In Figure (2)(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 described above. It is clear from the plot that the shapes are very different, and successfully exploiting them can lead to a good classification algorithm for the various activities. Also, when an abnormal event occurs (Figure (1)(b)), the trajectory, as represented by the shape, is significantly deformed and can be identified. This is precisely the main idea of the paper.

Another important aspect of the paper on which the algorithm hinges is the assumption that the rotation matrices obtained for the different instances of the activities are numerically close together. Thus, the set of training examples would cover the space of these matrices, which can then be applied on the test sequences. This is important because it ensures that the distinction between the different activities occurs based on the shape only, and not the rotation parameter. To prove this point, we plot the two rows of the different instances of the rotation matrices, obtained from the training examples, as two 3D plots in Figure (2)(b), i.e. the red circles represent $[r1, r2, r3]$ and the blue squares $[r4, r5, r6]$ in equation (4).[2] From the numerical values in the plots, it is clear that the difference between the different instances of $\mathbf{R}$ is small.

5 Experimental Results

We now present the results of our method for recognizing activities in a real-life problem. We consider an airport surveillance situation and present the details of our method using this example. A description of the experimental data used as input to our algorithm was given in the previous section. The video obtained in this situation is of low resolution and given this input, only certain kinds of activities can be monitored. However, it can provide important indications for areas of interest to focus attention on. While the resolution of the video is not sufficient to infer about the interaction between different individuals and/or objects, as in [2], it is of sufficient quality to infer about the activities carried out by groups of people or objects or both. Given the video sequence of what the camera sees, there are two main activities that need to be recognized and verified. They are: i) passengers boarding or getting off the plane and ii) the luggage cart or the fuel tank arriving and leaving. Since we work with motion trajectories, we do not bother about the direction of motion. The motion trajectories are precomputed by a tracking algorithm and, for the purposes of this paper, we assume are available to us. An activity is detected as an abnormal one if the deviations from both of the above ones are greater than a certain threshold. Other than

[1] Whether this is truly an abnormal event is a matter that probably cannot be resolved using purely vision techniques, as we discussed in the Introduction. It is possible that a passenger deviates from the normal path for a perfectly valid reason, e.g. to talk to someone. However, such a conclusion can only be drawn using higher level logical reasoning modules. But a vision algorithm can give precious input to such higher level reasoning modules which can draw the final conclusion. Vision algorithms can assist in simplifying the logical reasoning algorithms, thus leading to greater accuracy of the overall recognition system.

[2] Two separate plots are required to show the 6 components; however, we use the same axes for both the plots.

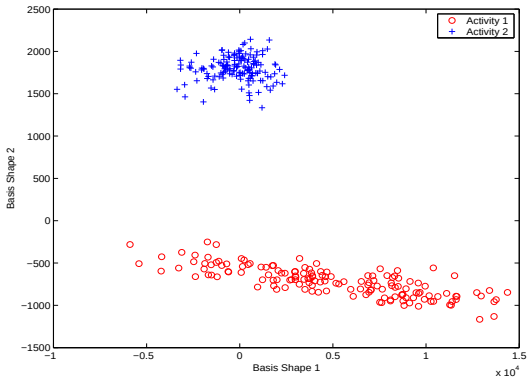


Figure 3: Plot of the projections of the various instances of the two activities, as available in the training data, onto the rotated basis shapes.

the two activities described above, there is random movement of airport personnel. Since there is no pattern to their motion, we cannot hope to learn and recognize their activities. Such personnel can be identified by their clothing, or by other means, and neglected for the purposes of the activity analysis method.

Training for Normal Activities

During the training phase, different instances of the above activities are considered. In our video sequences, we had a few instances of people boarding or deplaning, as well as the luggage cart or fuel tank approaching or receding from the aircraft. The trajectories for these cases were considered for the training set. More examples were created by perturbing these trajectories to simulate similar behavior. The trajectories, formed by tracking the different objects (passengers and vehicles), are sampled at P points in order to form the matrix $\mathbf{W}$ in (1). Once $\mathbf{W}$ is formed, the SVD procedure can be applied to obtain the rotation matrices and basis shapes as explained in Section 2, and plots of which are shown in Figure 2. The entire procedure is extremely fast and takes just a few seconds in a MATLAB implementation with 150 training examples. The projections of the different instances of the trajectories onto the rotated basis shapes can be obtained using equation (8). Thus we obtain the various values of $l_{m,i}$, where $m = 1, ..., 150$ and $i = 1, 2$. The plot of the various values of $l_{m,1}$ and $l_{m,2}$ for all m is shown in Figure 3, thus showing the clear demarcation between the two activities. Since this is done for a large number of training examples, it can be reasonably expected that the projections for the test sequence will lie in one of the above two clusters, if the activity is normal.

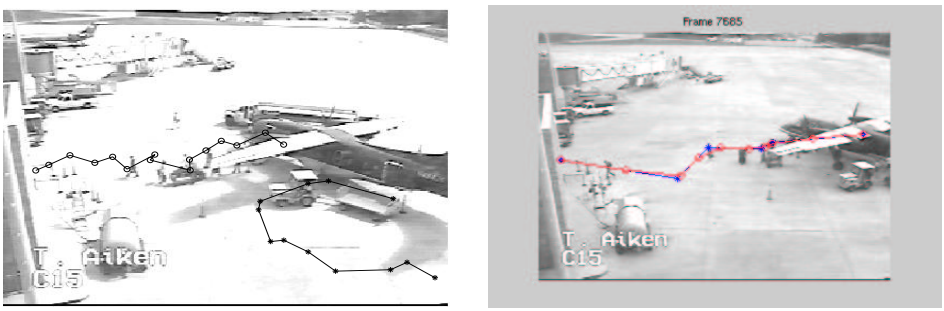
Identifying Normal Activities

Given a test video sequence with unknown activities, we obtain the motion trajectories over the entire sequence of the video. The problem is to verify the activity of each object, described by its trajectory, with respect to the training classes. The trajectories are sampled and the matrix $\mathbf{W}$ with just two rows (similar to equation (3)) is formed with the sampled points. The projections of this matrix with the rotated basis shapes are obtained according to (8). In Figure (4)(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 $m = 1, ..., 150$. Another test case is considered, that of the motion of the luggage cart. Its projections on the two sets of rotated basis shapes is shown in Figure (4)(b). These projections when represented in a two-dimensional plot yields two clusters similar to Figure (3). The plots in Figure (4) can be used to distinguish between the two activities, given just their motion trajectories. This can be done 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. We can thus automatically verify whether each of the different tasks, like passengers boarding a plane or luggage loaded into the cargo hold and the cart departing, were completed successfully or not.

Identifying Abnormal Activities

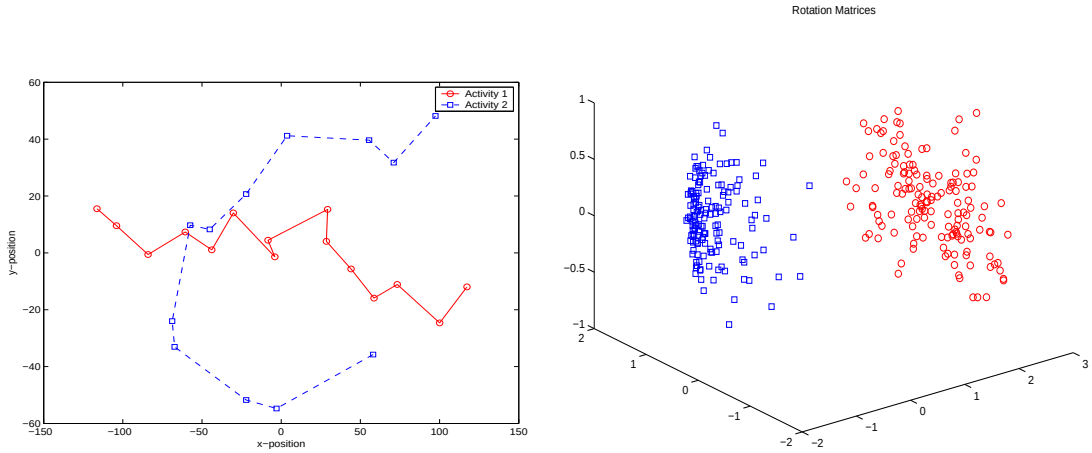
The next task is to determine if either of these activities was not completed successfully. By this we mean the detection of the case shown in Figure (1)(b). Such a decision can be made in isolation to identify an abnormal behavior, or to trigger an alarm in higher level reasoning modules in order to find out whether a prohibited activity has really occurred. Since the testing is done for each object at a time, the process can identify the suspicious individual or object. Again, since we did not have real video sequences of such behavior, we simulated it by pulling certain points away from the normal path.

Similar to the above example, the trajectory is obtained from a tracking algorithm and sampled to form the two rows of the $\mathbf{W}$ matrix. The projections onto the rotated basis shapes $\mathbf{R}_m S_1$ and $\mathbf{R}_m S_2$, for $m = 1, ..., 150$, are computed. Figures (5)(a) and (b) plot the projections for the abnormal activity and a normal one on both sets of rotated basis shapes. The deviations from the rotated basis shapes for activity one are larger than the deviations from the rotated basis shapes for activity two. We know that activity one is modeled primarily by the projections onto the learned basis shapes for that activity, while the the projections onto the basis shape of activity two models the residuals. The fact that the projections of the abnormal activity on the principal component is large indicates that it is a significant deviation



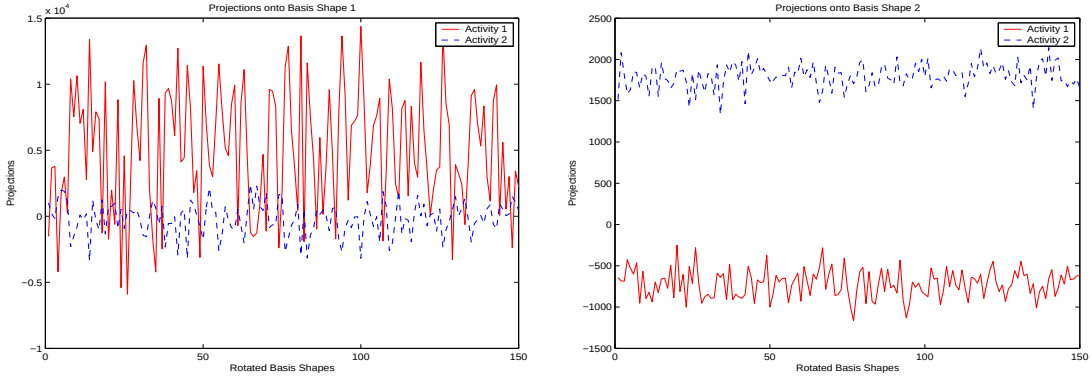
(a) (b)

Figure 1: (a): Two examples of normal activities represented by their trajectories. (b): An example of an abnormal activity where the average trajectory is distorted to simulate an abnormal behavior.



(a) (b)

Figure 2: (a) Plot of the centered shapes formed from the average trajectories of the two activities. (b) Plot of the first and second rows of the rotation matrices.

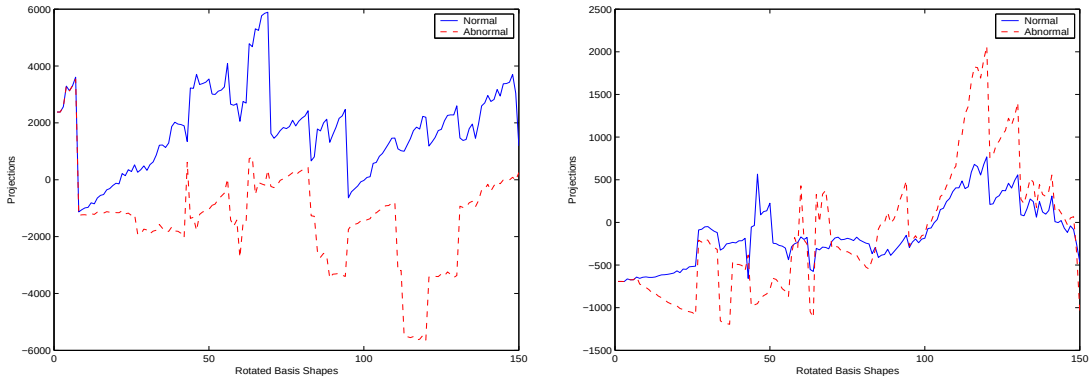


(a) (b)

Figure 4: 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).

from normalcy.

The Receiver Operating Characteristic (ROC) of the activity detection algorithm, as explained in Section 3, is shown in Figure 6. The plots are obtained through sim-



(a) (b)

Figure 5: Projections of the abnormal activity and a normal one on the rotated basis shapes for the first activity are shown in (a), while the projections on the rotated basis shapes for the second activity are shown in (b).

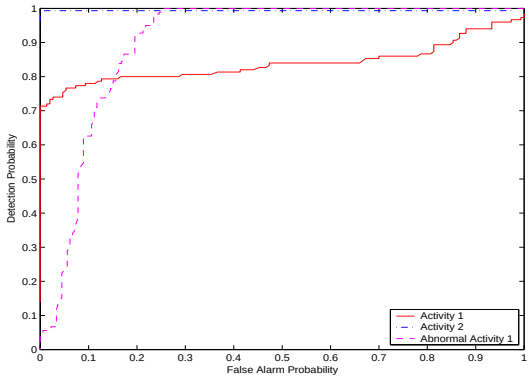


Figure 6: ROC plots for the two normal activities and the abnormal one.

ulations 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 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.

6 Conclusion

We have proposed a method for representing the activity of a dynamic configuration of objects by the shape formed by the trajectories of these objects. The advantage of this method is that it provides an elegant way of treating the interactions between the different objects. We consider a surveillance application using low resolution video, where each moving object can be modeled as a point. Using ideas on the estimation of non-rigid 3D shapes, we show that we can estimate our model parameters (3D shape and rotation) from a set of training examples showing different instances of the activities. This is done by extending the well-known factorization theorem to the domain of activity recognition. Given an unknown activity, the projection of the tracked points onto the model parameters can be used to automatically identify the activity, as well as detect any abnormal behavior. We presented detailed results of our method on a real life video surveillance problem involving the activities that occur near an aircraft after it arrives at its gate. One of our future areas of research is to automatically identify the number of such significant activities.

References

[1] W.E.L. Grimson, L. Lee, R. Romano, and C. Stauffer, "Using adaptive tracking to classify and monitor activities in a site," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998, pp. 22–31.

[2] S. Hongeng and R. Nevatia, "Multi-agent event recognition," in *IEEE International Conference on Computer Vision*, 2001, pp. II: 84–91.

[3] D. Mumford, "Mathematical theories of shape: Do they model perception?," *SPIE*, vol. 1570, pp. 2–10, 1991.

[4] R. Basri, L. Costa, D. Geiger, and D.W. Jacobs, "Determining the similarity of deformable shapes," *Vision Research*, vol. 38, pp. 2364–2385, 1998.

[5] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active shape models: Their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, January 1995.

[6] D.G. Kendall, D. Barden, T.K. Carne, and H. Le, *Shape and Shape Theory*, John Wiley and Sons, 1999.

[7] Authors, "Activity recognition using the dynamics of the configuration of interacting objects," in *Submitted to CVPR*, 2003.

[8] L. Torresani and C. Bregler, "Space-time tracking," in *European Conference on Computer Vision*, 2002.

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

[10] J. Costeira and T. Kanade, "A multibody factorization method for independent moving objects," *International Journal on Computer Vision*, vol. 29, no. 3, September 1998.