Activity Recognition Using the Dynamics of the Configuration of Interacting Objects *

Namrata Vaswani, Amit Roy Chowdhury, Rama Chellappa
Dept. of Electrical and Computer Engineering,
Center for Automation Research,
University of Maryland, College Park, MD 20742, USA
{namrata,amitrc,rama}@cfar.umd.edu

Abstract

Monitoring activities using video data is an important surveillance problem. A special scenario is to learn the pattern of normal activities and detect abnormal events from a very low resolution video where the moving objects are small enough to be modeled as point objects in a 2D plane. Instead of tracking each point separately, we propose to model an activity by the polygonal 'shape' of the configuration of these point masses at any time t, and its deformation over time. We learn the mean shape and the dynamics of the shape change using hand-picked location data (no observation noise) and define an abnormality detection statistic for the simple case of a test sequence with negligible observation noise. For the more practical case where observation (point locations) noise is large and cannot be ignored, we use a particle filter to estimate the probability distribution of the shape given the noisy observations upto the current time. Abnormality detection in this case is formulated as a change detection problem. We propose a detection strategy that can detect both 'drastic' and 'slow' abnormalities. Our framework can be directly applied for object location data obtained using any type of sensors - visible, radar, infra-red or acoustic.

1 Introduction

Monitoring activities from video data is an important surveillance problem. A special scenario is to learn the pattern of normal activities and detect abnormal events from very low resolution video where the moving objects are small enough to be modeled as point objects in a 2D plane. 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 patterns and co-occurrence statistics of different objects from extended observation periods and using these to detect unusual events. In [2], the authors used Bayesian networks to represent multi-agent events. The above approaches use the motion tracks of individual objects and their interaction with other objects in the scene. Instead of tracking point objects separately and then learning their interactions, we propose a different approach which models 'group activity' using Kendall's statistical shape theory [3]. A 'group activity' is represented by the polygonal 'shape' formed by joining the locations of the point objects (henceforth referred to as 'points') or *'landmarks'* at any time t and its deformation over time. This provides a compact global framework to jointly model the motion of all the moving objects performing an activity (group activity). We are able to identify "spatial" abnormalities, e.g. deviations from the normal path, as well as "temporal" abnormalities [1], e.g. sudden stopping for prolonged periods of time when the normal activity should be continuous motion. The advantage of using the shape of the configuration of objects is that it is invariant to translation and in-plane rotation of the camera (assuming orthographic projections).

Shape is defined as all the geometric information that remains when location, scale and rotational effects are filtered out [4]. Some of the earliest works in shape theory are Fourier descriptors [5] and extended Gaussian image model [6] both of which model shape of continuous curves. Also, there exists a huge body of work in the vision community on shape tracking, analysis and similarity [7, 8, 9, 10]. Statistical shape theory [3] for the shape formed by discrete points or landmarks began in the late 1970s and has evolved into practical statistical approaches for analyzing objects using probability distributions of shape. Of late, it has been applied to some problems in image analysis, object recognition and image morphing (Chapters 11 and 12 of [4]). All these examples, however, model the shape of a single object in static images. Our work presents an approach for extending this method for modeling the dynamics of the shape formed by the locations of a group of objects performing an activity.

* Partially supported by the DARPA/ONR Grant N00014-02-1-0809

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

For a "static shape activity" (explained in section 3), the mean shape of the object configuration remains constant with time and this is the case that we deal with in this paper. Consider as an example of "static shape activity", the video sequence of passengers getting out of a plane and moving towards the terminal (see figure 1 (a)). All passengers are supposed to follow the same path from the plane to the terminal. If one were to look at the shape formed by connecting the locations of all the passengers at any time instant it would look similar, except for deformations due to variations in the path taken by each individual. Suspicious activity in this example would be any person walking in an unexpected direction thus causing the shape of the configuration of passengers to change (figure 1(b)) or a person stopping in between which would also cause shape to change when the person behind the stopped person goes ahead of him.

Given a training sequence from a "static shape activity", we use the observed object configurations from a sequence of frames to learn the mean "activity shape". We define a tangent coordinate system at the mean shape as described in [4]. The tangent hyper-plane is a linear vector space that approximates the nonlinear shape space in the vicinity of the mean shape. The shape dynamics in tangent space is modeled using a Gauss-Markov model as discussed in our earlier work [11]. For the case of observations with negligible observation noise, we evaluate tangent coordinates of the shape of the test sequence and use log likelihood to detect abnormality.

In this paper, we consider the more practical (and difficult) case of large observation noise in the observed point locations. We now have a partially observed and nonlinear dynamical system [12] from which we need to estimate the shape (actually its posterior distribution) and also detect abnormality. This problem fits into the framework of particle filtering. Particle filtering is a sequential Monte Carlo method that was first introduced in [13] as an approach to non-linear, non-Gaussian Bayesian state estimation. Particle filters (PF) have been used in computer vision for shape based tracking of *a single* object using various representations of shape [14, 15]. [16] uses particle filtering to track multiple moving objects but it uses separate state vectors for each object and defines data association events to associate the state and observation vectors. But in our work, we represent the combined state of all moving objects using shape (tangent) coordinates. We use a PF only to estimate (filter out) the 'shape' of the configuration of moving objects from noisy observations of their locations.

Abnormality detection in this case is formulated as a change detection problem. Most algorithms for change detection are for linear systems. [17] is a reference for change detection in nonlinear systems using PFs but it assumes an abrupt change and known parameters after the change. In many situations an abnormality is a slow change and its parameters are unknown. We propose in this paper a change detection strategy that can deal with both 'drastic' (or abrupt) and 'slow' changes with change parameters unknown.

The rest of the paper is organized as follows. In section 2, we give a brief review of statistical shape theory and particle filtering. Section 3 describes the 'shape activity' model that we introduced in [11] and how to detect abnormality in the fully observed case. In section 4, we describe a particle filtering approach to estimate the posterior distribution of the shape from noisy observations of the configuration and a change detection strategy to detect abnormality. Experimental results are presented in section 5 and conclusions in section 6.

2 Preliminaries

2.1 Statistical Shape Theory

We briefly review the basic tools for statistical shape analysis as described by Dryden and Mardia in [4]. We use Kendall's representation of a shape configuration in m dimensional space as the $k \times m$ matrix formed by the locations of k landmark points on each specimen. For $m = 2$ dimensional shape a more convenient representation is a k dimensional complex vector with real and imaginary parts representing the x and y coordinates of the point. The mapping from configuration space tangent coordinates for shape involves the following steps:

Translation Normalization: In order to make the shape invariant to translation, the complex vector of raw location data (Y_{raw}) can be centered by subtracting out the mean of the vector, i.e.

$$Y = CY_{raw} \;\; where \;\; C = I_k - \frac{1_k {1_k}^T}{k}, \qquad (1)$$

I_k is a $k \times k$ identity matrix and 1_k is a k dimensional vector of ones.

Scale Normalization: *Preshape* is the geometric information that remains after location and scaling information has been filtered out. It is obtained by normalizing Y by its Euclidean norm, $s = ||Y||$, i.e.

$$z_Y = \frac{Y}{s}. \qquad (2)$$

Distance between shapes: A concept of distance between shapes is required to fully define the non-Euclidean shape metric space. The shape space is non-Euclidean (it is a spherical manifold) because of the scaling to norm one. The *full Procrustes distance* [4] of a centered complex configuration Y_1 from Y_2 is given by the Euclidean distance between the full Procrustes fit of the preshape of Y_1, (z_{Y_1}),

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

onto the preshape of Y_2, (z_{Y_2}). *Full Procrustes fit* is chosen to minimize

$$d(Y_2, Y_1) = ||z_{Y_2} - z_{Y_1}\beta e^{j\theta} - (a + jb)1_k||. \quad (3)$$

Full Procrustes distance, $d_F(Y_2, Y_1)$ is this minimum distance i.e. $d_F(Y_2, Y_1) = \inf_{\beta,\theta,a,b} d(Y_2, Y_1)$. Since the preshapes z_{Y_1} and z_{Y_2} have already been normalized for translation and scale, the translation value that minimizes $d(Y_1, Y_2)$, $\hat{a} + j\hat{b} = 0$, and the scale, $\hat{\beta} = |z^*_{Y_1} z_{Y_2}|$ is very close to one. The rotation angle, $\hat{\theta} = arg(z^*_{Y_1} z_{Y_2})$.

For a population of similar shapes, a full Procrustes mean shape ($\hat{\mu}$) is obtained by minimizing (over μ) the sum of squares of full Procrustes distances from each observation Y_i in the population to the unknown mean shape, μ, i.e.

$$[\hat{\mu}] = \arg\inf_{\mu} \sum_{i=1}^{n} {d_F}^2(Y_i, \mu). \quad (4)$$

For 2D shapes, the full Procrustes mean $\hat{\mu}$ can be found as the eigenvector corresponding to the largest eigenvalue of the matrix $S = \sum_{i=1}^{n} z_{Y_i} z^*_{Y_i}$ [18]. Obtaining the full Procrustes mean and aligning all preshapes in the dataset to it (by finding their full or partial Procrustes fit to the mean) is known as *Generalized Procrustes Analysis*. Partial Procrustes fit is obtained by setting $\beta = 1$ and solving only for the rotation angle in (3) to align the preshape to the mean. (See chapter 3 of [4] for details).

Shape Variability in Tangent Space: To examine the structure of shape variability from the average shape, we define a linearized space (tangent space) about the mean shape and consider variation from the mean in this linearized space. The preshape formed by k points lies on a complex hypersphere of unit radius. The aligned preshapes (after generalized Procrustes analysis) of a dataset of similar shapes would lie close to each other and to their Procrustes mean on this hypersphere. The tangent hyperplane to the hyper-sphere at the mean is an approximate linear space to represent this dataset and in this space, standard linear multivariate analysis techniques can be applied.

The partial Procrustes tangent coordinates [4] of a preshape (z), taking the Procrustes mean, μ, as the pole for the tangent projection, are obtained by projecting the partial Procrustes fit (w.r.t. μ) of a preshape, into the tangent space at the mean. They are evaluated as [4]

$$\begin{aligned} \theta(z, \mu) &= arg(z^*\mu) \\ v(z, \mu) &= [I_k - \mu\mu^*]e^{j\theta}z \end{aligned} \quad (5)$$

where z is the preshape. Note that the tangent coordinates lie in a $2k-4$ dimensional real hyperplane (two dimensions reduced due to X and Y translation normalization, one due to scale and one due to rotation normalization).

The inverse of the above mapping (tangent space to preshape space) is

$$z(v, \theta, \mu) = [(1 - v^*v)^{1/2}\mu + v]e^{j\theta} \quad (6)$$

The configuration is given by scaling the preshape by its scale (s), $Z = sz$.

2.2 Particle filtering (PF)

Let the state process $X = \{X_t\}$ be an $\mathcal{R}^{n_x}$-valued Markov process with a Feller transition kernel [12] $K_t(x_t, dx_{t+1})$ (where $\{x_t\}$ is a realization of the random process X_t). Let the observation process $Y = \{Y_t\}$ be an $\mathcal{R}^{n_y}$-valued stochastic process defined as $Y_t = h(X_t, t) + w_t$. The initial state distribution is denoted by $\pi_0(x)$ and the observation likelihood at time t given the state by $g_t(y_t|x_t)$. The particle filter [12] recursively approximates the optimal posterior distribution of the state at any time t given the past observations, by Monte Carlo sampling. It works for any non-linear, non-Gaussian dynamical system for which π_0, $K_t(x_t, dx_{t+1})$ is known and can be sampled from and $g_t(y_t|x_t)$ is known.

The filter [12] starts with sampling n times from the initial state distribution $\pi_0(x)$ to approximate it by $\pi^n_0(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_0^{(i)}}(x)$. Now assuming that the distribution of X_{t-1} given observations upto time $t-1$ has been approximated as $\pi^n_{t-1|t-1}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{t-1}^{(i)}}(x)$, the prediction step samples the new state $\bar{x}_t^{(i)}$ from the distribution $K_{t-1}(x_{t-1}^{(i)}, .)$. Thus the emperical distribution of this new cloud of particles, $\pi^n_{t|t-1}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\bar{x}_t^{(i)}}(x)$ is the probability distribution of X_t given observations upto time $t-1$. For each particle, its weight is proportional to the likelihood of the observation given that particle, i.e. $w_t^{(i)} = \frac{n g_t(y_t|\bar{x}_t^{(i)})}{\sum_{i=1}^{n} g_t(y_t|\bar{x}_t^{(i)})}$. $\bar{\pi}^n_{t|t}(x) = \frac{1}{n}\sum_{i=1}^{n} w_t^{(i)} \delta_{\bar{x}_t^{(i)}}(x)$ is an estimate of the probability distribution of the state given observations uptil time t. We resample n times with replacement from $\bar{\pi}^n_{t|t}(x)$ to obtain the emperical estimate $\pi^n_{t|t}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_t^{(i)}}(x)$. Note that both $\bar{\pi}^n_{t|t}$ and $\pi^n_{t|t}$ approximate $\pi_{t|t}$ but the resampling step increases the sampling efficiency by eliminating samples with very low weights.

3 'Shape' Activity Model

We use Dryden and Mardia's statistical shape theory ideas (described above to represent the shape of "an" object) to model the shape of the configuration of a group of moving objects and its deformations over time. The notion of separating the motion of a deforming shape into motion of an average shape and its deformations described by Soatto

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

and Yezzi in [19] can be extended to "shape activities". We define a "static shape activity" as one in which the average shape formed by the moving points remains constant with time and the deformation process is stationary. A "dynamic shape activity" on the other hand has a time varying average shape and/or a non-stationary shape deformation process.

Kendall's shape analysis methods (discussed above) describe the shape of a fixed number of landmarks and so when the number of point objects is not fixed with time, we resample the curve obtained by connecting the object locations at time t to represent it by a fixed number of points, k. The order in which the object locations are joined is kept the same (shape is not invariant to change in ordering of the points).

The complex vector formed by these k points (x and y coordinate forming the real and imaginary parts) is then centered using equation (1) to give the observation vector sequence, $\{Y_t\}$. We assume in this section that hand-picked or accurately measured object location data is available (negligible observation noise). The observation vector is normalized for scale (to obtain the preshape) and generalized Procrustes analysis (equation (4)) is performed on this sequence of pre-shapes to obtain the Procrustes mean shape, μ. The preshapes are aligned to μ and tangent coordinates at μ evaluated using equation (5). The complex tangent coordinate vector is rewritten as a real vector of twice the complex dimension.

3.1 Shape Dynamics in Tangent Space

Let the vector of tangent coordinates be represented by $v_t \in \mathcal{R}^{2k-4}$. The origin of the tangent hyperplane is chosen to be the tangent coordinate of μ and hence the data projected in tangent space has zero mean by construction. The time correlation between the tangent coordinates is learnt by fitting a stationary *Gauss Markov model* as described in our earlier work [11], i.e.

$$\begin{aligned} E[v_t] &= 0 \\ v_t &= Av_{t-1} + n_t, \end{aligned} \tag{7}$$

where n_t [1] is a zero mean i.i.d. Gaussian process and is independent of v_{t-1}. The details of evaluating the covariance matrix of v_t, Σ_v, the autoregression matrix A and covariance of noise Σ_n (assuming stationarity and ergodicity) are discussed in [11]

Based on the stationary Gauss Markov model described above we have,

$$\begin{aligned} f^0(v_t) &\sim \mathcal{N}(0, \Sigma_v), \ \forall t \\ f^0(v_{t+1}|v_t) &\sim \mathcal{N}(Av_t, \Sigma_n). \end{aligned} \tag{8}$$

[1] Note that to simplify notation, we do not distinguish between a random process and its realization in the rest of the paper.

Thus any $L+1$ length sequence, $\{v_{t-L}, ...v_{t-1}, v_t\}$, will have a joint Gaussian distribution.

3.2 Abnormality Detection: Fully Observed Case

We have assumed in this section that the noise in the shape of the observations is negligible compared to the system noise, n_t, and hence we have a fully observed dynamical model. For such a test observation sequence, we can evaluate the tangent coordinates (v_t) directly from the observations (Y_t) using equations (2) followed by (5).

The following hypothesis is used to test for abnormality. A given test sequence is said to be generated by a *normal activity iff* the probability of occurrence of its tangent coordinates using the pdf defined by (8) is large (greater than a certain threshold). Thus the distance to activity statistic for an '$L+1$' length observation sequence ending at time t, $d_{L+1}(t)$, is the negative log likelihood (without the constant terms) of the tangent coordinates of the observation i.e.

$$\begin{aligned} d_{L+1}(t) &= v_{t-L}^T \Sigma_v^{-1} v_{t-L} \\ &+ \sum_{\tau=t-L+1}^{t} (v_\tau - Av_{\tau-1})^T \Sigma_n^{-1} (v_\tau - Av_{\tau-1}) \end{aligned} \tag{9}$$

We test for abnormality at any time t by evaluating $d_{L+1}(t)$ for the past $L+1$ frames. In the results section, we refer to this as the 'log likelihood metric' (even though it is not actually a 'metric').

4 Partially Observed 'Shape' Activity Model

In the previous section, we defined an abnormality detection statistic for the case of negligible observation noise (fully observed system). But, when noise in the observations (projected in shape space) is comparable to the system noise, the above model will fail (See figure 2(c)). This is because tangent coordinates estimated directly from this very noisy observation data would be highly erroneous. Observation noise in the point locations will be large in most practical applications especially with low resolution video. In this case, we have to solve the joint problem of *filtering* out the actual configuration (Z_t) and the corresponding shape from the noisy observations ($Y_t = Z_t + w_t$) and also *detecting* abnormality (as a change in shape). Since Z_t is now unknown, so is the corresponding v_t and we thus have a partially observed non-linear dynamical system [12] with the following system (state transition) and observation model.

The system model includes the shape space dynamics (the Gauss-Markov model on tangent coordinates) and also

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

the dynamics of the scale and rotation [2]. Even though we are interested only in shape dynamics, modeling the rotation and scale dynamics as a first order stationary process helps to filter out sudden changes in scale/rotation caused by observation noise, which would otherwise get confused as sudden changes in shape space. Also, this dynamical model on scale, rotation (or translation) could model random motion of a camera due to its being inside a UAV (unmanned air vehicle) or any other unstable platform.

The observation model is a mapping from state space (tangent coordinates for shape, scale and rotation) back to configuration space, with noise added in configuration space.

4.1 System Model

The state vector, X_t is composed of $X_t = [v_t^T, \theta_t, \mathbf{s}_t]^T$ where v_t are the tangent coordinates of the unknown configuration Z_t, $\theta_t = arg(Z_t^* \mu)$ is the rotation normalization angle, and $\mathbf{s}_t = ||Z_t||$ is the scale. The transition model for shape (v_t) is discussed in section 3.1. The scale parameter at time t is assumed to follow a Rayleigh [3] distribution about its past value. The rotation angle is modeled by a uniform distribution with the previous angle as the mean. We have

$$\begin{aligned} v_t &= A v_{t-1} + n_t, \quad n_t \sim \mathcal{N}(0, \Sigma_n) \\ s_t &= r_t s_{t-1}, \quad r_t \sim Rayleigh(\sqrt{2/\pi}) \\ \theta_t &= \theta_{t-1} + u_t, \quad u_t \sim \text{Unif}(-a, a) \end{aligned} \tag{10}$$

with initial state distribution

$$\begin{aligned} v_0 &\sim \mathcal{N}(0, \Sigma_v) \\ s_0 &\sim \textit{Rayleigh}\,(\bar{s}_0) \\ \theta_0 &\sim \textit{Unif}\,(\bar{\theta}_0 - a_0, \bar{\theta}_0 + a_0) \end{aligned} \tag{11}$$

The model parameters A, Σ_n, Σ_v are learnt using a single training sequence of a normal activity and assuming stationarity for v_t as described in 3.1. The parameter a is learnt as $a = \max_t |\theta_t - \theta_{t-1}|$. Note that in this paper we have assumed a stationary system model for v_t. But in general, the framework described here is applicable even if Σ_v, A, Σ_n are time varying (non-stationary process).

4.2 Observation Model

In our current implementation, we assume that independent Gaussian noise with variance σ_{obs}^2 is added to the actual location of the points, i.e. [4]

$$\begin{aligned} Y_t &\sim \mathcal{N}(Z_t, \sigma_{obs}^2 I_{2k}) \quad \textit{where} \\ Z_t &= h(X_t) = s_t[(1 - v_{t_c}^* v_{t_c})^{1/2} \mu + v_{t_c}] e^{-j\theta_t} \end{aligned} \tag{12}$$

where $h(X_t)$ is the function given by equation (6) followed by scaling by s_t.

In general, both σ_{obs}^2 and μ can be time varying and the observation noise need not be i.i.d. in all the point object locations. Also, to take care of outliers, one could allow a small probability (p_{out}) of any point j occurring anywhere in the image with equal probability (uniform distribution).

4.3 Particle Filter

We use the state transition kernel given in (10) and the observation likelihood given by (12) in the particle filtering framework described in section 2.2. The PF provides at each time t, an n sample emperical estimate of the distribution of the state at time t given observations upto time $t-1$ (prediction) and the distribution of the state given observations upto time t, $\pi_{t|t}^n(v_t, s_t, \theta_t)$ (update). For abnormality detection, only the marginal of shape, $\pi_{t|t}^n(v_t)$ is used.

4.4 Abnormality Detection

We test for abnormality based on the following hypothesis. A test sequence of observations, $\{Y_t\}$ is said to be generated by a *normal activity iff*
(a) It is "correctly tracked" by the particle filter trained on the dynamical model learnt for a normal activity. We test this by thresholding the distance between the observation and its prediction based on past observations i.e. for normalcy,

$$(Y_t - E[Y_t | Y_{0:t-1}])^2 = (Y_t - E_{\pi_{t|t}}[h(X_t)])^2 < \gamma \tag{13}$$

and
(b) The expectation under $\pi_{t|t}(v_t)$ of the negative log-likelihood of normalcy of the tangent coordinates (expectation under $\pi_{t|t}$ of $d_1(t)$ from equation (9)) is below a certain normalcy threshold, η, i.e.

$$E \stackrel{\triangle}{=} E_{\pi_{t|t}}[-log f^0(v_t)] < \eta \tag{14}$$

The PF estimate $\pi_{t|t}^n$ will approximate $\pi_{t|t}$ correctly only if the observations are "correctly tracked" by the PF and hence only in the "correctly tracked" case, E can be estimated using the PF distribution. Also, note that E is actually the Kullback Leibler distance between the pdf corresponding to $\pi_{t|t}$ and the normal activity pdf of tangent coordinates, f^0,

[2] In our current implementation, we have not modeled translation dynamics (we use a translation normalized observation vector) assuming that observation noise does not change the centroid location too much.

[3] Rayleigh distribution chosen to maintain non-negativity of the scale parameter

[4] $v_{t_c} \in \mathcal{C}^{k-2}$ is the complex version of $v_t \in \mathcal{R}^{2k-4}$ (inverse of operation described in the paragraph just before section 3.1)

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

plus the entropy of $v_t|Y_{0:t}$. Hence this statistic is referred to as the K-L metric [5] in the results section.

The expression for E is approximated by E_n as follows

$$E_n \triangleq E_{\pi^n_{t|t}}[-logf^0(v_t)] = \frac{1}{n}\sum_{i=1}^{n} {v_t^{(i)}}^T \Sigma_v^{-1} v_t^{(i)} + C \quad (15)$$

(where $C \triangleq \log\sqrt{(2\pi)^{2k-4}|\Sigma_v|}$).

Now, a 'drastic' abnormality will cause the PF to lose track and hence will get detected using (a). If the abnormality is a 'slow' one (say a person slowly moving away in a wrong direction), the PF will not lose track. But a systematically increasing bias is introduced in the tangent coordinates (they no longer remain zero mean) and hence the expected negative log likelihood of normalcy will be large in this case causing (b) to be violated. Since the PF does not lose track in this case, the PF distribution estimates $\pi_{t|t}$ correctly and hence E can be estimated using a PF in this case.

5 Experimental Results

A video sequence of passengers getting out of a plane and walking towards the terminal (figure 1) was used as an example of "static shape activity" to test our algorithm. Since the number of passengers vary over time, the polygon formed by joining their locations (in the same order always) is resampled to obtain a fixed number of landmarks. We have tested the performance of the algorithm on simulated 'spatial' and 'temporal' abnormalities [1], since we do not have real sequences with abnormal behavior. 'Spatial' abnormality (shown in figure 1(b)) is simulated by making one person deviate from his original path. This simulates the case of a person deciding to not walk towards the terminal. 'Temporal' abnormality is simulated by fixing the location of one person thus simulating a stopped person (which can be a suspicious activity too). When the person behind the stopped person goes ahead of him, the loop formed causes the shape to change. Note that since we are using the shape of discrete points, ordering matters and it is for this reason that a stopped person gets detected as an abnormal shape.

We first show results for the case of low (negligible) observation noise, using the log likelihood metric defined in section 3.2. Given a test sequence, at every time instant t we apply the log likelihood metric to the past L frames with $L = 20$ i.e. $d_{20}(t) = -logf^0(v_{t-19}, v_{t-18}, ...v_t)$. Reducing L will detect abnormality faster but will reduce reliability. In figure 2(a), the cyan dashed line plot is for the case of zero observation noise (hand-picked points). The blue circles ('o') plot shows the metric for a normal activity with $\sigma^2_{obs} = 4$ ($\sigma = 2$ pixel) Gaussian noise added to the hand-picked points, while the green stars ('*') plot is for a spatial abnormality (also with the same amount of observation noise) introduced at $t = 5$ for 40 frames. 2(b) shows the same plots for a temporal abnormality (plotted with red triangles). The spatial abnormality gets detected (visually) around $t = 20$ while the temporal one takes a little longer. Some of the lag in both cases is because of $L = 20$. In 2(c) we show the same plots but with $\sigma^2_{obs} = 81$. The metric now confuses normal and abnormal behavior, as discussed in section 4.

In figure 3, we show results for 9 pixel observation noise ($\sigma^2_{obs} = 81$) but with the observation noise now incorporated into the dynamic model (partially observed dynamic model as discussed in section 4). We show plots for the more difficult case of 'slow abnormality' where the tracking errors are small even for the abnormal activity. Hence the K-L metric (expected log likelihood) is needed to distinguish between normal and abnormal behavior. Figure 3(a) shows the plot for a spatial abnormality (green stars, '*') introduced at $t = 5$ which gets detected around $t = 7$ while as shown in 3(b), the temporal abnormality (red triangles) takes a little longer to get detected. The K-L metric plots for two instances of normal activity with the same amount of noise added are shown in both (a) and (b) with blue circles ('o') and magenta crosses ('x').

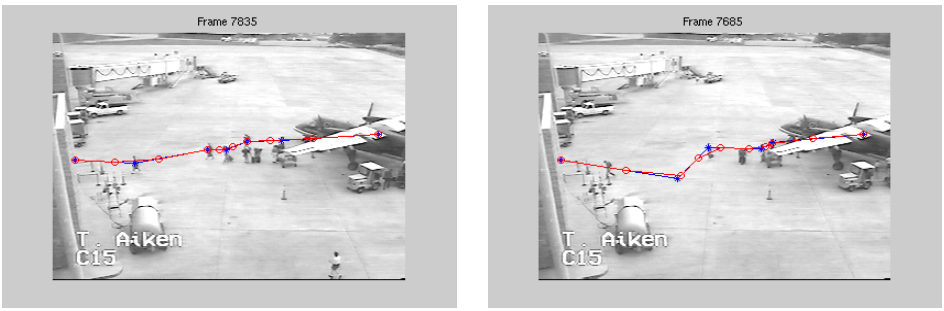
Figure 4 shows the Receiver Operating Characteristic (ROC) plots [20] for spatial abnormality (hypothesis H_1) versus normal activity (H_0). ROC plots the probability of abnormality detection (P_D) against the probability of a false alarm (P_F) [20]. The plots were generated by varying the normalcy threshold (η) and counting the number of times the abnormality gets detected in a normal (for P_F) and an abnormal sequence (for P_D) , for a given threshold. The three plots in the figure are for allowing different amounts of delay Δ_t for detection of the abnormality. As can be seen from the plots, if one were to allow only $\Delta_t = 5$, the maximum detection probability for $P_F \leq 0.2$ will be 0.85 while allowing a delay of $\Delta_t = 10$, increases this probability to 1. Actually for change detection problems, the ROC is a plot of the mean detection delay (assuming that the change will eventually get detected always) against the mean time between false alarms.

6 Conclusion

In this paper, we have looked at the problem of representing activities in low resolution video data where the moving objects are small enough to be modeled as point masses. Instead of representing the activity by the motion tracks of each individual object, we have proposed a compact global framework to model the activity using Kendall's shape theory. The activity is represented by the shape of by the configuration of the interacting objects, and its deforma-

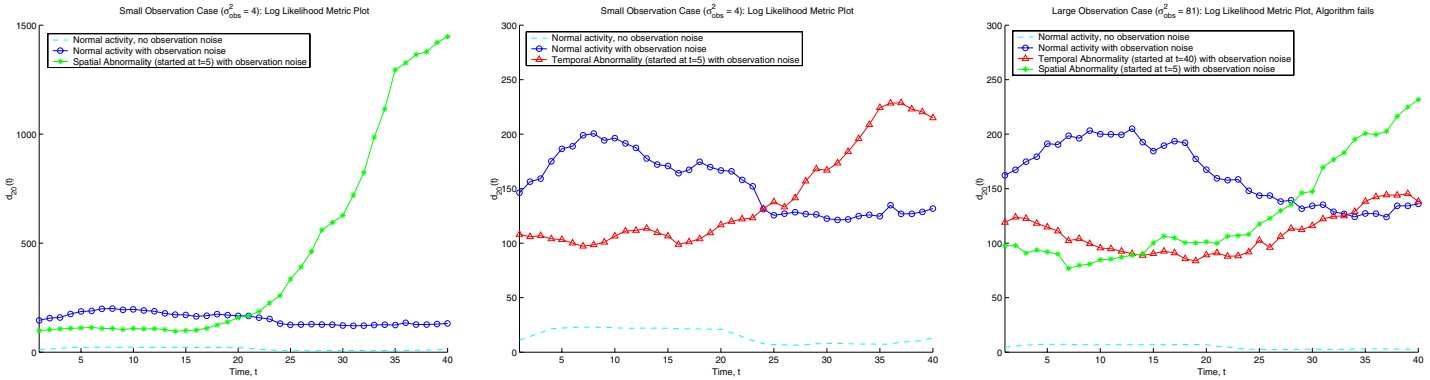
[5] even though it is not actually a 'metric'

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE



(a) (b)

Figure 1: (a): A 'normal activity' frame with shape contour superimposed, (b): Contour distorted by spatial abnormality. Note that the normal shape here appears to be almost a straight line, but that is just coincidence; our framework can deal with any kind of polygon formed by the point objects (landmarks). Also, the shape of landmarks does not distinguish open and closed polygons.



(a) (b) (c)

Figure 2: Plots of the log likelihood metric ($d_{20}(t)$) for normal and abnormal activities : (a) & (b) compare normal activity with spatial and temporal abnormality, respectively, for the case of small observation noise ($\sigma^2_{obs} = 4$). (c) shows the failure of the algorithm for large observation noise ($\sigma^2_{obs} = 81$). Note that the abnormality was introduced at $t = 5$.

tion over time. For test observation sequences with non-negligible observation noise, we have proposed to use a particle filter to estimate the posterior distribution of the shape given the observations. 'Drastic' abnormalities get detected because they cause the PF to lose track while for detecting 'slow' abnormalities for which the PF does not lose track, we have proposed to use the expected log likelihood of normalcy as the change detection statistic. Since our shape based algorithm models objects as point masses, the observations could as well be obtained using any kind of sensors - visible, radar, infrared or acoustic.

As part of future work, we intend to extend our framework to "dynamic shape activities" where the mean shape does not remain fixed. Also, we are working on using the prediction of current state based on past observations provided by the PF to improve the measurement model for tracking the observations. We intend to quantify the algorithm's robustness to model uncertainty and its sensitivity to rate of shape deformation over time. Finally, we hope to extend the algorithm to use observations from multiple and possibly moving sensors.

Acknowledgement

The authors would like to acknowledge Fumin Zhang of the ECE dept. at the University of Maryland, College Park for interesting discussions with him on the problem.

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.

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

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE

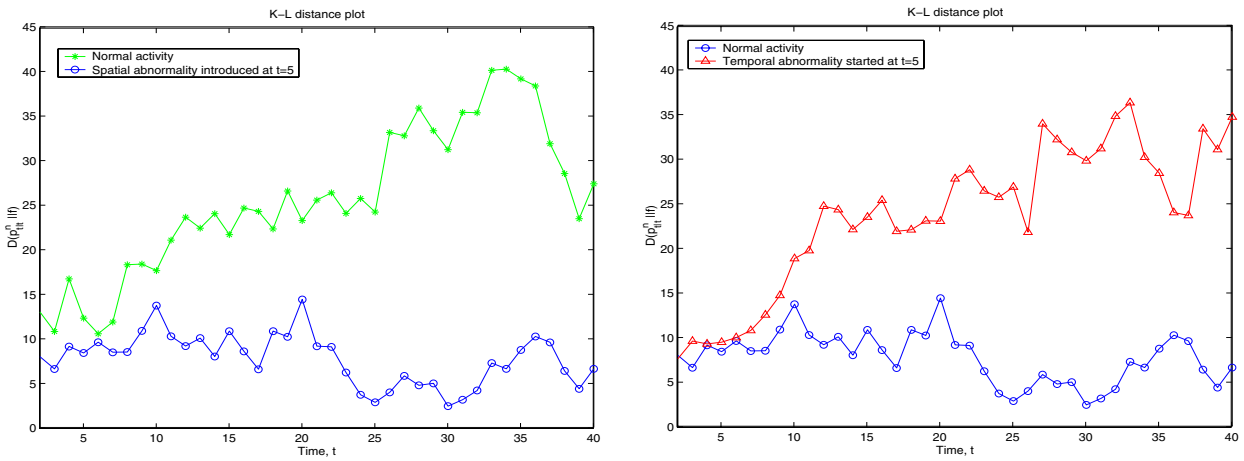


Figure 3: Plots of the K-L metric which works in large observation noise ($\sigma^2_{obs} = 81$) : (a) & (b) compare normal activity with spatial and temporal abnormality, respectively. Note that the abnormality was introduced at $t = 5$.

[4] I.L. Dryden and K.V. Mardia, *Statistical Shape Analysis*, John Wiley and Sons, 1998.

[5] C.T. Zahn and R.Z. Roskies, "Fourier descriptors for plane closed curves," *IEEE Transactions on Computers*, vol. C-21, pp. 269–281, 1972.

[6] B.K.P. Horn, "Extended gaussian images," *Proceedings of the IEEE*, vol. 72, pp. 1671–1686, 1984.

[7] I. Cohen, N. Ayache, and P. Sulger, "Tracking points on deformable objects using curvature information," in *European Conference on Computer Vision*, 1992, pp. 458–466.

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

[9] 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.

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

[11] N. Vaswani, A. RoyChowdhury, and R. Chellappa, "Statistical shape theory for activity modeling," in *International Conference on Acoustics, Speech and Signal Processing*, 2003.

[12] A. Doucet, N. deFreitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[13] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/nongaussian bayesian state estimation," *IEE Proceedings-F (Radar and Signal Processing)*, pp. 140(2):107–113, 1993.

[14] H. Moon, R. Chellappa, and A. Rosenfeld, "3d object tracking using shape-encoded particle propagation," *IEEE International Conference on Computer Vision*, 2001.

[15] J.P. MacCormick and A. Blake, "A probabilistic contour discriminant for object localisation," *IEEE International Conference on Computer Vision*, January 1998.

[16] D. Schulz, W. Burgard, D. Fox, and A. Cremers, "Tracking multiple moving targets with a mobile robot using particle filters and statistical data association," in *Proc. of the IEEE International Conference on Robotics and Automation*, 2001.

[17] B. Azimi-Sadjadi and P.S. Krishnaprasad, "Change detection for nonlinear systems: A particle filtering approach," in *American Control Conference*, 2002.

[18] J.T. Kent, "The complex bingham distribution and shape analysis," in *Journal of the Royal Statistical Society, Series B*, 1994, pp. 56:285–299.

[19] S. Soatto and A.J. Yezzi, "Deformotion: Deforming motion, shape average and the joint registration and segmentation of images," in *European Conference on Computer Vision*, 2002, p. III: 32 ff.

[20] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., 1991.

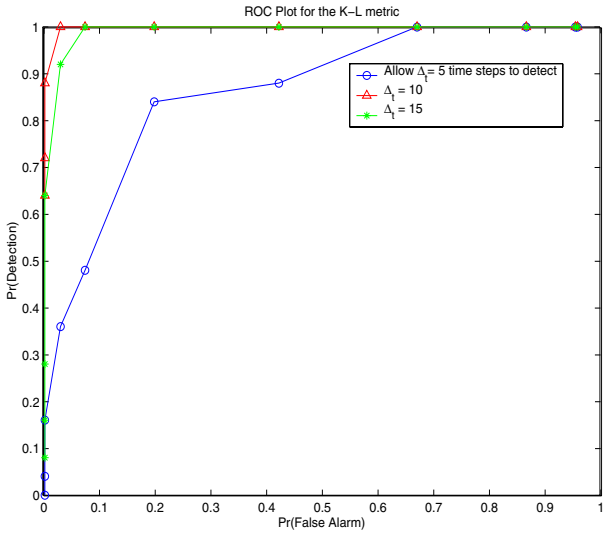


Figure 4: ROC plot using the K-L metric

Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
1063-6919/03 $17.00 © 2003 IEEE