Activity Recognition Using the Dynamics of the Configuration of Interacting Objects

Namrata Vaswani, Amit Roy Chowdhury, Rama Chellappa

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' formed by joining the locations 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 metric 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 a probability density function (pdf) for the actual shape given the noisy observations upto the current time. To detect abnormality, we propose to compare the distance of this estimated pdf from the pdf learnt earlier for a normal activity, using Kullback-Leibler distance. The approach can be directly applied for object location data obtained using sensors such as 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 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 to analyze the event. Instead of tracking point objects separately and then learning their interactions, we propose a different approach to the problem using Kendall's statistical shape theory [3]. We model an activity by the polygonal 'shape' formed by joining the locations of these point objects (henceforth referred to as 'points') at any time t and its deformation over time. This provides a compact global framework to model the motion of interacting moving objects over time. We are able to identify "spatial" abnormalities, e.g. deviations from the normal path, as well as "temporal" abnormalities, e.g. sudden stopping for prolonged periods of time when the normal activity should be continuous motion.

Shape is defined as all the geometric information that remains when location, scale and rotational effects are filtered out [4]. One of the earliest works in shape theory was that of Zahn and Roskies who used Fourier descriptors to model shape [5]. Another method for shape matching is the extended Gaussian image (EGI) model [6] in which the surface normal vector information for any object is mapped onto a unit sphere. Also, there exists a huge body of work in the vision community on shape tracking, analysis and similarity [7, 8, 9, 10, 11]. Statistical shape theory [3] which began in the late 1970s 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 [12], object recognition and image morphing (Chapter 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 to modeling the shape formed by the locations of a group of moving objects over time.

Consider as an example, 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. We learn the mean 'shape' (after removing location, scale and rotation) of the polygon formed by the locations of the passengers in any frame using hand picked location data (no observation noise). The dynamics of shape change is learnt by projecting the shape at any time t into the hyper-plane tangent to the (spherical) shape space at the mean and defining a Gauss Markov model in this tangent space (explained in sections 2.1 and and 3.1). The projection to tangent space from *fig-*

ure space [1] is nonlinear (equations (2), (5)) and hence we now have a nonlinear dynamical system in figure space. We first define a log likelihood metric in tangent space to detect abnormality for the simple case of a test sequence with negligible observation noise (fully observed case [13]). Next we consider the more practical (and difficult) case of large observation noise in the observed point locations. We now have a partially observed nonlinear dynamical system [13] from which we need to estimate the actual shape and also detect abnormality. Due to the nonlinearity and nonGaussianity (explained in section 4) of the system, we cannot use a Kalman filter for this problem. But this problem fits into the framework of particle filtering. Particle filtering [13] also known as the sequential Monte Carlo method [14] or condensation algorithm [15] was first introduced in [16] as an approach to non-linear, non-Gaussian Bayesian state estimation. Particle filters have been used in computer vision for shape based tracking of *a single* object using various representations of shape [17, 18]. In this work, we attempt to use a particle filter to track the 'shape' formed by the locations of a *group* of moving objects.

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 and how to estimate its parameters using hand-picked data and also how to detect abnormalility in the fully observed case. In section 4, we describe a particle filtering approach to track the actual shape from noisy observations (partially observed case) and detect abnormality. We use the dynamic model learnt from the hand-picked (or accurately observed) data as the system model for the particle filter. Experimental results are presented in section 5 and conclusions in section 6.

2 Preliminaries

2.1 Statistical Shape Theory

In this section 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 projection from the original figure space to Kendall's shape space and then to the hyperplane tangent to the (spherical) shape space 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.

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

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

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

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, $\hat{\theta} = arg(z^*_{Y_1} z_{Y_2})$.

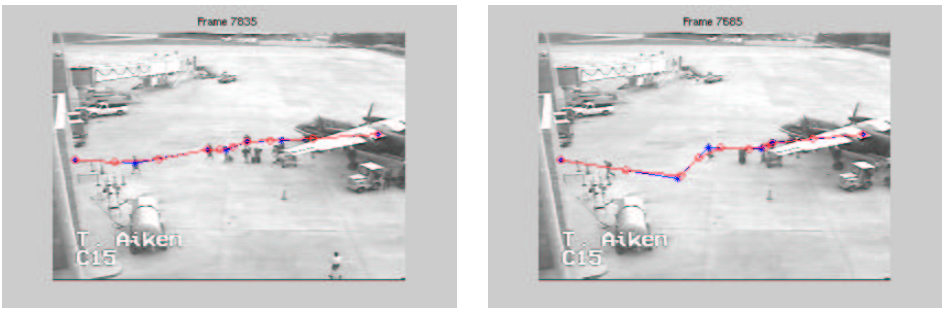
For a population of similar shapes, a full Procrustes mean shape ($\hat{\mu}$) is obtained by minimizing (over μ) the sum of squares of full Procrustes distances from each observation Y_i in the population to the unknown mean shape, μ, i.e.

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

For 2D shapes, the full Procrustes mean $\hat{\mu}$ can be found as the eigenvector corresponding to the largest eigenvalue of the matrix $S = \sum_{i=1}^{n} z_{Y_i} z^*_{Y_i}$ [19]. 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,

[1] Space defined by the original unnormalized point configuration



(a) (b)

Figure 1: (a): A 'normal activity' frame with shape contour superimposed, (b): Contour distorted by spatial abnormality

we define a linearized space (tangent space) about the mean shape and consider variance in the linearized space. The preshape formed by k points lies on a $k-1$ dimensional (because of translationnormalization) complex hypersphere of unit radius (due to scale normalization). 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 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]

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

where z is the preshape. Note that, the tangent coordinates are perpendicular to the Procrustes mean shape (by construction) and hence lie in a $k-2$ dimensional hyperplane.

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

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

The unnormalized shape is given by scaling the preshape by its scale (s), $Y = sz$.

2.2 Particle filtering

Let the state process $X = \{X_t\}$ be an $\mathcal{R}^{n_x}$-valued Markov process with a Feller transition kernel [13] $K_t(x_t, dx_{t+1})$ (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 $g_t(y_t|x_t)$. A particle filter [13] is a recursive algorithm that approximates by Monte Carlo sampling, the optimal posterior distribution of the state at any time t given the past observations. 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 [13] starts with sampling n times from the initial state distribution $\pi_0(x)$ to approximate it by $\pi_0^n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{x_0^{(i)}}(x)$. Now assuming that the distribution of X_{t-1} given observations upto time $t-1$ has been approximated as $\pi_{t-1|t-1}^n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{x_{t-1}^{(i)}}(x)$, 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, $\pi_{t|t-1}^n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{x}_t^{(i)}}(x)$ is the pdf 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. $w_t^{(i)} = \frac{n g_t(y_t|\bar{x}_t^{(i)})}{\sum_1^n g_t(y_t|\bar{x}_t^{(i)})}$. The measure $\bar{\pi}_{t|t}^n(x) = \frac{1}{n}\sum_{i=1}^n w_t^{(i)}\delta_{\bar{x}_t^{(i)}}(x)$ is the pdf of the state given observations uptil time t. We resample n times with replacement from $\bar{\pi}_{t|t}^n(x)$ to obtain $\pi_{t|t}^n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{x_t^{(i)}}(x)$.

3 Fully Observed 'Shape' Activity Model

We attempt to use Dryden and Mardia's statistical shape theory ideas (described above to represent the shape of "an" object) to model the 'shape' formed by the locations of a group of moving objects and its deformations over time. In describing the motion of a deforming shape, one needs to separate the effect of the global motion of the shape from its deformations. Extending Soatto and Yezzi's idea of static and dynamic deformable shapes [20], we define a "static shape activity" as one in which the shape formed by the moving points remains almost constant with time (except for small deformations). In this case, there is not

much information in the global motion parameters (translation, scale and rotation of the shape) and the activity can be represented by the mean "shape" and its allowed range of deformations. The shape deformation process in this case is stationary and ergodic. A "dynamic shape activity" on the other hand is represented by the time varying pattern of deformation and/or global motion (non-stationary process).

As an example of a "static shape activity", in our experiments we have looked at the "activity" of passengers getting out of a plane and walking towards the terminal. Since Kendall's shape analysis methods (discussed above) are for a fixed number of points, we resample the curve connecting the passenger locations at time t to represent it by a fixed number of points, k. 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 observed point 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 can be aligned to μ and projected into the tangent space (hyperplane) at μ using equation (5). The vector of tangent coordinates that we get is a complex k-dimensional vector. We string the real and imaginary parts of this vector to obtain a $2k$-dimensional real vector. But as explained earlier, the tangent coordinates actually lie in a $k-2$ dimensional complex space (which is equivalent to $2k-4$-dim real space).

3.1 Dynamical Model 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 one step *Gauss Markov model* , i.e.

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

where n_t [2] is a zero mean i.i.d. Gaussian process and is independent of v_{t-1}.

Since the activity is assumed to be stationary and ergodic, we can evaluate the covariance matrix of v_t, Σ_v for any time t as [21]

$$\Sigma_v = E[v_t v_t^T] = \frac{1}{T}\sum_{t=1}^{T} v_t v_t^T. \tag{8}$$

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

Also a minimum mean square error (MMSE) estimate of A (using stationarity assumption) can be evaluated as [21]

$$\begin{aligned} \hat{A} &= \Sigma_{v,1} * \Sigma_v^{-1} \quad \text{where} \\ \Sigma_{v,1} &= E[v_t v_{t-1}^T] = \frac{1}{T-1}\sum_{t=2}^{T} v_t v_{t-1}^T. \end{aligned} \tag{9}$$

Using $A = \hat{A}$ and ergodicity assumption, the noise covariance matrix can be calculated

$$\begin{aligned} \Sigma_n &= E[(v_t - Av_{t-1})(v_t - Av_{t-1})^T] \\ &= \frac{1}{T-1}\sum_{t=2}^{T}(v_t - Av_{t-1})(v_t - Av_{t-1})^T. \end{aligned} \tag{10}$$

Given a training sequence, we can use the above equations to estimate Σ_v, A and Σ_n. Based on the stationary Gauss Markov model described above we have,

$$\begin{aligned} v_t &\sim \mathcal{N}(0, \Sigma_v), \ \forall t \\ v_{t+1}|v_t &\sim \mathcal{N}(Av_t, \Sigma_n). \end{aligned} \tag{11}$$

Thus any L length sequence, $\{v_{t-L+1}, ...v_{t-1}, v_t\}$, would have a jointly normal distribution with pdf

$$\begin{aligned} f(v_{t-L+1}, ...v_t) &\overset{(a)}{=} f(v_{t-L+1}) f(v_{t-L+2}|v_{t-L+1}) ... f(v_t|v_{t-1}) \\ &\overset{(b)}{=} \frac{1}{\sqrt{(2\pi)^{2k-4}|\Sigma_v|}} \left(\frac{1}{\sqrt{(2\pi)^{2k-4}|\Sigma_n|}}\right)^{L-1} \times \\ &\exp\left(-\frac{v_{t-L+1}^T \Sigma_v^{-1} v_{t-L+1} + \sum_{\tau=t-L+2}^{t} (v_\tau - Av_{\tau-1})^T \Sigma_n^{-1} (v_\tau - Av_{\tau-1})}{2}\right) \end{aligned} \tag{12}$$

where (a) follows from the Markovian assumption and (b) follows from equations (11).

3.2 Abnormality Detection

We have assumed in this section that the noise in the observations when projected into tangent space is negligible compared to the system noise, n_t. For a test sequence, in this case, we can evaluate the tangent space projections (v_t) directly from the observations (Y_t) using equations (2) followed by (5).

We use the following hypothesis to test for abnormality. A given test sequence is said to be generated from the *normal activity iff* the probability of occurrence of its tangent projections (in the tangent plane generated by the normal activity mean μ) using the pdf given by (12) is large (greater than a certain threshold). Thus the distance to activity metric for an L frame sequence ending at time t, $d_L(t)$, is the log likelihood (without the constant terms) of the tangent coordinates of the observation i.e.

$$\begin{aligned} d_L(t) &= v_{t-L+1}^T \Sigma_v^{-1} v_{t-L+1} \\ &+ \sum_{\tau=t-L+2}^{t} (v_\tau - Av_{\tau-1})^T \Sigma_n^{-1} (v_\tau - Av_{\tau-1}) \end{aligned} \tag{13}$$

We test for abnormality at any time t by evaluating d_L for the past L frames. In the rest of the paper, we refer to this as the 'log likelihood metric'.

Now, if one looks at the eigenvalues of Σ_v, there are 5-6 dimensions of "almost" zero variance (eigenvalues much smaller than the rest). One could choose these directions to represent the Approximate Null Space (ANS) of the data. If data from the same activity is projected in these dimensions it will be very close to the origin with very high probability (follows from Chebyshev's inequality [21]) while as discussed in [22], this will not happen in general for data from any activity outside the 'normal activity class'. We use this idea to analyze tangent space data projected along the ANS using the same log likelihood metric as defined above but applied only to the 6-dimensional ANS space data. The difference between the values of the log likelihood metric for normal and abnormal activity is now more pronounced and also computed at a reduced computational cost.

4 Partially Observed 'Shape' Activity Model

In the previous section, we defined an abnormality detection metric for the case of negligible observation noise (accurately observed point locations data). But, when noise in the observations (projected in tangent 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 shape (Z_t) from the noisy observations ($Y_t = Z_t + w_t$) and also *detecting* abnormality. Since Z_t is now unknown, so is v_t and we thus have a partially observed non-linear dynamical system [13] with the following system (state transition) and observation model.

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 shape Z_t (equation 5), θ_t is the global rotation angle, $\theta_t = arg(Z_t^* \mu)$, and $\mathbf{s}_t$ is the global scale, $s_t = ||Z_t||$. The transition model for 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 mean, i.e.

$$\begin{aligned} v_t &\sim \mathcal{N}(Av_{t-1}, \Sigma_n) \\ s_t &\sim \textit{Rayleigh}\,(s_{t-1}) \\ \theta_t &\sim \textit{Unif}\,(\theta_{t-1} - a, \theta_{t-1} + a) \end{aligned} \tag{14}$$

with initial state distribution

$$\begin{aligned} v_0 &\sim \mathcal{N}(0, \Sigma_v) \\ s_0 &\sim \textit{Rayleigh}\,(\bar{s}_0) \\ \theta_0 &\sim \textit{Unif}\,(\bar{\theta}_0 - a_0, \bar{\theta}_0 + a_0) \end{aligned} \tag{15}$$

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 $a = \max_t |\theta_t - \theta_{t-1}|$. Due to lack of training data, the training sequence values of s_0 and θ_0 are taken as estimates of $\bar{s}_0$ and $\bar{\theta}_0$.

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

We assume that independent Gaussian noise with variance σ^2_{obs} is added to the actual location of the points, i.e. [4]

$$\begin{aligned} Y_t &\sim \mathcal{N}(Z_t, \sigma^2_{obs} I_{2k}) \quad \textit{where} \\ Z_t &= h(X_t) = s_t[(1 - v_{tc}^* v_{tc})^{1/2} \mu + v_{tc}] e^{-j\theta_t} \end{aligned} \tag{16}$$

where $h(X_t)$ is the function given by equation (6) followed by scaling by s_t. Also in general, both σ^2_{obs} and μ can be time varying.

To take care of outliers, we 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 (14) and the observation likelihood given by (16) in the particle filtering framework described in section 2.2. The particle filter provides at each time t, an n point δ function estimate of the distribution of the state variables at time t given observations upto time $t-1$, $\pi^n_{t|t-1}(v_t, s_t, \theta_t | \{Y_{0:t-1}\})$ (prediction) and the distribution of the state given observations upto time t, $\pi^n_{t|t}(v_t, s_t, \theta_t | \{Y_{0:t}\})$ (update).

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*

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

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

(a) The pdf of the tangent coordinates (v_t) given the observations upto time t, $f(v_t|Y_{0:t})$ is close to the normal activity pdf, $f(v_t)$ (given by equation (12) for $L = 1$). We measure closeness using the Kullback-Leibler distance [23] with $f(v_t|Y_{0:t})$ being the true pdf. Thus we have for a normal activity (η being a normality threshold),

$$D(f(v_t|Y_{0:t})||f(v_t)) < \eta. \tag{17}$$

and
(b) it is "correctly tracked" (i.e. the estimated pdf of the state given the observations, $\pi^n_{t|t}(v_t)$ is a good approximation to $f(v_t|Y_{0:t})$) by the particle filter trained on the dynamic model learnt for a normal activity.
Note that using a particle filter, we can approximate $D(f(v_t|Y_{0:t})||f(v_t))$ by $D(\pi^n_{t|t}(v_t)||f(v_t))$ only if the observations are "correctly" tracked. We work under the assumption that a normal activity sequence is (almost) always "correctly tracked".

Now, there can be two kinds of abnormalities. If the abnormality is a very 'drastic' one, it will not be "correctly tracked" by the particle filter (which is trained on a normal activity) and thus violate (b). In this case we ignore the value of $D(\pi^n_{t|t}(v_t)||f(v_t)$ since it is not a correct estimate of the actual K-L distance. On the other hand, for a 'slow' abnormality (say a person slowly walking in a wrong direction), the particle filter will be able to "correctly track" the observations i.e. the estimate $\pi^n_{t|t}(v_t)$ and hence $D(\pi^n_{t|t}(v_t)||f(v_t)$ will be a valid approximation. In this case, the abnormality gets detected by $D(\pi^n_{t|t}(v_t)||f(v_t) > \eta$.
Now, the expression for $D(\pi^n_{t|t}(v_t)||f(v_t))$ is

$$\begin{aligned} D(\pi^n_{t|t}(v_t)||f(v_t)) &= \int_{\mathcal{R}^{2k-4}} \pi^n_{t|t}(v_t) \log \frac{\pi^n_{t|t}(v_t)}{f(v_t)} dv_t \\ &= C + \frac{1}{n}\sum_{i=1}^{n} {v_t^{(i)}}^T \Sigma_v^{-1} v_t^{(i)} \\ &= C + E_{\pi^n_{t|t}} v_t^T \Sigma_v^{-1} v_t \end{aligned} \tag{18}$$

($C \triangleq -\frac{1}{n}\sum_{i=1}^{n} \log n + \frac{1}{n}\sum_{i=1}^{n} \log\sqrt{(2\pi)^{2k-4}|\Sigma_v|}$), which is nothing but a constant plus the expectation (under the estimated pdf of $v_t|Y_{0:t}$) of the log-likelihood metric in the fully observed case (equation (13) with $L = 1$). We refer to this as the 'K-L metric' in the results section. [5]

The other question is, how do we determine if the observations are "correctly tracked"? If the expectation of mean square error between the observation (Y_t) and the estimated shape, under the pdf $\pi^n_{t|t}(v_t, s_t, \theta_t)$, is very large when compared to the observation noise variance, the observations would be incorrectly tracked (the estimated pdf $\pi^n_{t|t}$ does not approximate the true one). Also, if the observation has occurred, its probability of occurrence should be greater than zero. If the estimated probability of the observation $\hat{P}(Y_t|Y_{0:t-1})$ is very small, then also the observation is incorrectly tracked.

Thus, to summarize, an activity sequence is *abnormal* if either it is incorrectly tracked *or* the K-L metric defined above exceeds a threshold. It is *normal* if it is "correctly tracked" *and* the K-L metric is less than the threshold.

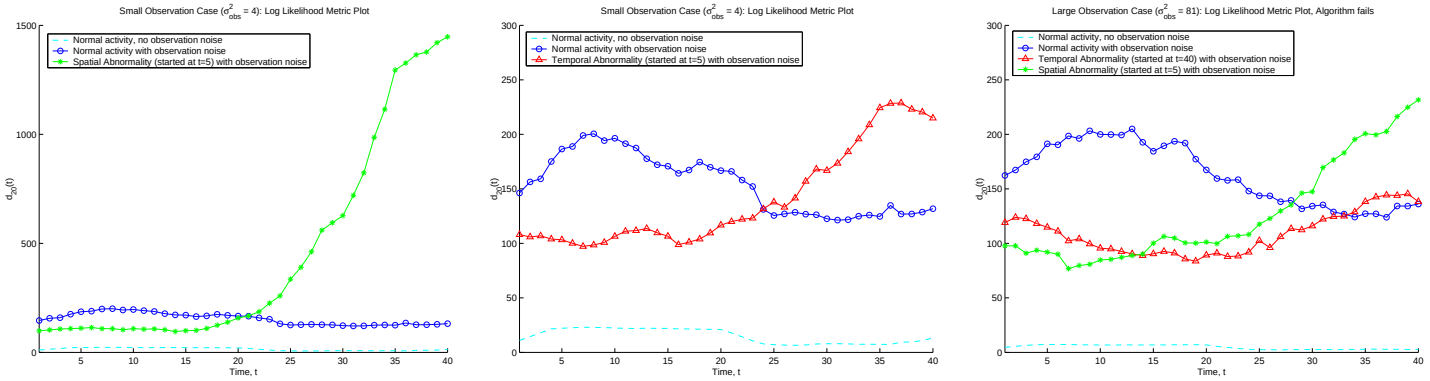
5 Experimental Results

We use a video sequence of passengers getting out of a plane and walking towards the terminal as an example of a "static shape activity" to test our algorithm. We test the performance of the algorithm on simulated spatial and temporal abnormalities, since we do not have real sequences with abnormal behavior. Spatial abnormality (shown in figure 1(b)) is simulated by making one particle deviate from its original path and then move back. This simulates the case of a person deciding to not walk towards the terminal. Temporal abnormality is simulated by fixing the location of a particle thus simulating a stopped person (which can be a suspicious activity too).

We first show results for the case of low (negligible) observation noise, using the log likelihood metric defined in 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. $d_{20}(t) = -log f(v_{t-19}, v_{t-18}, ...v_t)$. 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) 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 the 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, especially for the temporal abnormality.

In figure 3, we show results for 9 pixel observation noise ($\sigma^2_{obs} = 81$) but with the noise now incorporated into the dynamic model (particle filtering). We show plots for the more difficult case of 'slow abnormality' where the tracking errors are small ('correctly estimated' pdf) even for the abnormal activity. Hence the K-L distance metric is needed to distinguish between normal and abnormal behavior. (a)

[5] Note that both 'log-likelihood metric' and 'K-L metric' is an abuse of the word 'metric' since neither satisfy the properties of a metric. In fact since we do not know anything about the statistics of an abnormality (except that it is "not normal"), we cannot define a 'metric' in the rigorous sense.



(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 $t = 5$.

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 ($t = 10$) to get detected visually. 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 and magenta crosses ('x').

Figure 4 shows the Receiver Operating Characteristic (ROC) plots [21] for the case of high observation noise (using the K-L metric). ROC plots the probability of abnormality detection (P_D) against the probability of a false alarm (P_F) for the binary hypothesis testing problem described in this paper. The plots were generated by simulations, by varying the normality threshold (η) and counting the number of times the abnormality gets correctly (for P_D) or wrongly (for P_F), for a given threshold. The three plots in the figure are for allowing different amounts of delay Δ_t for detection of 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. In fact for $\Delta_t = 10$, we can achieve a detection probability P_D of 0.95 even with only allowing $P_F \leq 0.05$. If the allowed delay is increased to $\Delta_t = 15$, the curve approaches the ideal value of $P_D = 1$ for $P_F = 0$. However, in most surveillance applications it may not be possible to allow very large delays.

6 Conclusion

In this paper, we have looked at the problem of representing activity in low resolution video data where 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 formed by the locations of the interacting objects, and its deformation over time. We have learnt the dynamical model of shape change from noiseless (hand-picked) observation data and defined an abnormality metric for the simple case of test data with negiligible observation noise. For the more practical and difficult case of a test sequence with large observation noise, we have proposed to use a particle filter to estimate the probability distribution of the shape and defined a Kullback Leibler metric for detecting abnormalities. Experimental results have been shown for different kinds of abnormalities. Since the 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 our future work, we intend to apply our algorithm to "dynamic shape activities". We also intend to quantify the algorithm's robustness to model uncertainty and its sensitivity to rate of shape deformation over time.

References

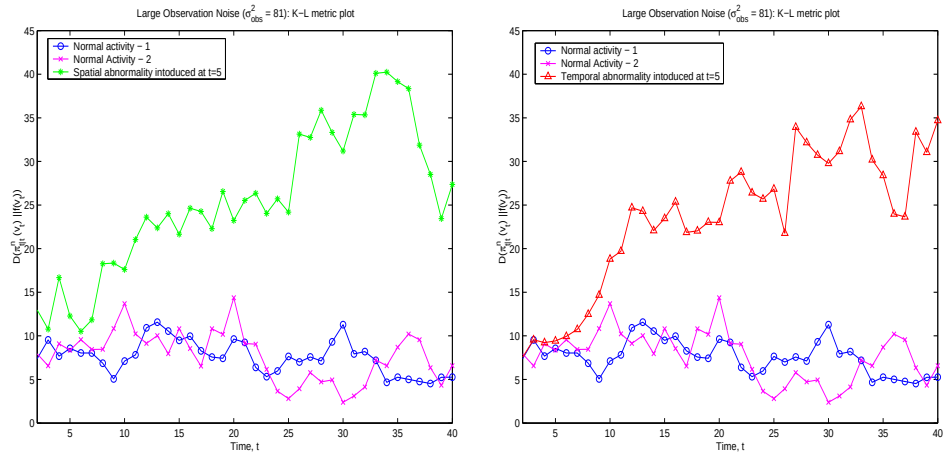
[1] W.E.L. Grimson, L. Lee, R. Romano, and C. Stauffer, "Using adaptive tracking to classify and monitor activities in a site," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998, pp. 22–31.

[2] S. Kurakake and R. Nevatia, "Description and tracking of moving articulated objects," in *International Conference on Pattern Recognition*, 1992, pp. I:491–495.

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

[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.



(a) (b)

Figure 3: Plots of the K-L metric ($D(\pi^n_{t|t}||f)$ 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 $t = 5$.

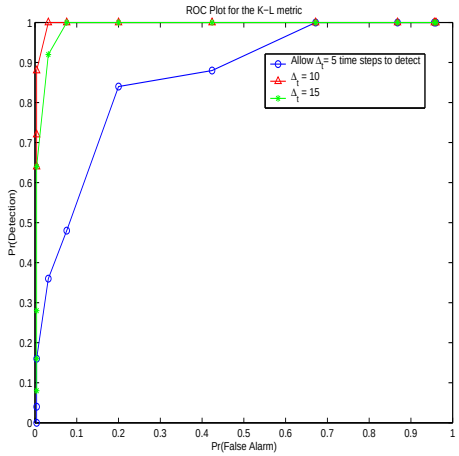


Figure 4: ROC plot using the K-L distance metric

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

[10] 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.

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

[12] I.L. Dryden, "Statistical shape analysis in high-level vision," in *IMA Workshop on Image Analysis and High Level Vision Modeling*, 2000.

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

[14] P. Fearnhead, "Sequential monte carlo methods in filter theory," in *PhD Thesis, Merton College, Universoty of OXford*, 1998.

[15] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," *European Conference on Computer Vision*, pp. 343–356, 1996.

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

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

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

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

[20] 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.

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

[22] N. Vaswani, "A linear classifier for gaussian class conditional distributions with unequal covariance matrices," in *International Conference on Pattern Recognition*, 2002.

[23] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley Series, 1991.