Epitomic Representation of Human Activities

Naresh P. Cuntoor and Rama Chellappa*
Center for Automation Research
University of Maryland College Park, MD 20742
{cuntoor, rama}@cfar.umd.edu

Abstract

We introduce an epitomic representation for modeling human activities in video sequences. A video sequence is divided into segments within which the dynamics of objects is assumed to be linear and modeled using linear dynamical systems. The tuple consisting of the estimated system matrix, statistics of the input signal and the initial state value is said to form an epitome. The system matrices are decomposed using the Iwasawa matrix decomposition to isolate the effect of rotation, scaling and projective action on the state vector. We demonstrate the usefulness of the proposed representation and decomposition for activity recognition using the TSA airport surveillance dataset and the UCF indoor human action dataset.

1. Introduction

Recent years have seen burgeoning literature in modeling activities ranging from simple, periodic activities such as walking and running ([1]) to more complex ones that involve an underlying semantic structure ([12], [23]). We are interested in modeling complex human activities performed both in indoor and outdoor scenarios such as office and home environment, surveillance and monitoring in airport and urban settings. A brief review of related work is presented next without attempting to provide an exhaustive survey.

Domain knowledge about activities can be readily incorporated using manually specified ontologies. The Video Event Representation Language (VERL) and Video Event Markup Language (VEML) were developed as a formalism to capture such ontologies [8]. Manual annotation and exhaustive construction of ontologies, however, can be tedious.

Several statistical approaches using hidden Markov models (HMMs) and dynamic Bayesian networks (DBNs) ([12], [16]) have been proposed. HMMs and DBNs have proven successful models for activities in which the graphical structure is known. If the structure is not known apriori or if the appearance varies drastically, it poses a formidable challenge to HMMs and DBNs. Hidden Markov model (HMMs) were used to model primitives of activities in [12]. Temporal sequencing between primitives was captured using stochastic context free grammar.

Features of motion in activities can be selected depending on activities of interest. Actions were modeled as a sequence of *dynamic instants* that are points of maximum curvature along the motion trajectory [20]. Similarly, in [22], changes in velocity curve profiles of actions were used to segment actions in videos streams. Reliance on features such as curvature can limit the domain of applicability. For example, in an airport tarmac surveillance scenario, trajectories formed by passengers walking from gate to aircraft can follow a straight-line path. Another limitation is that trajectories may contain only a few points of high curvature making it possible to encode only a few activities.

Neural networks were used to learn the distribution of motion trajectories, followed by vector quantization to cluster trajectories into a known number of classes [14]. This approach was developed further in [11] for robust tracking and anomaly detection using a fast fuzzy k-means algorithm. Trajectories were resampled to create data vectors of equal length for clustering. Subsequently, the time taken to complete an activity was modeled at the next level. In [21] factor graphs was employed to classifying surveillance-type data. These techniques accumulate motion trajectories in the scene to produce an intuitively appealing map of trajectories. In a highway surveillance scenario, for instance, it reveals lanes of traffic observed in the scene.

A formal approach for modeling trajectories as a shape (in Kendall's shape space) formed by moving landmarks was presented in [23]. Shape is defined as the geometrical information that remains after filtering out the effects of translation, rotation and scale. Procrustes distance between shapes was used to check for anomalous trajectories and dynamics were modeled in the shape's tangent space using

*This work was funded (in Part) by the U.S. Government VACE program.

1-4244-1180-7/07/$25.00 ©2007 IEEE

a first order Gauss-Markov model. This can model small changes about the mean activity shape. It may be necessary to preserve distinctions arising from factors such as rotation to distinguish between activities. For example, shape representation fails to distinguish between trajectories of persons embarking an aircraft from those of persons disembarking the aircraft. The proposed epitomic representation also attempts to extract geometrical aspects such as rotation, scaling and translation. Unlike [23], we retain this information when constructing a mid-level representation.

Factorization approach was used in [6] to model activities based on rank constraints. Its effectiveness was demonstrated for activities such as passengers embarking an aircraft. As the authors note, however, it cannot be used to recognize ground crew movement because of drastic variations across samples.

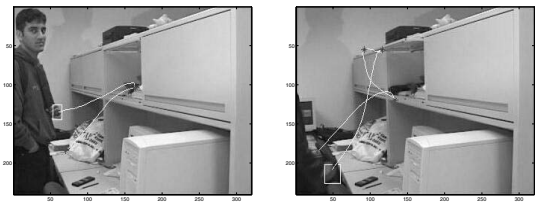
Image epitomes were introduced in [15], which modeled blocks of pixels using Gaussian distribution. This idea was extended to videos in [5], where cubes of pixels (i.e., cubic patches in xyt volume) were modeled using Gaussian distribution. The epitomic representation is attractive for its efficiency, extensibility and modularity.

In [24], activities were modeled as a sequence of linear dynamical systems. Switching instants between dynamical systems are identified when approximation errors from the trained database exceeds a threshold. Our method also models activities using linear dynamical systems. But the focus of application is different as outlined next.

We propose an epitomic model for activities, using kinematics of objects within short-time intervals. Given a video sequence, moving objects are detected and their motion trajectories are automatically extracted. Assuming that the dynamics is linear within the segment, it is modeled using linear systems $x_{k+1} = Fx_k + u_k$. Here, x_k is the state vector denoting position and velocity, F is a square matrix and u_k is the input signal. An epitome is defined as the tuple $(x_0; F; \mu, \Sigma)$, where x_0 is the initial state, F is the matrix and μ, Σ represent mean and covariance of the input signal $u_k, t \in [k_0, k_0 + T]$, for a segment of length T.

The system matrices are decomposed using the Iwasawa matrix decomposition [13] that yields three factors representing the effect of rotation, scaling and projective action on the state vector. The efficacy of the decomposition for key frame detection is demonstrated using the projective component. Also, it is used to geodesic distances between activities, which can be physically interpreted.

The rest of the paper is organized as follows. Low-level video processing is discussed in section 2. Section 3 motivates an epitomic representation for activities. Star diagrams are presented in section 4 as a way to visualize epitomes. In section 5, the Iwasawa decomposition is described. Section 6 describes the notion of distance for comparing epitomes. Section 7 illustrates the usefulness of the proposed method for activity recognition and key frame detection using the Transportation Security Administration (TSA) airport tarmac surveillance dataset and the University of Central Florida (UCF) indoor dataset of human actions. Section 8 concludes the paper.



(a) (b)

Figure 1. UCF Dataset: (a) Sample image from the UCF human action dataset. White line shows the hand trajectory; (b) Motion trajectory for *pick up object from desk*.

2. Low level video processing

In our experiments we have used existing algorithms for object detection and tracking with slight modifications. They are briefly summarized here. The background in each RGB color channel is modeled using single independent Gaussian distributions at every pixel [4]. The background model is reinitialized at regular intervals to handle changes in lighting. Motion trajectories are obtained using the KLT algorithm [17] whose feature points are initialized at detected locations of motion blobs by the background subtraction component. The trajectories are smoothed using a median filter. Low level processing in the UCF dataset follows a different procedure (section 7).

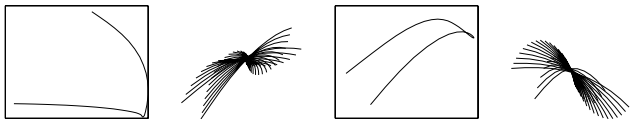
3. Epitome model for Activities

A video sequence is divided into segments of length T. Moving objects are detected and their short-time motion trajectories are obtained as described in the previous section. The kinematics of motion within segments is assumed to be linear so that linear systems can be used in modeling (1). The estimated tuples $(x_0; F; (\mu, \Sigma))$ are called activity epitomes, where x_0 represents the initial state, $F \in GL(n, \mathbb{R})$ ($GL(n, \mathbb{R})$ is a Lie group of all invertible matrices with real entries) is an invertible $n \times n$ system matrix and (μ, Σ) represents the statistics of input signal $u_k, k \in [0, T]$ that is assumed to be i.i.d. Gaussian distribution (Gaussian assumption is not essential and other distributions can be used).

Assuming full state output, kinematics of motion in each video segment can be written as:

$$x_{k+1} = Fx_k + u_k, \quad (1)$$

$$y_k = x_k, \quad (2)$$



(a) (b) (c) (d)

Figure 2. UCF Dataset: (a) Motion trajectory for *pick up object from desk* and (b) its star diagram; (c)Motion trajectory for *pick up an umbrella from cabinet* and (d) its star diagram.

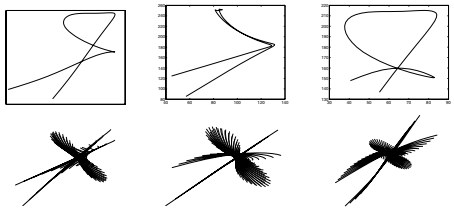


Figure 3. UCF Dataset: Top row shows trajectories of the hand opening the cabinet door. Bottom row shows the corresponding star diagrams.

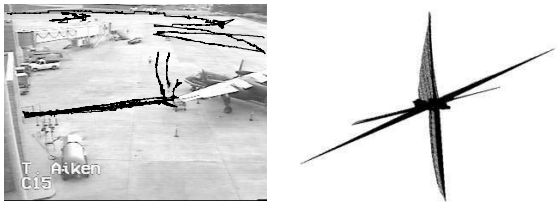
where $x_k \in \mathbb{R}^4$ is the state vector with initial value x_0, F is the system matrix of size 4×4 and u_k is the input signal. The state vector represents the 2-D position and velocity. The model in (1) is a state-space representation of the familiar Newton's laws in mechanics given by $m\ddot{q}(t)+k_1\dot{q}(t)+k_2q(t) = u(t)$, which describes the motion of bodies subjected to conservative forces.

We use the least squares method to estimate th elements of F and the statistics of u (mean, covariance). The $n \times n$ matrix is written as a vector in $\mathbb{R}^{n^2}$ and the x_k values are suitably re-arranged so that a desired structure can be imposed on F. For example, it is often useful to assume a block diagonal structure of F to represent a second-order dynamical model in which the x and y components of state evolution are decoupled.

The signal u_k can have a useful role in modeling activities, besides providing a way to deal with uncertainty. It can be used to synthesize parts of the activity for handling occlusions. Also it provides a way of visualizing the data.

4. Star Diagrams

Star diagrams provide a visual summary of activities and demonstrate similarities between subtrajectories though overall trajectories may have significant differences because of semantic ambiguity and context dependency. Semantic ambiguity refers to incomplete textual description of an activity whose motion trajectories can vary significantly across samples [12]. Computation of star diagrams using the estimated epitomes is described next. Short segments of trajectories are synthesized using the estimated initial state and input signal parameters (mean and variance). The mean of each segment is subtracted out. Subsequent segments are initialized using the endpoints of the previous segments. A temporally overlaid plot of the synthesized trajectory segments are called star diagrams.



(a) (b)

Figure 4. TSA airport surveillance dataset. (a) Motion trajectories of passengers embarking; (b) its star diagram

Two aspects of star diagrams merit emphasis. Since the mean of each segment is subtracted, spatial context is largely lost. This can be recovered using x_0 in the epitome. The ability to enunciate similarities at subtrajectory level, which accompanies the loss of context, is an asset for activity recognition and key frame detection. Second, epitomes can be reliably estimated using fewer number trajectories compared to a model that is based on the overall trajectory. This is useful when viewing direction changes, which in turn causes an apparent change in motion trajectories. The number of samples per viewing direction may be limited.

Figs. 1 shows a sample image from the UCF human action dataset along with the extracted trajectory of the hand performing a picking up action. Figs. 2(a) and (c) show two instances of picking up action to illustrate large changes in appearance. In fig. 2(a), an object lying in the desk is being picked up whereas in fig. 2(c), an object lying in the cabinet shelf is picked up. The star diagrams in the two cases (fig. 2(b) and (d)) differ by a rotation angle. This is captured using the Iwasawa matrix decomposition. Fig. 3 shows trajectories for opening the cabinet door and star diagrams synthesized using each of the trajectories. Though the appearance of trajectories in the three cases varies, similarities are preserved in the star diagram. In these illustrations, gradual rotations of the state vector and abrupt changes characterize the activities. Such changes motivate the decomposition of motion as discussed in the next section.

Fig. 4 shows star diagrams for activities on an airport tarmac. In fig. 4(a), activities include passengers embarking an airplane, movement of ground crew and a truck. The star diagram (fig. 4(b)) shows a compact representation using epitomes that were learnt using trajectories in fig. 4(a).

5. Iwasawa decomposition

Star diagrams provide a visual representation of epitomes. Motivated by the structure inherent in them, we propose a decomposition of epitomes into three components

using the Iwasawa matrix decomposition. The three components are rotation, scaling and translation of the state vector.

Definition 1 *Let $F \in GL(n, \mathbb{R})$. Then there exist unique matrices K, A, N, such that $F = KAN$, where (i) K is an orthogonal matrix, (ii) A is a diagonal matrix with positive diagonal entries, and (iii) N is a unit upper triangular matrix, i.e., all the diagonal elements are unity. This is called the Iwasawa matrix decomposition [13].*

It may be worth noting that the Iwasawa decomposition applies globally to the entire group manifold [10]. If $F \in SL(n, \mathbb{R})$ (i.e, det $F = +1$), the matrices take the following form for $n = 2$:

$$K = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \quad A = \begin{pmatrix} \sqrt{a_1} & 0 \\ 0 & \frac{1}{\sqrt{a_1}} \end{pmatrix}$$
$$\text{and } N = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix}, \tag{3}$$

where $\beta \in \mathbb{R}$, $a_1 \in \mathbb{R}^+$ and $\theta \in \mathbb{R}/\pi\mathcal{Z}$. Each of the three components are a one-parameter family that decouple the effect of transformation F on the state of the moving object. This is used to define distances between F's in the next section.

5.1. Special Case: $n = 2$

Consider $F \in SL(2, \mathbb{R})$ with $F = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix}$, where $f_{ij} \in \mathbb{R}$, $i, j \in \{1, 2\}$. By hypothesis, $f_{11}f_{22} - f_{12}f_{21} = 1$. From (3), the components of the decomposition can be calculated as follows:

$$\begin{aligned} a_1 &= f_{11}^2 + f_{21}^2 \\ \cos\theta &= \frac{f_{11}}{\sqrt{f_{11}^2 + f_{21}^2}} \\ \sin\theta &= \frac{-f_{21}}{\sqrt{f_{11}^2 + f_{21}^2}} \\ \beta &= \frac{1}{f_{11}}\left(\frac{f_{21} + f_{12}(f_{11}^2 + f_{21}^2)}{f_{11}^2 + f_{21}^2}\right) \\ &= \frac{1}{f_{11}}\left(\frac{f_{21}(f_{11}f_{22} - f_{12}f_{21}) + f_{12}(f_{11}^2 + f_{21}^2)}{f_{11}^2 + f_{21}^2}\right) \\ &= \frac{f_{11}f_{12} + f_{21}f_{22}}{f_{11}^2 + f_{21}^2} \end{aligned} \tag{4}$$

5.2. General Case

Let $F \in GL(n, \mathbb{R})$. Form

$$M = F^T F, \tag{5}$$

Table 1. Computing the Iwasawa matrix decomposition of an invertible matrix F

Let $M = F^T F$
Compute the Cholesky decomposition $M = R^T R$
Form the diagonal matrix $A = diag(R)$
Compute the unit upper triangular matrix $N = A^{-1}R$
Compute $K = FR^{-1}$

where M is symmetric, positive definite. (5) becomes

$$\begin{aligned} M &= (KAN)^T(KAN) && (6) \\ &= N^T A^T K^T K A N && (7) \\ &= N^T A^T A N, && (8) \end{aligned}$$

using $F = KAN$ and $K^T K = I$. Let

$$R \overset{def}{=} AN \tag{9}$$

Clearly, R is an upper triangular matrix. (8) becomes

$$M = R^T R \tag{10}$$

The factorization in (10) is computed using the Cholesky decomposition.

From definition 1, we know that N is a unit upper triangular matrix. The diagonal matrix A is formed by extracting the diagonal elements of R so that

$$N = A^{-1}R. \tag{11}$$

Since $F = KAN$ (definition 1),

$$K = FN^{-1}A^{-1} = FR^{-1}. \tag{12}$$

The steps are summarized in table 1.

5.3. Geometric interpretation and SVD

The Iwasawa decomposition yields $F = KAN$ as described in definition 1. The singular value decomposition (SVD) of F also yields three factors with $F = U\Sigma V^T$, where U and V are orthonormal matrices, and Σ is a diagonal matrix of singular values. In other words, SVD represents two rotations and scaling. It may be interesting to see the connection between the two decompositions.

SVD provides a co-ordinate basis whereas Iwasawa decomposition produces three factors which belong to subgroups generated by three linearly independent traceless matrices. The three factors belong to a maximal compact subgroup (K), a maximal Abelian subgroup (A) and a maximal nilpotent subgroup (N). The K component of $F = KAN$ decomposition captures rotation due to F. Both the diagonal matrices Σ and A in the two decompositions contain positive real entries and reflect scaling of the state vector. The N component in KAN decomposition, which is in reduced row-echelon form captures a projective action of F. The perceptual significance of N is described next.

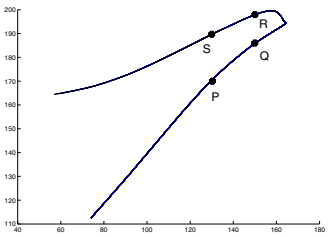


Figure 5. Illustrating the effectiveness of the N component of the Iwasawa decomposition using a sample trajectory from the UCF dataset.

5.4. Detecting key frames using the N component

From (3), we know that the N component has a unit upper triangular structure. In particular, for $F \in GL(2,\mathbb{R})$, the N component has one free element $\beta \in \mathbb{R}$. The sign changes of β reflects significant changes in direction of motion. The following example is used to show the usefulness of $sign(\beta)$.

Let $x_k = (x_k^{(x)}, x_{k-1}^{(x)}, x_{k-1}^{(y)}, x_k^{(y)})$ represent the state, where $(x_k^{(x)}, x_k^{(y)})$ denotes the position of the object at time k. Figure 5 shows a trajectory from the UCF dataset for picking up an object. The discrete time second-order decoupled system can be written as

$$x_{k+1} = \begin{pmatrix} f_1^{(x)} & f_2^{(x)} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & f_1^{(y)} & f_2^{(y)} \end{pmatrix} x_k + \begin{pmatrix} u_k^{(x)} \\ 0 \\ 0 \\ u_k^{(y)} \end{pmatrix} \quad (13)$$

For simplicity, consider the x component of motion, i.e., the top left 2×2 block of the system matrix in (13). The variation of the N component is discussed as the object traces the path $PQRS$ in three consecutive epitomes. The x and y co-ordinates of the points are denoted using superscripts, e.g., $P = (P^{(x)}, P^{(y)})$. The effect of N component is given by

$$x_{k+1}^{(x)} = \begin{pmatrix} 1 & \beta^{(x)} \\ 0 & 1 \end{pmatrix} x_k^{(x)} + \begin{pmatrix} u_k^{(x)} \\ 0 \end{pmatrix}, \quad (14)$$

where the superscript on the truncated state vector denotes motion of the x component. Using (14), the variation in $\beta^{(x)}$ as the object moves across the points is as follows

Case 1 *P to Q: Since $P^{(x)} < Q^{(x)}$, $\beta^{(x)} > 0$.*

Case 2 *Q to R: Since $P^{(x)} = Q^{(x)}$, $\beta^{(x)} = 0$.*

Case 3 *R to S: Since $P^{(x)} > Q^{(x)}$, $\beta^{(x)} < 0$.*

So the zero crossings of $\beta^{(x)}$ (and $\beta^{(y)}$) denote perceptually significant changes. The frames corresponding to the zero crossings are said to be key frames.

6. Distance between Epitomes

It is necessary to define a notion of distance between activity epitomes for many applications including activity recognition and clustering. Euclidean distance between $F_1 \in \mathbb{R}^{n\times n}$ and $F_2 \in \mathbb{R}^{n\times n}$ may be defined, in which F_1 and F_2 are thought of as vectors in $\mathbb{R}^{n^2}$. Alternatively, Frobenius distance between F_1, F_2 can be used, which is defined as follows:

$$\begin{aligned} d_f(F_1,F_2) &= ||F_1 - F_2||_F^2 \\ &= tr((F_1-F_2)(F_1-F_2)^T). \end{aligned} \quad (15)$$

This does not take the geometry of the space into account and it ignores the group structure of matrices. De Cock and De Moor described various ways of computing subspace angles between linear systems [7]. It is defined as the principle angle between the column spaces generated by the observability matrices of two models. Subspace angles have been used to measure similarity between dynamical models in [3] for recognizing humans based on gait.

Distance measures based on subspace angles (or principal angles) are widely accepted for comparing matrices that represent dynamics. A commonly used distance (e.g., in [3]) is Martin distance d_M^2 defined as [18]:

$$d_M^2 = -\log \prod_{i=1}^{n} \cos^2(\theta_i), \quad (16)$$

where $\theta_i, i = 1,\ldots,n$ are subspace angles. Subspace angles capture differences that are caused by rotation of subspaces while ignoring changes in other factors such as scaling and translation. Finsler distance [25] uses the largest principal angle between the subspaces unlike the Martin distance that uses all n values in (16). This is best suited when the signal is scalar-valued and generated by a strictly second-order stationary process. In our case, the number of outputs of the system is 4 (2D position and velocity). When the number of inputs and outputs associated with the system is more than one, the distance is not guaranteed to be non-negative. The natural metric in Finsler space is the Finsler-Minkowski metric.

Finsler geometry [2] is concerned with integrals of the form

$$\int_a^b f(q^1,\ldots,q^n;\frac{dq^1}{dt},\ldots,\frac{dq^n}{dt})dt.$$

The function $f(q^1,\ldots,q^n;\frac{dq^1}{dt},\ldots,\frac{dq^n}{dt})$ is positive unless all $\frac{dq^i}{dt}$ are zero. Here q stands for position and $\frac{dq}{dt}$ for velocity so that F denotes speed of the moving object. The integral measures the total distance traveled.

An infinitesimal Riemannian metric Δ (such as the Finsler-Minkowski metric) is the analog of the familiar dx quantity in Euclidean spaces. Just as the distance between

points in Euclidean space is obtained by integrating with respect to dx, the geodesic distance on a smooth manifold is obtained by integrating with respect to Δ.

Let $F \in G$. The geodesic distance $D : G \times G \rightarrow \mathbb{R}$ between matrices F_0 and F_1 can by calculated using the norm induced by the inner product Δ as follows:

$$D(F_0, F_1) = \min\{I_\Delta(F(\cdot)) : F \in C^1([0,1], G), F(0) = F_0, F(1) = F_1\}, \quad (17)$$

with $I_\Delta(F(\cdot)) = \int_0^1 \Delta(F, \dot{F})dt$. The Finsler-Minskowsi metric has the additional property that $\Delta(F, \dot{F}) = \Delta(F^{-1}\dot{F})$ [9].

More precisely, the minimum in (17) should be replaced by an infimum and existence of geodesics has to be established. It is known that geodesics exist as long as the manifold is complete ([2], [19]).

Evaluating the metric Δ and the integral in (17) is difficult except in certain special cases. In particular, the distances can be computed in the case of components obtained using the Iwasawa matrix decomposition as illustrated using examples below.

We illustrate the computation of geodesic distances for special cases (that are relevant to the KAN decomposition). $A = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}$. A simple calculation shows that the norm $\Delta(A)$ is

$$\Delta(A^{-1}\dot{A}) = \frac{|\dot{a_1}|}{a_1} + \frac{|\dot{a_2}|}{a_2} \quad (18)$$

The shortest path distance from the identity matrix I to A is $D(I, A) = |\log a_1| + |\log a_2|$.

Consider $N \in Aff(1)$ (also known as the "$ax + b$" group), i.e., $N = \begin{pmatrix} \rho & \beta \\ 0 & 1 \end{pmatrix}$. The Finsler-Minkowski metric becomes $\Delta(N^{-1}\dot{N}) = \frac{|\dot{\rho}|+|\dot{\beta}|}{\rho}$. For a unit upper triangular matrix, $\alpha = 1$ so that the shortest path distance minimizes $\int_0^1 \dot{\beta}d\beta$.

For $K = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$, the Martin distance in (16) is used.

7. Experiments

We demonstrate the usefulness of epitomic representation and the Iwasawa matrix decomposition using indoor and outdoor datasets.

7.1. UCF human actions dataset

The UCF dataset consists of 60 trajectories of common activities. We divide these into 7 classes: open door, pick up, put down, close door,erase board, pour water into cup and pick up object and put down elsewhere. The hand trajectories are obtained in a two step process in which the hand is first detected using a skin detector and then tracked using the mean-shift procedure. The resulting trajectories are smoothed out using anisotropic diffusion so that corners and sharp changes are retained. A detailed description of the steps is available in [20]. Figures 2(a) and (b) shows sample images from the dataset along with the trajectories. Typically, most of the activities in the dataset last for a few seconds i.e., of the order of 100 frames. We follow Rao *et al.*'s ([20]) division of the dataset into gallery and probe sets to ensure that the results remain comparable. The leave-one out method is used in the reporting the recognition scores.

Table 2. Results: Recognition rate (%) using UCF dataset.

Expt. set	UCF results [20]	Subspace angles	Our method
Open door	50	56	72
Pick up	-	50	61
Put down	-	55	72
Close door	53	60	73
Erase board	75	50	75
Pour water	33	67	100
Pick up and put down	56	44	90

7.1.1 Experiment 1: (Activity Recognition)

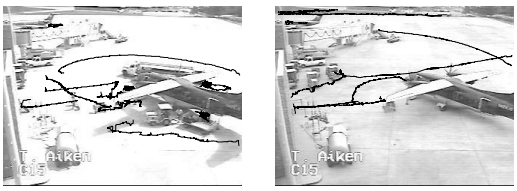
An extracted motion trajectory from the gallery set is divided into segments of 20 frames with overlap. The state vector is formed using the instantaneous position along the trajectory. An epitome $(x_0; F; \mu_u, \Sigma_u)$ is estimated for every segment as described in section 2. All 16 elements of $F \in GL(4, \mathbb{R})$ are estimated using least squares (section 2). The Iwasawa decomposition is used to find the K, A, N components of the estimated F matrices.

The distance from a test video sequence to those in the gallery is computed as follows. Epitomes of the test video sequence as before and its K, A, N components are computed. The distance between components are computed separately using the metrics as described in section 6. The recognition rates are summarized in Table 2.

Comparison of results: Recognition rates obtained using the proposed method are compared with those reported in [20] in which points of maximum curvature are used to compare activities. Also, we computed recognition scores using subspace angles between matrices [7], [3].

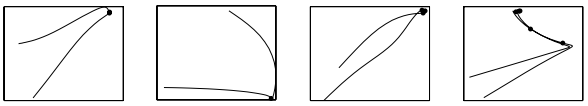
7.1.2 Experitment 2: (Key frame detection)

Activities such as picking up objects and opening a cabinet door have distinctive points along the trajectory that contain



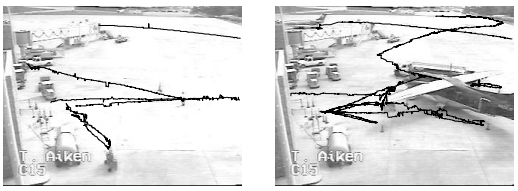
(a) (b) (c) (d)

Figure 6. Motion trajectories for two blocks of TSA data.



(a) (b) (c) (d)

Figure 7. UCF dataset: Dots represent key frames detected using zero crossings of β in the N component. Activities are (a)-(c) Pick up an object, (b) Open door.



(a) (b) (c) (d)

Figure 8. Motion trajectories for two blocks of TSA data.

Table 3. Recognizing passenger and ground crew trajectories in the TSA aiport surveillance dataset

	Detection rate (%)	False alarm (%)
Passengers	100	23
Ground crew	64	0

a sharp change in motion. These points denote perceptually siginificant time instants when the object is picked up or when the door is opened. The N component is used to detect these time instants (section 5.4). Sample results of key frame detection are given in figure 7.

7.2. TSA Airport Tarmac Surveillance Dataset

The TSA dataset consists (figures 6 and 8) of airport surveillance videos captured by a stationary camera that operates at approximately 30 frames per second. The image size is 320×240. It contains approximately $230,000$ frames or 120 minutes of video data. Activities include movement of ground crew personnel, vehicles, planes and passengers embarking and disembarking. Motion trajectories are extracted as described in section 2. Figure 6 shows sample motion trajectories in ten thousand frames. For each of the blocks in figures 6 and 8, dominant activities for each block are summarized below:

- Fig. 6(a): Luggage cart activity activity near aircraft. Ground crew movement near aircraft and to the gate.
- Fig. 6(b): Luggage cart and ground crew activity near the aircraft. The luggage cart exits. A truck crosses the scene.
- Fig. 8(a): Ground crew walk across scene and back to the gate. A truck crosses the scene.
- Fig. 8(b): Ground crew walk from gate to the aircraft and back. Another person walks across the scene and back to gate.

7.2.1 Experiment 3: (Activity Recognition)

Each motion trajectory is modeled separately so that at any given time there are as many epitomes as the number of objects in the scene. We test the proposed method for activity recognition. In figures 6 and 8, (b) shows the closest match for blocks of activity in (a). There are fifteen such blocks of data (of which only four are shown in figures 6 and 8 due to space constraints). Fourteen of these blocks were correctly matched. In the wrongly matched of video sequences, a truck driving away slowly was confused with a human (ground crew person) walking along a similar path. This confusion arises from modeling moving objects as point trajectories.

7.2.2 Experiment 4: (Clustering)

The contribution of individual activities to the overall recognition, however, is not clear in the above experiment. We focus on trajectories of humans to demonstrate the efficacy of the proposed decomposition and distance computation. There are two classes of humans in the dataset: passengers and ground crew personnel. It is necessary to distinguish between these two classes for applications such as anomaly detection. Unexpected motion patterns of passengers may be considered anomalous unlike the same motion pattern involving ground crew personnel. We use the K component of the epitome to distinguish between these two classes

since the motion of passengers is tightly clustered in this space. The K component represents rotation of the state as shown in (3). Recognition rates are summarized in table 3.

7.2.3 Experiment 5: Key frame detection

We test the efficacy of the N component for identifying the completion of luggage transfer between an aircraft and luggage cart. In the dataset there are eighteen instances of luggage cart in the scene. Of these four instances correspond to movement across the scene without meeting an aircraft. In the remaining fourteen instances (or seven roundtrips to the aircraft), the luggage cart goes to the aircraft to transfer luggage. The goal is to identify trajectories that correspond to these exchanges and their time instants. Motion blobs are detected and tracked as described in section 2. Dynamics is modeled using a decoupled second-order system (section 5.4) and sign changes are $\beta^{(x)}$ to detect luggage transfer times between luggage cart and aircraft.

Results: Of the 14 cases in which luggage carts were present, 12 were correctly identified. The average error in localizing the time at which the transfer occurs was 18 frames. The following objects caused three false alarms: (i) Movement of fuel truck and (ii) Two cases of truck moving to the terminal, stopping briefly before exiting the scene.

8. Summary

We presented an epitomic representation for modeling activities using piecewise linear dynamical segments. An epitome is said to be a tuple consisting of the estimated system matrix, initial state and input signal statistics. The Iwasawa matrix decomposition was used to factorize the system matrix into three components that represent the effect of rotation, scaling and projective action on the state vector. Its usefulness for activity recognition was demonstrated using both indoor and outdoor video datasets. As part of future work, we plan to explore the shape of star diagrams as a feature for activities.

References

[1] J.K. Aggarwal and Q. Cai. Human motion analysis:a review. *CVIU*, 73(3):428–440, 1999. 1

[2] D. Bao, S. S. Chern, and Z. Shen. *An Introduction to Riemann-Finsler Goemetry*. Springer, New York, 2000. 5, 6

[3] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. In *Proc. CVPR*, 2001. 5, 6

[4] A. Azabeyajani C. Wren, T. Darrell, and A. Pentland. Pfinder: Real time tracking of the human body. *IEEE Trans. PAMI*, 19:780–785, 1997. 2

[5] V. Cheung, B. Frey, and N. Jojic. Video epitomes. In *Proc. CVPR*, 2005. 2

[6] A. K. Roy Chowdhury and R. Chellappa. A factorization approach for activity recognition. In *Proc. IEEE Workshop on Event Mining*, volume 4, pages 41–46, Madison, WI, USA, 2003. 2

[7] K. De Cock and B. De Moor. Subspace angles and distances between arma models. In *Proceedings of MTNS*, 2000. 5, 6

[8] A. R. J. Francois, R. Nevatia, J. Hobbs, and R. C. Bolles. Verl: An ontology framework for representing and annotating video events. *IEEE Multimedia*, 12(4):76–86, 2005. 1

[9] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhauser, New York, 1999. 6

[10] S. Helgason. *Differential geometry, lie groups and symmetric spaces*. Academic Press, New York, 1978. 4

[11] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Trans. PAMI*, 24(9):1450–1464, 2006. 1

[12] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. PAMI*, 23:852–872, 2000. 1, 3

[13] K. Iwasawa. On some types of topological groups. *Annals of Mathematics*, 50:507–558, 1949. 2, 4

[14] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14(8):609–615, 1996. 1

[15] N. Jojic, B. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *Proc. ICCV*, 2003. 2

[16] D. Koller and U. Lerner. *Sequential Monte Carlo Methods in Practice*, chapter Sampling in Factored Dynamic Systems, pages 445–464. Springer, 2001. 1

[17] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *IJCAI*, pages 674–679, 1981. 2

[18] R. Martin. A metric for arma processes. *IEEE Trans. on Signal Processing*, 48(4):1164–1170, 2000. 5

[19] A. Mielke. Geometry, dynamics and mechanics. *Finite elastoplasticity, Lie groups and geodesics on SL(d), Springer-Verlag. Eds. Weinstein and Holmes*, 2002. 6

[20] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *IJCV*, 50(2):203–226, 2002. 1, 6

[21] C. Stauffer. Learning a factorized segmental representation of far-field tracking data. In *Proc. IEEE Workshop on Event Mining*, volume 7, pages 115–121, 2004. 1

[22] T. Syeda-Mahmood. Segmenting actions in velocity curve space. In *Proc. ICPR*, 2002. 1

[23] N. Vaswani, A. Roy Chowdhury, and R. Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. *IEEE Trans. IP*, 14(10):1603–1616, 2005. 1, 2

[24] D. Del Vecchio, R. M. Murray, and P. Perona. Decomposition of human motion into dynamics based primitives with application to drawing tasks. *Automatica*, 39(12):2085–2098, 2003. 2

[25] A. Weinstein. Almost invariant submanifolds for compact group actions. *J. Eur. Math. Soc.*, 2(1):53–86, 2000. 5