Action Recognition in Video by Covariance Matching of Silhouette Tunnels

Kai Guo, Prakash Ishwar, and Janusz Konrad
Department of Electrical and Computer Engineering, Boston University
8 Saint Mary's St., Boston, MA USA 02215 {kaiguo,pi,jkonrad}@bu.edu

***Abstract*—Action recognition is a challenging problem in video analytics due to event complexity, variations in imaging conditions, and intra- and inter-individual action-variability. Central to these challenges is the way one models actions in video, i.e., action representation. In this paper, an action is viewed as a temporal sequence of *local* shape-deformations of centroid-centered object silhouettes, i.e., the shape of the centroid-centered object *silhouette tunnel*. Each action is represented by the empirical covariance matrix of a set of 13-dimensional normalized geometric feature vectors that capture the shape of the silhouette tunnel. The similarity of two actions is measured in terms of a Riemannian metric between their covariance matrices. The silhouette tunnel of a test video is broken into short overlapping segments and each segment is classified using a dictionary of labeled action covariance matrices and the nearest neighbor rule. On a database of 90 short video sequences this attains a correct classification rate of 97%, which is very close to the state-of-the-art, at almost 5-fold reduced computational cost. Majority-vote fusion of segment decisions achieves 100% classification rate.**

***Keywords*-video analysis; action recognition; silhouette tunnel; covariance matching; generalized eigenvalues;**

I. Introduction

The proliferation of network cameras in the last few years has led to surveillance video overload; cameras produce data at rates far exceeding the capacity of human operators managing video surveillance networks. Thus, of interest are automatic or semi-automatic surveillance-video analysis methods. Of the many facets of video analysis, action recognition stands out as particularly important. For example, the ability to recognize that a person is running away from the scene of an accident or that a car is being driven in an erratic manner, can help alert law enforcement in real time or be useful in post-event video forensics. Action recognition also finds application in video retrieval, video indexing and detection of abnormal behavior.

Despite a significant effort by the computer vision and image processing communities, action recognition is still a challenging problem on the account of event complexity often present in the video (e.g., clutter, occlusions), variations in the imaging conditions (e.g., illumination, viewpoint, resolution) and only approximate repeatability of the same action by different individuals (e.g., no two individuals walk in exactly the same manner). Central to these challenges is the way one models actions in a video sequence, i.e., action representation. Some of the widely-used action representations are: static features based on limb shapes [1], [2], geometric models of objects [3], [4], motion/optical-flow patterns induced by moving objects [5], [6], and spatio-temporal features extracted from space-time video volume [7], [8], [9]. While some of these representations rely on pixel intensity, others are based on binary masks (often called silhouettes) or motion fields associated with moving objects. Experience to date has shown that action representation based on pixel intensities is not robust; differently dressed people performing the same action may be considered to act differently. While action recognition based on motion fields has been quite successful, it requires the additional, but not so simple, step of motion estimation. However, the dynamic nature of an action captured by a motion field is largely captured by object's silhouette evolving in time, i.e., a binary mask of moving object changing its shape in time, that we shall call a *silhouette tunnel*. Silhouette tunnels, also known as *object tunnels* [10], [11] or *activity tubes* [12], have been extensively studied in the literature with applications in video compression, summarization, frame-rate conversion, etc. Although silhouette tunnels do not capture motion inside objects, the moving silhouette boundary leaves a very distinct signature of occurring activity. Furthermore, a silhouette tunnel is void of color, texture and background characteristics, making it an appropriate representation for action in x-y-t space regardless of photometric properties of the moving object. To date several action recognition methods have been based on silhouettes [7], [13], [14], [15].

In particular, Gorelick *et al.* [7] developed a method that extracts shape properties of a silhouette tunnel by solving a Poisson equation (measurement of average length of a random walk from an interior point to silhouette tunnel boundary). An action classification based on this approach was shown to be remarkably accurate suggesting that the method is capable of extracting highly-discriminative information. However, the procedures used to extract spatio-temporal fea-

This material is based upon work supported by the US National Science Foundation (NSF) under awards CNS–0721884 and (CAREER) CCF–0546598. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

tures are fairly complicated and computationally-demanding.

Collins and Gross [16] have also used silhouettes to identify human actions, but their method extracts key frames from the query video sequence and matches them with key frames from training videos. The classification is performed by the nearest-neighbor rule based on normalized correlation scores. This method is conceptually simple and easy to implement, but it is based on 2-D silhouettes without considering the dynamics of silhouette evolution.

The dynamic nature of video has been also exploited by Bobick and Davis [13] who proposed a motion energy image (MEI), that represents where motion has occurred in an image sequence, and motion history image (MHI), that is a scalar field depicting how recently the motion occurred. Together, MEI and MHI act as a two-component version of a temporal template, and are compared with known actions in a database to determine the best action match.

Gait recognition [15], [14] is a specific class of action recognition problems, with a specific focus on humans. Gait recognition techniques can be divided into two broad categories: model-based and feature-based. Model-based approaches [17], [18] build a model with static and dynamic body parameters. Such approaches perform well at the expense of computational complexity. Feature-based techniques do not rely on the assumption of any specific body models. Different features can be extracted to represent the gait, such as the angular transform, radial integration transform (RIT) and circular integration transform (CIT).

In this paper, we propose a new framework for action recognition in video sequences. The proposed framework is general and applies to human actions as well as animals, man-made objects, etc. Like in some prior methods, we use silhouette tunnels to characterize actions, but we introduce a new metric to characterize and compare such tunnels. Our first contribution is the selection of features that accurately capture shape evolution in space-time. We propose 13 geometric attributes related to object shape and "life-span". Our second contribution is the application of covariance descriptor to these features in order to measure similarity between silhouette tunnels (actions). This choice has been inspired by recent covariance tracking methods [19], [20] that showed remarkable performance and reasonable computational complexity. Our third contribution is the development of an action recognition framework, including the majority-rule fusion that permits aggregation of action recognition results from short video segments. In comparison with the method of Gorelick *et al.* [7] our approach applied segment-by-segment shows almost identical performance but at a 5-fold reduced computational cost. However, the inclusion of our majority-based fusion improves the recognition rate to 100% without significant impact on computational complexity. Furthermore, our method is simpler conceptually and straightforward to implement as it does not require solving partial differential equations needed in the approach of Gorelick *et al.*

II. Overview of the problem and methodology

The goal of our work is automatic annotation of object "action" in fixed-perspective video footage using a dictionary of previously-annotated samples. In its full generality, this is a very hard problem because a scene may be composed of multiple objects which interact in intricate ways. We therefore focus on the key subproblem in which the video footage contains actions related to a single object. Such footage may potentially be obtained from preprocessing steps involving activity detection followed by tracking and isolation of object trajectories; tasks which may be difficult or impossible to perform in the vicinity of intersecting or overlapping object trajectories.

We need to first clarify what we mean by the terms object and action before we can describe our approach to detect and recognize them. The objects of focus in this work are humans but our approach can handle objects which are composed of one or multiple semi-rigid parts. By an action we mean a sequence of frequent, roughly repetitive changes in object shape and position[1] (see Fig. 1). Perceived changes in object position and shape may be due to changes that are intrinsic to the object or/and changes induced by the camera. Changes induced by the camera are not part of object action. We therefore assume that camera-induced global motion, such as dynamic pan, tilt, and zoom, as well as global chromatic and photometric changes, such as dynamic white balance and automatic exposure gain control, have been compensated for in the dictionary and test samples.

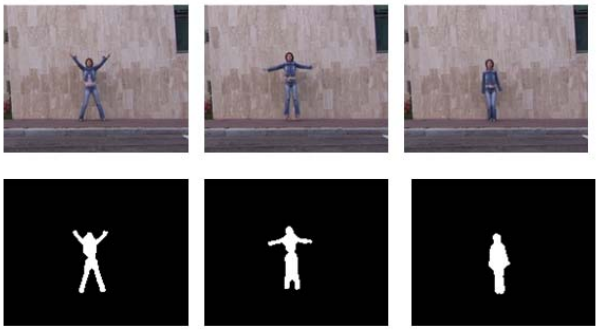

Figure 1. Example of a human action sequence: Three frames from a "jumping-jack" action sequence (top row) and corresponding silhouettes (bottom row) from the Weizmann Human Action Database (see Section IV).

Let us now consider changes in object shape and position associated with action. Changes in object position are related to object motion whereas changes in object shape are related to the relative movement of parts composing the object, e.g., limbs of a human. Although changes in object position may be indicative of the type of action, e.g., running versus walking, they may be unreliable: people walk at different paces and the video frame rate of samples in the dictionary may

[1]By object position we mean the location of the object centroid.

be different from that of the test sample. On the other hand, changes in object shape are quite indicative of action type; for example, crouching can be easily distinguished from walking by analyzing limb movements. Even running and walking have different shape deformation characteristics. Due to these considerations, we first remove object motion, as described towards the end of the next paragraph, and base our action recognition algorithm on changes in object shape.

Objects which undergo similar repetitive changes in shape over time, i.e., objects performing similar actions, can have very different chromatic, photometric, and textural properties in different scenes. Action recognition algorithms should be relatively invariant to these properties. One approach for developing algorithms with these invariance properties is to base them directly on the sequence of 2-D silhouettes of the moving and deforming object (see Fig. 1). Simple background subtraction techniques such as in [21], [22] and more-advanced spatio-temporal video segmentation methods based on level-sets [10] are capable of producing an object silhouette sequence from a raw video action sequence. Under ideal conditions, each frame in the silhouette sequence would contain a white mask (white = 1) which exactly coincides with the 2-D silhouette of the moving and deforming object against a "static" black background (black = 0). A sequence of such object silhouettes in time forms a spatio-temporal volume in x-y-t space that we refer to as a silhouette tunnel. As described in the previous paragraph, since changes in object position are of secondary importance for action recognition, we need to remove object motion. We can do this by aligning the centroids of object silhouettes in the background-subtracted sequence to the same spatial coordinates.

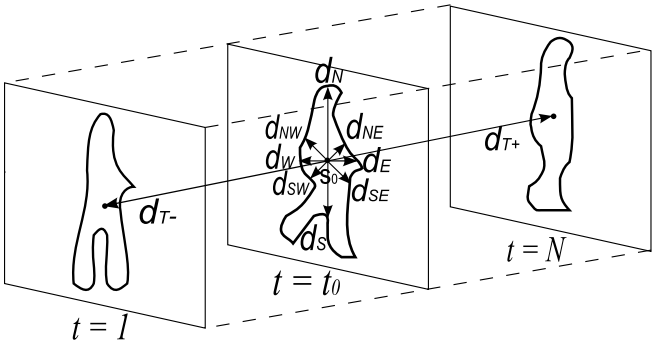

Figure 2. Each point $\mathbf{s}_0 = (x_0, y_0, t_0)^T$ of a silhouette tunnel within an N-frame action segment has a 13-dimensional feature vector associated with it: 3 position features x_0, y_0, t_0, and 10 shape features given by distance measurements from (x_0, y_0, t_0) to the tunnel boundary along 10 different spatio-temporal directions shown in the figure.

One final important aspect of object action that one needs to take cognizance of is the *repetitive* nature of shape changes which characterize an action. Many interesting actions such as walking and running consist of multiple, roughly periodic, "repetitions" of action segments that have similar shape deformations that define their essential character (see the cartoon illustration of an action segment in Fig. 2). These "repetitive" action segments within the same sequence and segments across different sequences corresponding to the same action will be similar in shape and duration but there will also be a good degree of variability. From the recognition viewpoint, this motivates the need for a dictionary of labeled sample action segments which are representative of the statistical variability that one expects to encounter within a given application context. From a processing viewpoint, this motivates the need to break a silhouette sequence into a set of (potentially overlapping) successive action segments each of which contains roughly one "period" of the action so that each segment can be individually classified fairly reliably.

Our overall framework for action recognition can be summarized as follows: We start with a raw test video sequence containing a single object whose position and shape change with time. Camera-induced global artifacts are assumed to have been compensated for in the raw video. We refer to this raw video sequence as the *action sequence*. The action sequence is subjected to background subtraction and centroid alignment to obtain what we refer to as the *silhouette sequence*. The silhouette sequence is then broken into a sequence of *overlapping* N-frame-long *action segments* where N is assumed to be large enough to contain roughly one complete "cycle" that is representative of the action. Each test action segment is then compared with the segments in a dictionary of previously-labeled action segments to find the most similar one. The test action segment is assigned the label of the most similar dictionary segment. Since individual segment decisions are expected to be somewhat noisy on the account of issues discussed in the introduction, we propose an additional step to filter out this decision noise. We propose to fuse the decisions of all action segments in an action sequence using the majority rule to arrive at the final decision for the entire test action sequence as illustrated in Fig. 3. This improves the reliability by overcoming misclassifications in up to one-half of the test action segments.

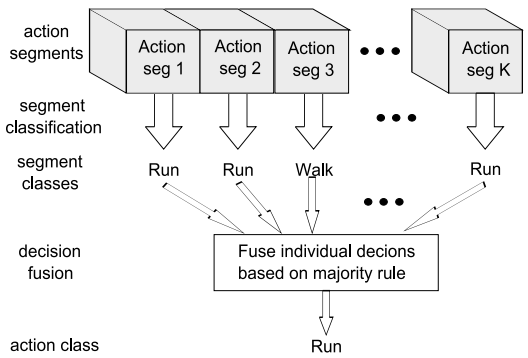

Figure 3. Action classification of a group of action segments by majority vote applied to segment decisions.

The key ingredient needed for successful action recognition at the segment level is the metric used for measuring how close a pair of silhouette tunnels are in terms of their shape. This metric must 1) allow reliable discrimination of different actions, 2) be easily computable so that a real-time operation is attainable, and 3) be invariant to spatial scaling. In the next section, we describe the novel object shape similarity features and the metric based on these features that we have developed that posses the three aforementioned properties.

III. Silhouette tunnel shape representation and comparison

There is an extensive body of literature devoted to the representation and comparison of shapes of volumetric objects. A variety of approaches have been explored ranging from deterministic mesh models used in the graphics community to statistical models, both parametric (e.g., ellipsoidal models) and non-parametric (e.g., Fourier descriptors). Our goal is to develop a fast and simple algorithm which can reliably discriminate between different shape classes without the need for human intervention. Our goal is not to accurately reconstruct any given shape per se. The ease of computation is an important factor in our work.

We first sketch the overall approach and then get into the technical details. Given an action segment, we first extract a rich collection of 13-dimensional feature vectors which describe the geometry of the silhouette tunnel. This initial collection of feature vectors provides a shape representation for the silhouette tunnel which is overcomplete; the tunnel can be completely reconstructed using this representation and there are more components in this representation than in the boundary of the tunnel. We view this step as a "feature-expansion" step in which the silhouette tunnel is embedded within a higher-dimensional feature space. The second step is a dimension reduction step: we obtain a simplified representation for the shape of the silhouette tunnel by "fusing" the overcomplete collection of feature vectors into a 13×13 empirical "shape" covariance matrix of the feature vectors in the collection. We view this shape covariance matrix as a reduced shape representation of the silhouette tunnel. The shape covariance matrix thus obtained is not invariant to spatial scaling (zoom). We obtain a scale-invariant shape covariance matrix by normalizing the individual elements of the shape covariance matrix by suitable factors or equivalently by normalizing the feature vectors before computing the covariance matrix. Finally, we compare the shape similarity between a silhouette tunnel in a dictionary action segment and a test action segment by measuring the distance between their normalized shape covariance matrices. The distance between two covariance matrices is measured using a Riemannian metric based on their generalized eigenvalues which respects the manifold structure of covariance matrices. We now explain each of these steps in detail.

A. Shape feature vectors

Let $\mathbf{s} = (x, y, t)^T$ denote the horizontal, vertical, and temporal coordinates of a pixel. Let $\mathcal{A}$ denote the set of coordinates of all pixels belonging to an action segment which is W pixels wide, H pixels tall, and N frames long, i.e., $\mathcal{A} := \{(x, y, t)^T : x \in [1, W], y \in [1, H], t \in [1, N]\}$. Let $\mathcal{S}$ denote the subset of pixel-coordinates in $\mathcal{A}$ which belong to the silhouette tunnel. With each pixel located at $\mathbf{s}$ within the silhouette tunnel, we associate the following 13-dimensional feature vector $\mathbf{f}(\mathbf{s})$ that captures certain shape characteristics of the tunnel:

$$\begin{aligned}\mathbf{f}(x, y, t) :=&[x, y, t, d_E, d_W, d_N, d_S,\\ &d_{NE}, d_{SW}, d_{SE}, d_{NW}, d_{T+}, d_{T-}]^T,\end{aligned} \tag{1}$$

where $(x, y, t)^T \in \mathcal{S}$ and d_E, d_W, d_N, and d_S are Euclidean distances from (x, y, t) to the nearest silhouette boundary point to the right, to the left, above and below the pixel, respectively. Similarly, d_{NE}, d_{SW}, d_{SE}, and d_{NW} are Euclidean distances from (x, y, t) to the nearest silhouette boundary point in the four diagonal directions, while d_{T+} and d_{T-} are similar measurements in the temporal direction. Clearly, these 10 distance measurements capture silhouette tunnel shape as "seen" from location $(x, y, t)^T$. Fig. 2 depicts these features graphically.

There is one shape feature vector $\mathbf{f}$ associated with each pixel of a silhouette tunnel, and thus there are a large number of feature vectors. The collection of all feature vectors $\mathcal{F}(\mathcal{S}) := \{\mathbf{f}(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$ is an overcomplete representation of the shape of the silhouette tunnel because $\mathcal{S}$ is completely determined by $\mathcal{F}$ and $\mathcal{F}$ contains additional data which are redundant. It is instructive to see how individual feature components change with the change of pixel location. Fig. 4 depicts each of the 13 features for a single silhouette frame (x-y slice of a silhouette tunnel for a fixed value of t) as an intensity image, where higher brightness means larger value of that feature. In this figure, the origin of the coordinate system is in the left-top corner of the image. Note that the intensity of the x-component image increases linearly from left to right inside the silhouette whereas the intensity of the y-component image increases from top to bottom. However, the intensity of the t-component image is spatially-constant since all pixels in the same frame have the same value of t. Similarly, the d_W image has lower values (dark) at the left of the silhouette since it measures the distance to the left silhouette boundary whereas the d_{T_-} and d_{T_+} images are very bright (large distance) in the torso and darker (shorter distance) within the limb areas. This is to be expected since the position of the torso is largely unchanged across time after centroid alignment whereas legs and arms move significantly. This potentially shortens the temporal distance to the silhouette tunnel boundary.

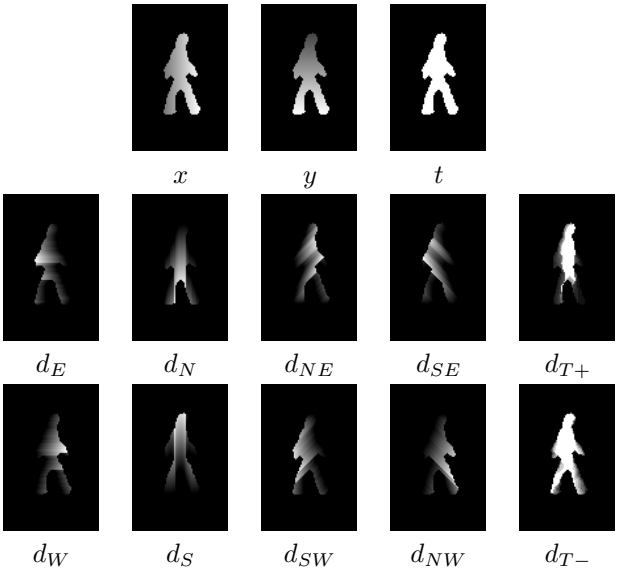

Figure 4. Individual components of feature vectors $\mathbf{f}(x, y, t)$ depicted as intensity images with t fixed and (x, y) variable. The origin is at the top left corner and brighter points denote larger values.

The feature vectors combine absolute pixel locations with the relative distances of pixels to tunnel boundaries measured along 10 different directions. Moreover, since each silhouette tunnel is likely to contain a different number of pixels, it will also contain a different number of feature vectors. Therefore, it is unclear how to measure the similarity between two silhouette tunnels based on the entire collection $\mathcal{F}(\mathcal{S})$ of these feature vectors.

B. Shape covariance matrix

Recently, Tuzel *et al.* [19], [20] proposed to compare two sets of feature samples by computing and comparing their empirical covariance matrices. Their application context was object tracking but this approach easily extends to our problem of comparing shapes of silhouette tunnels. The advantage of this approach lies in the fact that completely different features, such as pixel locations and luminance gradients, can be combined in one feature vector and used jointly for comparison. Moreover, since feature properties are captured in a single covariance matrix, feature sets of different sizes can be easily and efficiently compared.

We capture the shape properties of silhouette tunnel $\mathcal{S}$ by a 13×13 shape covariance matrix $C_{\mathcal{S}}$ defined as follows. Let $\mathbf{S} = (X, Y, T)^T$ denote a random location vector which is uniformly distributed over $\mathcal{S}$, i.e., the probability mass function of $\mathbf{S}$ is equal to zero for all locations $\mathbf{s} \notin \mathcal{S}$ and is equal to $1/|\mathcal{S}|$ at all locations in $\mathcal{S}$, where $|\mathcal{S}|$ denotes the volume of the silhouette tunnel. Then, $C_{\mathcal{S}} := \text{cov}(\mathbf{F})$ where $\mathbf{F} := \mathbf{f}(\mathbf{S})$. More explicitly,

$$C_{\mathcal{S}} := \text{cov}(\mathbf{F}) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{s} \in \mathcal{S}} (\mathbf{f}(\mathbf{s}) - \boldsymbol{\mu}_F)(\mathbf{f}(\mathbf{s}) - \boldsymbol{\mu}_F)^T \qquad (2)$$

where $\boldsymbol{\mu}_F = \mathbb{E}[\mathbf{F}] = \sum_{\mathbf{s} \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \mathbf{f}(\mathbf{s})$ is the mean feature vector. Thus, $C_{\mathcal{S}}$ is an empirical covariance matrix of the collection of vectors $\mathcal{F}(\mathcal{S})$. It captures the *second-order* empirical statistical properties of the collection. Note, that the volume of a silhouette tunnel $|\mathcal{S}|$ is typically more than 10^4, often more than 10^5. Since a covariance matrix is symmetric, only $(13^2 + 13)/2 = 91$ of its entries are independent thus affording a low-dimensional representation of all feature samples, independently of their number.

C. Normalization for spatial scale-invariance

The shape covariance matrix $C_{\mathcal{S}}$ in (2) computed from the 13 features in (1) is not invariant to *spatial* scaling of the silhouette tunnel, i.e., two silhouette tunnels $\mathcal{S}$ and $\mathcal{S}'$ that have identical shape but differ in spatial scale will have different covariance matrices. To illustrate the problem, ignoring integer-valued constraints, let $a > 0$ be a spatial scale factor and let $\mathcal{S}' := \{(ax, ay, t)^T : (x, y, t)^T \in \mathcal{S}\}$ be a silhouette tunnel obtained from $\mathcal{S}$ by stretching the horizontal and vertical dimension (but not time) by the factor a. Then $|\mathcal{S}'| = a^2|\mathcal{S}|$. Consider the covariance between the x-coordinate and the distance to the top boundary d_N (both are spatial features) for both $\mathcal{S}$ and $\mathcal{S}'$. These are respectively given by $\text{cov}(X, D_N)$ and $\text{cov}(X', D'_N)$ where $X' = aX$ and $D'_N = aD_N$. Consequently, $\text{cov}(X', D'_N) = a^2\text{cov}(X, D_N)$. An identical relationship holds for the covariance between any pair of spatial features. The covariance between any spatial feature and any temporal feature for $\mathcal{S}'$ will be a times that for $\mathcal{S}$ (instead of a^2) and the covariance between any pair of temporal features for $\mathcal{S}'$ and $\mathcal{S}$ will be equal. To see how the shape covariance matrix can be made invariant to spatial scaling of the silhouette tunnel, observe that $\text{cov}(X'/\sqrt{|\mathcal{S}'|}, D'_N/\sqrt{|\mathcal{S}'|}) = \text{cov}(X/\sqrt{|\mathcal{S}|}, D_N/\sqrt{|\mathcal{S}|})$. Thus to obtain a spatially scale-invariant shape covariance matrix, we must divide every spatial feature by the square root of the volume of the silhouette tunnel before computing the empirical covariance matrix using (2).

A similar approach can be used for temporal scaling which can arise due to frame rate differences between the test and dictionary action segments. However, since most cameras run at either 30 or 60 frames/fields per second, in this work we assume that the frame rates are identical and the segment size N is the same for the test and dictionary action segments. By construction, the shape covariance matrix is automatically invariant to spatio-temporal translation of the silhouette tunnel. It is, however, not invariant to rotation of the silhouette tunnel about the horizontal, vertical, and temporal axes. Rotations about the temporal axis by multiples of 45° have the effect of permuting the spatial components of the feature vector. In this work, however, we will assume that the test and dictionary silhouette tunnels have roughly the same spatial orientation. Rotations about the horizontal and vertical axes are less of a problem in practice because they may not correspond to meaningful

real-world silhouette tunnels.

D. Shape similarity metric

So far, we have identified which features to extract, argued that we can compactly characterize them using a covariance matrix, and shown how to normalize the covariance matrix to assure scale invariance. In order to compare the shapes of two silhouette tunnels (which describe an action), we need a metric defined in the space of covariance matrices that would tell us if the two covariance matrices describe similar shapes (actions) or not.

The set of all covariance matrices is not a Euclidean space because it is not closed under multiplication with negative scalars. Covariance matrices do, however, lie on a Riemannian manifold. Förstner and Moonen [23] proposed the following distance measure between two $d \times d$ covariance matrices:

$$\rho(C, C') := \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C, C')}, \tag{3}$$

where $\lambda_k(C, C')$ are the generalized eigenvalues of C and C', i.e.,

$$\lambda_k C \mathbf{u}_k = C' \mathbf{u}_k,$$

where $\mathbf{u}_k \neq \mathbf{0}$ is the k-th generalized eigenvector. This distance measure captures the manifold structure of covariance matrices and satisfies the metric axioms of positivity, symmetry, and triangle inequality. It has been used successfully in object tracking and face localization [19], [20]. We adopt this metric for shape comparison.

IV. Experimental results

In order to test the efficiency and performance of the proposed method, we conducted a series of experiments on the Weizmann Human Action Database available online[2] [7]. The database contains 90 low-resolution video sequences (180×144 pixels, 50fps) that show 9 different people with each person performing 10 different actions, such as jumping, walking, running, skipping, etc. These video sequences are typically 80 to 120 frames long. For each video sequence a binary sequence of 2-D silhouettes is also available (Fig. 1). As described earlier, in a preprocessing step, we align the centroids of individual 2-D silhouettes to compensate for any global object motion present. This typically changes the spatial dimensions by 10 to 20 pixels.

We measured the performance of our action recognition algorithm by the following cross-validation test. First, we divided each of the 90 silhouette sequences into overlapping N-frame long action segments (we tried both $N = 8$ and $N = 20$) with a 4-frame overlap.[3] We refer to this as the action segment database. Then, we selected one action segment (the query or test action segment) and deleted from the action segment database all those action segments which came from the same video sequence as the query segment. Finally, among the remaining action segments in the database (the action segment dictionary), we found an action segment that best matches our query action segment using the proposed distance measure (3). This is the nearest-neighbor classification rule. The query action segment was considered to be correctly classified if it had the same action label as the best-matched segment. We repeated the procedure for all query action segments in the database and calculated the correct classification rate (CCR) as the percentage of query action segments which were correctly classified. We attained a CCR of 97.05% for action segments of length $N = 8$. Table I shows the action "confusion" matrix for the entire database. The element in row i and column j of the matrix indicates the percentage of action i segments which were classified as action j. The sum of all elements in every row is 100%. Note that most of the errors are for "skipping" (action 7) which tends to be confused with "running" (action 5). A careful examination of these two actions indeed confirms that they have similar dynamics, potentially leading to confusion especially when the action segments are very short.

For comparison, Table II shows the action confusion matrix for the method proposed in [7] also applied to 8-frame action segments of the same database. This method is based on solving a Poisson equation. The overall CCR attained by this method is 97.83%. Although the overall CCR of this method is 0.78% higher than that of the proposed method, this comes at the expense of significantly increased computational complexity. As reported by Gorelick *et al.* [7], the overall processing time of their simulations performed in Matlab for a *particular* 50-frame pre-segmented silhouette sequence of spatial resolution 110×70 is 30 seconds on a 3GHz Pentium. This includes solving the Poisson equation, extracting features and computing moments for all action segments in the sequence. Since we did not have a 3GHz Pentium 4 available, we ran our simulations on a slower 2.8GHz Pentium 4 also in Matlab. Our algorithm takes 6 seconds *on average* on the Weizmann Human Action Database. This includes division into action segments, extraction of 13-dimensional feature vector sets for each segment, and computation of all segment feature covariance matrices (2). Clearly our algorithm has a 5-fold lower computational complexity while giving only slightly lower recognition performance. Furthermore, the simplicity of our algorithm makes it attractive for practical implementations.

In the above experiments, silhouette sequences were divided into overlapping action segments of length $N = 8$ to match the experiments carried out in [7]. The video frame rate of sequences in the database is 50fps. Thus any 8-frame segment extends over roughly 1/6-th of a second.

[2] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

[3] Thus there are typically 20 to 30 action segments.

Table I
ACTION CONFUSION MATRIX FOR THE PROPOSED METHOD ON 8-FRAME SEGMENTS (CCR = 97.05%).

	bend	jack	jump	sjump	run	side	skip	walk	wave1	wave2
bend	98.6	0	0	0	0	0	0	0	0	1.4
jack	0	100	0	0	0	0	0	0	0	0
jump	0		96.1	0	0	0	3.9	0	0	0
sjump	0	0	0	99.3	0	0	0	0	0	0.7
run	0	0	0	0	94.0	1.2	2.4	2.4	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	1.0	0	10.3	0	86.7	2.1	0	0
walk	0	0	0	0	1.3	0	0	98.7	0	0
wave1	0	0	0	0	0	0	0	0	98.0	2.0
wave2	0	0	0	0	0	0	0	0	4.9	95.1

Table II
ACTION CONFUSION MATRIX FOR METHOD FROM [7] ON 8-FRAME SEGMENTS (CCR = 97.83%).

	bend	jack	jump	sjump	run	side	skip	walk	wave1	wave2
bend	99.1	0	0	0	0	0	0	0	0	0.9
jack	0	100	0	0	0	0	0	0	0	0
jump	0		89.2	0	0	0	10.8	0	0	0
sjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	98.0	0	2.0	2.4	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	0	0	2.1	0	97.1	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0.9	0	0.9	0	0	0	0	94.8	3.5
wave2	0	0.9	0	0	0	0	0	0	1.9	97.2

Table III
ACTION CONFUSION MATRIX FOR THE PROPOSED METHOD ON 20-FRAME SEGMENTS (CCR = 98.68%).

	bend	jack	jump	sjump	run	side	skip	walk	wave1	wave2
bend	98.6	0	0	0.9	0	0	0	0	0.9	0
jack	0	100	0	0	0	0	0	0	0	0
jump	0		97.3	0	0	0	2.7	0	0	0
sjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	100	0	0	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	1.0	0	7.0	0	91.6	1.4	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0	0	0	0	0	99.2	0.8
wave2	0	0	0	0	0	0	0	0	0.9	99.1

This may be too short to provide adequate information for classification, since 1/6-th of a second may be insufficient to cover one complete period of the repetitive-structure of an action. Therefore, we can expect that as the length of video segments increases, more discriminative information will become available and this can potentially lead to improved classification performance. Table III shows the action confusion matrix for our method when 20-frame segments are used. 20 frames typically cover at least one full period of an action captured at 50fps. The overall CCR in this case is 98.68%, which improves the classification performance over 8-frame segments by 1.63%.

Thus far, we have described experiments in which individual action segments are classified using an action dictionary and the nearest-neighbor rule according to the metric (3). Due to event complexity, imaging conditions or action repeatability errors, our segment-by-segment decisions are occasionally erroneous. However, when we fuse the decisions from all individual action segments in a given silhouette sequence by using the majority rule to generate the final decision (Fig. 3), the CCR turns out to be always 100%. Admittedly, the use of majority-rule fusion may not be straightforward in streaming video since it is not obvious which video segments should be fused together (i.e., the temporal boundaries of fusion may not be known beforehand), but even a simple strategy of fusing a fixed

number of segments is likely to improve classification rates.

V. Concluding remarks

In this paper, we proposed a new approach to action recognition in video based on a pre-computed and labeled dictionary of object silhouette tunnel segments. The proposed approach involves a novel action comparison method based on a set of shape features of the silhouette tunnels and a metric based on their empirical covariance matrices. This comparison method, applied segment by segment, has been shown to be both effective and efficient. Our experimental results indicate that the proposed method has similar classification performance to that of a recent method of Gorelick *et al.* but at significantly reduced computational complexity. We have also proposed the majority rule to fuse action decisions resulting from segment-by-segment classification; the outcome of fusion is a classification of the complete video as a specific action. Our ongoing research is directed towards addressing numerous pending challenges that accompany real-world video such as object occlusion and clutter, detecting the temporal boundaries of action segments, managing action segment dictionaries, and handling "endless" streaming video.

References

[1] J. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Jun. 2007.

[2] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. European Conf. Computer Vision*, 2006.

[3] G. Mori and J. Malik, "Estimating body configurations using shape context matching," in *Proc. European Conf. Computer Vision*, Jan. 2002.

[4] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Jun. 2004.

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Computer Vision*, 2003.

[6] H. Sidenbladh and M. J. Black, "Learning the statistics of people in images and video," *Intern. J. Comput. Vis.*, vol. 54, no. 1–3, pp. 181–207, Aug. 2003.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[8] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *Proc. IEEE Int. Conf. Computer Vision*, 2005.

[9] E. Shechtman and M. Irani, "Space-time behavior-based correlation or how to tell if two underlying motion fields are similar without computing them?" *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 11, pp. 2045–2056, Dec. 2007.

[10] M. Ristivojević and J. Konrad, "Space-time image sequence analysis: object tunnels and occlusion volumes," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 364–376, Feb. 2006.

[11] J. Konrad, "Videopsy: Dissecting visual data in space-time," *IEEE Comm. Mag.*, vol. 45, no. 1, pp. 34–42, Jan. 2007.

[12] Y. Pritch, A. Rav-Acha, and S. Peleg, "Non-chronological video synopsis and indexing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 11, pp. 1971–1984, Nov. 2008.

[13] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[14] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 12, pp. 1505–1518, 2003.

[15] D. Ioannidis, D. Tzovaras, I. G. Damousis, S. Argyropoulos, and K. Moustakas, "Gait recognition using compact feature extraction transforms and depth information," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3 (part 2), pp. 623–630, 2007.

[16] R. T. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2002, pp. 366–371.

[17] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, 2004.

[18] D. Cunado, M. S. Nixon, and J. N. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Comput. Vis. Image Underst.*, vol. 90, no. 1, pp. 1–41, 2003.

[19] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. European Conf. Computer Vision*, May 2006.

[20] ——, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[21] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density for visual surveillance," *Proc. IEEE*, vol. 90, pp. 1151–1163, 2002.

[22] J. McHugh, J. Konrad, V. Saligrama, and P.-M. Jodoin, "Foreground-adaptive background subtraction," *IEEE Signal Process. Lett.*, vol. 16, no. 5, pp. 390–393, May 2009.

[23] W. Förstner and B. Moonen, "A metric for covariance matrices," Dept. of Geodesy and Geoinformation, Stuttgart University, Tech. Rep., 1999.