Action Recognition in Video by Sparse Representation on Covariance Manifolds 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

Abstract. A novel framework for action recognition in video using empirical covariance matrices of bags of low-dimensional feature vectors is developed. The feature vectors are extracted from segments of silhouette tunnels of moving objects and coarsely capture their shapes. The matrix logarithm is used to map the segment covariance matrices, which live in a nonlinear Riemannian manifold, to the vector space of symmetric matrices. A recently developed sparse linear representation framework for dictionary-based classification is then applied to the log-covariance matrices. The log-covariance matrix of a query segment is approximated by a sparse linear combination of the log-covariance matrices of training segments and the sparse coefficients are used to determine the action label of the query segment. This approach is tested on the Weizmann and the UT-Tower human action datasets. The new approach attains a segment-level classification rate of 96.74% for the Weizmann dataset and 96.15% for the UT-Tower dataset. Additionally, the proposed method is computationally and memory efficient and easy to implement.

Keywords: video analysis; action recognition; silhouette tunnel; covariance manifold; sparse linear representation

1 Introduction

Algorithms for recognizing human actions in a video sequence are needed in applications such as video surveillance, where the goal is to look for typical and anomalous patterns of behavior, and video search and retrieval in large, potentially distributed, video databases such as YouTube. Developing algorithms for action recognition in video that are not only accurate but also efficient in terms of computation and memory-utilization is challenging due to the complexity of the task and the sheer size of video.

The action recognition problem, in its full generality, is challenging due to the complexity of the scene (multiple interacting moving objects, clutter, occlusions, illumination variability, etc.), the camera (imperfections, motion and shake, and viewpoint),

* This material is based upon work supported by the US National Science Foundation (NF) under awards CANS–0721884 and (CAREER) CC–0546598, and National Geostatics-Intelligence Agency (NAG) under award HM1582-09-1-0037. 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 NF and NGA.

and the complexity of actions (non-rigid objects and intra- and inter- class action variability). Even when there is only a single uncluttered and unoccluded object [1] and the camera and illumination conditions are perfect (a typical assumption in the literature), the complexity and variability of actions makes action recognition a difficult problem.

The accuracy and efficiency of an action recognition algorithm critically depends on 1) *how actions are modeled and represented* and 2) *how distances between action representations are measured for classification*. To date, various action models and representations have been proposed, from those based on Hidden Markov Models [14, 12], through interest-point models [10, 4, 9, 8] which are sparse (relative to the number of pixels) yet highly discriminative, e.g., corners and SIFT features, and local motion models, e.g., kinematic characteristics from optical flow [1] and 3D local steering kernels [11], to silhouette tunnel shape models [6, 7]. Similarly, various metrics have been proposed to measure distances between action representations, from the Hausdorff distance between sets of action feature vectors in Euclidean space extracted from multiple action instances (e.g., see [6]) to the matrix cosine similarity measure (Frobenius inner product) between matrices of action feature vectors [11]. The methods developed to-date are either computationally and/or memory intensive and/or their accuracy varies significantly across different data sets.

In [7] we developed a nearest-neighbor (NN) supervised classification algorithm for human action recognition using a labeled dictionary of empirical feature-covariance matrices. These were obtained from bags of low-dimensional feature vectors extracted from the object silhouette tunnels and coarsely captured their shape. A Riemannian metric on the manifold of covariance matrices was used for determining nearest neighbors. In this paper, we apply the recently developed sparse linear representation framework for dictionary-based classification [13] to the matrix logarithm of the feature-covariance matrices as an alternative to NN-classification. We report the performance of this new approach on the Weizmann human action dataset [6] and the UT-tower dataset [3] provided by the ICPR 2010 "Aerial View Activity Classification Challenge". We also compare its performance with the method we previously developed in [7] that uses the same action representation (covariance matrix of silhouette shape features) but a different classification rule (NN-classifier).

2 Framework

We view action recognition as a supervised classification problem where the goal is to classify a query video segment using a dictionary of previously labeled training video segments. Video segments are typically high dimensional, e.g., a 20-frame video segment with a 128×128 frame resolution is, roughly, a 3×10^5-dimensional vector, whereas the number of training video segments is meager in comparison. It is therefore impractical to learn the global structure of training video segments by building classifiers directly in high-dimensional space. Graphical models, which attempt to capture global dependencies through local structure, are powerful; but training classifiers based on these models is challenging.

[1] Such footage may be obtained by detecting, tracking, and isolating object trajectories.

2.1 Action representation using low-dimensional feature-covariance matrices

We adopt a "bag of dense local feature vectors" modeling approach wherein a video segment is represented by a *dense* set of low-dimensional local feature vectors which describe the action. The local features, described in detail in Sec. 3, coarsely capture the shape of an object's silhouette tunnel (see Fig. 3). The advantage of this approach is that even a single video segment provides a very large number of local feature vectors (one per pixel) from which their statistical properties can be reliably estimated. However, the dimensionality of a bag of dense local feature vectors is still very high as there are as many feature vectors as pixels. This motivates the need for dimensionality reduction.

Estimating the distribution of the local feature vectors, though ideal, is computation-intensive and may not lead to a lower-dimensional representation. On the other hand, the mean feature-vector, which is low-dimensional, can be learned reliably and rapidly but may not be sufficiently discriminative. In the recent work [7] we discovered that if the features are well-chosen, then the feature-covariance matrix, which captures the second-order statistical properties of a bag of feature vectors, provides a remarkably discriminative representation for action recognition. In addition to their simplicity and effectiveness, covariance matrices have low storage and processing requirements. The action representation based on the covariance matrix of a bag of low-dimensional local feature vectors that coarsely capture the shape of an object's silhouette tunnel is depicted in Fig. 1. The operator which transforms an input video segment into an output feature-covariance matrix representation is denoted by Ψ.

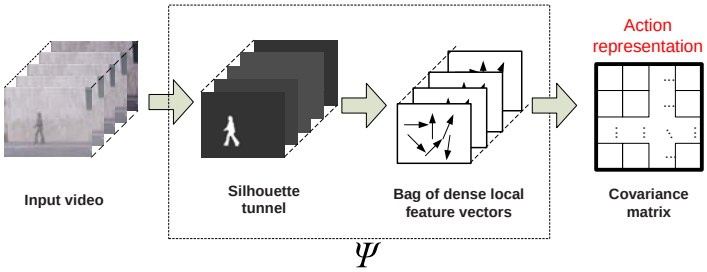

Fig. 1: Transformation of a video segment into a feature covariance matrix representation.

2.2 Classification on a covariance manifold

The set of all covariance matrices of a specified size do not form a vector space (they are not closed under multiplication by negative scalars); they form a Riemannian manifold. Classification problems on covariance manifolds can be converted into vector-space classification problems via the matrix logarithm: if $C = UDU^T$ is the eigendecomposition of the covariance matrix C, where D is the diagonal matrix of eigenvalues, then $\log(C) := U \log(D) U^T$, where $\log(D)$ is the diagonal matrix whose diagonal entries are the natural logarithms of the corresponding entries of D. The matrix logarithm maps the Riemannian manifold of symmetric non-negative definite matrices to the vector space of symmetric matrices [2].

Recently, in [13] Wright et al. developed a powerful framework (closely related to compressive sensing) for supervised classification in vector spaces based on finding a

sparse linear approximation of a query vector in an overcomplete dictionary of training vectors. The key idea underlying this approach is that a query vector can typically be well approximated by a sparse linear combination of training vectors belonging to the same class as the query but cannot be as well approximated by training vectors coming from a different class. A sparse linear representation can be obtained by solving an l^1-minimization problem described in Sec. 4. Locations of large non-zero coefficients in the sparse linear approximation are likely to indicate the label of the query vector. This approach has been successfully applied to many vision tasks such as face recognition, image super-resolution, and image denoising. We extend this approach to action recognition by applying it to (column) vectorized log-covariance matrices that we refer to as samples. Specifically, we approximate the log-covariance matrix of a query segment by a sparse linear combination of log-covariance matrices of all training segments.

The overall framework for action recognition is depicted in Fig. 2.

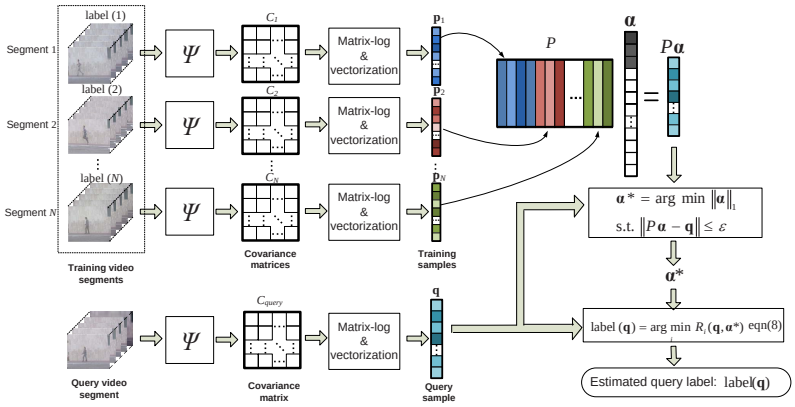

Fig. 2: Overview of the proposed action recognition framework (see Secs. 2.2 and 4).

3 Silhouette tunnel shape features

In this section, we describe the low-dimensional local features that we use to describe actions. The sequence of 2-D silhouettes of a moving and deforming object (see Fig. 3) is particularly attractive for action recognition because (i) it accurately captures object dynamics, (ii) it can be reliably, robustly, and efficiently computed in real-time using state-of-the-art background subtraction techniques, and (iii) it is largely invariant to chromatic, photometric, and textural properties of objects which are independent of their actions. 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. Action recognition may then be viewed as recognizing the *shape* of the silhouette tunnel. There is an extensive body of literature devoted to the representation and comparison of shapes of volumetric objects. Our goal is to reliably discriminate between shapes; not to accurately reconstruct them. Hence a coarse, low-dimensional representation of shape would suffice. We capture the shape

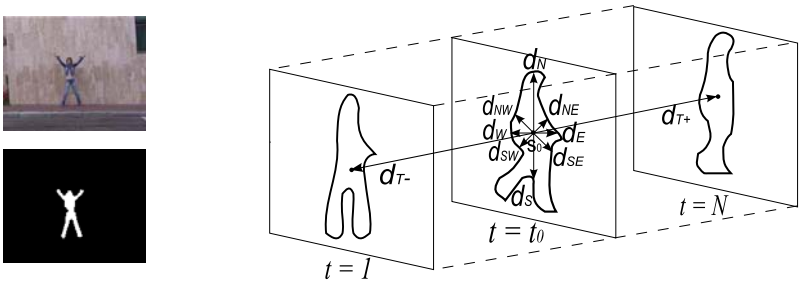

Fig. 3: *Left:* One frame of the "jumping-jack" human action sequence (top row) and the corresponding silhouette (bottom row) computed using background subtraction from the Weizmann human action dataset. *Right:* Each point $\mathbf{s}_0 = (x_0, y_0, t_0)^T$ of a silhouette tunnel within an N-frame action 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.

of a silhouette tunnel by the empirical covariance matrix of a bag thirteen-dimensional local shape features (described below) from our previous work [7].

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 (video) 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:

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

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. Fig. 3 depicts these features graphically. The 13×13 "shape" covariance matrix representation $C_{\mathcal{S}}$ of silhouette tunnel $\mathcal{S}$ in the action segment $\mathcal{A}$ is given by

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

where $\boldsymbol{\mu}_F = \sum_{\mathbf{s} \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \mathbf{f}(\mathbf{s})$ is the mean feature vector. Note that the size of an action segment $|\mathcal{A}|$ is typically on the order of 10^5 whereas a 13×13 covariance matrix, being symmetric, has only 91 independent entries. This provides a low-dimensional representation of the feature vectors no matter how numerous they may be.

4 Classification via sparse linear representation

In this section, we first explain how the log-covariance matrix of a query action segment can be approximated by a sparse linear combination of log-covariance matrices of all

training action segments by solving an l^1-minimization problem. We then discuss how the locations of large non-zero coefficients in the sparse linear approximation can be used to determine the label of the query.

The logarithm of a 13×13 covariance matrix C is a 13×13 symmetric matrix $\log(C)$ which has only 91 independent entries (elements on and above the main diagonal). We use $\mathbf{p} \in \mathbb{R}^{91}$ to denote the (column) vectorized matrix of the entries in $\log(C)$ that are on or above the main diagonal. For convenience of exposition, we will refer to such column vectorized log-covariance matrices as simply 'samples'. Let $\mathbf{p}_{i,j} \in \mathbb{R}^{91}$ denote the j-th training sample in the i-th class where $i = 1, \ldots, K$, and $j = 1, \ldots, n_i$. Thus there are K action classes, n_i training samples in action class i, and the total number of training samples is given by $M = \sum_{i=1}^{K} n_i$. We can stack up all the training samples from class i, column by column, to form the $91 \times n_i$ matrix $P_i := [\mathbf{p}_{i,1}\ \mathbf{p}_{i,2}\ \cdots\ \mathbf{p}_{i,n_i}]$. The $91 \times M$ matrix of all training samples is then given by $P := [P_1\ P_2\ \cdots\ P_K]$.

A given query sample $\mathbf{q}$ can be expressed as a linear combination of training samples by solving the matrix-vector equation given by

$$\mathbf{q} = P\boldsymbol{\alpha}, \tag{3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^M$ is the vector of coefficients. Since $M \gg 91$, the system (3) is underdetermined and has a solution except in the highly unlikely circumstance in which there are less than 91 linearly independent samples across all classes and $\mathbf{q}$ is outside of their span. If a solution to (3) exists, it is necessarily nonunique unless additional prior information, e.g., sparsity, restricts the set of feasible $\boldsymbol{\alpha}$.

We seek a sparse solution to (3) where, under ideal conditions, the only non-zero coefficients in $\boldsymbol{\alpha}$ are those which correspond to the class of the query sample. If (3) has a solution $\boldsymbol{\alpha}^*$ with $r < 91/2$ non-zero coefficients and every set of 91 columns of P is linearly independent, then $\boldsymbol{\alpha}^*$ is the unique sparsest solution to (3) (see [5]) which can be found, in principle, by solving the following NP-hard optimization problem:

$$\boldsymbol{\alpha}^* = \arg\min\ \|\boldsymbol{\alpha}\|_0, \qquad s.t.\ \ \mathbf{q} = P\boldsymbol{\alpha}, \tag{4}$$

where $\|\boldsymbol{\alpha}\|_0$ is the so-called l^0-norm: the number of non-zero entries in $\boldsymbol{\alpha}$. A key result in the theory of compressive sensing (see [5]) is that if the optimal solution $\boldsymbol{\alpha}^*$ is sufficiently sparse, then solving the l^0-minimization problem (4) is equivalent to solving the following l^1-minimization problem

$$\boldsymbol{\alpha}^* = \arg\min\ \|\boldsymbol{\alpha}\|_1, \qquad s.t.\ \ \mathbf{q} = P\boldsymbol{\alpha}. \tag{5}$$

Unlike (4), this problem is a convex optimization problem that can be solved in polynomial time. In practice, estimates of $\mathbf{p}_{i,j}$ may be noisy and (3) may not hold exactly. In practice one therefore solves the following ϵ-robust l^1-minimization problem

$$\boldsymbol{\alpha}^* = \arg\min\ \|\boldsymbol{\alpha}\|_1, \qquad s.t.\ \ \|P\boldsymbol{\alpha} - \mathbf{q}\|_2 \leq \varepsilon. \tag{6}$$

It turns out that even when not all sets of 91 columns of P are linearly independent, the solution $\boldsymbol{\alpha}^*$ to (6) is still very sparse in the sense that its components, arranged in decreasing order of magnitude, decay very rapidly.

We now discuss how the locations of large non-zero components of $\boldsymbol{\alpha}^*$ can be used to determine the label of the query. Each component of $\boldsymbol{\alpha}^*$ weights the contribution of its corresponding training sample to the representation of the query sample. Ideally, the sparse non-zero coefficients should only be associated with training samples that come from the same class as the query sample. In practice, however, non-zero coefficients will be spread across more than one action class. To decide the label of the query sample, we follow Wright et al. [13] and use a reconstruction residual error (RRE) measure to decide the query class. Let $\boldsymbol{\alpha}_i^* := [\alpha_{i,1}^* \ \alpha_{i,2}^* \ \cdots \ \alpha_{i,n_i}^*]^T$ denote the coefficients associated with training samples from class i, i.e., columns of P_i. The RRE measure of class i is then defined as:

$$R_i(\mathbf{q}, \boldsymbol{\alpha}^*) := \|\mathbf{q} - P_i \boldsymbol{\alpha}_i^*\|_2. \tag{7}$$

To the query sample $\mathbf{q}$ we assign the class label that leads to the minimum RRE, i.e.,

$$\text{label}(\mathbf{q}) := \arg\min_i R_i(\mathbf{q}, \boldsymbol{\alpha}^*). \tag{8}$$

5 Some practical considerations and the overall algorithm

One important aspect of human action recognition is the repetitive nature of actions. Many actions, such as walking, running and jumping, consist of multiple, roughly periodic, "repetitions" of shorter action segments which describe the essential action characteristics. Long video sequences of the same action may exhibit large differences due to action variability. In addition, the frame-boundaries where one action ends and another begins may not be available in some practical scenarios. This motivates the need to break a long query video sequence into a sequence of overlapping action segments and classify each segment. Short overlapping action segments can also increase the number and diversity of the training set so that the action can be classified more reliably. Ideally, the duration of an action segment should be long enough to contain at least one "period" of an action. The typical period of many moderately-paced human actions is on the order of 0.4-0.8 seconds. For a camera operating at 25 frames per second (fps), this corresponds to an action segment which contains 10–20 frames.

The motion of the centroid of an object's silhouette across frames is of secondary importance for action recognition. It is the sequence of deformations of the silhouettes about their centroids that is crucial. We can remove the motion of the centroids by aligning them to the same spatial coordinates. It is also possible to make the silhouette tunnel shape covariance matrix $C_{\mathcal{S}}$ invariant to spatial scaling (e.g., due to zoom) and temporal scaling (e.g., due to temporal subsampling) by normalizing the feature vectors before computing $C_{\mathcal{S}}$ via (2). We refer to [7] for the details.

The overall framework for action recognition can be summarized as follows (see Figs. 1 and 2). We start with a raw query video sequence which has only one moving object. We compute the silhouette sequence by background subtraction and then parse it into a sequence of overlapping N-frame-long segments (we used 8-frame segments with a 4-frame overlap in our experiments). We map the silhouette tunnel of each N-frame-long action segment to its shape covariance matrix, take its logarithm and column-vectorize the upper-triangular portion. To classify each action segment, we

solve the l^1-minimization problem (6) to obtain a sparse linear representation and then use (7) and (8). Since individual segment decisions are expected to be somewhat noisy, we perform an additional step to filter out this decision noise. We fuse the decisions of all action segments in an action sequence using the majority rule to arrive at the final decision for the entire query video sequence. This improves the reliability by overcoming misclassifications in up to one-half of the test action segments.

6 Experimental results

In this section, we report the results of performance evaluation of the proposed method on two publicly-available datasets: the Weizmann human action dataset [2] and the UT-tower human action dataset [3]. Although the KTH dataset [3] has been widely used to test the performance of action recognition methods, we omit it in our tests since it does not include silhouette sequences that are needed for a fair comparison.

6.1 Weizmann human action dataset

This dataset consists of 90 low-resolution video sequences (180×144 pixels) that show 9 different people each performing 10 different actions. For each video sequence, a binary sequence of 2-D silhouettes is also available. As described in Sec. 5, we parse all silhouette sequences into overlapping 8-frame long silhouette segments with a 4-frame overlap. We refer to the resulting collection of segments as the silhouette *segment* dataset. Performance-evaluation is based on the leave-one-out cross validation (LOOCV) test. For each query silhouette segment from the segment dataset, we first remove all those segments which come from the same silhouette sequence as the query segment. Then, based on the remaining segments in the segment dataset, we determine the action label of the query segment using the proposed method. Details of the experimental setup can be found in [7]. The correct classification rate (CCR) is defined as the percentage of query segments that are correctly classified. Since the CCR is based on classifying individual segments, we call it SEG-CCR. In practice, however, we are usually interested in classifying a complete video sequence containing an action; not just one of its segments. Since segments provide time-localized action information, in order to obtain classification for the complete sequence, we apply the majority rule (dominant label wins) to the decisions obtained from individual segments of the video sequence as described in Sec. 5. In this case, we calculate a sequence-level CCR, that we call SEQ-CCR, defined as the percentage of query sequences that are correctly classified.

The proposed method attained a SEG-CCR of 96.74% and a SEQ-CCR of 100%. Table 1 shows the action "confusion" matrix based on SEG-CCR values. 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%. The confusion matrix indicates that while some actions, such as 'bend' and 'run', are more confusing, others, such as 'walk' and 'side', are easier to distinguish.

[2] http://www.wisdom.weizmann.ac.il/∼vision/SpaceTimeActions.html
[3] http://www.nada.kth.se/cvap/actions/

Table 1: Action confusion matrix: Weizmann human action dataset, 8-frame segments with 4-frame overlap, SEG-CCR = 96.74%.

	bend	jack	jump	sjump	run	side	skip	walk	wave1	wave2
bend	91.9	1.3	0	0.7	0	0	0	0	4.1	2.0
jack	0	99.4	0	0.6	0	0	0	0	0	0
jump	0	0	95.1	0	0	2.0	2.9	0	0	0
sjump	0	0.8	0	96.7	0	2.5	0	0	0	0.7
run	0	0	0	1.2	91.6	0	1.2	6.2	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	1.0	0	4.2	0	92.7	2.1	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0.6	0	0	0	0	99.4	0
wave2	0	1.4	0	0	0	0	0	0	1.4	97.2

Table 2: LOOCV CCR comparison of the proposed method with state-of-the-art methods: Weizmann human action dataset, 8-frame segments with 4-frame overlap.

Method	**Proposed**	Guo *et al.* [7]	Gorelick *et al.* [6]	Niebles *et al.* [9]	Ali *et al.* [1]	Seo *et al.* [11]
SEG-CCR	**96.74%**	97.05%	97.83%	-	95.75%	-
SEQ-CCR	**100%**	100%	-	90%	-	96%

Table 2 compares the performance of the proposed method with some of the state-of-the-art action recognition methods, including our previous method [7] based on NN-classification on the feature-covariance manifold. It is clear that the proposed algorithm is very close in performance to our previous method and also approaches the performance of Gorelick et al.'s method [6].

6.2 UT-tower human action dataset

The UT-tower action dataset is used in the "Aerial View Activity Classification Challenge" at the ICPR 2010 Contest on Semantic Description of Human Activities (SDHA). The dataset consists of 108 video sequences with a frame resolution of 360×240 pixels and a frame rate of 10fps. The contest requires classifying video sequences into one of 9 categories of human actions. Each of the 9 actions is performed 2 times by 6 individuals for a total of 12 video sequences per action category. Ground truth action labels, bounding boxes, and foreground masks for each video sequence are provided. Only the acting person is included in the bounding box. In addition to the challenges associated with the low resolution of objects of interest in this dataset – the average height of human figures is about 20 pixels – there are additional challenges, such as camera jitter, shadows, and blurry visual cues (see [3] for details).

We conducted experiments using the same procedures as for the Weizmann dataset including LOOCV. The method proposed here attains a SEG-CCR of 96.15% and a SEQ-CCR of 97.22%. Table 3 shows the confusion matrices of SEG-CCR and SEQ-CCR values. Since the UT-Tower dataset is new and no action recognition results are publicly available for this dataset at the time of writing of this paper, in Table 4 we only compare the performance of the proposed method with our previous method [7].

Table 3: Action confusion matrices: UT-Tower human action dataset, 8-frame segments with 4-frame overlap.

	SEG-CCR=96.15%									SEQ-CCR=97.22%								
	point	stand	dig	walk	carry	run	wave1	wave2	jump	point	stand	dig	walk	carry	run	wave1	wave2	jump
point	88.0	6.0	6.0	0	0	0	0	0	0	91.7	0	8.3	0	0	0	0	0	0
stand	4.4	94.2	1.4	0	0	0	0	0	0	16.7	83.3	0	0	0	0	0	0	0
dig	2.0	1.5	96.0	0	0.5	0	0	0	0	0	0	100	0	0	0	0	0	0
walk	1.4	0	0	98.6	0	0	0	0	0	0	0	0	100	0	0	0	0	0
carry	0	0	0	0	99.5	0.5	0	0	0	0	0	0	0	100	0	0	0	0
run	0	0	0	0	0	100	0	0	0	0	0	0	0	0	100	0	0	0
wave1	0	0	0.5	0	0	0	94.1	5.4	0	0	0	0	0	0	0	100	0	0
wave2	0	0	0	0	0	0	7.5	92.5	0	0	0	0	0	0	0	0	100	0
jump	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	100

All video sequences except those for pointing and standing are classified without error. Standing is sometimes confused with pointing whereas pointing is occasionally confused with standing. Both of these action categories are essentially static poses and are sufficiently similar to even cause confusion in human observers on account of the low resolution of the dataset.

Table 4: LOOCV CCR comparison of the proposed method with our previous method: UT-Tower human action dataset, 8-frame segments with 4-frame overlap.

Method	**Proposed**	Guo *et al.* [7]
SEG-CCR	**96.15%**	93.53%
SEQ-CCR	**97.22%**	96.30%

The proposed method is also time-efficient and easy to implement. Our experimental platform was Intel Centrino (CPU: T7500 2.2GHz + Memory: 2GB) with Matlab 7.6. The computation of 13-dimensional feature vectors and the calculation of log-covariance matrices can be efficiently implemented on this platform, costing together about 4.3 seconds per silhouette sequence with spatial resolution of 111×81 and length of 89 frames. This method is also memory efficient since the training and query sets essentially store 13×13 log-covariance matrices instead of video data. Given a query sequence with 20 query segments and a training set with 1239 training segments, it takes about 4.5 seconds to classify all query segments (solving 20 l^1-norm minimization problems), i.e., about 0.22 seconds per query segment.

7 Concluding remarks

In this paper, we proposed a new approach to action recognition in video based on sparse linear representations of log-covariance matrices of silhouette shape features. The proposed method is motivated by Wright et al.'s work [13] that has been successfully applied in the context of face recognition. The salient characteristic of our method is the fact that it uses log-covariance matrices to represent actions in a vector space.

Our experimental results on the Weizmann dataset indicate that the classification performance of the proposed method is similar to that of recent successful methods, such as Gorelick's method [6] and our previous method [7]. At the same time, its computational complexity is relatively low in both feature extraction, on account of feature simplicity, and classification, owing to efficiencies in solving the l^1 minimization. On the challenging UT-Tower dataset, the proposed method outperforms our previous approach based on the same features and NN classification.

Acknowledgment

The authors would like to thank Prof. Pierre Moulin from the ECE Department at UIUC for suggesting the application of the recently-developed sparse linear representation framework for dictionary-based classification to log-covariance matrices as an alternative to NN-classification.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans. Pattern Anal. Machine Intell. 32(2), 288–303 (Feb 2010)
2. Arsigny, V., Pennec, P., Ayache, X.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic resonance in medicine 56(2), 411–421 (2006)
3. Chen, C.C., Ryoo, M.S., Aggarwal, J.K.: UT-Tower Dataset: Aerial View Activity Classification Challenge. http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html (2010)
4. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE Int'l Workshop VS-PETS (2005)
5. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Comm. Pure Appl. Math 59, 797–829 (2004)
6. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Machine Intell. 29(12), 2247–2253 (Dec 2007)
7. Guo, K., Ishwar, P., Konrad, J.: Action recognition from video by covariance matching of silhouette tunnels. In: Proc. Brazilian Symp. on Computer Graphics and Image Proc. (Oct 2009)
8. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learing realistic human actions from movies. In: Proc. IEEE Conf. Computer Vision Pattern Recognition (Jun 2008)
9. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In: Intern. J. Comput. Vis. (Mar 2008)
10. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proc. Int. Conf. Pattern Recognition (Jun 2004)
11. Seo, H.J., Milanfar, P.: Action recognition from one example. IEEE Trans. Pattern Anal. Machine Intell. submitted
12. Starner, T., Pentland, A.: Visual recognition of american sign language using hidden markov models. In: IEEE Int. Conf. on Automatic Face and Gesture Recognition (1995)
13. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Machine Intell. 31(2), 210–227 (Feb 2009)
14. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time sequential images using hidden markov model. In: Proc. IEEE Conf. Computer Vision Pattern Recognition (Jun 1992)