# Action Recognition from One Example

Hae Jong Seo, *Student Member, IEEE,*
and Peyman Milanfar, *Senior Member, IEEE*

**Abstract**

We present a novel action recognition method based on space-time locally adaptive regression kernels and the matrix cosine similarity measure. The proposed method uses a single example of an action to find similar matches. It does not require prior knowledge about actions; foreground/background segmentation, or any motion estimation or tracking. Our method is based on the computation of novel space-time descriptors from a query video, which measure the likeness of a voxel to its surroundings. Salient features are extracted from said descriptors and compared against analogous features from the target video. This comparison is done using a matrix generalization of the cosine similarity measure. The algorithm yields a scalar resemblance volume, with each voxel indicating the likelihood of similarity between the query video and all cubes in the target video. Using nonparametric significance tests and non-maxima suppression, we detect the presence and location of actions similar to the query video. High performance is demonstrated on challenging sets of action data containing fast motions, varied contexts, and even when multiple complex actions occur simultaneously within the field of view. Further experiments on the Weizmann and KTH datasets demonstrate state-of-the-art performance in action categorization, despite the use of only a single example.

**Index Terms**

Action Recognition, Space-time descriptor, correlation and regression analysis

## I. Introduction

A huge number of videos (e.g., BBC[1],Youtube[2]) are available online today and the number is rapidly growing. Human actions constitute one of the most important parts in movies, TV shows, and consumer-generated videos. Analysis of human actions in videos is considered a

[1]http://www.bbcmotiongallery.com

[2]http://www.youtube.com

very important problem in computer vision because of such applications as human-computer interaction, content-based video retrieval, visual surveillance, analysis of sports events and more. The term "action" refers to a simple motion pattern as performed by a single subject, and in general lasts only for a short period of time, namely just a few seconds. *Action* is often distinguished from *activity* in the sense that action is an individual atomic unit of activity. In particular, human action refers to physical body motion. Recognizing human actions from video is a very challenging problem due to the fact that physical body motion can look very different depending on the context: for instance, similar actions with different clothes, or in different illumination and background can result in a large appearance variation; or, the same action performed by two different people may look quite dissimilar in many ways.

### *A. Problem Specification*

We present a novel approach to the problem of human action recognition as a video-to-video matching problem. Here, recognition is generally divided into two parts: category classification and detection/ localization. The goal of action classification is to classify a given action query into one of several pre-specified categories (for instance, 6 categories from KTH action dataset [3]: boxing, hand clapping, hand waving, jogging, running, and walking). Meanwhile, action detection is meant to separate an action of interest from the background in a target video (for instance, spatiotemporal localization of a walking person). This paper tackles both action detection and category classification problems simultaneously by searching for an action of interest within other "target" videos with only a *single* "query" video. In order to avoid the disadvantages of learning-based methods which require a large number of training examples, we focus on a sophisticated feature representation with an efficient and reliable similarity measure which also allows us to avoid the difficult problem of explicit motion estimation.

In general, the target video may contain actions similar to the query, but these will typically appear in completely different context (See Fig. 1 Left.) Examples of such differences can range from rather simple optical or geometric differences (such as different clothes, lighting, action speed and scale changes); to more complex inherent structural differences such as for instance a hand-drawn action video clip (e.g., animation) rather than a real human action.

### B. Related work

Over the last two decades, many studies have attempted to tackle this problem and made impressive progress. Approaches can be categorized on the basis of *action representation*; namely, appearance-based representation [4], [5], [6], [7], shape-based representation [8], [9], [10], [2], optical-flow-based representation [11], [12], [13], [14], interest-point-based representation [3], [15], [16], [17], [18], [19], and volume-based representation [20], [21], [22], [1], [23], [24], [25]. We refer the interested reader to [26], [27], [28] and references therein for a good summary.

As examples of the interest-point-based approach which has gained a lot of interest, Niebles et al. [16], [15] considered videos as spatiotemporal bag-of-words by extracting space-time interest points and clustering the features, and then used a probabilistic Latent Semantic Analysis (pLSA) model to localize and categorize human actions. Yuan et al. [29] also used spatiotemporal features as proposed by [17]. They extended the naive Bayes nearest neighbor classifier [30], which was developed for object recognition, to action recognition. By modifying the efficient searching method based on branch-and-bound [31] for the 3-D case, they provided a very fast action detection method. However, the performance of these methods can degrade due to 1) the lack of enough training samples; 2) misdetections and occlusions of the interest points since they ignore global space-time information.

Shechtman and Irani [1] recently employed a three dimensional correlation scheme for action detection. They focused on subvolume matching in order to find similar motion between the two space-time volumes, which can be computationally heavy. Ke et al. [23] presented an approach which uses boosting on 3-D Haar-type features inspired by similar features in 2-D object detection [32]. While these features are very efficient to compute, many examples are required to train an action detector in order to achieve good performance. They further proposed a part-based shape and flow matching framework [33] and showed good action detection performance in crowded videos. Recently, Kim et al. [24] generalized canonical correlation analysis to tensors and showed very good accuracy on the KTH action dataset, but their method requires a manual alignment process for camera motion compensation. Ning et al. [25] proposed a system to search for human actions using a coarse-to-fine approach with a five-layer hierarchical space-time model. These volumetric methods do not require background subtraction, motion estimation, or complex models of body configuration and kinematics. They tolerate variations in appearance,

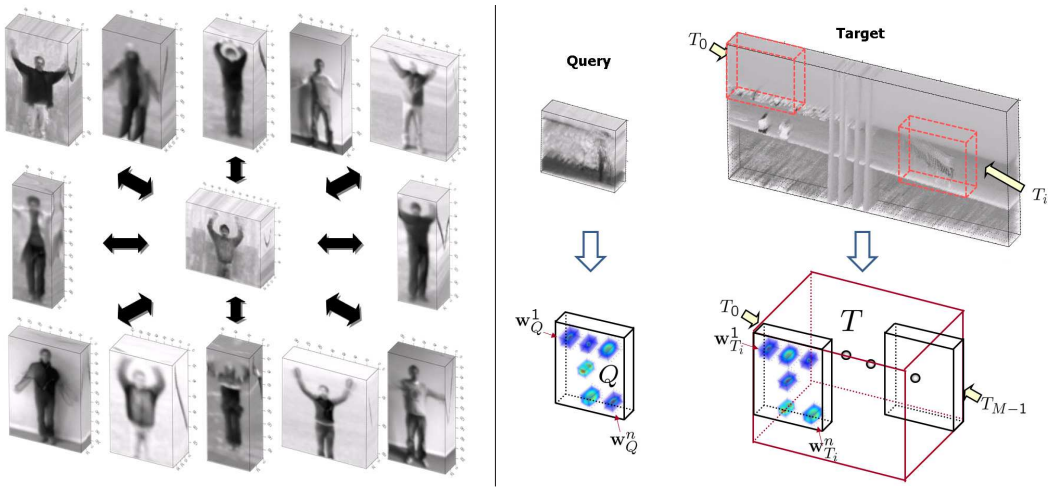


Fig. 1. Left: A hand-waving action and possibly similar actions, Right: Action detection problem (a) Given a query video $Q$, we wish to detect/localize actions of interest in a target video $T$. $T$ is divided into a set of overlapping cubes (b) space-time local steering kernels (3-D LSKs) capture the geometric structure of underlying data.

scale, rotation, and movement to some extent.

Methods such as those in [33], [1], [25], [34] which aim at recognizing actions based solely on one query are very useful for applications such as video retrieval from the web (e.g., viewdle[3],videosurf[4]). In these methods, a single query video is provided by users and every gallery video in the database is compared with the given query, posing a video-to-video matching problem.

### C. Overview of the Proposed Approach

In this paper, our contributions to the action recognition task are mainly two-fold. First, we propose a novel feature representation that is derived from space-time local (steering) regression kernels (3-D LSKs) which capture the underlying structure of the data quite well, even in the presence of significant distortions and data uncertainty. Second, we generalize a training-free nonparametric detection scheme to 3-D, which we developed earlier for 2-D object detection [35]. We report state-of-the art performance on action category classification by using the resulting nearest neighbor classifier. In order to achieve better classification performance, we apply space-time saliency detection [36], [37] to larger videos in order to automatically crop to a short action

[3]http://www.viewdle.com

[4]http://www.videosurf.com

clip.

We propose to use 3-D LSKs for the problems of detection/localization of actions of interest between a query video and a target video. The key idea behind 3-D LSKs is to robustly obtain local space-time geometric structures by analyzing the radiometric (voxel value) differences based on estimated space-time gradients, and use this structure information to determine the shape and size of a canonical kernel (descriptor). The motivation to use these 3-D LSKs is the earlier successful work on adaptive kernel regression for image denoising, interpolation [38], deblurring [39], and superresolution [40]. The 3-D LSKs implicitly contain information about the local motion of the voxels across time, thus requiring no explicit motion estimation.

Referring to Fig. 2, by denoting the target video ($T$), and the query video ($Q$), we compute a dense set of 3-D LSKs from each. These densely computed descriptors are highly informative, but taken together tend to be over-complete (redundant). Therefore, we derive features by applying dimensionality reduction (namely PCA) to these resulting arrays, in order to retain the most salient characteristics of the 3-D LSKs. The feature collections from $Q$ and $T_i$ (a chunk of the target which is the same size as the query; See Fig. 1 Right) form feature volumes $\mathbf{F}_Q$ and $\mathbf{F}_{T_i}$. We compare the feature volumes $\mathbf{F}_{T_i}$ and $\mathbf{F}_Q$ from the $i^{th}$ cube of $T$ and $Q$ to look for matches. Inspired in part by many studies [41], [42], [43], [44], [45], [46] which took advantage of cosine similarity over the conventional Euclidean distance, we employ *Matrix Cosine Similarity* (MCS) as a similarity measure which generalizes the notion of cosine similarity between two vectors [47], [48], [49]. The optimality properties of this approach are described in [35] within a naive Bayes framework.

In general, it is assumed that the query video is smaller than target video. However, this is not true in practice and a query video may indeed include a complex background which deteriorates recognition accuracy. In order to deal with this problem, it is necessary to have a procedure which automatically segments from the query video a small cube that only contains a valid human action. For this, we employ space-time saliency detection [36]. This idea not only allows us to extend the proposed detection framework to action category classification, but also improve both detection and classification accuracy by automatically removing irrelevant background from the query video. Fig. 2 shows an overview of our proposed framework for action detection and category classification.

[50] introduced a space-time local self-similarity descriptor for action detection and showed

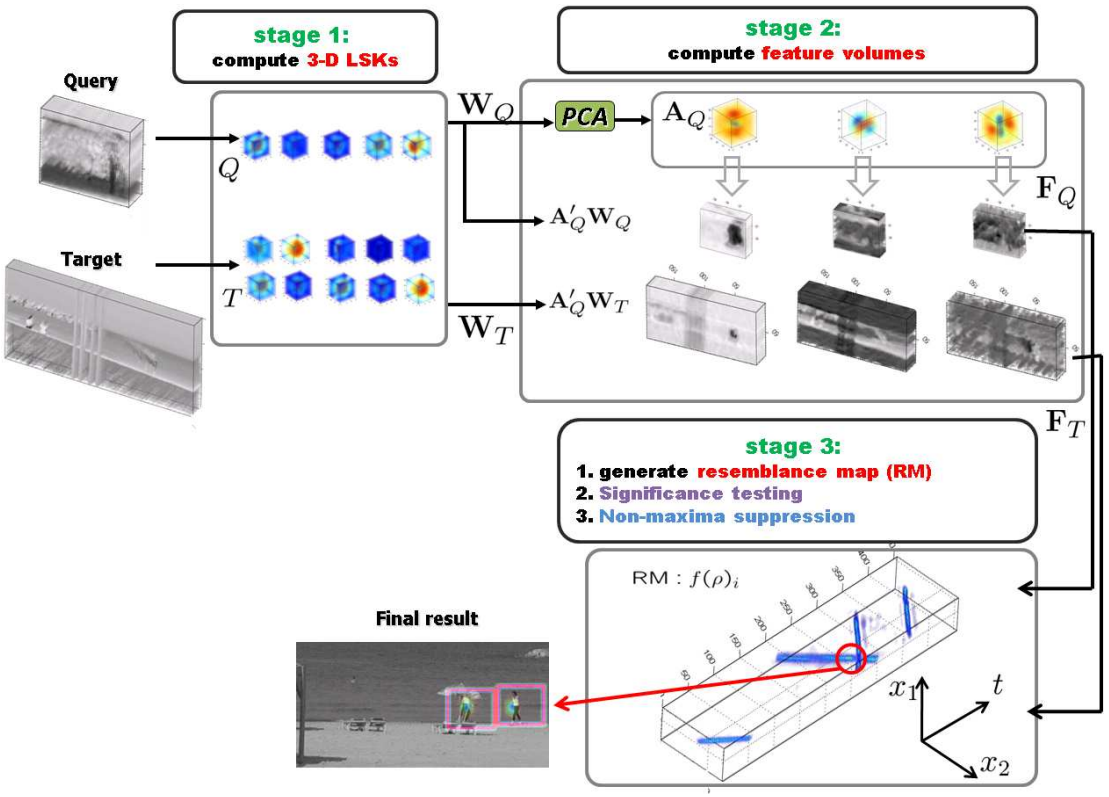


Fig. 2. System overview of action detection framework (There are broadly three stages.)

performance improvement over related earlier previous approach as [1]. It is worth mentioning that this (independently derived) local space-time self-similarity descriptor is a special case of 3-D LSK and is also related to a number of other local data adaptive metrics such as Optimal Space-Time Adaptation (OSTA) [51] and Non-Local Means (NLM) [52] which have been used very successfully for video restoration in the image processing community.

As a related action representation, Ali and Shah [13] very recently proposed kinematic features (divergence, vorticity, symmetric and anti-symmetric optical flow and so forth) based on optical flows. By applying PCA to these features, they extracted dominant kinematic features and used them for action recognition along with the multiple instance learning approach [53]. Our action representation is somewhat similar to theirs in the sense that we both use PCA to extract feature sets, but their method requires learning, and is focused on estimated motion while our method does not involve learning, and uses 3-D LSKs which extract both shape and flow information implicitly, and at the same time.

The proposed action detection method is also distinguished from our earlier 2-D work in [35] proposed for object detection, in the following respects; 1) action detection addressed in this paper is considered to be more challenging than static (2-D) object detection due to additional problems such as variations in individual motion and camera motion, 2) we use space-time local steering kernels which capture both *spatial* and *temporal* geometric structure, 3) while [35] assumed that a query image is always smaller than a target and only contains an object of interest, we relax this assumption to deal with more realistic scenarios by incorporating space-time saliency detection [36], and 4) while [35] focused on detection tasks, in this paper, we further achieved state-of-the art action classification performance as well as high detection accuracy. Therefore, a nontrivial extension of 2-D framework to 3-D for action recognition, and the careful examination of the proposed approach on challenging action datasets are the aims of this paper.

Before we begin a more detailed description, we highlight some key aspects of the proposed framework.

- We propose a novel feature representation derived from densely computed 3-D LSKs. Since the calculation of 3-D LSKs is stable in the presence of uncertainty in the data [38], our approach is robust even in the presence of noise. In addition, normalized 3-D LSKs provide a certain invariance to illumination changes (see Fig. 6.)
- As opposed to [1] which filtered out "non-informative" descriptors in order to reduce the time complexity, we automatically obtain the most salient feature volumes by applying Principal Components Analysis (PCA) to a collection of 3-D LSKs. The proposed method is feasible in practice because the dimension of features after PCA is significantly reduced (e.g., from say $3 \times 3 \times 7 = 64$, to 3 or 4), even though the descriptors are densely computed.
- The proposed method is tolerant to modest deformations (i.e., $\pm 20\%$ scale change (in space-time), $\pm 15$ degree rotation change) of the query and can detect multiple actions that occur simultaneously in the field of view using multiple queries.
- From a practical standpoint, it is important to note that the proposed framework operates using a single example of an action of interest to find similar matches; does not require any prior knowledge (learning) about actions being sought; and does not require any pre-processing step or segmentation of the target video. Since we do not require background subtraction, the proposed method can work with cluttered scenes with dynamic backgrounds.

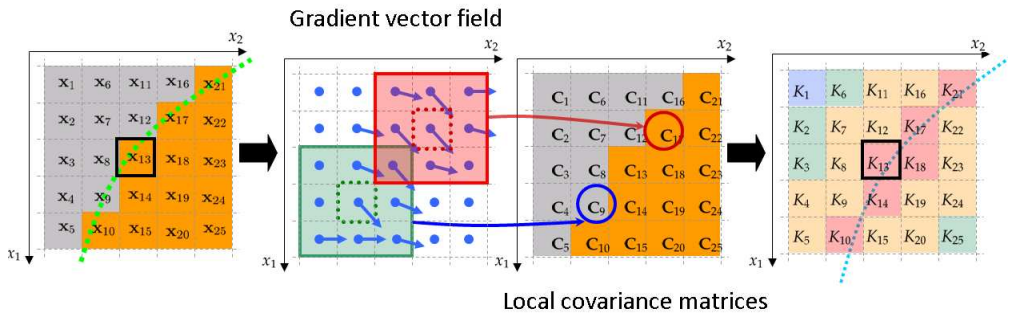


Fig. 3. Graphical description of how LSK values centered at pixel of interest $\mathbf{x}_{13}$ are computed in an edge region. Note that each pixel location has its own $\mathbf{C} \in \mathbb{R}^{2\times 2}$ computed from gradient vector field within a local window. In $K$ values, red means higher values (higher similarity).

## II. Technical Details

As outlined in the previous section, our approach to detect actions consists broadly of three stages (see Fig 2.) Below, we describe each of these steps in detail. In order to make the concepts more clear, we first briefly describe the local steering kernels in 2-D. For extensive detail on this subject, we refer the reader to [38], [35].

### A. Local Steering Kernel as a descriptor

*1) Local Steering Kernel in 2-D (LSK):* The key idea behind LSK is to robustly obtain the local structure of images by analyzing the radiometric (pixel value) differences based on estimated gradients, and to use this structure information to determine the shape and size of a canonical kernel. The local steering kernel is defined as follows:

$$K(\mathbf{x}_l - \mathbf{x}_i) = \frac{\sqrt{\det(\mathbf{C}_l)}}{h^2} \exp\left\{\frac{(\mathbf{x}_l - \mathbf{x}_i)^T \mathbf{C}_l (\mathbf{x}_l - \mathbf{x}i)}{-2h^2}\right\}, \quad l = 1, \cdots, P, \tag{1}$$

where $\mathbf{x}_i = [x_1, x_2]_i^T$ is a pixel of interest, $\mathbf{x}_l = [x_1, x_2]_l^T$ are a local neighboring pixels, $h$ is a global smoothing parameter[5], $P$ is the total number of samples in a local analysis window around a sample position at $\mathbf{x}_i$, and the matrix $\mathbf{C}_l \in \mathbb{R}^{(2\times 2)}$ is a covariance matrix estimated from a collection of first derivatives along spatial axes. More specifically, the covariance matrix

[5] $h$ is set to 2 for all experiments.

$\mathbf{C}_l$ can be first naively estimated as $\mathbf{J}_l^T\mathbf{J}_l$ with

$$\mathbf{J}_l = \begin{bmatrix} z_{x_1}(\mathbf{x}_1), & z_{x_2}(\mathbf{x}_1) \\ \vdots & \vdots \\ z_{x_1}(\mathbf{x}_P), & z_{x_2}(\mathbf{x}_P) \end{bmatrix},$$

where $z_{x_1}(\cdot)$ and $z_{x_2}(\cdot)$ are the first derivatives along $x_1-$, and $x_2-$ axes. For the sake of robustness, we compute a more stable estimate of $\mathbf{C}_l$ by invoking the singular value decomposition (SVD) of $\mathbf{J}_l$ with regularization as [38], [35]

$$\mathbf{C}_l = \gamma \sum_{q=1}^{2} a_q^2 \mathbf{v}_q \mathbf{v}_q^T \in \mathbb{R}^{(2\times 2)}, \tag{2}$$

with

$$a_1 = \frac{s_1 + \lambda'}{s_2 + \lambda'}, \quad a_2 = \frac{s_2 + \lambda'}{s_1 + \lambda'}, \quad \gamma = \Big(\frac{s_1 s_2 + \lambda''}{P}\Big)^{\alpha}, \tag{3}$$

where $\lambda'$ and $\lambda''$ are parameters[6] that dampen the noise effect and keep the denominators of $a_q$'s from being zero, and $\alpha$ is a parameter [7] that restricts $\gamma$. The singular values $(s_1, s_2)$ and the singular vectors $(\mathbf{v}_1, \mathbf{v}_2)$ are given by the compact SVD of $\mathbf{J}_l = \mathbf{U}_l\mathbf{S}_l\mathbf{V}_l^T = \mathbf{U}_l \mathrm{diag}[s_1, s_2]_l[\mathbf{v}_1, \mathbf{v}_2, ]_l^T$. Fig. 3 illustrates how the covariance matrices and respective LSK values are computed.

Indeed, the covariance matrix $\mathbf{C}_l$ modifies the shape and size of the local kernel in a way which robustly encodes the local geometric structures. The shape of the LSK's is not simply a Gaussian, despite the simple definition in (1) above. It is important to note that this is because for each pixel $\mathbf{x}_l$ in the vicinity of $\mathbf{x}_i$, a different matrix $\mathbf{C}_l$ is used, therefore leading to a far more complex and rich set of possible shapes for the resulting LSKs. The same idea is valid in 3-D as well, as we describe below.

*2) Space-Time Local Steering Kernel (3-D LSK):* Now, we introduce the time axis to the data model so that $\mathbf{x}_l = [x_1, x_2, t]_l^T$: $x_1$ and $x_2$ are the spatial coordinates, and $t$ is the temporal coordinate. Similar to the 2-D case, the covariance matrix $\mathbf{C}_l$ can be naively estimated as $\mathbf{J}_l^T\mathbf{J}_l$ with

$$\mathbf{J}_l = \begin{bmatrix} z_{x_1}(\mathbf{x}_1), & z_{x_2}(\mathbf{x}_1), & z_t(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ z_{x_1}(\mathbf{x}_P), & z_{x_2}(\mathbf{x}_P), & z_t(\mathbf{x}_P) \end{bmatrix},$$

[6] $\lambda'$ and $\lambda''$ are set to 1 and $10^{-8}$ respectively, and they are fixed for all experiments.

[7] $\alpha$ is set to 0.29 and fixed for all experiments.

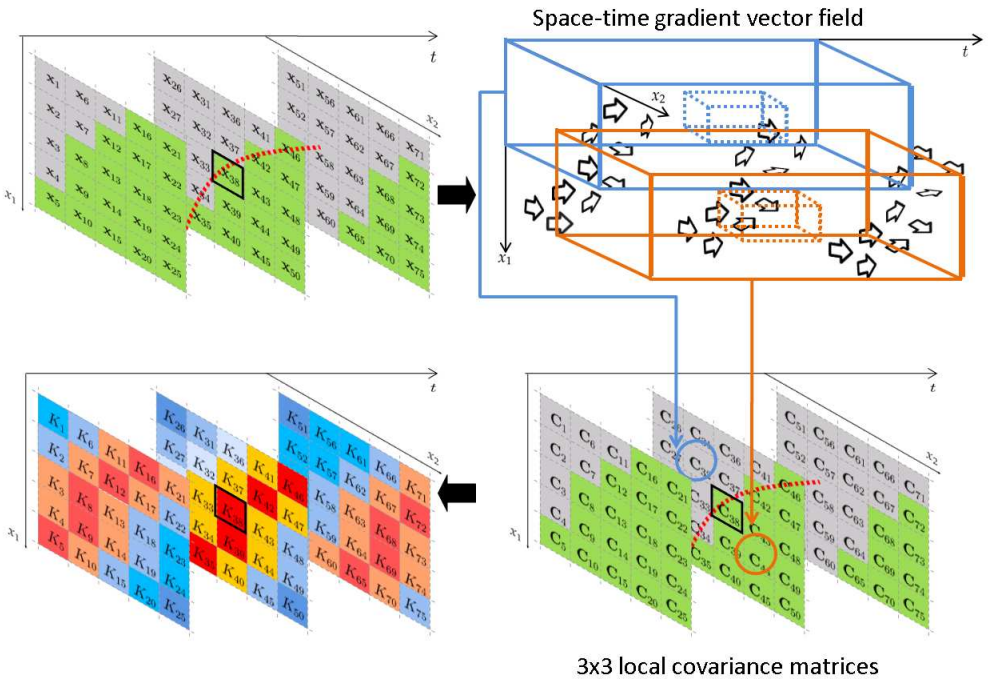


Fig. 4. Graphical description of how 3-D LSK values centered at voxel of interest $\mathbf{x}_{38}$ are computed in a space-time edge region. Note that each voxel location has its own $\mathbf{C} \in \mathbb{R}^{3\times 3}$ computed from space-time gradient vector field within a local space-time window.

where $z_{x_1}(\cdot), z_{x_2}(\cdot)$, and $z_t(\cdot)$ are the first derivatives along $x_1-, x_2-$, and $t-$ axes, and $P$ is the total number of samples in a *space-time* local analysis window (or cube) around a sample position at $\mathbf{x}_i$. Again, $\mathbf{C}_l$ is estimated by invoking the singular value decomposition (SVD) of $\mathbf{J}_l$ with regularization as [40]:

$$\mathbf{C}_l = \gamma \sum_{q=1}^{3} a_q^2 \mathbf{v}_q \mathbf{v}_q^T \in \mathbb{R}^{(3\times 3)}, \tag{4}$$

with

$$a_1 = \frac{s_1 + \lambda'}{\sqrt{s_2 s_3} + \lambda'}, \quad a_2 = \frac{s_2 + \lambda'}{\sqrt{s_1 s_3} + \lambda'}, \quad a_3 = \frac{s_3 + \lambda'}{\sqrt{s_1 s_2} + \lambda'}, \quad \gamma = \left(\frac{s_1 s_2 s_3 + \lambda''}{P}\right)^{\alpha}, \tag{5}$$

where $\lambda'$ and $\lambda''$ are parameters[8] that dampen the noise effect and restrict $\gamma$ and the denominators of $a_q$'s from being zero. As mentioned earlier, the singular values ($s_1, s_2$, and $s_3$) and the singular vectors ($\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$) are given by the compact SVD of $\mathbf{J}_l = \mathbf{U}_l \mathbf{S}_l \mathbf{V}_l^T = \mathbf{U}_l \mathrm{diag}[s_1, s_2, s_3]_l [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]_l^T$.

[8] $\lambda', \lambda'', \alpha$, and $h$ are set to the same values as 2-D LSKs and fixed for all experiments.

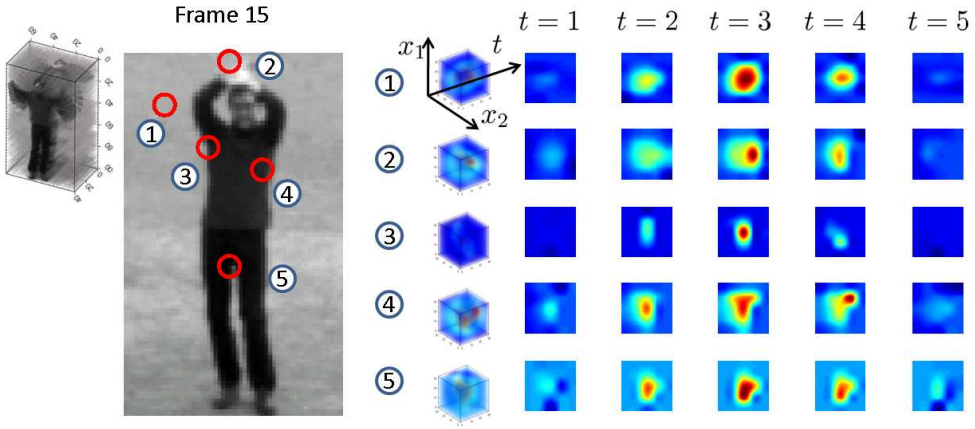


Fig. 5. Examples of 3-D LSKs capturing 3-D local underlying geometric structure in various regions. In order to compute 3-D LSKs, 5 frames (frame 13 to frame 17) were used.

Fig. 4 illustrates how 3-D LSKs are computed in a space-time region. In the 3-D case, orientation information captured in 3-D LSK contains the motion information implicitly [40]. It is worth noting that a significant strength of using this implicit framework (as opposed to the direct use of estimated motion vectors) is the flexibility it provides in terms of smoothly and adaptively changing descriptors. This flexibility allows the accommodation of even complex motions, so long as their magnitudes are not excessively large.

Fig. 5 shows examples of 3-D local steering kernels capturing 3-D local underlying geometric structure in various space-time regions. As can be seen in (1), the values of the kernel $K$ are based on the covariance matrices $\mathbf{C}_l$ along with their space-time locations $\mathbf{x}_l$. Intuitively, $\mathbf{C}_l$'s computed from the local analysis window are similar to one another in the motion-free region (see Fig. 5 [1]). On the other hand, in the region where motion exists (see Fig. 5 [2,3,4,5]), the kernel size and shape depend on both $\mathbf{C}_l$ and its space-time location $\mathbf{x}_l$ in the local space-time window. Thus, high values in the kernel are yielded along the space-time edge region whereas the rest of kernel values are near zero.

In what follows, at a position $\mathbf{x}_i$, we will essentially be using (a normalized version of) the function. $K(\mathbf{x}_l - \mathbf{x}_i)$ as descriptors, representing a video's inherent local space-time geometry. To be more specific, the 3-D LSK function $K(\mathbf{x}_l - \mathbf{x}_i)$ is densely calculated and normalized as follows

$$W_I^i = \frac{K(\mathbf{x}_l - \mathbf{x}_i)}{\sum_{l=1}^{P} K(\mathbf{x}_l - \mathbf{x}_i)}, \tag{6}$$

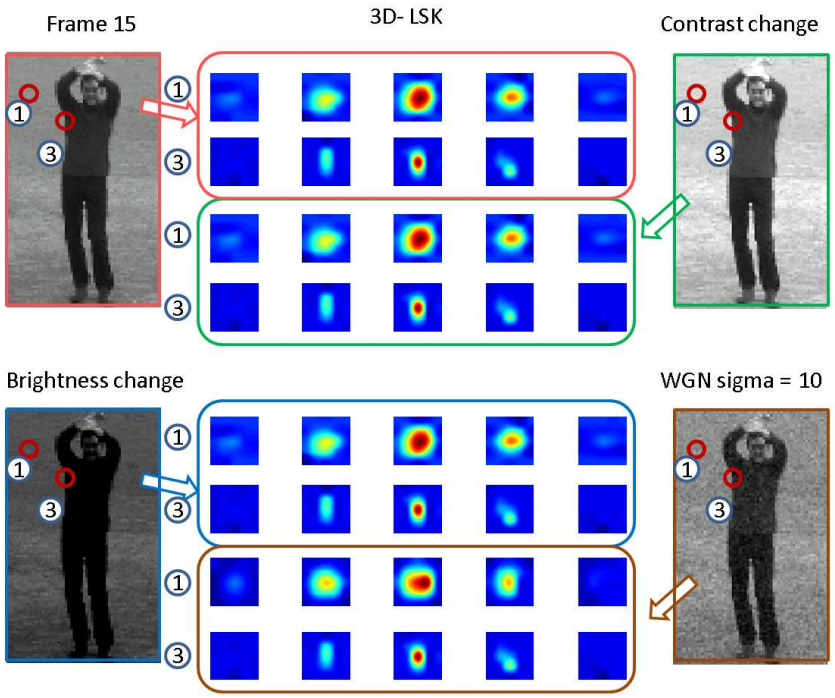


Fig. 6. Invariance and robustness of 3-D LSK weights $W(\mathbf{x}_l - \mathbf{x})$ in various challenging conditions. Note that WGN means White Gaussian Noise.

where $I$ can be $Q$ or $T$ for query or target, respectively[9].

Normalization of this kernel function yields invariance to brightness change and robustness to contrast change as shown in Fig. 6 (as was similarly shown for 2-D LSKs in [35].)

Fig 7 shows that a collection of 3-D LSKs reveals global space-time geometric information. It is interesting to note that 3-D LSKs[10] seem related to "spatiotemporal descriptors based on 3-D gradients" introduced in [56]. However, our method is quite different in that their descriptor is based on histograms of oriented 3D spatiotemporal gradients while our LSKs are based on the similarity between a center voxel and surrounding voxels in a space-time neighborhood measured with the help of gradients. Our descriptors capture higher-level contextual information than the histogram of space-time gradients. Furthermore, we extract salient characteristics of 3-D LSKs by further applying Principal Component Analysis (PCA) as described in the following section.

[9]Note that videos here are gray scale. The case of color is worth treating independently and is discussed in [35]

[10]HoG [54] and HoF [55] are also related to our 2-D LSKs ($x_1 - x_2$ axes) and 2-D LSKs (either $x_1 - t$ axes or $x_2 - t$ axes).

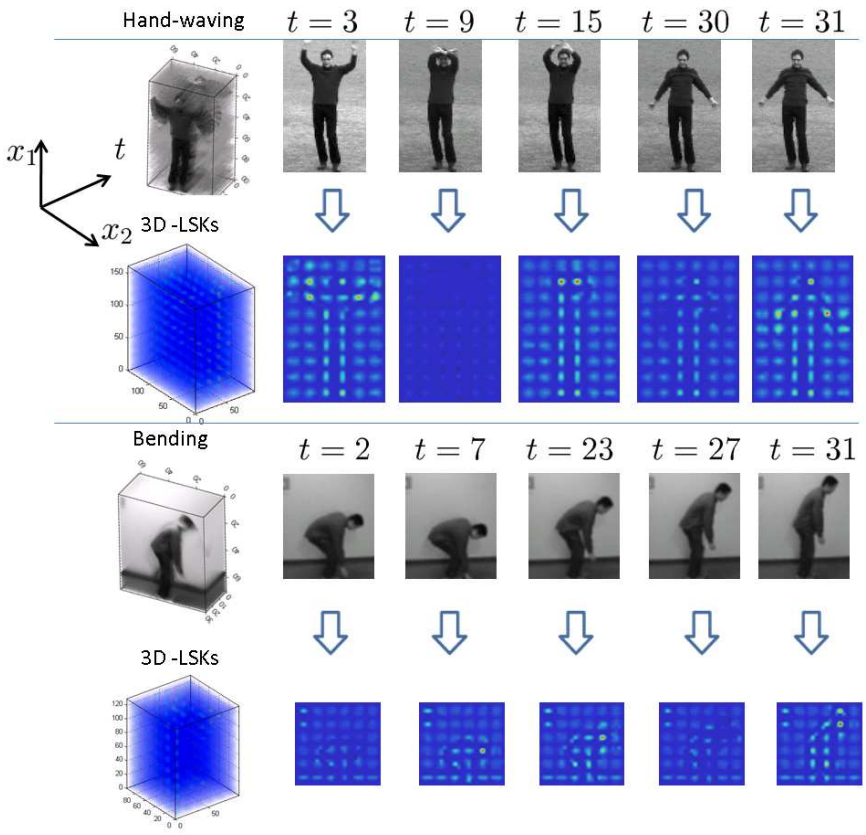


Fig. 7. 3D-LSKs computed from two actions such as hand-waving and bending are shown. For graphical description, we only computed 3-D LSKs at non-overlapping $5 \times 5 \times 5$ cubes, even though we compute 3-D LSKs densely in practice.

### *B. Feature representation*

It has been shown in [35] that the normalized LSKs in 2-D follow a power-law (i.e., a long-tail) distribution. That is to say, the features are scattered out in a high dimensional feature space, and thus there basically exists no dense cluster in the descriptor space. The same principle applies to 3-D LSK. In order to illustrate and verify that the normalized 3-D LSKs also satisfy this property, we computed an empirical bin density ($100$ bins) of the normalized 3-D LSKs (using a total of $50,000$ 3-D LSKs) computed from $90$ videos of the Weizmann action dataset ([2]) using the K-means clustering method (See Fig. 8.) The utility of this observation becomes clear in the next paragraphs.

In the previous section, we computed a dense set of 3-D LSKs from $Q$ and $T$. These densely computed descriptors are highly informative, but taken together tend to be over-complete (redundant). Therefore, we derive features by applying dimensionality reduction (namely PCA)

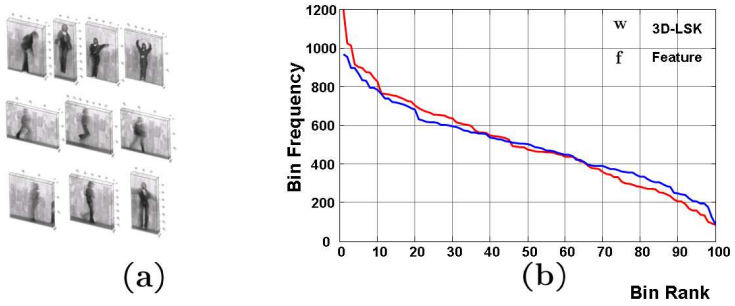


Fig. 8. (a) Some example video sequences (Weizman dataset) where 3-D LSKs were computed. (b) Plots of the bin density of 3-D LSKs and their corresponding low-dimensional features.

to these resulting arrays, in order to retain only the salient characteristics of the 3-D LSKs. As also observed in [57], [30], an ensemble of local features with even little discriminative power can together offer significant discriminative power. However, both quantization and informative feature selection on a long-tail distribution can lead to a precipitous drop in performance. Therefore, instead of any quantization and informative feature selection, we focus on reducing the dimension of 3-D LSKs using PCA. [11]

This idea results in a new feature representation with a moderate dimension which inherits the desirable discriminative attributes of 3-D LSK. The distribution of the resulting features sitting on the low dimensional manifold also tends to follow a power-law distribution as shown in Fig. 8 (b) and this allows us to the use *Matrix Cosine Similarity* (MCS) measure which will be illustrated in Section II-C. The optimality property and justification of MCS can be found in [35].

In order to organize $W_Q$ and $W_T$, which are densely computed from $Q$ and $T$, let $\mathbf{W}_Q$,$\mathbf{W}_T$ be matrices whose columns are vectors $\mathbf{w}_Q$,$\mathbf{w}_T$, which are column-stacked (rasterized) versions of $W_Q$,$W_T$ respectively:

$$\begin{aligned} \mathbf{W}_Q &= [\mathbf{w}_Q^1, \cdots, \mathbf{w}_Q^n] \in \mathbb{R}^{P \times n}, \\ \mathbf{W}_T &= [\mathbf{w}_T^1, \cdots, \mathbf{w}_T^{n_T}] \in \mathbb{R}^{P \times n_T}, \end{aligned} \tag{7}$$

where $n$ and $n_T$ are the number of cubes where 3-D LSKs are computed in the query $Q$ and

[11] Ali and Shah [13] also pointed out that interest point descriptor-based action recognition methods have a limitation in that useful pieces of global information may be lost.

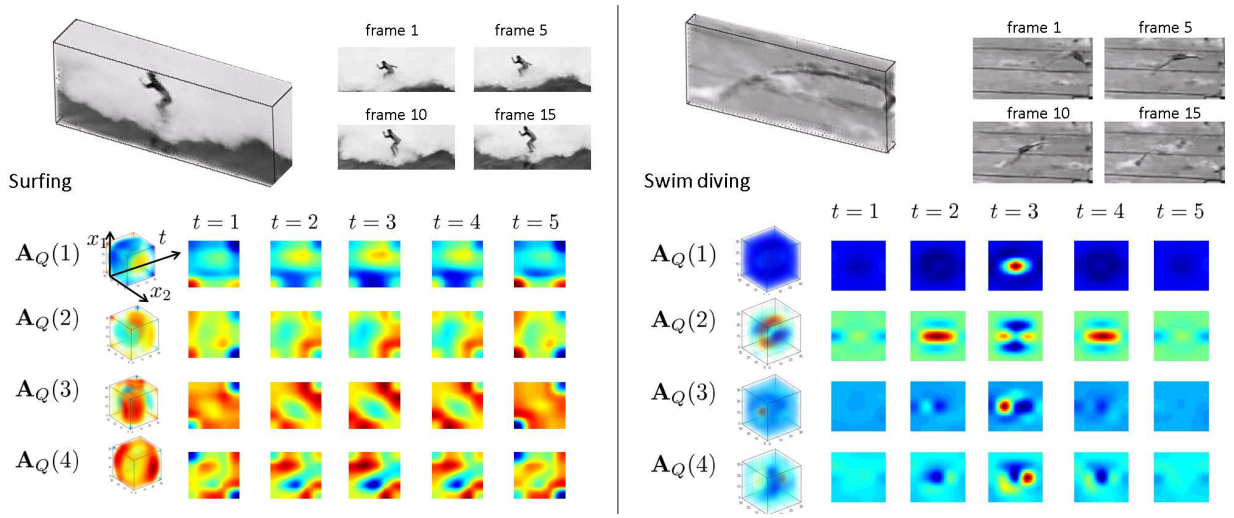


Fig. 9. Examples of top 4 principal components in $\mathbf{A}_Q$ for various actions such as surfing and diving. Note that these eigenvectors reveal geometric characteristic of queries in both space and time domain, and thus they are totally different from linear 3-D Gabor filters.

the target $T$ respectively.

As described in Fig. 2, the next step is to apply PCA to $\mathbf{W}_Q$ and retain the first (largest) $d$ principal components[12] which form the columns of a matrix $\mathbf{A}_Q \in \mathbb{R}^{P\times d}$. Next, the lower dimensional features are computed by projecting $\mathbf{W}_Q$ and $\mathbf{W}_T$ onto $\mathbf{A}_Q$:

$$\mathbf{F}_Q = [\mathbf{f}_Q^1, \cdots, \mathbf{f}_Q^n] = \mathbf{A}_Q^T \mathbf{W}_Q \in \mathbb{R}^{d\times n}, \quad \mathbf{F}_T = [\mathbf{f}_T^1, \cdots, \mathbf{f}_T^{n_T}] = \mathbf{A}_Q^T \mathbf{W}_T \in \mathbb{R}^{d\times n_T}. \tag{8}$$

Figs. 9 and 10 illustrate that the principal components $\mathbf{A}_Q$ learned from different actions such as surfing, diving, hand waving, and bending actions are quite distinct from one another. Fig. 11 shows what the features $\mathbf{F}_Q, \mathbf{F}_T$ look like for a walking action and a jumping action. In order to show where actions appear, we drew red ovals around each action in the target video. These examples illustrate (as quantified later in the paper) that the derived feature volumes have a good discriminative power even though we do not involve any learning over a set of training examples.

It is worth noting that a similar approach was also taken by [58] where PCA was applied to interest point descriptors such as SIFT, leading to enhanced performance. Very recently, [13]

[12]Typically, $d$ is selected to be a small integer such as 3 or 4 so that 80 to 90% of the information in the LSKs would be retained. (i.e., $\frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{P}\lambda_i} \geq 0.8$ (to 0.9) where $\lambda_i$ are the eigenvalues.)

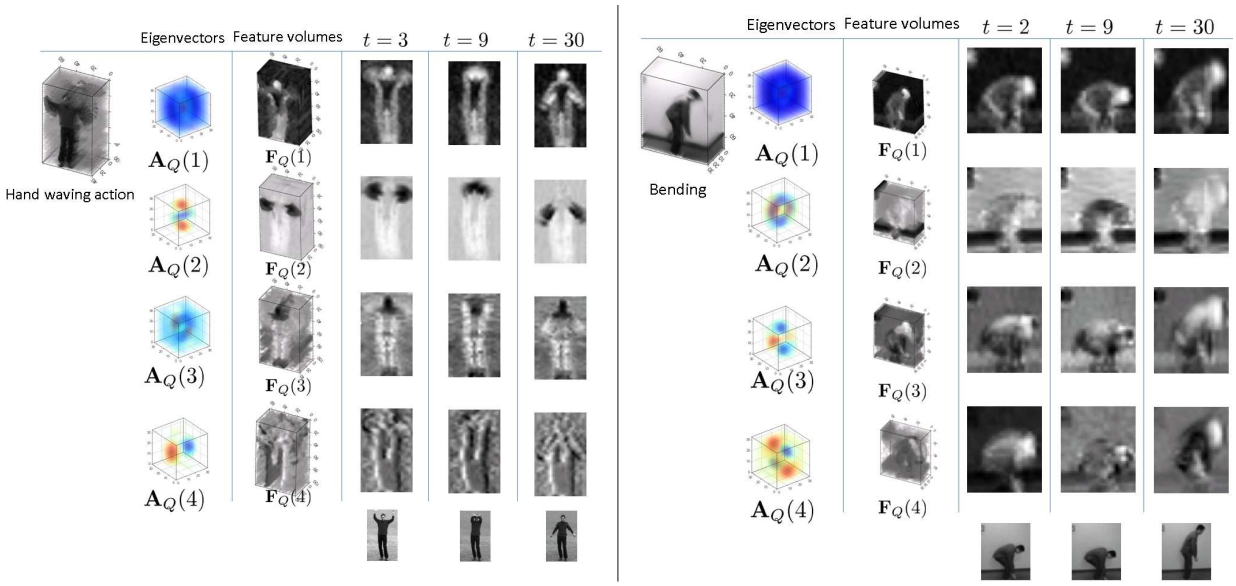


Fig. 10. $\mathbf{A}'_Q s$ and $\mathbf{F}'_Q s$ of hand waving action and bending action.

proposed a set of kinematic features that extract different aspects of motion dynamics present in the optical flow. They obtained bags of kinematic modes for action recognition by applying PCA to a set of kinematic features. We differentiate our proposed method from [13] in the sense that 1) motion information is implicitly contained in 3-D LSK while [13] explicitly compute optical flow; 2) background subtraction was used as a pre-processing step, while our method is fully automatic, 3) [13] employed multiple instance learning for action classification while our proposed method deals with both action detection and classification from a single example.

### *C. Detecting Similar Actions using the Matrix Cosine Measure*

*1) Matrix Cosine Similarity:* The next step in the proposed framework is a decision rule based on the measurement of a *distance* between the computed feature volumes $\mathbf{F}_Q, \mathbf{F}_{T_i}$. We were motivated by earlier works such as [46], [41], [42], that have shown the effectiveness of correlation-based similarity.

The *Matrix Cosine Similarity* (MCS) between two feature matrices $\mathbf{F}_Q, \mathbf{F}_{T_i}$ which consist of a set of feature vectors can be defined as the Frobenius inner product between two normalized matrices as follows:

$$\rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) = < \overline{\mathbf{F}}_Q, \overline{\mathbf{F}}_{T_i} >_F = \mathrm{trace}\left(\frac{\mathbf{F}_Q^T \mathbf{F}_{T_i}}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F}\right) \in [-1, 1], \tag{9}$$

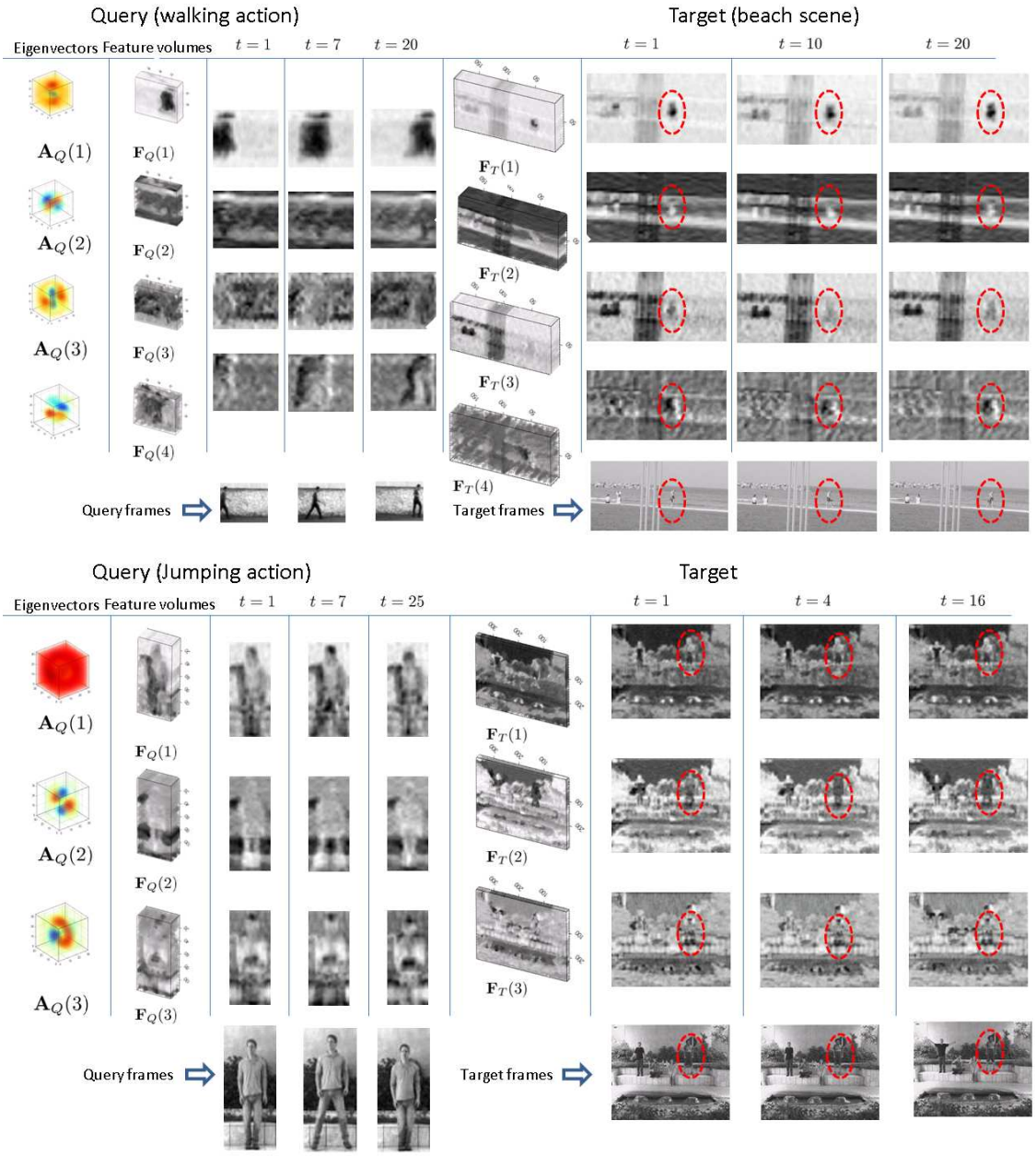


Fig. 11. $\mathbf{A}_Q$ is learned from a collection of 3-D LSKs $\mathbf{W}_Q$, and Feature row vectors of $\mathbf{F}_Q$ and $\mathbf{F}_T$ are computed from query $Q$ and target video $T$ respectively. Eigenvectors and feature vectors were transformed to volume and up-scaled for illustration purposes. Note that even though there are cluttered background in $T$ (jumping action), our feature representation can capture only salient characteristics of actions.

where, $\overline{\mathbf{F}}_Q = \frac{\mathbf{F}_Q}{\|\mathbf{F}_Q\|_F} = \frac{1}{\|\mathbf{F}_Q\|_F}[\mathbf{f}_Q^1, \cdots, \mathbf{f}_Q^n]$ and $\overline{\mathbf{F}}_{T_i} = \frac{\mathbf{F}_{T_i}}{\|\mathbf{F}_{T_i}\|_F} = \frac{1}{\|\mathbf{F}_{T_i}\|_F}[\mathbf{f}_{T_i}^1, \cdots, \mathbf{f}_{T_i}^n]$. Equation (9) can be rewritten as a weighted sum of the vector cosine similarities $\rho(\mathbf{f}_Q, \mathbf{f}_{T_i}) = \frac{\mathbf{f}_Q^T \mathbf{f}_{T_i}}{||\mathbf{f}_Q|| ||\mathbf{f}_{T_i}||}$ ([46], [41], [42]) between each pair of corresponding feature vectors (i.e., columns) in $\mathbf{F}_Q, \mathbf{F}_{T_i}$

as follows:

$$\rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) = \sum_{\ell=1}^{n} \frac{\mathbf{f}_Q^{\ell\,T} \mathbf{f}_{T_i}^{\ell}}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F} = \sum_{\ell=1}^{n} \rho(\mathbf{f}_Q^{\ell}, \mathbf{f}_{T_i}^{\ell}) \frac{\|\mathbf{f}_Q^{\ell}\| \|\mathbf{f}_{T_i}^{\ell}\|}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F}. \tag{10}$$

The weights are represented as the product of $\frac{\|\mathbf{f}_Q^{\ell}\|}{\|\mathbf{F}_Q\|_F}$ and $\frac{\|\mathbf{f}_{T_i}^{\ell}\|}{\|\mathbf{F}_{T_i}\|_F}$ which indicate the relative importance of each feature in the feature sets $\mathbf{F}_Q, \mathbf{F}_{T_i}$. We see here an advantage of the MCS in that it takes account of the strength and angle similarity of vectors at the same time. Hence, this measure not only generalizes the cosine similarity naturally, but also overcomes the disadvantages of the conventional Euclidean distance which is sensitive to outliers.[13]

It is worth noting that [1] proposed 3-D volume correlation score (global consistency measure between query and target cube) by computing a weighted average of local consistency measures. The difficulty with that method is that local consistency values should be explicitly computed from each corresponding subvolume of the query and target video. Furthermore, the weights to calculate a global consistency measure are based on a sigmoid function, which is somewhat ad-hoc. Here, we claim that our MCS measure is better motivated, more general, and effective than their global consistency measure for action detection as we also allude to in section III-A.

The next step is to generate a so-called resemblance volume (RV), which will be a volume of voxels, each indicating the likelihood of similarity between the $Q$ and $T_i$. As for the final test statistic comprising the values in the resemblance volume, we use the *proportion* of shared variance ($\rho_i^2$) to that of the "residual" variance $(1 - \rho_i^2)$. More specifically, RV is computed as follows:

$$\mathrm{RV} : f(\rho_i) = \frac{\rho_i^2}{1 - \rho_i^2}. \tag{11}$$

The resemblance volume generated from $f(\rho_i)$ provides better contrast and dynamic range in the result $(f(\rho_i) \in [0, \infty])$. More importantly, from a quantitative point of view, we note that $f(\rho_i)$

[13]We compute $\rho(\mathbf{F}_Q, \mathbf{F}_{T_i})$ over $M$ (a possibly large number of) target cubes and this can be efficiently implemented by column-stacking the matrices $\mathbf{F}_Q, \mathbf{F}_{T_i}$ and simply computing the (vector) cosine similarity between two long column vectors as follows:

$$\begin{aligned} \rho_i \equiv \rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) &= \sum_{\ell=1}^{n} \frac{\mathbf{f}_Q^{\ell\,T} \mathbf{f}_{T_i}^{\ell}}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F} = \sum_{\ell=1}^{n} \sum_{j=1}^{d} \frac{f_Q^{(\ell,j)} f_{T_i}^{(\ell,j)}}{\sqrt{\sum_{\ell=1}^{n} \sum_{j=1}^{d} |f_Q^{(\ell,j)}|^2} \sqrt{\sum_{\ell=1}^{n} \sum_{j=1}^{d} |f_{T_i}^{(\ell,j)}|^2}}, \\ &= \rho(\mathrm{colstack}(\mathbf{F}_Q), \mathrm{colstack}(\mathbf{F}_{T_i})) \in [-1, 1], \end{aligned}$$

where $f_Q^{(\ell,j)}, f_{T_i}^{(\ell,j)}$ are elements in the $\ell^{th}$ vector $\mathbf{f}_Q^{\ell}$ and $\mathbf{f}_{T_i}^{\ell}$ respectively, and $\mathrm{colstack}(\cdot)$ means an operator which column-stacks (rasterizes) a matrix.

is essentially the Lawley-Hotelling trace statistic [59], [60], which is used as an efficient test statistic for detecting correlation between two data sets. Furthermore, historically, this statistic has been suggested in the pattern recognition literature as an effective means of measuring the separability of two data clusters (e.g. [57].)

*2) Non-Parametric Significance Test:* If the task is to find the most similar cube ($T_i$) to the query ($Q$) in the target video, one can choose the cube which results in the largest value in the RV (i.e., $\max f(\rho_i)$) among all the cubes, no matter how large or small the value is in the range of $[0, \infty]$. This, however, is unwise because there may not be *any* action of interest present in the target video. We are therefore interested in two types of significance tests. The first is an overall test to decide whether there is any sufficiently similar action present in the target video at all. If the answer is yes, we would then want to know how many actions of interest are present in the target video and where they are. Therefore, we need two thresholds: an overall threshold $\tau_o$ and a threshold $\tau$ to detect the (possibly) multiple occurrences of similar actions in the target video.

In a typical scenario, we set the overall threshold $\tau_o = 1$, which is about $50\%$ of variance in common[14] (i.e., $\rho^2 = 0.5$). In other words, if the maximal $f(\rho_i)$ is just above $1$, we decide that there exists at least one action of interest. The next step is to choose $\tau$ based on the properties of $f(\rho_i)$. When it comes to choosing the $\tau$, there is need to be more careful. If we have a basic knowledge of the underlying distribution of $f(\rho_i)$, then we can make predictions about how this particular statistic will behave, and thus it is relatively easy to choose a threshold which will indicate whether the pair of features from the two videos are sufficiently similar. But, in practice, we do not have a very good way to model the distribution of $f(\rho_i)$. Therefore, instead of assuming a type of underlying distribution, we employ the idea of nonparametric testing. Namely, we compute an empirical probability density function (PDF) from the samples $f(\rho_i)$ in the given resemblance volume, and set $\tau$ so as to achieve, for instance, a 99 percent significance level in deciding whether the given values are in the extreme (right) tails of the distribution. This approach is based on the assumption that in the target video, most cubes do not contain the action of interest (in other words, an action of interest is a relatively rare event), and therefore, the few matches will result in values which are in the tails of the distribution of $f(\rho_i)$.

[14]This in effect represents an unbiased choice reflecting our lack of prior knowledge about whether any similar actions are present at all.

*3) Non-maxima Suppression:* After the two significance tests with $\tau_o, \tau$ are performed, we employ the idea of non-maxima suppression [61] for the final detection. We take the volume region with the highest $f(\rho_i)$ value and eliminate the possibility that any other action is detected within some radius[15] of the center of that volume again. This enables us to avoid multiple false detections of nearby actions already detected. Then we iterate this process until the local maximum value falls below the threshold $\tau$.

## III. Experimental Results

In this section, we demonstrate the performance of the proposed method with comprehensive experiments on three datasets; namely, the general action dataset [1], the Weizmann action dataset [2], and the KTH action dataset [3]. The general action dataset is used to evaluate detection performance of the proposed method, while the Weizmann action dataset and the KTH action dataset are employed for action categorization. Comparison is made with state-of-the-art methods that have reported their results on these datasets.

### *A. Action Detection*

In this section, we show several experimental results on searching with a short query video against a (typically longer and larger ) target video. Our method detects the presence and location of actions similar to the given query and provides a series of bounding cubes with resemblance volume embedded around detected actions. Note again that no background/foreground segmentation is required in the proposed method. Our proposed method can also handle modest variations in rotation (up to $\pm 15$ degree), and spatial and temporal scale change (up to $\pm 20\%$). Given $Q$ and $T$, we spatially blur and downsample both $Q$ and $T$ by a factor of 3 in order to reduce the time-complexity. We then compute 3-D LSK of size $3 \times 3$ (space) $\times 7$ (time) as descriptors so that every space-time location in $Q$ and $T$ yields a 63-dimensional local descriptor $\mathbf{W}_Q$ and $\mathbf{W}_T$ respectively. The reason why we choose a lager time axis size than space axis of the cube is that we focus on detecting similar actions regardless of different appearances. Thus we give a higher priority to temporal evolution information than spatial appearance. We end up with

[15]The size of this exclusion region will depend on the application at hand and the characteristics of the query video.

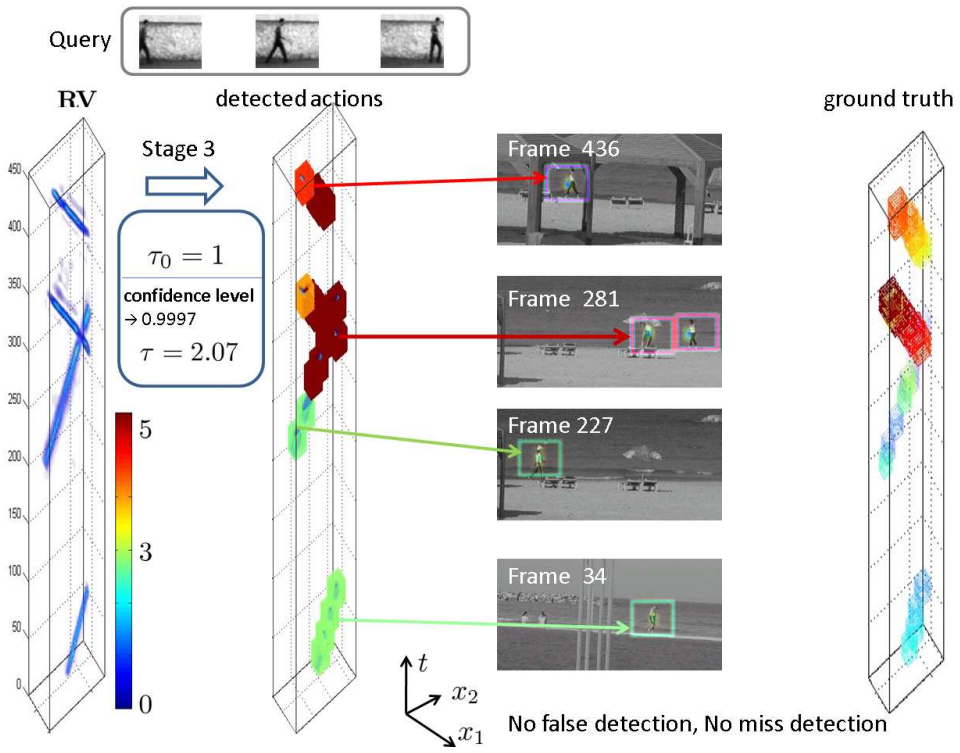


Fig. 12. Results of searching for walking persons on the beach. Top: a short walk query, Left: resemblance volume (RV), Middle: detected walking actions after two significance tests ($\tau_0 = 1, \tau = 2.07$) followed by non-maxima suppression. Red areas among detected volumes mean higher resemblance to the given query. Right: ground truth volume. Note that colors in the ground truth volume are used to distinguish individual actions from each other. This figure is better illustrated in color.

$\mathbf{F}_Q, \mathbf{F}_T$ by reducing dimensionality from 63 to $d = 4$ and then, we obtain RV by computing the MCS measure between $\mathbf{F}_Q, \mathbf{F}_T$.

We applied our method to 4 different examples [1]: for detecting 1) walking people, 2) ballet turn actions, 3) swim jump actions and 3) multiple actions in one video.

Fig. 12 shows the results of searching for instances of walking people in a target beach video (456 frames of $180 \times 360$ pixels). The query video contains a very short walking action moving to the right (14 frames of $60 \times 70$ pixels) and has a background context which is not the beach scene. In order to detect walking actions in either directions, we used two queries ($Q$ and its mirror-reflected version) and generated two RVs. By voting the higher score among values from two RVs at every space-time location, we arrived at one RV which includes correct locations of walking people in the correct direction. Red color represents higher resemblance while blue color denotes lower resemblance values. The threshold $\tau$ for each test example was determined

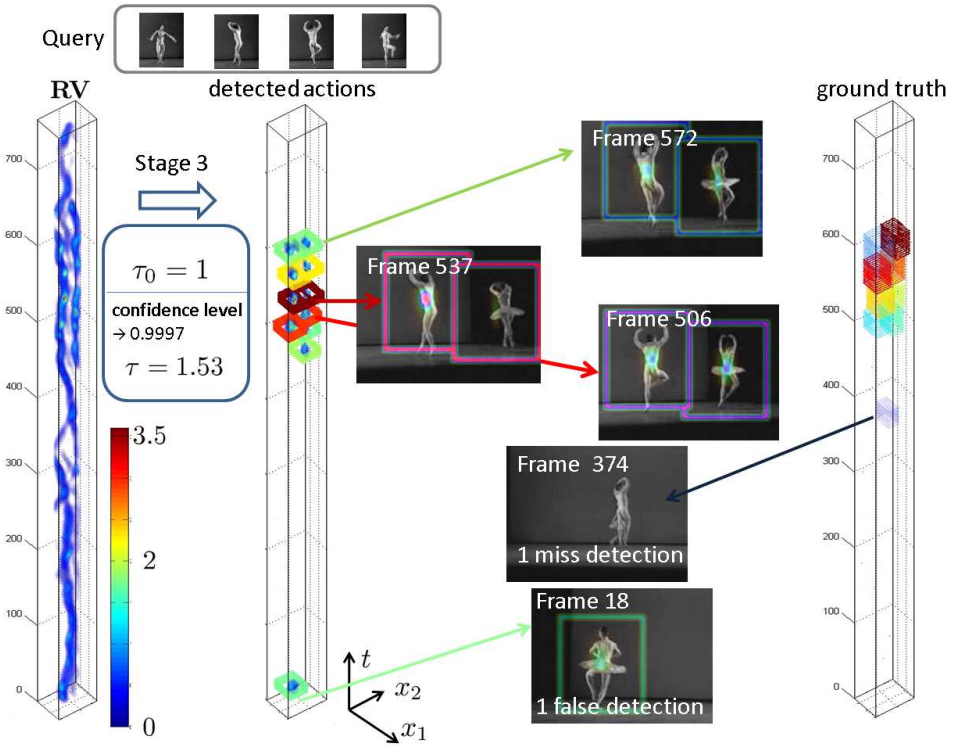


Fig. 13. Results of searching ballet turn on the ballet video. Top: a short ballet turn query, Left: resemblance volume (RV), Middle: detected turning actions after two significance tests ($\tau_0 = 1, \tau = 1.53$) followed by non-maxima suppression. Right: ground truth volume. Note that colors in the ground truth volume are used to distinguish individual actions from each other. This figure is better illustrated in color.

by the 99.97 percent confidence level[16].

Fig. 13 shows the results of detecting ballet turning action in a target ballet video (766 frames of $144 \times 192$ pixels). The query video contains a single turn of a male dancer (13 frames of $90 \times 110$ pixels). Note that this video contains very fast moving parts and contains large variability in spatial scale and appearance (the female dancer wearing a skirt) as compared to the given query $Q$. After two significance testing with $\tau_0 = 1, \tau = 1.53$, most of the turns of the two dancers (a male and a female) with two false positives and one miss were detected. However, if we set the confidence level to 0.998 instead of 0.9997, all of the turns are detected with more false positives.

Fig. 14 shows the results of detecting diving action in a target Olympic relay-match video

[16]Compared to object detection [35], there are typically many more samples in the computed resemblance volume in 3-D. Therefore, the candidate regions are more rare, and thus the confidence levels need to be somewhat higher.

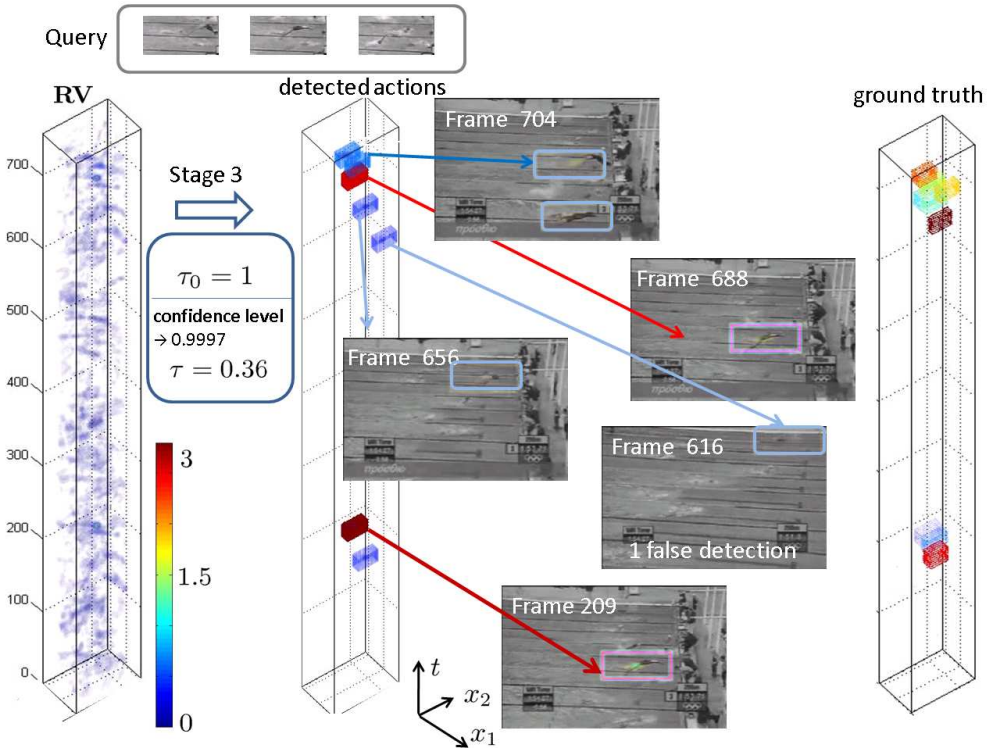


Fig. 14. Results of searching for diving on the Olympic swim-relay match video. Top: a short dive query, Left: resemblance volume (RV), Middle: detected diving actions after two significance tests ($\tau_0 = 1, \tau = 0.36$) followed by non-maxima suppression. Right: ground truth volume. Note that colors in the ground truth volume are used to distinguish individual actions from each other. This figure is better illustrated in color.

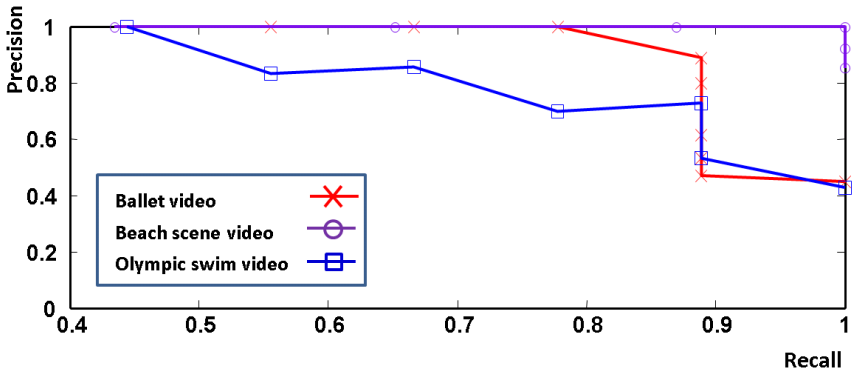


Fig. 15. Precision-Recall curves of our method for three different actions (walk, ballet turn, and dive) shown in Fig. 12, 13, and 14. Note that other state-of-the art action detection methods in [1], [50], [25] did not provide any quantitative performance on these examples.

(757 frames of $240 \times 360$ pixels). This target video was severely MPEG compressed. The query video contains a single dive into a pool (16 frames of $70 \times 140$ pixels). Most of the dives with

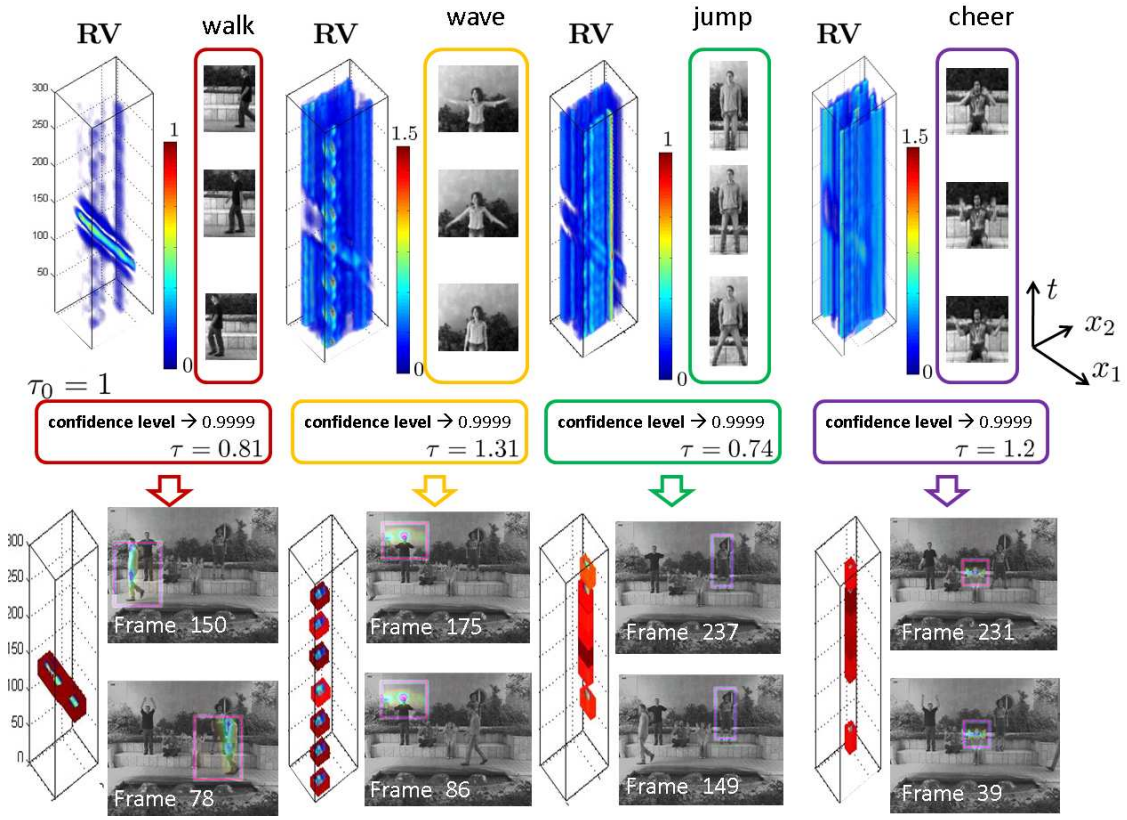


Fig. 16. Results of searching multiple actions in one video. Top: RVs with respect to four different action queries such as walk, wave, jump, and cheer, Bottom: detected actions after two significance tests ($\tau_0 = 1$, confidence level: 0.9999) followed by non-maxima suppression. This figure is better illustrated in color.

a few false positives were detected even though this video contains the severe noise and various actions such as a variety of swim styles, flips under the water, and splashes of water.

It is worth noting here that training-free action detection methods [1], [50], [25] which also tested on this dataset only presented qualitative results with either empirically chosen threshold values or no description about how the threshold values are determined. On the other hand, the threshold values are automatically chosen in our algorithm with respect to the confidence level. Unlike [1], [50], [25], we provide the precision-recall curves[17] in Fig. 15 for the quantitative evaluation. For these experiments, we used the entire frames while [1], [50], [25] used a part of video frames. The detection result of the proposed method on this video outperforms those in

[17]Precision = $\frac{TP}{TP+FP}$, and Recall = $\frac{TP}{nP}$ where $TP$ is true positives, $FP$ is false positives, and $nP$ is the total number of positives in the dataset.

[1], [25] which had a number of miss detections and false positives, and compares favorably to that in [50] in terms of visual detection accuracy.

Fig.16 shows the results of detecting 4 different actions (walk, wave, cheer, and jump) which occur simultaneously in a target video (300 frames of 288 $\times$ 360 pixels). Four query videos were matched against the target video independently. Most of the actions were detected although one of two cheer actions on the target video was missed because it contains head shaking as well while the query video does not have any head motion.

In all the above examples, we used the same parameters. It is evident, based on all the results above, that the proposed training-free action detection based on 3-D LSK works well and is robust to modest variations in spatiotemporal scale.

### *B. Action Category Classification*

As opposed to action detection, action category classification aims to classify a given action query into one of several pre-specified categories. In earlier discussion on action detection, we assumed that in general the query video is smaller than the target video. Now we relax this assumption, and thus we need a preprocessing step which selects a valid human action from the query video. This idea allows us not only to extend the proposed detection framework to action category classification, but also improves both detection and classification accuracy by removing unnecessary background from the query video.

Once the query video is cropped to a short action clip, the cropped query is searched against each labeled video in the database, and the value of the resemblance volume (RV) is viewed as the likelihood of similarity between the query and each labeled video. Then we classify a given query video as one of the predefined action categories using a nearest neighbor (NN) classifier.

*1) Action Cropping in Videos:* In this section, we introduce a procedure which automatically extracts from the query video a small cube that only contains a valid action. Space-time saliency detection [36] can provide such a mechanism (see Fig. 17.) We downsample each frame of query video $Q$ to a coarse spatial scale $(64 \times 64)$ in order to reduce the time-complexity[18]. We then compute 3-D LSK of size 3 $\times$ 3 $\times$ 3 as features and generate feature matrices $\mathbf{F}_i$ in a (3 $\times$ 3 $\times$ 7) local space-time neighborhood. We generated space-time saliency maps $S$ by computing

[18]We do not downsample the video in the time domain.

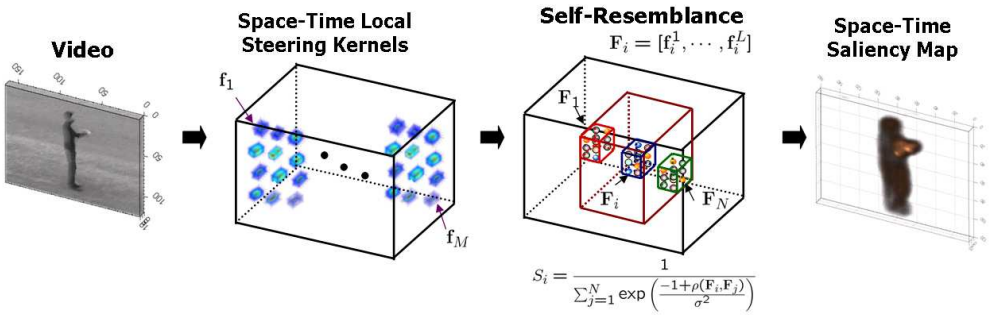


Fig. 17. Graphical overview of space-time saliency detection system.

self-resemblance measure as shown in Fig. 18. Then, we use again the idea of non-parametric significance testing to detect space-time proto-objects. Namely, we compute an empirical PDF from all the saliency values and set a threshold so as to achieve, a 95 $\%$ significance level[19] in deciding whether the given saliency values are in the extreme (right) tails of the empirical PDF. The approach is based on the assumption that in the video, a salient action is a relatively rare event and thus results in values which are in the tails of the distribution of saliency map values. After making a binary map by thresholding the space-time saliency map, a morphological filter is applied. More specifically, we dilate the binary object map with a disk shape of size $5 \times 5$. Proto-objects are extracted from corresponding locations of the original video. Fig. 18 shows that the space-time saliency detection method successfully detects only salient human actions in both the Weizmann dataset [2] and the KTH dataset [3]. Next, we crop the valid human action region by fitting a 3-D rectangular box to space-time proto-objects.

*2) Weizmann Action Data Set:* The Weizmann action dataset contains 10 actions (bend, jumping jack, jump forward, jump in place, jump sideways, skip, run, walk, wave with two hands, and wave with one hand) performed by 9 different subjects. This dataset contains videos with static cameras and simple background, but it provides a good testing environment to evaluate the performance of the algorithm when the number of categories are large compared to the KTH dataset (a total of 6 categories). The testing was performed in a "leave-one-out" setting, *i.e.*, for each run the videos of 8 subjects are labeled and the videos of the remaining subject are used for testing (query). We applied the automatic action cropping method introduced in the previous section to the query video. Then the resulting short action clip is matched against the remaining

[19]We select a somewhat loose confidence level here since we do not wish to miss the relevant action in the query.

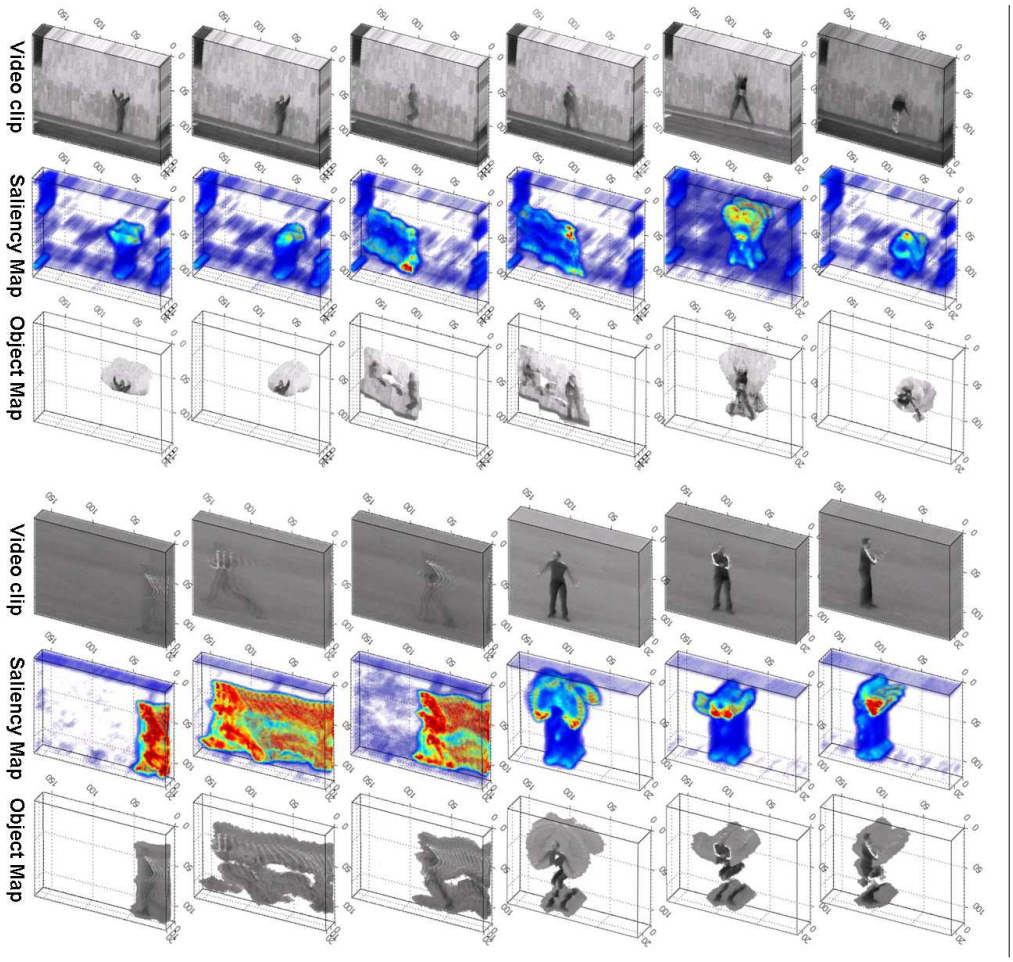


Fig. 18. Found space-time proto-objects from the Weizmann [2] and the KTH dataset [3]

labeled videos using the proposed method. We classify each testing video as one of the 10 action types by 3-NN (nearest neighbor) as similarly done in [25]. The results are reported as the average of nine runs. We were able to achieve a recognition rate of 96% for all ten actions. The recognition rate comparison is provided in Table I as well. The proposed method which is training-free performs favorably against state-of-the-art methods [15], [62], [21], [63], [13], [64], [65], [56] which largely depend on training[20]

[20]It is worth noting that different groups employed different experimental methodologies. There are broadly two main evaluation methods: 1) leave-one-out and 2) split-data-equally. The split-data-equally means that the a collection of video sequences are divided into two equal sets randomly: one for training examples and the other for testing (query). Since our method does not involve any training, we adopted the leave-one-out in this paper. Other recent papers [66], [67] also used the leave-one-out setting.

TABLE I

COMPARISON OF AVERAGE RECOGNITION RATE ON THE WEIZMANN DATASET ([2])

| Our approach | 3-NN | 2-NN | 1-NN |
|---|---|---|---|
| Recognition rate | **96**% | **90**% | **90**% |

| Method | Juenjo et al. [62] | Liu et al. [21] | Klaser et al. [56] | Schindler and Gool [68] |
|---|---|---|---|---|
| Recognition rate | 95.33% | 90% | 84.3% | 100% |
| Method | Niebles et al. [15] | Ali et al. [13] | Sun et al. [69] | Fathi and More [70] |
| Recognition rate | 90% | 95.75% | 97.8% | 100% |
| Method | Jhuang et al. [63] | Batra et al. [64] | Bregonzio et al. [66] | Zhang et al. [65] |
| Recognition rate | 98.8% | 92% | 96.6% | 92.89% |

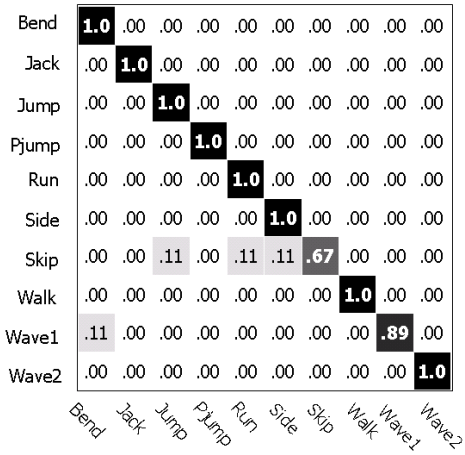


Fig. 19. Average confusion matrix for the Weizmann action dataset. (Here, 3-NN was used as similarly done in [25].)

We further provide the results using 1-NN and 2-NN for comparison. We observe that these results also compare favorably to several state-of-the-art methods even though our method involves no training phase, and requires no background/foreground segmentation. As an added bonus, our method provides localization of actions as a side benefit. Fig. 19 shows the confusion matrix for our method. Note that our method is mostly confused by similar action classes, such as skip with jump, run, and side.

*3) KTH Action Data Set:* In order to further quantify the performance of our algorithm, we also conducted experiments on the KTH dataset. The KTH action dataset contains six types of

human actions (boxing, hand waving, hand clapping, walking, jogging, and running), performed repeatedly by 25 subjects in 4 different scenarios: outdoors ($c_1$), outdoors with camera zoom ($c_2$), outdoors with different clothes ($c_3$), and indoors ($c_4$). This dataset seems more challenging than the Weizmann dataset because there are large variations in human body shape, view angles, scales, and appearance. The "leave-one-out" cross validation is again used to measure the performance. Fig. 20 shows the confusion matrices from our method for each scenario and the average confusion matrix across all scenarios. We were able to achieve a recognition rate of $95.66\%$ on these six actions. The recognition rate comparison with competing methods is provided in Table II as well. It is worth noting that our method outperforms all the other state-of-the-art methods and is fully automatic. Table III further shows that our scenario-wise recognition rates are consistently higher than those reported in [25] and [63].

TABLE II

DETAILED COMPARISON OF RECOGNITION RATE ON THE KTH DATASET. $Avg$ IS THE AVERAGE ACROSS 4 SCENARIOS.

| Methods | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $Avg$ |
|---|---|---|---|---|---|
| Our Approach | $\mathbf{97.33}\%$ | $\mathbf{92.67}\%$ | $\mathbf{95.3}\%$ | $\mathbf{97.32}\%$ | $\mathbf{95.66}\%$ |
| Ning et al. [25] (3-NN) | 95.56 % | 82.41 % | 90.66 % | $94.72\%$ | $92.09\%$ |
| Jhuang et al. [63] | 96.0 % | 89.1 % | 89.8 % | $94.8\%$ | $91.7\%$ |

TABLE III

COMPARISON OF AVERAGE RECOGNITION RATE ON THE KTH DATASET

| Our approach | 3-NN | 2-NN | 1-NN |
|---|---|---|---|
| Recognition rate | **95.66**% | **93**% | **89**% |

| Method | Kim et al. [24] | Ning et al. [25] | Klaser et al. [56] | Schindler and Gool [68] |
|---|---|---|---|---|
| Recognition rate | $95.33\%$ | $92.31\%$ (3-NN) | $91.4\%$ | $92.7\%$ |
| Method | Ali et al. [13] | Niebles et al. [15] | Liu and Shah [71] | Sun et al. [69] |
| Recognition rate | $87.7\%$ | $81.5\%$ | $94.2\%$ | $94\%$ |
| Method | Dollar et al. [72] | Wong et al. [73] | Rapantzikos et al. [67] | Laptev et al. [55] |
| Recogniton rate | $81.17\%$ | $84\%$ | $88.3\%$ | $91.8\%$ |

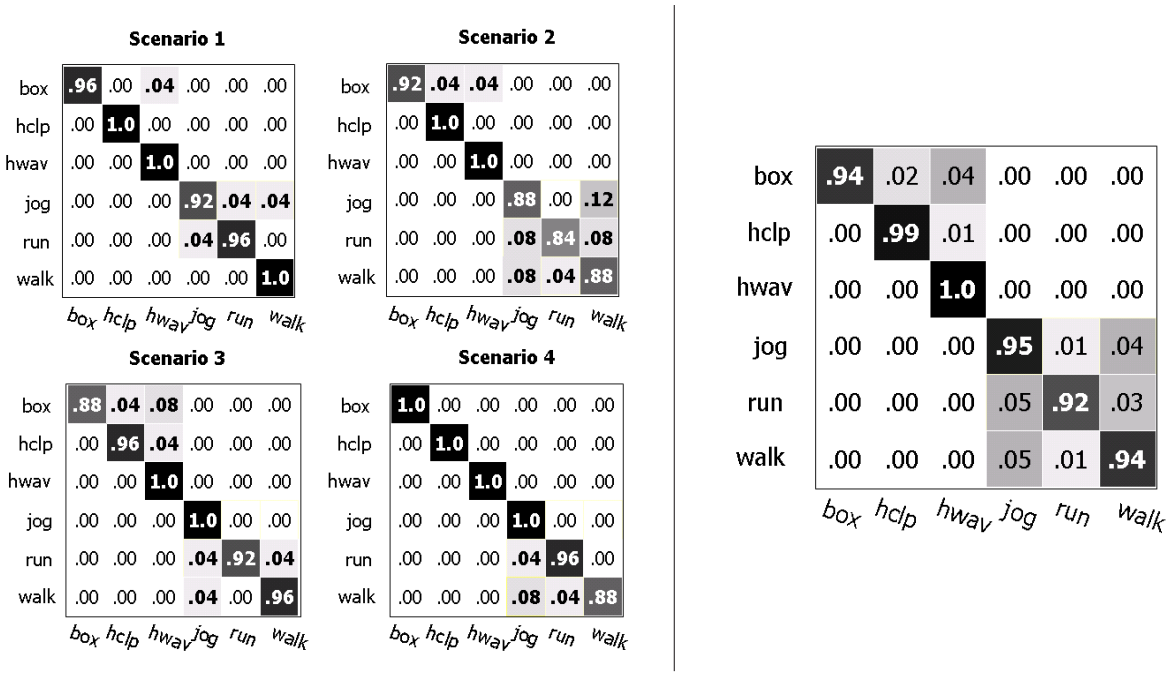


Fig. 20. Left: Tables of confusion matrix for the KTH action dataset in each scenario, Right: Average confusion matrix for the KTH action dataset across all scenarios (Here, 3-NN was used as similarly done in [25].)

*4) Discussion:* Our system is designed with recognition accuracy as a high priority. A typical run of the action detection system implemented in Matlab takes a little over 1 minute on a target video $T$ (50 frames of $144 \times 192$ pixels, Intel Pentium CPU 2.66 Ghz machine) using a query $Q$ (13 frames of $90 \times 110$). Most of the run-time is taken up by the computation of MCS (about 9 seconds, and 16.5 seconds for the computation of 3-D LSKs from $Q$ and $T$ respectively, which needs to be computed only once.) There are many factors that affect the precise timing of the calculations, such as query size, complexity of the video, and 3-D LSK size. By applying coarse-to-fine search [74] or branch and bound [31] can be applied to speed up the method. As another way of reducing time-complexity, we could use look-up table instead of computing the local covariance matrix $\mathbf{C}$ at every pixel. A multi-scale approach [75], [35] can also be applied to improve efficiency. Even though our method is stable in the presence of moderate amount of camera motion, our system can benefit from camera stabilization methods as done in [76], [77] in case of large camera movements.

In the Weizmann dataset and the KTH dataset, target videos contain only one type of action. However, as shown in Fig 16, target video may contain multiple actions in practice. In this case,

simple nearest neighbor classifiers can possibly fail. Therefore, we might benefit from contextual information to increase accuracy of action recognition systems as similarly done in [78]. In fact, there is a broad agreement in the computer vision community about the valuable role that context plays in any image understanding task [79], [80].

## IV. Conclusion and Future Directions

In this paper, we have proposed a novel action recognition algorithm by employing *space-time local steering kernels* which robustly capture underlying space-time data structure; and by using a training-free nonparametric detection scheme based on *Matrix Cosine Similarity*. The proposed method can automatically detect in the target video the presence, the number, as well as location of actions similar to the given query video. In order to increase the detection accuracy and further deal with action classification, we employed action cropping method based on space-time saliency detection. Challenging sets of real-world human action experiments demonstrated that the proposed approach achieves a high recognition accuracy and improves upon other state-of-the-art methods. Unlike most state-of-the-art methods that involve training, background/foreground segmentation, and manual aligning of actions, the proposed method operates using a *single* example of an action of interest to find similar matches; does not require any prior knowledge (learning) about actions being sought; and does not require any segmentation or pre-processing step of the target video. Since local regression kernels in 2-D and in 3-D were originally designed for image (video) restoration, the proposed framework should become useful in jointly addressing the problems of enhancement and recognition where there might be a degraded query or target. By computing local regression kernels from images (video) at once, we may be able to not only detect objects (actions) of interest, but enhance images (videos) at the same time. These aspects of the work are the subject of ongoing research.

## V. Acknowledgment

This work was supported in part by AFOSR Grant FA 9550-07-01-0365

## References

[1] 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 Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 11, pp. 2045–2056, November 2007.

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," *IEEE Conference on Pattern Recognition (ICPR)*, June 2004.

[4] T. Darrell and A. Pentland, "Classifying hand gestures with a view-based distributed representation," *In Advances in Neural Information Processing Systems*, vol. 6, pp. 945–952, 1993.

[5] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time sequential image using hidden markov model," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1992.

[6] H. Jiang, M. Crew, and Z. Li, "Successive convex matching for action detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[7] T. Starner and A. Pentland, "Visual recognition of American sign language using hidden Markov model," *International Workshop on Automatic Face and Gesture Recognition*, 1995.

[8] C. Carlsson and J. Sullivan, "Action recognition by shape matching to key frame," *Workshop on Models Versus Examplars in Computer Vision*, 2001.

[9] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[10] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[11] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, no. 3, pp. 1257–1265, March 2001.

[12] J. Little and J. Boyd, "Recognizing people by their gait: The shape of motion," *Journal of Computer Vision Research*, 1998.

[13] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

[14] Y. Yacoob and M. Black, "Parameterized modeling and recognition of activities," *Computer Vision and Image Understanding*, vol. 73, pp. 232–247, 1999.

[15] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories usig spatial-temporal words," *International Journal of Computer Vision (IJCV)*, vol. 79, no. 3, pp. 299–318, March 2008.

[16] J. Niebles and L. Fei-Fei, "A hierarchical models of shape and appearance for human action classification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007.

[17] Z. Laptev and T. Lindeberg, "Space-time interest points," *IEEE International Conference on Computer Vision (ICCV)*, October 2003.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[19] A. Oikonomopoulous, I. Patras, and M. Pantic, "Spationtemporal saliency for human action recognition," *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[20] T. Mahmood, A. Vasilescu, and S. Sethi, "Recognition of action events from multiple video points," *IEEE Workshop on Detection and Reconition of Events in Video, (ICCV)*, 2001.

[21] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[22] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," *ACM Multimedia*, 2007.

[23] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[24] T. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.

[25] H. Ning, T. Han, D. Walther, M. Liu, and T. Huang, "Hierarchical space-time model enabling efficient search for human actions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 6, pp. 808–820, 2009.

[26] C. Cedras and M. Shah, "Motion based recognition: A survey," *Image and Vision Computing*, vol. 13, pp. 129–155, 1995.

[27] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, pp. 428–440, 1999.

[28] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, November 2008.

[29] J. Yuan, Z. Liu, and Y. Wu, "Discriminative subvolume seaerch for efficient action detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[30] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[31] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[32] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision (IJCV)*, vol. 57, no. 2, pp. 137–154, 2004.

[33] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[34] C. Yeo, P. Ahammad, K. Ramchandran, and S. S. Satry, "High-speed action recognition and localization in compressed domain videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 1006–1015, Aug 2008.

[35] H. J. Seo and P. Milanfar, "Training-free, generic object detection using locally adaptive regression kernels," *Accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[36] ——, "Static and space-time visual saliency detection by self-resemblance," *Submitted to Journal of Vision*, May 2009.

[37] ——, "Nonparametric bottom-up saliency detection by self-resemblance," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1st International Workshop on Visual Scene Understanding (ViSU09)*, Apr 2009.

[38] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing (TIP)*, vol. 16, no. 2, pp. 349–366, February 2007.

[39] ——, "Deblurring using regularized locally-adaptive kernel regression," *IEEE Transactions on Image Processing (TIP)*, vol. 17, pp. 550–563, April 2008.

[40] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," *IEEE Transactions on Image Processing (TIP)*, vol. 18, no. 9, pp. 1958–1975, September 2009.

[41] Y. Fu, S. Yan, and T. S. Huang, "Correlation metric for generalized feature extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 12, pp. 2229–2235, 2008.

[42] Y. Fu and T. S. Huang, "Image classification using correlation tensor analysis," *IEEE Transactions on Image Processing (TIP)*, vol. 17, no. 2, pp. 226–234, 2008.

[43] C. Liu, "The Bayes decision rule induced similarity measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 6, pp. 1086–1090, 2007.

[44] ——, "Clarification of assumptions in the relationship between the Bayes decision rule and the whitened cosine similarity measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 6, pp. 1116–1117, 2008.

[45] D. Lin, S. Yan, and X. Tang, "Comparative study: Face recognition on unspecific persons using linear subspace methods," *IEEE International Conference on Image Processing (ICIP)*, 2005.

[46] Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant analysis in correlation similarity measure space," *IEEE International Conference on Machine Learning (ICML)*, vol. 227, pp. 577–584, 2007.

[47] J. W. Schneider and P. Borlund, "Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 11, pp. 1586–1595, 2007.

[48] P. Ahlgren, B. Jarneving, and R. Rousseau, "Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 6, pp. 550–560, 2003.

[49] J. Rodgers and W. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[50] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[51] J. Boulanger, C. Kervrann, and P. Bouthemy, "Space-time adaptation for patch-based image sequence restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, pp. 1096–1102, June 2005.

[52] A. Buades, B. Coll, and J. M. Morel, "Nonlocal image and movie denoising," *International Journal of Computer Vision (IJCV)*, vol. 76, no. 2, pp. 123–139, 2008.

[53] Y. Chen, J. Bi, and J. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transcations on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, no. 12, pp. 1931–1947, 2006.

[54] N. Dalai and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[55] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[56] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," *British Machine Vision Conference*, sep 2008. [Online]. Available: http://lear.inrialpes.fr/pubs/2008/KMS08

[57] R. Duda, P. Hart, and D. Stork, *Pattern Classification, 2nd Edition*. New York: John Wiley and Sons Inc, 2000.

[58] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[59] M. Tatsuoka, *Multivariate Analysis*. Macmillan, 1988.

[60] T. Calinski, M. Krzysko, and W. Wolynski, "A comparison of some tests for determining the number of nonzero canonical correlations," *Communication in Statistics, Simulation and Computation*, vol. 35, pp. 727–749, 2006.

[61] F. Devernay, "A non-maxima suppression method for edge detection with sub-pixel accuracy," *Technical report, INRIA*, no. RR-2724, 1995.

[62] I. Junejo, E. Dexter, I. Laptev, and P. Perez, "Cross-view action recognition from temporal self-similarities." *European Conference Computer Vision (ECCV)*, October 2008.

[63] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recogntion," *IEEE International Conference on Computer Vision(ICCV)*, October 2007.

[64] D. Batra, T. Chen, and R. Sukthankar, "Space-time shapelets for action recogntion," *IEEE Workshop on Motion and video Computing (WMVC)*, January 2008.

[65] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," *European Conference on Computer Vision (ECCV)*, 2008.

[66] M. Bregonzio, S. Gong, and T. Xiang, "Recognising actions as clouds of space-time interest points," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[67] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spationtemporal feature points for action recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[68] K. Schindler and L. Gool, "Action snippets: How many frames does human action recognition require," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[69] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[70] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[71] J. Liu and M. Shah, "Learning human actions via information maximization," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[72] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *In proceeding of Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, October 2005.

[73] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural informtion," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007.

[74] A. Bandapadhay and J. Fu, "Searching parameter spaces with noisy linear constraints," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998.

[75] Y. Ke, "Volumetric features for video event detection," *Ph.D thesis, Computer Science Department, Carnegie Mellon University*, 2008.

[76] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, no. 8, pp. 873–890, August 2001.

[77] P. T. Veit, F. Cao, and P. Bouthemy, "Probabilistic parameter-free motion detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2004.

[78] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[79] S. K. Divvala, D. Hoiem, J. H. Hays, A. Efros, and M. Hebert, "An empirical study of context in object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[80] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends Cognitive Science*, November 2007.