Space-Time Behavior-Based Correlation — OR — How to Tell If Two Underlying Motion Fields Are Similar without Computing Them?

Eli Shechtman, *Student Member*, *IEEE Computer Society*, and Michal Irani, *Member*, *IEEE*

Abstract—We introduce a behavior-based similarity measure that tells us whether two different space-time intensity patterns of two different video segments could have resulted from a similar underlying motion field. This is done directly from the intensity information, without explicitly computing the underlying motions. Such a measure allows us to detect similarity between video segments of differently dressed people performing the same type of activity. It requires no foreground/background segmentation, no prior learning of activities, and no motion estimation or tracking. Using this behavior-based similarity measure, we extend the notion of two-dimensional image correlation into the three-dimensional space-time volume and thus allowing to correlate dynamic behaviors and actions. Small space-time video segments (small video clips) are "correlated" against the entire video sequences in all three dimensions (x, y, and t). Peak correlation values correspond to video locations with similar dynamic behaviors. Our approach can detect very complex behaviors in video sequences (for example, ballet movements, pool dives, and running water), even when multiple complex activities occur simultaneously within the field of view of the camera. We further show its robustness to small changes in scale and orientation of the correlated behavior.

Index Terms—Space-time analysis, motion analysis, action recognition, motion similarity measure, template matching, video correlation, video indexing, video browsing.

1 Introduction

Different people with similar behaviors induce completely different space-time intensity patterns in a recorded video sequence. This is because they wear different clothes, and their surrounding backgrounds are different. What is common across such sequences of the same behaviors is the underlying induced motion fields. This observation was used in [9], where low-pass filtered optical-flow fields (between pairs of frames) were used for action recognition. However, dense, unconstrained, and nonrigid motion estimation is highly noisy and unreliable. Clothes worn by different people performing the same action often have very different spatial properties (different color, texture, and so forth). Uniform-colored clothes induce local aperture effects, especially when the observed acting person is large (which is why Efros et al. [9] analyze small people "at a glance"). Dense flow estimation is even more unreliable when the dynamic event contains unstructured objects like running water, flickering fire, and so forth.

In this paper, we introduce an approach for measuring the degree of consistency (or inconsistency) between the implicit underlying motion patterns in two video segments, *without explicitly computing those motions*. This is done *directly from the space-time intensity (gray scale) information in those two video volumes*. In fact, this "behavioral similarity" measure between two video segments answers the following question: given two completely different space-time intensity patterns (two video segments), could they have been induced by the same (or similar) space-time motion fields? Such a behavioral similarity measure can therefore be used to detect similar behaviors and activities in video sequences despite differences in appearance due to different clothing, different backgrounds, different illuminations, and so forth.

Our behavioral similarity measure requires *no* prior foreground/background segmentation (which is often required in action-recognition methods, for example, [3], [4], [24], [27]). It requires no prior modeling or learning of activities and is therefore *not* restricted to a small set of predefined activities (as opposed to that in [2], [4], [5], [7], [26]). Although [2], [9], [14], [21], [26] require explicit motion estimation or tracking, our method does not. By avoiding explicit motion estimation, we avoid the fundamental hurdles of optical flow estimation (aperture problems, singularities, and so forth). Our approach can therefore handle video sequences of very complex dynamic scenes where motion estimation is extremely difficult, such as scenes with flowing/splashing water, complex ballet movements, and so forth.

Our method is *not* invariant to large geometric deformations of the video template. However, it is not sensitive to small deformations of the template (including small changes in scale and orientation).

We use this measure to extend the notion of traditional two-dimensional (2D) image correlation into a three-dimensional (3D) space-time video-template correlation. The behavioral similarity measure is used here for "correlating"

• *The authors are with the Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, 76100 Rehovot, Israel. E-mail: {eli.shechtman, michal.irani}@weizmann.ac.il.*

Manuscript received 28 July 2006; revised 20 Dec. 2006; accepted 5 Feb. 2007; published online 23 Feb. 2007.
Recommended for acceptance by S. Soatto.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0566-0706.
Digital Object Identifier no. 10.1109/TPAMI.2007.1119.

a small "video query" (a small video clip of an action) against a large video sequence in all three dimensions (x, y, z) for detecting all video locations with high behavioral similarity. This gives rise to various applications based on action detection using a simple example clip (such as video browsing and indexing, unsupervised video clustering based on behaviors [28], "Intelligent fast forward" [29], "Do as I Do" action synthesis [9], and so forth).

Space-time approaches to action recognition, which also perform direct measurements in the space-time intensity video volume, have been previously suggested in [3], [8], [16], [19], [20], [27], [28]. Slices of the space-time volume (such as the xt-plane) were used in [20] for gait recognition. This approach exploits only a small portion of the available data and is limited to cyclic motions. In [3], [27], actions were represented using space-time shape generated by a moving silhouette of a human figure. Informative features were extracted from the space-time shapes using the solution of the Poisson equation inside the shapes [3] or from local geometry of their surface [27]. Although these papers showed promising action recognition performances, they are restricted to cases where figure-background segmentation can be obtained. In [28], empirical distributions of space-time gradients collected from an entire video clip are used. In [8], a video sequence is characterized by the distribution of space-time feature prototypes (cluster centers) extracted at temporal interest points from "cuboids" of space-time gradients.

As such, the work in [28] and [8] are restricted to a single action in the field of view of the camera at any given time and do not capture the geometric structure of the action parts (neither in space, nor in time). In [15], [16], a sparse set of "space-time corner" points are detected and used to characterize the action while maintaining scale invariance. Since there are so few such points in a typical motion, the method may be prone to occlusions and to misdetections of these interest points. Some actions do not include such points at all (see more details in [14] and in [8]). It is therefore also limited to a single action in the field of view of the camera. Niebles et al. [19] extended the method by Dollár et al. [8] by unsupervised learning (similar to the "bag of features" in images) of action classes and showed an application for detecting multiple actions in the field of view. However, this method is based on a complex learning phase from multiple examples of each action, and as in [8], their method does not capture the geometric structure of the action parts.

An approach for registering two video clips (including action clips) was presented in [25]. Their local similarity measure was based on normalized cross-correlation of space-time gradients of corresponding small space-time patches (ST-patches) across the two sequences. Although this measure is invariant to local changes in contrast and brightness, it relies on the similarity of the local structure of the two clips and is therefore *not* invariant to nonlinear local changes in appearance or to more complex local changes in appearance and local deformations. In contrast, our measure is based on the similarity (consistency) of the underlying motion fields and is therefore completely invariant to the appearance of the compared actions or moving objects.

In the area of motion analysis, many studies have utilized eigendecomposition of the spatio-temporal "Structure Tensor" (or "scatter matrix") for accurate estimation of the optical flow [1], [13], nonlinear filtering [23], extraction of space-time interest points [16], extraction of other space-time operators [17], and motion segmentation [10]. In this work, we explore a different kind of analysis using the "Structure Tensor" for a different application.

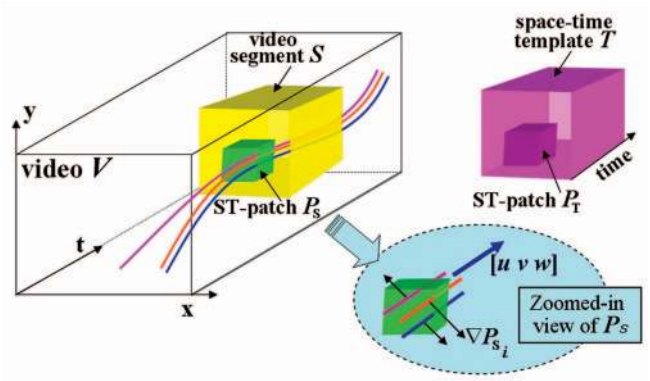

Fig. 1. Overview of framework and notations.

Because our approach captures dense spatio-temporal geometric structure of the action, it can therefore be applied to small video templates. Multiple such templates can be correlated against the same video sequence to detect multiple different activities. A shorter version of this paper appeared in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '05) [22]. To our best knowledge, this is the first work that shows an ability to detect multiple different activities that occur simultaneously in the field of view of the camera, without any prior spatial or temporal segmentation of the video data, and in the presence of cluttered dynamic backgrounds.

The rest of this paper will present our approach for measuring the degree of consistency between two underlying motion fields (without computing any motion) and its applications to action analysis. However, the consistency measure itself may be useful in many other applications that require comparing motion fields (for example, motion segmentation, dynamic texture analysis, and so forth).

1.1 Overview of the Approach and Notations

Fig. 1 provides a graphical view of the notations used in the paper. A small space-time template T (= a very small video clip, for example, $30 \times 30 \times 30$) is "correlated" against a larger video sequence V (for example, $200 \times 300 \times 1,000$) in all three dimensions (x, y, and t). This generates a space-time "behavioral correlation surface" $C(x, y, t)$ or, more precisely, a space-time "behavioral correlation volume" (*not* shown in the figure). Peaks within this correlation volume are locations in the video sequence V with similar behavior to the template T.

Each value in the correlation volume $C(x, y, t)$ is computed by measuring the degree of "behavioral similarity" between two *video segments*: the space-time template T and a video segment $S \subset V$ (of the same dimensions as T), centered around the point $(x, y, t) \in V$. The behavioral similarity between two such video segments, T and S, is evaluated by computing and integrating local consistency measures between small ST-patches (for example, $7 \times 7 \times 3$) within these video segments. Namely, for each point $(i, j, k) \in S$, a small ST-patch $P_S \subset S$ centered around (i, j, k) is compared

against its corresponding[1] small ST-patch $P_T \subset T$ (see Fig. 1). These local scores are then aggregated to provide a global correlation score for the entire template T at this video location. (This is similar to the way correlation of image templates is sometimes performed. However, here, the small patches P also have a temporal dimension, and the similarity measure between patches captures the similarity of the implicit underlying motions.)

We will start by exploring unique properties of intensity patterns induced in small ST-patches P within the video data (Section 2). Step by step, we will develop the consistency measure between two such ST-patches (P_T and P_S) (Sections 3 and 4). These local scores are then aggregated into a more global behavior-based correlation score between two video segments (T and S), which in turn leads to the construction of a correlation volume of the video query T relative to the entire large video sequence V (Section 6). Examples of detecting complex activities (pool dives, ballet dances, and so forth) in a real noisy video footage are shown in Section 7.

2 Properties of a Space-Time Intensity Patch

We will start by exploring unique properties of intensity patterns induced in small space-time patches of video data. In short, we will refer to a small "space-time patch" as an **ST-patch**. If an ST-patch P is small enough (for example, $7 \times 7 \times 3$), then all pixels within it can be assumed to move with a single uniform motion. This assumption is true for most of ST-patches in real video sequences. (It is very similar to the assumption used in [18] for optical flow estimation, but in our case, the patches also have a temporal dimension.) A very small number of patches in the video sequence will violate this assumption. These are patches located at motion discontinuities, as well as patches that contain an abrupt temporal change in the motion direction or velocity.

A locally uniform motion induces a local brush of straight parallel lines of color (or intensity) within the ST-patch P. All the color (intensity) lines within a single ST-patch are oriented in a single space-time direction (u, v, w) (see zoomed in part in Fig. 1). The orientation (u, v, w) can be different for different points (x, y, t) in the video sequence. It is assumed to be uniform only locally, within a small ST-patch P centered around each point in the video. Examining the *space-time gradients* $\nabla P_i = (P_{x_i}, P_{y_i}, P_{t_i})$ of the intensity at each pixel within the ST-patch P $(i = 1 \dots n)$, then these gradients will all be pointing to directions of maximum change of intensity in space time (Fig. 1). Namely, these gradients will all be perpendicular to the direction (u, v, w) of the brush of color/intensity lines

$$\nabla P_i \begin{bmatrix} u \\ v \\ w \end{bmatrix} = 0. \tag{1}$$

Different space-time gradients of different pixels in P (for example, ∇P_i and ∇P_j) are not necessarily parallel to each other. However, they all reside in a single 2D plane in the space-time volume that is perpendicular to (u, v, w). Note that (1) does *not* require for the frame-to-frame displacements to be infinitesimally small, only uniform within P. However, it cannot handle very large motions that induce temporal aliasing. These issues are addressed in Section 6.

Stacking these equations from all n pixels within the small ST-patch P, we obtain

$$\underbrace{\begin{bmatrix} P_{x_1} & P_{y_1} & P_{t_1} \\ P_{x_2} & P_{y_2} & P_{t_2} \\ & \cdots & \\ & \cdots & \\ P_{x_n} & P_{y_n} & P_{t_n} \end{bmatrix}_{n\times 3}}_{\mathbf{G}} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}, \tag{2}$$

where n is the number of pixels in P (for example, if P is $7 \times 7 \times 3$, then $n = 147$). Multiplying both sides of (2) by $\mathbf{G^T}$ (the transposed of the gradient matrix $\mathbf{G}$) yields

$$\mathbf{G}^{\mathrm{T}}\mathbf{G} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times 1}. \tag{3}$$

$\mathbf{G}^{\mathrm{T}}\mathbf{G}$ is a 3×3 matrix. We denote it by $\mathbf{M}$:

$$\mathbf{M} = \mathbf{G}^{\mathrm{T}}\mathbf{G} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y & \Sigma P_x P_t \\ \Sigma P_y P_x & \Sigma P_y^2 & \Sigma P_y P_t \\ \Sigma P_t P_x & \Sigma P_t P_y & \Sigma P_t^2 \end{bmatrix}, \tag{4}$$

where the summation is over all pixels within the ST-patch. Therefore, for all small ST-patches containing a single uniform motion, the matrix $\mathbf{M_{3\times 3}}$ (also called the "Gram matrix" of $\mathbf{G}$) is a *rank-deficient matrix*: $rank(\mathbf{M}) \leq 2$. Its smallest eigenvalue is therefore zero ($\lambda_{min} = 0$), and (u, v, w) is the corresponding eigenvector. Note that the size of $\mathbf{M}$ (3×3) is independent from the size of the ST-patch P. This matrix, is also known as the space-time "Structure Tensor" or "scatter matrix" [1], [10], [13], [16], [17], [23].

Now, if there exists an ST-patch for which $rank(\mathbf{M}) = 3$, then this ST-patch cannot contain a single uniform motion (that is, there is no single $[u\ v\ w]$ vector that is perpendicular to all space-time intensity gradients). In other words, this ST-intensity patch was induced by *multiple independent motions*. Note that this observation is reached by examining $\mathbf{M}$ alone, which is directly estimated from color or intensity information. No motion estimation is required. As mentioned above, $rank(\mathbf{M}) = 3$ happens when the ST-patch is located at spatio-temporal motion discontinuity. Such patches are also known as "space-time corners" [15], [16] or patches of "no coherent motion" [13]. These patches are typically rare in a real video sequence.

3 Consistency between Two ST-Patches

Similar rank-based considerations can assist in telling us whether *two* different ST-patches, P_1 and P_2, with completely different intensity patters, could have resulted from a similar motion vector (that is, whether they are motion consistent). Once again, this is done directly from the underlying intensity information within the two patches, without explicitly computing their motions, thus avoiding aperture problems that are so typical of small patches.

We say that the two ST-patches P_1 and P_2 are *motion consistent* if there exists a common vector $\mathbf{u} = [u\ v\ w]^T$ that

1. What we mean by "corresponding" is the following: If (i, j, k) is the location of a pixel within the segment S, then accordingly, the pixel at position (i, j, k) within segment T will be referred to as the "corresponding pixel," and the ST-patch centered around it would be the "corresponding patch."

satisfies (2) for both of them, that is, $\mathbf{G_1 u} = \mathbf{0}$ and $\mathbf{G_2 u} = \mathbf{0}$. Stacking these together, we get

$$\mathbf{G}_{12}\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{bmatrix}_{2n\times 3}\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{2n\times 1}, \quad (5)$$

where matrix $\mathbf{G_{12}}$ contains all the space-time intensity gradients from both ST-patches P_1 and P_2.

As before, we multiply both sides by $\mathbf{G_{12}^T}$, yielding

$$\mathbf{M}_{12}\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times 1}, \quad (6)$$

where $\mathbf{M_{12}} = \mathbf{G_{12}^T G_{12}}$ (the Gram matrix) is a 3×3 rank-deficient matrix: $rank(\mathbf{M_{12}}) \leq 2$.

Now, given two different space-time intensity patches, P_1 and P_2 (each induced by a single uniform motion), if the combined matrix $\mathbf{M_{12}}$ is *not rank deficient* (that is, $rank(\mathbf{M_{12}}) = 3 <=> \lambda_{min}(\mathbf{M_{12}}) \neq 0$), then these two ST-patches *cannot be motion consistent*.

Note that $\mathbf{M_{12}} = \mathbf{M_1} + \mathbf{M_2} = \mathbf{G_1^T G_1} + \mathbf{G_2^T G_2}$ and is based purely on the intensity information within these two ST-patches, avoiding explicit motion estimation.

Moreover, for our higher level purpose of space-time template correlation, we currently assumed that P_1 and P_2 are of the same size (n). However, in general, there is no such limitation in the above analysis.

4 Handling Spatio-Temporal Ambiguities

The rank-3 constraint on $\mathbf{M_{12}}$ *for detecting motion inconsistencies is a sufficient but not a necessary condition.* Namely, if $rank(\mathbf{M_{12}}) = 3$, then there is no single image motion that can induce the intensity pattern of both ST-patches P_1 and P_2 and, therefore, they are not motion consistent. However, the other direction is not guaranteed: There can be cases in which there is no single motion that can induce the two space-time intensity patterns P_1 and P_2, yet $rank(\mathbf{M_{12}}) < 3$. This can happen when each of the two ST-patches contains only a degenerate image structure (for example, an image edge) moving in a uniform motion. In this case, the space-time gradients of each ST-patch will reside on a line in the space-time volume, all possible (u, v, w) vectors will span a 2D plane in the space-time volume and, therefore, $rank(\mathbf{M_1}) = 1$ and $rank(\mathbf{M_2}) = 1$. Since $\mathbf{M_{12}} = \mathbf{M_1} + \mathbf{M_2}$, therefore, $rank(\mathbf{M_{12}}) \leq 2 < 3$, regardless of whether there is or is no motion consistency between P_1 and P_2.

The only case in which the rank-3 constraint on $\mathbf{M_{12}}$ is both sufficient and necessary for detecting motion inconsistencies is when both matrices $\mathbf{M_1}$ and $\mathbf{M_2}$ are each of rank-2 (assuming each ST-patch contains a single motion); namely—when both ST-patches P_1 and P_2 contain non-degenerate image features (corner-like).

In this section, we generalize the notion of the rank constraint on $\mathbf{M_{12}}$ *to obtain a sufficient and necessary motion-consistency constraint for both degenerate and nondegenerate ST-patches.*

If we examine all possible ranks of the matrix $\mathbf{M}$ of an individual ST-patch P, *which contains a single uniform motion*, then $rank(\mathbf{M}) = 2$ when P contains a corner-like image feature, $rank(\mathbf{M}) = 1$ when P contains an edge-like image feature, and $rank(\mathbf{M}) = 0$ when P contains a uniform-colored image region.

This information (about the *spatial* properties of P) is captured in the 2×2 upper left minor $\mathbf{M}^{\diamondsuit}$ of the matrix $\mathbf{M}$ (see (4))

$$\mathbf{M}^{\diamondsuit} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y \\ \Sigma P_y P_x & \Sigma P_y^2 \end{bmatrix}.$$

This is very similar to the matrix of the Harris detector [12], but the summation here is over the 3D ST-patch and not a 2D image patch.

In other words, for an ST-patch with a single uniform motion, the following rank condition holds: $rank(\mathbf{M}) = rank(\mathbf{M}^{\diamondsuit})$. Namely, when there is a single uniform motion within the ST-patch, the added temporal component (which is captured by the third row and the third column of $\mathbf{M}$) does not introduce any increase in rank.

This, however, does not hold *when an ST-patch contains more than one motion*, that is, when the motion is not along a single straight line. In such cases, the added temporal component introduces an increase in the rank, namely, $rank(\mathbf{M}) = rank(\mathbf{M}^{\diamondsuit}) + 1$. (The difference in rank cannot be more than 1, because only one column/row is added in the transition from $\mathbf{M}^{\diamondsuit}$ to $\mathbf{M}$.) Thus,

One patch. Measuring the *rank-increase* Δr between $\mathbf{M}$ and its 2×2 upper left minor $\mathbf{M}^{\diamondsuit}$ reveals whether the ST-patch P contains a single or multiple motions

$$\Delta r = rank(\mathbf{M}) - rank(\mathbf{M}^{\diamondsuit}) = \begin{cases} 0 & single\ motion \\ 1 & multiple\ motions. \end{cases} \quad (7)$$

Note that this is a generalization of the rank-3 constraint on $\mathbf{M}$, which was presented in Section 2. (When the rank $\mathbf{M}$ is 3, then the rank of its 2×2 minor is 2, in which case, the rank-increase is one.) In fact, "space-time corners" [16] can be seen as a special case of ST-patches with a rank-increase of one (see Appendix B). The constraint (7) holds both for degenerate and nondegenerate ST-patches.

Following the same reasoning for two different ST-patches (similar to the way the rank-3 constraint of a single ST-patch was generalized in Section 3 for two ST-patches), we arrive at the following sufficient and necessary condition for detecting motion inconsistency between two ST-patches:

Two patches. Measuring the *rank-increase* Δr between $\mathbf{M_{12}}$ and its 2×2 upper left minor $\mathbf{M_{12}^{\diamondsuit}}$ reveals whether the two ST-patches, P_1 and P_2, are motion consistent with each other:

$$\Delta r = rank(\mathbf{M}_{12}) - rank\left(\mathbf{M}_{12}^{\diamondsuit}\right) = \begin{cases} 0 & consistent \\ 1 & inconsistent. \end{cases} \quad (8)$$

This is a generalization of the rank-3 constraint on $\mathbf{M_{12}}$ presented in Section 3. In Appendix A, we prove constraint (8) and show that it holds both for degenerate and nondegenerate ST-patches.

5 Continuous Rank-Increase Measure Δr

The straightforward approach to estimate the rank-increase from $\mathbf{M}^{\diamondsuit}$ to $\mathbf{M}$ is to compute their individual ranks and then take the difference, which provides a binary value (0 or 1). The rank of a matrix is determined by the number of nonzero eigenvalues it has.

However, due to noise, eigenvalues are never zero. Applying a threshold to the eigenvalues is usually data dependent, and a wrong choice of a threshold would lead to wrong rank values. Moreover, the notion of motion consistency between two ST-patches (which is based on the rank-increase) is often not binary: If two motions are very similar but not identical—are they consistent or not? We would therefore like to have a *continuous measure* of motion consistency between two ST-patches. This motivated us to develop the following continuous notion of rank-increase.

Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of the 3×3 matrix $\mathbf{M}$. Let $\lambda_1^\diamondsuit \geq \lambda_2^\diamondsuit$ be the eigenvalues of its 2×2 upper left minor $\mathbf{M}^\diamondsuit$. From the *Interlacing Property* of eigenvalues in symmetric matrices ([11, p. 396]), it follows that $\lambda_1 \geq \lambda_1^\diamondsuit \geq \lambda_2 \geq \lambda_2^\diamondsuit \geq \lambda_3$. This leads to the following observations:[2]

$$\lambda_1 \geq \frac{\lambda_1 \cdot \lambda_2 \cdot \lambda_3}{\lambda_1^\diamondsuit \cdot \lambda_2^\diamondsuit} = \frac{det(\mathbf{M})}{det(\mathbf{M}^\diamondsuit)} \geq \lambda_3 \tag{9}$$

and

$$1 \geq \frac{\lambda_2 \cdot \lambda_3}{\lambda_1^\diamondsuit \cdot \lambda_2^\diamondsuit} \geq \frac{\lambda_3}{\lambda_1} \geq 0.$$

We define the continuous rank-increase measure Δr to be

$$\Delta r = \frac{\lambda_2 \cdot \lambda_3}{\lambda_1^\diamondsuit \cdot \lambda_2^\diamondsuit} \tag{10}$$

$0 \leq \Delta r \leq 1$. The case of $\Delta r = 0$ is an ideal case of no rank-increase, and when $\Delta r = 1$, there is a clear rank-increase. However, the above continuous definition of Δr allows to handle noisy data (without taking any threshold) and provides varying degrees of rank-increases for varying degrees of motion-consistencies.[3]

Fig. 8 in Appendix B shows the continuous rank-increase measure computed for *individual* ST-patches of a single sequence of a walking person. Note that within a *single* sequence, patches with high individual rank-increase measure (7) are sparse and correspond to space-time corners, local regions of motion discontinuity, and/or temporal aliasing (very fast motion). This is also the case when comparing two aligned sequences of the same action (that is, pairs of patches across the two sequences with high *joint* rank-increase measure (8) are sparse). However, when the two sequences are misaligned (or contain different actions), there are many pairs of patches with high *joint* rank-increase measure. This is employed in Section 6.

2. It is easy to show that the term $\frac{det(\mathbf{M})}{det(\mathbf{M}^\diamondsuit)}$ is the "pure temporal" eigenvalue that was derived in [17] using a Galilean diagonalization of the matrix $\mathbf{M}$. It is shown there that this diagonalization compensates for the local constant velocity and the "pure temporal" eigenvalue encodes information about the nonlinearity of the local motion.

3. Providing some intuition for this choice of Δr: In an "ideal" case, there are two clusters of eigenvalues—small ones and large ones, where the values within each cluster are similar, and there is a large gap between them (a ratio of several orders of magnitude, much larger than the ratios within each cluster). The large values are related to the contrast of the ST-patch and the small ones to the noise level. When there is no "clear" rank-increase, the denominator has one extra large eigenvalue than the numerator and the ratio in (10) is thus very small. For example, if $rank(\mathbf{M_{12}}) = 2$ and $rank(\mathbf{M_{12}^\diamondsuit}) = 2$, there will be two large values in the denominator and only one in the numerator. When there is a "clear" rank-increase, the numerator and denominator contain the same number of eigenvalues from both clusters, and the ratio thus tends to 1. In any other case, this ratio attains intermediate values. This observation was verified in our empirical evaluations.

6 Correlating a Space-Time Video Template

A space-time video template T consists of many small ST-patches. It is "correlated" against a larger video sequence by checking its consistency with every video segment centered around every space-time point (x, y, t) in the large video. A good match between the video template T and a video segment S should satisfy two conditions:

1. It should bring into "motion-consistent alignment" as many ST-patches as possible between T and S (that is, minimize their *joint* rank-increase measure (8).
2. It should maximize the alignment of motion discontinuities within the template T with motion discontinuities within the video segment S. Such discontinuities may also result from space-time corners and very fast motion. Namely, it should maximize the alignment of patches with high individual rank-increase measure (7), where the joint rank-increase test (8) will not apply.

A good global template match should minimize the number of local **in**consistent matches between the *linear patches* (with a single linear motion) and should also minimize the number of matches between linear patches in one sequence with *nonlinear patches* (with multiple motions or motion discontinuity) in the other sequence. The following measure captures the degree of *local **in**consistency* between a small ST-patch $P_1 \in T$ and an ST-patch $P_2 \in S$, according to the abovementioned requirements:

$$m_{12} = \frac{\Delta r_{12}}{min(\Delta r_1, \Delta r_2) + \epsilon}, \tag{11}$$

where ϵ avoids division by 0 (for example, $\epsilon = 10^{-5}$).

This measure yields low values (that is, "consistency") when P_1 and P_2 are motion consistent with each other (in which case, $\Delta r_{12} \approx \Delta r_1 \approx \Delta r_2 \approx 0$). It also provides low values when *both* P_1 and P_2 are patches located at motion discontinuities within their own sequences (in which case, $\Delta r_{12} \approx \Delta r_1 \approx \Delta r_2 \approx 1$). m_{12} will provide high values (that is, "**in**consistency") in all other cases.

Our empirical evaluations on both real and synthetic data show that for two ST-patches P_1 and P_2, $\Delta r_{12} \geq min(\Delta r_1, \Delta r_2)$. In addition, for two identical patches, the following holds $\Delta r_{11} = \Delta r_1$. This follows from (10) and the fact that the eigenvalues of a matrix multiplied by a scalar are multiplied by the same scalar. We therefore assume in our algorithm that the above measure is bounded below by 1, and the lowest measure is attained when a patch is compared to itself.

To obtain a *global **in**consistency measure* between the template T and a video segment S, the average value of m_{12} in T is computed: $\frac{1}{N}\Sigma m_{12}$, where N is the number of space-time points (and, therefore, also the number of ST-patches) in T. Similarly, a *global consistency measure* between the template T and a video segment S can be computed as the average value of $\frac{1}{m_{12}}$, where $0 \leq \frac{1}{m_{12}} \leq 1$. The autoconsistency of a patch (consistency with itself) is 1 and, therefore,

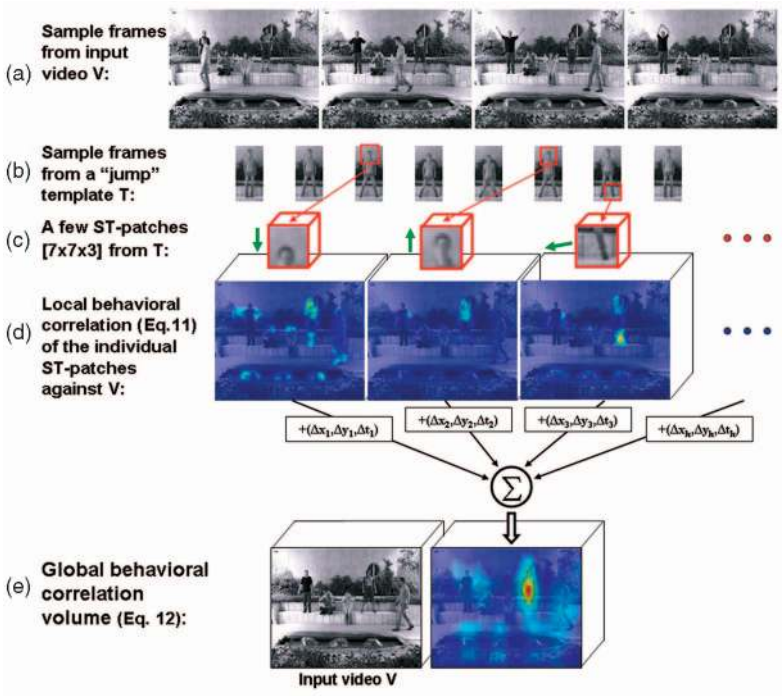

Fig. 2. From local patch consistency to global template correlation. This figure illustrates how the global template correlation volume is constructed from many local correlation volumes of small ST-patches that construct the template sequence. (a) A few sample frames from an input video $\mathbf{V}$ with multiple actions. (b) A few representative frames from a "jump" template $\mathbf{T}$. (c) Three small $[7 \times 7 \times 3]$ ST-patches were chosen for illustration. The direction of motion is shown by the green arrows. (d) The correlation volumes resulting from "correlating" (11) each of the ST-patches against the entire video $\mathbf{V}$. Each volume highlights regions that are consistent with the left/down/up motions of the patches correspondingly. (e) These volumes are shifted according to the patch locations inside the template (denoted by $+(\Delta x_k, \Delta y_k, \Delta t_k)$) and summed up to get the global template correlation volume (12). All false locations are pruned and we are left with the true location.

the global autocorrelation of a template (video segment) with itself is also 1.

6.1 Weighting the Gradients

We would like the consistency measure to be able to compare ST-patches of very different contrast. Suppose P_1 has much larger contrast than P_2 (larger space-time gradients), then the matrix $\mathbf{M_{12}}$ will be mostly affected by $\mathbf{M_1}$ and not much by $\mathbf{M_2}$. In order to make the consistency measure invariant to contrast, we normalize the gradients at each point by the mean magnitude of the local gradients (+ some small constant) in a small space-time window (we used the patch size). This is equivalent to replacing the matrix $\mathbf{M} = \mathbf{G^T G}$ by $\tilde{\mathbf{M}} = \mathbf{G^T W^{-1} G}$, where $\mathbf{W}$ is a diagonal matrix with the mean local gradient magnitudes on its diagonal.

6.2 Weighting the Patches

Not all patches in a template should necessarily have equal importance when their local consistency measure are summed into the global consistency measure. One may want to weight patch contributions differently, depending on the specific task at hand. For example, if the dynamic information is more important than the static one, more weight can be assigned to patches with high temporal derivatives or normal flow. One type of patches that should be given low weights in most practical situations are uniform patches (almost no derivatives). A uniform patch has high motion consistency with *any* other patch. This might lead to high behavioral correlation values in uniform textureless regions, which is usually not desired. We therefore applied a simple weighting function: $w_{12} = min(f(|\nabla P_1|), f(|\nabla P_2|))$, where f is a Sigmoid function, which gives a low weight if any of the patches is uniform with almost no space-time gradients (in our experiments, we use a Sigmoid function with a threshold of 15 gray levels and a width of 10 gray levels). The final template correlation score becomes

$$C(T,S) = \frac{\Sigma w_{12} \frac{1}{m_{12}}}{\Sigma w_{12}}, \tag{12}$$

which is the global consistency measure we used in our experiments. The above weighting function can be further generalized by giving more weight to spatially corner-like patches than to the more ambiguous edge-like patches. Fig. 2 illustrates the averaging process of ST-patches into one global consistency measure.

A space-time template T (for example, $30 \times 30 \times 30$) can thus be "correlated" against a larger video sequence (for example, $200 \times 300 \times 1,000$) by sliding it in all three dimensions (x, y, and t) while computing its consistency with the underlying video segment at every video location. This generates a space-time "correlation surface" (or more precisely, a space-time "correlation volume"). Peaks within this correlation volume are locations in the large video sequence where similar behavior to that depicted by the template is detected. To allow flexibility to small changes in scale and orientation, we correlate the template and the video at half of their original resolution. Examples of such correlation results can be found in Figs. 3, 4, and 5.

Fig. 3. Walking on the beach. (a) $T = a\ short\ walk\ clip$. (b) $V = the\ longer\ beach\ video\ against\ which\ T\ was\ "correlated."$ (c) Peaks of space-time correlation C superimposed on V (see text). For video sequences, see [30].

Fig. 4. Ballet example. (a) $T = a\ single\ turn\ of\ the\ man\text{-}dancer$. (b) $V = the\ ballet\ video\ against\ which\ T\ was\ "correlated."$ (c) Peaks of space-time correlation C superimposed on V (see text). For video sequences, see [30].

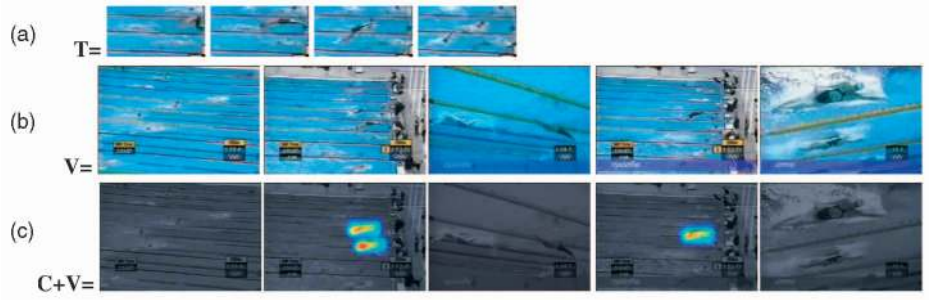

Fig. 5. Swim-relay match. (a) $T = a\ single\ dive\ into\ the\ pool$. (b) $V = the\ swim\text{-}relay\ video\ against\ which\ T\ was\ "correlated."$ (c) Peaks of the space-time correlation C superimposed on V (see text). For video sequences, see [30].

6.3 Computational Efficiency

In regular image correlation, the search space is 2D (the entire image). In the presented space-time correlation, the search space is 3D (the entire video sequence), and the local computations are more complex (for example, eigenvalue estimations). As such, special care must be taken of computational issues. The following observations allow us to speedup the space-time correlation process significantly:

1. The local matrices $\mathbf{M}_{3\times3}$ (4) can be computed and stored ahead of time for all pixels of all video sequences in the database and, separately, for the space-time templates (the video queries). The only matrices that need to be estimated online during the space-time correlation process are the combined matrices $\mathbf{M_{12}}$ (6), which result from comparing ST-patches in the template with ST-patches in a database sequence. This, however, does not require any new gradient estimation during runtime, since $\mathbf{M_{12}} = \mathbf{M_1} + \mathbf{M_2}$ (see end of Section 3).
2. Eigenvalue estimation, which is part of the rank-increase measure (10), is computationally expensive when applied to $\mathbf{M_{12}}$ at every pixel. The following observations allow us to approximate the rank-increase measure without resorting to eigenvalue computation.

   $det(\mathbf{M}) = \lambda_1 \cdot \lambda_2 \cdot \lambda_3$, and $det(\mathbf{M}^{\diamondsuit}) = \lambda_1^{\diamondsuit} \cdot \lambda_2^{\diamondsuit}$. The rank-increase measure of (10) can be rewritten as

$$\Delta r = \frac{\lambda_2 \cdot \lambda_3}{\lambda_1^{\diamondsuit} \cdot \lambda_2^{\diamondsuit}} = \frac{det(\mathbf{M})}{det(\mathbf{M}^{\diamondsuit}) \cdot \lambda_1}.$$

   Let $\|\mathbf{M}\|_F = \sqrt{\sum M(i,j)^2}$ be the Frobenius norm of the matrix $\mathbf{M}$. Then, the following relation holds between $\|\mathbf{M}\|_F$ and λ_1 [11]:

$$\lambda_1 \leq \|\mathbf{M}\|_F \leq \sqrt{3}\lambda_1.$$

   The scalar $\sqrt{3}(\approx 1.7)$ is related to the dimension of $\mathbf{M}(3\times3)$. The rank-increase measure Δr can therefore be approximated by

$$\Delta\hat{r} = \frac{det(\mathbf{M})}{det(\mathbf{M}^{\diamondsuit}) \cdot \|\mathbf{M}\|_F}. \tag{13}$$

   $\Delta\hat{r}$ requires no eigenvalue computation, is easy to compute from $\mathbf{M}$, and provides the following bounds on the rank-increase measure Δr of (10): $\Delta\hat{r} \leq \Delta r \leq \sqrt{3}\Delta\hat{r}$. Although less precise than Δr, $\Delta\hat{r}$ provides sufficient separation between "rank-increases" and "no rank-increases." We use this approximated measure to speedup our space-time correlation process.
3. If we were to compute the entire correlation volume,[4] then the overall runtime of a $144 \times 180 \times 200$ video sequence and a $50 \times 25 \times 20$ query would be

4. Since the correlation volume is smooth, it is enough to compute it for every other pixel in all frames and then interpolate.

30 minutes on a Pentium 4 3.0 GHz processor using our nonoptimized implementation (mostly in Matlab). However, since we are searching only for correlation peaks, it is not necessary to estimate the entire correlation volume, and the process can be significantly sped up using a coarse-to-fine search.

This is done by constructing space-time Gaussian pyramids [6] from the original sequence V and the template T. Namely, each coarser pyramid level is generated by blurring and subsampling the previous pyramid level *both in time and in space*. A full search is performed in the coarsest resolution level to find several peaks of behavior correlation above some predefined threshold. The top K peaks are propagated to the next higher space-time resolution level, and a new search is performed only in a small space-time region around each peak to refine the locations.[5] This search process proceeds similarly to the next levels until the final search in the finest resolution level yields the exact locations of the highest correlation peaks in the original sequence.

Another speedup relates to the number of patches in the template that are computed and contribute to the final correlation score (12). Instead of taking overlapping patches around all pixels in the template T (and their matching patches from the video segment S), it is possible to take only a subset of patches that represent the template well enough. This subset is chosen in a sparse space-time grid of locations in the template.

Complexity reduction. Let N be the size (in pixels) of the video V, M is the size of the template T, R is the reduction in the size of the coarsest resolution level in the space-time pyramid with L levels relative to the original level, r is the reduction of pixels in the sparse grid in the template, K is the maximal number of peaks propagated from one pyramid level to the next, and W is the size of the neighborhood. The complexity of exhaustively computing the full "correlation volume" using all template patches in the finest level is $O(NM)$. In contrast, the complexity of the coarse-to-fine search described above is $O(NM/(R^2r^2) + LKW)$. We found that the following parameters gave adequate results: $R = 8$, $r = 4$, $L = 2$, $W = 10^3$, and $K = 10$. For finding 10 highest peaks in the above example ($144 \times 180 \times 200$ video and $50 \times 25 \times 20$ query), we got a reduction of more than two orders of magnitude in the computational complexity, thus *reducing the search time*[6] *to 10 seconds*, which is close to real time (equivalent to 20 frames/sec), using our nonoptimized implementation. In general, coarse-to-fine search algorithms have some probability of misdetection. However, in our experiments, we found this probability to be very low.

In order to reduce effects of temporal aliasing due to fast motion, the video sequences were first spatially blurred. Spatial Gaussian blurring of size $[5 \times 5]$ with $\sigma = 0.8$ was applied to the two input sequences ($\mathbf{V}$ and $\mathbf{T}$) prior to processing. The size of the ST-patches P (4) was $[7 \times 7 \times 3]$, using weighted sums of gradients with Gaussian weights ($\sigma_{space} = 1.5$ and $\sigma_{time} = 0.8$) instead of regular sums.

5. This refinement search can also be done in a cascade of grids fashion—by searching first in a sparse grid of space-time locations (for example, every three pixels and every two frames) and then refine the search on a denser grid of locations around the highest peaks.

6. The preprocessing time of the large video sequence took additional 15 seconds, but it is computed only once in case of multiple queries and can be done offline in some applications.

7 Results

One possible application of our space-time correlation is to detect "behaviors of interest" in a video database. A behavior of interest can be defined via one (or more) example video clip (a "video query"). Such video queries serve as space-time correlation templates, which are matched against a different (typically longer and larger) video sequence. Our approach seeks for video locations with similar underlying motion fields (both of the figure and of the background) without segmentation or motion estimation. Please view the video clips (databases and queries) of the following experiments at www.wisdom.weizmann.ac.il/~vision/BehaviorCorrelation.html [30].

Fig. 3 shows the results of applying our method to detect all instances of walking people in a beach video. The space-time template T was a very short walk clip (14 frames of 60×70 pixels) of a different man recorded elsewhere. Fig 3a shows a few sampled frames from T. Fig 3b shows a few sampled frames from the long beach video V (460 frames of 180×360 pixels). The template T was "correlated" twice with V—once as is, and once its mirror reflection, to allow detections of walks in both directions. Fig. 3c shows the peaks of the resulting space-time correlation surface (volume) $C(x, y, t)$ superimposed on V. *Red* denotes highest correlation values; *blue* denotes low correlation values. Different walking people with different clothes and different backgrounds were detected. Note that *no background-foreground segmentation* was required. The behavioral-consistency between the template and the underlying video segment is *invariant to the differences in spatial appearance* of the foreground moving objects and of their backgrounds. It is sensitive only to the underlying motions.

Fig. 4 shows the analysis of a ballet footage downloaded from the Web ("Birmingham Royal Ballet"). The space-time template T contains a single turn of a man dancer (13 frame of 90×110 pixels). Fig. 4a shows a few sampled frames from T. Fig. 4b shows a few frames from the longer ballet clip V (284 frames of 144×192 pixels), against which T was "correlated." Peaks of the space-time correlation volume C are shown superimposed on V (Fig. 4c). Most of the turns of the two dancers (a man and a woman) were detected, despite the variability in spatial scale relative to the template (up to 20 percent, see more details about robustness to variations in Section 8). Note that this example contains very fast moving parts (frame to frame).

Fig. 5 shows the detecting dives into a pool during a swimming relay match. This video was downloaded from the Web site of the 2004 Olympic Games and was severely MPEG compressed. The video query T is a short clip (70×140 pixels $\times 16$ frames) showing one dive (shown slightly enlarged in Fig 5a for visibility). It was correlated against the one-minute long video V (757 frames of 240×360 pixels, Fig 5b). Despite the numerous simultaneous activities (a variety of swim styles, flips under the water, and splashes of water) and despite the severe noise, the space-time correlation was able to separate most of the dives from other activities (Fig 5c). One dive is missed due to partial occlusion by the Olympic logo at the bottom right of the frame. There is also one false detection, due to a similar motion pattern occurring in the water. It is unlikely to assume that optical flow estimation would produce anything meaningful on such a noisy sequence, with so much background clutter, splashing water, and so forth. Also, it is unlikely that any segmentation method would be able to separate foreground and background objects

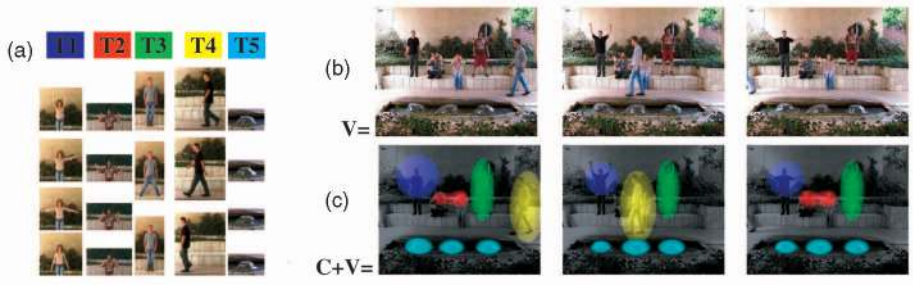

Fig. 6. Detecting multiple activities. (a) $T_1, \ldots T_5 = five\ different\ short\ video\ templates.$ (b) $V = the\ video\ against\ which\ T\ was\ "correlated."$ (c) Ellipses with colors corresponding to the five activities are displayed around the peaks detected in all five correlation surfaces $C_1, \ldots, C_5$ (see text). For video sequences, see [30].

here. Nevertheless, the space-time correlation method was able to produce reasonable results.

Fig. 6 shows the detection of five different activities that occur simultaneously: "walk," "wave," "clap," "jump," and "fountain" (with flowing water). Five small video queries were provided $(T_1, \ldots, T_5)$, one for each activity (Fig 6a). These were performed by different people and backgrounds than in the longer video V. A short subclip from the rightmost fountain was used as the fountain query T_5. Fig. 6c shows the peaks detected in each of the five correlation volumes $C_1, \ldots, C_5$. Space-time ellipses are displayed around each peak, with its corresponding activity color. All activities were correctly detected, including the flowing water in all three fountains.

In all the above examples, a threshold was applied highlighting the peaks. The threshold was chosen to be 0.7-0.8 of the highest peak value detected. In these various examples, it is evident that the correlation volume behaves smoothly around the peaks. The size of the basin of attraction occupied about half the size of the human figure, and the peak in each basin was usually unique. These properties enable us to use efficient optimization tools when searching for the maxima (as was suggested at the end of Section 6).

8 Robustness to Local and Global Deformations

Although not invariant to changes in scale and rotation, our method is quite robust to some degree of space-time deformations between the video segments. Deformations can be of two types: Global parametric variations, which include spatial scaling, temporal scaling (changes in the speed of an action), slight rotations, and affine shears (due to different viewpoints). Local nonparametric variations are due to the different shapes of people and deviations in their performances of the same actions.

In order to evaluate the robustness of our approach to such variations, we performed the following empirical tests. We took a short clip of a walking person and correlated it against other clips of the same person walking but with different clothes and background, as well as to other walking people. In addition, the "Walk" clip was matched also to clips of other actions of the same person and other people (most of the clips were taken from the database in [3]). These clips included actions such as "Run," "Jump aside," "Skip," "Jump ahead" (all in the same direction of the "walk"), and "Nonhorizontal" actions including "Jumping jack" and "Waving" actions. The last video clip contained "Sea waves" that were moving horizontally at the speed of the reference "Walk" action. After computing the measures for the original clips with respect to the reference "Walk" clip, a series of simple parametric deformations were applied to each of these clips, each time evaluating their measure with respect to the reference clip. These transformations included: shifts in x, y, and t, spatial scale, temporal scale, and rotations. After each global scaling or rotation transformation, an exhaustive search was performed to find the best matching score between the reference "Walk" sequence and the transformed sequence.

The final scores for all of these tests are summarized in the graphs in Fig. 7. These graphs show that the "Walk" actions (marked in blue) can be easily distinguished from the other actions. The results show that our method is robust to a range of deformations of the action clip (of size[7] $45 \times 80 \times 40$): a vertical shift of up to 10 percent of the template height (8 pixels), a horizontal shift of up to 11 percent of template width (5 pixels), 10 percent shift in time (4 frames), 10 percent spatial scale, 10 percent temporal scale, and a rotation of at least 10 degrees. These values are determined for the worst case, that is, for the most similar action to the deformed "Walk" action.

There are several reasons for this robustness: First, this method relies on the correspondence of the average motions extracted from small *regions* (ST-patches) and not on *pixel-wise* correspondence between the structure within the patches (as assumed, for example, in [25]). The second reason is that natural motion fields are smooth and contain lower frequencies than the intensity information itself. The third reason is that we do not compare the true flow fields but measure a continuous measure for the possible motion-consistencies. Degenerate patches (for example, edges) can match a variety of space-time patterns with a common motion component so their contribution to the global template measure may not change significantly by small local shifts. These reasons allow for matching templates with substantial differences between them without the need to accurately align them beforehand (as was suggested in [25]).

The robustness of the method to local and global variations can be further improved by several simple modifications to the current method. For example, instead of checking motion consistency between two corresponding patches from the two videos, a patch from one sequence can be matched to *multiple* patches within a small neighborhood in the second sequence, seeking for the best local match among them.

7. The action clip was generated by selecting a 45×80 window surrounding the walking person for a duration of 40 frames. This generates a diagonal parallelogram in space time, that is contained in a bounding space-time box of $130 \times 80 \times 40$.

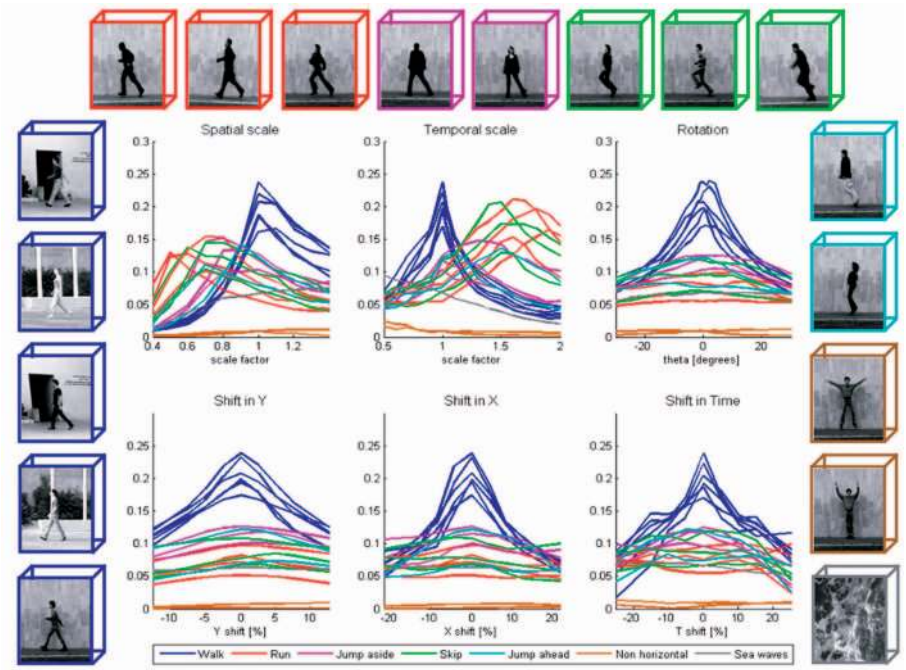

Fig. 7. Robustness to small global transformations. These graphs show the behavioral correlation measure of a reference video clip of a "Walk" action (lower left blue-framed sequence) correlated with other clips of different actions that were deformed using various parametric transformations. The other clips contain additional "Walk" actions of different people, clothes, backgrounds (all prealigned to the reference one), as well as other actions: "Run," "Jump aside," "Skip," "Jump ahead," and "Nonhorizontal" actions including "jumping jack" and "waving" actions. The last video clip contains "Sea waves" that are moving horizontally at the speed of the reference action. Each group of actions was color coded with the same color (see sample sequences surrounding the graphs). The x-axis of the bottom translation graphs is measured as the percentage from the size of the reference template that was $130 \times 80 \times 40$ in our example. The maximal possible correlation value is 1 (the autocorrelation of any sequence with itself). Note that some of the other actions (mainly, the "Run" and "Skip") obtain similar correlation values to the "Walk" actions, when they are scaled in time, meaning that, in many cases, a clip of a walking person cannot be distinguished from a *slowed down* clip of a running or a skipping person.

9 Conclusion

By examining the intensity variations in video patches, we can implicitly characterize the space of their possible motions. This is done without having to explicitly commit to a particular choice of flow vectors (which are likely to be erroneous for complex dynamic scenes). This allows us to identify whether two different space-time intensity patterns in two different video segments could have been induced by similar underlying motion fields. We use this to compare ("correlate") small video templates against large video sequences in order to detect all locations with similar dynamic behaviors, whereas being invariant to appearance, and without prior foreground/background segmentation. To our best knowledge, this is the first time multiple different behaviors/actions occurring simultaneously in the field of view are detected and in very complex dynamic scenes. Currently, our method is not invariant to large geometric deformations of the video template. However, it is *not* sensitive to small changes in scale and orientation and can be extended to handle large changes in scale by employing a multiscale framework (in space and in time). This is part of our future work.

Appendix A

Proof of the Rank-Increase Constraint

In Section 4, we stated an *iff* consistency constraint between two ST-patches P_1 and P_2. We would like to prove that there exists a common motion vector $\mathbf{u} = [u\ v\ w]^T$ to P_1 and P_2 that satisfies (6) *iff*

$$\Delta r = rank(\mathbf{M}_{12}) - rank(\mathbf{M}_{12}^{\diamondsuit}) = 0. \tag{14}$$

Note that *not* every vector $[u\ v\ w]^T$ corresponds to a valid motion vector: only vectors for which $w \neq 0$. The vector $[u\ v\ 0]^T$ corresponds to an *infinite* motion, whereas $w \neq 0$ to a *finite* motion, and $\begin{bmatrix}\tilde{u}\\ \tilde{v}\end{bmatrix} = \begin{bmatrix}u/w\\ v/w\end{bmatrix}$ is the common physical flow field vector.

Therefore, for *consistent* motions, we can always write

$$\mathbf{M}_{12}\begin{bmatrix}\tilde{u}\\ \tilde{v}\\ 1\end{bmatrix} = \mathbf{0}. \tag{15}$$

We will next show how this leads to the consistency constraint in (14), and vice versa (both directions).

A.1 Consistent Motions (15) $\Rightarrow \Delta r = 0$

Let $\mathbf{C_1}$, $\mathbf{C_2}$, and $\mathbf{C_3}$ denote the columns of M_{12}, then we can rewrite (15) as

$$\mathbf{M}_{12}\begin{bmatrix}\tilde{u}\\ \tilde{v}\\ 1\end{bmatrix} = \left[\mathbf{C_1}\middle|\,\mathbf{C_2}\middle|\,\mathbf{C_3}\right]\begin{bmatrix}\tilde{u}\\ \tilde{v}\\ 1\end{bmatrix} = \mathbf{0}.$$

Thus, the third column $\mathbf{C_3}$ is a linear combination of the first two columns.

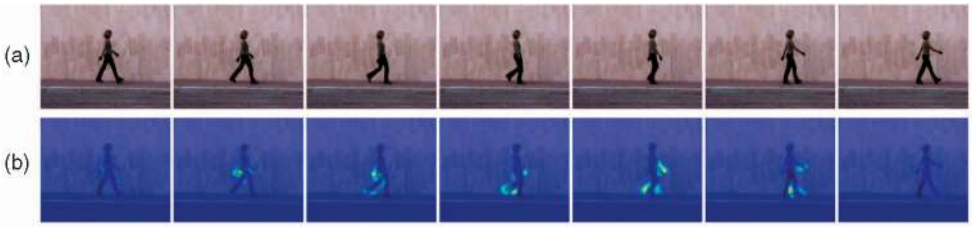

Fig. 8. The continuous rank-increase measure (single sequence). (a) Representative frames from a walk sequence. (b) The continuous rank-increase measure is computed for each individual ST-patch in the sequence and is overlayed on the sequence. High values (yellow-red) are obtained in regions where hands and legs motion is changing direction, regions of motion discontinuities (hands intersect the body, legs cross), or motion aliasing where the speed of limbs is large.

Due to the symmetry of $\mathbf{M_{12}}$

$$\begin{bmatrix} & \mathbf{M}_{12}^{\diamondsuit} & & m_{31} \\ & & & m_{32} \\ m_{31} & m_{32} & & m_{33} \end{bmatrix} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \mathbf{0}.$$

The third row $[m_{31}\ m_{32}\ m_{33}]$ (which is also equal to the third column $\mathbf{C}_3^T$) is therefore a linear combination of the first two rows of $\mathbf{M_{12}}$ (which are equal to $\mathbf{C}_1^T$ and $\mathbf{C}_2^T$). In particular, $[m_{31}\ m_{32}]$ is spanned by the rows of $\mathbf{M}_{12}^{\diamondsuit}$.

Therefore,

$$\begin{aligned} &rank(\mathbf{M}_{12}) = \\ &dim(span\{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\}) = dim(span\{\mathbf{C}_1, \mathbf{C}_2\}) \\ &= rank\left(\begin{bmatrix} \mathbf{M}_{12}^{\diamondsuit} \\ m_{31}\ m_{32} \end{bmatrix}\right) = rank\left(\mathbf{M}_{12}^{\diamondsuit}\right). \end{aligned}$$

□

A.2 $\Delta r = 0 \Rightarrow$ Consistent Motions (15)

Reversing the above reasoning—if $rank(\mathbf{M_{12}}) = rank(\mathbf{M}_{12}^{\diamondsuit})$, then the third column in $\mathbf{M_{12}}$ is spanned by the first two columns. Namely, $\exists a, b$ such that $\mathbf{C_3} = a\mathbf{C_1} + b\mathbf{C_2}$. Therefore,

$$\mathbf{M_{12}} \begin{bmatrix} -a \\ -b \\ 1 \end{bmatrix} = \mathbf{0},$$

that is, (15) is satisfied. □

Appendix B

Rank-Increase Patches—A Generalization of "Space-Time Corners"

In this appendix, we show the relation between the rank-increase measure and "space-time corners" [15], [16]. Fig. 8 shows the continuous rank-increase measure (Δr in (13)) computed for *individual* ST-patches of a sequence of a walking person. The computed measure is overlayed on top of the sequence. For ST-patches with a single motion (the upper body parts), the rank-increase values are very low. The high values are obtained in ST-patches containing more than one motion, for example, where hands and legs motion is changing its direction, in regions of motion discontinuities (hands intersect the body, legs cross) or in ST-patches with motion aliasing where the speed of limbs is large, relative to the camera frame rate. These patches are characterized by $rank(\mathbf{M}^{\diamondsuit}) = rank(\mathbf{M}) - 1$ (see (7)). "Space-time corners" are spatial corners whose direction of motion changes. Therefore, for "space-time corners," $rank(\mathbf{M}^{\diamondsuit}) = 2$ and $rank(\mathbf{M}) = 3$. Thus, ST-patches with high individual rank-increase are a generalization of the "space-time corners" as they include both spatial corner-like ST-patches with $rank(\mathbf{M}^{\diamondsuit}) = 2$ and also edge-like ST-patches with $rank(\mathbf{M}^{\diamondsuit}) = 1$, both with more than one linear motion.

The "space-time corners" were shown to be informative for action recognition as they appear at the same space-time locations in similar actions. ST-patches with high individual rank-increase can thus serve as *denser* features for action recognition in similar conditions to the ones in [16]. However, although denser, ST-patches with high rank-increase still suffer from the same disadvantages. First, they lack information about the direction of the motion as "space-time corners" do. Only by looking at the joint matrix $\mathbf{M_{12}}$ constructed from *two* different ST-patches can we capture the consistency between the motions in the two patches. Second, the rank-increase regions and the "space-time corners" depend on the background. When the background is textured, many ST-patches of high rank-increase will emerge on the contour of the moving body due to foreground-background motion discontinuities (this was exploited in [10] for motion segmentation). Third, these ST-patches depend on the speed of motion relative to the frame rate. Above some velocity value starts motion aliasing that raises the rank-increase. Fourth, these ST-patches are still too sparse. Moreover, there are actions with very few or no ST-patches with rank-increase (for example, an action of a person moving circularly his arm). See [14] for more details about these cases.

In this work, we *do not* use the rank-increase of individual patches for generating features. Our method employs patches with *low* rank-increase measure, and we measure consistency of *pairs* of patches by examining the rank-increase measure of their joint matrix.

Acknowledgments

The authors would like to thank Yoni Wexler, Oren Boiman, and the anonymous reviewers for their useful remarks on the first drafts of the paper. This work was supported in part by the Israeli Science Foundation (Grant 267/02) and by the Moross Laboratory at the Weizmann Institute of Science.

References

[1] J. Bigün and G. Granlund, "Optical Flow Based on the Inertia Matrix of the Frequency Domain," *Proc. Swedish Symp. Image Analysis (SSAB) Symp. Picture Processing*, Mar. 1988.

[2] M.J. Black, "Explaining Optical Flow Events with Parameterized Spatio-Temporal Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 1326-1332, June 1999.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *Proc. Int'l Conf. Computer Vision,* pp. 1395-1402, Oct. 2005.
[4] A. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, Mar. 2001.
[5] C. Bregler, "Learning and Recognizing Human Dynamics in Video Sequences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1997.
[6] Y. Caspi and M. Irani, "A Step towards Sequence-to-Sequence Alignment," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 682-689, June 2000.
[7] O. Chomat and J.L. Crowley, "Probabilistic Sensor for the Perception of Activities," *Proc. European Conf. Computer Vision,* 2000.
[8] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," *Proc. Second Joint IEEE Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* Oct. 2005.
[9] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. Int'l Conf. Computer Vision,* Oct. 2003.
[10] D. Feldman and D. Weinshall, "Motion Segmentation Using an Occlusion Detector," *Proc. ECCV Workshop Dynamical Vision,* 2006.
[11] G. Golub and C.V. Loan, *Matrix Computations,* third ed. Johns Hopkins Univ. Press, 1996.
[12] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Proc. Fourth Alvey Vision Conf.,* pp. 147-151, 1988.
[13] B. Jähne, H. Haußecker, and P. Geißler, *Handbook of Computer Vision and Application,* vol. 2. Academic Publishers, 1999.
[14] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," *Proc. Int'l Conf. Computer Vision,* pp. 166-173, 2005.
[15] I. Laptev, "On Space-Time Interest Points," *Int'l J. Computer Vision,* vol. 64, no. 2/3, pp. 107-123, 2005.
[16] I. Laptev and T. Lindeberg, "Space-Time Interest Points," *Proc. Int'l Conf. Computer Vision,* 2003.
[17] T. Lindeberg, A. Akbarzadeh, and I. Laptev, "Galilean-Diagonalized Spatio-Temporal Interest Operators," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 57-62, 2004.
[18] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Proc. Image Understanding Workshop,* pp. 121-130, 1981.
[19] J.C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Proc. British Machine Vision Conf.,* 2006.
[20] S.A. Niyogi and E.H. Adelson, "Analyzing and Recognizing Walking Figures in xyt," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1994.
[21] C. Rao, A. Yilmaz, and M. Shah, "View Invariant Representation and Recognition of Actions," *Int'l J. Computer Vision,* vol. 50, no. 2, pp. 203-226, 2002.
[22] E. Shechtman and M. Irani, "Space-Time Behavior Based Correlation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 405-412, June 2005.
[23] H. Spies and H. Scharr, "Accurate Optical Flow in Noisy Image Sequences," *Proc. Int'l Conf. Computer Vision,* vol. 1, pp. 587-592, July 2001.
[24] J. Sullivan and S. Carlsson, "Recognizing and Tracking Human Action," *Proc. European Conf. Computer Vision,* vol. 1, pp. 629-644, 2002.
[25] Y. Ukrainitz and M. Irani, "Aligning Sequences and Actions by Maximizing Space-Time Correlations," *Proc. European Conf. Computer Vision,* May 2006.
[26] Y. Yacoob and M.J. Black, "Parametrized Modeling and Recognition of Activities," *Computer Vision and Image Understanding,* vol. 73, no. 2, pp. 232-247, 1999.
[27] A. Yilmaz and M. Shah, "Actions Sketch: A Novel Action Representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 984-989, 2005.
[28] L. Zelnik-Manor and M. Irani, "Event-Based Analysis of Video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 123-130, Sept. 2001.
[29] L. Zelnik-Manor and M. Irani, "Statistical Analysis of Dynamic Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 9, pp. 1530-1535, Sept. 2006.
[30] www.wisdom.weizmann.ac.il/~vision/BehaviorCorrelation.html, 2007.

Eli Shechtman received the BSc degree in electrical engineering magna cum laude from Tel-Aviv University in 1996, and the MSc degree in mathematics and computer science from the Weizmann Institute of Science in 2003. He is currently finishing his PhD studies at the Weizmann Institute of Science, working on similarity measures in image and video data and their applications. He is about to start his postdoctoral studies jointly at the University of Washington and at Adobe Systems Inc. He was awarded the Weizmann Institute Dean's prize for MSc students, and received several awards for his PhD research work, including the J.F. Kennedy award (the highest award at the Weizmann Institute) and the Israeli Knesset (Israeli parliament) outstanding student award. He received the best paper award at European Conference on Computer Vision (ECCV) 2002 and a best poster award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004. He is a student member of the IEEE Computer Society.

Michal Irani received the BSc degree in mathematics and computer science in 1985 and the MSc and PhD degrees in computer science in 1989 and 1994, respectively, all from the Hebrew University of Jerusalem. From 1993 to 1996, she was a member of the technical staff in the Vision Technologies Laboratory at the David Sarnoff Research Center, Princeton, New Jersey. She is currently an associate professor at the Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Israel. Her research interests include computer vision and video information analysis. Her prizes and honors include the David Sarnoff Research Center Technical Achievement Award (1994), the Yigal Allon three-year fellowship for outstanding young scientists (1998), and the Morris L. Levinson Prize in Mathematics (2003). She also received the best paper awards at the European Conference on Computer Vision (ECCV) 2000 and 2002, the honorable mention for the Marr Prize at the IEEE Int'l Conference on Computer Vision (ICCV) 2001 and 2005, and a best poster award at the Computer Vision and Pattern Recognition (CVPR) 2004. She served as an associate editor of the *Transactions on Pattern Analysis and Machine Intelligence* in 1999-2003. She is a member of the IEEE.