Space-Time Behavior Based Correlation

Eli Shechtman Michal Irani
Dept. of Computer Science and Applied Math
The Weizmann Institute of Science
76100 Rehovot, Israel

Abstract

We introduce a behavior-based similarity measure which 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 2-dimensional image correlation into the 3-dimensional space-time volume, thus allowing to correlate dynamic behaviors and actions. Small space-time video segments (small video clips) are "correlated" against 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 (e.g., ballet movements, pool dives, running water), even when multiple complex activities occur simultaneously within the field-of-view of the camera.

1. Introduction

Different people with similar behaviors induce completely different space-time intensity patters 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 same behaviors is the underlying induced motion fields. This observation was used in [5], where low-pass filtered optical-flow fields (between pairs of frames) were used for action recognition. However, dense unconstrained and non-rigid 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, etc.) Uniform-colored clothes induce local aperture effects, especially when the observed acting person is large (which is why Efros et. al [5] 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, etc.

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 (grayscale) 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, etc.

Our behavioral similarity measure requires *no* prior foreground/background segmentation (which is often required in action-recognition methods, e.g., [2, 13]). It requires no prior modelling or learning of activities, and is therefore *not* restricted to a small set of predefined activities (as opposed to [14, 1, 3, 4, 2]). While [5, 14, 1] 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, etc.) 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, etc. 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 2-dimensional image correlation, into a 3-dimensional space-time video-template correlation. The behavioral similarity measure is used for “correlating” a small “video query” (a small video clip of an action) against a large video sequence in all three dimensions (x,y,t), for detecting all video locations with high behavioral similarity.

Space-time approaches to action recognition have been previously suggested [11, 15, 9], which also perform direct measurements in the space-time intensity video volume. Slices of the space-time volume (such as the X-T plane) were used in [11] for gait recognition. This approach exploits only a small portion of the available data, and is limited for cyclic motions. In [15] empirical distributions of space-time gradients collected from an entire video clip are used. As such, they 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 [9] 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. It is therefore also limited to a single action in the field-of-view of the camera.

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. To our best knowledge, this is the first work which 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.

Overview of the Approach & 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, e.g., $30 \times 30 \times 30$) is “correlated” against a larger video sequence V (e.g., $200 \times 300 \times 1000$) 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 surface are locations in the video sequence V with similar behavior to the template T.

Each value in the correlation surface $C(x, y, t)$ 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 *space-time patches* (e.g., $7 \times 7 \times 3$) within these video segments. Namely, for each point $(x, y, t) \in S$, a small space-time patch (ST-patch) $P_S \subset S$ centered around (x, y, t) is compared against its corresponding small space-time 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 sometime performed. However, here the small patches P also have a temporal dimension, and the similarity measure between patches captures similarity of the implicit underlying motions.)

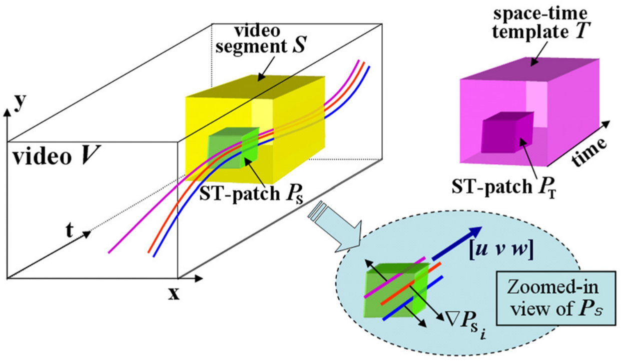

Figure 1. Overview of framework and notations.

We will start by exploring unique properties of intensity patterns induced in small space-time patches P within video data (Section 2). Step by step, we will develop the consistency measure between two such space-time 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 surface of the video query T relative to the entire large video sequence V (Section 6). Examples of detecting complex activities (pool dives, ballet dances, etc.) in 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. For short, we will refer to a small space-time patch as **ST-patch**. If a ST-patch P is small enough (e.g., $7 \times 7 \times 3$), 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 [10] 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 ($i = 1..n$), 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:

$$\nabla P_i \begin{bmatrix} u \\ v \\ w \end{bmatrix} = 0 \qquad (1)$$

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

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

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

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

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

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

$$\mathbf{M} = \mathbf{G^T G} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y & \Sigma P_x P_t \\ \Sigma P_y P_x & \Sigma P_y^2 & \Sigma P_y P_t \\ \Sigma P_t P_x & \Sigma P_t P_y & \Sigma P_t^2 \end{bmatrix} . \qquad (4)$$

where the summation is over all pixels within the space-time patch. Therefore, for all small space-time 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 has also been used in the past for optical flow estimation [8], non-linear filtering [12] and extraction of space-time interest points [9].

Now, if there exists a ST-patch for which $rank(\mathbf{M}) = 3$, then this ST-patch cannot contain a single uniform motion (i.e., 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" [9] or patches of "no coherent motion" [8]. These patches are typically very 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 (i.e., 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 two ST-patches P_1 and P_2 are *motion consistent* if there exists a common vector $\mathbf{u} = [u\ v\ w]^T$ that satisfies Eq. 2 for both them, i.e., $\mathbf{G_1 u} = \mathbf{0}$ and $\mathbf{G_2 u} = \mathbf{0}$. Stacking these together we get:

$$\mathbf{G_{12}} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \mathbf{G_1} \\ \mathbf{G_2} \end{bmatrix}_{2n\times 3} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{2n\times 1} \qquad (5)$$

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:

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

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* (i.e., $rank(\mathbf{M_{12}}) = 3 <=> \lambda_{min}(\mathbf{M_{12}}) \neq 0$), 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). But 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 which 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 which 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 space-time patches contains only a degenerate image structure (e.g., 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 isn't 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 & necessary motion-consistency constraint for both degenerate & non-degenerate 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, $rank(\mathbf{M}) = 0$ 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 Eq. 4):

$$\mathbf{M}^\diamondsuit = \left[\begin{array}{cc} \Sigma P_x^2 & \Sigma P_x P_y \\ \Sigma P_y P_x & \Sigma P_y^2 \end{array} \right].$$

This is very similar to the matrix of the Harris detector [7], but the summation here is over the 3-dimensional space-time patch, and not a 2-dimensional image patch.

In other words, for a 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 third column of $\mathbf{M}$) does not introduce any increase in rank.

This, however, does not hold *when a ST-patch which contains more than one motion*, i.e., 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:

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

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 1). The constraint (7) holds both for degenerate and non-degenerate ST-patches.

Following the same reasoning for two different ST-patches (similarly 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 $\mathbf{M}_{\mathbf{12}}^\diamondsuit$ reveals whether two ST-patches, P_1 and P_2, are motion-consistent with each other:

$$\Delta r = rank(\mathbf{M_{12}}) - rank(\mathbf{M}_{\mathbf{12}}^\diamondsuit) = \begin{cases} 0 & \textit{consistent} \\ 1 & \textit{inconsistent} \end{cases} \tag{8}$$

This is a generalization of the rank-3 constraint on $\mathbf{M_{12}}$ presented in Section 3. The constraint (8) holds both for degenerate and non-degenerate 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 values (0 or 1). The rank of a matrix is determined by the number of non-zero 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-

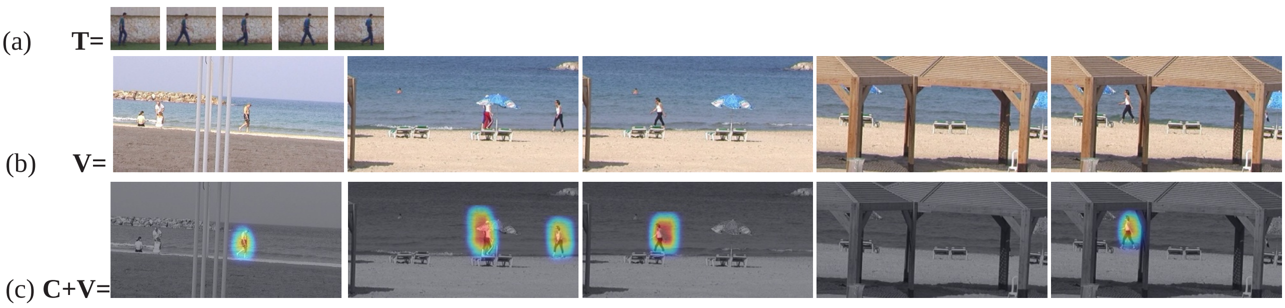

Figure 2. Walking on the beach. *(a) T = a short walk clip. (b) V = the longer beach video against which T was "correlated". (c) Peaks of space-time correlation C superimposed on V (see text). For video sequences see:* ***www.wisdom.weizmann.ac.il/∼vision/BehaviorCorrelation.html***

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 ([6] 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 two observations:

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

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.

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:
(i) It should bring into "motion-consistent alignment" as many ST-patches as possible between T and S.
(ii) 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.

A good global template match should minimize the number of local **in**consistent matches between regular patches (patches not containing motion discontinuity), and should also minimize the number of matches between regular patches in one sequence with motion discontinuity patches in the other sequence.

The following measure captures the degree of *local* **in***consistency* between a small ST-patch $P_1 \in T$ and a ST-patch $P_2 \in S$, according to the above-mentioned requirements:

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

where ϵ avoids division by 0. This measure yields low values (i.e., '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 (i.e., '**in**consistency') in all other cases.

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 $\frac{1}{m_{12}}$, i.e.: $C(T, S) = \frac{1}{N}\Sigma\frac{1}{m_{12}}$, which is the measure we used in our experiments.

A space-time template T (e.g., $30 \times 30 \times 30$) can thus be "correlated" against a larger video sequence (e.g., $200 \times 300 \times 1000$) 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 surface are locations in the large video sequence

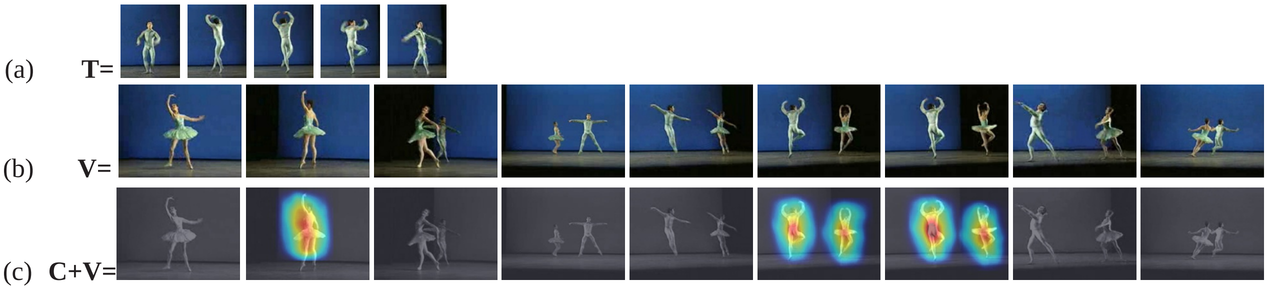

Figure 3. Ballet example. *(a) T = a single turn of the man-dancer. (b) V = the ballet video against which T was "correlated". (c) Peaks of space-time correlation C superimposed on V (see text). For video sequences see:* ***www.wisdom.weizmann.ac.il/~vision/BehaviorCorrelation.html**.*

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 peaks can be found in Figs. 2.c, 3.c, 4.c.

Computational efficiency:
In regular image correlation, the search space is 2D (the entire image). In the presented space-time correlation the search space is 3-dimensional (the entire video sequence), and the local computations are more complex (e.g., 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:

(i) The local matrices $\mathbf{M}_{3\times 3}$ (Eq. 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 which need to be estimated online during the space-time correlation process are the combined matrices $\mathbf{M_{12}}$ (Eq. 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 run-time, since $\mathbf{M_{12}} = \mathbf{M_1} + \mathbf{M_2}$ (see end of Section 3).

(ii) Eigenvalue estimation, which is part of the rank-increase measure (Eq. 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 Eq. (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 [6]: $\lambda_1 \leq \|\mathbf{M}\|_F \leq \sqrt{3}\lambda_1$. The scalar $\sqrt{3}$ (≈ 1.7) is related to the dimension of $\mathbf{M}$ (3×3). The rank-increase measure Δr can therefore be approximated by:

$$\Delta\hat{r} = \frac{det(\mathbf{M})}{det(\mathbf{M}^{\diamondsuit}) \cdot \|\mathbf{M}\|_F} \qquad (12)$$

$\Delta\hat{r}$ requires no eigenvalue computation, is easy to compute from $\mathbf{M}$, and provides the following bounds on the rank-increase measure Δr of Eq. (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'[1]. We use this approximated measure to speed-up our space-time correlation process.

(iii) The overall run-time for computing the entire "correlation volume" of a $144 \times 180 \times 200$ video sequence and a $60 \times 30 \times 30$ query, is 30 minutes on a Pentium 4, 2.4 GHz (since the correlation volume is smooth, it is enough to compute it for every other pixel and every other frame, and then interpolate). When searching only for correlation peaks, this process can be further sped-up using coarse-to-fine multi-grid search, thus *reducing the run-time to less than one minute* for the above example.

In our experiments the video sequences were of reduced spatial resolution, to reduce effects of temporal aliasing due to fast motion. Spatial blurring of size $[5 \times 5]$ with $\sigma = 0.75$ was applied before extracting the space-time gradients. The size of the space-time patches P (Eq. 4) was $[7 \times 7 \times 3]$, using weighted sums of gradients with Gaussian weights ($\sigma_{space} = 1.5$, and $\sigma_{time} = 1.0$) instead of regular sums.

7. Results

One possible application of our space-time correlation is to detect "behaviors of interest" in a video database. A

[1] In the analytic definition of the rank increase measure (Eq. 10), Δr attains high values in case of a uniform patch (where all eigenvalues might be equally small). In order to overcome this situation and to add numerical stability to the measure, we added a small constant to the Frobenius norm in the denominator of Eq. 12, that corresponds to 10 gray-level gradients.

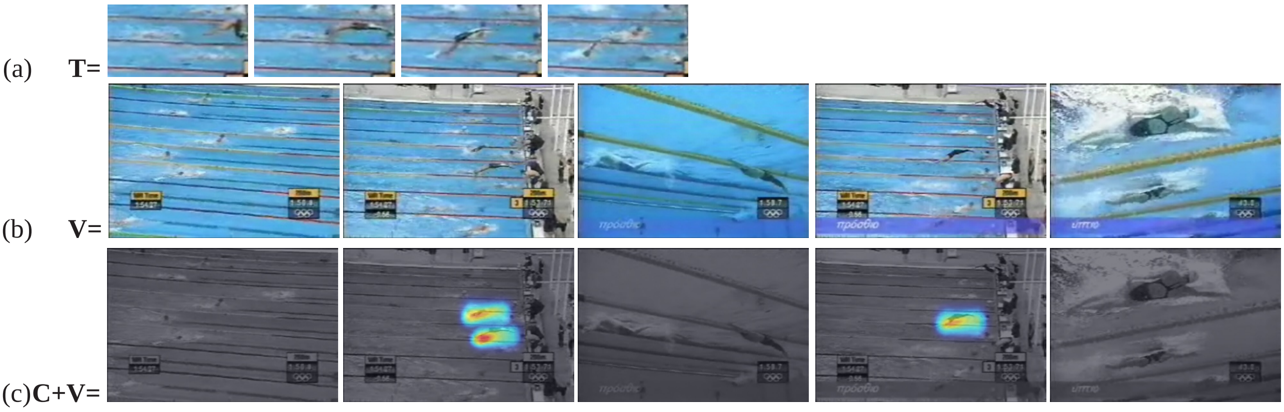

Figure 4. Swim-relay match. *(a) T = a single dive into the pool. (b) V = the swim-relay video against which T was "correlated". (c) Peaks of the space-time correlation C superimposed on V (see text). For video sequences see:* ***www.wisdom.weizmann.ac.il/∼vision/BehaviorCorrelation.html***

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. Please view the video clips (databases and queries) of the following experiments in: *www.wisdom.weizmann.ac.il/∼vision/ Behavior-Correlation.html*.

Fig. 2 shows 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 60x70 pixels) of a different man recorded elsewhere. Fig 2.a shows a few sampled frames from T. Fig 2.b shows a few sampled frames from the long beach video V (460 frames of 180x360 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 2.c 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 differences in spatial appearance* of the foreground moving objects and of their backgrounds. It is sensitive only to the underlying motions.

Fig. 3 shows 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 90x110 pixels). Fig. 3.a shows a few sampled frames from T. Fig. 3.b shows a few frames from the longer ballet clip V (284 frames of 144x192 pixels), against which T was "correlated". Peaks of the space-time correlation surface C are shown superimposed on V (Fig. 3.c). Most of the turns of the two dancers (a man and a woman) were detected, despite the variability in scale relative to the template (up to 20%). Note that this example contains very fast moving parts (frame-to-frame).

Fig. 4 shows detecting dives into a pool during a swimming relay match. This video was downloaded from the website of the 2004 Olympic Games, and was severely mpeg-compressed. The video query T is a short clip (70x140 pixels x16 frames) showing one dive (shown slightly enlarged in Fig 4.a for visibility). It was correlated against the one-minute long video V (757 frames of 240x360 pixels, Fig 4.b). Despite the numerous simultaneous activities (a variety of swim styles, flips under the water, splashes of water), and despite the severe noise, the space-time correlation was able to separate most of the dives from other activities (Fig 4.c). 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, etc. Also, it is unlikely that any segmentation method would be able to separate foreground and background objects here. Yet, the space-time correlation method was able to produce reasonable results.

Fig. 5 shows detection of five different activities which occur simultaneously: 'walk', 'wave' 'clap' 'jump', and 'fountain' (with flowing water). Five small video queries were provided ($T_1, .., T_5$), one for each activity (Fig 5.a). These were performed by different people&backgrounds than in the longer video V. A short sub-clip from the right-most fountain was used as the fountain-query T_5. Fig 5.c shows the peaks detected in each of the five correlation surfaces $C_1, .., C_5$. 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-

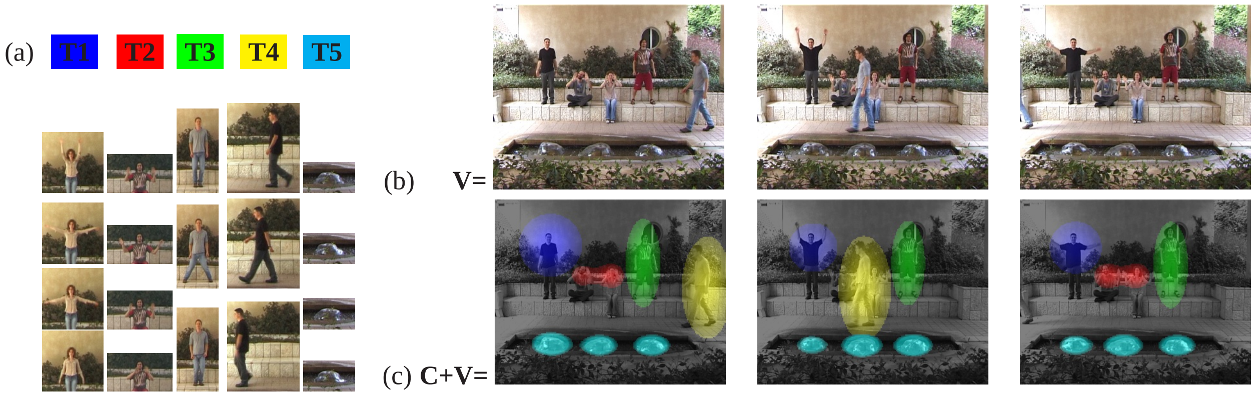

Figure 5. Detecting multiple activities. *(a) $T_1, ..T_5$ = five different short video templates. (b) V = the video against which T was "correlated". (c) Ellipses with colors corresponding to the 5 activities are displayed around the peaks detected in all 5 correlation surfaces $C_1, .., C_5$ (see text). For video sequences see:* ***www.wisdom.weizmann.ac.il/∼vision/BehaviorCorrelation.html****.*

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 Sec. 6).

8. Conclusion

By examining the intensity variations in video patches, we can implicitly characterize the space of their possible motions without having to explicitly choose arbitrary (wrong) ones of them as done in optical flow estimation. This allows us to identify whether two different 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 to detect all locations with similar dynamic behaviors, while being invariant to appearance, and without prior foreground/background segmentation. To our best knowledge, this is the first time multiple different behaviors/actions are detected simultaneously, 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 multi-scale framework (in space and in time). This is part of our future work.

Acknowledgments

We would like to thank Yoni Wexler and Oren Boiman for their useful remarks on the first draft of the paper. This work was supported in part by the Israeli Science Foundation (Grant No. 267/02) and by the Moross Laboratory at the Weizmann Institute of Science.

References

[1] M. J. Black. Explaining optical flow events with parameterized spatio-temporal models. In *CVPR*, 1999.

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001.

[3] C. Bregler. Learning and recognizing human dynamics in video sequences. *CVPR*, June 1997.

[4] O. Chomat and J. L. Crowley. Probabilistic sensor for the perception of activities. *ECCV*, 2000.

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, October 2003.

[6] G. Golub and C. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996.

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 147–151, 1988.

[8] B. Jähne, H. Haußecker, and P. Geißler. *Handbook of Computer Vision and Application*, volume 2. Academic Publishers, 1999.

[9] I. Laptev and T. Lindeberg. Space-time interest points. *ICCV*, 2003.

[10] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *IUworkshop*, pages 121–130, *IUW*, 1981.

[11] S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in xyt. *CVPR*, June 1994.

[12] H. Spies and H. Scharr. Accurate optical flow in noisy image sequences. *ICCV*, 1:587–592, July 2001.

[13] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *ECCV*, 2002.

[14] Y. Yacoob and M. J. Black. Parametrized modeling and recognition of activities. *CVIU*, 73(2):232–247, 1999.

[15] L. Zelnik-Manor and M. Irani. Event-based analysis of video. *CVPR*, 2001.