Actions as Space-Time Shapes

Lena Gorelick, Moshe Blank,
Eli Shechtman, *Student Member*, *IEEE*,
Michal Irani, *Member*, *IEEE*, and
Ronen Basri, *Member*, *IEEE*

Abstract—Human action in video sequences can be seen as silhouettes of a moving torso and protruding limbs undergoing articulated motion. We regard human actions as three-dimensional shapes induced by the silhouettes in the space-time volume. We adopt a recent approach [14] for analyzing 2D shapes and generalize it to deal with volumetric space-time action shapes. Our method utilizes properties of the solution to the Poisson equation to extract space-time features such as local space-time saliency, action dynamics, shape structure, and orientation. We show that these features are useful for action recognition, detection, and clustering. The method is fast, does not require video alignment, and is applicable in (but not limited to) many scenarios where the background is known. Moreover, we demonstrate the robustness of our method to partial occlusions, nonrigid deformations, significant changes in scale and viewpoint, high irregularities in the performance of an action, and low-quality video.

Index Terms—Action representation, action recognition, space-time analysis, shape analysis, poisson equation.

1 Introduction

Recognizing human action is a key component in many computer vision applications, such as video surveillance, human-computer interface, video indexing and browsing, recognition of gestures, analysis of sports events, and dance choreography.

Despite the fact that good results were achieved by traditional action recognition approaches, they still have some limitations. Many of them involve computation of optical flow [3], [11], whose estimation is difficult due to, e.g., aperture problems, smooth surfaces, and discontinuities. Others, [31], [7] employ feature tracking and face difficulties in cases of self-occlusions, change of appearance, and problems of reinitialization. Methods that rely on key frames (e.g., [9]) or eigenshapes of foreground silhouettes (e.g., [13]) lack information about the motion. Some approaches are based on periodicity analysis (e.g., [21], [24], [13]) and are thus limited to cyclic actions.

Some of the recent successful work done in the area of action recognition [10], [33], [17], [25], [16] have shown that it is useful to analyze actions by looking at a video sequence as a space-time volume (of intensities, gradients, optical flow, or other local features).

On the other hand, studies in the field of object recognition in 2D images have demonstrated that silhouettes contain detailed information about the shape of objects, e.g., [23], [1], [14], [8]. When a silhouette is sufficiently detailed people can readily identify the object, or judge its similarity to other shapes. One of the well-known shape descriptors is the Medial Axis Distance Transform [5], where each internal pixel of a silhouette is assigned a value reflecting its minimum distance to the boundary contour. The Medial Axis Transform opened the way to the advent of skeleton-based representations such as [26], [23]. Recently, [14] presented an alternative approach based on a solution to a Poisson equation. In this approach, each internal point is assigned with the mean time required for a particle undergoing a random-walk process starting from the point to hit the boundaries. In contrast to the distance transform, the resulting scalar field takes into account many points on the boundaries and, so, reflects more global properties of the silhouette. In addition, it allows extracting many useful properties of a shape, including part structure as well as local orientation and aspect ratio of the different parts simply by differentiation of the Poisson solution. Moreover, unlike existing pairwise comparison measures such as Chamfer and Hausdorff, which are designed to compute a distance measure between pairs of shapes, the Poisson-based descriptor provides description for single shapes and, so, it is naturally suitable for tasks requiring class modeling and learning.

Our approach is based on the observation that in video sequences a human action generates a *space-time shape* in the space-time volume (see Fig. 1). These shapes are induced by a concatenation of 2D silhouettes in the space-time volume and contain both the spatial information about the pose of the human figure at any time (location and orientation of the torso and limbs, aspect ratio of different body parts), as well as the dynamic information (global body motion and motion of the limbs relative to the body).

A similar approach was recently presented in [32], where human actions were presented as 3D spatio-temporal surfaces and analyzed using differential geometric surface properties. While our space-time volume representation is essentially derived from the same input (concatenation of silhouettes), it is robust to noise at the bounding contour of the extracted silhouettes as opposed to the local surface features in [32]. Our volumetric space-time features allow us to avoid the nontrivial and computationally expensive problem of surface parameterization (i.e., contour parameterization and frame-to-frame point correspondences) and surface-to-surface point correspondence between actions as described in [32].

Several other approaches use information that could be derived from the space-time shape of an action. Bobick and Davis [6] uses motion history images representation and [20] analyzes planar slices (such as x-t planes) of the space-time intensity volume. Note that these methods implicitly use only *partial* information about the space-time shape. Methods for 3D shape analysis and matching have been recently used in computer graphics (see survey in [28]). However, in their current form, they do not apply to space-time shapes due to the nonrigidity of actions, the inherent differences between the spatial and temporal domains, and the imperfections of the extracted silhouettes.

In this paper, we generalize a method developed for the analysis of 2D shapes [14] to deal with volumetric space-time shapes induced by human actions. This method exploits the solution to the Poisson equation to extract various shape properties that are utilized for shape representation and classification. We adopted some of the relevant properties and extend them to deal with space-time shapes (Section 2.1). The spatial and temporal domains are different in nature and therefore are treated differently at several stages of our method. The additional time domain gives rise to new space-time shape entities that do not exist in the spatial domain, such as a space-time "stick," "plate," and "ball." Each such type has different informative properties that characterize every space-time point. In addition, we extract space-time saliency at every point, which detects fast moving protruding parts of an action (Section 2.2).

Unlike images, where extraction of a silhouette might be a difficult segmentation problem, the extraction of a space-time shape from a video sequence can be simple in many scenarios. In video surveillance with a fixed camera as well as in various other settings, the appearance of the background is known. In these cases, using a simple change detection algorithm usually leads to satisfactory space-time shapes. Moreover, in cases of motion discontinuities, motion aliasing, and low-quality video, working with silhouettes may be advantageous over many existing methods ([3], [7], [10], [11], [15], [16], [17], [21], [20], [24], [25], [31], [33]) that compute optical flow, local space-time gradients, or other intensity-based features.

- *L. Gorelick, E. Shechtman, M. Irani, and R. Basri are with the Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, PO Box 26 Rehovot 76100, Israel.*
  E-mail: {lena.gorelick, eli.shechtman, michal.irani, ronen.basri}@weizmann.ac.il.
- *M. Blank is with EyeClick ltd., 5 Shoham St., Ramat Gan 52521, Israel.*
  E-mail: moshe.blank@gmail.com.

Manuscript received 18 June 2006; revised 11 Jan. 2007; accepted 14 May 2007; published online 5 June 2007.
Recommended for acceptance by B.S. Manjunath.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0455-0606.
Digital Object Identifier no. 10.1109/TPAMI.2007.70711.

0162-8828/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

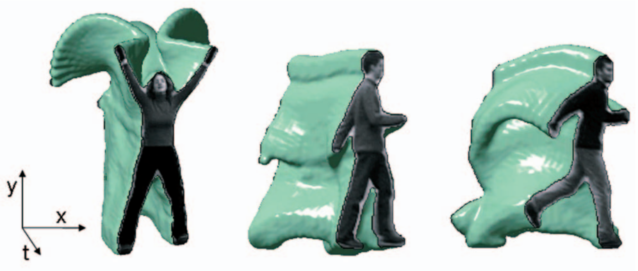

Fig. 1. Space-time shapes of "jumping-jack," "walk," and "run" actions.

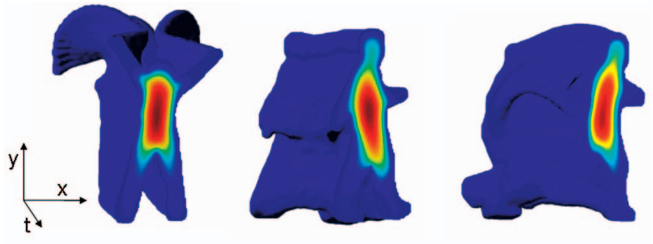

Fig. 2. The solution to the Poisson equation on space-time shapes of shown in Fig. 1. The values are encoded by the color spectrum from blue (low values) to red (high values).

Our method is fast, does not require prior video alignment and is not limited to cyclic actions. We demonstrate the robustness of our approach to partial occlusions, nonrigid deformations, imperfections in the extracted silhouettes, significant changes in scale and viewpoint, and high irregularities in the performance of an action. Finally, we report the performance of our approach in the tasks of action recognition, clustering, and action detection in a low-quality video (Section 3).

A preliminary version of this paper appeared in ICCV '05 [4].

2 Representing Actions as Space-Time Shapes

2.1 The Poisson Equation and Its Properties

Consider an action and its space-time shape S surrounded by a simple, closed surface. Below, we generalize the approach in [14] from 2D shapes in images to to deal with volumetric space-time shapes. We assign each space-time point within the shape with the mean time required for a particle undergoing a random-walk process starting from the point to hit the boundaries. This measure can be computed [14] by solving a Poisson equation of the form: $\Delta U(x, y, t) = -1$, with $(x, y, t) \in S$, where the Laplacian of U is defined as $\Delta U = U_{xx} + U_{yy} + U_{tt}$, subject to the Dirichlet boundary conditions $U(x, y, t) = 0$ at the bounding surface ∂S. In order to cope with the artificial boundary at the first and last frames of the video, we impose the Neumann boundary conditions requiring $U_t = 0$ at those frames [29]. The induced effect is of a "mirror" in time that prevents attenuation of the solution toward the first and last frames.

Note that space and time units may have different extents, thus when discretizing the Poisson equation we utilize space-time grid with the ratio $c_{ts} = h_t/h_s$ where (h_s, h_t) are the meshsizes in space and in time. Different values of c_{ts} affect the distribution of local orientations and saliency features across the space-time shape and, thus, allows us to emphasize different aspects of actions. In the following, we assume c_{ts} is given. (See more discussion in Section 3.1.)

Numerical solutions to the Poisson Equation can be obtained by various methods. We used a simple "w-cycle" of a geometric multigrid solver which is linear in the number of space-time points [29].

Fig. 2 shows a spatial cross-cut of the solution to the Poisson equation obtained for the space-time shapes shown in Fig. 1. High values of U are attained in the central part of the shape, whereas the external protrusions (the head and the limbs) disappear at relatively low values of U. The isosurfaces of the solution U represent smoother versions of the Dirichlet bounding surface and are perpendicular to the Neumann bounding surfaces (first and last frames) [14]. If we now consider the 3×3 Hessian matrix H of U at every internal space-time point, H will vary continuously from one point to the next and we can treat it as providing a measure that estimates locally the space-time shape near any interior space-time point. The eigenvectors and eigenvalues of H then reveal the local orientation and aspect ratio of the shape [14].

A 2×2 Hessian and its eigenvalues have been used before for describing 3D surface properties [2], [12], [32], [30]. This requires specific surface representations, e.g., surface normals, surface triangulation, surface parameterization, etc. Note, that converting our space-time binary masks to such surfaces is not a trivial task. In contrast, we extract local shape properties at every space-time point including internal points by using a 3×3 Hessian of the solution U without any surface representation.

2.2 Extracting Space-Time Shape Features

The solution to the Poisson equation can be used to extract a wide variety of useful local shape properties [14]. We adopted some of the relevant properties and extended them to deal with space-time shapes. The additional time domain gives rise to new space-time shape entities that do not exist in the spatial domain. We first show how the Poisson equation can be used to characterize space-time points by identifying space-time saliency of moving parts and locally judging the orientation and rough aspect ratios of the space-time shape. Then, we describe how these local properties can be integrated into a compact vector of global features to represent an action.

2.2.1 *Local Features*

Space-Time Saliency. Human action can often be described as a moving torso and a collection of parts undergoing articulated motion [7], [15]. Below, we describe how we can identify portions of a space-time shape that are salient both in space and in time.

In the space-time shape induced by a human action, the highest values of U are obtained within the human torso. Using an appropriate threshold, we can identify the central part of a human body. However, the remaining space-time region includes both the moving parts and portions of the torso that are near the boundaries, where U has low values. Those portions of boundary can be excluded by noticing that they have high gradient values. Following [14], we define

$$\Phi = U + \frac{3}{2}\|\nabla U\|^2, \tag{1}$$

where $\nabla U = (U_x, U_y, U_t)$.

Consider a sphere which is a space-time shape of a disk growing and shrinking in time. This shape has no protruding moving parts and, therefore, all of its space-time points are equally salient. Indeed, it can be shown that, in this case, Φ is constant. In space-time shapes of natural human actions, Φ achieves its highest values inside the torso and its lowest values inside the fast moving limbs. Static elongated parts or large moving parts (e.g., head of a running person) will only attain intermediate values of Φ. We define the space-time saliency features as a normalized variant of Φ

$$w_\Phi(x, y, t) = 1 - \frac{\log(1 + \Phi(x, y, t))}{\max\limits_{(x,y,t)\in S}(\log(1 + \Phi(x, y, t)))}, \tag{2}$$

which emphasizes fast moving parts. Fig. 3 illustrates the space-time saliency function w_Φ computed on the space-time shapes of Fig. 1.

For actions in which a human body undergoes a global motion (e.g., a walking person), we compensate for the global translation of the body in order to emphasize motion of parts relative to the torso. This is done by fitting a smooth trajectory (2nd order polynomial) to the centers of mass collected from the entire sequence and then by aligning this trajectory to a reference point (similarly to figure-centric stabilization in [11]). This essentially is equivalent to

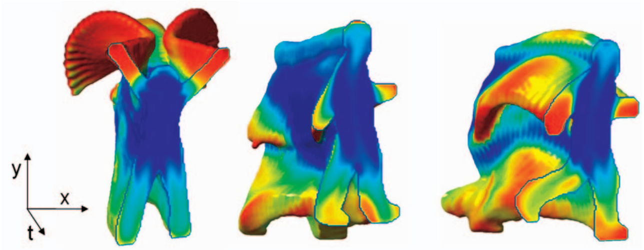

Fig. 3. **Examples of the local space-time saliency features**—w_Φ. The values are encoded by the color spectrum from blue (low values) to red (high values).

redirecting the low-frequency component of the action trajectory to the temporal axis. Linear fitting would account for global translation of a shape in the space-time volume. We chose however to use second order fitting to allow also acceleration. A third order polynomial would overcompensate and attenuate the high frequency components as well, which is undesired.

Space-Time Orientations. We use the 3×3 Hessian H of the solution to the Poisson equation to estimate the local orientation and aspect ratio of different space-time parts. Its eigenvectors correspond to the local principal directions and its eigenvalues are related to the local curvature in the direction of the corresponding eigenvectors and therefore inversely proportional to the length [14]. Below, we generalize this approach to space-time.

Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of H. Then, the first principal eigenvector corresponds to the shortest direction of the local space-time shape and the third eigenvector corresponds to the most elongated direction. Inspired by earlier works [22], [18] in the area of perceptual grouping, and 3D shape reconstruction, we distinguish between the following three types of local space-time structures:

- $\lambda_1 \approx \lambda_2 \gg \lambda_3$—corresponds to a space-time "stick" structure. For example, a small moving object generates a slanted space-time "stick," whereas a static object has a "stick" shape in the temporal direction. The informative direction of such a structure is the direction of the "stick" which corresponds to the third eigenvector of H.
- $\lambda_1 \gg \lambda_2 \approx \lambda_3$—corresponds to a space-time "plate" structure. For example, a fast moving limb generates a slanted space-time surface ("plate"), and a static vertical torso/limb generates a "plate" parallel to the y-t plane. The informative direction of a "plate" is its normal which corresponds to the first eigenvector of H.
- $\lambda_1 \approx \lambda_2 \approx \lambda_3$—corresponds to a space-time "ball" structure which does not have any principal direction.

Using the ratio of the eigenvalues at every space-time point we define three continuous measures of "plateness" $S_{pl}(x,y,t)$, "stickness" $S_{st}(x,y,t)$, and "ballness" $S_{ba}(x,y,t)$, where

$$S_{pl} = e^{-\alpha\frac{\lambda_2}{\lambda_1}}, \quad S_{st} = (1 - S_{pl})e^{-\alpha\frac{\lambda_3}{\lambda_2}}, \quad S_{ba} = (1 - S_{pl})\left(1 - e^{-\alpha\frac{\lambda_3}{\lambda_2}}\right). \tag{3}$$

Note that $S_{pl} + S_{st} + S_{ba} = 1$ and the transition between the different types of regions is gradual (we use $\alpha = 3$).

We then identify regions with vertical, horizontal, and temporal "plates" and "sticks." Let $\mathbf{v}(x,y,t)$ be the informative direction (of a "plate" or a "stick") computed with Hessian at each point. The deviations of the informative direction from the principal axes directions can be measured by $D_j(x,y,t) = |\mathbf{v} \cdot \mathbf{e}_j|$ with $\mathbf{e}_j$, $j \in \{1,2,3\}$ denoting the unit vectors in the direction of the principal axes (x, y, and t). Eventually, we define the orientation local features to be

$$w_{i,j}(x,y,t) = S_i(x,y,t) \cdot D_j(x,y,t), \tag{4}$$

where $i \in \{pl, st\}$ and $j \in \{1,2,3\}$. We have found the isotropic "ball" features to be redundant and, therefore, did not use them.

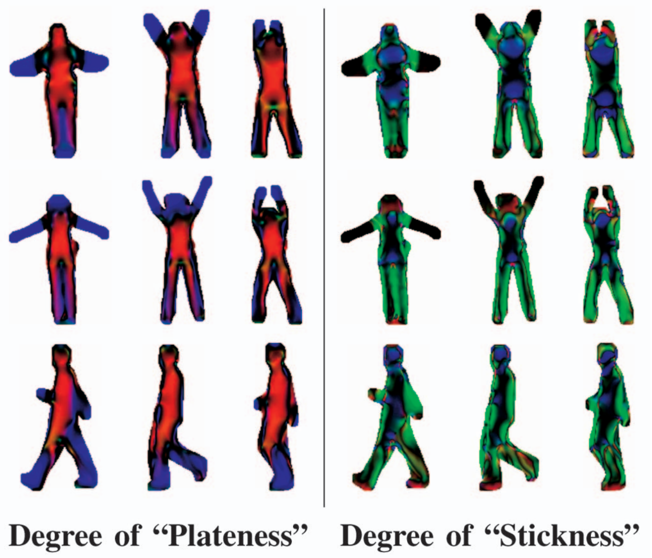

Fig. 4. **Space-time orientations of plates and sticks** for "jumping-jack" (first two rows) and "walk" (last row) actions. The first two rows illustrate three sample frames of two different persons performing the "jumping-jack" action. In the third row, we show a person walking. The left three columns show a schematic representation of normals where local plates were detected. The right three columns show principal directions of local sticks. In all examples, we represent with the blue, red, and green colors regions with temporal, horizontal, and vertical informative direction accordingly. The intensity denotes the extent to which the local shape is a plate or a stick. For example, fast moving hands of a "jumping-jack" are identified as plates with normals oriented in temporal direction (appear in blue on the left). Whereas slower moving legs are identified as vertical sticks (appear in green on the right). Note the color consistency between the same action of two different persons, despite the dissimilarity of their spatial appearance.

Fig. 4 demonstrates examples of space-time shapes and their orientation measured locally at every space-time point.

2.2.2 *Global Features*

In order to represent an action with global features, we use weighted moments of the form

$$m_{pqr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w(x,y,t) g(x,y,t) x^p y^q t^r dx dy dt, \tag{5}$$

where $g(x,y,t)$ denotes the characteristic function of the space-time shape, $w(x,y,t)$ is a one of the seven possible weighting functions: $w_{i,j}(x,y,t)$ (4) or $w_\Phi(x,y,t)$ (2). Note that $0 \leq w(x,y,t) \leq 1 \forall (x,y,t)$.

In the following section, we demonstrate the utility of these features in action recognition and classification experiments.

3 Results and Experiments

For the first two experiments (action classification and clustering), we collected a database of 90 low-resolution (180×144, deinterlaced 50 fps) video sequences showing nine different people, each performing 10 natural actions such as "run," "walk," "skip," "jumping-jack" (or shortly "jack"), "jump-forward-on-two-legs" (or "jump"), "jump-in-place-on-two-legs" (or "pjump"), "gallop-sideways" (or "side"), "wave-two-hands" (or "wave2"), "wave-one-hand" (or "wave1"), or "bend." To obtain space-time shapes of the actions, we subtracted the median background from each of the sequences and used a simple thresholding in color-space. The resulting silhouettes contained "leaks" and "intrusions" due to imperfect subtraction, shadows, and color similarities with the background (see Fig. 5 for examples). In our view, the speed of global translation in the real world (due to different view points or, e.g., different step sizes of a tall versus a short person) is less informative for action recognition than the shape and speed of the limbs relative to the torso. We therefore compensate for the

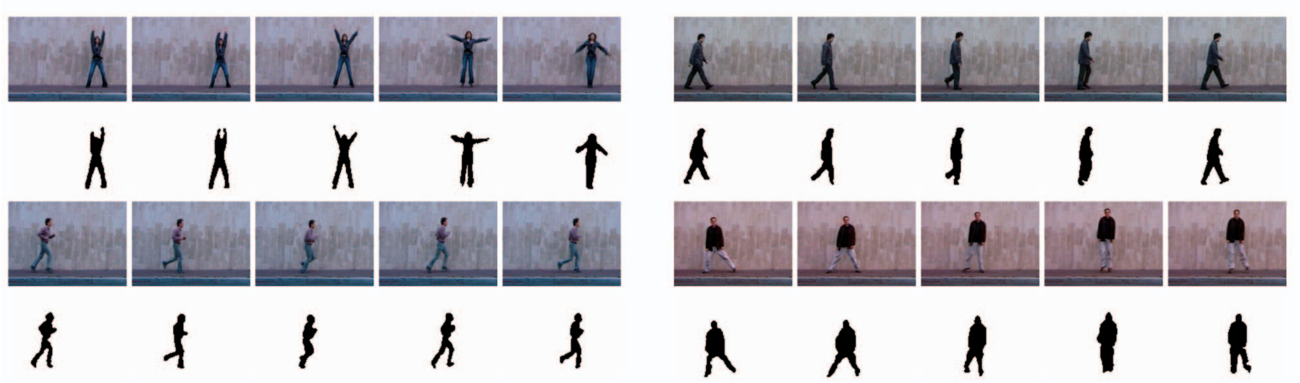

Fig. 5. **Examples of video sequences and extracted silhouettes** from our database.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	100	0	0	0	0	0	0	0	0	0
a2	0	98.0	2.0	0	0	0	0	0	0	0
a3	0	2.9	97.1	0	0	0	0	0	0	0
a4	0	0	0	100	0	0	0	0	0	0
a5	0	0	10.8	0	89.2	0	0	0	0	0
a6	0	0	0	0	0	100	0	0	0	0
a7	0	0	0	0	0	0	100	0	0	0
a8	0	0	0	0.9	0	0.9	0	94.8	3.5	0
a9	0	0	0	0.9	0	0	0	1.9	97.2	0
a10	0	0	0	0	0	0	0	0	0.9	99.1

(a)

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	82.4	0.8	2.4	0	14.4	0	0	0	0	0
a2	2.0	34.7	51.0	0	10.2	0	2.0	0	0	0
a3	1.4	40.6	43.5	0	8.7	0	5.8	0	0	0
a4	0	0	0	95.5	0	0.8	0	0.8	1.5	1.5
a5	13.8	13.8	26.2	0	29.2	0	16.9	0	0	0
a6	0	0	0	12.8	0	84.9	0	0	1.2	1.2
a7	3.2	4.8	23.8	0	17.5	0	50.8	0	0	0
a8	0	0	0	0.9	0	7.0	0	29.6	62.6	0
a9	0	0	0	0.9	0	0.9	0	44.3	51.9	1.9
a10	0	0	0	4.5	0	0	0	3.6	5.4	86.6

(b)

Fig. 6. (a) Action confusion in classification experiment using our method. (b) Action confusion in classification experiment using the method in [33]. (a1-“walk,” a2-“run,” a3-“skip,” a4-“jack,” a5-“jump,” a6-“pjump,” a7-“side,” a8-“wave1,” a9-“wave2,” and a10-“bend”).

translation of the center of mass by aligning the silhouette sequence to a reference point (see Section 2.2.1). The database, as well as the extracted silhouettes, are available for download at [27].

For each sequence, we solved the Poisson equation using meshsizes $(h_s = 1, h_t = 3)$ and computed seven types of local features: “stick” and “plate” features, measured at three directions each (as in (4)), and the saliency features (as in (2)). In order to treat both the periodic and nonperiodic actions in the same framework as well as to compensate for different length of periods, we used a sliding window in time to extract space-time cubes, each having eight frames with an overlap of four frames between the consecutive space-time cubes. Moreover, using space-time cubes allows a more accurate localization in time while classifying long video sequences in realistic scenarios. We centered each space-time cube about its space-time centroid and brought it to a uniform scale in space preserving the spatial aspect ratio. Note that the coordinate normalization above does not involve any global video alignment. We then computed global space-time shape features with spatial moments up to order $m_s = 2$ and time moments up to order $m_t = 2$ (i.e., with $p + q \leq m_s$ and $r \leq m_t$ in (5)), giving rise to a $7 \times (m_t + 1) \times (m_s + 1)(m_s + 2)/2 = 7 \times 18 = 126$ feature vector representation per space-time cube. (The maximal order of moments was chosen empirically by testing all possible combinations of m_s and m_t between 1 and 5.)

3.1 Action Classification

For every video sequence, we perform a leave-one-out procedure, i.e., we remove the entire sequence (all its space-time cubes) from the database while other actions of the same person remain. Each cube of the removed sequence is then compared to all the cubes in the database and classified using the nearest neighbor procedure (with Euclidian distance operating on normalized global features). Thus, for a space-time cube to be classified correctly, it must exhibit high similarity to a cube of a different person performing the same action. Indeed, for correctly classified space-time cubes, the distribution of the person labels, associated with the retrieved nearest neighbor cubes, is fully populated and nonsparse, implying that our features emphasize action dynamics, rather than person shape characteristics.

The algorithm misclassified 20 out of 923 space-cubes (2.17 percent error rate). Fig. 6a shows action confusion matrix for the entire database of cubes. Most of the errors were caused by the “jump” action which was confused with the “skip.” This is a reasonable confusion considering the small temporal extent of the cubes and partial similarity between dynamics of these actions.

We also ran the same experiment with *ordinary* space-time shape moments (i.e., substituting $w(x, y, t) = 1$ in (5)). The algorithm misclassified 73 out of 923 cubes (7.91 percent error rate) using moments up to order $m_s = 4$ in space and $m_t = 7$ in time resulting in $(m_t + 1) \times (m_s + 1)(m_s + 2)/2 - 4 = 116$ features (where -4 stands for the noninformative zero moment and the first-order moments in each direction). Further experiments with all combinations of maximal orders between 2 and 9 yielded worse results. Note that space-time shapes of an action are very informative and rich as is demonstrated by the relatively high classification rates achieved even with ordinary shape moments.

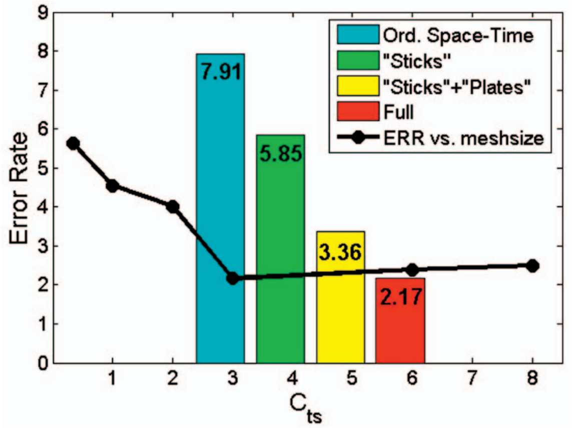

Fig. 7. **Evaluation of our method in different settings:** sensitivity to the meshsize ratio c_{ts} (black line), contribution of different features to the overall performance (color-coded bars).

In Fig. 7 (black line), we evaluate the sensitivity of our method to the meshsize ratio c_{ts} by plotting the classification error rate as a function of c_{ts}. As can be seen, the method is quite robust to this parameter. We found that $c_{ts} = 3$ works best for our collection of human actions. We used deinterlaced sequences of 180×144, 50 fps, with average person size (width) of 12 pixels. We expect the optimal ratio c_{ts} to grow linearly with the change in frame rate or the size of the person performing the action. Moreover, in the same Fig. 7 (color-coded bars), we demonstrate how each of the local shape features contributes to the overall classification performance by evaluating our method in three different settings: using moments extracted from "stick" features only, "stick" and "plate" features only and using all of them ("stick," "plate," and "salience" features). These are compared to the performance obtained with ordinary space-time moments.

For comparison with our method, we applied the method of [33] to our database using the original implementation obtained from the authors. We used the same sliding window size of eight frames every four frames. The algorithm with the best combination of parameters (16 equally spaced bins, 3 pyramid levels) misclassified 336 out 923 cubes (36.40 percent error rate). The confusion matrix in Fig. 6b shows that most of the errors of the method of [33] occur between "run" and "skip," "side" and "skip," and "wave1" and "wave2" actions. The latter can be easily explained since location of a movement is not grasped by looking at histograms alone. Moreover, only absolute values of the gradient are taken in [33] and, thus, two motions performed in opposite directions will be similar.

3.2 Action Clustering

In this experiment, we applied a common spectral clustering algorithm [19] to 90 unlabelled action sequences. We defined the distance between any two sequences to be a variant of the Median Hausdorff Distance

$$D_H(s^1, s^2) = \underset{j}{\text{median}}\left(\min_i \|c_i^1 - c_j^2\|\right) + \underset{i}{\text{median}}\left(\min_j \|c_i^1 - c_j^2\|\right), \quad (6)$$

where $\{c_i^1\}$ and $\{c_j^2\}$ denote the space-time cubes belonging to the sequences s^1 and s^2 accordingly. In contrast to assigning a label to the entire space-time shape, separate classifying of the overlapping cubes allows more flexibility since it accounts explicitly for occasional occlusions and other imperfections in the space-time shape of the action. As a result, we obtained ten separate clusters of the 10 different actions with only four of the sequences erroneously clustered with other action sequences, (see Fig. 8).

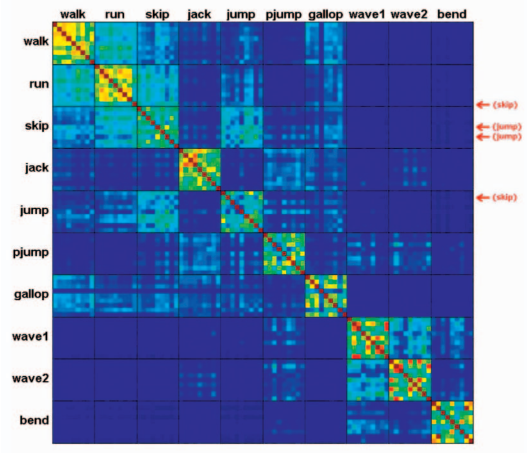

Fig. 8. **Results of spectral clustering.** Distance matrix, reordered using the results of spectral clustering. We obtained 10 separate clusters of the 10 different actions. The rows of the erroneously clustered sequences are marked with arrows and the label of the misclassified class.

3.3 Robustness

In this experiment, we demonstrate the robustness of our method to high irregularities in the performance of an action. We collected 10 test video sequences of people walking in various difficult scenarios in front of different nonuniform backgrounds (see Fig. 9 for a few examples). We show that our approach has relatively low sensitivity to partial occlusions, nonrigid deformations, and other defects in the extracted space-time shape. Moreover, we demonstrate the robustness of our method to substantial changes in viewpoint. For this purpose, we collected ten additional sequences, each showing the "walk" action captured from a different viewpoint (varying between 0 degree and 81 degree relative to the image plane with steps of 9 degree). Note, that sequences with angles approaching 90 degree contain significant changes in scale within the sequence. See the upper left sequence in Fig. 9, showing "walk" in the 63 degree direction. The rest of the sequences can be found at [27].

For each of the test sequences s, we measured its Median Hausdorff Distance to each of the action types a_k, $k \in \{1 \dots 9\}$ in our database

$$D_H(s, a_k) = \underset{i}{\text{median}}(\min_j \|c_i - c_j\|), \quad (7)$$

where $c_i \in s$ is a space-time cube belonging to the test sequence and $c_j \in a_k$ denotes a space-time cube belonging to one of the training sequences of the action a_k. We then classified each test sequence as the action with the smallest distance. Fig. 10a, shows for each of the test sequences the first and second best choices and their distances as well as the median distance to all the actions. The test sequences are sorted by the distance to their first best chosen action. All the test sequences in Fig. 10a were classified correctly as the "walk" action. Note the relatively large difference between the first (the correct) and the second choices (with regard to the median distance). Fig. 10b shows similar results for the sequences with varying viewpoints. All sequences with viewpoints between 0 degree and 54 degree were classified correctly with a large relative gap between the first (true) and the second closest actions. For larger viewpoints, a gradual deterioration occurs. This demonstrates the robustness of our method to relatively large variations in viewpoint.

3.4 Action Detection in a Ballet Movie

In this experiment, we show how given an example of an action we can use space-time shape properties to identify all locations with similar actions in a given video sequence.

We chose to demonstrate our method on the ballet movie example used in [25]. This is a highly compressed (111 Kbps, wmv format) $192 \times 144 \times 750$ ballet movie with effective frame rate of 15 fps, moving camera and changing zoom, showing performance

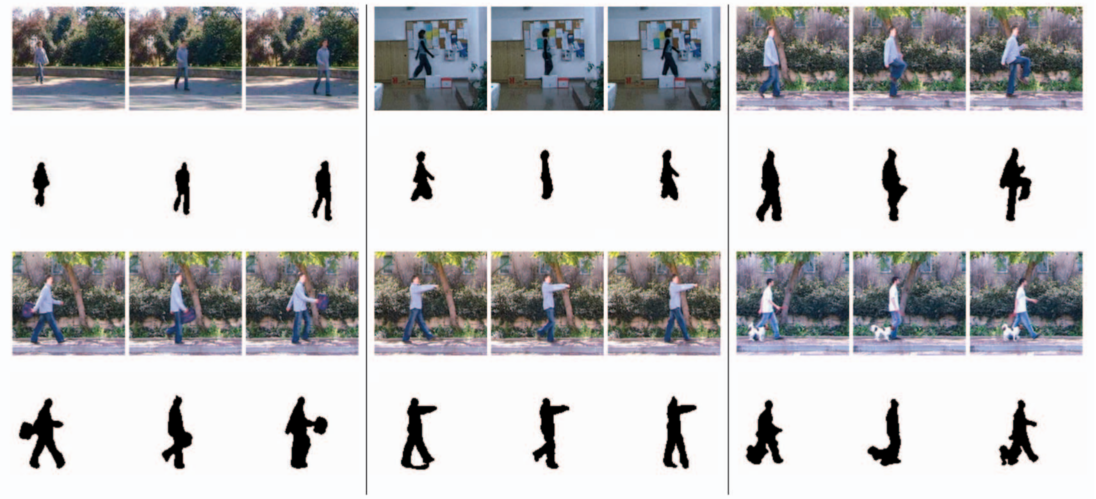

Fig. 9. **Examples of sequences used in robustness experiments.** We show three sample frames and their silhouettes for the following sequences (left to right): "Diagonal walk" (63 degree), "Occluded legs," "Knees up," "Swinging bag," "Sleepwalking," and "Walking with a dog."

Test Seq.	1^{st} *best*		2^{nd} *best*		*Med.*
Normal walk	walk	5.6	run	8.2	11.2
Walking in a skirt	walk	5.6	side	8.1	9.9
Carrying briefcase	walk	6.6	side	8.5	10.4
Limping man	walk	7.0	skip	8.8	10.3
Occluded legs	walk	8.2	skip	11.0	11.3
Knees Up	walk	8.3	side	9.6	10.1
Walking with a dog	walk	8.4	run	9.9	11.4
Sleepwalking	walk	8.4	run	9.8	12.1
Swinging a bag	walk	9.6	side	11.1	12.9
Occluded by a pole	walk	10.6	jack	11.6	12.5

(a)

Test Seq.	1^{st} *best*		2^{nd} *best*		*Med.*
Dir. 0^o	walk	8.3	run	10.8	12.6
Dir. 9^o	walk	7.9	side	9.9	12.2
Dir. 18^o	walk	8.2	side	10.2	12.1
Dir. 27^o	walk	8.2	side	9.7	11.5
Dir. 36^o	walk	8.3	side	10.3	11.7
Dir. 45^o	walk	9.0	side	10.7	11.6
Dir. 54^o	walk	9.1	side	10.6	11.3
Dir. 63^o	walk	11.1	side	11.6	12.9
Dir. 72^o	walk	11.3	pjump	12.1	12.9
Dir. 81^o	walk	12.6	pjump	12.7	13.3

(b)

Fig. 10. **Robustness experiment results.** The leftmost column describes the test action performed. For each of the test sequences, the closest two actions with the corresponding distances are reported in the second and third columns. The median distance to all the actions in the database appears in the rightmost column. (a) Shows results for the sequences with high irregularities in the performance of the "walk" action. (b) Shows results for the "walk" sequences with varying viewpoints.

of two (female and male) dancers. We manually separated the sequence into two parallel movies each showing only one of the dancers. For both of the sequences, we then solved the Poisson equation and computed the same global features as in the previous experiment for each space-time cube.

We selected a cube with the male dancer performing a "cabriole" pa (beating feet together at an angle in the air) and used it as a query to find all the locations in the two movies where a similar movement was performed by either a male or a female dancer. Fig. 11 demonstrates the results of the action detection by simply thresholding euclidian distances computed with normalized global features. These results are comparable to the results reported in [25]. Accompanying video material can be found at [27].

4 Conclusion

In this paper, we represent actions as space-time shapes and show that such a representation contains rich and descriptive information about the action performed. The quality of the extracted features is demonstrated by the success of the relatively simple classification scheme used (nearest neighbors classification and Euclidian distance). In many situations, the information contained in a single space-time cube is rich enough for a reliable classification to be performed, as was demonstrated in the first classification experiment. In real-life applications, reliable performance can be achieved by integrating information coming from the entire input sequence (all its space-time cubes), as was demonstrated by the robustness experiments.

Our approach has several advantages: First, it does not require video alignment. Second, it is linear in the number of space-time points in the shape. The overall processing time (solving the Poisson equation and extracting features) in Matlab of a $110 \times 70 \times 50$ presegmented video takes less than 30 seconds on a Pentium 4, 3.0 GHz. Third, it has a potential to cope with low-quality video data, where other methods that are based on intensity features only (e.g., gradients), might encounter difficulties.

As our experiments show, the method is robust to significant changes in scale, partial occlusions, and nonrigid deformations of the actions. While our method is not fully view invariant, it is

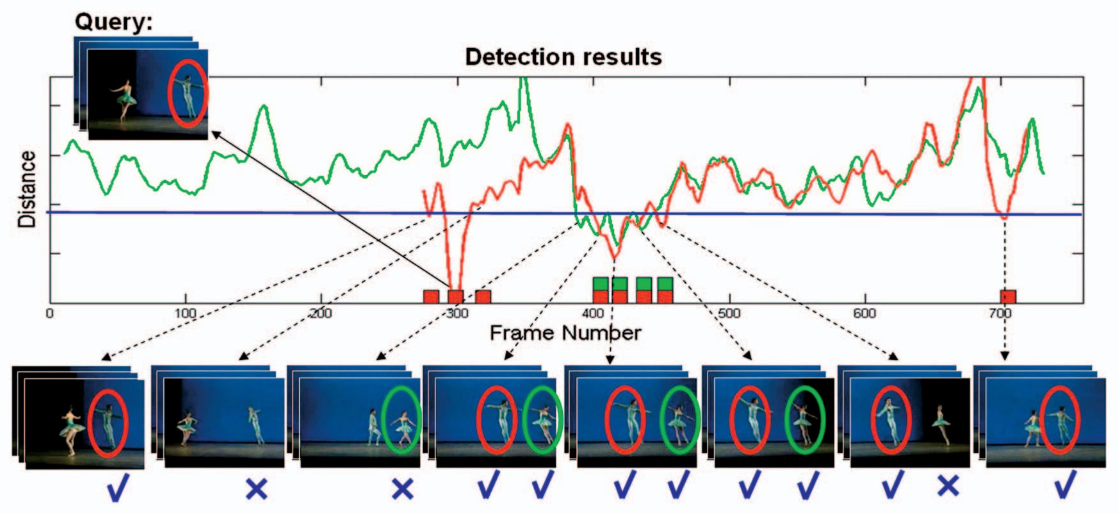

Fig. 11. **Results of action detection in a ballet movie.** The green and the red lines denote the distances between the query cube and the cubes of the female and the male dancers accordingly. The ground truth is marked with the green squares for the female dancer and the red squares for the male dancer. A middle frame is shown for every detected space-time cube. Correct detections are marked with blue "v," whereas false alarms and misses are marked with blue "x." The algorithm detected all locations with actions similar to the query except for one false alarm of the female dancer and two misses (male and female), all marked with blue "x." The two misses can be explained by the difference in the hand movement, and the false alarm—by the high similarity between the hand movement of the female dancer and the query. Additional "cabriole" pa of the male dancer was completely occluded by the female dancer, and therefore ignored in our experiment. Full video results can be found at [27].

however robust to large changes in viewpoint (up to 54 degree). This can be further improved by enrichment of the training database with actions taken from a few discrete viewpoints.

Acknowledgments

The authors thank Lihi Zelnik-Manor for helpful discussions and her assistance in adjusting the method in [33] for comparison. This work was supported in part by the Israel Science Foundation Grant No. 267/02, by the European Commission Project IST-2002-506766 Aim Shape, by the Binational Science foundation Grant No. 2002/254, and by a research grant from A.M.N. fund for the promotion of science, culture, and arts in Israel. The research was conducted at the Moross Laboratory for Vision and Motor Control at the Weizmann Institute of Science.

References

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 4, pp. 509-522, Apr. 2002.

[2] P.J. Besl and R.C. Jain, "Invariant Surface Characteristics for 3D Object Recognition in Range Images," *Computer Vision, Graphics, and Image Processing,* vol. 33, no. 1, pp. 33-80, 1986.

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

[4] 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, 2005.

[5] H. Blum, "A Transformation for Extracting New Descriptors of Shape," *Models for the Perception of Speech and Visual Form, Proc. Symp.,* pp. 362-380, 1967.

[6] 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.

[7] C. Bregler, "Learning and Recognizing Human Dynamics in Video Sequences," *Proc. Computer Vision and Pattern Recognition,* June 1997.

[8] S. Carlsson, "Order Structure, Correspondence and Shape Based Categories," *Proc. Int'l Workshop Shape, Contour, and Grouping,* p. 1681, 1999.

[9] S. Carlsson and J. Sullivan, "Action Recognition by Shape Matching to Key Frames," *Proc. Workshop Models versus Exemplars in Computer Vision,* Dec. 2001.

[10] O. Chomat and J.L. Crowley, "Probabilistic Sensor for the Perception of Activities," *Proc. European Conf. Computer Vision,* 2000.

[11] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. Int'l Conf. Computer Vision,* Oct. 2003.

[12] T. Fan, G. Medioni, and A. Nevatia, "Matching 3-D Objects Using Surface Descriptions," *Proc. IEEE Int'l Conf. Robotics and Automation,* vol. 3, no. 24-29, pp. 1400-1406, 1988.

[13] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, "Behavior Classification by Eigendecomposition of Periodic Motions," *Pattern Recognition,* vol. 38, no. 7, pp. 1033-1043, 2005.

[14] L. Gorelick, M. Galun, E. Sharon, A. Brandt, and R. Basri, "Shape Representation and Classification Using the Poisson Equation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, Dec. 2006.

[15] S.X. Ju, M.J. Black, and Y. Yacoob, "Cardboard People: A Parametrized Model of Aticulated Image Motion," *Proc. Second Int'l Conf. Automatic Face and Gesture Recognition,* pp. 38-44, Oct. 1996.

[16] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," *Proc. Int'l Conf. Computer Vision,* pp. 166-173, 2005.

[17] I. Laptev and T. Lindeberg, "Space-Time Interest Points," *Proc. Int'l Conf. Computer Vision,* 2003.

[18] G. Medioni and C. Tang, "Tensor Voting: Theory and Applications," *Proc. 12th Congres Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle,* 2000.

[19] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *Proc. Advances in Neural Information Processing Systems 14,* pp. 849-856, 2001.

[20] S.A. Niyogi and E.H. Adelson, "Analyzing and Recognizing Walking Figures in xyt," *Proc. Computer Vision and Pattern Recognition,* June 1994.

[21] R. Polana and R.C. Nelson, "Detection and Recognition of Periodic, Nonrigid Motion," *Int'l J. Computer Vision,* vol. 23, no. 3, 1997.

[22] E. Rivlin, S. Dickinson, and A. Rosenfeld, "Recognition by Functional Parts," *Proc. Computer Vision and Pattern Recognition,* pp. 267-274, 1994.

[23] T. Sebastian, P. Klein, and B. Kimia, "Shock-Based Indexing into Large Shape Databases," *Proc. European Conf. Computer Vision,* vol. 3, pp. 731-746, 2002.

[24] S. Seitz and C. Dyer, "View-Invariant Analysis of Cyclic Motion," *Int'l J. Computer Vision,* vol. 25, no. 3, pp. 231-251, Dec. 1997.

[25] E. Shechtman and M. Irani, "Space-Time Behavior Based Correlation," *Proc. Computer Vision and Pattern Recognition,* June 2005.

[26] K. Siddiqi, A. Shokoufandeh, S.J. Dickinson, and S.W. Zucker, "Shock Graphs and Shape Matching," *Proc. IEEE Int'l Conf. Computer Vision,* p. 222, 1998.

[27] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html, 2005.

[28] J. Tangelder and R. Veltkamp, "A Survey of Content Based 3D Shape Retrieval Methods," *Proc. Shape Modeling Int'l,* pp. 145-156, 2004.

[29] U. Trottenberg, C. Oosterlee, and A. Schuller, *Multigrid.* Academic Press, 2001.

[30] U. Weidenbacher, P. Bayerl, H. Neumann, and R. Fleming, "Sketching Shiny Surfaces: 3D Shape Extraction and Depiction of Specular Surfaces," *ACM Trans. Applied Perception,* vol. 3, no. 3, pp. 262-285, 2006.

[31] 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.

[32] A. Yilmaz and M. Shah, "Actions Sketch: A Novel Action Representation," *Computer Vision and Pattern Recognition,* vol. 1, pp. 984-989, 2005.

[33] L. Zelnik-Manor and M. Irani, "Event-Based Analysis of Video," *Computer Vision and Pattern Recognition,* pp. 123-130, Sept. 2001.