Actions as Space-Time Shapes

Moshe Blank Lena Gorelick Eli Shechtman Michal Irani Ronen Basri

Dept. of Computer Science and Applied Math.
Weizmann Institute of Science
Rehovot 76100, Israel

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 [9] 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, non-rigid deformations, significant changes in scale and viewpoint, high irregularities in the performance of an action and low quality video.

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. Some of the recent work done in the area of action recognition [7, 21, 11, 17] have shown that it is useful to analyze actions by looking at the video sequence as a space-time intensity volume. Analyzing actions directly in the space-time volume avoids some limitations of traditional approaches that involve the computation of optical flow [2, 8] (aperture problems, smooth surfaces, singularities, etc.), feature tracking [20, 4] (self-occlusions, re-initialization, change of appearance, etc.), key frames [6] (lack of information about the motion). Most of the above studies are based on computing local space-time gradients or other intensity based features and thus might be unreliable in cases of low quality video, motion discontinuities and motion aliasing.

Figure 1. **Space-time shapes** of "jumping-jack", "walking" and "running" actions.

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., [16, 1, 9, 5]. When a silhouette is sufficiently detailed people can readily identify the object, or judge its similarity to other shapes.

Our approach is based on the observation that the human action in video generates a *space-time shape* in the space-time volume (see Fig. 1). These space-time shapes contain both spatial information about the pose of the human figure at any time (location and orientation of the torso and the limbs, aspect ratio of the different body parts), as well as the dynamic information (global body motion and motion of the limbs relative to the body). Several other approaches use information that could be derived from the space-time shape of an action. [3] uses motion history images representation and [14] 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 [18]). However, in their current form, they do not apply to space-time shapes due to the non-rigidity 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 analysis of 2D shapes [9], 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 (Sec. 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 (Sec. 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.

Our method is fast and does not require prior video alignment. We demonstrate the robustness of our approach to partial occlusions, non-rigid deformations, imperfections in the extracted silhouettes 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 (Sec. 3).

2. Representing Actions as Space-Time Shapes

Below we generalize the approach in [9] to deal with volumetric 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. We assign every internal space-time point a value reflecting its relative position within the space-time shape. This is done by assigning each space-time point 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 [9] by solving a Poisson equation of the form:

$$\Delta U(x,y,t) = -1, \tag{1}$$

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 [19]. The induced effect is of a "mirror" in time that prevents attenuation of the solution towards the first and last frames.

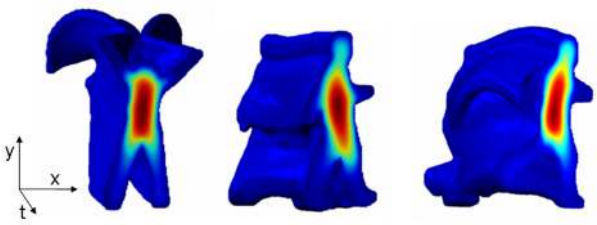

Figure 2. **The solution to the Poisson equation on space-time shapes** of "jumping-jack", "walking" and "running" actions. The values are encoded by the color spectrum from blue (low values) to red (high values). Note that points at the boundary attain zero values (Dirichlet boundary conditions).

Note that space and time units may have different extents, thus when discretizing the Poisson equation we utilize space-time grid with different meshsizes in space and in time. This affects the distribution of local orientations and saliency features across the space-time shape, and thus allows us to emphasize different aspects of actions. We found that a discretization scheme that uses space-time units with spatial extent twice as long as the temporal one (distance between frames) works best for most of human actions that we collected.

Fig. 2 shows a spatial cross-cut of the solution to the Poisson equation obtained for several space-time shapes in Fig. 1. The level sets of U represent smoother versions of the bounding surface with the external protrusions (fast moving limbs) disappearing already at relatively low values of U. Below we generalize the analysis in [9] to characterize actions as space-time shapes using measures that estimate locally the second order moments of a shape near any given point.

Consider first a space-time shape given by a conic, i.e., composed of the points (x,y,t) satisfying

$$P(x,y,t) = ax^2 + by^2 + ct^2 + dxy + eyt + fxt + g \leq 0. \tag{2}$$

In this case the solution to the Poisson equation takes the form

$$U(x,y,t) = -\frac{P(x,y,t)}{2(a+b+c)}. \tag{3}$$

The isosurfaces of U then contain a nested collection of scaled versions of the conic boundary, where the value of U increases quadratically as we approach the center. If we now consider the Hessian matrix of U we obtain at any given point exactly the same matrix, namely

$$H(x,y,t) = -\frac{1}{a+b+c}\begin{pmatrix} a & d/2 & f/2 \\ d/2 & b & e/2 \\ f/2 & e/2 & c \end{pmatrix}. \tag{4}$$

This matrix is in fact the second moment matrix of the entire 3D conic shape, scaled by a constant. The eigenvectors and eigenvalues of H then reveal the orientation of the shape and its aspect ratios.

For general space-time shapes described by more complicated equations the isosurfaces of U represent smoother versions of the boundaries and the Hessian varies continuously from one point to the next. The Hessian provides a measure that estimates locally the space-time shape near any space-time point inside the shape.

Numerical solutions to the Poisson Equation can be obtained by various methods. We used an efficient multigrid technique to solve the equation. The time complexity of such solver is linear in the number of space-time points. In all our experiments one multigrid "w-cycle" was sufficient to obtain an adequate solution. For more details see [19]. Finally, the solution obtained may be noisy near the boundaries due to discretization. To reduce this noise we apply as a post-processing stage a few relaxation sweeps enforcing $\triangle U = -1$ inside the space-time shape and $\triangle U = 0$ outside. This will smooth U near the boundaries and hardly affect more inner points.

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 [9]. 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. Next 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 [4, 10]. Below we describe how we can identify portions of a space-time shape that are salient both in space and in time.

In a 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 [9] we define

$$\Phi = U + \frac{3}{2}\|\nabla U\|^2 \qquad (5)$$

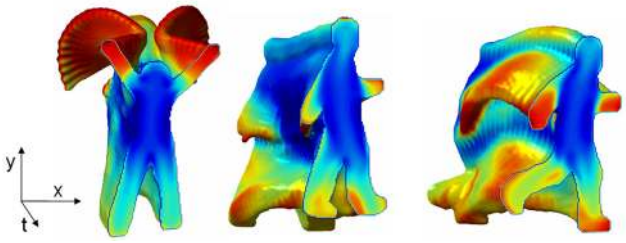

Figure 3. **Examples of the local space-time saliency features** - $\hat{\Phi}$. The values are encoded by the color spectrum from blue (low values) to red (high values).

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

Consider a space-time sphere which is a space-time shape of a disk growing and shrinking in time. This isotropic space-time shape has no protruding moving parts and therefore all its space-time points are equally salient. Indeed, $\Phi = \frac{r^2}{6}$ at all points inside the sphere, with r denoting the radius of the sphere. 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 use a normalized variant of Φ

$$\hat{\Phi}(x,y,t) = 1 - \frac{\log\left(1+\Phi(x,y,t)\right)}{\max\limits_{(x,y,t)\in S}\left(\log\left(1+\Phi(x,y,t)\right)\right)}, \qquad (6)$$

which emphasizes fast moving parts. Fig. 3 illustrates the space-time saliency function $\hat{\Phi}$ 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 function to the centers of mass collected from the entire sequence and considering only the deviations from this function (similarly to figure-centric stabilization in [8]).

Space-Time Orientations

The Poisson equation can be used to estimate the local orientation and aspect ratio of different space-time parts. This is done by constructing the Hessian H Eq. (4) that approximates locally the second order shape moments at any given point, and its eigenvectors correspond to the local principal directions. The eigenvalues of H are related to the local curvature in the direction of the corresponding eigenvectors and therefore inversely proportional to the length.

Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of H. Then the first principle 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 [15, 12] in the area of perceptual grouping,

and 3D shape reconstruction, we distinguish between the following 3 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.

We exploit the decomposition above to characterize each point with two types of local features. The first is related to the local shape structure, and the second relies on its most informative orientation. 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

$$\begin{aligned} S_{pl} &= e^{-\alpha\frac{\lambda_2}{\lambda_1}} \\ S_{st} &= (1-S_{pl})e^{-\alpha\frac{\lambda_3}{\lambda_2}} \\ S_{ba} &= (1-S_{pl})(1-e^{-\alpha\frac{\lambda_3}{\lambda_2}}). \end{aligned} \tag{7}$$

Note that $S_{pl} + S_{st} + S_{ba} = 1$ and the transition between the different types of regions is gradual.

The second type of local features identifies 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. Then the orientation measures are defined as:

$$\begin{aligned} D_1 &= e^{-\beta|\mathbf{v}\cdot\mathbf{e}_1|} \\ D_2 &= e^{-\beta|\mathbf{v}\cdot\mathbf{e}_2|} \\ D_3 &= e^{-\beta|\mathbf{v}\cdot\mathbf{e}_3|}, \end{aligned} \tag{8}$$

with $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ denoting the unit vectors in the direction of the principle axes - x,y and t (we used $\beta = 3$).

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:

$$m_{pqr} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y,t)g(x,y,t)x^p y^q t^r dxdydt \tag{9}$$

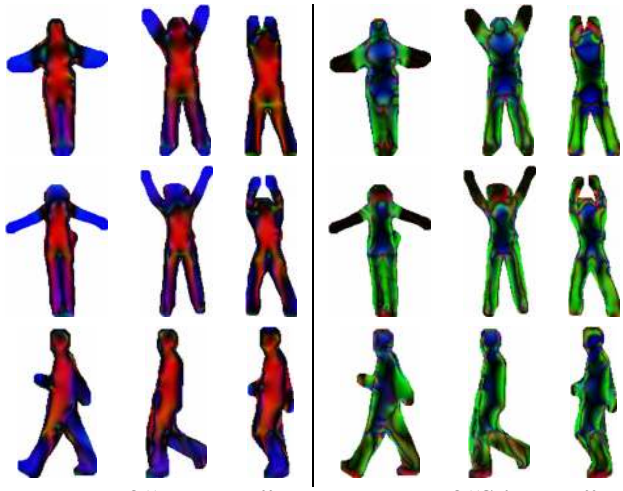

Figure 4. **Space-time orientations of plates and sticks** for "jumping-jack" (first two rows) and "walk" (last row) actions. The first two rows illustrate three sample frames of two different persons performing the "jumping-jack" action. In the third row we show a person walking. The left three columns 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.

where $g(x,y,t)$ denotes the characteristic function of the space-time shape, $w(x,y,t)$ is a weighting function. For each pair of a local shape type i and a unit vector $\mathbf{e}_j$, we substitute the weights w with the combined local feature

$$w(x,y,t) = S_i(x,y,t) \cdot D_j(x,y,t) \tag{10}$$

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 as global features. Note that $0 \le w(x,y,t) \le 1 \ \forall (x,y,t)$.

In addition to the above six types of weighting functions we also generate space-time saliency moments using $w(x,y,t) = \hat{\Phi}$ of Eq. (6).

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

3. Results and Experiments

The local and global space-time features presented in 2.2 are used for action recognition and classification.

For the first two experiments (action classification and clustering) we collected a database of 81 low-resolution (180×144, 25 fps) video sequences showing nine different people, each performing nine natural actions such as "running", "walking", "jumping-jack", "jumping-forward-on-two-legs", "jumping-in-place-on-two-legs", "galloping-sideways", "waving-two-hands", "waving-one-hand", "bending". 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). For actions in which a human body undergoes a global motion, we compensate for the translation of the center of mass in order to emphasize motion of parts relative to the torso by fitting a second order polynomial to the frame centers of mass.

For each sequence we solved the Poisson equation and computed seven types of local features $w(x, y, t)$ in Eq. (10) and Eq. (6). In order to treat both the periodic and non-periodic 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 10 frames with an overlap of 5 frames between the consecutive space-time cubes. 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. We then computed global space-time shape features with spatial moments up to order 5 and time moments up to order 2 (i.e., with $p+q \leq 5$ and $r \leq 2$ in Eq. (9)), giving rise to a 280 feature vector representation per space-time cube. Note, that coordinate normalization above does not involve any global video alignment/ registration.

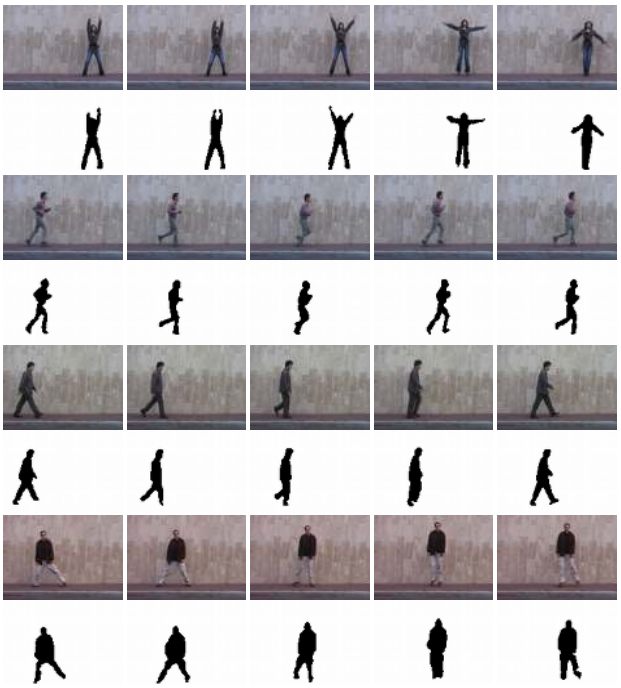

Figure 5. **Examples of video sequences and extracted silhouettes** from our database.

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. This way the possibility of high similarity due to *spatial* appearance purely, is minimized.

The algorithm misclassified 1 out of 549 space-time cubes (0.36% error rate). The correct classifications originated uniformly from all other persons in the database. We also ran the same experiment with *ordinary* space-time shape moments (i.e., substituting $w(x, y, t) = 1$ in Eq. (9)) of up to order 7 in space and in time. The algorithm misclassified 17 out of 549 cubes (3.10% error rate). Further experiments with all combinations of orders between 3 and 14 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.

To demonstrate the superiority of the space-time shape information over spatial information collected separately from each frame of a sequence we conducted an additional experiment. For each of the space-time cubes in our database we centered the silhouette in each frame about its spatial centroid and brought it to a uniform scale preserving the spatial aspect ratio. We then computed spatial shape moments of the silhouette in each of the frames separately and concatenated these moments into one feature vector for the entire space-time cube. Next, we used these moments to perform the same leave-one-out classification procedure. We tested all combinations of orders between 3 and 8 resulting in up to 440 features. The algorithm with the best combination misclassified 35 out 549 cubes (6.38% error rate). To explain why the space-time approach outperforms the spatial-per-frame approach consider for example the "run" and "walk" actions. Many successive frames from the first action may exhibit high spatial similarity to the successive

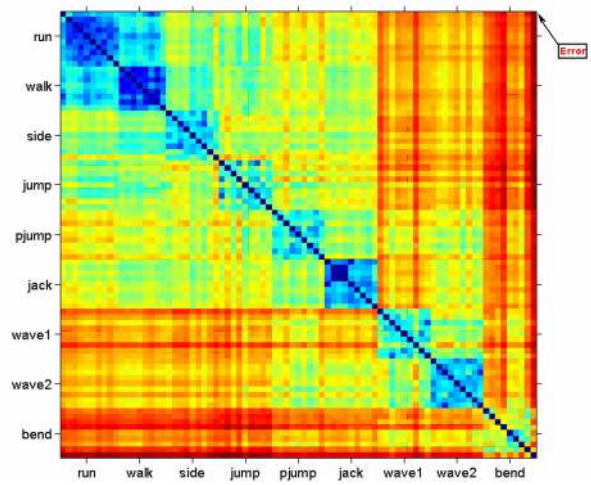

Figure 6. **Results of spectral clustering.** Distance matrix, reordered using the results of spectral clustering. We obtained nine separate clusters of the nine different actions. The row of the erroneously clustered "walk" sequence is marked with an arrow.

frames from the second one. Ignoring the dynamics within the frames might lead to confusion between the two actions.

3.2. Action Clustering

In this experiment we applied a common spectral clustering algorithm [13] to 81 unlabelled action sequences. We defined the distance between any two sequences to be a variant of the Median Hausdorff Distance:

$$D_H(s^1, s^2) = \underset{j}{\text{median}}(\min_i \|c_i^1 - c_j^2\|) + \underset{i}{\text{median}}(\min_j \|c_i^1 - c_j^2\|), \quad (11)$$

where $\{c_i^1\}$ and $\{c_j^2\}$ denote the space-time cubes belonging to the sequences s^1 and s^2 accordingly. As a result we obtained nine separate clusters of the nine different actions with only one "walk" sequence erroneously clustered with the "run" sequences. Fig. 6 shows the resulting distance matrix.

3.3. Robustness

In this experiment we demonstrate the robustness of our method to high irregularities in the performance of an action. We collected ten test video sequences of people walking in various difficult scenarios in front of different non-uniform backgrounds (see Fig. 7 for a few examples). We show that our approach has relatively low sensitivity to partial occlusions, non-rigid deformations and other defects in the extracted space-time shape. In addition, it is partially robust to changes in viewpoint, as is demonstrated by the "diagonal walk" example (30-40 degrees, see Fig. 7, upper left).

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:

$$D_H(s, a_k) = \underset{i}{\text{median}}(\min_j \|c_i - c_j\|) \quad (12)$$

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. All the test sequences except for one were classified correctly as the "walk" action. Fig. 8 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. Note that in the misclassified sequence the difference between the first and second (the correct) choices is small (w.r.t the median distance), compared to the differences in the other sequences.

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 [17]. This is a highly compressed (111Kbps, wmv format) $192 \times 144 \times 750$ ballet movie with effective frame rate of 15 fps, moving camera and changing zoom, showing performance 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

Test Sequence	1^{st} *best*		2^{nd} *best*		*Med.*
Normal walk	walk	7.8	run	11.5	15.9
Walking in a skirt	walk	8.8	run	11.6	16.0
Carrying briefcase	walk	10.0	gallop	13.5	16.7
Knees up	walk	10.5	jump	14.0	14.9
Diagonal walk	walk	11.4	gallop	13.6	15.1
Limping man	walk	12.8	gallop	15.9	16.8
Occluded legs	walk	13.4	pjump	15.0	15.8
Swinging bag	walk	14.9	jack	17.3	19.7
Sleepwalking	walk	15.2	run	16.8	19.9
Walking with a dog	run	17.7	walk	18.4	22.2

Figure 8. **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. Abbreviations: pjump = "jumping-in-place-on-two-legs", jump = "jumping-forward-on-two-legs", jack = "jumping-jack", gallop = "galloping-sideways".

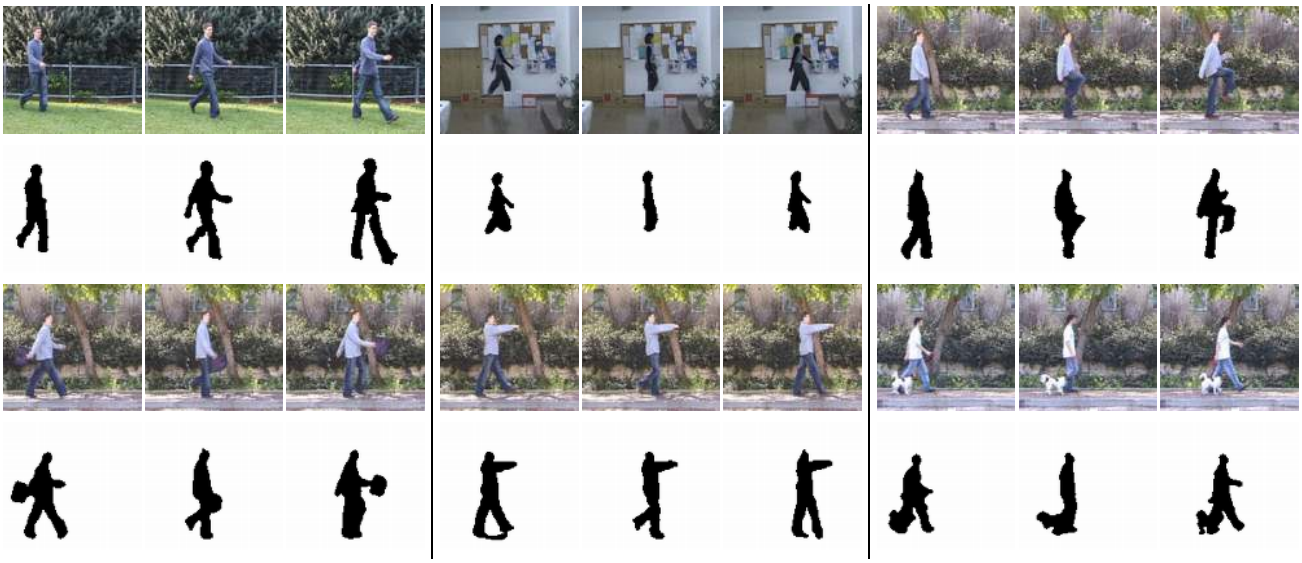

Figure 7. **Examples of sequences used in robustness experiments.** We show three sample frames and their silhouettes for the following sequenced (from left to right): "Diagonal walk", "Occluded legs", "Knees up", "Swinging bag", "Sleepwalking", "Walking with a dog".

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. 9 demonstrates the results of the action detection by simply thresholding euclidian distances computed with normalized global features. 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. The algorithm detected all locations with action 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. These results are comparable to the results reported in [17]. Accompanying video material can be found at `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`

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 $110\times70\times50$ pre-segmented 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. On the other hand by looking at the space-time shape only we ignore the intensity information inside the shape. In the future this method can be combined with intensity based features to further improve the performance. It is also possible to broaden the range of space-time features extracted with the Poisson equation in order to deal with more challenging tasks such as human gait recognition. Finally, this approach can also be applied with very little change to general 3D shapes representation and matching.

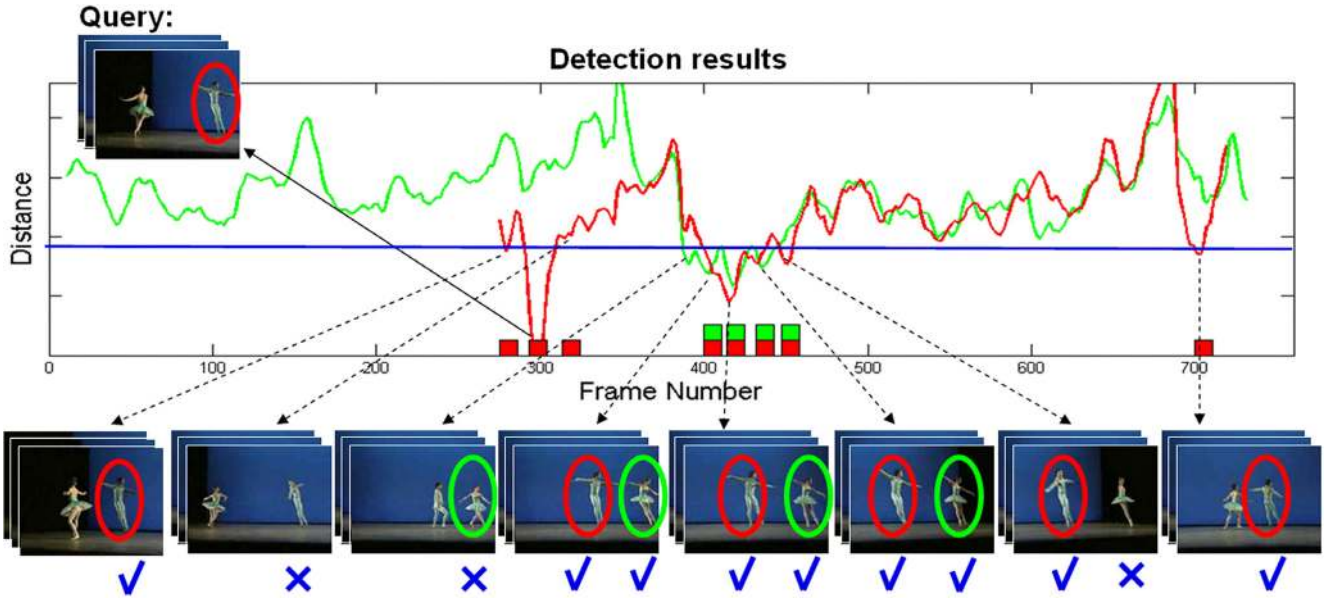

Figure 9. **Results of action detection in a ballet movie.** The green and the red lines denote the distances between the query cube and the cubes of the female and the male dancers accordingly. The ground truth is marked with the green squares for the female dancer and the red squares for the male dancer. 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". Full video results can be found at `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`.

Acknowledgements

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, and by the Binational Science foundation Grant No. 2002/254. 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. *PAMI*, 24(4):509–522, 2002.

[2] M. J. Black. Explaining optical flow events with parameterized spatio-temporal models. *CVPR*, 1:1326–1332, 1999.

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

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

[5] S. Carlsson. Order structure, correspondence and shape based categories. *International Workshop on Shape, Contour and Grouping , Springer Lecture Notes in Computer Science*, page 1681, 1999.

[6] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. *Workshop on Models versus Exemplars in Computer Vision*, December 2001.

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

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

[9] L. Gorelick, M. Galun, E. Sharon, A. Brandt, and R. Basri. Shape representation and recognition using the poisson equation. *CVPR*, 2:61–67, 2004.

[10] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parametrized model of aticulated image motion. *2nd Int. Conf. on Automatic Face and Gesture Recognition*, pages 38–44, October 1996.

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

[12] G. Medioni and C. Tang. Tensor voting: Theory and applications. *Proceedings of RFIA, Paris, France*, 2000.

[13] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *NIPS*, pages 849–856, 2001.

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

[15] E. Rivlin, S. Dickinson, and A. Rosenfeld. Recognition by functional parts. *CVPR*, pages 267–274, 1994.

[16] T. Sebastian, P. Klein, and B. Kimia. Shock-based indexing into large shape databases. *ECCV (3)*, pages 731–746, 2002.

[17] E. Shechtman and M. Irani. Space-time behavior based correlation. *In proceedings of CVPR*, June 2005.

[18] J. Tangelder and R. Veltkamp. A survey of content based 3d shape retrieval methods. *Proceedings Shape Modeling International*, pages 145–156, 2004.

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

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

[21] L. Zelnik-Manor and M. Irani. Event-based analysis of video. *CVPR*, pages 123–130, September 2001.