Features Extraction by using Poisson Equation

Mamta Bhardwaj, Hemant Tulsani, Shalini Rajput
Electronics and Communication Department, AIACTR, New Delhi - 110031
Email: mamta.bhardwaj1984@gmail.com

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 for analyzing 2D shapes and generalize it to deal with volumetric space-time action shapes. This 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 and classification. We show that these features are useful for action recognition, detection, and clustering. This method is fast, does not require video alignment, and is applicable in many scenarios where the background is known. Moreover, we demonstrate the robustness of our method to partial occlusions, no rigid deformations, significant changes in scale and viewpoint, high irregularities in the performance of an action, and low-quality video.

Index Terms- *Action representation, space-time analysis, shape analysis, Medial axis transform, Poisson equation.*

1. INTRODUCTION

Features extraction and recognition of 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 whose estimation is difficult due to, e.g. aperture problems, smooth surfaces, and discontinuities. Others employ feature tracking and face difficulties in cases of self-occlusions, change of appearance, and problems of re-initialization. Methods that rely on key frames Eigen shapes of foreground silhouettes lack information about the motion. Some approaches are based on periodicity analysis and are thus limited to cyclic actions. Some of the recent successful works done in the area of action recognition 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, 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 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 and 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 pair wise 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 modelling 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 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). Several other approaches use information that could be derived from the space-time shape of an action uses motion history images representation and 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. 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 the analysis of 2D shapes 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. The spatial and temporal domains are different in nature and therefore are treated differently at several stages of our method. 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 method that compute optical flow, local space-time gradients, or other intensity-based features.

2. MEDIAL AXIS TRANSFORM

Medial Axis (MA), also known as Centers of Maximal Disks, is a useful representation of a shape for image description and analysis. MA can be computed on a distance transform, where each point is labeled to its distance to the background. Recent algorithms allow one to compute Squared Euclidean Distance Transform (SEDT) in linear time in any dimension. While these algorithms provide exact measures, the only known method to characterize MA on SEDT, using local tests and Look-Up Tables (LUT), is limited to 2D and small distance values.

The medial axis of a shape provides a compact representation of its features and their connectivity. As a result, researchers have discovered and are still exploring its use in many fields, such as topology recognition for grid generation. The medial axis is defined when the shape is embedded in an Euclidean space and is endowed with a distance function. Therefore, an expedient route is to efficiently obtained followed by the medial axis construction. In 3D, a sphere is called *medial* if it meets *S*, the domain boundary, only tangentially in at least two points. The medial axis *M* is defined as the closure of the set of centers of all medial spheres. Informally, the medial axis of a surface in 3D is the set of all points that have more than one closest point on the surface. They are often called the medial axis transform (MAT) for that 3D bounded domain.

Also action representation by temporal templates is done in previous work but these methods implicitly use only partial information about the space time shape. When the object performs an action in 3D, the points on the outer boundary of the object are projected as 2D (x, y) contour in the image plane. A sequence of such 2D contours with respect to time generates a spatiotemporal volume (STV) in (x, y, t), which can be treated as 3D object in the(x, y, t) space. We analyze STV by using the differential geometric surface properties, such as peaks, pits, valleys and ridges, which are important action descriptors capturing both spatial and temporal properties. A set of motion descriptors for a given is called an action sketch. The action descriptors are related to various types of motions and object deformations. The first step in our approach is to generate STV by solving the point correspondence problem between consecutive frames. The correspondences are determined using a two-step graph theoretical approach. After the STV is generated, actions descriptors are computed by analyzing the differential geometric properties of STV. This method is analyzed using differential geometric surface properties while our space-time volume representation is essentially derived from the same input i.e. by concatenation of Silhouettes.

Fig1 given below gives geometries where *d* distance contours and medial axes (in thicker line) Results are same but Laplacian criteria is simpler and cheaper to calculate.

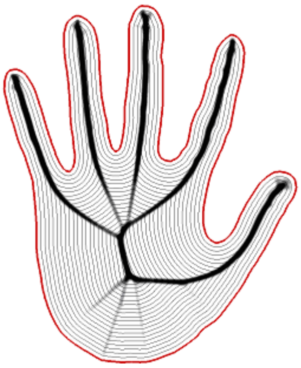


Fig1- Laplacian based medial axis criteria

3. ACTION AS SPACE TIME SHAPES

3.1 POISSON EQUATION

Consider a silhouette *S* surrounded by a simple, closed contour. A sensible approach to inferring properties of the silhouette is to assign to every internal point a value that depends on the relative position of that point within the silhouette. One popular example is the distance transform, which assigns to every point within the silhouette a value reflecting its minimal distance to the boundary contour, and which can be computed by solving the Eikonal equation ($\|\nabla u\|^2 = 1$).

An alternative approach is to place a set of particles at the point and let them move in a random walk until they hit the contour. Then we can measure various statistics of this random walk, such as the mean time required for a particle to hit the boundaries. This particular measure can be computed by solving a Poisson equation of the form ∇U (x, y, t)= -1 with (*x, y, t*) $\in S$, where the Laplacian of *U* is defined as $\nabla U =$

$\partial_{xx} + \partial_{yy} + U_{tt}$ 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 at those frames. The induced effect is of a "mirror" in time $U_T = 0$ that prevents attenuation of the solution toward the first and last frames.

Note that space and time unit may have different extents, thus with the ratio $c_{ts} = h_t / h_s$ where h_t, h_s are the mesh size in time and in space. Different values of c_{ts} affect the distribution of local orientation and saliency features across the space and thus allows us to emphasize different aspects of actions. In the following we assume c_{ts} is given.

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. Fig.2 shows a spatial cross-cut of the solution to the Poisson equation obtained for the space-time shapes shown in 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) 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 Eigen values of H then reveal the local orientation and aspect ratio of the shape. A 2×2 Hessian and its Eigen values have been used before for describing 3D surface properties This requires specific surface representations, e.g., surface normal, 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.

3.2 EXTRACTING SPACE-TIME FEATURES

The solution to the Poisson equation can be used to extract a wide variety of useful local shape properties. 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.

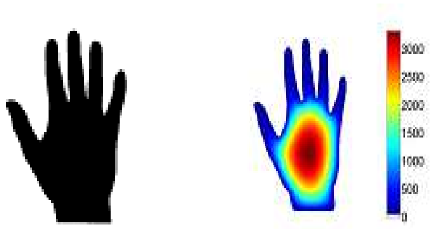


Fig2-Solution to Poisson Equation for silhouettes

3.2.1 LOCAL FEATURES

SPACE-TIME SALIENCY - Human action can often be described as a moving torso and a collection of parts undergoing articulated motion. 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 we define

$$\phi = U + 3/2 \|\nabla U\|^2 \qquad (1)$$

Where $\nabla U = (U_x, U_y, U_t)$

Consider a sphere which is a space-time shape of a risk 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.

$$w_\phi(x,y,t) = 1 - \frac{\log(1+\phi(x,y,t))}{\max_{(x,y,t)\in S}(\log(1+\phi(x,y,t)))} \qquad (2)$$

This emphasizes fast moving parts. 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 centres of mass collected from the entire sequence and then by aligning this trajectory to a reference point (similarly to figure centric stabilization in This essentially is equivalent to 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.

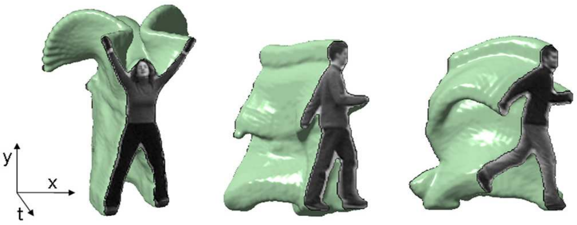
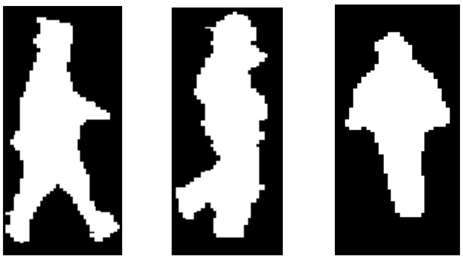
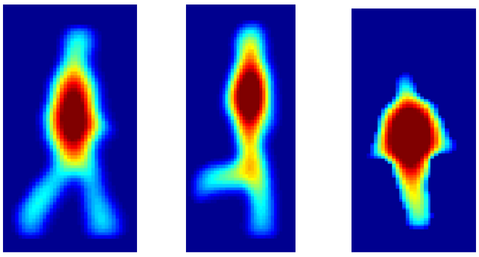


Fig 3 Action as space time shapes



(a) (b) (c)

Fig 4 Extracted Silhouettes Shapes (a) Walk action (b) Run action (c) Jack Action



(a) (b) (c)

Fig5 Poisson Equation Solution on Space-time Shapes. (a) Walk Action (b) Run Action (c) Jack Action.

3.2.2 SPACE –TIME ORIENTATION

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 eigen values are related to the local curvature in the direction of the corresponding eigenvectors and therefore inversely proportional to the length Below, we generalize this approach to space-time.

Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigen values 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 in the area of perceptual grouping, and 3D shape reconstruction, we distinguish between the following three types of local space-time structures:

- $\lambda_1 \approx \lambda_2$ □ λ_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.
- λ_1 □ $\lambda_2 \approx \lambda_3$ 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.

4. GLOBAL FEATURES

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

$$m_{pqr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{\infty}^{\infty} w(x,y,t)g(x,y,t)x^p y^q t^r d_x d_y d_t \qquad (3)$$

Where $w(x,y,t)$ is one of the seven weighting functions. $g(x,y,t)$ denotes the characteristic function of the space time shapes.

5. 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 neighbour 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 neighbour cubes, is fully populated and no sparse, implying that our features emphasize action dynamics, rather than person shape characteristics. The algorithm misclassified 40 out of 789 space-cubes (5.07 percent error rate). Elapsed time is 186.838248 seconds. 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 (4). The algorithm misclassified 73 out of 789 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 no informative 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.

6. RESULTS AND EXPERIMENTS

For action classification and clustering we collected a database of 90 low-resolution (180 × 144, de interlaced 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 colour similarities with the background. In our view, the speed of global translation in the real world (due to different viewpoints 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 translation of the center of mass by aligning the silhouette sequence to a reference point.

	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	10.2	0	2.0	0	0	0
a 3	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	6
a10	0	0	0	4.5	0	0	0	3.6	5.4	86.6

(a)

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	96	0	0	4	0	0	0	0	0	1
a2	0	102	0	0	0	0	0	0	0	0
a3	0	0	42	2	0	0	15	0	0	0
a4	0	0	0	79	0	0	0	0	0	0
a5	0	0	0	0	45	0	0	0	0	0
a6	0	0	0	0	0	51	0	2	0	0
a7	0	0	2	0	2	0	48	0	0	0
a8	0	0	0	0	0	0	0	99	0	0
a9	0	0	0	0	0	0	0	0	101	6
a10	0	0	0	0	0	0	0	0	6	86

(b)

Fig. 6 (a) Action confusion in classification experiment using the method in [16]. (a1-"bend," a2-"jack," a3-"jump," a4-"pjump," a5-run," a6-"side," a7-"skip," a8-"walk," a9-"wave1," and a10-"wave2"). (b) Action confusion in classification experiment using our method.

For each sequence, we solved the Poisson equation using mesh sizes h_s =1, $h_t = 3$ and computed seven types of local features: "stick" and "plate" features, measured at three directions each and the saliency features. In order to treat both the periodic and no 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 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 centred 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 $m_t = 2$ (The maximal order of moments was chosen empirically by testing all possible combinations of m_t and m_s between 1 and 5).

7. 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 neighbours 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 × 70 × 50 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. As our experiments show, the method is robust to significant changes in scale, partial occlusions, and no rigid deformations of the actions.

REFERENCES

[1] 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

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as Space-Time hapes,” Proc. Int’l Conf. Computer Vision, pp. 1395-1402, 2005. 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.

[3] 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.

[4] 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.

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

[6] M. Blank, L. Gorelick, E. Shecht man, M. Irani, and R. Basri, “Actions as Space-Time Shapes,” Proc. Int’l Conf. Computer Vision, pp. 1395-1402, 2005.

[7] S.A. Niyogi and E.H. Adelson, “Analyzing and Recognizing Walking Figures in x,y,t,” Proc. Computer Vision and Pattern Recognition, June 1994.

[8] R. Polana and R.C. Nelson, “Detection and Recognition of Periodic, Nonrigid Motion,” Int’l J. Computer Vision, vol. 23, no. 3, 1997.

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

[10] J. Tangelder and R. Veltkamp, “A Survey of Content Based 3D Shape Retrieval Methods,” Proc. Shape Modeling Int’l, pp. 145-156, 2004.

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

[12] C. Bregler, “Learning and Recognizing Human Dynamics in Video Sequences,” Proc. Computer Vision and Pattern Recognition, June 1997.

[13] S. Carlsson, “Order Structure, Correspondence and Shape Based Categories,” Proc. Int’l Workshop Shape, Contour, and Grouping, p. 1681, 1999.

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

[15] 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.

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