Statistical Analysis of Dynamic Actions

Lihi Zelnik-Manor and Michal Irani, *Member*, *IEEE*

Abstract—Real-world action recognition applications require the development of systems which are fast, can handle a large variety of actions without a priori knowledge of the type of actions, need a minimal number of parameters, and necessitate as short as possible learning stage. In this paper, we suggest such an approach. We regard dynamic activities as long-term temporal objects, which are characterized by spatio-temporal features at multiple temporal scales. Based on this, we design a simple statistical distance measure between video sequences which captures the similarities in their behavioral content. This measure is nonparametric and can thus handle a wide range of complex dynamic actions. Having a behavior-based distance measure between sequences, we use it for a variety of tasks, including: video indexing, temporal segmentation, and action-based video clustering. These tasks are performed without prior knowledge of the types of actions, their models, or their temporal extents.

Index Terms—Action recognition, video indexing, temporal segmentation.

1 Introduction

Dynamic activities can form a powerful cue for analysis of video information, including action-based video indexing, browsing, clustering, and segmentation. Previous work [23], [2], [11], [15], [14], [9], [17], [12], [20] has primarily focused on the recognition of sets of predefined actions, or assumed restricted imaging environments. For example, Ju et al. model and recognize articulated motions [11], Black and Yacoob treat facial expressions [2], and the approaches of Polana and Nelson [17], Cutler and Davis [6], Liu and Picard [12], and of Saisan et al. [20] are designed to detect periodic activities. These methods propose elegant approaches for capturing the important characteristics of these actions by specialized parametric models which usually give rise to high-quality recognition of the studied actions. Their construction, however, is usually done via an extensive learning phase, where many examples of each studied action are provided (often manually segmented and/or manually aligned).

Real-world applications, however, are unlikely to be restricted to recognition of prestudied carefully modeled actions. When dealing with general video data often there is no prior knowledge about the types of actions in the video sequence, their temporal and spatial extent, or their nature (periodic/nonperiodic). A desired application might be for the user who is viewing a movie (e.g., a sports movie), to point out an interesting video segment which contains an action of interest (e.g., a short clip which shows a tennis serve), and request the "system" to fast-forward to the next clip (or find all clips) where a "similar" action occurs. We refer to this as *"action-based video indexing"* or *"Intelligent Fast-Forward."* Another desired application is behavior-based temporal segmentation of long video sequences. Given a long video sequence containing a variety of actions one would like to detect the start-end points of the actions, without requiring any a priori knowledge of the types of actions or their temporal extents. Such applications require developing a notion of behavior-based similarity between video clips which is based on a less-specialized (but also less restrictive) approach to activity modeling.

In this paper, we develop such an approach. We design a similarity measure between video sequences which is based on behavior alone. That is, it preserves the temporal variations, while being insensitive to appearance changes such as varying clothing, lighting conditions, etc. This measure is nonparametric and can thus handle a wide range of dynamic behaviors. It may not be optimal for one specific action, but allows for general behavior-based analysis of video information containing unknown action types.

A preliminary version of this paper appeared in CVPR '01 [24].

2 Related Work

Approaches for modeling actions in nonparametric ways have been previously suggested (e.g., [17], [22], [20], [1], [5], [8]). Most of these were limited to recognition of periodic activities or dynamic textures [17], [22], [20], [1], however, some did aim at recognizing structured activities. Chomat and Crowley [5], represented actions by distributions of motion features at multiple *spatial* scales. Our measurements, on the other hand, are performed at multiple *temporal* scales. We therefore capture *temporal textures* as opposed to "moving spatial textures" which are captured by [5]. In a more recent work, Efros et al. [8] suggested a nonparametric approach to action recognition, based on comparing templates of dense optical flow-fields. This comparison is possible when the figure in the video is in coarse resolution (e.g., a few tens of pixels tall). When these conditions hold, the approach of [8] provides very good action recognition results. However, it cannot handle cases where the actors have different sizes, shapes, or are of different phases of the same motion.

Two highly related lines of work are that using Motion History Images (MHI) [3] and the work of Niyogi and Adelson [15], [14]. Both rely on finding the silhouette in each image and characterizing the action by the properties of the stacked silhouettes. The MHI work [3], [4] generated a 2D template image for each action by overlaying weighted silhouettes, while Niyogi and Adelson [15], [14] tried to characterize the shape of the XYT volume generated by the stacked set of silhouettes. The MHI approach relies on template matching and thus can detect occurrences of a previously learned action; however, it cannot be used for action-based temporal segmentation and clustering where the actions in the video are not known a priori. Both approaches are limited to actions which can be characterized by their silhouettes and cannot handle temporal textures such as flowing water.

3 What Is an Action?

Actions are long-term temporal objects, which usually extend over tens or hundreds of frames. Polana and Nelson [17] separated the class of temporal objects into three groups and suggested separate approaches for modeling and recognizing each: 1) *temporal textures* which have indefinite spatial and temporal extent (e.g., flowing water), see [16], 2) *activities* which are temporally periodic but spatially restricted (e.g., a person walking), see [17], and 3) *motion events* which are isolated actions that do not repeat either in space or in time (e.g., smiling). In this paper, we refer to *temporal actions/behaviors* as all of the above, and would like to treat all of them within a single framework. Nevertheless, most of our experiments focused on the latter two as we find those more interesting.

We next make a set of observations that will affect the design of our action modeling scheme. First, we observe that a representation which can handle the three types of actions/behaviors has to be general, i.e., it cannot make any hard assumptions such as stationarity (which is common in modeling temporal textures) or periodicity (which is common in modeling repetitive activities). Second, a behavior-based representation has to rely on motion-based features which are invariant to changes in appearance such as those caused by different clothes, changes in lighting conditions, variations in background, etc. Last, we note that actions are

- *L. Zelnik-Manor is with the Computer Vision Lab, Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125. E-mail: lihi@vision.caltech.edu.*
- *M. Irani is with the Department of Computer Science and Applied Math, Ziskind Building, Room 224, The Weizmann Institute of Science Rehovot, 76100 Israel. E-mail: michal.irani@weizmann.ac.il.*

Manuscript received 10 Nov. 2004; revised 12 Sept. 2005; accepted 27 Jan. 2006; published online 13 July 2006.
Recommended for acceptance by I.A. Essa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0602-1104.

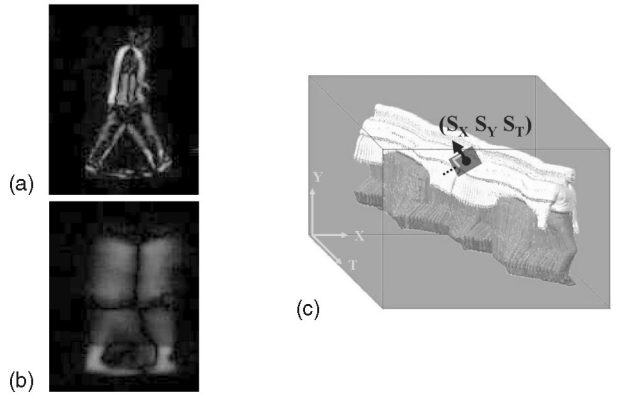

Fig. 1. Representation: The temporal derivative of a walk action at (a) high and (b) low temporal resolutions. (c) The space-time volume S corresponding to a left-to-right walk. The space-time gradient $(S_x, S_y, S_t) = (\frac{\partial S}{\partial x}, \frac{\partial S}{\partial y}, \frac{\partial S}{\partial t})$ is estimated at each space-time point (x, y, t).

characterized by multiple temporal scales. For example, in a sequence of a walking person, the high temporal resolutions will capture the motion of the limbs, whereas the low temporal resolutions will mostly capture the gross movement of the entire body (see Figs. 1a and 1b).

One should note that, although actions are captured at multiple temporal scales, a specific action is always performed at approximately the same speed and, thus, captured at the same temporal scales. For example, a single step of a walking person, viewed by two different video cameras of the same frame rate, will extend over the same number of frames in both sequences, regardless of the internal or external camera parameters. Similarly, a single step of two different people will extend roughly across the same number of frames. This observation implies that when comparing two actions one need not perform comparisons across temporal scales, but rather it suffices to compare actions at corresponding temporal scales. This facilitates the design of a sequence-to-sequence similarity measure.

A scenario which does not comply to the above assumption is when the same action is performed at significantly different speeds. In this case, comparing corresponding temporal scales might be insufficient to detect the similarity between the actions. Note, however, that, in this case, it is not straightforward to say whether indeed the two videos should be considered as capturing the same action or not. For example, are jogging and sprinting the same action? Is a slow dance equivalent to a fast one? If, for the application at hand, the answer is yes, then one would extend the proposed approach to handle those cases by comparing measurements across temporal scales as well.

4 Action Representation and Distance Measure

Based on the above observations, local space-time measurements at multiple *temporal scales* of the video sequence are taken as samples of a stochastic temporal process (the action) and are used to construct an empirical distribution associated with this action. Two actions are considered similar if they could have been generated by the same stochastic process, i.e., if their empirical distributions at corresponding temporal scales are similar. This is explained next.

For obtaining measurements at multiple temporal scales we first construct a *temporal pyramid* of the entire video sequence by blurring and subsampling the sequence along the temporal direction only. The temporal pyramid of a sequence S is thus a pyramid of sequences $S^1(=S), S^2, \ldots, S^L$, where the image frames in all the levels (sequences) within the pyramid are of the same size, and each sequence S^l has half the number of frames of the higher resolution sequence S^{l-1}.

For each level (sequence) S^l in the temporal pyramid, we estimate the local space-time intensity gradient (S^l_x, S^l_y, S^l_t) at all space-time points (See Fig. 1). We then ignore all space-time points for which the temporal derivative is below some threshold, thus using information only from spatio-temporal points which participate in the action. This step can be regarded as a very rough spatial segmentation when the sequence shows a person performing against a static background. For videos displaying temporal textures, such as flowing water, or when the camera is moving, changes occur in most of the space-time points and thus this step has little effect if at all.

The gradient is normal to the local spatio-temporal surface generated by the motion in the space-time sequence volume[1] (at temporal resolution l). Thus, the gradient *direction* captures the local surface orientation, which depends mostly on the local behavioral properties of the moving object, while its *magnitude* depends primarily on the local photometric properties of the moving object and is affected by its spatial appearance (e.g., contrast, color, texture of clothes, illumination, etc.). To preserve the behavioral (orientation) information alone and eliminate as much of the photometric component as possible (the magnitude), we normalize the spatio-temporal intensity gradients to be of length 1. To be invariant to negated contrasts between foreground and background (e.g., a person wearing dark/light clothes against a light/dark background) and to the direction of action (e.g., right-to-left or left-to-right), we further take the absolute value of the normalized space-time gradients. Our local space-time measurements are therefore:

$$(N^l_x, N^l_y, N^l_t) = \frac{(|S^l_x|, |S^l_y|, |S^l_t|)}{\sqrt{(S^l_x)^2 + (S^l_y)^2 + (S^l_t)^2}}, \quad (1)$$

where $l = 1, \ldots, L$ and usually $L = 3$ or 4.

Our action representation would ideally be the $3L$-dimensional distribution of the set of measurements associated with each space-time point across all temporal scales. For example, when using three temporal scales ($L = 3$) each space-time point is associated with a nine-dimensional vector: $[N^1_x, N^1_y, N^1_t, N^2_x, N^2_y, N^2_t, N^3_x, N^3_y, N^3_t]$. To enforce simultaneous occurrence of space-time measurements at multiple temporal scales we would like to construct a multi-dimensional ($9D$) distribution over the set of all vectors. However, multidimensional histograms (e.g., [21]), are computationally intensive and memory-consuming (e.g., for $L = 3$, and assuming 256 bins for each histogram dimension, the size of the multi-dimensional histogram is 256^9).

To handle this curse-of-dimensionality, we suggest a simplified representation, obtained by assuming that all the components of a spatio-temporal point are independent of each other. Taking this assumption, we associate with each action a set of $3L$ one-dimensional empirical distributions $\{p^l_k\}$, one for each component of the space-time measurements $(k = x, y, t)$ at each temporal scale $(l = 1, \ldots, L)$. The empirical distribution p^l_k of measurements N^l_k is represented by a discrete smoothed histogram h^l_k whose integral is normalized to 1. For example, when using three temporal scales an action is represented by a set of nine one-dimensional histograms: $\{h^1_x, h^1_y, h^1_t, h^2_x, h^2_y, h^2_t, h^3_x, h^3_y, h^3_t\}$. We then require the simultaneous occurrence of *distributions* of space-time measurements at multiple temporal scales. The "behavioral" distance between two sequences (S_1 and S_2) is measured by the distances between corresponding empirical distributions of all the components of the space-time measurements at all temporal scales using χ^2 divergence,[2] which are

1. This can be seen from the linear expansion of the brightness constancy equation: Let (dx, dy, dt) denote the local translation of the space-time point (x, y, t). Assuming local brightness constancy yields $S(x, y, t) = S(x + dx, y + dy, t + dt)$. First order Taylor Expansion provides $(S_x, S_y, S_t) \cdot (dx, dy, dt)^T = 0$, i.e., the space-time gradient is orthogonal to the local space-time displacement.

2. We have experimented with various distance measures, such as Jensen-Shannon and Intersection, and all provided similar results. We have selected the χ^2 as it extends naturally to multidimensional histograms.

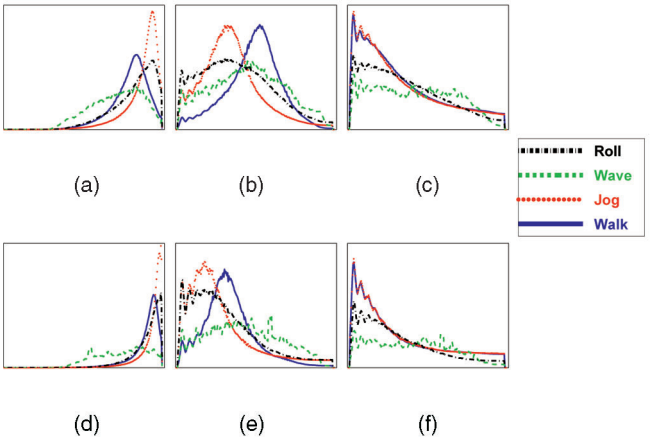

Fig. 2. Distributions of space-timemeasurements: 663 video clips, each 64 frame long, were taken out of the video sequence of Fig. 3, showing different people walking, running, waving, and rolling. The plots above show for each action the mean distribution of each space-time component in two scales. (a) h^0_t, (b) h^0_x, (c) h^0_y, (d) h^1_t, (e) h^1_x, and (f) h^1_y.

added to obtain a single (squared) distance measure between the two sequences:

$$D^2 = \frac{1}{3L}\sum_{k,l,i}\frac{\left[h^l_{1k}(i)-h^l_{2k}(i)\right]^2}{h^l_{1k}(i)+h^l_{2k}(i)} \quad \begin{array}{l} k \in \{x,y,t\} \\ l = 1,\ldots,L \\ i = 1,\ldots,number\ of\ bins. \end{array} \tag{2}$$

Clearly, the strong independence assumption taken above is wrong as the components N^l_k of each spatio-temporal point are dependent by construction. By ignoring this dependence, we implicitly make a new assumption: that the set of one-dimensional empirical distributions suffices to capture the differences between actions. One cannot prove when this claim holds. Instead, we examine the quality of this representation empirically by testing on a variety of action types as described in the next few sections.

To provide a better intuition on what can be captured by the suggested action representation Fig. 2 shows the mean empirical distributions (histograms) computed over multiple video clips of four different actions. Some actions differ in all of the components, e.g., waving and rolling, while others differ only in some components, e.g., walking and running. As will be shown in the following sections, these differences were enough to separate between these actions.

Since our representation is based on distributions, it will be the same for sequences displaying the same action, even when they are of different frame size or when they display a different number of repetitions of the action.

5 Applications

5.1 Action-Based Indexing (Detection)

Given a single example clip of an action of interest and a long test sequence, we wish to detect actions similar to the example action, having no prior information on the content of the test video. This is achieved by comparing the action-of-interest against all subsequences of the long video, with the same temporal length as the example clip (similar to the action detection in [6]). Subsequences with small distance to the given example clip are detected as representing the same action. When an action repeats multiple times consecutive subsequences will be detected as the same action, thus the final detection results can include video segments of various lengths.

5.2 Action-Based Temporal Segmentation

To temporally segment a streaming video, we compare every subsequence of length T to its consecutive subsequence of length T using the distance measure of (2). This results in a set of distance values where maxima points correspond to start-end points of actions. The number of frames T should be loosely related to the length of a single repetition of an action in the sequence.

5.3 Action-Based Clustering

The temporal segmentation scheme described above can be applied online as video streams in. However, when the entire video sequence is available (e.g., in a batch mode) and when the sequence contains multiple occurrences of actions (e.g., the same actions performed at different times), then our action-based distance measure can further be used for grouping actions into action-consistent clusters. We compare every subsequence of length T against *all* other subsequences of length T within the long video sequence to construct a distance matrix. We then use the normalized cut approach of [18] to cluster the data (the number of clusters is determined by the user).

6 Experiments and Results

To evaluate the quality of the suggested approach, we have performed a series of experiments on four video sequences including a variety of action types. The actual videos can be found at: http://www.wisdom.weizmann.ac.il/~vision/EventDetection.html.

- **Periodic activities—"Walk" sequence**. Fig. 3 shows results on a video sequence of several-minutes long (approximately 6,000 frames). The video was recorded outdoors by a stationary video camera. It contains four types of frequently occurring periodic activities: walking, running, hand-waving, and walking-in-place (performed by different people of both genders wearing different clothes for different lengths of time), and single occurrences of several other activities (e.g., rolling and other free activities). Most of the activities were performed parallel to the image plane, but several parts include walking in slightly diagonal directions and some on snake-like paths. Waving includes waving with a single hand or both hands (not necessarily having the same phase). Fig. 3 shows the high-quality results obtained for action-based indexing and clustering using $T = 64$.
- **Nonperiodic activities—"Punch-Kick-Duck" sequence**. Fig. 4 shows results on a video displaying both multiple repetitions and single occurrences of three actions: punch, kick, and duck.
- **Isolated nonperiodic activities—"Tennis" sequence**. Fig. 5 shows the result of applying action-based clustering to a 500-frame long tennis sequence recorded with a panning camera. The sequence was first stabilized to compensate for the camera-induced background motion using [10]. A sliding window of size $T = 10$ was applied to the stabilized sequence. The three detected clusters correspond to *strokes* (backhand and forehand), *hops*, and *steps* of the tennis player. Since our normalized local measurements are invariant to mirror reflections of the same action, the backhand and forehand strokes are clustered together into a single "strokes" class. In this sequence, the strokes appeared with no consecutive repetitions showing the capability of our approach to handle isolated nonperiodic actions.
- **Temporal textures—"Water" sequence**. Fig. 6 further displays the robustness of our approach to classification of temporal textures. The sequence is composed of a mixture of 11 video clips showing water flowing in four different types of motion. The transitions between consecutive clips is an additive fade-in/fade-out effect. Our clustering scheme separated the sequence into the four types of motion.

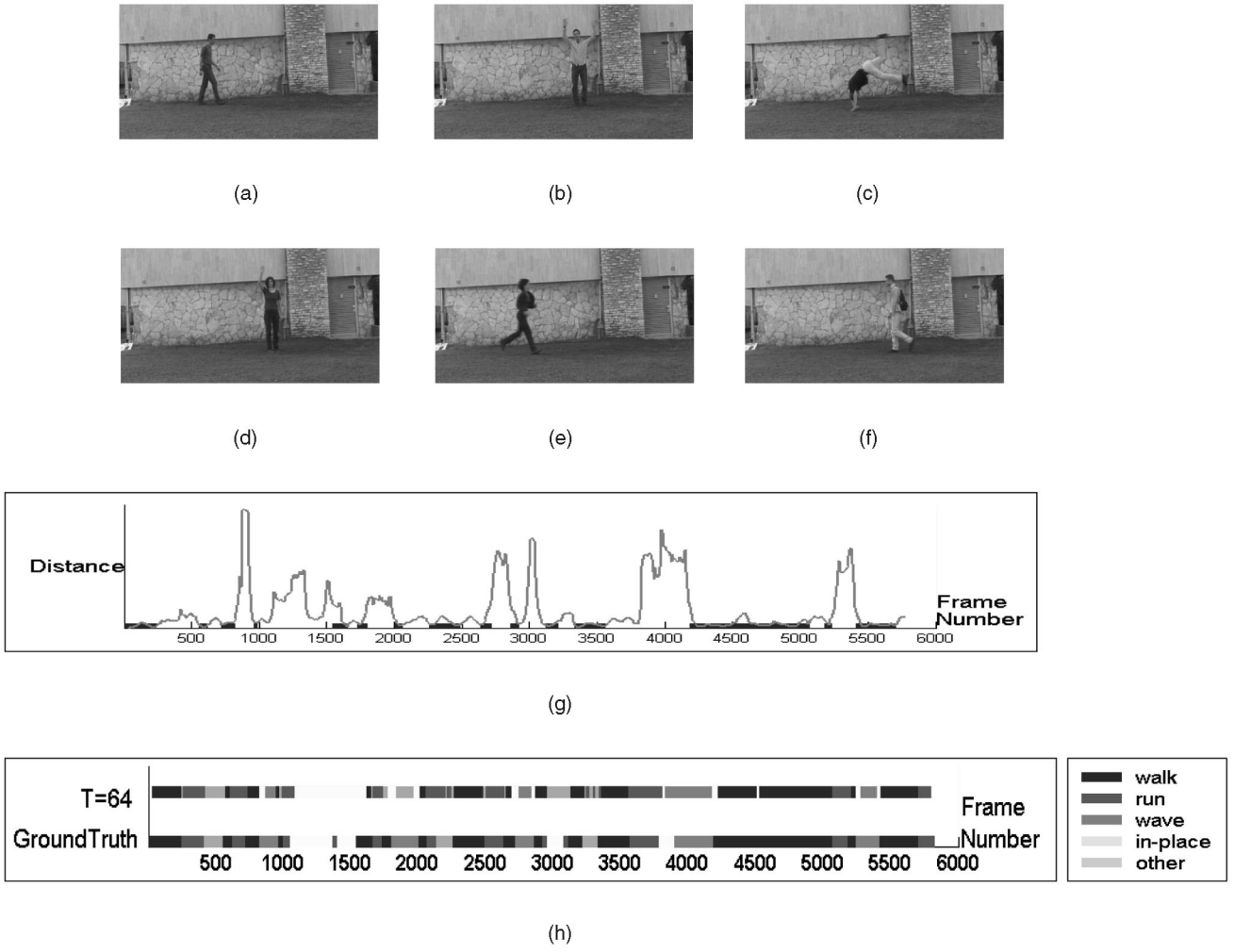

Fig. 3. Periodic activities: (a), (b), (c), (d), (e), and (f) Representative frames of the "Walk" sequence. (g) The measured distances (using (2)) between a single example clip showing a person walking and a sliding window shifted across the entire sequence. The blue bars mark ground-truth for walks. (h) Results of action-based clustering. Top row: All subsequences corresponding to the same cluster were assigned the same color. Bottom row: Manually marked ground-truth information.

7 Evaluating Robustness

We next evaluate the robustness of the suggested action recognition scheme by analyzing its invariance with respect to various parameters.

Invariance to Spatial and Temporal Translations. The spatial position of an object in the image plane has no effect on the local space-time gradients; thus, our distance measure is invariant to translations in the image plane. The suggested representation and distance measure are also invariant to temporal translations. Namely, two actions will be recognized as the same even if they are not temporally synchronized (i.e., if they are out of phase) since the statistics are collected over the entire space-time volume of a clip.

Varying Spatial Scales. Variations in spatial size due to moderate changes in zoom or in distance of the acting person from the video camera have only a little affect on the gradient orientations and, thus, can be handled by our approach. On the other hand, large changes in zoom or in the distance from the camera may change the observed spatial features and, therefore, will affect the gradient orientations. To overcome this problem, one can extend the suggested approach by constructing for each action a representation at multiple spatial scales.

Appearance Changes. Changes in appearance inflicted by different clothing or background will be handled nicely by the suggested approach, as the results show. This invariance will break when extreme changes are involved, for example, when one person wears a highly textured shirt and single color pants while the other wears a single color shirt and highly textured pants. This is since there will be very little gradient information in the single color body part. To overcome this, we first blur the sequences; thus, textured clothes are smeared into single color. We also noted that, in most cases, clothes are not perfectly homogeneous thus sufficient gradient information can be obtained.

Varying Temporal Scales. As discussed in Section 3, we assume that different instances of the same action occur at similar speeds. Nevertheless, our approach proved empirically to be robust to moderate changes in speed. For example, in the "punch-kick-duck" sequence of Fig. 4 the extent of a single repetition of an action varied between 25 and 40 frames. Nevertheless, we were able to correctly classify all of them. We have also tested this on a sequence showing a person walking (at 5, 6, and 7 kmh) and running (at 8 and 9 kmh) on a treadmill (results are not included due to lack of space). Although the difference in speed between the fastest walk and the slowest run was the same as between the two runs, or two walks, an accurate separation between walking and running was obtained.

Changes in Viewing Direction. Significant changes in viewing direction will induce different image motions and, thus, different gradients. Although in theory our distance measure should not be view-point invariant, we have empirically found it to be rather robust to it. This can be seen, for example, in the results obtained for

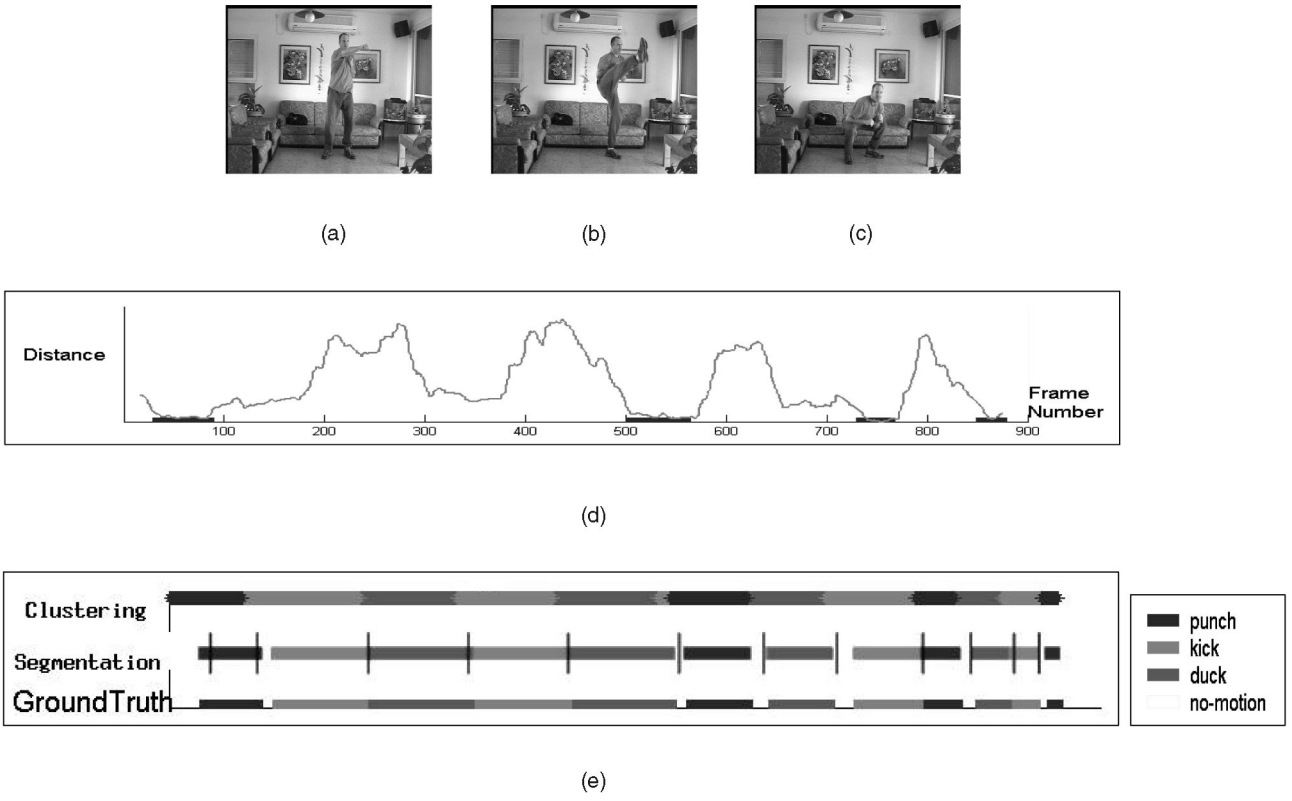

Fig. 4. Nonperiodic actions: (a), (b), and (c) Sample frames from the "punch-kick-duck" sequence. (d) The measured distances between a single punch clip and all other subsequences. The blue bars mark ground-truth for punches. (e) Results of temporal segmentation and clustering using a temporal windows of length $T = 32$ frames. For temporal segmentation, the detected cuts are marked by black vertical lines on top of the ground-truth values.

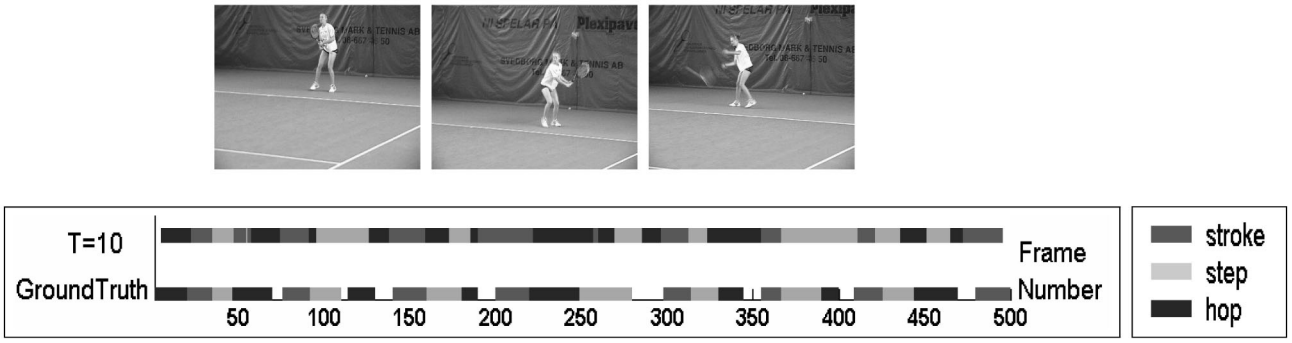

Fig. 5. Isolated nonperiodic actions: Representative frames and clustering result of the "Tennis" sequence.

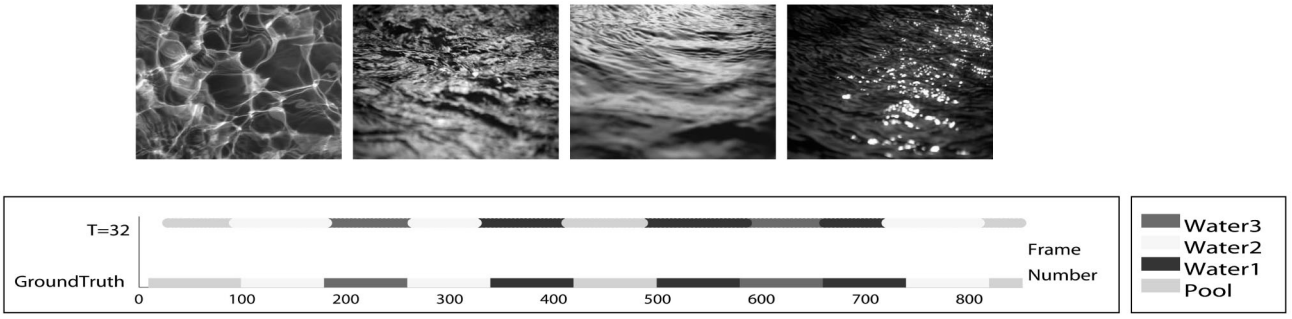

Fig. 6. Temporal textures: Representative frames and clustering result of the "Water" sequence.

the sequences of Figs. 3 and 4 where some of the activities were performed at various angles. An evaluation of robustness to viewpoint changes was additionally performed[3] on a video sequence showing different people walking at various angles with respect to the camera. Applying the clustering process to this sequence while setting the number of clusters to two separated between front view (i.e., walking toward the camera) and side view, although the side view included a high variability in angles (between 30 and 90 degrees to the camera optical axis).

3. Results are not included due to lack of space.

Robustness to Length of Temporal Window. The applications of Section 5 require setting a user-defined length T. We investigated the robustness to variations in T on the "punch-kick-duck" sequence. Using $T = 32$, high-quality results were obtained although the length of a single repetition of an action in this sequence varies between 25 and 40 frames. When using $T = 16$, high-quality clustering results were still obtained, however, when applying temporal segmentation some false temporal cuts were also detected. In most cases, these occurred at the short pauses between repeated occurrences of the same action. A very small temporal window of length $T = 8$ captures too little information about the action, hence, some of the correct cuts were missed, while other false cuts were detected, both when clustering and when segmenting.

Robustness to Changes in Number of Clusters. In the clustering process, we have one extra parameter: the number of clusters. We tested robustness to the choice of the number of clusters on the "punch-kick-duck" sequence. When the number of clusters is set to 2 (instead of the correct number 3) all "punches" and all "kicks" are grouped together into one cluster and the second cluster corresponds to all the "ducks." This result is intuitive since the "punches" and "kicks" are far more similar to each other than to the "ducks." Setting the number of clusters to the wrong number 4 resulted in separating the "punches" into two clusters. One contained all the clips in which the punching person stood frontal to the camera and the other clips with a small angle between the person and the camera.

8 Refining the Representation and Measure

When only a single example clip of an action is available, the action representation is constructed from it. When multiple example clips of the same action are available (either specified manually or obtained via the automatic clustering process), we can refine the action representation to emphasize the contribution of important space-time measurements at the important temporal scales, as learned from the examples.

We first rewrite the χ^2 distance measure using vector notation: $\chi^2 = (h_2 - h_1)^T [diag(h_2 + h_1)]^{-1} (h_2 - h_1)$, where $diag(h_2 + h_1)$ is a diagonal matrix whose ith diagonal entry is $h_2(i) + h_1(i)$ and can be viewed as a weight assigned to the ith histogram bin.

When multiple example clips of the same action type A are available, we compute the *mean* and *covariance* of all the corresponding distributions. The mean histogram $\overline{h_A}$ can be used as the action representation and the covariances cov_A indicate the reliability and the relative significance of the individual histogram bins. When estimating the distance measure between the action A (represented by $\overline{h_A}$) and any new incoming sequence with an empirical distribution h, the weights of bins should be replaced with ${cov_A}^{-1}$. This way, high weights are assigned to bins of low variance (which are more reliable) and low weights to bins of high variance (which are less reliable). The refined distance measure specialized for detecting actions similar to A is therefore: $D_A^2 = \left(h - \overline{h_A}\right)^T {cov_A}^{-1} \left(h - \overline{h_A}\right)$. This is actually the squared Mahalanobis distance [7], applied here to distributions (histograms).

9 Conclusions

The task of action recognition in video is an extremely difficult one. Many approaches thus focus on recognizing a highly limited set of actions, often just a single one like walking. However, real systems will have to extend the variety and number of action types they can handle. In this paper, we presented an approach which aims to go in that direction. By representing actions in a nonparametric way, we were able to utilize a single framework to recognize periodic and nonperiodic activities, isolated occurrences, and multiple-repetitions, as well as handling both "structured" video (e.g., showing people) and dynamic textures (e.g., flowing water). While this provides a "proof-of-concept" that a single system can handle such a wide variety of cases, better methods still need to be developed to allow for further invariance to changes in viewing direction, appearance, distance to the camera, etc.

Acknowledgments

This work was supported by the European Commission Project IST-2000-26001 VIBES and by the Israeli Ministry of Science Grant no. 1229.

References

[1] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Texture Mixing and Texture Movie Synthesis Using Statistical Learning," *IEEE Trans. Visualization and Computer Graphics,* 2001.

[2] M.J. Black and Y. Yacoob, "Recognizing Facial Expressions in Image Sequences Using Local Parametrized Models of Image Motion," *Int'l J. Computer Vision,* vol. 25, no. 1, pp. 23-48, 1997.

[3] A. Bobick and J. Davis, "The Representation and Recognition of Action Using Temporal Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, Mar. 2001.

[4] G. Bradski and J. Davis, "Motion Segmentation and Pose Recognition with Motion History Gradients," *Int'l J. Machine Vision and Applications,* vol. 13, no. 3, pp. 174-184, 2002.

[5] O. Chomat and J.L. Crowley, "Probabilistic Recognition of Activity Using Local Appearance," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1999.

[6] R. Cutler and L. Davis, "Robust Real-Time Periodic Motion Detection, Analysis, and Applications," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 781-796, Aug. 2000.

[7] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. New York: John Wiley and Sons, Inc., 2001.

[8] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," *Proc. Int'l Conf. Computer Vision,* vol. 2, pp. 726-733, Oct. 2003.

[9] D.M. Gavrila and L.S. Davis, "3-D Model-Based Tracking of Humans in Action: A Multi-View Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1996.

[10] M. Irani, B. Rousso, and S. Peleg, "Computing Occluding and Transparent Motions," *Int'l J. Computer Vision,* vol. 12, pp. 5-16, Feb. 1994.

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

[12] F. Liu and R.W. Picard, "Finding Periodicity in Space and Time," *Proc. Int'l Conf. Computer Vision,* Jan. 1998.

[13] C.W. Ngo, T.C. Pong, H. Zhang, and R.T. Chin, "Detection of Gradual Transitions through Temporal Slices Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 36-41, June 1999.

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

[15] S.A. Niyogi and E.H. Adelson, "Analyzing Gait with Spatiotemporal Surfaces," *Proc. Workshop Non-Rigid Motion and Articulated Objects,* Nov. 1994.

[16] R. Polana and R. Nelson, "Recognition of Motion from Temporal Texture," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1992.

[17] R. Polana and R. Nelson, "Detecting Activities," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1993.

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

[19] Y. Rui and P. Anandan, "Segmenting Visual Actions Based on Spatio-Temporal Motion Patterns," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2000.

[20] P. Saisan, G. Doretto, S. Soatto, and Y.N. Wu, "Dynamic Texture Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* Dec. 2001.

[21] B. Schiele and J.L. Crowley, "Probabilistic Object Recognition Using Multidimensional Receptive Field Histograms," *Proc. Int'l Conf. Pattern Recognition,* Aug. 1996.

[22] M. Szummer and R.W. Picard, "Temporal Texture Modeling," *Proc. Int'l Conf. on Image Processing,* Sept. 1996.

[23] 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.

[24] L. Zelnik-Manor and M. Irani, "Event-Based Analysis of Video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. II, pp. 123-130, Dec. 2001.

[25] H. Zhang, A. Kankanhali, and W. Smoliar, "Automatic Partitioning of Full-Motion Video," *Multimedia Systems,* 1993.