Space-Time Shapelets for Action Recognition

Dhruv Batra[1] Tsuhan Chen[1] Rahul Sukthankar[2,1]
batradhruv@cmu.edu tsuhan@cmu.edu rahuls@cs.cmu.edu

[1]Carnegie Mellon University [2]Intel Research Pittsburgh

Abstract

Recent works in action recognition have begun to treat actions as space-time volumes. This allows actions to be converted into 3-D shapes, thus converting the problem into that of volumetric matching. However, the special nature of the temporal dimension and the lack of intuitive volumetric features makes the problem both challenging and interesting. In a data-driven and bottom-up approach, we propose a dictionary of mid-level features called Space-Time Shapelets.[1] *This dictionary tries to characterize the space of local space-time shapes, or equivalently local motion patterns formed by the actions. Representing an action as a bag of these space-time patterns allows us to reduce the combinatorial space of these volumes, become robust to partial occlusions and errors in extracting spatial support. The proposed method is computationally efficient and achieves competitive results on a standard dataset [5].*

1. Introduction

Recognizing actions and activities recorded on a video sequence leads us closer to the ultimate goal of completely understanding the world captured or sampled in that space-time interval. At the same time, these are significant problems and stand-alone applications in their own right, holding potential solutions for improved and universal human-computer interaction, automated health and behaviour monitoring of the sick and elderly, and detection of suspicious activities in surveillance footage. Recent works [5, 11, 22, 24] have demonstrated success by working with the space-time volumes formed by the actions, and treating action recognition as volumetric matching. Of course, as Neumann *et al*. [15] and Boyer *et al*. [23] discuss, actions really reside in a four-dimensional space-time space, while we have access to a three-dimensional volume as a result of the projection being performed in the camera. The existence of resultant self-occlusions and the special nature of one of the dimensions is precisely what makes this problem challenging. Blank *et al*. [5] extend an existing 2-D shape representation technique [9] to 3-D space-time volumes, and then perform nearest neighbour classification on the feature vectors extracted from these volumes. However, this feature extraction involves quantifying and characterizing arbitrary qualities like "stick-ness", "plate-ness" and "ball-ness" of local structures. Bobick and Davis [6] propose an interesting way to combine spatial and motion information with Motion History Images (MHI) and Temporal Templates, but still work with images not volumes. Weinland *et al*. [23] extend their work to volumes but require multiple calibrated cameras. Our work will fit somewhere in the middle ground – we work with space-time volumes, but introduce a dictionary of mid-level local space-time shape/motion descriptors that we call Space-Time Shapelets, which not only scales down the combinatorial space of these volumes, but also makes us robust to partial occlusions and errors in extracting spatio-temporal support.

For the sake of completeness, we would like to refer to another class of approaches that extend the 2-D interest point feature matching paradigm [1, 20] to spatio-temporal interest points and descriptors [21, 13, 8]. However, there are two significant advantages of our approach: i) surveys [20, 14] have shown the non-repeatability of interest points causes major problems for any matching algorithm,[2] and in video slow moving actions or low quality sequences often result in very few or no interest points being found at all,[3] and ii) by working purely with the shape of the volume formed by the action, we are able to focus on the question: "*What* characterizes this action?", and to a large extent ignore the issue of "*Who* performed this action?". As a result, our method is more robust to lighting conditions,

[1] For a discussion on the historical use and interpretations of the term *shapelet*, please see Sabzmeydani and Mori [19].

[2] For an interesting discussion on the use of interest points vs. dense sampling please see Nowak *et al*. [16]

[3] Ke *et al*. [10] show examples of two very simple and commonly occurring motions that fail to produce any interest points using the software from Laptev and Lindeberg [12]

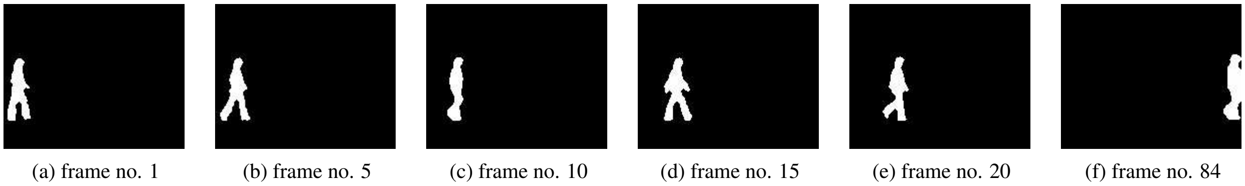


Figure 1: Video Sequence showing a person walking

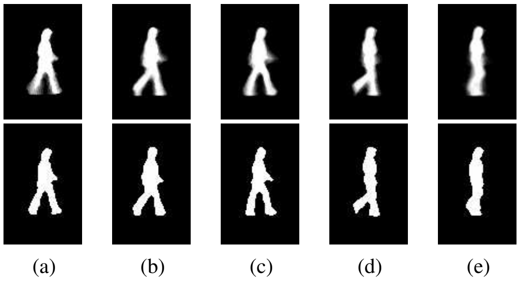


Figure 2: Key Poses: Top row shows the five cluster centers (means), the bottom row shows actual examples closest to cluster centers (pseudo-medians).

clothing worn by the person performing the action, etc. Of course, there exists a trade-off, because this robustness is achieved at the expense of an assumption that we have access to the space-time volume create by an action; but as Blank *et al.* [5] discuss, this is a reasonable assumption to make in many scenarios like surveillance where we have access to a "background" appearance model, and adaptive background subtraction is effective.

The remainder of this paper is organized as follows: Section 2 introduces the idea of Space-Time Shapelets and their application in our overall algorithm; Section 3 describes our experimental setup and results; Section 4 concludes with discussions, and future directions.

2. Space-Time Shapelets

Cognitive studies [4, 7] have shown the wealth of information that humans are able to extract from silhouettes alone: recognizing objects, labeling parts, comparing similarities to other shapes. This motivates the central question of this paper: what could we do with a video of silhouettes? Figure 1 shows a few frames from a video sequence of a person walking. If we were to define a shape as a bias-variance normalized version of these masks, then figure 2 shows the key articulations (using k-means clustering) achieved while performing this action. As useful as this formulation is, there are drawbacks we cannot hope to cope with: i) the space of these global shapes and key articulations is too large, and ii) there is very little generalization across actions, so we would have to recompute these key-positions every time that an action is added to the dataset. Motivated by the success of local shape descriptors [2, 3], we would prefer to describe and characterize local space-time structures rather than global shapes. Thus, in a sense, we wish to extract key articulations of local space-time shapes around *all* points in the video volume. In the same bottom-up data-driven manner as before, we extract all possible MxNxF volumes from the space-time volumes formed by our action dataset, and cluster them to extract Space-Time Shapelets.

Figure 3 shows the extracted Shape-Time Shapelets as 3-D volumes. In order to avoid viewing difficulties due to partial occupation of voxels by cluster centers, and ambiguities caused by shading effects, we have chosen not to display the cluster centers, however, as the following sections describe we work solely with the cluster centers (which are what we refer to by shapelets). There is nothing special about the data-points closest (in euclidean distance sense) to the cluster centers (which we term, *pseudo-medians*), and they are shown in the figure purely for illustrative purposes, to get an idea of what these shapelets might look like, if quantized. These are the local 3-D parts, overlapping chunks of which put together, make up our action volumes. However, this is just one way to interpret these Space-Time Shapelets. Another, more intuitive way is to interpret them as local edge-structures moving across our field of view over time. Figure 4 shows two of these pseudo-medians, only this time, they are accompanied by x-y time slices that make up these volumes. These time slices allow us to visualize the local edge motions captured by these 3-D volumes. Imagine focusing your attention on a small region in 2-D, or equivalently peeping through a window while an action is being performed on the other side. What shapes would one see passing through the field of view during a given time interval? And if we could characterize such structures, would we then be able to recognize the actions being performed on the other side of this "perception wall"? Even if we assume access to multiple such windows (as our algorithm does), this still seems a challenging task – but one that our proposed method accomplishes with a competitive ability.

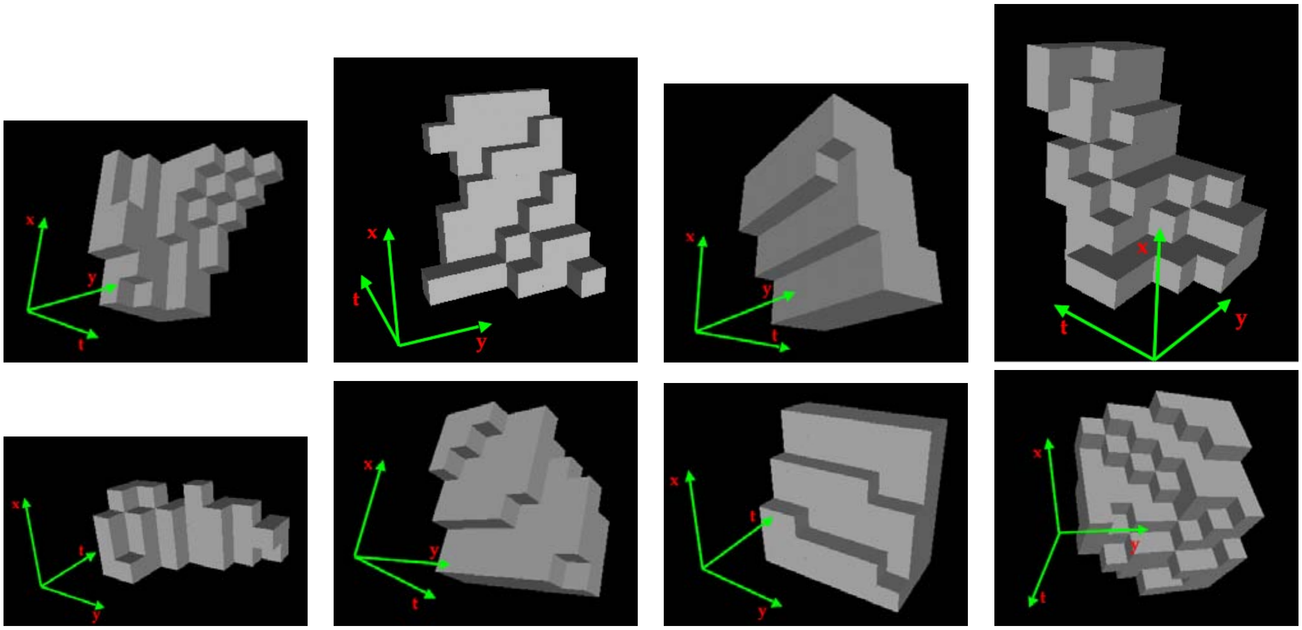


Figure 3: Space-Time Shapelets: Shown are the data-points closest to a few cluster centers (pseudo-medians), created from 7x7x7 volumes. The indicated temporal dimension makes it easier to visualize motion.

2.1. Feature Vectors

Once a dictionary of these Space-Time Shapelets has been established, we represent every voxel in the action volume by the following feature vector, which represents a distribution over the shapelets:

$$f_{\mathcal{D}}(\mathbf{x}) = \begin{bmatrix} p_1 & p_2 & \dots & p_n \end{bmatrix}^T, \tag{1}$$

where

$$\begin{aligned} p_i &= \text{Pr}\left(\text{Vol}(\mathbf{x}) \text{ belongs to shapelet i}\right), & \text{(2a)} \\ &= \text{Pr}\left(Sh_i \mid \text{Vol}(\mathbf{x})\right), & \text{(2b)} \\ &\propto \text{Pr}\left(\text{Vol}(\mathbf{x}) \mid Sh_i\right) \ \text{Pr}\left(Sh_i\right). & \text{(2c)} \end{aligned}$$

The first probability is modeled using an exponential kernel:

$$\text{Pr}\left(\text{Vol}(\mathbf{x}) \mid Sh_i\right) \propto \exp\{-d(\text{Vol}(\mathbf{x}), Sh_i)\}, \tag{3}$$

where, d is the euclidean distance operator, and the marginal is modeled using popularity at the end of the clustering process:

$$\text{Pr}\left(Sh_i\right) = \frac{\#\,\text{members in cluster i}}{\#\,\text{data points}}. \tag{4}$$

Here, Vol$(\mathbf{x})$ is the MxNxF volume centered at the point $\mathbf{x}$, flattened into a column vector. It should be noted that $\mathcal{D}$ indexes the dictionary being worked with.

2.2. Action Representation and Classification

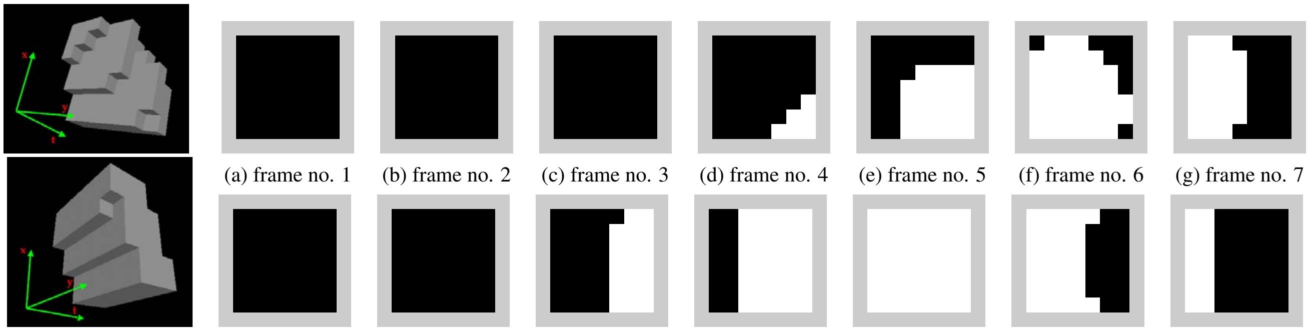
Following a bag-of-words model, we represent an action by a histogram over a dictionary of these space-time shapelets.

$$h_{\mathcal{D}}(V) = 1/n \sum_{\text{voxels}} f_{\mathcal{D}}(\mathbf{x}) \tag{5}$$

In practice, however, in order to reduce computation, the above summation is run only over those voxels that contain at least one foreground and one background voxel in their neighbourhoods, defined by Vol$(\mathbf{x})$. Finally, we use nearest-neighbour and logistic-regression classifiers for classification.

3. Experiments and Results

We test our approach on a recently introduced dataset for action recognition: the Weizmann dataset by Blank *et al.* [5], which consists of 81 video sequences ($180 \times 144, 25$ fps) of nine actions ("run", "walk", "jumping-jack", "jumping-forward", "jumping-in-place", "galloping-sidways", "two-handed-wave", "one-handed-wave"), performed by nine people. In order to create a dictionary of Space-Time Shapelets, we randomly chose half of these videos (while ensuring a uniform sample from each action), and clustered all possible 7x7x7 volumes that contained at least one foreground and background voxel. This collection of more than 100,000 343-dimensional feature vectors were clustered using a publicly available efficient C implemen-



(a) frame no. 1 (b) frame no. 2 (c) frame no. 3 (d) frame no. 4 (e) frame no. 5 (f) frame no. 6 (g) frame no. 7

Figure 4: Unrolling volumes: Each row depicts a shapelet as a volume, and then x-y time slices, or frames that make up these volumes. In the frames, white represents object pixels, black represents background, and gray pixels exist for illustration purposes to provide contrast.

Figure 5: Weizmann Dataset: The rows represent different actions, while the columns show different people performing those actions

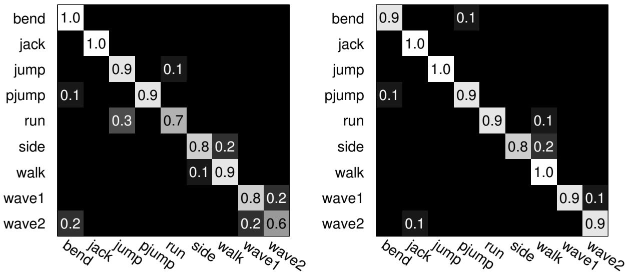
tation of k-means/x-means by Pelleg and Moore [17]. The number of cluster centers was automatically chosen using the x-means algorithm [18] maximizing a Bayesian Information Criterion (BIC) criterion within a range (10 - 50).

3.1. Sequence Classification

For our first experiment, we converted each action sequence into a histogram (as described by equation 5) which is then treated as feature vector for this volume. We then performed leave-one-out-cross-validation (LOOCV) to get the classification rate. Using the 1-Nearest Neighbour classifier, we achieved an accuracy of 82.7%. Figure 6a shows the resulting confusion matrix. We can see that the one-handed-wave and the two-handed-wave are often confused with each other; this is unsurprising considering that both hands in the wave produce similar motion patterns, and our bag-of-features representation is unable to represent the spatial structure, which is precisely the difference between these two actions. However, by using a logistic regression classifier, we were able to increase this accuracy to 91.4%. To this purpose, we trained nine one-vs-rest binary logistic units to model the probability of a class given the feature vector. At test time, we assigned a feature vector the label of the most confident of the nine models (i.e. $\arg\max \Pr$ (Class | feature)). The confusion matrix, using logistic regression is shown in figure 6b. We note that this classifier is able to discriminate between wave1 and wave2 better, and intuitively, we can understand that this is because the weights of the logistic classifier tune to ignore commonly observed motion patterns (and thus similarly populous bins), and focus on the discriminative motion patterns.

3.2. Action Localization

If the video sequences do not contain periodic actions, we can no longer classify entire volumes, and need to perform temporal localization of the action. In order to stay comparable to Blank *et al.* [5], we use a sliding window in time, 10 frames wide, with an overlap of 5 frames, between successive windows. In a similar fashion as the sequence classification, we represented every space-time windowed volume as a histogram over our dictionary of Space-Time Shapelets, and classified these histograms as one of the classes. We again used leave-one-out-cross-validation (LOOCV) to compute a final accuracy, however, this time all windows from the same sequences were also removed and the classifiers were trained on the remaining data. It should be noted that, since we do not apply any temporal smoothness, different subparts of a sequence could in fact be labeled as different actions. This, combined with the fact that we now have more instances to label, would



(a) Using 1-NN

	bend	jack	jump	pjump	run	side	walk	wave1	wave2
bend	1.0								
jack		1.0							
jump			0.9		0.1				
pjump	0.1			0.9					
run			0.3		0.7				
side						0.8	0.2		
walk						0.1	0.9		
wave1								0.8	0.2
wave2	0.2							0.2	0.6

(b) Using LR

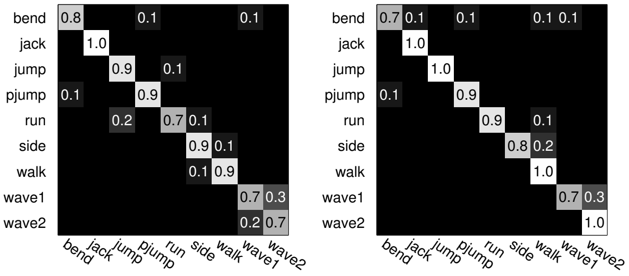
	bend	jack	jump	pjump	run	side	walk	wave1	wave2
bend	0.9			0.1					
jack		1.0							
jump			1.0						
pjump	0.1			0.9					
run					0.9		0.1		
side						0.8	0.2		
walk							1.0		
wave1								0.9	0.1
wave2		0.1							0.9

Figure 6: Sequence Classification results using two different classifiers. We can see improved discriminatory power between one-handed-wave (wave1) and two-handed-wave (wave2).

lead us to expect a worse performance than the sequence classification experiment. However, we found that the performance was actually comparable: the mean accuracy using 1-nearest neighbour was 84.1%, compared to 82.7% in the sequence classification experiment. One possible explanation might be that we now have more data, and an action window only has to be similar to another action *window* (and not the entire sequence) to be labeled that class. Figure 7a shows the confusion matrix achieved by using 1-nearest neighbour classifier. As in the previous experiment, the dominant cause of errors is wave1-wave2 confusion, which we address by using the logistic regression classifier. The mean accuracy with this classifier was 88.2% , and figure 7b shows the confusion matrix.

3.3. Robustness Experiment

In this experiment, we test our approach to establish robustness to viewpoint changes, partial occlusion, background clutter, and other errors cause by more realistic sequences and actions. Figure 8 shows a few examples sequences from the "robust sequences" in the Weizmann dataset [5]. These sequences consist of the walk action being perfomed under different conditions like: "walking in a skirt", "carrying a briefcase", "knees up", "limping walk", "occluded feet", "swinging bag", "sleepwalking", "walking with a dog", "walking past a pole", and nine walk sequences recorded from different viewpoints ($0°$, $9°$, $18°$, $27°$, $36°$, $45°$, $54°$, $72°$, $81°$) with respect to the image plane. Out of these 18 videos, our algorithm was able to label to all but the $72°$ and $81°$ viewpoint sequences correctly, which contain significant scale changes. These results are comparable to those reported by Blank *et al.* [5], who classify all sequences correctly.



(a) Using 1-NN

	bend	jack	jump	pjump	run	side	walk	wave1	wave2
bend	0.8			0.1				0.1	
jack		1.0							
jump			0.9		0.1				
pjump	0.1			0.9					
run			0.2		0.7	0.1			
side						0.9	0.1		
walk						0.1	0.9		
wave1								0.7	0.3
wave2								0.2	0.7

(b) Using LR

	bend	jack	jump	pjump	run	side	walk	wave1	wave2
bend	0.7	0.1		0.1			0.1	0.1	
jack		1.0							
jump			1.0						
pjump	0.1			0.9					
run					0.9		0.1		
side						0.8	0.2		
walk							1.0		
wave1								0.7	0.3
wave2									1.0

Figure 7: Action Localization results using two different classifiers.

4. Conclusions

We looked at the problem of action recognition as that of spatio-temporal volumetric matching. In order to describe these volumes in a discriminative, yet memory and time efficient manner, we represented these volumes as a bag-of-local-parts. Instead of manually defining these parts in a supervised manner, we decided to follow a data-driven approach by introducing Space-Time Shapelets which give us a dictionary of these local spatio-temporal parts, however, we achieved this dictionary in an unsupervised manner. We showed two different interpretations of these Space-Time Shapelets: i.e. the volumetric and local-motion interpretations. Bulk of the computation in our work is done at training time (e.g. extraction of local volume vectors over training data, clustering to create a Shapelet dictionary). At test time, the feature extraction for a $(180 \times 144 \times 50)$ pre-segmented sequence takes 3 - 6 seconds (which is an order of magnitude faster than the Poisson equation solution used by Blank *et al.* [5]) using our unoptimized Matlab® code on a 3.4 GHz Intel® Xeon™ machine. Finally, we evaluated this simple and efficient algorithm a standard action recognition dataset. Although our results are lower than those reported by Blank *et al.* [5], the deficiency can be primarily attributed to confusion between actions such as wave1 and wave2, where the local (shapelet) information is similar and where histogram-based methods are unable to represent global properties of the action. Thus, our results are representative of the trade-off between global and local representations: global representations tend to be richer, but local descriptors are more robust to occlusions and clutter. The Weizmann-robust-dataset gives preliminary indications that this is the case, and we plan to explore the issue in future work using more challenging datasets with significantly more occlusion.

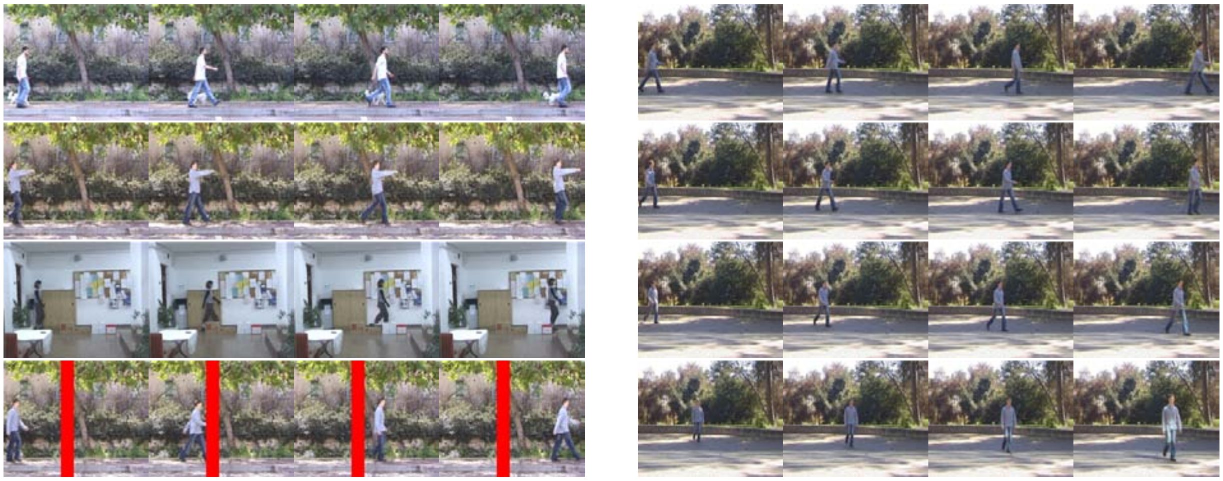


Figure 8: Difficult Sequences: On the left are the sequences – "walking with a dog", "sleepwalking", "occluded feet", "walking past a pole". On the right are sequences with view point changes of $9°$, $27°$, $45°$ and $81°$ (with respect to image plane).

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11):1475–1490, 2004.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[3] A. C. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, pages 607–614, 2001.

[4] I. Biederman. Human image understanding: recent research and a theory. In *WHMV*, pages 13–57, San Diego, CA, USA, 1986.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402. `www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`, 2005.

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

[7] F. Cutzu and M. J. Tarr. Representation of three-dimensional object similarity in human vision. volume 3016, pages 460–471. SPIE, 1997.

[8] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[9] L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt. Shape representation and classification using the poisson equation. *CVPR*, 02:61–67, 2004.

[10] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, volume 1, pages 166 – 173, October 2005.

[11] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007.

[12] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, page 432, Washington, DC, USA, 2003.

[13] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *WSCVMA*, 2004.

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[15] J. Neumann, C. Fermller, and Y. Aloimonos. Animated heads: From 3d motion fields to action descriptions. In *IWDM*, 2000.

[16] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

[17] D. Pelleg and A. Moore. `http://www.cs.cmu.edu/~dpelleg/kmeans.html`. Auton Lab K-means and X-means implementation.

[18] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, pages 727–734, 2000.

[19] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *CVPR*, 2007.

[20] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(2):151–172, 2000.

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[22] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR*, volume 1, pages 405–412, 2005.

[23] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2-3), 2006.

[24] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *CVPR*, pages 984–989, 2005.