'Dynamism of a Dog on a Leash' or Behavior classification by eigen-decomposition of periodic motions.

Roman Goldenberg, Ron Kimmel, Ehud Rivlin, and Michael Rudzsky

Computer Science Department, Technion—Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL

Abstract. Following Futurism, we show how periodic motions can be represented by a small number of eigen-shapes that capture the whole dynamic mechanism of periodic motions. Spectral decomposition of a silhouette of an object in motion serves as a basis for behavior classification by principle component analysis. The boundary contour of the walking dog, for example, is first computed efficiently and accurately. After normalization, the implicit representation of a sequence of silhouette contours given by their corresponding binary images, is used for generating eigen-shapes for the given motion. Singular value decomposition produces these eigen-shapes that are then used to analyze the sequence. We show examples of object as well as behavior classification based on the eigen-decomposition of the binary silhouette sequence.

1 Introduction

Futurism is a movement in art, music, and literature that began in Italy at about 1909 and marked especially by an effort to give formal expression to the dynamic energy and movement of mechanical processes. A typical example is the 'Dynamism of a Dog on a Leash' by Giacomo Balla, who lived during the years 1871-1958 in Italy, see Figure 1 [2]. In this painting one could see how the artist captures in one still image the periodic walking motion of a dog on a leash. Following Futurism, we show how periodic motions can be represented by a small number of eigen-shapes that capture the whole dynamic mechanism of periodic motions. Singular value decomposition of a silhouette of an object serves as a basis for behavior classification by principle component analysis. Figure 2 present a running horse video sequence and its eigen-shape decomposition. One can see the similarity between the overlapped eigen-shapes 2(c) and another futurism style painting "The Red Horseman" by Carlo Carra [2]. The boundary contour of the walking dog, for example, is computed efficiently and accurately by the fast geodesic active contours [15]. After normalization, the implicit representation of a sequence of silhouette contours given by their corresponding binary images, is used for generating eigen-shapes for the given motion. Singular value decomposition produces the eigen-shapes that are used to analyze the sequence. We show examples of object as well as behavior classification based on the eigen-decomposition of the sequence.

Fig. 1. ‘Dynamism of a Dog on a Leash’ 1912, by Giacomo Balla. Albright-Knox Art Gallery, Buffalo.

2 Related work

Motion based recognition received a lot of attention in the last several years. This is due to the general recognition of the fact that the direct use of temporal data may significantly improve our ability to solve a number of basic computer vision problems such as image segmentation, tracking, object classification, etc., as well as the availability of a low cost computer systems powerful enough to process large amounts of data.

In general, when analyzing a moving object, one can use two main sources of information to rely upon: changes of the moving object position (and orientation) in space, and object deformations.

Object position is an easy-to-get characteristic, applicable both for rigid and non-rigid bodies that is provided by most of the target detection and tracking systems, usually as a center of the target bounding box. A number of techniques [17], [16], [11], [26] were proposed for the detection of motion events and for the recognition of various types of motions based on the analysis of the moving object trajectory and its derivatives. Detecting object orientation is a more challenging problem which is usually solved by fitting a model that may vary from a simple ellipsoid [26] to a complex 3D vehicle model [18] or a specific aircraft-class model adapted for noisy radar images as in [9].

While object orientation characteristic is more applicable for rigid objects, it is object deformation that contains the most essential information about the nature of the non-rigid body motion. This is especially true for natural non-rigid objects in locomotion that exhibit substantial changes in their apparent view, as in this case the motion itself is caused by these deformations, e.g. walking, running, hoping, crawling, flying, etc.

There exists a large number of papers dealing with the classification of moving non-rigid objects and their motions, based on their appearance. Lipton et al. describe a method for moving target classification based on their static appearance [19] and using the skeletonization [13]. Polana and Nelson [24] used local motion statistics

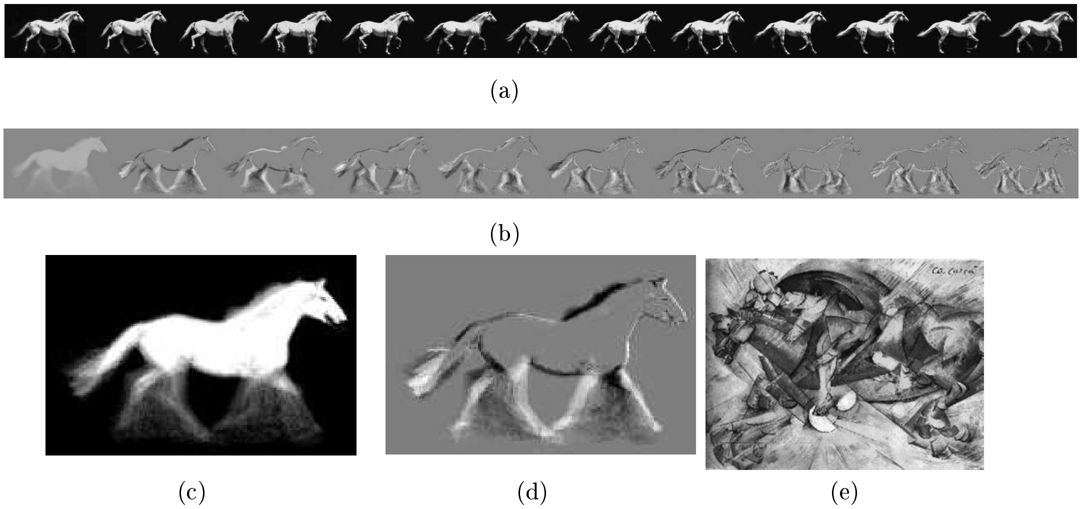


Fig. 2. (a) running horse video sequence, (b) first 10 eigen-shapes, (c,d) first and second eigen-shapes enlarged, (e) 'The Red Horseman', 1914, by Carlo Carra, Civico Museo d'Arte Contemporanea, Milan.

computed for image grid cells to classify various types of activities. An original approach using the temporal templates and motion history images (MHI) for action representation and classification was suggested by Davis and Bobick in [3]. Cutler and Davis [10] describe a system for real-time moving object classification based on periodicity analysis. It would be impossible to describe here the whole spectrum of papers done in this field and we refer the reader to the following surveys [5], [14] and [21].

The most related to our approach is a work by Yacoob and Black [28], where different types of human activities were recognized using a parameterized representation of measurements collected during one motion period. The measurements were eight motion parameters tracked for five body parts (arm, torso, thigh, calf and foot).

In this paper we concentrate on the analysis of the deformations of moving non-rigid bodies in an attempt to extract characteristics that allow us to distinguish between different types of motions and different classes of objects.

3 Our approach

Our basic assumption is that for any given class of moving objects, like humans, dogs, cats, and birds, the apparent object view in every phase of its motion can be encoded as a combination of several basic body views or configurations. Assuming that a living creature exhibits a pseudo-periodic motion, one motion period can be used as a comparable information unit. Then, by extracting the basic views from a large training set and projecting onto them the observed sequence of object views collected from one motion period, we obtain a parameterized representation of object's motion that can be used for classification.

Unlike [28] we do not assume an initial segmentation of the body into parts and do not explicitly measure the motion parameters. Instead, we work with the changing apparent view of deformable objects and use the parameterization induced by their form variability.

In what follows we describe the main steps of the process that include,

- Segmentation and tracking of the moving object that yield an accurate external object boundary in every frame.
- Periodicity analysis, in which we estimate the frequency of the pseudo-periodic motion and split the video sequence into single-period intervals.
- Frame sequence alignment that brings the single-period sequences above to a standardized form by compensating for temporal shift, speed variations, different object sizes and imaging conditions.
- Parameterization by building an eigen-shape basis from a training set of possible object views and projecting the apparent view of a moving body onto this basis.

3.1 Segmentation and Tracking

As our approach is based on the analysis of deformations of the moving body, the accuracy of the segmentation and tracking algorithm in finding the target outline is crucial for the quality of the final result. This rules out a number of available or easy-to-build tracking systems that provide only a center of mass or a bounding box around the target and calls for more precise and usually more sophisticated solutions.

Therefore we decided to use the geodesic active contour approach [4] and specifically the 'fast geodesic active contour' method described in [15], where the segmentation problem is expressed as a geometric energy minimization. We search for a curve C that minimizes the functional

$$S[\mathcal{C}] = \int_0^{L(\mathcal{C})} g(\mathcal{C}) ds,$$

where ds is the Euclidean arclength, $L(\mathcal{C})$ is the total Euclidean length of the curve, and g is a positive edge indicator function in a 3D hybrid spacial-temporal space that depends on the pair of consecutive frames $I^{t-1}(x,y)$ and $I^t(x,y)$. It gets small values along the spacial-temporal edges, i.e. moving object boundaries, and higher values elsewhere.

In addition to the scheme described in [15], we also use the background information whenever a static background assumption is valid and a background image $B(x,y)$ is available. In the active contours framework this can be achieved either by modifying the g function to reflect the edges in the difference image $D(x,y) = |B(x,y) - I^t(x,y)|$, or by introducing additional area integration terms to the functional $S(\mathcal{C})$:

$$S[\mathcal{C}] = \int_0^{L(\mathcal{C})} g(\mathcal{C}) ds + \lambda_1 \int_{inside(\mathcal{C})} |D(x,y) - c_1|^2 da + \lambda_2 \int_{outside(\mathcal{C})} |D(x,y) - c_2|^2 da,$$

where λ_1 and λ_2 are fixed parameters and $c1$, $c2$ are given by:

$$c_1 = average_{inside(\mathcal{C})}[D(x, y)]$$
$$c_2 = average_{outside(\mathcal{C})}[D(x, y)]$$

The latter approach is inspired by the 'active contours without edges' model proposed by Chan and Vese [6] and forces the curve $\mathcal{C}$ to close on a region whose interior and exterior have approximately uniform values in $D(x, y)$. A different approach to utilize the region information by coupling between the motion estimation and the tracking problem was suggested by Paragios and Deriche in [22].

Figure 3 shows some results of moving object segmentation and tracking using the proposed method.

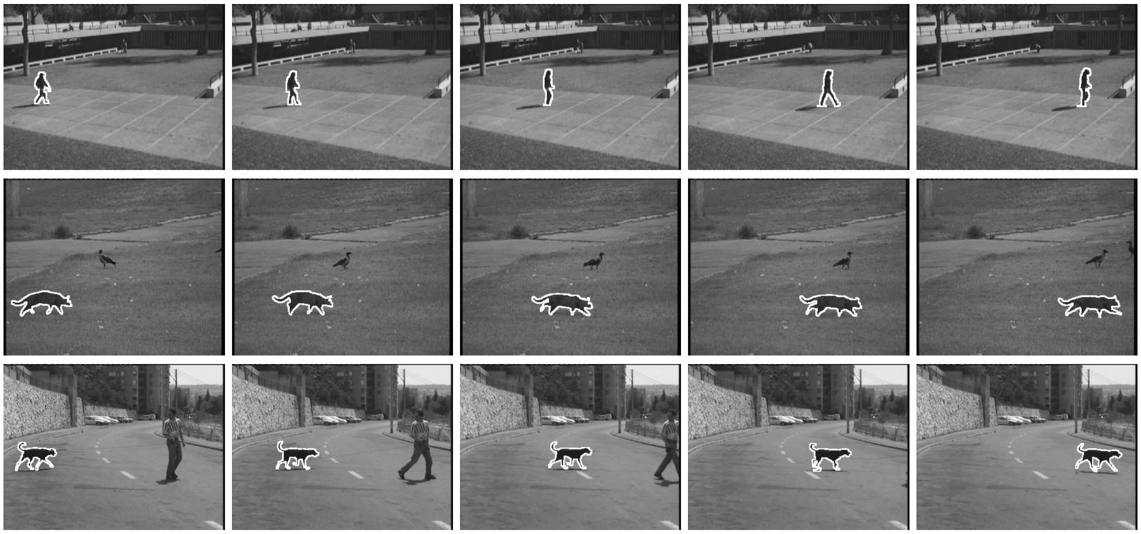


Fig. 3. Non-rigid moving object segmentation and tracking.

3.2 Periodicity analysis

Here we assume that the majority of non-rigid moving objects are self-propelled alive creatures whose motion is almost periodic. Thus, one motion period, like a step of a walking man or a rabbit hop, can be used as a natural unit of motion and extracted motion characteristics can by normalized by the period size.

The problem of detection and characterization of periodic activities was addressed by several research groups and the prevailing technique for periodicity detection and measurements is the analysis of the changing 1-D intensity signals along spatio-temporal curves associated with a moving object or the curvature analysis of feature point trajectories [23], [20], [25], [27]. Here we address the problem using global characteristics of motion such as moving object contour deformations and the trajectory of the center of mass.

By running frequency analysis on such 1-D contour metrics as the contour area, velocity of the center of mass, principal axes orientation, etc. we can detect the basic period of the motion. Figures 4 and 5 present global motion characteristics derived from segmented moving objects in two sequences. One can clearly observe the common dominant frequency in all three graphs.

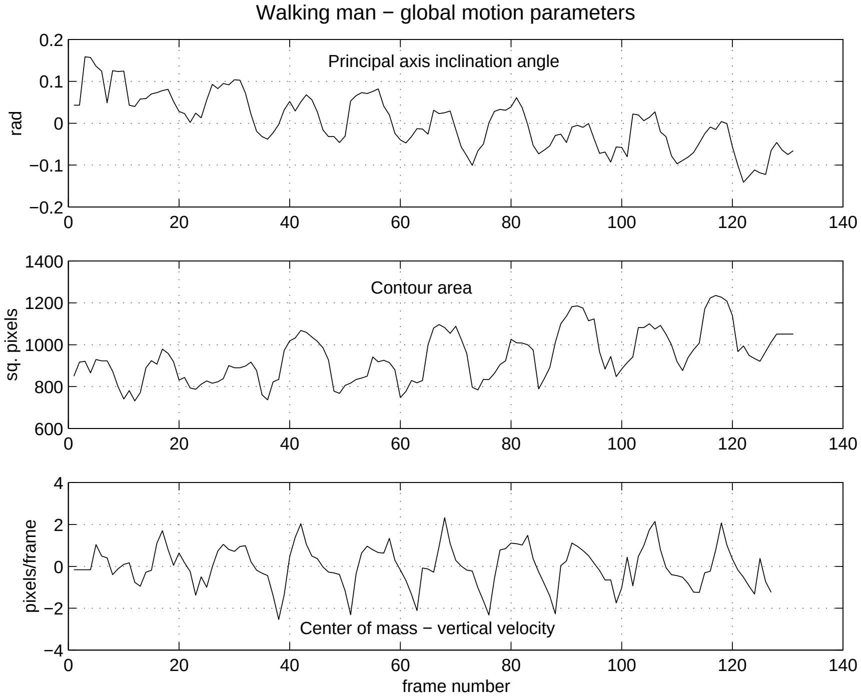


Fig. 4. Global motion characteristics measured for walking man sequence.

The period can also be estimated in a straightforward manner by looking for the frame where the external object contour best matches the object contour in the current frame. Figure 6 shows the deformations of a walking man contour during one motion period (step). Samples from two different steps are presented and each vertical pair of frames is phase synchronized. One can clearly see the similarity between the corresponding contours. An automated contour matching can be performed in a number of ways, e.g. by comparing contour signatures or by looking at the correlation between the object silhouettes in different frames. Figure 7 shows four graphs of inter-frame silhouette correlation values measured for four different starting frames taken within one motion period. It is clearly visible that all four graphs nearly coincide and the local maxima peaks are approximately evenly spaced. The period, therefore, can be estimated as the average distance between the neighboring peaks.

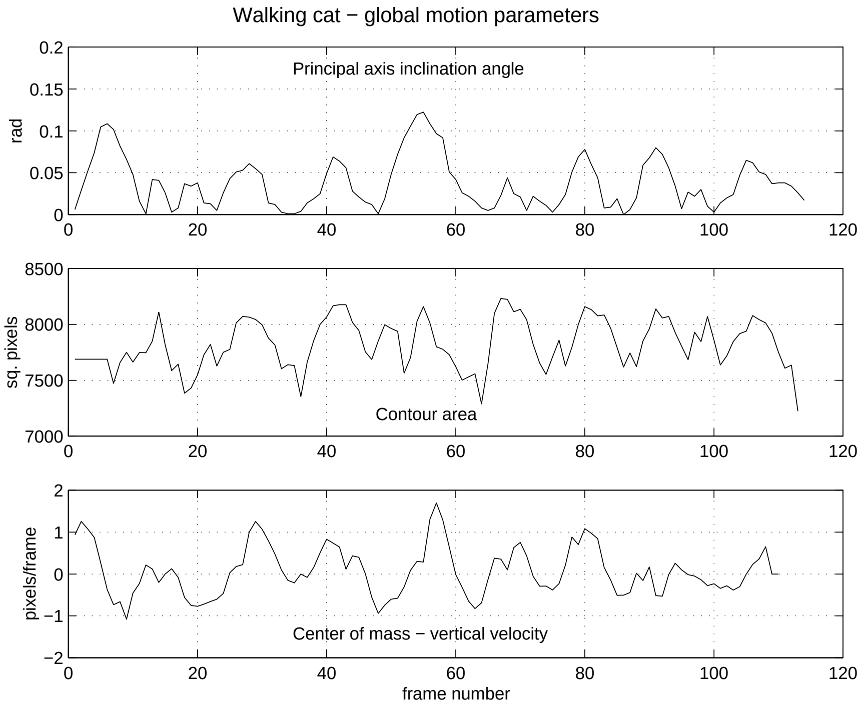


Fig. 5. Global motion characteristics measured for walking cat sequence.

3.3 Frame sequence alignment

One of the most desirable features of any classification system is the invariance to a set of possible input transformations. As the input in our case is not a static image, but a sequence of images, the system should be robust to both spacial and temporal variations.

Spatial alignment: Scale invariance is achieved by cropping a square bounding box around the center of mass of the tracked target silhouette and re-scaling it to a predefined size (see Figure 8).

One way to have orientation invariance is to keep a collection of motion samples for a wide range of possible motion directions and then look for the best match. This approach was used by Yacoob and Black in [28] to distinguish between different walking directions. Although here we experiment only with motions nearly parallel to the image plane, the system proved to be robust to small variations in orientation. Since we do not want to keep models for both left-to-right and right-to-left motion directions, the right-to-left moving sequences are converted to left-to-right by horizontal mirror flip.

Temporal alignment: A good estimate of the motion period allows us to compensate for motion speed variations by re-sampling each period subsequence to a

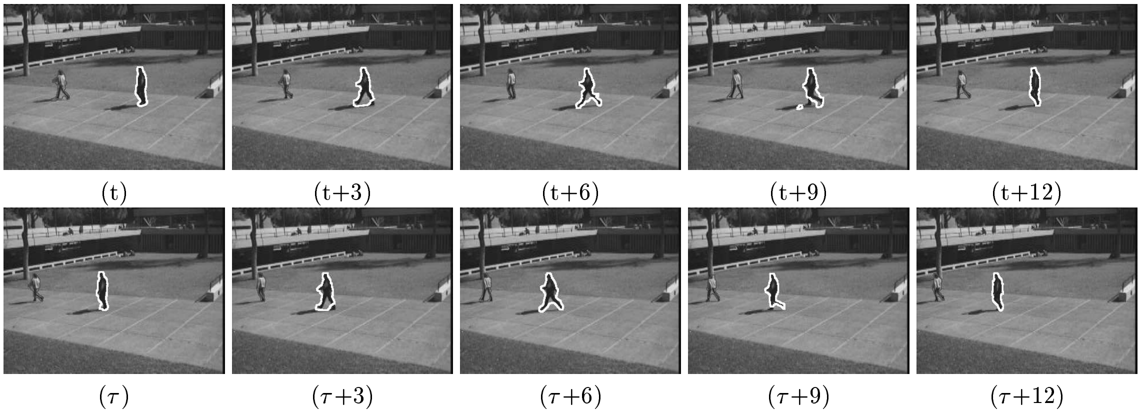


Fig. 6. Deformations of a walking man contour during one motion period (step). Two steps synchronized in phase are shown. One can see the similarity between contours in corresponding phases.

predefined duration. This can be done by interpolation between the binary silhouette images themselves or between their parameterized representation as explained below. Figure 9 presents an original and re-sampled one-period subsequence after scaling from 11 to 10 frames.

Temporal shift is another issue that has to be addressed in order to align the phase of the observed one-cycle sample and the models stored in the training base. In [28] it was done by solving a minimization problem of finding the optimal parameters of temporal scaling and time shift transformations so that the observed sequence is best matched to the training samples. Polana and Nelson [24] handled this problem by matching the test one-period subsequence to reference template at all possible temporal translations.

Assuming that in the training set all the sequences are accurately aligned, we find the temporal shift of a test sequence by looking for the starting frame that best matches the generalized (averaged) starting frame of the training samples, as they all look alike. Figure 10 shows (a) - the reference starting frame taken as an average over the temporally aligned training set, (b) - a re-sampled single-period test sequence and, (c) the correlation between the reference starting frame and the test sequence frames. The maximal correlation is achieved at the seventh frame, therefore the test sequence is aligned by cyclically shifting it 7 frames to the left.

3.4 Parameterization

In order to reduce the dimensionality of the problem we first project the object image in every frame onto a low dimensional base that represents all possible appearances of objects that belong to a certain class, like humans, four-leg animals, etc.

Let n be number of frames in the training base of a certain class of objects and M be a training samples matrix, where each column corresponds to a spatially aligned image of a moving object written as a binary vector. In our experiments we use 50×50 normalized images, therefore, M is a $2500 \times n$ matrix. Matrix M

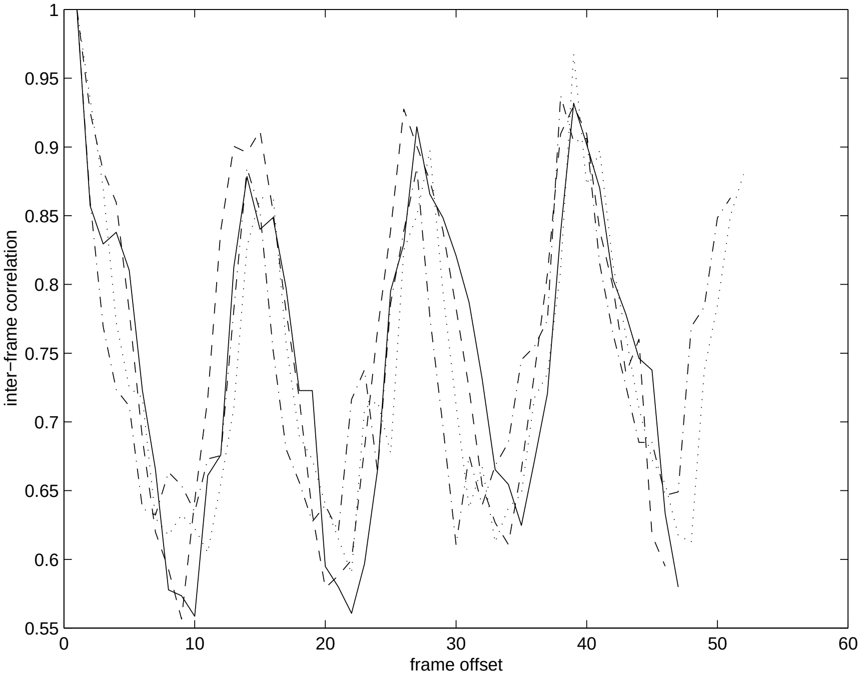


Fig. 7. Inter-frame correlation between object silhouettes. Four graphs show the correlation measured for four initial frames.

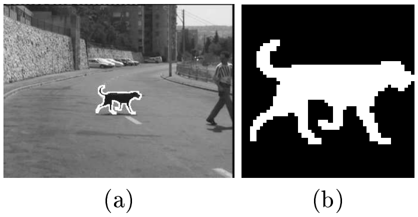


Fig. 8. Scale alignment. A minimal square bounding box around the center of the segmented object silhouette (a) is cropped and re-scaled to form a 50×50 binary image (b).

is decomposed using Singular Value Decomposition as $M = U\Sigma V^T$, where U is an orthogonal matrix of principal directions and the Σ is a diagonal matrix of singular values. The principal basis $\{U_i, i = 1..k\}$ for the training set is then taken as k columns of U corresponding to the largest singular values in Σ. Figure 11 presents a principal basis for the training set formed of 800 sample images collected from more than 60 sequences showing dogs and cats in motion. The basis is built by taking the $k = 20$ first principal component vectors.

We assume that by building such representative bases for every class of objects and then finding the basis that best represents a given object image in a minimal distance to the feature space (DTFS) sense, we can distinguish between various object classes. Figure 12 shows the distances from more than 1000 various images of people, dogs and cats to the feature space of people and to that of dogs and cats.

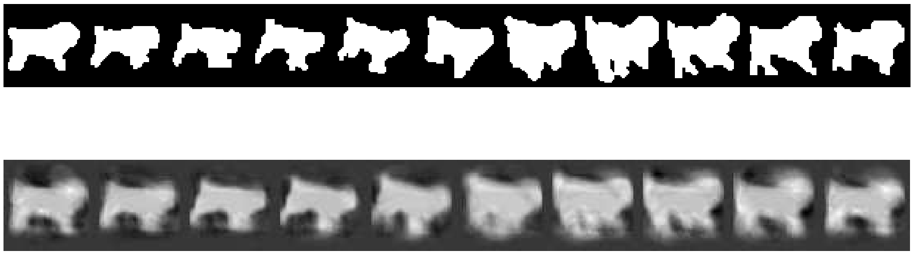


Fig. 9. Temporal alignment. Top: original 11 frames of one period subsequence. Bottom: re-sampled 10 frames sequence.

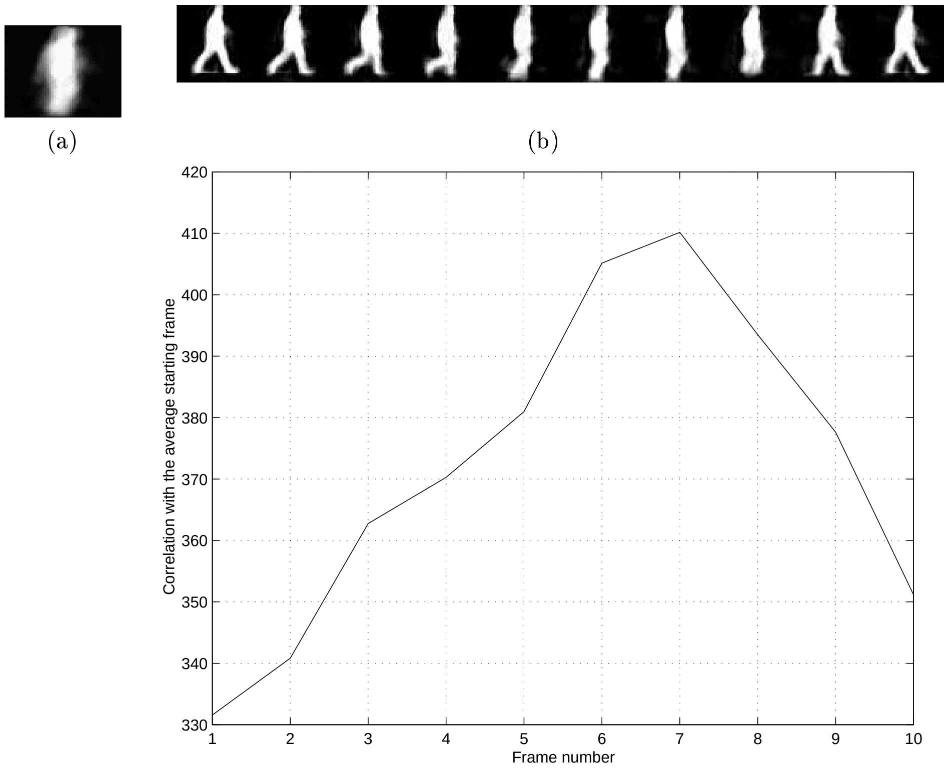
In all cases, images of people were closer to the people feature space than to the animals' feature space and vise a versa. This allows us to distinguish between these two classes. A similar approach was used in [12] for the detection of pedestrians in traffic scenes.

If the object class is known (e.g. we know that the object is a dog), we can parameterize the moving object silhouette image I in every frame by projecting it onto the class basis. Let B be the basis matrix formed from the basis vectors $\{U_i, i = 1..k\}$. Then, the parameterized representation of the object image I is given by the vector $\overline{p}$ of length k as $\overline{p} = B^T \overline{v_I}$, where $\overline{v_I}$ is the image I written as a vector.

The idea of using a parameterized representation in motion-based recognition context is certainly not a new one. To name a few examples we mention again the work of Yacoob and Black [28]. Cootes et al. [8] used similar technique for describing feature point locations by a reduced parameter set. Baumberg and Hogg [1] used PCA to describe a set of admissible B-spline models for deformable object tracking. Chomat and Crowley [7] used PCA-based spatio-temporal filter for human motion recognition.

Figure 13 shows several normalized moving object images from the original sequence and their reconstruction from a parameterized representation by back-projection to the image space. The numbers below are the norms of differences between the original and the back-projected images. These norms can be used as the DTFS estimation.

Now, we can use these parameterized representations to distinguish between different types of motion. The reference base for the activity recognition consists of temporally aligned one-period subsequences, whereas the moving object silhouette in every frame of these subsequences is represented by its projection to the principal basis. More formally, let $\{I_f : f = 1..T\}$ be a one-period, temporally aligned set of normalized object images, and $\overline{p_f}, f = 1..T$ a projection of the image I_f onto the principal basis B of size k. Then, the vector P of length kT formed by concatenation of all the vectors $\overline{p_f}, f = 1..T$, represent a one-period subsequence. By choosing a basis of size $k = 20$ and the normalized duration of one-period subsequence to be



(a) (b)

(c)

Fig. 10. Temporal shift alignment: (a) - average starting frame of all the training set sequences, (b) - temporally shifted single-cycle test sequence, (c) - the correlation between the reference starting frame and the test sequence frames

$T = 10$ frames, every single-period subsequence is represented by a feature point in a 200-dimensional feature space.

In the following experiment we processed a number of sequences of dogs and cats in various types of locomotion. From these sequences we extracted 33 samples of walking dogs, 9 samples of running dogs, 9 samples with galloping dogs and 14 samples of walking cats. In Figure 14 we depict the resulting feature points projected for visualization to the 3-D space using the three first principal directions . One can easily observe four separable clusters corresponding to the four groups.

Another experiment was done over the ‘people’ class of images. Figure 15 presents feature points corresponding to several sequences showing people walking and running parallel to the image plane and running at oblique angle to the camera. Again, all three groups lie in separable clusters.

The classification can be performed, for example, using the k-nearest-neighbor algorithm. We conducted the ‘leave one out’ test for the dogs set above, classifying every sample by taking them out from the training set one at a time, and the three-nearest-neighbors strategy resulted in 100% success rate.

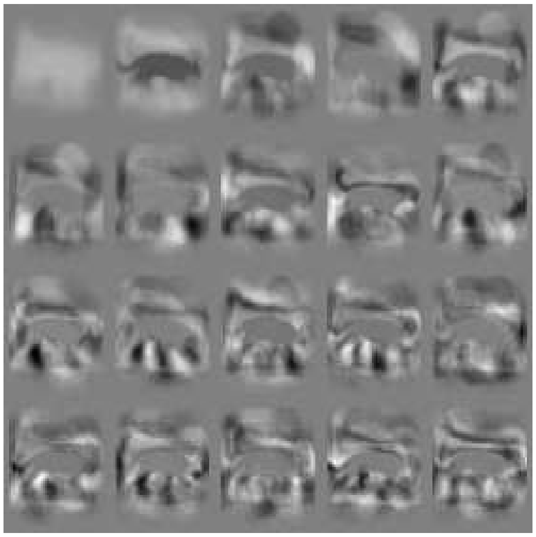


Fig. 11. The principal basis for the 'dogs and cats' training set formed of 20 first principal component vectors.

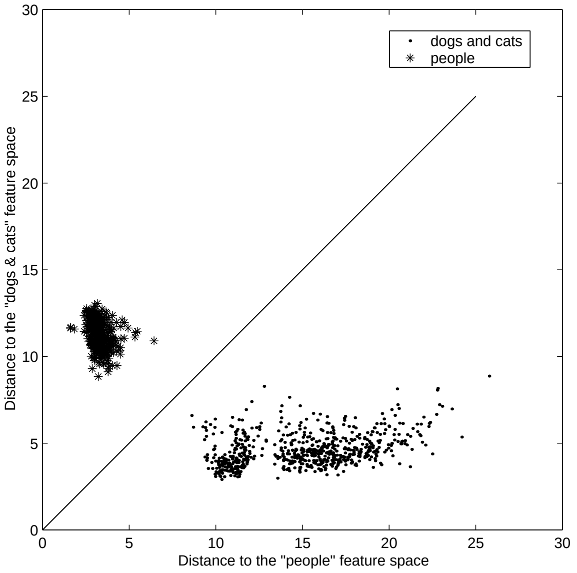


Fig. 12. Distances to the 'people' and 'dogs and cats' feature spaces from more than 1000 various images of people, dogs and cats.

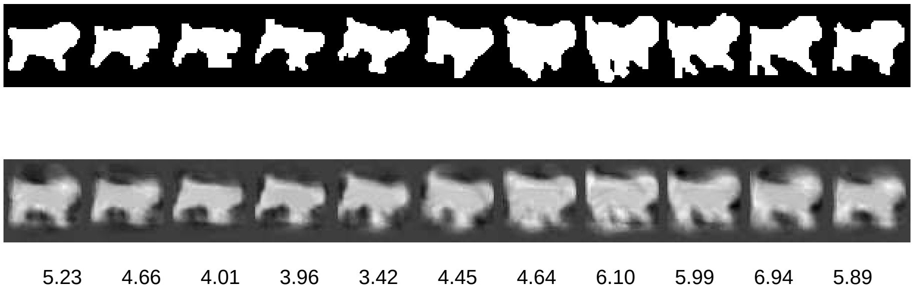


Fig. 13. Image sequence parameterization. Top: 11 normalized target images of the original sequence. Bottom: the same images after the parameterization using the principal basis and back-projecting to the image basis. The numbers are the norms of the differences between the original and the back-projected images.

4 Concluding remarks

We presented a new framework for motion-based classification of moving non-rigid objects. The technique is based on the analysis of changing appearance of moving objects and is heavily relying on high accuracy results of segmentation and tracking by using the fast geodesic contour approach. The periodicity analysis is then performed based on the global properties of the extracted moving object contours, followed by video sequence spatial and temporal normalization. Normalized one-period subsequences are parameterized by projection onto a principal basis extracted from a training set of images for a given class of objects. A number of experiments show the ability of the system to analyze motions of humans and animals, to distinguish between these two classes based on object appearance, and to classify various type of activities with a class, such as walking, running, galloping. The 'dogs and cats' experiment demonstrate the ability of the system to discriminate between these two very similar by appearance classes by analyzing their locomotion.

References

1. A Baumberg and D Hogg. An efficient method for contour tracking using active shape models. In *In Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, Austin, 1994.
2. J R Beniger. The Arts and New Media site. In *www.usc.edu/schools/annenberg/asc/projects/comm544/*, University of South California, Annenberg School for Communication.
3. A Bobick and J Davis. The representation and recognition of action using temporal templates. *IEEE Trans. on PAMI*, 23(3):257–267, 2001.
4. V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *IJCV*, 22(1):61–79, 1997.
5. C Cedras and M Shah. Motion-based recognition: A survey. *IVC*, 13(2):129–155, March 1995.
6. T F Chan and L A Vese. Active contours without edges. *IEEE trans. on Image Processing*, 10(2):266–277, February 2001.
7. O Chomat and J Crowley. Recognizing motion using local appearance, 1998.
8. T F Cootes, C J Taylor, D H Cooper, and J Graham. Active shape models: Their training and application. *CVIU*, 61(1):38–59, January 1995.
9. N J Cutaia and J A O'Sullivan. Automatic target recognition using kinematic priors. In *Proceedings of the 33rd Conference on Decision and Control*, pages 3303–3307, Lake Buena Vista, FL, December 1994.

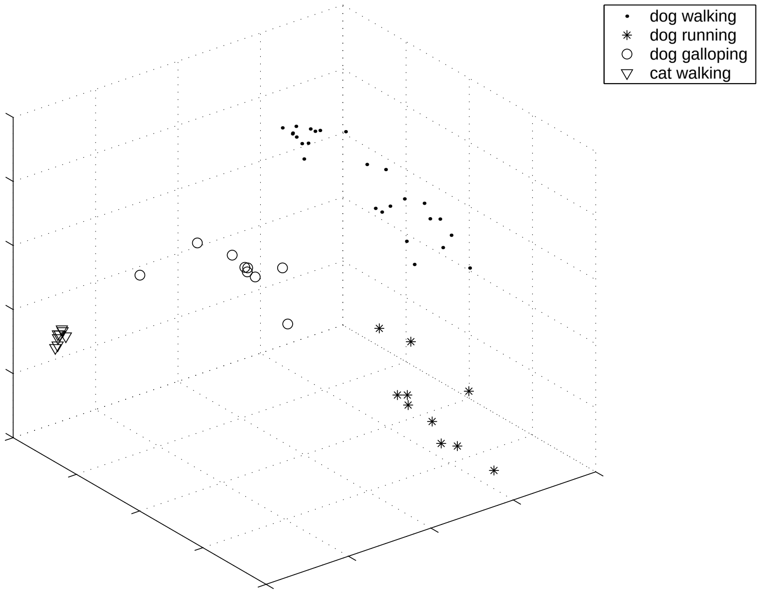


Fig. 14. Feature points extracted from the sequences with walking, running and galloping dogs and walking cats and projected to the 3-D space for visualization

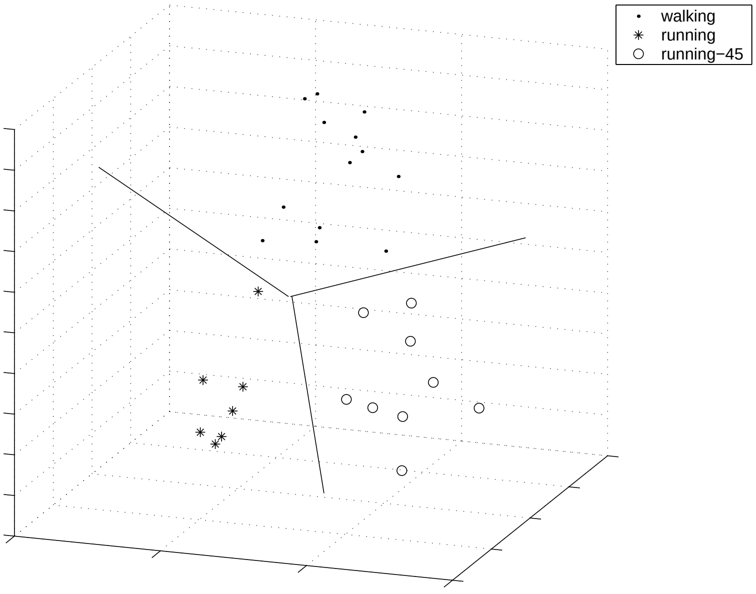


Fig. 15. Feature points extracted from the sequences showing people walking and running parallel to the image plane and at 45 degrees angle to the camera. Feature points are projected to the 3-D space for visualization

10. R Cutler and L Davis. Robust real-time periodic motion detection, analysis, and applications. *PAMI*, 22(8):781–796, August 2000.
11. S A Engel and J M Rubin. Detecting visual motion boundaries. In *Proc. Workshop on Motion: Representation and Analysis*, pages 107–111, Charleston, S.C., May 1986.
12. U Franke, D Gavrila, S Gorzig, F Lindner, F Paetzold, and C Wohler. Autonomous driving goes downtown. *IEEE Intelligent System*, 13(6):40–48, 1998.
13. H Fujiyoshi and A Lipton. Real-time human motion analysis by image skeletonization. In *Proc. of the Workshop on Application of Computer Vision*, October 1998.
14. D M Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, January 1999.
15. R Goldenberg, R Kimmel, E Rivlin, and M Rudzsky. Fast geodesic active contours. *IEEE Trans. on Image Processing*, 10(10):1467–75, October 2001.
16. K Gould, K Rangarajan, and M Shah. Detection and representation of events in motion trajectories. In *Advances in Image Processing and Analysis, chapter 14. SPIE Optical Engineering Press*, June 1992. Gonzalez and Mahdavieh (Eds.).
17. K Gould and M Shah. The trajectory primal sketch: a multi-scale scheme for representing motion characteristics. In *Proc. Conf. on Computer Vision and Pattern Recognition, San Diego, CA*, pages 79–85, 1989.
18. D Koller, K Daniilidis, and H-H Nagel. Model-based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision*, 10:257–281, 1993.
19. A Lipton, H Fujiyoshi, and R Patil. Moving target classification and tracking from real-time video. In *In Proc. IEEE Image Understanding Workshop*, pages 129–136, 1998.
20. F Liu and R W Picard. Finding periodicity in space and time. In *Proc. of the 6th Int. Conf. on Computer Vision*, pages 376–383, Bombay, India, 1998.
21. D M Moeslund and E Granum. A survey of computer vision-based human motion capture. *CVIU*, 81(3):231–268, March 2001.
22. N Paragios and R Deriche. Geodesic active regions for motion estimation and tracking. In *Proc. of the 7th Int. Conf. on Computer Vision*, pages 688–694, Kerkyra, Greece, 1999.
23. R Polana and R C Nelson. Detecting activities. *Journal of Visual Communication and Image Representation*, 5:172–180, 1994.
24. R. Polana and R.C. Nelson. Detection and recognition of periodic, nonrigid motion. *IJCV*, 23(3):261–282, June 1997.
25. S M Seitz and C R Dyer. View invariant analysis of cyclic motion. *Int. Journal of Computer Vision*, 25(3):231–251, December 1997.
26. J M Siskind and Q Morris. A maximum-likelihood approach to visual event classification. In *Proceedings of the Fourth European Conference on Computer Vision*, pages 347–360, Cambridge, UK, April 1996.
27. P S Tsai, M Shah, K Keiter, and T Kasparis. Cyclic motion detection for motion based recognition. *Pattern Recognition*, 27(12):1591–1603, December 1994.
28. Y Yacoob and M J Black. Parameterized modeling and recognition of activities. *CVIU*, 73(2):232–247, February 1999.