PERIODIC EVENT DETECTION AND RECOGNITION IN VIDEO

Vivek E P, Erik Pogalin, Arnold W M Smeulders

Intelligent Systems Lab Amsterdam, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, Netherlands
vivek@science.uva.nl, A.W.M.Smeulders@uva.nl

ABSTRACT

Periodicity attracts special attention in human cognition. Hence it is important to consider that in automatic analysis of motion events. This paper presents a method for representing periodic events with which events can be compared irrespective of their duration. The effectiveness of such a representation is verified with event classification.

Index Terms**—** motion analysis, action recognition, visual periodicity, periodic events, event representation

1. INTRODUCTION

Machine analysis of motion events is important to applications such as surveillance and video indexing. There exists several different approaches for video-based motion analysis [1, 2]. Motion events can be broadly classified as periodic and aperiodic events. The events which are repetition of some basic motion sequence over time can be categorized as periodic events. Human actions such as walking and running are the most common examples of such events. Detection and analysis of periodic events has been a topic of interest in motion analysis. In their work, Polana and Nelson [3] presented a Fourier analysis based method to analyze periodic events. Using the estimated frequency, the method represent periodic events with features extracted from spatio-temporal solids corresponding to single cycle of the event and classification is done based on such a representation. The disadvantage of this method is that it requires temporal scaling and phase correction for handling events with different frequency and phase. Cutler and Davis [4] proposed a correlation based approach for identifying periodic activities. The method works by segmenting the object from background and finding the correlation between the extracted object segments over time. Lattice representation of periodic events were used for motion based object classification. The limitation is that the object has to be segmented in each frame for detecting periodicity and the motion classification relies on temporal behavior alone. The work by Briassouli and Ahuja [5] analyzes sequences with multiple moving objects and extracts the periods of individual objects. The method works by projecting the frame-images obtained after background subtraction onto x and y axes. This method is restricted to periodicity analysis and was not applied to motion classification. Polana and Nelson [3] verified performance of their approach with a data set containing 6 actions performed by 2 subjects with some danger of over training due to limited number of subjects. Instead of using specific human body models, they used a general approach for learning the event models based on features extracted from a set of labeled videos. In our approach, a similar data driven method is used for modeling events and we considered a much larger data set (9 actions performed by 9 subjects) for our experiments.

The approach presented in this paper makes use of the concept of visual periodicity [6] for detecting periodic behavior in motion sequences. Unlike previous approaches this method represents an event by extracting the temporal and spatial nature of the motion independently. The event classification is done using the spatial features alone. The advantage of such a notion is that the events can be compared irrespective of their phase and duration . The rest of the paper is organized as follows: Section 2 presents some of the recent methods used for motion analysis. Temporal analysis of events plays a key role in the proposed method. Section 3 describes visual periodicity analyzer which is used in our approach. The proposed method for event classification is described in section 4 with experimental results in section 5.

2. ACTION RECOGNITION : RECENT DEVELOPMENTS

The topic of motion analysis and action recognition has evolved in many different directions. As a predominant class of motion events, some of the methods considered human actions as a specific motion class and addressed the problem of activity analysis with explicit models of human body parts [7, 8]. A more general approach is to model actions with pose primitives. Thurau and Hlavac [9] used non-negative matrix factorization (NMF) for determining pose primitives from a training set of different activities. Motion classification is done by representing events with pose histograms . Goldenberg et al. [10] presented a similar approach with pose primitives being determined using singular value decomposition(SVD). Weinland and Boyer [11] used a key-pose based embedding technique for action recognition. The work also discussed different methods for selecting key pose silhouettes. Souvenir and Babbs [12] proposed a method for view-invariant action recognition. The method learns low-dimensional manifold corresponding to human actions as a function of view point. The Radon transform of the key-pose silhouette is used for representing actions.

Visualization of actions as space-time volumes has lead to the development of a different class of methods which works with sptio-temporal features. Niebles et al [13] presented an unsupervised approach for learning human actions based on a probabilistic model for latent topic analysis of videos. The method used histograms of spatio-temporal features for representing action sequences. In their work Gorelick et al [14] proposed the use of Poisson equation in combination with space time shapes for action classification. The salience and orientation of solution to Poisson equation are used as local space time descriptors. The space time shapes are represented with weighted moments derived from local descriptors. Most of these methods tend to ignore repetitive behavior present in many of the motion events. The method presented in this paper makes use of this attribute of events for recognizing them.

3. VISUAL PERIODICITY ANALYSIS

The method proposed by Pogalin et al [6] for visual periodicity detection works by aligning the object windows extracted using a suitable tracking algorithm. The periodic behavior of the object in the scene is analyzed with PCA [15], which groups together the input data that are spatially correlated. PCA captures the periodic variations in intensity and shape of the object with unobserved variables. The approach is particularly interesting as the analysis is done by splitting the data into spatial and temporal components.

Let $Y = [y_{t1}, y_{t2} \dots y_{tN}]$ be a $D \times N$ matrix that represents the input video data (object windows) with N frames, each frame having D pixels. Each vector y_{tn} is formulated by the row wise concatenation of pixels from frame $V[x, tn]$. The data can be reconstructed optimally ($\hat{Y}$) as a weighted combination of Q-dimensional ($Q << D$) vectors of unobserved variable $U = [u_1, u_2 ... u_N]$ and a set of D-dimensional orthonormal basis vectors $W = [w_1, w_2 ... w_Q]$. This is given by:

$$\hat{Y} = WU + \bar{Y} \tag{1}$$

Here $\bar{Y}$ is the set of mean vectors. With PCA, the weight vectors w_q are given by the eigenvectors of the covariance matrix and the variation contained in each eigenvector is indicated by the corresponding eigenvalue. The value of Q is chosen by the percentage of variance need to be retained in the reconstructed data.

Periodicity analysis is done based on the fact that, while reconstructing the video data with PCA, the spatial behavior of motion is captured in eigenvectors w_q, where as unobserved variable u_q captures the temporal behavior. The frequency of the periodic event is estimated by combining the spectrum of each component of the unobserved variable.

Let $P_q(f)$ be the estimated spectrum corresponding to unobserved variable u_q and λ_q^* be fraction of total variance retained in u_q. Then the combined spectrum is given as

$$\bar{P}(f) = \sum_{q=1}^{Q} \lambda_q^* P_q(f) \tag{2}$$

The dominant frequency components in the combined spectrum are obtained by using the approach in [16]. Peaks in spectrum are detected using a dilation operation and frequencies lower than frequency resolution of the periodogram are discarded during further processing. Starting from the lowest frequency, each peak in the spectrum is checked against others for its harmonicity. A frequency is called harmonic if it can be expressed as the linear combination of the existing fundamental frequencies. A fundamental is required to have a higher peak than its harmonics. Since multiple fundamentals can exists in the spectrum, the fundamental together with its harmonics having highest total energy is used to represent the dominant frequency component in the data. Let $E(f)$ be the spectral energy at frequency f and f_k^0 be a fundamental with harmonics f_k^i. The dominant frequency f_{est}, of the spectrum is obtained as:

$$f_{est} = \arg\max_{f_k^0} \left\{ E(f_k^0) + \sum_i E(f_k^i) \right\} \tag{3}$$

In case of periodic motion events, the frequency thus estimated gives dominant frequency of the event. As a temporal attribute, frequency does not carry spatial behavior of object motion. The following section presents a method for analyzing motion events by extracting spatial behavior of the object motion.

4. PERIODIC EVENT RECOGNITION

Figure 1 shows the different stages in processing a given video for periodic motion classification. In the first step, the object under motion is tracked with a suitable tracker which can localize the object within a window. This prepossessing is useful in scenarios where periodic motion is overlapped with translation, which needs to be negated for periodicity analysis. For the experiments presented in this paper the Foreground-Background tracker [17] was used for tracking. From the object windows thus obtained, the temporal behavior (frequency spectrum) of the event is analyzed by means of visual periodicity analyzer discussed in Section 3. This frequency information is then used for representing the motion event by extracting only the spatial nature of motion. Such a representation could be advantageous in motion classification as the phase and duration of the motion sequences will not be present in it.

4.1. Event Representation

An object motion is characterized by its position in space over time. By discarding the temporal information present in an event, it is possible to obtain a spatial signature of the event. As videos are 2D representation of 3D space, the spatial signature of events obtained from videos can be considered as 2D images. For a pure periodic event (motion events without aperiodic components like unidirectional translation) the spatial signature will be the same irrespective of its duration or initial and final positions of the object. Thus the use of such a representation can be effective for comparing periodic events.

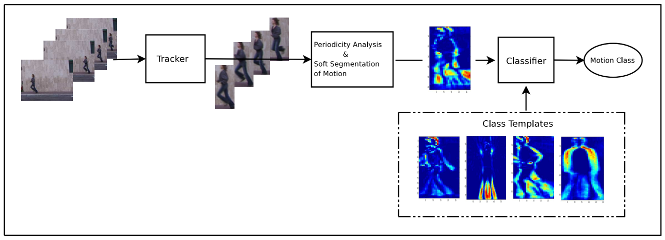

Fig. 1. Periodic motion analysis

Periodicity analysis discussed in section 3 determines fundamental frequency and its harmonics which are dominant in the event (eqn 3). By performing Fourier analysis of the object windows over time and selecting the coefficients corresponding to the dominant frequencies, it is possible to obtain the spatial representation corresponding to the event. Let $V[x,t]$ be the object window over time and let F be the set of dominant frequencies with f^0 being the fundamental frequency.

$$Z[x,\omega] = \frac{1}{N}\sum_{t=1}^{t=N} V[x,t]\exp\left(-j\omega t\right) \tag{4}$$

$$X[x] = \sum_{f\in F}\frac{\bar{P}(f)}{\bar{P}(f^0)}|Z[x,f]| \tag{5}$$

$X[x]$ can be seen as a weighted point set with weights indicating spatial behavior of the event. Figure 2 shows $X[x]$ obtained with four different periodic events.

4.2. Event Classification

The 2D representation of events can be compared with Earth Mover's Distance (EMD) [18] which is a measure of dissimilarity between weighted point sets. Let $A = \{(p_1,w_1),(p_2,w_2),\ldots,(p_m,w_m)\}$ and $B = \{(q_1,u_1),(q_2,u_2),\ldots,(p_n,w_n)\}$ be two point sets, where pi, q_i are points in R^d and w_i, u_i are the corresponding weights in R^+. Then EMD between A and B is defined as

$$E_d(A,B) = \frac{\min_{F\in\mathcal{F}}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}D(p_i,q_i)}{\min\{W_A,W_B\}} \tag{6}$$

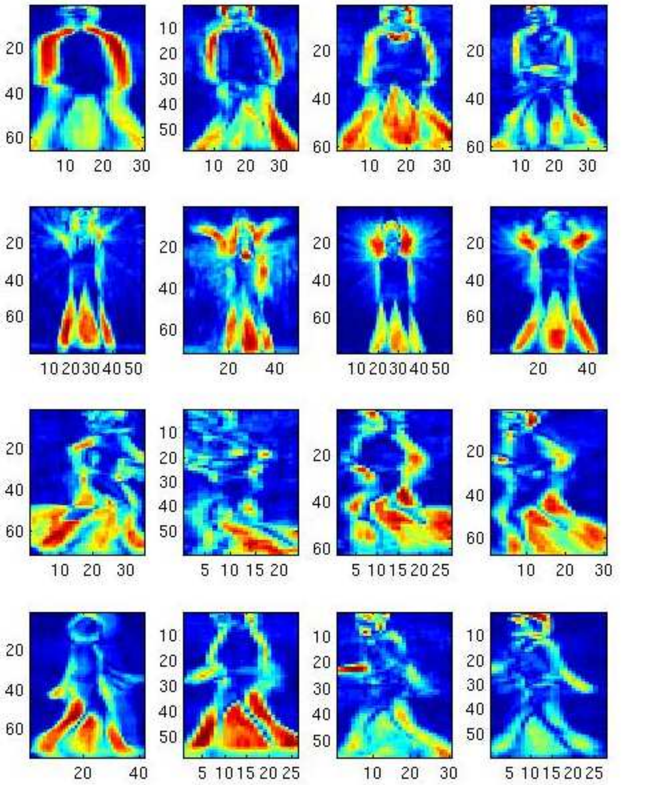

Fig. 2. Weighted point set representation of four different motion events(gallop sideways, jumping jack, run and walk) performed by four subjects

where $D : R^d \times R^d \rightarrow R$ is the distance measure, W_A and W_B are the total weights of A and B, $F = \{fij\}$ is a feasible flow satisfying following conditions:

a $f_{ij} \geq 0, i = 1,\ldots m, j = 1,\ldots,n$

b $\sum_{j=1}^{m} f_{ij} \leq w_i, i = 1,\ldots,m$

c $\sum_{i=1}^{n} f_{ij} \leq u_j, j = 1,\ldots,n$

d $\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} = min\{W_A,W_B\}$

5. EXPERIMENTS

The proposed method is evaluated with Weizmann human action data set [10] consisting of nine periodic events(discarding bending action in the data set). Performance evaluation is done using 1-NN rule with EMD [18] as the distance metric. The data set consisting of nine actions performed by nine different subjects is split into training (5 samples per event) and testing (4 samples per event) sets at random. Each instance of an event in the test set is classified by finding the most similar instance from the training data. As the direction of motion

	side	jack	run	walk	pjump	jump	skip	wave2	wave1
side	0.62	0.01	0.02	0.32	0.00	0.00	0.00	0.00	0.00
jack	0.00	0.91	0.04	0.02	0.02	0.00	0.00	0.00	0.00
run	0.08	0.00	0.62	0.10	0.00	0.00	0.18	0.00	0.00
walk	0.17	0.00	0.17	0.47	0.02	0.03	0.12	0.00	0.00
pjump	0.01	0.00	0.00	0.07	0.32	0.40	0.18	0.00	0.00
jump	0.00	0.00	0.00	0.00	0.35	0.60	0.02	0.00	0.00
skip	0.01	0.00	0.36	0.09	0.22	0.00	0.30	0.00	0.00
wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.03
wave1	0.00	0.02	0.00	0.00	0.05	0.00	0.00	0.00	0.91

Table 1. Confusion matrix with 1-NN classification

	proposed	Thurau et al [9]	Gorelick et al [14]
side	62%	98%	100%
jack	91%	100%	100%
run	62%	97%	98%
walk	47%	99%	100%
pjump	32%	63%	100 %
jump	60%	87%	89%
skip	30%	94%	97%
wave2	96%	98%	97 %
wave1	91 %	71%	94 %

Table 2. Recognition rates for different methods (using different estimation techniques)

affects the representation of an event (eg walking from left to right and right to left), the dissimilarity is taken as the minimum of EMD obtained in two different ways: 1) directly 2) by reflecting one of the samples with respect to a vertical axis through its center of mass. The experiment is repeated 100 times and the average confusion matrix is shown in Table 1. The confusion matrix shows that events with predominant leg motion (pjump, skip and walk) are getting misclassified more often compared to events with unique hand movement (hand waving and jumping jack). Table 2 shows comparison of the proposed method with some of the existing methods which use a different classification strategy. Though the recognition rate of the proposed method is found to be low in some cases, that was not our main purpose. We aim for a method capable of comparing events which are of different duration and we aim for a method which does not require heavy interaction segmentation of the object or temporal scaling of the event. The event recognizer requires only the event label and the initial object window to be tracked by the tracker. Under these circumstances the approach can be scaled to handle very large data sets where the methods in the references can not due the amount of interaction.

6. REFERENCES

[1] Jake K. Aggarwal and Q. Cai, "Human motion analysis: a review," *Computer Vision Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[2] Ronald Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[3] Ramprasad Polana and Randal C. Nelson, "Detection and recognition of periodic, nonrigid motion," *International Journal Computer Vision*, vol. 23, no. 3, pp. 261–282, 1997.

[4] Ross Cutler and Larry S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions Pattern Analysis Machine Intelligence*, vol. 22, no. 8, pp. 781–796, 2000.

[5] Alexia Briassouli and Narendra Ahuja, "Extraction and analysis of multiple periodic motions in video sequences," *IEEE Transactions Pattern Analysis Machine Intelligence*, vol. 29, no. 7, pp. 1244–1261, 2007.

[6] E. Pogalin, A.W.M. Smeulders, and A.H.C Thean, "Visual quasi-periodicity," in *CVPR*, 2008.

[7] Vasu Parameswaran and Rama Chellappa, "View invariance for human action recognition," *International Journal Computer Vision*, vol. 66, no. 1, pp. 83–101, 2006.

[8] Yaser Sheikh, Mumtaz Sheikh, and Mubarak Shah, "Exploring the space of a human action," in *ICCV*, 2005, pp. 144–149.

[9] C. Thurau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," in *CVPR*, 2008.

[10] Roman Goldenberg, Ron Kimmel, Ehud Rivlin, and Michael Rudzsky, "Behavior classification by eigendecomposition of periodic motions," *Pattern Recognition*, vol. 38, no. 7, pp. 1033–1043, 2005.

[11] Daniel Weinland and Edmond Boyer, "Action recognition using exemplar-based embedding," in *CVPR*, 2008.

[12] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *CVPR*, 2008.

[13] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[14] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *IEEE Transactions Pattern Analysis Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[15] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience Publication, 2000.

[16] Fang Liu and Rosalind W. Picard, "Finding periodicity in space and time," in *ICCV*, 1998, p. 376.

[17] Hieu T. Nguyen and Arnold W. Smeulders, "Robust tracking using foreground-background texture discrimination," *International Journal Computer Vision*, vol. 69, no. 3, pp. 277–293, 2006.

[18] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.