FULL ACTION INSTANCES FOR MOTION ANALYSIS

Stergios Poularakis[1], *Alexia Briassouli*[1,2*], *Ioannis Kompatsiaris*[2]

(1) University of Thessaly, Greece. (2) Informatics and Telematics Institute

ABSTRACT

Motion analysis is an important component of surveillance, video annotation and many other applications. Current work focuses on the tracking of moving entities, the representation of their actions and the classification of sequences. A wide range of methods are available for the characterization and analysis of human activity. This work presents an original approach for the detailed characterization of activity in a video sequence. A novel framework for encoding and extracting representative, repeating segments of activities is presented, resulting in "Full Action Instances". We focus on the analysis of human activities, however the proposed algorithm can be extended to more general categories of action that contains repetitive components, due to its general design.

1. INTRODUCTION

In recent years, significant research attention has been devoted to video motion analysis for the efficient representation, storage, transmission and access of multimodal data. The analysis of human activities has received considerable attention [1], with a vast range of methods focused specifically on this problem, using for example models of the human body, constraints of human kinematics [2], [3] [4], characteristic features of the human motions [5], their periodicity [6]. In this work, we present a novel approach for extracting repeating segments of a human activity without employing any domain-specific information. These segments are referred to as "Full Action Instances" (FAIs) and, although they are used in the analysis of human motion in this work, they can be extended any domain, since their extraction does not involve any information specific to human activity.

2. FULL ACTION INSTANCES (FAI), ENERGY OF TEMPORAL VIDEO DERIVATIVE EL_T

In this section, we present the background and main components of the proposed system. A "Full Action Instance" (FAI) is defined as the basic pattern that is repeated during the length of the action, denoted as P. This means that descriptions like $K = \{P, P, ..., U, ..., P\}$ are not allowed, whereas $K = \{U, P, ..., P, U\}$, $U \subseteq P$ are valid. The FAI is constructed based on the temporal derivatives of each frame i. In particular, we consider a video sequence of N frames with luminance $L(x, y, t)$. The temporal derivative of frame i are given by $L_{ti} = L_i(x, y, t + 1) - L_i(x, y, t)$ and its average energy (per frame) is:

$$E\{L_{ti}\} = \sum_{x,y} |L_{ti}|^2. \quad (1)$$

The sequence of these temporal-derivative energy signals $EL_t = [E\{L_{t1}\}, E\{L_{t2}\}, ..., E\{L_{tN}\}]$ is a one-dimensional vector that reflects the temporal variations of the activity's energy. Fig. 1 shows a frames of a girl jumping and the corresponding EL_t. The repetitions of the jumping are clearly present in the EL_t signal and the FAI can be derived from it.

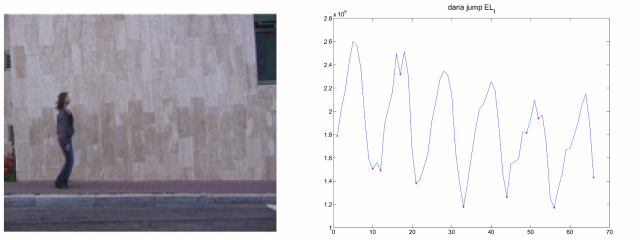


Fig. 1. Frames of a girl jumping and the corresponding EL_t.

From Fig. 1 it can be seen that the beginning and end of each repeating activity pattern corresponds to minima of the EL_t signals, so EL_t minima can be used to localize the FAIs. In order to verify this experimentally, EL_t's are extracted for ten motion categories, namely Walk, Jump, Run, Pjump (jumping in place), Bend, Jack (jumping jacks), Side (sideways walking), Skip, Wave (two different waving videos): in all cases, FAI endpoints coincide with EL_t minima and have a bell-shape, even for activities that are very different from each other.

In practice the EL_t signal also contains local minima that do not correspond to the endpoints of an FAI (Fig. 2). In those cases, the direct matching of all local minima to FAI endpoints would obviously lead to the erroneous localization of repeating activity patterns. This is where the main contribution of our work lies: we propose a simple yet effective approach to correctly find EL_t minima that correspond to actual

*Corresponding author.

FAI endpoints. The proposed method is based on the correct encoding and association of each bell's geometric characteristics, as described in detail in Sec. 3. The results of Sec. 4 show that, indeed, the proposed framework can localize FAIs with good accuracy.

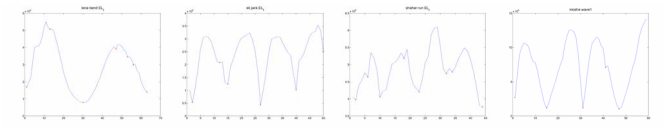


Fig. 2. EL_t for various activities; local minima do not always correspond to FAI endpoints.

3. EXTRACTION OF FAIS FROM EL_T

The goal of the proposed framework is to efficiently and reliably extract each motion's individual FAIs from the EL_t. The minima of the EL_t are not sufficient for finding the correct beginning and end of each FAI, since the EL_t curve may also contain several local minima that are not FAI endpoints. In this section we present an algorithm that overcomes this issue with very good results.

The main observation on which the proposed technique is based is that each sequence of two consecutive local minima in an EL_t may contain either a part of, or an entire FAI bell. The problem then becomes that of determining if a segment of EL_t is an entire FAI bell or not. This is achieved by encoding the local minima and the corresponding curve segments and assigning combinations of the resulting "codewords" to FAI bells or segments according to a set of rules that we design. The resulting algorithm, whose input is the EL_t and output the detected bells, has three main steps: (1) Extract local minima and maxima and separate into groups, (2) Find "group codes", (3) Combine groups to detect FAI bells.

Step 1: In the first step, all local minima and maxima in the EL_t are found and separated into groups of two successive local minima m_1, m_2 and the local maximum M between them, so that each group is represented as $O = (m_1, M, m_2)$. Each point in the group is represented by two coordinates, one corresponding to time t, and one to the magnitude of the EL_t at that time $y(t) = EL_t$.

Step 2: A codeword is then assigned to each group, according to a set of rules based on the fact that each group O either corresponds to a segment of a bell, or an entire bell. We have determined that there are four possible kinds of groups, encoded as in Table 1.

In order to assign codewords, one first finds the difference of the coordinates each group's points. The points' coordinates are $m_1 = (t_{m1}, y_{m1})$, $M = (t_M, y_M)$, $m_2 = (t_{m2}, y_{m2})$ and the corresponding differences $y_1 = y_M - y_{m1}$, $y_2 = y_M - y_{m2}$, $\Delta t = t_{m2} - t_{m1}$. The algorithm for computing each group's codeword is then given by Table 2. In these rules, the terms $>>, <<, \cong$ and "small enough" need to be further specified. Δt is compared to an application-dependent threshold T, so that "small enough" means $\Delta t < T$. This threshold is application dependent because it should be determined by the duration of each FAI. In this work, after extensive experimentation, $T = 6$ was found to be the optimal value, limiting FAI duration to be at least 6 frames long. In order to determine $>>, <<, \cong$, we consider the ratio:

$$\lambda = \frac{min(y_1, y_2)}{max(y_1, y_2)}. \quad (2)$$

Here, if $\lambda > Th$, then $y_1 \cong y_2$, and if $\lambda < Th$, then $max(y_1, y_2) >> min(y_1, y_2)$. We experimentally found that $Th = 0.6$ leads to accurate and consistent results.

Step 3: In the last step, the bell estimates are formed by combining the group codewords appropriately. In some cases, each pair of codewords can be mapped to one codeword, as in the examples of Figs. 3(a), (b), whereas in other cases no mapping is possible (Fig. 3 (c)).

All possible combinations with the result of the mappings are shown in Table 3. The first group is represented as (x, y), the second as (z, w) and the group resulting from their combination as (A, B). In cases where no combination is possible, we set $(A, B) = (-, -)$. Explanations about Table 3 follow:

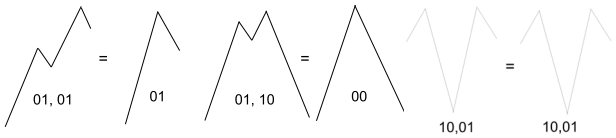
When $O = (O_{m1}, O_M, O_{m2})$, $\Omega = (\Omega_{m1}, \Omega_M, \Omega_{m2})$ are combined, they give a new group $V = (V_{m1}, V_M, V_{m2})$ with $V_{m1} = O_{m1}$, $V_M = (\frac{t_{O_{m1}} + t_{O_{m1}}}{2}, max(y_{O_{m1}}, y_{O_{m2}}))$, and $V_{m2} = \Omega_{m2}$. The new group V will be combined with its

Table 1. Possible Groups

Code	Form	Description
00		Full bell
01		Segment, open on the right
10		Segment, open on the left
11		Small bell segment

Table 2. Algorithm for codeword computation

1. If $y_1 >> y_2$, then code = 01
2. If $y_1 << y_2$, then code = 10
3. If $y_1 \cong y_2$, then:
   3a. If Δt small enough, then code = 11
   3b. else code = 00.



(a) (b) (c)

Fig. 3. (a), (b) Mappings of pairs of groups to a new group. (c) No new mapping possible.

neighboring group. C=1 denotes the detection of a new full bell. In this case, the first group (x, y) is stored as full bell and the second, (z, w), is combined with the group that follows. D=1 denotes an ambiguous case, where we should recompute the new team $V = (V_{m1}, V_M, V_{m2})$ according to new values V_{m1}, V_M, V_{m2}, extracted by changing the threshold Th.

The rules about these combinations can be represented using Boolean functions, where the input is (x, y, z, w) and the output is (A, B, C, D):

$$A = x \cdot z,\ B = y \cdot w,\ C = \bar{y} \cdot \bar{z},\ D = \bar{x} \cdot y \cdot \bar{w} + x \cdot z \cdot w \quad (3)$$

By applying these rules to the groups created from the local minima and maxima of the EL_t, we extract the FAIs for the activities observed. Thus, each instantiation of an FAI is fully characterized by the time instants at which it begins and ends. This information can be exploited in numerous manners, for example for the estimation of an activity's periodicity, or for the detailed characterization of the repeating motion in an activity (from the FAI bell).

4. EXPERIMENTS

Experiments were conducted with videos of nine people performing the ten different activities mentioned in Sec. 2[1].

4.1. Minima of EL_t are endpoints of FAIs

The assumption that the endpoints of EL_t are endpoints of FAIs makes intuitive sense, but is also validated experimentally. Table 4 shows the percentages of the actual FAI endpoints that match local minima. We consider that there is a match when the endpoints deviate by $0 - 6$ frames. The restriction of this deviation to 0 frames leads to a low matching percentage. However this restriction is too strict and actually non-realistic, since the EL_t is estimated from derivatives, so its minima correspond to pairs of frames. The results of Table 4 for deviations above 0 are very precise, indicating that EL_t minima can be reliably considered as FAI endpoints. In practice we choose the maximum acceptable error for the localization of FAI endpoints to be equal to three frames.

[1]The videos were taken from http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html

Table 3. Possible Groups

(x,y)	(z,w)	(A, B)	C	D	Figures
0 , 0	0 , 0	- , -	1	0	
0 , 0	0 , 1	- , -	0	1	
0 , 0	1 , 0	0 , 0	0	1	
0 , 0	1 , 1	0 , 0	0	0	
0 , 1	0 , 0	0 , 0	0	1	
0 , 1	0 , 1	0 , 1	0	0	
0 , 1	1 , 0	0 , 0	0	1	
0 , 1	1 , 1	0 , 1	0	1	
1 , 0	0 , 0	- ,-	1	0	
1 , 0	0 , 1	- ,-	1	0	
1 , 0	1 , 0	1 ,0	0	0	
1 , 0	1 , 1	1 ,0	0	1	
1 , 1	0 , 0	0 ,0	0	0	
1 , 1	0 , 1	0 ,1	0	0	
1 , 1	1 , 0	1 ,0	0	0	
1 , 1	1 , 1	1 ,1	0	1	

4.2. Density of Local Minima

As mentioned in Sec. 2 and Sec. 3, an initial, naive approach to the detection of FAI endpoints would be to consider them equal to the local minima of EL_t. In practice this approach would introduce many errors, as it would be sensitive to local noise. In order to verify this experimentally, we estimate the density of local minima in the EL_t signals for all ten videos, in the presence of increasing noise. We define the density of local minima as:

$$\delta_{min} = \frac{\text{number of local minima}}{\text{number of FAI endpoints}}. \quad (4)$$

This quantity is essentially a measure of the expected difficulty in correctly detecting the FAI endpoints, since a more dense presence of local minima increases the probability of error. Table 5 shows that, indeed, for higher amounts of noise, δ_{min} increases, rendering a more elaborate method for localizing the endpoints, like that of Sec. 3, necessary. For reasons

Table 4. Matching accuracy of EL_t minima with FAI endpoints for maximum error from 0 to 6 frames.

Error	0	1	2	3	4	5	6
Match %	44.4	90.3	96.9	97.8	98.8	98.9	99.6

Table 5. Average Density of Local Minima and Average Performance (%) for noiseless and noisy data.

σ_{Noise}	Avg. δ_{min}	Avg. Perf. (%)
0	1.69	89.33
1	1.72	89.08
3	1.8	88.83
5	2.02	88.59
7	2.22	86.6
9	2.44	88.83
11	2.64	83.37
13	2.75	76.43
15	2.97	75.19
17	3.11	73.45
19	3.11	65.76

of space, we present the value of δ_{min} for increasing noise averaged over the ten videos, and for every second value of the noise variance. Detailed results for all videos are provided in the webpage http://inf-server.inf.uth.gr/ briassou/FAI.html.

4.3. Detection of FAIs

In this section, detection results for the FAIs using the encoding method of Sec. 3 are presented. The EL_t for each activity is initially estimated (Sec. 2), and ground truth for its local minima is found manually. We measure performance as the percentage of FAI bell endpoints that are correctly detected. In order to examine the robustness of the proposed method, we add noise of increasing variance to the data and measure the FAI detection performance in that case. The results for the noiseless data and for data in the presence of additive Gaussian noise with increasing variance from 1 to 20 are shown in the last column of Table 5. For reasons of space, we have tabulated results for every second σ_{Noise} increase (i.e. $\sigma_{Noise} = 1, 3, 5...$), averaged over the ten videos. Detailed analytical results for all videos are provided in the webpage http://inf-server.inf.uth.gr/ briassou/FAI.html. As expected, the FAI detection performance decreases with the increase of the additive noise. This makes intuitive sense considering the corresponding increase in the density of the local minima. Nevertheless, the performance degradation is not very high, and becomes more noticeable for particularly high values of noise. Consequently, the proposed approach is robust and can be used reliably in realistic practical applications.

5. CONCLUSIONS

In this work, an original method for the characterization of repeating human activities is presented. The repeating parts of human activities,"Full Action Instances" (FAIs) are extracted from the changepoints in the energy of the data's temporal derivative. An algorithm for the accurate extraction of FAIs is proposed based on the geometric characteristics of the FAI segments. Experiments demonstrate that the system leads to reliable and robust detection results even in the presence of noise. The proposed algorithm does not make any assumptions about the kind of activity taking place, and is thusly not limited to the analysis of human motions. Future work includes the extension of its application to more kinds of repeating motions, not necessarily strictly periodic or human-generated.

Acknowledgments:

The research leading to these results has received funding from the European Communitys Seventh Framework Programme FP7/2007-2013 under grant agreement FP7-214306 - JUMAS.

6. REFERENCES

[1] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding: CVIU*, vol. 73, no. 1, pp. 82–98, 1999.

[2] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, pp. 90–102, 1999.

[3] S. Ju, M. Black, and Y. Yaccob, "Cardboard people: a parameterized model of articulated image motion," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 1996, p. 3844.

[4] T. Zhao, T. S. Wang, and H. Y. Shum, "Learning a highly structured motion model for 3d human tracking," in *Proc. Asian Conf. Computer Vision*, 1996, p. 3844.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[6] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, "Behavior classification by eigendecomposition of periodic motions," *Pattern Recognition*, vol. 38, pp. 1033.