Proc. of Int. Conf. on Pattern Recognition (ICPR), Quebec City, August 2002.

Automatic Recognition of Unpredictable Events in Videos

Longin Jan Latecki
Computer and Information Science Department
Temple University
Philadelphia, PA 19122, USA
latecki@joda.cis.temple.edu

Daniel de Wildt
Department of Applied Mathematics
University of Hamburg
Bundesstr. 55, 20146 Hamburg, Germany
uni@dewildt.de

Abstract

The presented approach allows us to recognize frames in video sequence that are significantly different from the previous and/or following frames. In this way we are able to detect unpredictable events in videos. We map a video sequence to a polygonal trajectory by mapping each frame to a feature vector and joining the vectors representing consecutive frames by line segments. Shape analysis of the obtained polygonal curve allows us to detect frames representing unpredictable events. We demonstrate the performance of our approach on surveillance videos.

1 Motivation

First we describe what we mean under "predictability" of frames in an image stream. If frames are predictable, they are not as important as the ones that are unpredictable. We will rank these frames lower, since they can be inferred them from the previous frames. For example, imagine the following situation. A person enters a camera view field, walks from left to right in front of the camera, disappears, and nothing else happens. In this case, the two groups of frames showing the person entering and exiting the camera view field are unpredictable. The frames showing the person walking form a predictable group of frames, since we see a constant motion of a "large" object from left to right and the motion of body parts can be neglected. Clearly, the two groups of frames showing the constant image before the person entered the camera view field and after he/she exited it are predictable. The concept of predictability was motivated in [2] by techniques that cameramen use.

We need a mathematical framework in which the concept of predictability is represented by some mathematical structure. We will map each frame in an image stream to a feature point in a high dimensional space in such a way that frames belonging to the same group of predictable frames are collinear points. One of the feature dimensions is the frame number (or time stamp), in order to be able to recognize object motions. This already guarantees us that the frames showing a constant image form a collinear group of points in the feature space. In this paper we define further features such that a group of frames maps to a set of collinear points in case of predictable changes, and this is not the case for unpredictable changes.

Further we would like that the level of unpredictability is also reflected by some mathematical concept. Intuitively the more unpredictable the change in the video sequence, the more curved the corresponding trajectory in the feature space should be. We will introduce a measure that allows to measure the curveness of the parts of trajectory and their linearity in a robust manner in the presence of noise.

Noise for a image stream is distinct from pixel noise. The image stream generated by a fixed camera looking at people walking on the street may be considered to have a stationary component and a visual noise component, due to body parts movement and due to changing colors of people's clothes. The passing of a single car would be definitely the un-noisy information contained in the image stream. Since we expect the video signal to be noisy in this sense, we need to incorporate in our measure a filtering step that will enhance the linearity of linear parts as well as the curveness the parts with a significant curvature.

The curveness measure that we will introduce allows to rank video frames by their relevance that depends on the context, i.e., a given frame can be of high relevance in one context but of low relevance in the other, since we would like that the rank of the frames reflects their relevance to the content of the video clip. The context is represented by part of the corresponding trajectory to which we apply our measure, and the relevance is

represented by the value of our curveness measure.

2 Mapping an Image Stream to a Trajectory in the Feature Space

In order to obtain a real-time system, we decided to use a set of simple features extracted from images. We wanted that an object motion that can be approximatively regarded as translational motion results in nearly collinear set of points in the feature space.

We assign the set of 37 features to each image in a video sequence in the following way:

In the YUV color space that is used in MPEG encoding, for each of the 3 components, we define 4 histogram buckets by dividing the components in 4 intervals. Each bucket contributes 3 feature vector components: the pixel count, and the x and y coordinates of the centroid of the pixels in the bucket. That is 36 components, and we add the time (frame index) to get 37 components. This mapping produces a trajectory that is a polygonal curve in $\mathbb{R}^{37}$. The distance in this feature space is simply the Euclidean distance.

As the camera translates or pans smoothly without seeing new things, the centroid components change linearly and the trajectory of feature points is linear. If the camera suddenly decelerates, the trajectory has a large curveness, because the centroids decelerate.

This feature space was introduced in [2]. An alternative mapping is also presented in [2], where a statistical model for each frame is generated using a hidden Markov model (HMM) technique. Then a distance measure between two frames is based on the probability that each frame could have been generated by the model of the other.

In order to compensate for rapid changes in feature values that appear due to image and video noise, we applied morphological opening followed by closing to each coordinate in the feature space. In our experimets, the filter base of width 7 turned out to be optimal. This means for a given feature coordinate that we consider it values for a given frame together with the values for 3 preceding and 3 following frames. We will denote such filter *Morph3*, and similar for different width of the filters.

3 Curveness Computation

In Section 2 we described a mapping of a video sequence to a trajectory that is a polyline. Since the polyline may be noisy, in the sense that it is not linear but only nearly linear for the video stream segments where nothing of interest happens, i.e., the segments are predictable, and the parts of high curvature are difficult to detect locally, it is necessary to filter it.

In section 4 we will describe a filtering operation which we call discrete curve evolution. The goal is to simplify the polyline so that its sections become linear when the corresponding video stream segments are predictable. We achieve this by iterated removal of the vertices that represent the most predictable video frames. In the geometric language for the polyline trajectory, these vertices are the most linear ones. Consequently, the remaining vertices of the simplified polyline are frames that have a higher non-predictability value than the deleted ones.

We apply the process of the discrete curve evolution to parts of polygonal trajectory. We divide polyline P into overlapping parts T_i, each having 25 consequtive points (representing 25 frames). Since we work with movies with 25 frames/s, each part T_i represents one second of the video. The parts T_i are shifted every 10 frames. Thus, each intersection $T_i \cap T_{i+1}$ contains 15 points. We apply discrete curve evolution to each part T_i until five points remain, including the beginning and the endpoint that are fixed. Based on these five points, we assign a curveness measure to T_i. We denote the five points a, b, c, d, e with a and e being the endpoints. The *curveness measure* is given by the following formula:

$$C(T_i, P) = |d(a,b) + d(b,c) + d(c,d) + d(d,e) - d(a,e)| \quad (1)$$

For a planar polygon $abcde$, C can be viewed as the difference between the side lengths and the length of the basis ae. C maps each of the overlapping parts T_i to a non-negative real number. The domaim of fuction C is the sequence $T_1, ..., T_n$ of all parts of a polyline P. The key feature of this mapping is that the higher value of $C(T_i, P)$ is the more unpredictable is the sequence of frames in T_i. Hence, the curveness measure allows us to rank parts of videos by their relevance.

The extraction of significantly unpredictable parts in videos is a simple process now. In order to identify significantly unpredictable parts of surveillance videos, we use maxima of the function C that are above a certain threshold value (which we determined experimentally). Our experimental results are presented in Section 5.

4 Discrete Curve Evolution

Our approach to simplification of video polylines is based on a novel process of discrete curve evolution applied in the context of shape similarity of planar objects in [3]. However, here we will use a different relevance measure of vertices.

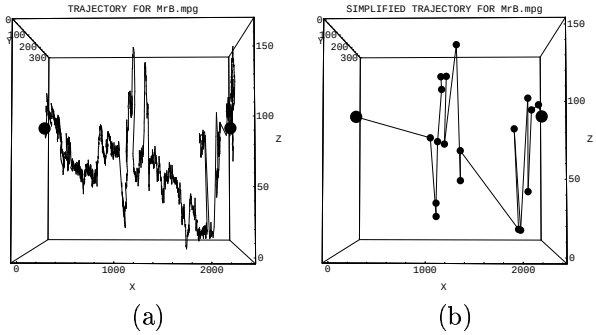

Figure 1. (a) Video trajectory with 2379 vertices for a Mr. Bean's video clip. (b) A simplified polygon with 20 most relevant frames (black dots).

Let P be a polyline (that does not need to be simple). We will denote the vertices of P by $Vertices(P)$. A *discrete curve evolution* produces a sequence of polylines $P = P^0, ..., P^m$ such that $|Vertices(P^m)| \leq 3$, where $|\ .\ |$ is the cardinality function. Each vertex v in P^i (except the first and the last) is assigned a relevance measure $K(v, P^i) \in \mathbb{R}_{\geq 0}$. The relevance measure $K(v, P^i)$ that we used for our experiments is defined below. The process of *discrete curve evolution* is very simple:

- At every evolution step $i = 0, ..., m-1$, a polygon P^{i+1} is obtained after the vertices whose relevance measure is minimal have been deleted from P^i.

Our relevance measure $K(v, P^i)$ that determines the order of vertex deletion depends on vertex v and its two neighbor vertices u, w in P^i. It is given by the formula

$$K(v, P^i) = K(u, v, w) = |d(u, v) + d(v, w) - d(u, w)| \quad (2)$$

where d is the Euclidean distance function in R^{37}.

Observe that the relevance measure is not a local property with respect to the polygon P, although its computation is local in P^i for every vertex v. This implies that the relevance of a given video frame v is context dependent, where the context is given by the adaptive neighborhood of v, since the neighborhood of v in P^i can be different than its neighborhood in P. Observe also that our relevance measure implies that the length change between P^i and P^{i+1} is minimal if P^{i+1} is obtained from P^i by deleting a single vertex.

Fig. 1 from [2] illustrates the curve simplification produced by the discrete curve evolution for a polygonal curve obtained from 80s of one of Mr. Bean's video clips. Although we only see a 3D projection, we can observe that the most relevant vertices of the curve and the general shape are preserved.

5 Experimental Results

We performed a large number of experimental results to verify the proposed technique using many different kinds of surveillance video clips. Due to the limited space, we illustrate our results on two video clips. These video clips as well as the results can be viewed on our home page [5] under Information. Both clips present a 15 seconds of an indoor video. In the first one (seciurity1.mpg) a stationary camera shows an empty office room, then one of the authors enters his office, and after a short time, he exits the office. Here we have two unpredictable events, when the author enters and when he exits the room.

The second video clip (Mov3.mpg) is taken by a hand held camera. It shows a student sitting behind his desk. Beside a small body part movements, there are three significant events, he raises his right hand twice and he raises his left hand once.

In both video clips, our system with the same threshold value recognizes the significant events as can be seen in Figures 2 and 3. As we said in Section 2, the best results are optained for morphological filger Morph3.

6 Conclusions

In this work, we have proposed and implemented a system for automatically detecting unpredictable events in videos. Our approach is based on key frame extraction presented in [2]. A different set of image features was used for key frame extraction in [4].

We divide a video sequence into overlapping parts, for each part we extract three most relevant key frames. We use these key frames to assign a curveness measure to each part. The higher is the curveness value of a given part, the more unpredictable events it present. This feature is the basis for extracting unpredictable events in surveillance videos, which we demonstrated in our experimental results.

References

[1] DeMenthon, D.F., Kobla, V., M., and Doermann, D., "Video Summarization by Curve Simplification", ACM Multimedia 98, Bristol, England, pp. 211-218, September 1998.

[2] DeMenthon, D.F., Latecki, L.J., Rosenfeld, A., and Vuilleumier Stückelberg, M., "Relevance Ranking

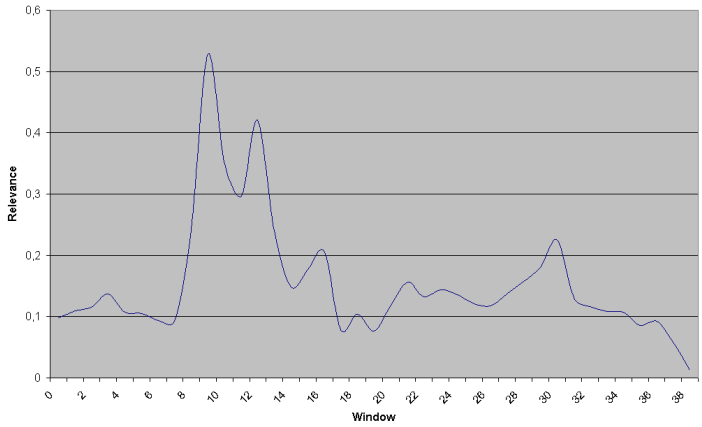

Figure 2. The graph of the curveness function for the first video clip (seciurity1.mpg). Relevance denotes here the values of the curveness function. One can see two clear maxima.

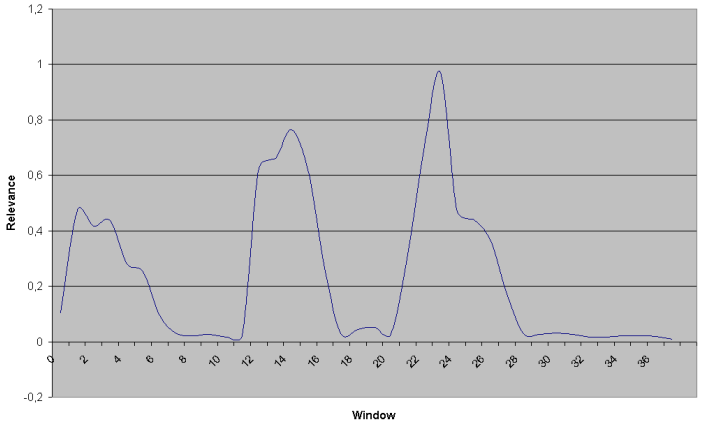

Figure 3. The graph of the curveness function for the second video clip (Mov3.mpg). Relevance denotes here the values of the curveness function. One can see three clear maximas.

and Smart Fast-Forward of Video Data by Polygon Simplification", Int. Conf. on Visual Information Systems, pp. 49-61, November 2000.

[3] L. J. Latecki and R. Lakämper. Shape Similarity Measure Based on Correspondence of Visual Parts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(10), pp. 1185-1190, 2000.

[4] L. J. Latecki, D. de Wildt, and J. Hu. Extraction of Key Frames from Videos by Optimal Color Composition Matching and Polygon Simplification. *Proc. of the Workshop on Multimedia Signal Processing*, Cannes, France, October 2001.

[5] Latecki, L.J. and de Wildt, D., www.videokeyframes.de

[6] Smith, M.A., and Kanade, T., "Video Skimming for Quick Browsing Based on Audio and Image Characterization", Proc. of CVPR, 1997.

[7] Yeung, M.M, and Yeo, B.L., "Time-Constrained Clustering for Segmentation of Video into Story Units", Proc. of ICPR, 1996.