Automatic spatial and temporal organization of long range video sequences from low level motion features

Alberto Quintero Delgado, Yannick Benezeth, Désiré Sidibé
Université de Bourgogne
ajquinterod@gmail.com, {yannick.benezeth, dro-desire.sidibe}@u-bourgogne.fr

Abstract

In this paper, we address the analysis of activities from long range video sequences. We present a method to automatically extract spatial and temporal structure from a video sequence from low level motion features. The scene layout is first extracted, with a set of regions that have homogeneous activities called Motion Patterns. These regions are then analyzed and the recurrent temporal motifs are extracted for each Motion Patterns. Preliminary results show that our method can accurately extract important temporal motifs from video surveillance sequences.

1. Introduction

Most of public areas are now monitored with surveillance cameras. The manual analysis of all this information is clearly impossible. Nowadays, most of these CCTV are never screened and are recorded to be later used as evidence. However, even the search for specific events in a large collection of video is also a very time consuming and tedious task that can be impossible if the amount of video is too large. It is conceivable to browse several hours or days of videos if there exists some metadata about the content of the video.

We present in this paper a method to extract, without any supervision, areas of homogeneous activities called *Motion Patterns* (MP). Then, the activity for each MP is modeled infering the cycles, periodicity and the ordered atomic activities that compose the recurrent motif. This information can be very useful to easilly browse very long video sequences, to determine the statistics of events, to predict future activities and thus possibly detect unusual ones.

Many video content anaysis methods use conventional tracking-based methods to recover individual moving object trajectories [7]. Specific information about moving objects such as position, velocity or acceleration is then used for high-level tasks. However, it is broadly accepted that object tracking is ill-suited for videos with a large number of moving objects such as in crowded scenes. Other video representation have been proposed in the litterature based on background subtraction [1], optical flow [4] or spatio-temporal features [6].

These low-level features are latter used with supervised or unsupervised learning methods to recognize a specific activity or identify an usual behavior. Supervised approaches can perform quite well when accurate labels are provided however it can be difficult to explicitely define positive and negative classes in complex scenes when several activities occur simultaneously. To take into account the sequential nature of activities and the possible causal dependencies between them, it is necessary to use more complex models such as Hidden Markov Model [5], Bayesian network [3] or Probabilistic Topic Models [8].

In this paper, we propose a much simpler method that nevertheless obtain global scene states useful for a later information mining. Causal dependencies between atomic activities are not considered because the scene is reduced to a set of regions with homogenous activities. We propose to extract the cycles, the periodicity and the set of atomic activities that composes the recurrent activity of each MP.

2. Methodology

Our proposed approach is based on three steps: 1) computing low level motion features, 2) extracting the spatial organization of the scene (Motion Patterns), and finally, 3) using the obtained MPs to extract the recurrent activity patterns or motifs in each of them.

2.1. Motion features

The first step is to compute the low level motion features for the whole video sequence. We compute dense optical flow using the method developed in [2]. The flow matrix is then divided into a grid and just one value is kept in each of its cells:

$$C_i(x_c, y_c) = \begin{cases} \theta, & \text{if } \bar{\rho} \geq \rho_t \\ -1, & \text{otherwise} \end{cases}, \quad (1)$$

where C_i is a cell in the grid $\mathcal{G}$, x_c and y_c are the coordinates of the center of C_i, θ is the dominant angle in that cell, $\bar{\rho}$ is the average magnitude of all the flow vectors in C_i and ρ_t is a threshold.

2.2. Spatial organization

After extraction of low level motion features, the temporal evolution of the motion feature in each cell C_i is represented as a histogram Φ_i. We then compute a similarity matrix $S \in \mathcal{R}^{m \times m}$, where m is the number of cells, between the activities in different cells. $S_{ij} = d_{ij}$, where d_{ij} is the dissimilarity between the histograms Φ_i and Φ_j.

In order to find spatial organization of activities in the video, a hierarchical clustering is applied on the normalized Laplacian matrix L computed from the similiraty matrix S. The number of clusters is automatically obtained from the eigendecomposition of L. We set the number of clusters so that 95% of the total variance is retained. The obtained clusters define the Motion Patterns, and all the cells belonging to the same class are said to be part of the MP represented by that class. The result is a mask image for each MP found in the scene as shown in Fig.1 (top right image).

2.3. Temporal organization

In this section the long video sequence is divided into small video clips, and each cell is represented by $C_{i,x_c,y_c}(t)$, with $t = 1, \ldots, N$ and N the number of frames in the clip. The signal $C_{i,x_c,y_c}(t)$ is a description of the motion activity in cell C_i over time. We then compute a similarity matrix based on the similarity between the signals representing the cells in each short video clips. The same clustering method as in Section 2.2 is applied to the similarity matrix to produce a label for every clip. Each label is represented as a string, *i.e.* we convert integers to characters, and we apply a substring matching technique to find the most recurrent substring (MRS) *i.e.* the motif in the video sequence.

3. Experiments

Fig. 1 shows an example of obtained results. The top right image in the figure shows the recovered MPs in the video sequence. The recovered MPs are shown with different colors while areas with no activity are shown in black. The plot in the bottom shows the assigned class to each of the small clip from the long video sequence for the MP 3. In this case, just the average motif recovered was matched and it can be seen highlighted in different colors for each occurence. Three activities, labeled A, B and C, occur periodically in this MP. Similar results are obtained for other MPs.

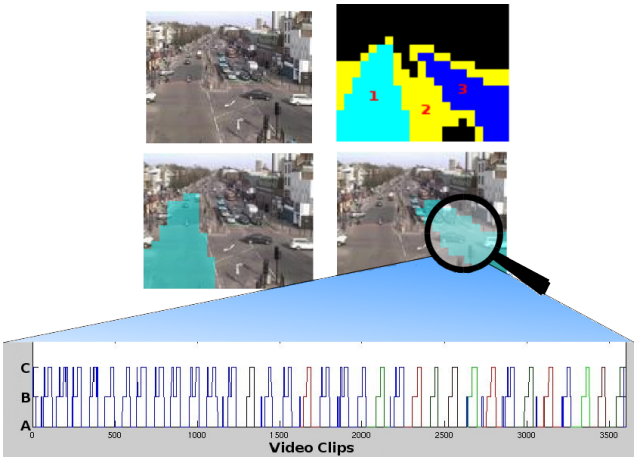

Figure 1. Example of results; Top left image is the original frame, top right image shows the recovered MP, and bottom image shows the recovered MRS in the MP#3.

4. Conclusion

This paper presents a method for recovering the spatial and temporal organization of long video sequences. Areas of homogeneous activities, Motion Patterns, are first extracted and the activity for each MP is modeled to infer the characteristics of the recurrent activity in the MP.

The approach shows promizing results on videos with a dense activity flow. Our future work is to improve the method, in particular the recurrent motif extraction step, for video scenes with sparse flow of activities.

References

[1] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans on PAMI*, 23(3):257–267, 2001. 1

[2] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. pages 25–36. Springer, 2004. 1

[3] S. Calderara, R. Cucchiara, and A. Prati. A distributed outdoor video surveillance system for detection of abnormal people trajectories. *in ICDSC*, pages 364–371, 2007. 1

[4] P.-M. Jodoin, Y. Benezeth, and Y. Wang. Meta-tracking for video scene understanding. *IEEE int. conf. on AVSS*, 2013. 1

[5] D. Kuettel, M. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. *IEEE Conf. on CVPR*, pages 1951–1958, 2010. 1

[6] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. *int. conf. on Multimedia*, pages 357–360, 2007. 1

[7] C. Stauffer and W.E.L.Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on PAMI*, 22(8):747–757, 2000. 1

[8] J. Varadarajan and J. Odobez. Topic models for scene analysis and abnormality detection. *in ICCV Workshop*, pages 1338–1345, 2009. 1