Event Detection and Classification in Video Surveillance Sequences

Vasileios Chasanis and Aristidis Likas

Department of Computer Science, University of Ioannina,
45110 Ioannina, Greece.
{vchasani,arly}@cs.uoi.gr

Abstract. In this paper, we present a system for event recognition and classification in video surveillance sequences. First, local invariant descriptors of video frames are employed to remove background information and segment the video into events. Next, visual word histograms are computed for each video event and used to define a distance measure between events. Finally, machine learning techniques are employed to classify events into predefined categories. Numerical experiments indicate that the proposed approach provides high event detection and classification rates.

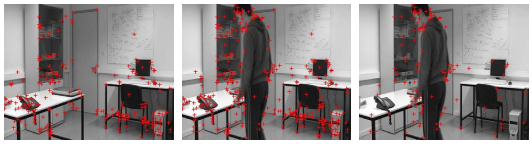
Key words: Video surveillance, Event detection, Dynamic time warping

1 Introduction

Video surveillance has received many attention over the last years and is a major research topic in computer vision [4]. Typically, the framework of a video surveillance system involves the following stages: background substraction, environment modeling, object detection, classification and tracking of moving objects and descriptions of behaviors/events. The goal of video surveillance systems is to detect and characterize events as activities using unsupervised or supervised techniques.

In [2], a method is presented that integrates audio and visual information for scene analysis in a typical surveillance scenario, using only one camera and one monaural microphone. In [8], a video behavior modeling method is proposed for online normal behavior recognition and anomaly detection. For each video segment, blobs are detected that correspond to scene events. These scene events are clustered into groups using a gaussian mixture model producing a behavior representation for the video segment.

In our approach, local invariant descriptors are employed to remove background information. Then, by analyzing the number of foreground descriptors, we automatically segment the video surveillance sequence into segments/events, which describe some activity taking place in the room under surveillance. Each video segment/event is represented either by a single (summary) visual word



(a) (b) (c)

Fig. 1. Video frame a) of the background and the location of the extracted descriptors b) of an event with its descriptors c) of an event with unmatched descriptors.

histogram or by multidimensional signal corresponding to the visual word histograms of its own frames. In the second case, Dynamic Time Warping distance [7] is employed to define a proper event dissimilarity metric. Finally, supervised and unsupervised techniques are implemented either to classify or to cluster events into categories.

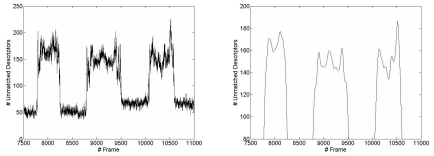
The rest of the paper is organized as follows: In Section 2, the procedure of background substraction is described. In Section 3, the proposed event detection algorithm is presented. In Section 4, we define an event dissimilarity metric and in Section 5 we present numerical experiments for video event classification and clustering into categories. Finally, in Section 6, we provide some conclusions.

2 Background Substraction

For each frame of the video surveillance sequence, SIFT descriptors are extracted as proposed in [6]. In this work, we concentrate on different individual activities performed in an indoor environment, captured by using a standing camera. Thus, background remains the same and object/event detection relies on foreground detection modules. In order to remove descriptors that correspond to background objects, we compare the descriptors of each frame of the video surveillance sequence with a set of pre-computed descriptors corresponding to frames describing only the background using the comparison approach proposed in [6]. In Fig. 3(a), we present a video frame of the background and the location of the extracted descriptors. In Fig. 3(b) and Fig. 3(b), we present a video event frame with the corresponding SIFT descriptors and the descriptors that do not match with those of the background, respectively.

3 Video Segmentation into Events

After we have subtracted the descriptors corresponding to background, we wish to identify unique events in the video sequence. In our surveillance problem a video event is defined as the time interval where a person performs an activity. Thus, it is expected that when someone enters the room under surveillance, new descriptors will appear that do not correspond to background. In our method



(a) (b)

Fig. 2. a) Normal and b) smoothed signal of the number of unmatched descriptors of a video surveillance sequence.

we analyze a vector that corresponds to the number of "unmatched" descriptors between each frame and the background. In Fig. 2(a), we present the sequence of "unmatched" descriptors of a video surveillance sequence.

In order to detect the beginning and the end of a video event, this vector H is smoothed by:

$$L_t = \sum_{n=-\infty}^{\infty} H_n \cdot K_\sigma(t-n), \tag{1}$$

where K_σ is a normalized discretized gaussian kernel with zero mean and standard deviation σ. Furthermore, we discard low values of the smoothed signal to remove noise (background descriptors that have not been removed). In Fig. 2(b) we present the final smoothed signal for the sequence of Fig. 2(a).

3.1 Event Representation

After we have segmented the video into N events, we represent each video frame of the event or the whole event with a visual word histogram. More specifically, for each video event $E_i,\ i = 1, \ldots, N$ a different number of descriptors is computed that describe certain objects or interest points in the event. Suppose we are given a video event E_i and its corresponding set of n video frames $F = \{f_1, \ldots, f_n\}$. For each video frame $f_j,\ j = 1, \ldots, n$, a set of SIFT descriptors D_{f_j} is extracted using the algorithm presented in [6]. Then, all the sets of descriptors are concatenated to describe the whole event

$$D_{E_i} = D_{f_1} \bigcup \ldots \bigcup D_{f_n}. \tag{2}$$

To extract *visual words* from the descriptors, the sets of descriptors for all N video events $D_V = D_{E_1} \bigcup D_{E_2} \bigcup \ldots \bigcup D_{E_N}$ is clustered into K groups $\{C_1, C_2, \ldots, C_K\}$ using the k-means algorithm, where K denotes the total visual words vocabulary size. To construct the visual word histogram (bag of visual words) for video frame f_t, each element of the set of descriptors D_{f_t} is assigned to one of the K visual words (clusters), thus resulting into a vector containing the frequency of each visual word in the video frame. Thus, given that frame f_t

has D descriptors $d_{t_1}, \ldots, d_{t_D}$, the visual word histogram VHF_t for this video frame is defined as:

$$VHF_t(l) = \frac{\#\{d_{t_j} \in C_l,\ j = 1, \ldots, D\}}{|D|},\ l = 1, \ldots, K. \tag{3}$$

Similarly, a visual word histogram VHE_i of an event i is constructed by assigning each descriptor of set D_{E_i} to one of the K visual words (clusters).

4 Event Dissimilarity

In order to proceed with video event classification an event dissimilarity metric must be defined. In our approach we consider two approaches. In the first one, to compute a distance value between two events E_i and E_l we compare their corresponding visual word histograms VHE_i and VHE_l. In the second approach, we compare the visual word histograms VHF of their frames. More specifically, suppose that we are given events $E_i = \{f_1^i, \ldots, f_{n_i}^i\}$ and $E_l = \{f_1^l, \ldots, f_{n_l}^l\}$. Since $n_i \neq n_l$, we have to define a proper dissimilarity metric to compare these two events. In our approach, we use Dynamic Time Warping (DTW) distance, which is employed to compare two events with different number of frames. Dynamic Time Warping (DTW) is a well-known technique to find an optimal alignment between two given time-independent sequences [7].

4.1 Event Dissimilarity Metric

Each frame $f_j^i,\ j = 1, \ldots, n_i$ of event E_i is represented with a visual word histogram VHF_j^i as defined in equation (3). Thus, event E_i is represented by a K-dimensional signal of length n_i:

$$VE_i = \begin{pmatrix} VHF_1^i(k=1) & \ldots & VHF_{n_i}^i(k=1) \\ \vdots & \ldots & \vdots \\ VHF_1^i(k=K) & \ldots & VHF_{n_i}^i(k=K) \end{pmatrix}, \tag{4}$$

where K the size of the vocabulary size employed to create the visual word histograms in Section 3.1. Each row k of matrix VE_i represents the frequency of "visual word" k in the time interval of the event.

In order to compute the distance between two video segments/events E_i and E_l we compute the average DTW distance of their K-multidimensional signals. More specifically

$$D(E_i, E_l) = \frac{1}{K} \sum_{k=1}^{K} DTW(VHF^i(k), VHF^l(k)), \tag{5}$$

where $VHF^i(k), VHF^l(k)$ are the k-th rows of the K-dimensional signals VE_i, VE_l representing segments/events E_i and E_l, respectively.

Fig. 3. Sample frames of the background and the five categories of events.

Table 1. Classification and Clustering results for the first video sequence.

K	1-NN		3-NN		5-NN		SVM		Hierarchical Clustering	
	DTW	EV	DTW	EV	DTW	EV	DTW	EV	DTW	EV
10	80%	85%	80%	85%	65%	65%	75%	65%	80%	45%
20	90%	90%	95%	90%	90%	80%	95%	95%	95%	90%
50	95%	95%	95%	95%	95%	90%	100%	95%	100%	90%
100	95%	90%	100%	100%	10%	95%	100%	95%	100%	100%
200	95%	90%	100%	100%	100%	100%	100%	100%	100%	100%

5 Experimental Results

5.1 Video surveillance sequence

The video sequence we used comprises of more than 25000 frames and contains different individual and not overlapping activities performed in an indoor environment captured by a standing camera. In this video sequence, 20 activities/events are performed that are divided in five categories, as presented in Fig. 3. The result of the automatic segmentation was optimal, since no over-segmentation or under-segmentation was performed and all 20 events were detected as unique.

5.2 Classification Results

To classify the 20 events into 5 categories we carried out two experiments. In the first one, we used the nearest neighbor classifier [3] and in the second one we used Support Vector Machines [1]. We implemented the nearest neighbor classifier with 1, 3, and 5 nearest neighbors for both dissimilarity measures defined in Section 4. Comparison between the visual word histograms of events is referred as EV and comparison between the visual word histograms of the frames of the events is referred as DTW. In Table 1 we present the numerical results of the experiments for different number of visual words K. The classification accuracy was estimated using the leave-one-out (LOO) approach [3].

In the second experiment, Support Vector Machine (SVM) classifiers [1] were employed using the leave one out (LOO) scheme again. In our approach, we employed the typical radial basis function (RBF) kernel and the parameters C, γ were selected through cross-validation. In Table 1, we present the numerical results for the two compared approaches of Section 4 and for different number of visual words K. It can be observed that DTW distance gives results slightly superior to the ones obtained by the other dissimilarity metric.

5.3 Clustering Results

We have also employed an unsupervised method for grouping the video events into categories. More specifically, we performed agglomerative hierarchical clustering [5], setting the number of cluster to five and using the Ward criterion to select the clusters to be merged at each iteration. In Table 1 we present the clustering accuracy for the two approaches of Section 4 using a different number of visual words K. It can be observed that DTW distance provides better results for a small number of visual words.

6 Conclusions

In this paper, we have presented a method for video event detection and classification in video surveillance sequences. For each video frame, local invariant descriptors were computed and compared to a pre-computed set of descriptors from the background framer of the surveillance room. In this way, a number of "unmatched" descriptors was identified that describe foreground objects. By analyzing the number of "unmatched" descriptors, the video sequence was segmented into segments/events. Each video event was represented either by a single (summary) visual word histogram or by a K-dimensional signal corresponding to the visual word histograms of its frames. Thus, two different approaches were followed in order to compare video events. Unsupervised and supervised learning methods were employed to cluster and classify the events into certain categories. Numerical results presented in this paper indicate that our approach achieves high detection, classification and clustering rates.

References

1. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
2. M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Transactions on Multimedia*, 9(2):257–267, 2007.
3. R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.
4. W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics*, 34:334–352, 2004.
5. A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.
6. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
7. H. Sakoe. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49, 1978.
8. T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):893–908, 2008.