Computer Vision and Image Understanding 116 (2012) 320–329

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Multi-scale and real-time non-parametric approach for anomaly detection and localization ☆

Marco Bertini *, Alberto Del Bimbo, Lorenzo Seidenari

Università degli Studi di Firenze – MICC, Firenze, Italy

ARTICLE INFO

Article history:
Received 14 March 2011
Accepted 1 September 2011
Available online 25 October 2011

Keywords:
Video surveillance
Anomaly detection
Space–time features

ABSTRACT

In this paper we propose an approach for anomaly detection and localization, in video surveillance applications, based on spatio-temporal features that capture scene dynamic statistics together with appearance. Real-time anomaly detection is performed with an unsupervised approach using a non-parametric modeling, evaluating directly multi-scale local descriptor statistics. A method to update scene statistics is also proposed, to deal with the scene changes that typically occur in a real-world setting. The proposed approach has been tested on publicly available datasets, to evaluate anomaly detection and localization, and outperforms other state-of-the-art real-time approaches.

1. Introduction and previous work

The real-world surveillance systems currently deployed are primarily based on the performance of human operators that are expected to watch, often simultaneously, a large number of screens (up to 50 [2]) that show streams captured by different cameras. One of the main tasks of security personnel is to perform proactive surveillance to detect suspicious or unusual behavior and individuals [3] and to react appropriately. As the number of CCTV streams increases, the task of the operator becomes more and more difficult and tiring: after 20 min of work the attention of an operator degrades [4]. Operators usually take into account specific aspects of activity and human behavior in order to predict possible perilous events [2], although often they can not explain their own criteria used to detect an unusual situation [3], or do not recognize unusual behaviors because they have not gathered enough knowledge of the environment and of the common behaviors they have to watch [5].

Video analytics techniques that automatically analyze video streams to warn, possibly in real-time, the operators that unusual activity is taking place, are receiving much attention from the scientific community in recent years. The detection of unusual events can be used also to guide other surveillance tasks such as human behavior and action recognition, target tracking, and person and car identification; in this latter case it is possible to use pan-tilt-zoom cameras to capture high resolution images of the subjects that caused the anomalous events.

Anomaly detection is the detection of patterns that are unusual with respect to an established normal behavior in a given dataset, and is an important problem studied in several diverse fields [6]. Approaches to anomaly detection require the creation of a model of normal data, so to detect deviations from the model in the observed data. Three broad categories of anomaly detection techniques can be considered, depending on the approach used to learn the model: supervised [7–14], semi-supervised [15,16] or unsupervised [17–28]. In this work we follow an unsupervised approach, based on the consideration that anomalies are rare and differ amongst each other with unpredictable variations.

The model can be learned off-line as in [7,8,10,29] or can be incrementally updated (as in [19,20,22,26]) to adapt itself to the changes that may occur over time in the context and appearance of a setting. Our approach continuously updates the model, to gather knowledge of common events and to deal with changes in "normal" behavior, e.g. due to variations in lighting and scene setting.

Most of the methods for identifying unusual events in video sequences use trajectories [8–10,13,15–17,23,28–30] to represent the activities shown in a video. In these approaches objects and persons are tracked and their motion is described by their spatial location. Blob features have been used in [20,27,31], without tracking the blobs. The main drawback of tracking-based approaches is the fact that only spatial deviations are considered anomalies, thus abnormal appearance or motion of a target that follows a "normal" track is not detected.

Optical flow has been used to model typical motion patterns in [11,19,21,22,31], but, as noted in [29], this measure also may become unreliable in presence of extremely crowded scenes; to solve this issue a dense local sampling of optical flow has been adopted

☆ An earlier version of this paper appeared in a conference proceeding [1].

* Corresponding author.
E-mail address: bertini@dsi.unifi.it (M. Bertini).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2011.09.009

in [12,19]. Local spatio-temporal descriptors have been successfully proposed in [32,33] to recognize human actions, while more simple descriptors based on spatio-temporal gradients have been used to model motion in [18,29] for anomaly detection. Dynamic textures have been used to model multiple components of different appearance and dynamics in [25,34].

Another issue that is common to both tracking and blob-based approaches is the fact that it is very difficult to cope with crowded scenes, where precise segmentation of a target is impossible. It is also important to consider that trajectory based methods rely on a long chain of algorithms (blob detection, data association, tracking, ground plane trajectory extraction) each of which may fail, leading to the failure of the whole anomaly detection system. Instead, approaches that are purely pixel-based, learning a scene representation independently of the explicit modeling of object motion, allow to skip the chain of intermediate decisions required by the chain of algorithms, and detect an event directly from the representation of frames.

Some recent works consider the fact that, in some cases, an event can be regarded as anomalous if it happens in a specific context; for example the interaction of multiple objects may be an anomaly even if their individual behavior, if considered separately, is normal. These works consider the scene [27,22,29], typically modeled with a grid of interest points, or the co-occurrence of behaviors and objects [14,21,25,28] like persons and vehicles.

In this work we propose a multi-scale non-parametric approach that detects and localize anomalies, using dense local spatio-temporal features that model both appearance and motion of persons and objects. Real-time performance is achieved using a careful modeling of dense sampling of overlapping features. Using these features it is possible to cope with different types of anomalies and crowded scenes. The proposed approach addresses the problem of high variability in unusual events and, using a model updating procedure, deals with scene changes that happen in real world settings. The spatial context of the spatio-temporal features is used to recognize contextual anomalies.

The rest of this paper is structured as follows: scene representation, spatio-temporal descriptor and feature sampling are described in Section 2; in Section 3 is presented the real-time anomaly detection method, with multi-scale integration, context modeling and model updating procedure; finally experimental results, obtained using standard datasets are discussed in Section 4. Conclusions are drawn in Section 5

2. Scene representation

Modeling crowd patterns is one of the most complex contexts for detection of anomalies in video surveillance scenarios. Describing such statistics is extremely complex since, as stated in Section 1, the use of trajectories does not allow to capture all the possible anomalies that may occur, e.g. due to variations of scene appearance and the presence of unknown objects moving in the scene; this is due to the fact that object detection and tracking are often unfeasible both for computational issues and for occlusions. On the other hand, global crowd descriptors are not able to describe anomalous patterns which often occur locally (e.g. a cyclist or a person moving in an unusual direction among a crowd). The most suitable choice in this context is to observe and collect local space–time descriptors.

2.1. Feature sampling

Surveillance scenes are typically captured using low frame rate cameras or at a distance, leading to a short temporal extent of actions and movements (often just 5–10 frames). Therefore, it is necessary to sample these features densely in order to obtain as complete as possible coverage of the scene statistics. This approach is also motivated by the good performance obtained using dense sampling in object recognition [35] and human action recognition [36].

The solution adopted in this work is to use spatio-temporal features that are densely sampled on a grid of cuboids that overlap in space and time. Fig. 1 shows an example of spatial, temporal and spatio-temporal overlaps of cuboids, and an example of application of overlapping spatio-temporal cuboids to a video. This approach permits localization of an anomaly both in terms of position on the frame and in time, with a precision that depends on the size and overlap of cuboids; it also models the fact that certain parts of the scene are subject to different anomalies, illumination conditions, etc., and is well suited for the typical surveillance setup where a fixed camera is observing a scene over time. Considering the position of the cuboids on the grid it is also possible to evaluate the context of an anomaly, inspecting the nearby cuboids. Moreover, it makes it possible to reach real-time processing speed, since it does not require spatio-temporal interest point localization. In our previous work [1] we have investigated how the overlap affects the performance of the system, and determined that a 50% spatial overlap provides the best performance, detecting more abnormal patterns without raising false positives, because spatial localization of the anomaly is improved. On the other hand temporal overlap does not provide an improvement and, instead, may increase false detections.

2.2. Spatio-temporal descriptors

To compute the representation of each spatio-temporal volume extracted on the overlapping regular grid, we define a descriptor based on three-dimensional gradients computed using the luminance values of the pixels (Fig. 1). Each cuboid is divided in subregions. Each subregion is described by spatio-temporal image gradient represented in polar coordinates as follows:

$$M_{3D} = \sqrt{G_x^2 + G_y^2 + G_t^2}, \tag{1}$$

$$\phi = \tan^{-1}\left(G_t / \sqrt{G_x^2 + G_y^2}\right), \tag{2}$$

$$\theta = \tan^{-1}(G_y/G_x), \tag{3}$$

where G_x, G_y and G_t are computed using finite difference approximations:

$$G_x = L_{\sigma_d}(x+1,y,t) - L_{\sigma_d}(x-1,y,t), \tag{4}$$

$$G_y = L_{\sigma_d}(x,y+1,t) - L_{\sigma_d}(x,y-1,t), \tag{5}$$

$$G_t = L_{\sigma_d}(x,y,t+1) - L_{\sigma_d}(x,y,t-1). \tag{6}$$

L_{σ_d} is obtained by filtering the signal I with a Gaussian kernel of bandwidth σ_d to suppress noise; in all the experiments we have used $\sigma_d = 1.1$, a value which proved to be effective in representing space–time patches in our previous work in human action recognition [37]. We compute two separate orientation histograms by quantizing ϕ and θ and weighting them by the magnitude M_{3D}.

It can be observed that if the overlap of cuboids precisely matches the subregions of nearby cuboids we can reuse the computations of these subregions for different cuboids' descriptors (Fig. 2). Using a number of spatial subregions that is a multiple of the overlap reduces the computational cost of the descriptors [38]: considering that a 50% overlap of cuboids is optimal then it is convenient to use an even number of spatial regions, since it is possible to reuse 50% or, depending on the position of the cuboid, 75% of the descriptors of nearby cuboids.

Therefore, we have divided the cuboid in 8 subregions, two along each spatial direction and two along the temporal direction.