Non-parametric Anomaly Detection Exploiting Space-time Features

Lorenzo Seidenari
seidenari@dsi.unifi.it

Marco Bertini
bertini@dsi.unifi.it

Dipartimento di Sistemi e Informatica - University of Florence
Viale Morgagni 65 - 50134, Florence, Italy
http://www.micc.unifi.it

ABSTRACT

In this paper a real-time anomaly detection system for video streams is proposed. Spatio-temporal features are exploited to capture scene dynamic statistics together with appearance. Anomaly detection is performed in a non-parametric fashion, evaluating directly local descriptor statistics. A method to update scene statistics, to cope with scene changes that typically happen in real world settings, is also provided. The proposed method is tested on publicly available datasets.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing; H.5.1 [**Multimedia Information Systems**]: Video

General Terms

Algorithms, Experimentation

Keywords

Anomaly detection, surveillance, local descriptors, action recognition, spatio-temporal interest points

1. INTRODUCTION

Video surveillance is becoming one of the most active domains in automatic video analysis and understanding. The video surveillance systems currently deployed rely primarily on human operators that have to watch the streams of several cameras, usually simultaneously. One of the basic tasks of these operators is to understand if some unusual behaviour is happening in the scene and then react appropriately. The growing number of CCTV cameras being deployed makes the systems based on human operators unscalable: monitoring is expensive, tiring (the attention of an operator degenerates after 20 minutes [7]) and thus ineffective. A practical solution is the deployment of automatic methods that analyze the video streams and warn, in real time, the operators when some unusual activity is taking place. Once an anomaly has been detected it is possible to perform higher level video analysis, such as behaviour and human action recognition, or exploit pan-tilt-zoom cameras to capture higher resolution images, e.g. to perform face logging and recognition.

Anomaly detection approaches require to build a model of normal data and then to attempt to detect deviations from this model in the observed data. The creation of this model can be based on supervised [1, 3, 5, 8] or unsupervised approaches [2, 4, 9, 10, 17, 19]. Given the fact that anomalies are rare and, by their very nature, have unpredictable variations, in this work we follow an unsupervised approach.

The model can be learned off-line as in [3, 5, 8, 12] or incrementally updated (as in [10]) to cope with the changes that happen over time within the visual context of a scene. Our approach continuously updates the model, to deal with changes in lighting and setting of a scene.

Most of the methods for identifying unusual events in video sequences use trajectories [1, 5, 8, 9, 12, 19] 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; only spatial deviations are considered anomalies, thus the abnormal behaviour related to the appearance or the motion of a target that follows a "normal" track is not detected. Optic flow has been used to model typical motion patterns in [10, 17] but, as noted in [12], this measure may become unreliable in presence of extremely crowded scenes. Local spatio-temporal descriptors have been successfully proposed in [6, 13] to recognize human actions, while more simple descriptors based on spatio-temporal gradients have been used to model motion in [2, 12] for anomaly detection.

In this work we propose a non-parametric approach that detects and localize anomalies in real-time, using local spatio-temporal features that model both appearance and motion of persons and objects, to deal with different types of anomalies. This approach addresses both the problem of high variability in unusual events and the need of dealing with scene changes that happen in real world settings. The paper is structured as follows: in Sect. 2 is presented the anomaly detection method; the local spatio-temporal descriptor is described in Sect. 3; finally experimental results, obtained using standard datasets, and conclusions are discussed in Sect. 4.

MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

2. NON-PARAMETRIC ANOMALY DETECTION

Our system is able to learn from a normal data distribution fed as a training set but can also start without any knowledge of the scene, learning and updating the model over time. The model can always be updated with a very simple procedure. Despite the simple formulation of this approach our system is able to model complex scenes, including both dynamic and static appearance patterns.

2.1 Semi-supervised detection

In anomaly detection tasks a certain amount of normal data is usually available; our system can exploit this data as a training set to bootstrap itself and run in a semi-supervised fashion. Our system can also be run online with no previous knowledge of the scene, since a model update procedure is provided.

To jointly capture scene motion and appearance statistics we extract pixel cuboids on a regular, slightly overlapped spatio-temporal grid. Cuboids are represented with a robust space-time descriptor described in Sect. 3. To decide if an event is anomalous a way to estimate normal descriptor statistics is needed. Moreover since no assumptions are made on the scene geometry or topology, it is important to describe this normal descriptors distribution locally w.r.t. the frame. Therefore, given a certain amount of training frames for each cell in our grid, space-time descriptors are collected and stored using a structure for fast nearest-neighbour search, providing local estimates of anomalies; an overview of this schema is shown in Fig. 1. The training stage is very straightforward, since we do not use any parametric model to learn the local motion and appearance; instead we represent the scene normality directly with descriptor instances.

A simple way to decide if an event happening at a certain time and in a certain frame location has to be considered anomalous is to perform a range query on the training set data structure to look for neighbours. Once an optimal radius for each image location is learned, all patterns for which the range query does not return any neighbour are considered anomalies. The problem with this technique is the intrinsic impossibility to select a-priori a correct value for the radius. This happens for two reasons: firstly, each scene location undergoes different dynamics, for example a street will mostly contains fast unidirectional motion generated by cars and other vehicles, while a walkway will have less intense motion and more directional variation; moreover a static part of the scene, like the side of a parking lot, will mostly contain static information. Secondly, we want to be able to update our model dynamically by adding data which has to be considered normal given the fact that we observed that kind of pattern for a sufficient amount of time; therefore, since that scene statistics has to evolve over time, the optimal radius will evolve too. Finally, we also would like to select a value that encodes the system sensitivity, i.e. the probability that the observed pattern is not generated from the underlying scene descriptors distribution.

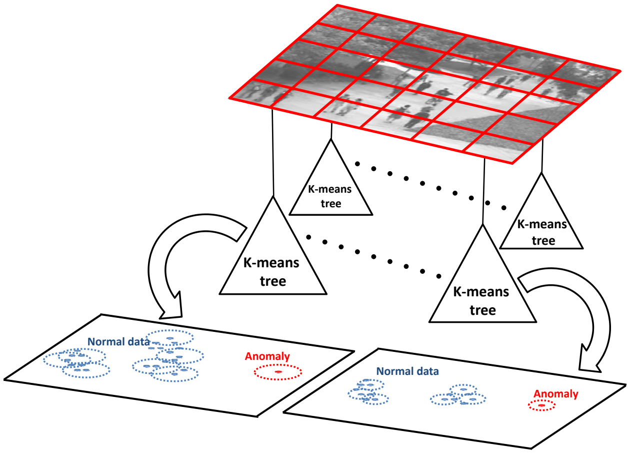

Figure 1: System overview. Each cell features are stored on efficient k-means tree based indexes. Planes underneath represent a simplified view of the high dimensional feature space; dashed circles are plotted at the optimal radius value.

To estimate the optimal radius for each data structure we compute CDF_i, the empirical cumulative distribution of nearest-neighbour distances of all interest point in the structure of cell i. Given a probability p_a below which we consider an event anomalous, we choose the radius $\hat{r}_i$ for cell i such that:

$$\hat{r}_i = CDF_i^{-1}\left(1 - p_a\right). \quad (1)$$

The anomaly probability p_a can be set to $10^{-3}, 10^{-4}, 10^{-5}\dots$ depending on the user needs to obtain a more or less sensitive system. This optimal radius formulation allows easy data-driven parameter selection and model update.

2.2 Model update

Since this kind of anomaly detection applications are thought to be run for a long time, it is very likely that a scene will change its appearance over time; very simple examples are the event of a snowstorm, the presence of some material in a yard for maintenance purposes or the placement of new temporary structures. It is therefore very urgent to provide a way to update our model. Again we propose a very straightforward data-driven technique.

Together with the data-structure for each grid cell, we keep a list of anomalous patterns. On a regular basis this list is inspected and new data is incorporated by applying the following procedure. We exploit the same range query approach presented in the previous section, to look for normality in the abnormality list. If some event happens very frequently it is likely that it will a have certain amount of neighbours in feature space, while true anomalous event will still be outliers. After an optimal radius is estimated for the anomalous pattern list, we discard all outliers in this list and incorporate all other data in the cell i training set. The optimal radius $\hat{r}_i$ for the updated cell is then recomputed.

Even if it is not required, since they can be used with default values, two parameters of the system can be tuned to adapt them to a particular scenario: grid density and overlap of cuboids. Reducing the cuboids overlap can increase the detection performance, while using a more or less dense spatio-temporal grid can serve also as a system adaptation for a specific camera resolution or frame rate. These two parameters are directly bound to physical and technical system properties (e.g. camera resolution and computer processing speed) that the user can easily access to figure out a proper configuration. Instead, the system automatically computes the optimal radius parameter, that is a quantity that is extremely task, scene and time dependent.

3. SPACE-TIME FEATURES

Space-time volumes extracted on the regular grid are represented as in the following. To compute the representation of each volume we define a descriptor based on three-dimensional gradients. Each volume is divided in 18 subregions (three along each spatial direction and two along the temporal); each subregion is described by spatio-temporal image gradient represented in polar coordinates as follows

$$\begin{aligned} M_{3D} &= \sqrt{G_x^2 + G_y^2 + G_t^2}, & (2)\\ \phi &= \tan^{-1}(G_t / \sqrt{G_x^2 + G_y^2}), & (3)\\ \theta &= \tan^{-1}(G_y / G_x) & (4) \end{aligned}$$

where G_x, G_y and G_t are respectively computed using finite difference approximations: $L_{\sigma_d}(x+1,y,t) - L_{\sigma_d}(x-1,y,t)$, $L_{\sigma_d}(x,y+1,t) - L_{\sigma_d}(x,y-1,t)$ and $L_{\sigma}(x,y,t+1) - L_{\sigma_d}(x,y,t-1)$. L is obtained by filtering the signal I with a Gaussian kernel of bandwidth σ_d. We compute two separated orientation histograms quantizing ϕ and θ, weighting them by the magnitude M_{3D}. We do not apply a re-orientation of the 3D neighbourhood, since rotational invariance,which is invaluable in object detection and recognition, is not desired in a human behavior and scene modelization context. The ϕ (with range, $-\frac{\pi}{2}, \frac{\pi}{2}$) and θ ($-\pi$,π) are quantized in four and eight bins, respectively. The overall dimension of the descriptor is thus $3 \times 3 \times 2 \times (8+4) = 216$. This construction of the three-dimensional histogram is inspired, in principle, by the approach proposed by Scovanner *et al.* [21], where they construct a weighted three-dimensional histogram normalized by the solid angle value (instead of quantizing separately the two orientations) to avoid distortions due to the polar coordinate representation. However, we have found that our method is computationally less expensive, equally effective in describing motion information given by appearance variation, and showing a state-of-the-art performance (see Tab. 1).

4. EXPERIMENTAL RESULTS

4.1 Descriptor evaluation

The descriptor is initially tested in an action recognition problem on two standard dataset. KTH dataset contains videos of 25 people performing 6 different actions in 4 recording conditions; Weizmann is made of 93 videos of 9 actors performing 10 different actions. KTH is considered more challenging because of illumination and scale variation and for the amount of actors involved in the recording. A bag-of-words system is used for this test, using k-means clustering for the dictionary creation and SVM with χ^2 kernel as a classifier. Table 1 compares the average accuracy obtained by our descriptor with state-of-the-art descriptors: the performance is above or in line with the other approaches, but without requiring any tuning of descriptor parameters.

Method	**KTH**	**Weizmann**
Our method	90.38	92.30
Rapantzikos *et al.* [20]	88.3	-
Laptev *et al.* [14]	91.8	-
Dollár *et al.* [6]	81.2	-
Wong and Cipolla [24]	86.62	-
Scovanner *et al.* [21]	-	82.6
Niebles *et al.* [18]	83.33	90
Liu *et al.* [15]	-	90.4
Kläser *et al.* [11]	91.4	84.3
Willems *et al.* [23]	84.26	-

Table 1: Comparison of our method to state-of-the-art.

4.2 System evaluation

We tested our approach on UCSD[1] anomaly dataset which provides a frame-by-frame local anomaly annotation. Videos are recorded at a resolution of 238 × 158 and 10 fps. A subset of the dataset has also spatial binary frame masks to enable spatial accuracy performance evaluation. This dataset mostly contains sequences of pedestrians in walkways; annotated anomalies, that are not staged, consider non-pedestrian entities (bikers, skaters, small carts) accessing the walkway, pedestrians moving in anomalous motion patterns or in non walkway regions. The dataset is split in two sets of sequences, each of which is recorded from a different camera and corresponds to a different scene. The first subset contains 34 training video samples and 36 testing video samples, while the latter contains 16 training video samples and 14 testing video samples for a global amount of 100. Each sequence lasts around 200 frames, for a total dataset duration of $\sim$ 33 minutes. Fig. 2 reports the precision-recall curve for this dataset created varying the p_a parameter from 10^{-2} to 10^{-5}, showing a good performance. Fig. 3 shows a qualitative comparison of anomaly localization of our approach with other state-of-the-art approaches, while Fig. 4 shows other examples of anomaly localization of our approach. Another test has been performed using a dataset created by us, recorded in a parking lot with wall-mounted pan-tilt-zoom cameras. The video has a resolution of 320 × 240 and 6 fps, and is publicly available on the ViSOR[2] site [22]. The video has been recorded during a single day and is composed by a sequence of more than 5 hours, without anomalies, and some staged anomalous behaviors like people fighting, running or waving hands. The performance obtained on this dataset is reported in Table 2. We used 8×8 grids of 5 frames long cuboids; in our experiments we have seen that increasing the grid density (e.g. 16×16) improves performances and reduces noise at the cost of a heavier computation thus compromising partially the real-time behavior of our system; cuboids longer than 5 frames instead have a worst performance, probably due to the fact that events span a very little time frame, considering also the low frame rate.

Conclusions

In this paper we have presented a non-parametric anomaly detection approach that can be executed in real-time in a completely unsupervised manner. Spatio-temporal features

[1] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm
[2] http://www.openvisor.org

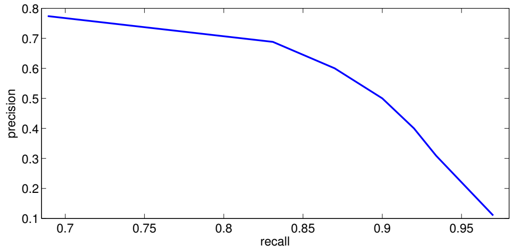

Figure 2: Precision-Recall curve for the UCSD dataset.

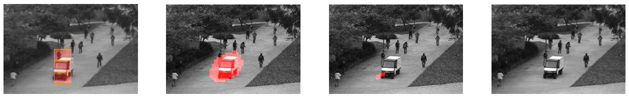

Figure 3: Qualitative comparison with methods presented in [16, 17]: our method, mixture of dynamic textures, social force and mixture of principal components analyzers, social force only.

True Pos.	False Pos.	False Neg.	Precision	Recall
218	38	86	.85	.72

Table 2: Precision and recall on MICC Dataset.

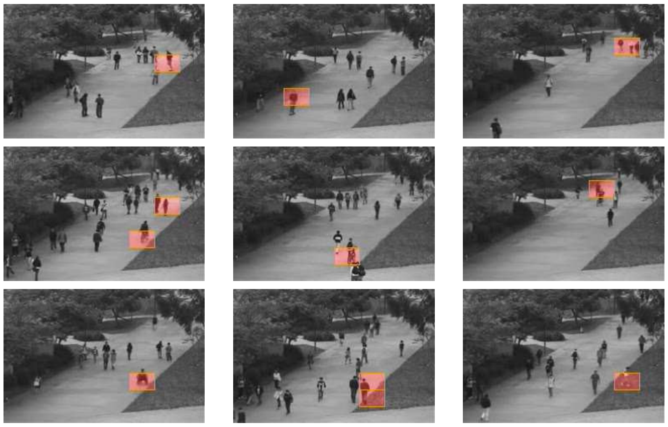

Figure 4: Anomaly detection results on UCSD dataset. Detected anomalies are skaters, bikers and trolleys or vehicles (see Fig. 3). Our system also detects a wheelchair and people walking off the walkway as anomalous patterns.

that capture appearance and motion information have been used to capture the scene dynamics. Our future work will deal with the expansion of the system to include appearance context.

5. REFERENCES

[1] P. Antonakaki, D. Kosmopoulos, and S. J. Perantonis. Detecting abnormal human behaviour using multiple cameras. *Signal Processing*, 89(9):1723 – 1738, 2009.

[2] O. Boiman and M. Irani. Detecting irregularities in images and in video. *International Journal of Computer Vision*, 74(1):17–31, Aug. 2007.

[3] C. Brax, L. Niklasson, and M. Smedberg. Finding behavioural anomalies in public areas using video surveillance data. In *Proc. of 11th International Conference on Information Fusion*

[4] M. Breitenstein, H. Grabner, and L. Van Gool. Hunting Nessie: Real time abnormality detection from webcams. In *Proc. of ICCV'09 WS on Visual Surveillance*, 2009.

[5] S. Calderara, C. Alaimo, A. Prati, and R. Cucchiara. A real-time system for abnormal path detection. In *Proceedings of 3rd IEE International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*, London, UK, Dec. 2009.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. of VSPETS*, 2005.

[7] N. Haering, P. Venetianer, and A. Lipton. The evolution of video surveillance: an overview. *Machine Vision and Applications*, 19(5-6):279–290, Oct. 2008.

[8] I. Ivanov, F. Dufaux, T. M. Ha, and T. Ebrahimi. Towards generic detection of unusual events in video surveillance. In *Proc. of AVSS*, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[9] F. Jiang, Y. Wu, and A. Katsaggelos. A dynamic hierarchical clustering method for trajectory-based unusual video event detection. *IEEE Transactions on Image Processing*, 18(4):907–913, Apr. 2009.

[10] J. Kim and K. Grauman. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *Proc. of CVPR*, 2009.

[11] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-Gradients. In *Proc. of BMVC*, 2008.

[12] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proc. of CVPR*, 2009.

[13] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. of CVPR*, 2008.

[15] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *Proc. of CVPR*, 2008.

[16] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Proc. of CVPR*, 2010.

[17] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Proc. of CVPR*, 2009.

[18] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[19] C. Piciarelli and G. Foresti. On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters*, 27(15):1835 – 1842, 2006. Vision for Crime Detection and Prevention.

[20] K. Rapantzikos, Y. Avrithis, and S. Kollia. Dense saliency-based spatiotemporal feature points for action recognition. In *Proc. of CVPR*, 2009.

[21] P. Scovanner, S. Ali, and M. Shah. A 3-Dimensional SIFT descriptor and its application to action recognition. In *Proc. of ACM Multimedia*, 2007.

[22] R. Vezzani and R. Cucchiara. Video surveillance online repository (ViSOR): an integrated framework. *Multimedia Tools and Applications*, in press, 2010.

[23] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. of ECCV*, 2008.

[24] S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *Proc. of ICCV*, 2007.