Dense Spatio-temporal Features For Non-parametric Anomaly Detection And Localization

Lorenzo Seidenari, Marco Bertini, Alberto Del Bimbo
Dipartimento di Sistemi e Informatica - University of Florence
Viale Morgagni 65 - 50134, Florence, Italy
{seidenari,bertini,delbimbo}@dsi.unifi.it

ABSTRACT

In this paper we propose dense spatio-temporal features to capture scene dynamic statistics together with appearance, in video surveillance applications. These features are exploited in a real-time anomaly detection system. Anomaly detection is performed using a non-parametric modelling, evaluating directly local descriptor statistics, and an unsupervised or semi-supervised approach. 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 and compared to other state-of-the-art approaches.

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

Currently, the video surveillance systems that are deployed rely primarily on human personnel that have to watch, usually simultaneously, the streams of several cameras. One of the main objectives of these operators is to identify if some unusual event is happening in the scene and then react appropriately. The growing number of CCTV cameras being deployed makes the systems based on human operators not scalable: monitoring is expensive and tiring (after 20 minutes of work the attention of an operator degrades [9]) and thus becomes ineffective. To solve these issues video analytics techniques that automatically analyze the video streams and warn, in real time, the operators when some unusual activity is taking place, are getting a large attention in the scientific community. Once an anomaly has been detected it is possible to perform higher level video analysis, such as target tracking, 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 variations in the observed data from this model. The model can be learned using supervised [2, 5, 7, 10] or unsupervised approaches [1, 4, 6, 11, 12, 16, 17]. Given the fact that anomalies are rare, differing amongst each other with unpredictable variations, in this work we follow an unsupervised approach.

The model can be learned off-line as in [5, 7, 10, 13] or incrementally updated (as in [1, 12]) to adapt itself w.r.t. the changes that happen over time within the visual context of a setting. Our approach continuously updates the model, to deal with changes in "normal" behaviour, e.g. due to variations in lighting and scene setting.

Most of the methods for identifying unusual events in video sequences use trajectories [2, 7, 10, 11, 13, 17] 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; the main drawbacks of tracking-based approaches are 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, and the fact that it is very difficult to cope with crowded scenes. Optic flow has been used to model typical motion patterns in [1, 12, 16] but, as noted in [13], also this measure may become unreliable in presence of extremely crowded scenes; to solve this issue a dense local sampling of optic flow has been adopted in [1]. Local spatio-temporal descriptors have been successfully proposed in [8, 14] to recognize human actions, while more simple descriptors based on spatio-temporal gradients have been used to model motion in [4, 13] for anomaly detection.

In this work we propose a non-parametric approach that detects and localize anomalies in real-time, using dense local spatio-temporal features that model both appearance and motion of persons and objects. Using these features it is possible to deal with different types of anomalies and with crowded scenes. 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: the local spatio-temporal descriptor is described in Sect. 2; in Sect. 3 is presented the anomaly detection method; finally experimental results,

ARTEMIS'10, October 29, 2010, Firenze, Italy.

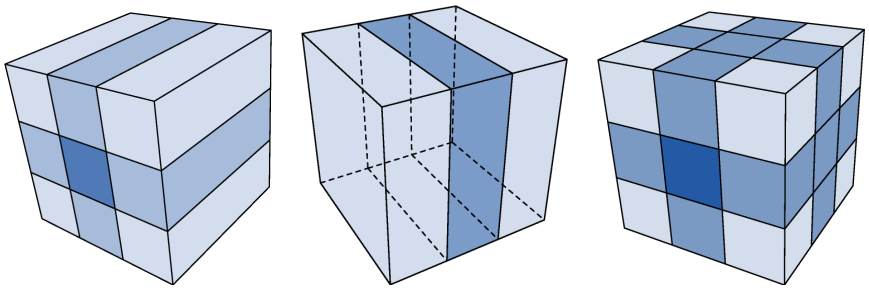

Figure 1: Examples of overlapping cuboids: i) spatial overlap, ii) temporal overlap, iii) spatio-temporal overlap.

obtained using standard datasets, and conclusions are discussed in Sect. 4.

2. SPATIO-TEMPORAL FEATURES

Detecting abnormal situations in video-surveillance scenarios has often to deal with the modelling of crowd patterns. Describing such statistics is extremely complex since object detection and tracking is often unfeasible both for computational issues and for occlusions; moreover as stated in Sec. 1, the use of trajectories does not allow to capture variations of scene appearance and the presence of unknown objects moving in the scene. Global crowd descriptors are not able to describe anomalous patterns which often occurs locally (e.g. a biker or a person moving in an unusual direction among a crowd). The most suitable choice in this context is to observe and collect very short local space-time patches. Due to the short temporal extension (5-10 frames) of actions and movements, especially if the scene is filmed at a distance as typical in surveillance, is necessary to sample this features overlapped both in time and space so to obtain an almost complete coverage of the scene statistics.

The spatio-temporal features used in the system are densely sampled using a grid of cuboids that overlap in space and time. Fig. 1 shows an example of spatial, temporal and spatio-temporal overlaps of cuboids. This approach allows to precisely localize an anomaly both in terms of position on the frame and in time; it models also 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. In addition it makes it possible to reach real-time processing speed, since it does not require to perform spatio-temporal interest point localization.

The spatio-temporal volumes extracted on the overlapping regular grid are represented as in the following. To compute the representation of each volume we define a descriptor based on three-dimensional gradients computed using the luminance values of the pixels (Fig. 2). 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

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

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

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

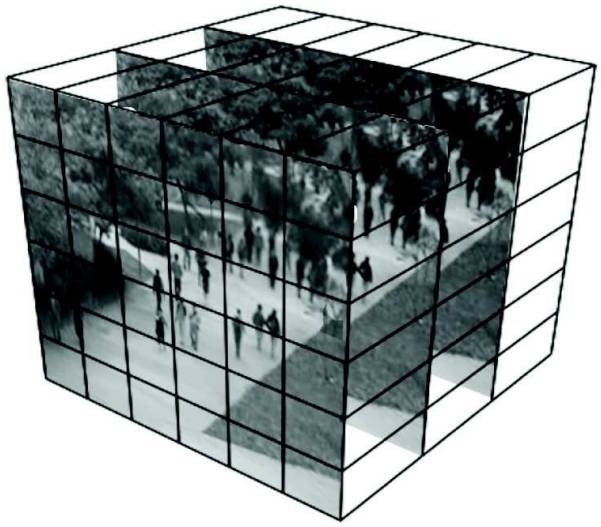

Figure 2: Example of cuboids extraction.

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 (in all the experiments we have used $\sigma_d = 1.1$). We compute two separated orientation histograms quantizing ϕ and θ, weighting them by the magnitude M_{3D}. This descriptor is robust w.r.t. illumination and lighting changes, as required in a surveillance context in which a video could be recorded over a large extent of time. We do not apply a re-orientation of the 3D neighbourhood, since rotational invariance, otherwise useful in object detection and recognition tasks, is not desirable 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.* [18], 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 appear-

ance variation, and showing an accuracy of human action recognition that is above or in line with other state-of-the-art descriptors [3], but without requiring tuning of descriptor parameters.

3. 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 "normal behaviour" profile dynamically. 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 and crowded scenes, including both dynamic and static appearance patterns.

3.1 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 on-line 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. 2. In order to decide if an event is anomalous there is need of a method to estimate normal descriptor statistics. 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. 3. 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 location of the video stream 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 variations of the direction; 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

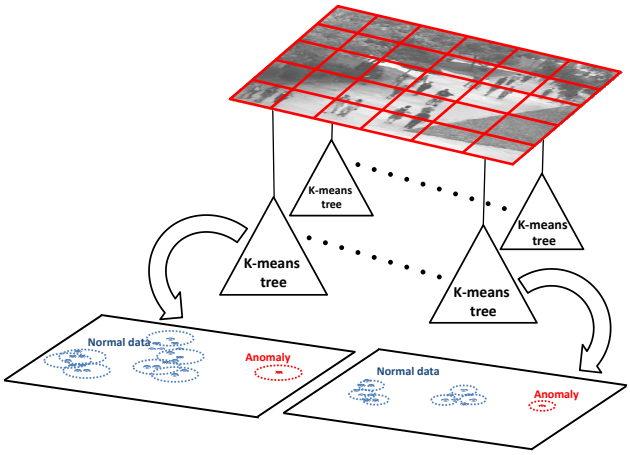

Figure 3: System overview. Each cell features are stored on efficient k-means tree based indexes. Planes underneath represent a simplified view of the high dimensional feature space; dashed circles are plotted at the optimal radius value.

probability that the observed pattern is not generated from the underlying scene descriptors distribution.

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 the cell i of the overlapping grid. Given a probability p_a below which we consider an event anomalous, we choose the radius $\hat{r}_i$ for cell i as:

$$\hat{r}_i = CDF_i^{-1}(1 - p_a)\,. \tag{4}$$

The anomaly probability p_a can be set to $10^{-2}, 10^{-3}, 10^{-4}$, $10^{-5}, \ldots$ depending on the user's need to obtain a more or less sensitive system. This optimal radius formulation allows easy data-driven parameter selection and model update.

3.2 Model update

Since the applications for anomaly detection in video surveillance are designed to be executed 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 cars that enter and exit a parking lot or the placement of temporary structures in a setting. 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 overlapping 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 subsection, 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 the estimation of an optimal radius 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.

4. EXPERIMENTAL RESULTS

4.1 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 using fixed cameras that overlook pedestrian walkways. This dataset mostly contains sequences of pedestrians in walkways; annotated anomalies, that are not staged, are non-pedestrian entities (bikers, skaters, small carts) accessing the walkway and 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. We tested our approach on the first split of the UCSD dataset. Each anomalous frame in the testing set is annotated; for each cuboid classified as anomalous, we flag as anomalous each region of the frames from which it was created; frames that contain at least one anomalous region are considered anomalous.

In the first set of experiments we evaluated the best parameters for dense sampling and overlapping of the spatio-temporal descriptors. Results are reported using the ROC curve and the Equal Error Rate (EER) - that is the rate at which both false positives and misses are equal. Fig. 5 reports the ROC curves while varying the parameters of size and spatial overlapping of cuboids (reported in terms of pixels), with p_a varying from 10^{-5} to 10^{-2}. Reducing the cuboid size allows a more precise localization of anomalies but, as a drawback, may increase the amount of false detections, see Fig. 1 for an example of this behavior of our system. We also tested different cuboid temporal extensions in the range 5-12 frames, finding that the best performance is obtained with a value of 8. The EER curve in Fig. 4 shows that the best results are obtained for cuboids of 40 × 40 pixels, while the ROC curve shows that a 50% spatial overlap achieves the lowest EER (i.e. intersection of the ROC curve with the dashed line), while temporal overlap has no substantial benefit; spatial overlap helps to detect more abnormal patterns without raising false positives since it improves the spatial localization of the anomaly, while a dense sampling in time, derived from temporal overlap, increases the false positive rate leading to a slightly degraded performance.

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

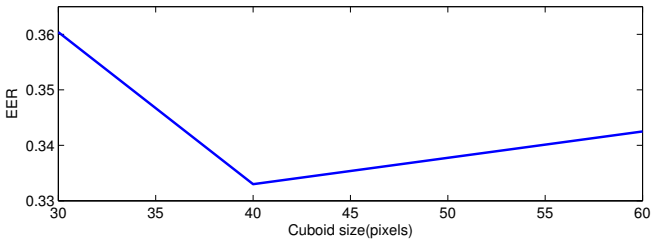

Figure 4: Equal Error Rate varying cuboid size.

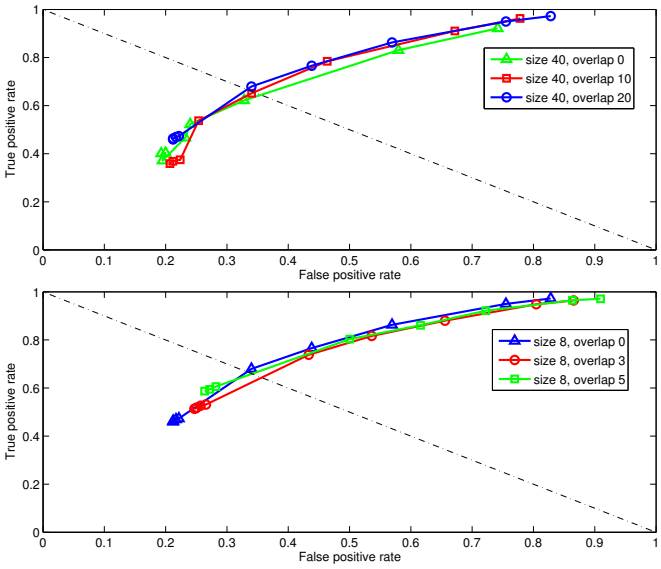

Figure 5: ROC curves for the UCSD dataset, while varying size and spatial overlap of cuboids (top) and temporal extent and overlap (bottom).

Since the approach aims at real-time processing, we have evaluated the impact of the dense sampling of cuboids, computing the average number of processed frames per second while varying the spatial overlap of cuboids. Fig. 6 shows that even with 50% spatial overlap, the system is able to process 17 frames per second, despite the fact that no code optimization, like parallelization, has been adopted. Cuboid size does not affect the computation time since smaller cuboids imply an increased number of descriptors which are faster to compute while bigger cuboids generate fewer but slower to compute descriptors. This results were obtained on a 2.6 GHz CPU with 3 GB of RAM.

Since in video surveillance the precision of the alarms is important, because a human operator may be disturbed by a high number of false alarms, in Fig. 7 we report the precision-recall curve for the UCSD dataset, created varying the p_a parameter from 10^{-5} to 10^{-2}, showing a good performance; considering low probabilities p_a for the anomalies reduces the recall, while raising the precision, and viceversa. In particular the break-even point at 0.70 of precision and recall is obtained for $10^{-4} \leq p_a \leq 10^{-3}$. Fig. 9 shows a qualitative comparison of anomaly localization of our approach with other state-of-the-art approaches, while Fig. 10 shows other examples of anomaly localization of our approach.

We compare our system with results of other state-of-the-

Figure 9: Qualitative comparison of anomaly localization with other methods: our method, mixture of dynamic textures [15], social force and mixture of principal components analyzers [12,15], social force only [16].

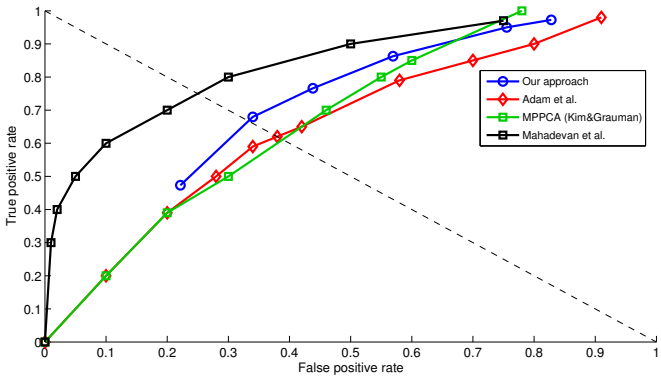

Figure 8: ROC curve to compare of our method with state-of-the-art approaches. The dashed diagonal is the EER line.

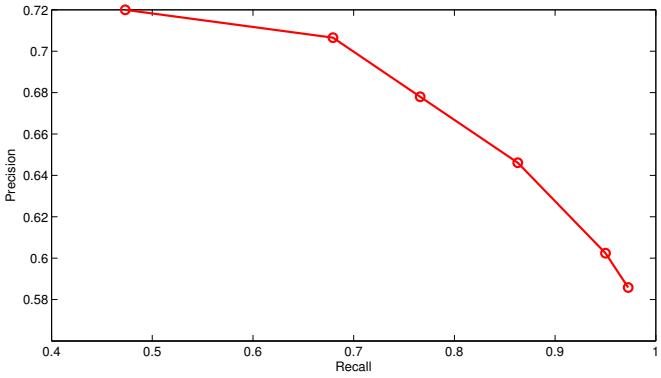

Figure 7: Precision-Recall curve for the UCSD dataset.

art approaches, as they are reported in [15]: MPCCA [12], Adam *et al.* [1] and Mahadevan *et al.* [15]. Fig. 8 shows that our approach obtains the second best result after the method proposed in [15], but it has to be noted that this approach is not suitable for real-time processing since it takes 25 seconds to process a single frame on a computer with a computational power comparable to the machine used in our experiments.

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. We have also provided a straightforward procedure to dynamically update the learned model, to deal with scene changes that happen in real-world surveillance scenarios. Dense and overlapping spatio-temporal features, that model appearance and motion information, have been used to capture the scene dynamics, allowing the detection of anomalies, such as carts and bicycles ona pedestrian walkway in a challenging crowded scene which cannot be modeled using trajectories or pure motion statistics (optical flow).

A comparison on a publicly available dataset shows that our method executes in real-time and achieves the best performance with respect to existing state-of-the-art real-time solutions [1, 12]. Our future work will deal with the expansion of the system to model contextual appearance, motion and temporal information.

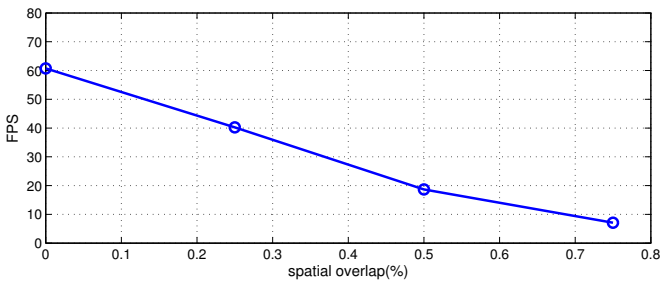

Figure 6: Number of frames per second (FPS) processed while varying the spatial overlap of cuboids.

5. REFERENCES

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed- location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560, March 2008.

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

[3] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. Recognizing human actions by fusing spatio-temporal appearance and motion descriptors. In *Proc. of International Conference on Image Processing (ICIP)*, Cairo, Egypt, November 2009.

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

[5] 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*, 2008.

Figure 10: Anomaly detection results on UCSD dataset. Detected anomalies are skaters, bikers and trolleys or vehicles (see Fig. 9). Our system also detects a wheelchair and people walking off the walkway as anomalous patterns.

Figure 11: Spatial anomaly localization varying the spatial cuboid extension (20, 30, 40, 60 pixels).

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *Proc. of CVPR*, San Francisco, CA, USA, 2010.

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

[17] 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.

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