

UNIVERSITÀ
DEGLI STUDI
FIRENZE

FLORE
Repository istituzionale dell'Università degli Studi di Firenze

Scene and crowd behaviour analysis with local space-time descriptors

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

Original Citation:

Scene and crowd behaviour analysis with local space-time descriptors / M. Bertini;A. Del Bimbo;L. Seidenari. - STAMPA. - (2012), pp. 1-6. (Intervento presentato al convegno Proc. of 5th International Symposium on Communications, Control, and Signal Processing (ISCCSP) nel 2012-May) [10.1109/ISCCSP.2012.6217857].

Availability:

This version is available at: 2158/656341 since:

Publisher:

IEEE

Published version:

DOI: 10.1109/ISCCSP.2012.6217857

Terms of use:

Open Access

La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf)

Publisher copyright claim:

(Article begins on next page)

19 March 2024

Proceedings of the 5th International Symposium on Communications, Control and Signal Processing,
ISCCSP 2012, Rome, Italy, 2-4 May 2012

SCENE AND CROWD BEHAVIOUR ANALYSIS WITH LOCAL SPACE-TIME DESCRIPTORS

Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari

Media Integration and Communication Center, Università degli Studi di Firenze
Viale Morgagni 65 - 50134 Firenze, Italy
bertini, delbimbo, seidenari@dsi.unifi.it

ABSTRACT

In this paper we propose a local space-time descriptor to be employed for behaviour analysis in video-surveillance applications. We show how this local video representation is able to extract scene semantics in both a supervised (behaviour recognition) and semi-supervised (anomaly detection) setup. Our approach yields state-of-the art performance on two publicly available datasets and is not computationally intensive.

Index Terms— Anomaly detection, local descriptors, behaviour recognition, spatio-temporal descriptors

1. INTRODUCTION AND PREVIOUS WORK

The performance of the real-world surveillance systems currently deployed is essentially dependent on that of the human operators whose duty is to watch, typically simultaneously, a large number of screens showing the video streams captured by different cameras. One of the main tasks of this security staff is to detect suspicious or unusual behaviour of individuals and crowds, so to react appropriately.

Video analysis techniques that automatically interpret video streams to warn the operators, possibly in real-time, that unusual activity or certain human action 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 behaviour and action recognition, target tracking, and person and car identification (e.g. using pan-tilt-zoom cameras to capture high resolution images of the subjects).

Some interesting research contributions to human action recognition have proposed methods to automatically detect and recognize classes of human activity [1–4]. However, building a general recognition and classification system for this type of dynamic facts has proven to be very challenging, because of the many variations in setting, persons and actions that may be observed: setting variations can be caused by clutter or background changes, scene illumination changes and camera motion; person appearance may change in size, shape and posture; actions that are semantically equivalent can manifest differently or partially. All these issues reflect in the difficulty of defining effective descriptors of spatio-temporal facts.

A possible solution is to use local representations, that have shown better performance for videos in unconstrained scenes. In fact, they are less sensitive to partial occlusions and noise and overcome some limitations of holistic models, such as the necessity of performing background subtraction and target tracking. In this approach, body movements and scene changes are described from the observations of spatio-temporal interest points, computing robust local descriptors of the local patches either around salient points or in correspondence of grids used for dense sampling. Laptev [5] proposed an extension to the Harris-Förstner corner detector for the spatio-temporal case; interesting parts were extracted from voxels surrounding local maxima of spatio-temporal corners, i.e. locations of videos which exhibit strong variations of intensity both in spatial and temporal directions. The extension of the scale-space theory to the temporal dimension permitted to define a method for automatic scale-selection. Laptev's descriptors were used in [6] to define a set of codewords so to have each video shot of a human action represented by a histogram of dynamic visual words. Dollár *et al.* [7] proposed a different descriptor than Laptev's, by looking for locally periodic motion. While this method produces a denser sampling of the spatio-temporal volume, it does not provide automatic scale-selection. Despite of it, experimental results have shown that it improves w.r.t. [6]. In [8] Willems *et al.* extended SURF feature to time and defined a new scale-invariant spatio-temporal detector and descriptor that showed high efficiency. More simple descriptors based on spatio-temporal gradients have been used to model motion in [9, 10] for anomaly detection. Dynamic textures have been used to model multiple components of different anomaly appearance and dynamics in [11, 12].

In this paper we propose a local space-time descriptor to be employed for behaviour recognition and anomaly detection in video-surveillance applications. The descriptor can be used in real-time applications and obtains state-of-the art performance on two publicly available datasets .

978-1-4673-0276-0/12/$31.00 ©2012 IEEE

2. THE SPATIO-TEMPORAL DESCRIPTOR

2.1. Feature sampling

Similarly to the approaches used in scene and object detection, local spatio-temporal descriptors can be computed around sparsely or densely sampled points. In the first case, in our approach, the spatio-temporal interest points are detected at video local maxima of the Dollár's detector applied over a set of spatial and temporal scales. Multiple scales are used to capture the essence of motion activity. Linear filters are separately applied to the spatial and temporal dimension: the spatial scale permits to detect visual features of higher or lower detail, while the temporal scale allows to detect *action primitives* performed at different speeds. The response function is defined as $R = \left(I * g_\sigma * h_{ev}\right)^2 + \left(I * g_\sigma * h_{od}\right)^2$ where $I(x, y, t)$ is a sequence of gray-level images, $g_\sigma(x, y)$ is the spatial Gaussian filter with kernel σ, h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters in the temporal dimension defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$, where $\omega = 4/\tau$. They provide a strong response to temporal intensity changes, specifically to periodic motion patterns. In the experiments we used $\sigma = \{2, 4\}$ as spatial scales and $\tau = \{2, 4\}$ as temporal scales. In the second case we sample the features on a grid, in order to obtain a coverage of the scene statistics that is as complete as possible. This approach is also motivated by the good performance obtained using dense sampling in object recognition [13] and human action recognition [14]. When considering the task of anomaly detection 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. This overlap has three effects: it performs a more precise localization of an anomaly both in terms of position and time; it takes into account the fact that certain parts of the scene are subject to different anomalies and illumination conditions; it is well suited for the typical fixed camera setup used in surveillance deployments. In our previous work [15] 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. Preliminary experiments have instead shown that temporal overlap does not provide an improvement and may even increase false detections.

2.2. Descriptor computation

Spatio-temporal volumes are taken around spatio-temporal interest points selected using one the approaches presented in the previous subsection, and divided into equally sized regions: S_x, S_y for the spatial dimensions x, y and T for the temporal dimension t.

For each region a descriptor that accounts for the variation of appearance and motion is obtained as follows. Image gradients on x, y and t direction are computed as:

$$\begin{aligned} G_x &= I(x+1, y, t) - I(x-1, y, t), \\ G_y &= I(x, y+1, t) - I(x, y-1, t), \\ G_t &= I(x, y, t+1) - I(x, y, t-1). \end{aligned}$$

and orientations are computed for each pixel:

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

The orientation histograms of ϕ and θ are derived, by weighting the contribution of each pixel with the gradient magnitude $M_G = \sqrt{G_x^2 + G_y^2 + G_t^2}$ (for ϕ and θ). The orientations ϕ (with range $-\frac{\pi}{2}, \frac{\pi}{2}$) and θ ($-\pi$,π) are quantized in four and eight bins respectively. The descriptor is obtained by concatenating the histograms of ϕ and θ of each region. The total dimension is $S_x \times S_y \times T \times (8 + 4)$. The choice of S_x, S_y has an influence on the speed of computation of the descriptor when using dense sampling with overlap. In fact, using a number of spatial subregions that is a multiple of the overlap reduces the computational cost of the descriptors: 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. The choice of T is dependent on the frame rate of the video: typically surveillance videos are captured with low framee rate, so that there is no need to have a fine grained temporal subdivision. Therefore, we have divided the cuboid in 8 subregions ($S_x = S_y = 2$), two along each spatial direction and two along the temporal direction. This choice increases the speed of the system of about 50%, with respect to a division of cuboids in $3 \times 3 \times 2$ regions [15]. The final size of the descriptor is then $2 \times 2 \times 2 \times (8 + 4) = 96$.

This construction of the three-dimensional histogram is inspired, in principle, by the approach proposed by Scovanner *et al.* [16], where they construct a weighted three-dimensional histogram normalized by the solid angle value (instead of separately quantizing 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 shows an accuracy of human action recognition that is above or in line with other state-of-the-art descriptors [17], but without requiring tuning of descriptor parameters.

3. BEHAVIOUR AND ANOMALY MODELLING

3.1. Supervised approach for behaviour recognition

The proposed descriptor can be used to represent video sequences using a bag of spatio-temporal visual words, follow-

ing the successful results achieved in object and scene classification. The basic idea of the bag-of-visual-words (BoVW) approach is to represent visual content as an unordered collection of (visual) words. To this end, it is necessary to define a visual vocabulary from the local features extracted in the video sequences, performing a quantization of the original feature space. The visual vocabulary is generated by clustering of a set of interest points and each cluster is treated as a visual word. In particular, we use the k-means algorithm because of its simplicity and convergence speed. By mapping the features extracted from a video to the vocabulary, we can represent it by the frequency histogram of visual words. Then, this histogram is fed to a classifier to predict the action category.

In our work classification is performed using non-linear SVMs with the χ^2 kernel [18]. For multi-class classification, we use the *one-vs-one* approach.

3.2. Unsupervised approach for anomaly detection

Considering that anomalies are rare and differ amongst each other with unpredictable variations, in this work we follow an unsupervised approach. Our technique is inspired by the one proposed in [19], where the proposed scene representation is global and static, based on global histograms of oriented gradients of single frames. Instead, in our approach, we represent the scene using local spatio-temporal features with dense sampling and we exploit the idea of the adaptive threshold in order to learn, over time, local models for different portions of the scene. Another significant difference with respect to [19] is the use of pure data instead of clusters, to avoid to corrupt data distribution and to produce a more accurate estimation of the distance threshold used to detect anomalies. Also the model update procedure is different: since we are not applying any clustering procedure to data, our model update can be performed just by analyzing the detected anomalies stored over time; therefore it can be performed more frequently, without the need to operate either in detection mode or in maintenance mode.

In order to decide if an event is anomalous we need a method to estimate normal descriptor statistics. Since no assumptions are made on the scene setup, it is important to define this normal descriptor distribution locally with respect to 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. The training stage is very straightforward, since we do not use any parametric model to learn the local motion and appearance. A simple way to decide if an event happening at a certain time and location of the video stream should be considered anomalous, is to perform a range query on the training set data structure, looking 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 main issue with this technique is the intrinsic impossibility of selecting *a priori* a correct value for the radius. This happens for several reasons: firstly, each scene location undergoes different dynamics, for example a street will mostly contains fast unidirectional motion generated by vehicles, a walkway will have less intense motion and more variations of the direction and, instead, 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 should be considered normal given the fact that we observed that kind of pattern for a sufficient amount of time; therefore, scene statistics will evolve over time and 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.

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

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

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. After setting such value p_a, optimal radii are estimated for each cell with likely different values. This optimal radius formulation allows easy data-driven parameter selection and model update.

4. EXPERIMENTAL RESULTS

4.1. Experiments on crowd behaviours

In this experiment we analyze the ability of our descriptor to classify crowd behaviours, in particular normal pedestrian behaviours and panic. We tested our approach on the crowd behaviour dataset created by the University of Minnesota (UMN)[1]. The dataset comprises the videos of 11 different scenarios in different indoor and outdoor scenes. Each video consists of an initial part of normal behaviour and ends with sequences of panicking and escaping. Videos are recorded at 30 FPS with a resolution of 320×240. Examples of these behaviours are shown in Fig. 1. We split the dataset in four sub-sets segmenting different scenes which include different indoor and outdoor settings. For each video of each sub-set we extracted clips, with a length of 100 frames, from each "normal" and "panic" parts, obtaining a total of 88 video clips. Experiments are carried with a 4-fold cross-validation per scene, avoiding to classify a video using a model trained on clips from the same scene. We test our approach evaluating different interest points detection strategies. In particular

[1] http://mha.cs.umn.edu/proj_events.shtml

we tested both the detector based (sparse) approach and a dense sampling approach. The dense sampling has shown to perform best with a 97.43% accuracy while the detector based approach accuracy is 85.89%. This is probably due to the fact that in this dataset persons are relatively small w.r.t. the frame size and the detector finds too few interest points.

4.2. Experiments on anomaly detection

We tested our approach on the UCSD[2] anomaly dataset presented in [12], which provides frame-by-frame local anomaly annotation. The dataset consists of two subsets, corresponding to different scenes using fixed cameras that overlook pedestrian walkways: one (called Peds1) contains videos of people moving towards and away from the camera, with some perspective distortion; the other (called Peds2) shows pedestrian movement parallel to the camera. Videos are recorded at 10 FPS with a resolution of 238×158 and 360×240, respectively. This dataset mostly contains sequences of pedestrians in walkways; annotated anomalies, that are not staged, are non-pedestrian entities (cyclists, skaters, small carts) accessing the walkway and pedestrians moving in anomalous motion patterns or in non-walkway regions. The first subset contains 34 training video samples and 36 testing video samples, while the latter contains 16 training video samples and 12 testing video samples. Each sequence lasts around 200 frames, for a total dataset duration of $\sim$ 33 minutes. 10 videos of the Peds1 subset have manually generated pixel-level binary masks, which identify the regions containing anomalies. 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. We follow the evaluation procedure reported in [12]: in the frame level evaluation an abnormal frame is considered correctly detected if at least one pixel of the frame is detected as anomalous; in the pixel level evaluation an abnormal frame is considered correctly detected if at least the 40% of the anomalous pixels are detected correctly and considered a false positive otherwise. A "lucky guess" happens when a region different from the one that generated the anomaly is detected as anomalous in the same frame. The frame level detection evaluation does not takes into account this phenomenon. In our previous work [15] we evaluated the best parameters for dense sampling and overlapping of the spatio-temporal descriptors: the best results were obtained for cuboids of 40×40 pixels, with 8 frames of depth, a spatial overlap of 50% and no temporal overlap. In these experiments we used the same parameters.

We compare our system with other state-of-the-art approaches, whose results are reported in [12]: MPCCA [20], Adam *et al.* [21], Mehran *et al.* [22] and Mahadevan *et al.* [12]. 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. 2, Fig. 3 and Tab. 1 report the results for anomaly detection in Peds1 and Peds2. Fig.4 and Tab. 1 report results for anomaly localization on Peds1. Our performance at the frame level is close to Social Force for Peds1 but is far superior in the localisation task to all other methods except [12]. Our approach, with the use of multiple scales and contextual queries, obtains the second best result in temporal and spatial anomaly detection after the method proposed in [12], 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 one used in our experiments. The good results in anomaly localization imply that we are not taking advantage of lucky guesses, but that we accurately localise the abnormal behaviours in space and time. Fig. 5 shows a qualitative comparison of anomaly localization of our approach with state-of-the-art off-line approach [12].

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

	Localisation	UCSPed1	UCSPed2	Average
Single scale	27%	34%	32%	33%
Multiscale	28%	32%	31%	32%
Context	**29%**	**31%**	**30%**	**30%**
MDT (Mahadevan *et al.*)	45%	25%	25%	25%
MPPCA (Kim *et al.*)	18%	40%	30%	35%
Social Force (Mehran *et al.*)	21%	31%	42%	37%
Adam *et al.*	24%	38%	42%	40%

Table 1. Summary of quantitative system performance and comparison with state-of-the-art. EER is reported for frame level anomaly detection on Peds1 and Peds2 datasets. Localisation performance is presented as detection rate at EER.

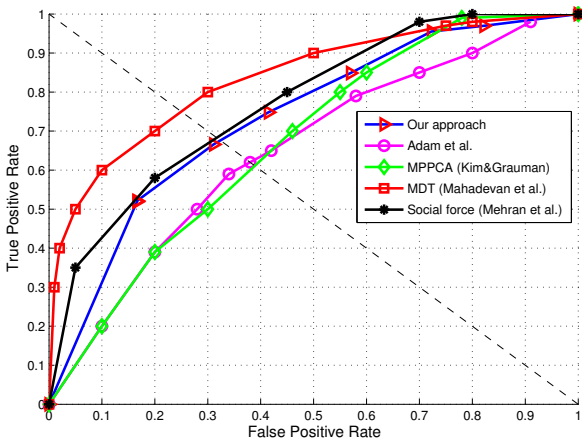


Fig. 2. ROC curve to compare our method with state-of-the-art approaches on the Peds1 dataset. The dashed diagonal is the EER line.

5. CONCLUSIONS

In this paper we have presented a spatio-temporal descriptor that can be used for crowd behavior recognition and non-parametric anomaly detection. The descriptor has been tested



Fig. 1. Sample frames from the UMN dataset; top row shows *normal crowd behaviour* and bottom row shows *crowd panic* frames.

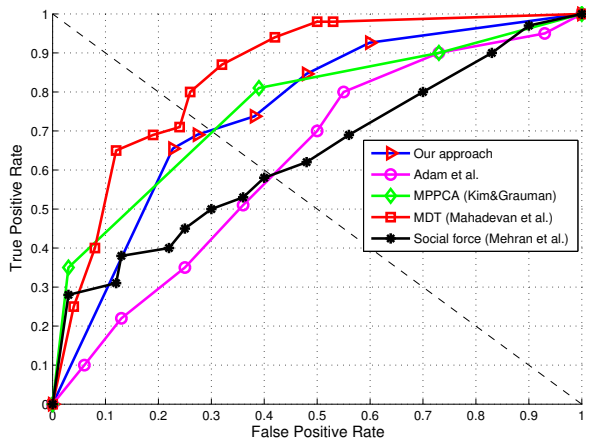


Fig. 3. ROC curve to compare our method with state-of-the-art approaches on the Peds2 dataset. The dashed diagonal is the EER line.

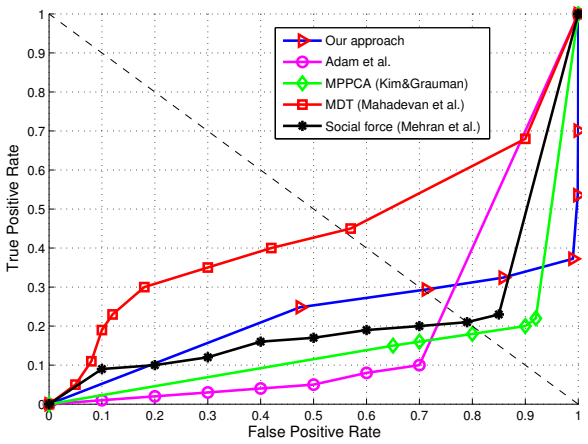


Fig. 4. ROC curve to compare the localisation accuracy of our method with state-of-the-art approaches using Peds1 dataset. The dashed diagonal is the EER line (note that the plot of a random classifier is not diagonal in this case, but close to zero).

in combination with dense and overlapping spatio-temporal volumes and with sparse sampling, to capture the scene dynamics, allowing the detection of different types of anomalies and crowd behaviours. The proposed descriptor is capable of handling challenging crowded scenes that cannot be modeled using trajectories or pure motion statistics (e.g. optical flow).

6. REFERENCES

[1] T.B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. in press, 2010.

[3] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[4] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[5] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[6] C. Schuldt, Ivan Laptev, and Barbara Caputo, "Recog-

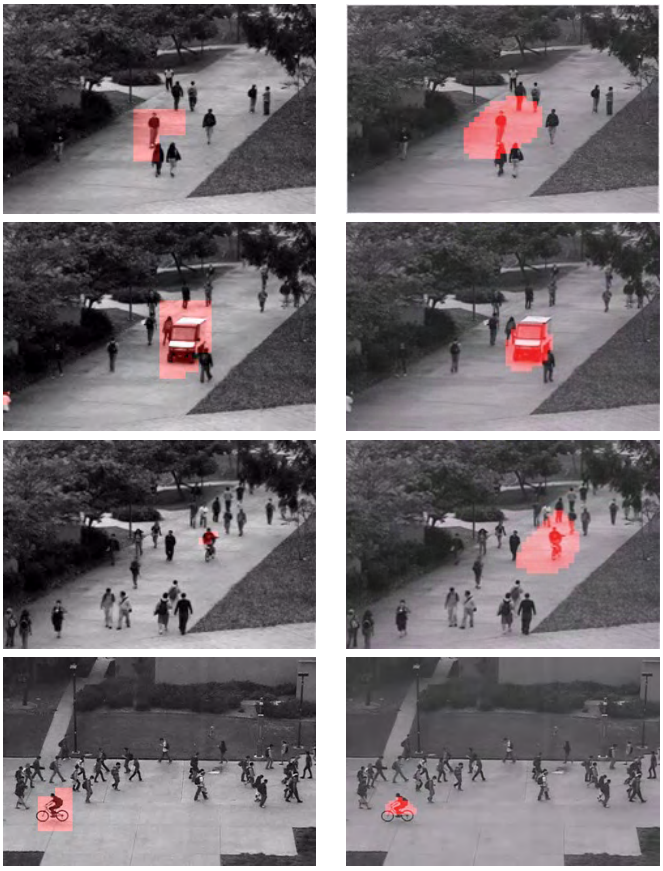


Fig. 5. Anomaly localisation results (top) compared with the best performing method [22] (bottom) on the UCSD dataset.

nizing human actions: a local SVM approach," in *Proc. of ICPR*, 2004.

[7] Piotr Dollar, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of VSPETS*, 2005.

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

[9] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision (IJCV)*, vol. 74, no. 1, pp. 17–31, Aug. 2007.

[10] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1446 –1453.

[11] Fan Jiang, Ying Wu, and A.K. Katsaggelos, "Detecting contextual anomalies of crowd motion in surveillance video," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 1117 –1120.

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

[13] Frederic Jurie and Bill Triggs, "Creating efficient codebooks for visual recognition," in *Proc. of ICCV*, 2005.

[14] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev, and Cordelia Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. of British Machine Vision Conference (BMVC)*, sep 2009, p. 127.

[15] Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo, "Dense spatio-temporal features for non-parametric anomaly detection and localization," in *Proc. of ACM Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, Oct 2010, pp. 27–32.

[16] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-Dimensional SIFT descriptor and its application to action recognition," in *Proc. of ACM Multimedia*, 2007.

[17] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra, "Effective codebooks for human action categorization," in *Proc. of ICCV International Workshop on Video-oriented Object and Event Classification (VOEC)*, Kyoto, Japan, September 2009.

[18] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[19] 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.

[20] 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.

[21] 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*, vol. 30, no. 3, pp. 555–560, March 2008.

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