Real-Time Detection of Unusual Regions in Image Streams

R. Schuster, R. Mörzinger and W. Haas
JOANNEUM RESEARCH
Steyrergasse 17, 8010 Graz, Austria
rene.schuster@joanneum.com

H. Grabner and L. Van Gool
Computer Vision Laboratory, ETH Zurich
Sternwartstrasse 7, 8092 Zurich, Switzerland
grabner@vision.ee.ethz.ch

ABSTRACT

Automatic and real-time identification of unusual incidents is important for event detection and alarm systems. In today's camera surveillance solutions video streams are displayed on-screen for human operators, e.g. in large multi-screen control centers. This in turn requires the attention of operators for unusual events and urgent response.

This paper presents a method for the automatic identification of unusual visual content in video streams real-time. In contrast to explicitly modeling specific unusual events, the proposed approach incrementally learns the usual appearances from the visual source and simultaneously identifies potential unusual image regions in the scene. Experiments demonstrate the general applicability on a variety of large-scale datasets including different scenes from public web cams and from traffic monitoring. To further demonstrate the real-time capabilities of the unusual scene detection we actively control a Pan-Tilt-Zoom camera to get close up views of the unusual incidents.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Tracking*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms

Algorithms, Security

1. INTRODUCTION

Unsupervised and real-time detection of unusualness in video streams is a prerequisite for identifying interesting and critical situations in surveillance scenarios, see Fig. 1. Human operators have limited attention spans as well as cognitive limits in respect to how many different video streams can be observed simultaneously. In large multi-screen control centers it would therefore be beneficial to automatically activate or switch to the camera view(s) of interest. Obviously, for this task real-time performance is needed. A second application is video summarization. As example one might think of having a short video documentation of unusual situations observed over a weekend (e.g. on a construction site).

Figure 1: ***[St. Gallen 22/04/2010 05:24]*** **Real-time detection of unusual incidents in image streams is essential for generating immediate alerts.**

Abnormality detection is a classical task in computer vision and hence many different approaches were proposed over the last two decades, e.g. [5, 6, 7]. Most approaches build and partly also update a model of normality. This model represents typical situations or behaviors observed in the data stream. One definition of normality is if very *similar* content has been observed at least once in the past [8].

Based on this paradigm approaches have been proposed recently which store (cluster) the observed data. By using the concept of meaningful nearest neighbors outliers can be detected and used for scene based abnormality detection [1] or video summarization [3]. The benefit is that no scene-specific, manually tuned similarity threshold for the classifier has to be set. A purely data driven approach has the advantage that it works on different scenes without human intervention. Further it allows to automatically and permanently adapt to changes.

Our approach extends ideas from meaningful nearest neighbors abnormality detection [1] and additionally focuses on the localization of regions with "unusual" content in an image. In fact, instead of modeling the whole scene at once, we propose to use an overlapping grid of small abnormality detectors, similar as in classifier based background models [4]. For updating the local abnormality detectors, we propose an efficient, yet simple, and therefore real-time method.

The remainder of the paper is structured as follows. Sec. 2 describe details of our approach for unusual region detection. Sec. 3 presents experimental results, demonstrating the applicability of the approach on a variety of datasets. Additionally, we show a real-time application for active camera control. Finally, Sec. 4 concludes the paper.

MM'10, October 25–29, 2010, Firenze, Italy.

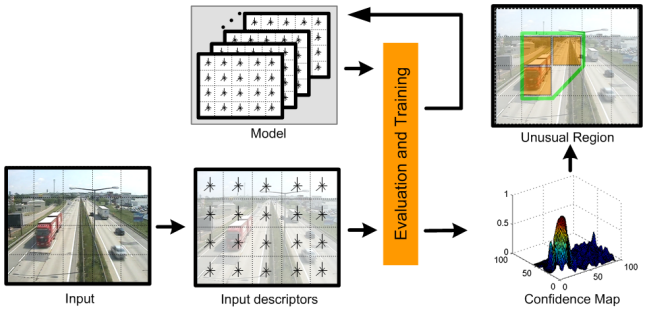


Figure 2: For evaluation descriptors are calculated in the current input and compared to the model of observations. The resulting confidence map is segmented into an unusual region. The training of the model is done in parallel.

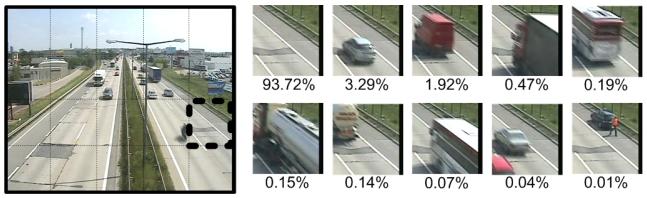


Figure 3: For a single highlighted block the image patches related to the cluster centers of the local model are shown with their observation percentage.

2. UNUSUAL REGION DETECTION

Inspired by the recent work of [1] we use a purely data driven approach and build a model of usualness by storing representative clusters of observed data. The overall process is shown in Fig. 2 and described in the following. For accurate localization the input image I is divided into small partly-overlapping blocks where individual abnormality detection is performed. Each block i is represented by a local descriptor $\mathbf{x}_i \in \mathbb{R}^n$ and compared to the local model of usualness M_i. The local models consist of relevant visual observations from the past, i.e. a number of representative cluster centers are stored for each block as shown in Fig 3. For evaluation new observations are compared to previous representative observations. In order to keep track of a changing environment the local models are incrementally updated. To prevent a currently observed unusual region to become usual in the next frame, the training is done with a time delay. In a post-processing step the responses of the individually evaluated blocks are combined to regions.

Feature Descriptor. As feature representation for the observed visual content we use *Histograms of Oriented Gradients*(HoG) [2]. In order to account for changes in illumination and contrast, the gradient strengths are locally normalized in each block. In fact, we use 9 rectangular cells (i.e. R-HOG) and 9 bin histogram per cell and concatenate them to a 81-dimensional feature descriptor $\mathbf{x}_i$.

Model and Training. Our model of observations consists of a set of local models M_i. For each M_i observed data from the past is stored in maximum N representative cluster centers. Each cluster center $m_{i,j}$ $(j = 1..N)$ is a feature descriptor. An associated observation percentage $f_{i,j}$ which represents the number of observations in the cluster

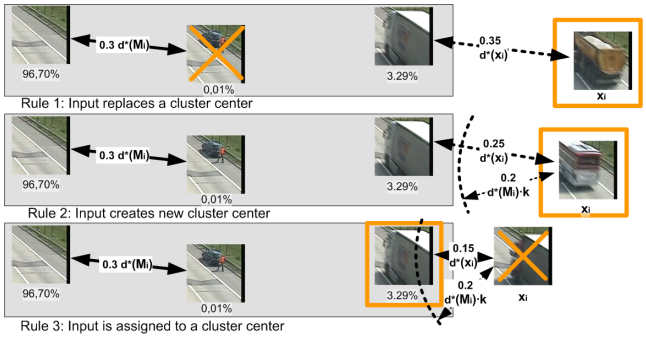


Figure 4: Three different cases during training. An observed input creates a new cluster center, replaces or gets assigned to an existing cluster center.

is stored. A local model is updated with the current input observation $\mathbf{x}_i$. In the proposed approach we augment the concept of meaningful nearest neighbours and the on-line clustering approach of [1]. For achieving real-time performance we consider practical implementation issues and propose the following rules for training and evaluation. Our approach is based on two distances. First, the best match, i.e. the minimum distance from $\mathbf{x}_i$ to all other cluster centers in M_i is computed as

$$d^\star(\mathbf{x}_i) = \min_{m_{i,j} \in M_i} d(\mathbf{x}_i, m_{i,j}). \tag{1}$$

where the distance function $d(\cdot,\cdot)$ compares two descriptors. Second, the closest pair of cluster centers in each M_i is determined, i.e. the pair with the minimum inner model distance

$$d^\star(M_i) = \min_{\substack{m_{i,1}, m_{i,2} \in M_i \\ m_{i,1} \neq m_{i,2}}} d(m_{i,1}, m_{i,2}). \tag{2}$$

As depicted in Fig. 4, we consider three different rules for updating the model in the following sequence:

Rule 1. $\mathbf{x}_i$ replaces a $m_{i,j}$ if $d^\star(\mathbf{x}_i) > d^\star(M_i)$. The $m_{i,j}$ with the lower $f_{i,j}$ of the closest pair is merged into the other cluster of this pair. The $x_{i,j}$ of the merged $m_{i,j}$ remains unchanged whereas the new $f_{i,j}$ is computed by adding the individual observation percentages. This is done in view of achieving an uniform distribution of the cluster centers, while keeping the complexity of M_i constant.

Rule 2. $\mathbf{x}_i$ extends the model by creating a new $m_{i,j}$ if $d^\star(M_i) \geq d^\star(\mathbf{x}_i) \geq k\, d^\star(M_i)$ where $k = \frac{|M_i|}{N}$ is the ratio of the current model size to the maximum size. Consequently only as many cluster centers as needed will be created which depends on the variety of the observations.

Rule 3. $\mathbf{x}_i$ is assigned to a $m_{i,j}$ if $d^\star(\mathbf{x}_i) < k\, d^\star(M_i)$. The $f_{i,j}$ of the nearest neighbour $m_{i,j}$ is increased.

Our approach has a two-fold ability to unlearn: implicitly by merging cluster centers and explicitly by removing cluster centers in case the elapsed time to the last observation exceeds a certain threshold. This cleans up very rare appearances which might degenerate the model.

Evaluation. The evaluation of a current observation $\mathbf{x}_i$ is based on the score for unusualness

$$s(\mathbf{x}_i) = \begin{cases} 1 & \text{if } d^\star(\mathbf{x}_i) > d^\star(M_i) \\ \frac{1}{f_{i,j}} & \text{otherwise} \end{cases} \tag{3}$$

where j denotes the closest cluster center. If the current observation cannot be matched with one of the existing cluster centers the maximum score for unusualness is reported. If a close cluster center is found, the score is indirectly proportional to its observation percentage. The final detection regions are composed of all blocks with score $s(\mathbf{x}_i) > 0.5$.

Real-time performance optimization. Once a model is well trained, the processing time can be reduced without significant loss in accuracy by skipping input frames for training. The fact that the number of cluster centers for each block is flexible takes into account the variance in image content across different image regions. Since only the minimum necessary cluster centers are stored, the effort for comparing and storing is limited. A further optimization concerns evaluation. Here, the computational expensive calculation of the nearest neighbor is limited to cases where the input descriptor has a significant difference to the previous input descriptor, in our experiments set to 1% of the distance space.

3. EXPERIMENTS AND RESULTS

In this section we demonstrate the general applicability of the proposed approach on a variety of datasets and the real-time performance in an active camera tracking framework. As part of a video track, this paper comes with a 4:30 minutes long video - please watch it for more illustrative results.

3.1 Unsupervised Unusual Region Detection

Data. For our experiments we apply the detector to several different realistic large-scale datasets including videos with 25 fps frame rate: *Street*(4 days), *Highway*(20 h), *Car plant*(16 h), low frame rate: *Brest*[1](4 days), *Billboard*[2](12 h) and the previously published *Time Square* [1](50 days).

Parameter setting. As parameters we used a grid of semi-overlapping blocks with size 48×48 pixels and a maximum number of 120 cluster centers. As distance function the computationally efficient normalized cross-correlation is used.

Computational Performance. When analyzing every single frame of a video stream in SD resolution the average processing time is real-time (20 fps) including video acquisition on a 2.66 Ghz Intel Dual Core. The memory consumption of a model with 6160 blocks and an average number of 40 cluster centers is around 130 MB.

Evaluation and Results. A manual assessment of unusualness for quantitative evaluation is highly subjective and may not be significant since it clearly depends on the definition of task and the evaluator. For this reason and due to lack of existing benchmark data (ground truth), we limit the evaluation to a qualitative investigation and interpretation of the detection results. Fig. 5 shows sample results from the dataset. Unusual regions indicated by polygons are detected even in cluttered scenes. The fact that the approach produces plausible results on multiple extremely different datasets demonstrates its general applicability.

A comparison of the results obtained in the *Times Square* dataset with results from a related work [1] shows that we are not only able to detect similar unusual appearances but we are also able to localize them quite accurately. Limitations of the approach in terms of 'interestingness' are depicted in the sample frames in the bottom row (n-p) of Fig. 5, where appearances that seem usual for the *Times Square* are wrongly detected. The main reason for undesired detections are changing shades and configurations of salient edges. These detections do not indicate interesting appearances, yet they have actually never be seen before on that location. A typical output during initial model training is shown in (m). The duration depends on the amount of activity in the scene, for *Times Square* the model learned the normality within 2000 frames.

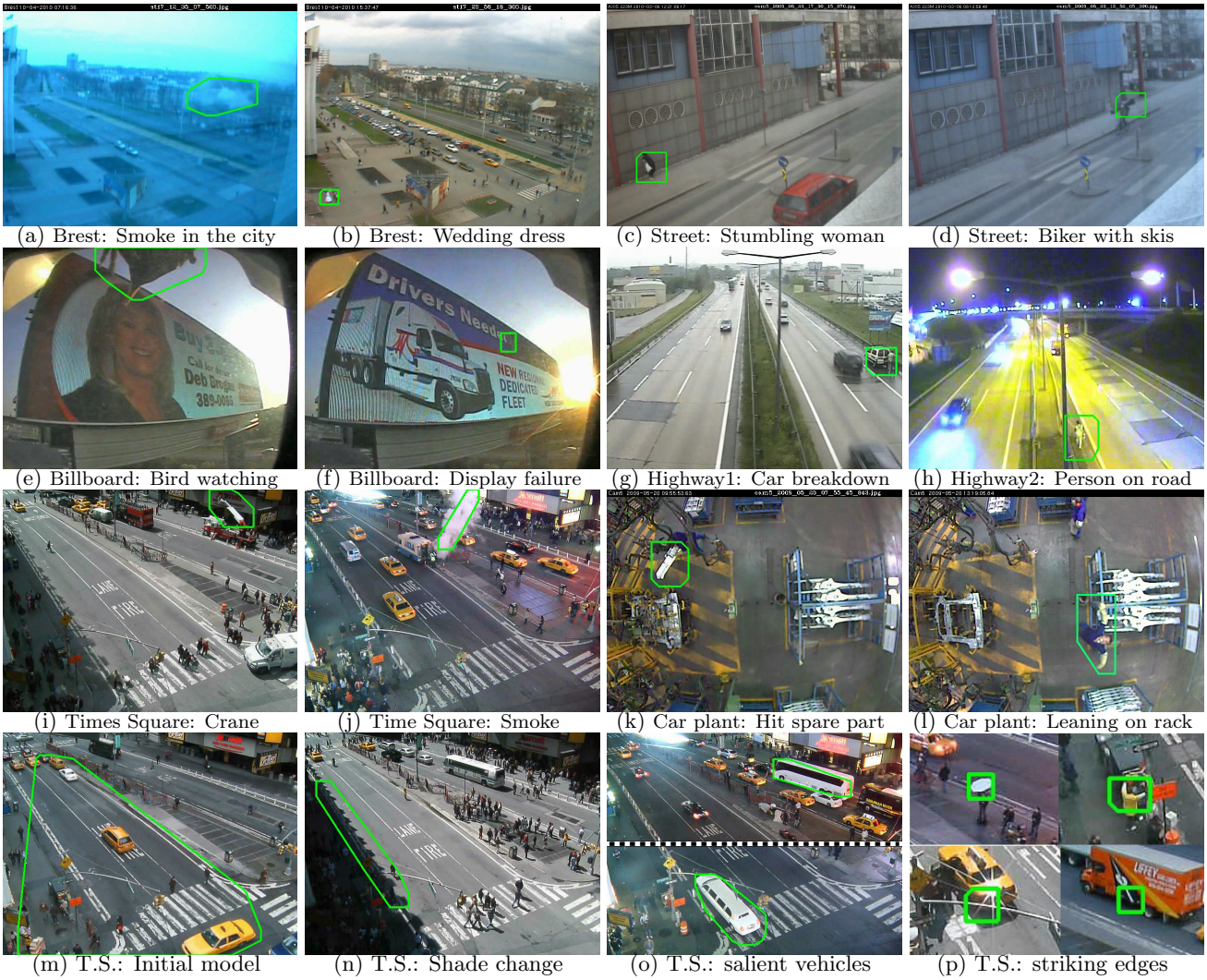
[1] `http://camera.brest.by/view/index.shtml`
[2] `http://216.203.115.186:5001/view/index.shtml`

Figure 6: Unusual regions identified in a low resolution static camera (left) used to steer an active camera automatically for making close-ups of the scene (right).

3.2 Active Camera Tracking by Unusualness

Practical surveillance applications demand real-time processing performance as an important factor. For demonstrating the real-time processing capabilities, our approach is used within an active camera tracking application. The overall system consists of a static camera and an active camera with (partly) overlapping fields of view. The goal is to immediately steer the active camera using pan-tilt-zoom operations to the place of interest if an unusual situation is observed in the view of the static camera.

Camera Network. For identifying the single focus of attention, all blocks evaluated as unusual are segmented into a single convex hull. Calibration information is used for transforming the position of a detected unusual region in the static camera to absolute pan and tilt angles. Zoom is adapted in regard to the size of the unusual region. In case observations in the static camera are identified as usual, the active camera focuses on the area where unusualness was most frequently detected in the past. Fig. 6 shows exemplary results of close-ups of unusual scenes captured from the active camera (Axis 213 PTZ). The unusual appearances of a person sitting on the sidewalk and off-limits crossing is detected in the static view and the active view is immediately steered to the detected region using pan-tilt operations. This shows that unusual regions can be detected fast enough, i.e. in real-time so that it is possible to control active cameras accordingly.



(a) Brest: Smoke in the city (b) Brest: Wedding dress (c) Street: Stumbling woman (d) Street: Biker with skis

(e) Billboard: Bird watching (f) Billboard: Display failure (g) Highway1: Car breakdown (h) Highway2: Person on road

(i) Times Square: Crane (j) Time Square: Smoke (k) Car plant: Hit spare part (l) Car plant: Leaning on rack

(m) T.S.: Initial model (n) T.S.: Shade change (o) T.S.: salient vehicles (p) T.S.: striking edges

Figure 5: Exemplary results of unusual regions detected in 6 different datasets. Common appearances such as moving cars and pedestrians are classified as usual. (Best viewed magnified and in color.) ***Please also watch the video contribution submitted with this paper.***

4. CONCLUSION

A fully automatic approach for the detection and localization of unusual regions in image streams is presented. The unusualness detector classifies regions as unusual if a similar visual appearance has not or rarely been observed in the past. Promising results on 6 very different datasets and real-time active camera tracking triggered by unusualness are presented. In the future we plan to extend the feature representation with motion descriptors for recognizing unusual object movement.

5. ACKNOWLEDGMENTS

This research was supported by the European Community's 7^{th} Framework Programme under grant agreement no FP7-ICT-216465 SCOVIS. The authors further would like to thank Georg Thallinger and Helmut Neuschmied for their support and feedback.

6. REFERENCES

[1] M. Breitenstein, H. Grabner, and L. V. Gool. Hunting nessie: Real time abnormality detection from webcams. In *Proc. IEEE WS on Visual Surveillance*, 2009.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[3] L. V. Gool, M. Breitenstein, S. Gammeter, H. Grabner, and T. Quack. Mining from large image sets. In *Proc. ACM Int. Conf. on Image and Video Retrieval*, 2009.

[4] H. Grabner and H. Bischof. On-line boosting and vision. In *Proc. CVPR*, volume 1, pages 260–267, 2006.

[5] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *BMVC*, 1996.

[6] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 2000.

[7] X. Wang, K. Ma, G. Ng, and W. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *Proc. CVPR*, 2008.

[8] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *Proc. CVPR*, 2004.