Video Parsing for Abnormality Detection

Borislav Antić and Björn Ommer
Interdisciplinary Center for Scientific Computing, University of Heidelberg, Germany
{borislav.antic,bommer}@iwr.uni-heidelberg.de

Abstract

Detecting abnormalities in video is a challenging problem since the class of all irregular objects and behaviors is infinite and thus no (or by far not enough) abnormal training samples are available. Consequently, a standard setting is to find abnormalities without actually knowing what they are because we have not been shown abnormal examples during training. However, although the training data does not define what an abnormality looks like, the main paradigm in this field is to directly search for individual abnormal local patches or image regions independent of another.

To address this problem we parse video frames by establishing a set of hypotheses that jointly explain all the foreground while, at same time, trying to find normal training samples that explain the hypotheses. Consequently, we can avoid a direct detection of abnormalities. They are discovered indirectly as those hypotheses which are needed for covering the foreground without finding an explanation by normal samples for themselves. We present a probabilistic model that localizes abnormalities using statistical inference. On the challenging dataset of [15] it outperforms the state-of-the-art by 7% *to achieve a frame-based abnormality classification performance of* 91% *and the localization performance improves by* 32% *to* 76%.

1. Introduction

Object and behavior recognition in videos of crowded scenes is one of the primary challenges of computer vision. The problem becomes even more challenging when unusual objects or suspicious behaviors are to be detected. Finding such abnormalities in videos is crucial for applications ranging from automatic quality control to visual surveillance. However, while detecting normal objects is already difficult due to a large within-class variability, abnormality detection poses the additional problem that there exist infinitely many ways for an object to appear in unusual context (irregular object instance) or to behave abnormally (unusual activity). Thus it is simply impossible to learn a model for everything that is abnormal or irregular. Consequently, recent work on abnormality detection [15] has established benchmark datasets where the training data contains only normal visual patterns and a discriminative approach cannot be employed to directly localize irregularities. So the question is: how can we find an abnormality, if we *do not know what to search for*? Despite this fundamental problem, the main paradigm to abnormality detection is currently to classify each local image patch individually, e.g., [4, 26] or to detect abnormal image regions separately [27]. However, deciding locally and independently about the abnormality of each individual image region is an ill-posed problem.

We can avoid this issue by abandoning the standard approach of object detection which aims at finding each object in a scene that is independent of the others. Typically, abnormality detection is based on videos from a stationary camera (e.g. surveillance videos) where powerful background subtraction algorithms [25] can provide foreground/background segregation. We then need to find a set of normal object hypotheses that together explain all the foreground pixels. Therefore, object hypotheses have to be distributed over the scene so that all the foreground is covered while extending into the background as little as possible. To solve this problem all hypotheses have to be jointly placed within the scene and their spatial configuration has to be determined, i.e., it has to be decided what instance from the set of positive training samples provides the best fit. All the object hypotheses which are needed to explain the foreground but which themselves cannot be explained by an instance from the set of normal training samples are then abnormal objects. By parsing the scene and *jointly* inferring all required object hypotheses we can *indirectly* discover objects which must have been present in the scene without actually knowing what to look for.

To make this problem feasible we follow a two-stage approach. First, a shortlist of hypotheses is computed that has a low-false negative and high-false positive rate, i.e. a superset of all hypotheses that might eventually be needed is computed. Background subtraction rules out all hypotheses in the background and a discriminative background classifier is used to retain only those hypotheses that are very

In Proceedings of the 13th International Conference on Computer Vision (ICCV'11), ©2011 IEEE

unlikely to be background. Based on the shortlist of candidate hypotheses the problem of video parsing and explaining foreground with object hypotheses becomes a discrete optimization problem. We have to find a subset of the original hypotheses that are needed and sufficient for covering all the foreground. Therefore, the presence of hypotheses and their correspondence to the exemplars of normal objects from the training data have to be jointly inferred for all hypotheses. Correspondences are established between spatiotemporal patterns to capture not just the appearance of objects but also the change thereof which is caused by the behavior of objects. Our probabilistic approach not only labels objects as being irregular but also infers a per-pixel abnormality probability which allows to segment abnormal objects from the scene without having seen training data for them.

The experimental setup that we follow is that of abnormality detection in highly crowded scenes based on the novel dataset of Mahadevan *et al*. [15]. The challenging videos feature walkways crowded by walking pedestrians. Abnormality is not staged but consists of events that occur naturally such as unusual objects (e.g. cars on walkways) or unusual behavior (e.g. people cycling across walkways). Videos are of low resolution (pedestrians have a height between 10 and 30 pixels) and objects are heavily overlapping so that learning models of visual patterns becomes a challenging problem. Moreover, the training data features only normal patterns with large within-class variability, whereas the test set consists of normal and abnormal instances. To extend the future utility of this benchmark database we have completed the pixel-wise ground-truth annotation of the test set which was previously only available for a small subset of all frames. In the experimental evaluation, our approach has significantly outperformed all current methods for abnormality detection on this dataset.

2. Related work

We mainly focus on the work related to abnormality detection in videos since a comprehensive overview over the general topics of object recognition, tracking, and action recognition is beyond the scope of this paper and can be found in Ommer *et al*. [18]. The main paradigm for abnormality detection in videos is to extract semi-local features [14, 21] and to learn a model on the normal samples from the training data. Abnormality is estimated by measuring how bad the model fits. The degree of supervision in these models varies to a great extent. Some of the approaches [6, 22] are based on a set of constraints that are introduced to specify the normality, whereas the methods [27, 3, 1, 24] are unsupervised approaches which directly determine normal patterns. The approach by Adam *et al*. [1] can be deemed local since the attention is directed to individual activities occurring in a local area. While this approach provides good results when it comes to implementation and efficiency, its performance suffers from the incapability of the model to incorporate temporal aspects of relationships among activities. Xiang and Gong [26] proposed a method that automatically recognizes behavior and detects abnormalities without applying any manual labeling. Zhong *et al*. [27] detect objects by thresholding a motion filter and they propose an unsupervised method that integrates the prototypical image features and classifies a group of behavior patterns either as normal or abnormal. Kim and Grauman [10] proposed a method to detect abnormalities in a video sequence based on a space-time Markov random field model. This model dynamically adapts to abnormal activities that consists of unpredictable variations. Some of the current methods for the detection of abnormal behavioral patterns are based on unsupervised one-class learning approaches. These namely include topic models such as in [24, 8] or Markov random field models [10]. Other methods utilize supervised approaches for classifying events and patterns such as [20, 9]. Mahadevan *et al*. [15] present a method for unusual behavioral pattern detection in crowded scenes that is based on mixtures of dynamic textures. Their approach also includes jointly performed modeling of the dynamics and appearance of a scene as well as detection of temporal abnormalities which are represented as low-probability occurrences and unusual spatial activities which are dealt with using the discriminative saliency property. The provided dataset contains low-resolution video sequences of crowds with occlusion.

We propose to parse video frames by jointly inferring all objects that are needed to explain the foreground that has been observed. Instead of detecting abnormal image regions independently as in current approaches, abnormalities are discovered indirectly after establishing a complete interpretation of the foreground, as a subset of all hypotheses that are necessary and sufficient for explaining all pixels of the foreground map. A similar problem appears in object tracking where the goal is to link object detections into possible tracks and then find a subset of tracks that provides a mutually consistent covering. That can be achieved by solving a constrained optimization problem [19]. Previous approaches related to scene parsing differ in that a parametric scene [23, 2] or object model [11, 7] or a non-parametric exemplar-based representation for objects [13, 16] can be constructed. In contrast to these methods, we are not provided any training samples of the abnormalities we are searching for but we can leverage a foreground/background segregation.

3. Abnormality Detection by Jointly Explaining All Scene Constituents

Abnormalities cannot be searched for directly, since the class of abnormalities is infinitely large and so there are no

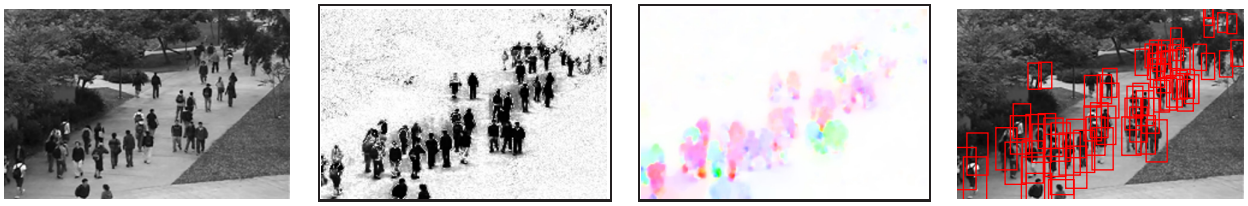


Figure 1. Initialization of the video parsing method. From left to right are: source frame, foreground probability map, optical flow map and the set of hypotheses of the shortlist.

(or not enough) training samples available. However, in case of a stationary camera the foreground/background segregation becomes feasible due to background subtraction. The foreground mask renders it then possible to turn the abnormality detection problem into a task of video parsing. The goal is thus to explain all the foreground using object hypotheses and to explain each hypothesis using a sample from the set of normal training examples. The underlying statistical inference problem has to be tackled jointly for all hypotheses, since hypotheses can explain each other away. Abnormalities are then those hypotheses that are required to explain the foreground but which themselves cannot be explained by normal training samples.

3.1. Initialization

To parse a novel frame in a video, several pieces of information have to be gathered. In the first place, background subtraction is performed. This is possible because we assume the stationary camera model for our videos. Background subtraction facilitates further processing by ignoring irrelevant parts of the video. For background subtraction, we follow [25] and assume that each video frame can be decomposed into a background image corrupted by sparse foreground pixels. The matrix of successive video frames $U = [\tilde{u}_{t-\tau+1} \cdots \tilde{u}_t]$ is thus a sum of a low-rank matrix $B = [\tilde{b}_{t-\tau+1} \cdots \tilde{b}_t]$ that corresponds to the background and a sparse matrix $F = [\tilde{f}_{t-\tau+1} \cdots \tilde{f}_t]$ which corresponds to the foreground. This decomposition can be achieved efficiently by convex optimization [25]. The foreground label $f_j \in \{0, 1\}$ for each pixel j is selected according to the foreground probability computed from the raw foreground pixel value $\tilde{f}_j$,

$$P(f_j = 1) := 1 - \exp\Big(-\frac{\|\tilde{f}_j\|}{\sigma_f}\Big). \tag{1}$$

Secondly, the optical flow vectors v_j are computed by the method [12]. Velocity of an object hypothesis i is then calculated as a weighted average of the optical flow vectors over the support $\mathcal{S}_i$ of the hypothesis i

$$v_i = \frac{\sum_{j \in \mathcal{S}_i} P(f_j = 1) \cdot v_j}{\sum_{j \in \mathcal{S}_i} P(f_j = 1)} \tag{2}$$

To initialize the subsequent parsing and abnormality detection, a shortlist of candidate object hypotheses is computed. From a large number of object hypotheses that could be established in a video frame most hypotheses are not compatible with the foreground mask as they would be located in the background. Now we can efficiently evaluate an appearance based classifier on candidate object hypotheses in the foreground to obtain a shortlist of relevant hypotheses. Since the training data does not contain abnormal instances but only the background and normal foreground, it is important to note that this is basically an inverted background detector, i.e., a discriminative SVM classifier that is trained to distinguish the background from anything else that deviates from it. A vector of spatiotemporal derivatives $d_i = [\frac{\partial \tilde{u}_j}{\partial x}, \frac{\partial \tilde{u}_j}{\partial y}, \frac{\partial \tilde{u}_j}{\partial t}]_{j \in \mathcal{S}_i}$ is used as a feature vector in the SVM classifier. The features capture both the appearance (spatial patterns) and behavior (temporal patterns) in a video domain that is crucial for good performance in abnormality detection. The SVM classifier uses a linear kernel and produces a probabilistic output [5] as an estimate of the probability of the background class $P(o_i = 0|d_i)$. The classifier is trained in a batch mode on samples from training videos.

The resulting shortlist of object hypotheses is set to have a high recall and low precision. This opportunistic prefiltering retains a reasonable number of hypotheses (on the order of 10 to 10^2) without losing any relevant ones. However, all of these hypotheses have been found independently of each other. Therefore, there will be spurious hypotheses that can be explained away by others. Moreover, abnormalities can only be discovered once the foreground has been explained by a set of mutually compatible object hypotheses. Abnormal hypotheses are then the ones which cannot be described by the object model that has been learned during training, but which are nevertheless needed to explain the foreground that cannot be explained by other hypotheses. The initialization stage is illustrated in Fig. 1.

3.2. Model Formulation

Given the initialization, the task of scene parsing is as follows. Select a subset of the initial set of hypotheses that explains all the foreground and explain each object hy-

pothesis using the object model (e.g., which training samples correspond to a particular query hypothesis) that has been learned during training. The activation/deactivation of candidate hypotheses and their explanation with the object model have to be solved jointly for all hypotheses since they are mutually competing. The main inference process that parsing is based upon is that of *explaining away* as we will see later. Object hypotheses are necessary for explaining the foreground if they cannot be explained away by others. If such a necessary hypothesis fits to the object model that has been learned from the training videos that only contain normal patterns then this is a normal instance, otherwise we have found an abnormality. Since the model is inherently probabilistic a probability of abnormality is provided. The graphical model of our video parsing approach is shown in Fig. 2.

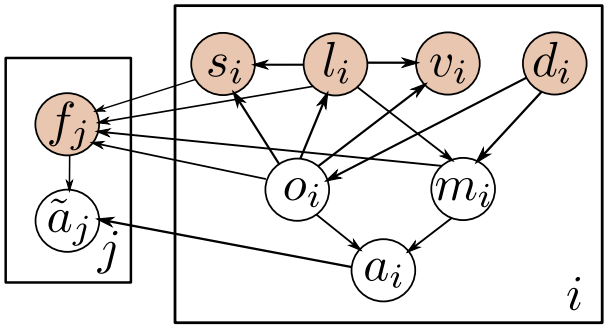


Figure 2. Graphical model of the proposed video parsing method for abnormality detection.

The initialization provides a set of object hypotheses, where each hypothesis has a location $l_i \in \mathbb{R}^2$, a scale $s_i \in \mathbb{R}$, an overall appearance descriptor $d_i \in \mathcal{D}$ that lives in feature space $\mathcal{D}$, and a velocity $v_i \in \mathbb{R}^2$. After the initialization, all object hypotheses are assumed to be required, i.e. the indicator variable $o_i \in \{0, 1\}$ is initialized as $o_i = 1$. Our goal is now to find a subset of all hypotheses that is necessary and sufficient for explaining all pixels of the foreground mask $f_j \in \{0, 1\}$. Moreover, we aim at explaining each hypothesis based on a normal object sample from the training data. Thus, for each hypothesis i the best exemplar $m_i \in \mathcal{M}$ from the training data $\mathcal{M}$ is sought (see Fig.3). For abnormal objects all exemplars will obviously have high matching costs. Consequently, the probability that sample m_i is matched to the i-th hypothesis in a query frame depends on how similar they are in appearance, $\Delta(d_i, d_{m_i})$. Δ is the distance in the feature space $\mathcal{D}$. Moreover, each visual pattern has a particular probability to occur at a specific location, e.g. cars are more likely to drive on roads than on sidewalks, whereas pedestrians are more likely to walk on sidewalks. The probability that the training sample m_i will be matched to the hypothesis i is given by

$$P(m_i|l_i, d_i) \propto P(m_i|l_i) \cdot P(m_i|d_i) \tag{3}$$
$$\propto \frac{\exp\Big(-\beta_l \cdot \|l_i - l_{m_i}\|\Big)}{Z(l_i)} \times \frac{\exp\Big(-\beta_d \cdot \Delta(d_i, d_{m_i})\Big)}{Z(d_i)},$$

where $Z(\cdot)$ is the partition function. The probability of hypothesis i being an actual object (and not a spurious detection) depends on the observed properties of the hypothesis (descriptor d_i, location l_i, scale s_i, and velocity v_i) and is given by

$$P(o_i = 1|d_i, l_i, s_i, v_i) \propto P(o_i = 1|d_i) \tag{4}$$
$$\times\, p(l_i|o_i = 1) \cdot p(s_i|o_i = 1, l_i) \cdot p(v_i|o_i = 1, l_i)$$

Here $P(o_i|d_i)$ is the SVM appearance classifier from Sec. 3.1, while $p(l_i|o_i)$, $p(s_i|o_i, l_i)$ and $p(v_i|o_i, l_i)$, for $o_i = 1$, are nonparametric models of location, scale and velocity of normal objects in the image. Otherwise, if $o_i = 0$, location, scale and velocity have uniform distribution.

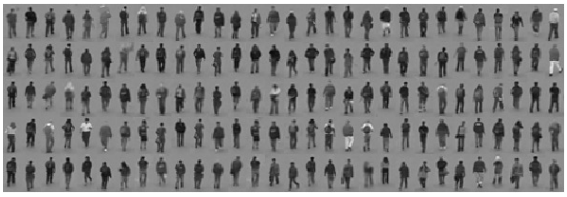


Figure 3. A subset of exemplars obtained from the training data.

Finally, we need to estimate the foreground probability of a pixel j. This probability depends on all hypotheses i that cover the pixel. Let $\mathcal{S}_i$ be the support of the ith hypothesis, i.e., the set of all pixels that are covered by it. Then $\{i : j \in \mathcal{S}_i\}$ is the set of all hypotheses that contain pixel j. We assume that the probability the pixel is background, given all the hypotheses, can be expressed as a product of probabilities of the pixel being background, given each hypothesis alone. We also account for a possibility that the pixel is foreground even if all hypotheses claim it is background. This is modeled by the leak probability p_0. The foreground probability is therefore given by

$$P(f_j = 1|\{o_i, m_i, l_i, s_i\}_{i:j\in\mathcal{S}_i}) \tag{5}$$
$$= (1 - p_0)\Big[1 - (1 - p_0) \prod_{i:j\in\mathcal{S}_i} \big(1 - P(f_j = 1|o_i, m_i, l_i, s_i)\big)\Big]$$

The first factor $(1 - p_0)$ allows the pixel to be background (i.e. $P(f_j = 0|\{o_i, m_i, l_i, s_i\}_{i:j\in\mathcal{S}_i}) = p_0$) even if there is a hypothesis that asserts the pixel is foreground (i.e. $P(f_j = 1|o_i, m_i, l_i, s_i) = 1$ for some i). The second factor $(1 - p_0)$ allows the pixel to be foreground (i.e. $P(f_j = 1|\{o_i, m_i, l_i, s_i\}_{i:j\in\mathcal{S}_i}) \approx p_0$) even if all hypotheses assert

that the pixel is background (i.e. $P(f_j = 0|o_i, m_i, l_i, s_i) = 1$ for all i).

To obtain the foreground probability of a pixel based on a training sample m_i, the foreground probability map $P(f^{m_i} = 1)$ of the training sample is pasted into the query frame at the location l_i. Thus we have to shift and scale it from the reference frame of the training sample into that of the current frame and obtain

$$P(f_j = 1|o_i, m_i, l_i, s_i) = o_i \cdot \mathbf{1}(j \in \mathcal{S}_i) \cdot P(f^{m_i}_{s_i^{-1}(l_j - l_i)} = 1) \tag{6}$$

Here $\mathbf{1}(\cdot)$ is the indicator function and if $o_i = 0$ or $j \notin \mathcal{S}_i$ then the hypothesis i does not explain the pixel j.

3.3. Inference by Foreground Parsing

The goal is now to estimate which of the hypotheses are actually needed to explain the foreground and to find a matching training sample for each hypothesis. For abnormal hypotheses Eq. 4 will yield low probabilities. If foreground $f_j = 1$ is observed and asserted by the hypothesis i, and no other hypothesis can be found that could explain the presence of the foreground at that pixel, then the probability of the hypothesis i increases. This statistical inference is also called *explaining away* in the literature, since for an observed variable f_j different hypotheses i that share the same pixel j become statistically dependent so that the absence of one hypothesis can dictate the presence of another.

To infer the unknown variables o_i and m_i, we have to find the joint configuration $\{\hat{o}_i, \hat{m}_i\}_i$ that maximizes the posterior probability

$$\begin{aligned}\{\hat{o}_i, \hat{m}_i\}_i &= \operatorname*{argmax}_{\{o_i, m_i\}_i} P(\{o_i, m_i\}_i | \{d_i, l_i, s_i, v_i\}_i, \{f_j\}_j) \\ &= \operatorname*{argmax}_{\{o_i, m_i\}_i} \prod_i \Big(P(o_i|d_i, l_i, s_i, v_i) \cdot P(m_i|d_i, l_i) \Big) \\ &\quad \times \prod_j P(f_j | \{o_i, m_i, l_i, s_i\}_{i:j \in \mathcal{S}_i}) \end{aligned} \tag{7}$$

To solve the given problem we follow an alternating optimization approach. In each iteration we fix the parameters of all but one hypothesis i and then maximize over its parameters (o_i, m_i). Each iteration is actually a search in the space $\{0, 1\} \times \mathcal{M}$ where the variables (o_i, m_i) live

$$\begin{aligned}&\operatorname*{argmax}_{o_i, m_i} P\big(o_i, m_i | \{d_{i\prime}, l_{i\prime}, s_{i\prime}, v_{i\prime}\}_{i\prime}, \{f_j\}_j, \{o_{i\prime}, m_{i\prime}\}_{i\prime \neq i}\big) \\ &= \operatorname*{argmax}_{o_i \in \{0,1\}, m_i \in \mathcal{M}} P(o_i|d_i, l_i, s_i, v_i) \cdot P(m_i|d_i, l_i) \\ &\times \prod_{j \in \mathcal{S}_i} P(f_j | \{o_{i\prime}, m_{i\prime}, l_{i\prime}, s_{i\prime}\}_{i\prime : j \in \mathcal{S}_{i\prime}}). \end{aligned} \tag{8}$$

Typically, only a few rounds of iterations are needed to converge to a locally optimal solution.

3.4. Detecting Abnormalities

Finally, the ith hypothesis is an abnormality, $a_i = 1$, if this hypothesis is necessary to explain the observed foreground, $\hat{o}_i = 1$, but it has a low probability according to Eq. 4 and if no matching training sample can be found, i.e., the best estimate $\hat{m}_i$ for a matching sample (obtained from Eq. 7) is unlikely to explain this hypothesis,

$$\begin{aligned}&P(a_i = 1|o_i = 1, m_i = \hat{m}_i) \\ &\propto P(o_i \neq 1|d_i, l_i, s_i, v_i) \cdot P(m_i \neq \hat{m}_i|d_i, l_i)\end{aligned} \tag{9}$$

Similarly, pixel j is part of an abnormal object, $\tilde{a}_j = 1$, if it is in the foreground, $f_j = 1$, and if any of the hypotheses that extend over this pixel, $\{i : j \in \mathcal{S}_i\}$, is abnormal,

$$\begin{aligned}&P(\tilde{a}_j = 1|f_j = 1, \{a_i\}_{i:j \in \mathcal{S}_i}) \\ &\propto P(f_j = 1) \cdot \max_{i:j \in \mathcal{S}_i} P(a_i = 1|o_i, m_i)\end{aligned} \tag{10}$$

4. Experimental Evaluation

We evaluate our approach on the challenging abnormality datasets *Ped1* and *Ped2* that have been recently proposed by Mahadevan *et al*. [15]. The video sequences feature a pedestrian walkway acquired by a stationary camera with low resolution (pedestrians have a height between 10 and 30 pixels). The crowd density varies and there are numerous sequences that are very crowded and with severe occlusions. Abnormalities are not staged but are naturally occurring events such as i) objects that are unusual in the present surroundings (e.g. cars on walkways) or ii) objects that behave irregularly such as people cycling across walkways or walking over the surrounding grass. Other abnormalities include skaters, small carts, and wheelchairs. The training data contains only normal objects and actions, so that no model for abnormalities can be learned. We concentrate mainly on the *Ped1* dataset as a larger, more difficult one of the two benchmark sets, which also features some perspective distortion and a scale variability of more than one octave. The standard experimental protocol uses 34 clips for training and 36 for testing in the *Ped1* dataset, and 16 clips for training and 14 for testing in the *Ped2* dataset.

There exist two evaluation methodologies: abnormality detection on a frame level and pixel-accurate detection. In the first, a frame is labeled as abnormal if it contains one or more abnormalities. Repeating the detection for multiple thresholds yields then an ROC curve. In the second experiment abnormality detections are compared to pixel level ground-truth masks. To obtain an ROC curve, Mahadevan *et al*. consider frames as abnormal if at least 40% of all truly abnormal pixels are detected. A shortcoming of the current datasets is that pixel-wise ground-truth is only available for a small number of test sequences. To improve the utility of the benchmark datasets, we have completed the pixel-wise

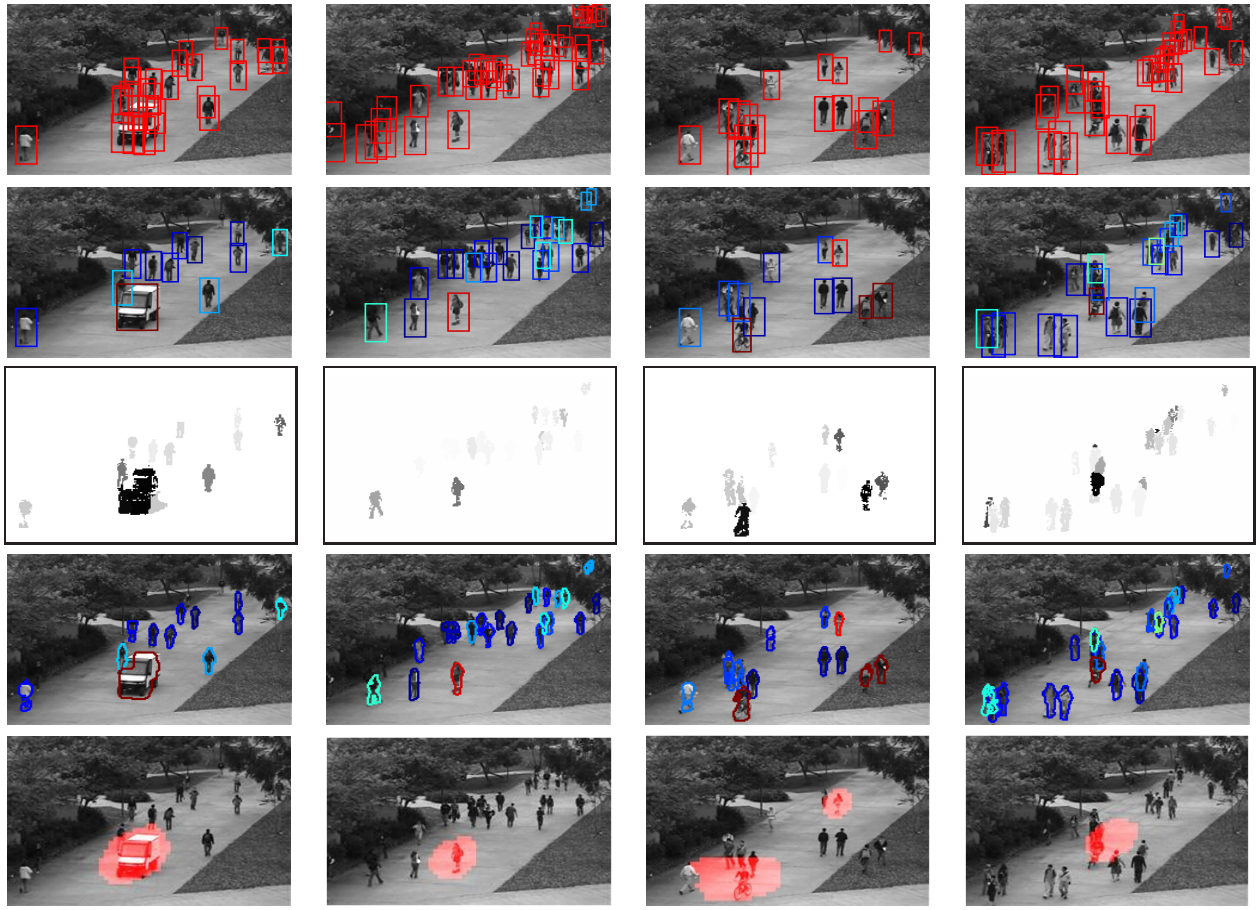


Figure 4. Columns show detection results on different frames. Row i) our initial shortlist, row ii) hypotheses and abnormality probability a_i, row iii) per-pixel probability $\widetilde{a}_j$, row iv) best fitting model m_i, row v) result by [15]. Best viewed in color.

ground-truth annotation for all of the test videos in the *Ped1* dataset and will make it publicly available at the address `http://hci.iwr.uni-heidelberg.de/COMPVIS/research/abnormality`.

4.1. Comparing with the State-of-the-Art

We compare our approach with the state-of-the-art abnormality detection methods on the *Ped1* and *Ped2* benchmark datasets. The methods include the mixture of dynamic textures [15], the social force model [17], the mixture of optical flow [10], the optical flow monitoring method [1], and a combination of [17] and [10] that was investigated in [15].

In all of the experiments our approach significantly outperforms all the other approaches. Our per-frame labeling on the *Ped1* dataset (Fig. 6) achieves an EER of 18%, which is an improvement of 7% over [15], and an improvement of 22% over [10]. We also compare the area under the ROC curve (AUC), which is a more robust measure, as it does not depend on only a single spot on the curve. We achieve an AUC of 91% compared to 84% of [15]. Per-frame labeling on the *Ped2* dataset (Fig. 7a) yields an EER of 14%, which is an improvement of 11% over [15], and it also results in an AUC of 92% compared to 85% of [15]. Nevertheless, our current MATLAB implementation is approximately twice as fast in the prediction phase (5-10 secs per frame) as the currently best performing approach [15].

In order to estimate the abnormality of pixels, we follow a direct probabilistic approach where the variables $\widetilde{a}_j$ are obtained directly by statistical inference. The abnormality masks are then compared to the pixel-level ground truth masks as in [15]. Our approach improves the AUC in this experiment (Fig. 7b) by 32% achieving 76% average performance compared to 44% by the currently best approach [15]. In that paper, the detection performance at the point of equal error was also reported where we achieve a 23% gain yielding a detection rate of 68% compared to 45% by [15]. Note that in the previous experiment [15] have compared error rates while they are measuring detection rates. This is why we report the standard and more robust AUC in all cases. To enhance the future utility of this dataset, Fig. 7c

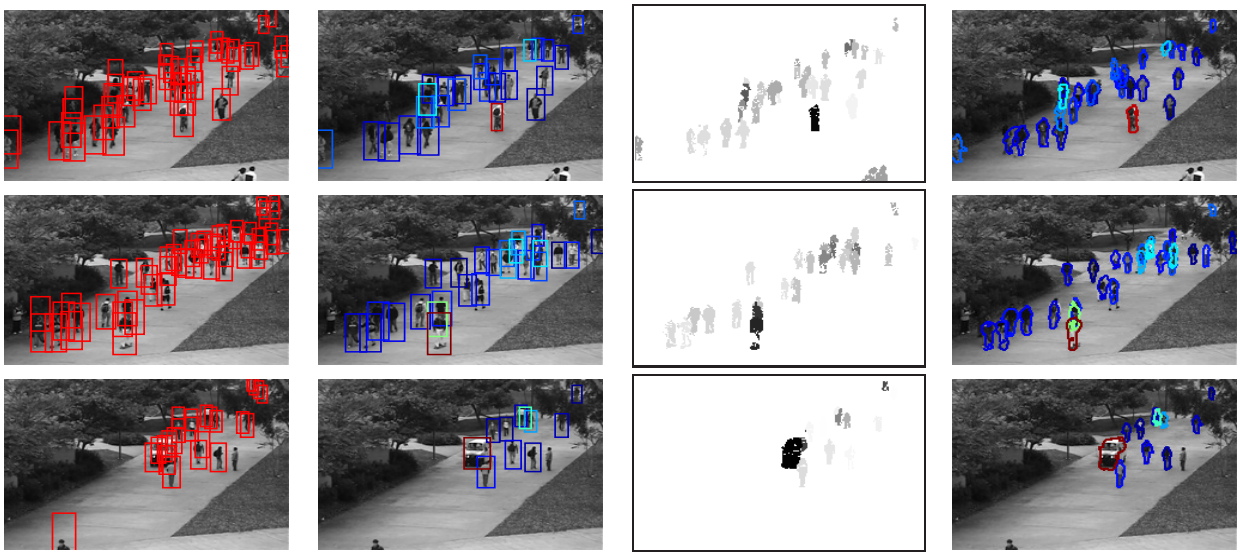


Figure 5. Rows show additional detection results on various frames. Column i) our initial shortlist, column ii) hypotheses and abnormality probability a_i, column iii) per-pixel probability $\widetilde{a}_j$, column iv) best fitting model m_i. Best viewed in color.

reports the detection performance of the fully labeled *Ped1* test set that we have assembled.

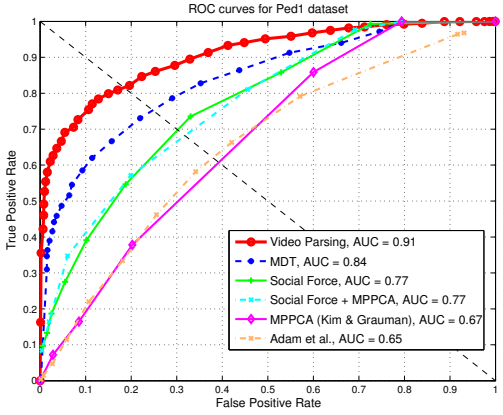


Figure 6. Per-frame abnormality detection results for the *Ped1* dataset. We achieve a 7% gain in AUC over the state-of-the-art.

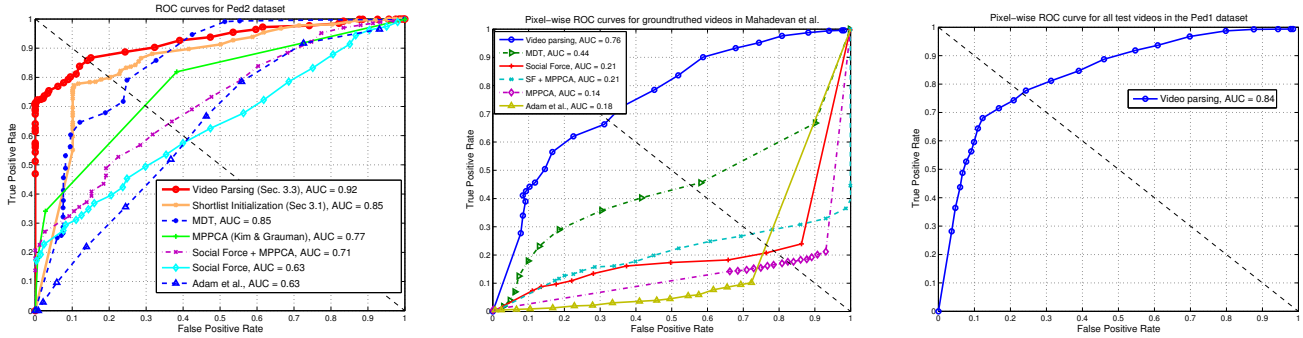
Fig. 4 compares the abnormality localization of our approach with that of [15]. The columns show results on different frames. i) Row one visualizes our initial shortlist with its spurious detections, ii) row two shows the hypotheses and their abnormality probability a_i (ranging from blue for normal to red for abnormal) after optimization, iii) row three displays the pixel-level abnormality $\widetilde{a}_j$, and iv) row four explains each hypothesis by the best fitting model m_i and for abnormalities all connected abnormal pixels are grouped. The comparison between our localization of abnormalities in row iv) with the currently best performing approach [15] in row v) further explains the significant performance gain we achieve in Fig. 7b. Further detection results of our video parsing approach are shown in Fig. 5.

5. Conclusion

To avoid the ill-posed problem of directly detecting abnormalities and classifying individual image regions independently from another as abnormal, we have proposed a scene parsing approach. All object hypotheses that are needed to explain the foreground of a video frame are jointly inferred. At the same time, each hypothesis seeks to be explained by a normal training example. In our probabilistic model, sets of hypotheses are jointly explaining the foreground while they are also able to explain each other away, simultaneously. Thus we are not detecting hypotheses individually but we find a layout that jointly describes the scene. Abnormalities are then discovered indirectly as those hypotheses which are needed to explain the scene but which themselves cannot be explained by the normal training samples. Our parsing approach has demonstrated its potential by significantly improving the state-of-the-art performance on a challenging benchmark dataset.

Acknowledgement

This work has been supported by the German Research Foundation (DFG) within the program "Spatio-/Temporal Graphical Models and Applications in Image Analysis", grant GRK 1653, and by the Excellence Initiative of the German Federal Government.



(a) (b) (c)

Figure 7. a) Per-frame abnormality detection results for the *Ped2* dataset. We achieve a 7% gain in AUC over the state-of-the-art. The orange curve illustrates the performance of our approach after the shortlist initialization (Sec. 3.1), whereas the red curve depicts the performance after the explaining away procedure (Sec.3.3). b) Pixel-wise abnormality detection for the labeling provided by Mahadevan *et al*. We observe a 32% improvement in the AUC. c) Pixel-wise abnormality detection for our fully labeled test set.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *PAMI*, 30, 2008. 2, 6

[2] N. Ahuja and S. Todorovic. Connected segmentation tree: A joint representation of region layout and hierarchy. In *CVPR*, 2008. 2

[3] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *CVPR*, 2008. 2

[4] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *ICCV*, pages 462–469, 2005. 1

[5] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, May 2011. 3

[6] H. Dee and D. Hogg. Detecting inexplicable behaviour. In *BMVC*, 2004. 2

[7] S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*, 2007. 2

[8] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *ICCV*, 2009. 2

[9] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *ICCV*, pages 128–135, 2009. 2

[10] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, 2009. 2, 6

[11] I. Kokkinos and A. L. Yuille. HOP: Hierarchical object parsing. In *CVPR*, 2009. 2

[12] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss. Human-assisted motion annotation. In *CVPR*, pages 1–8, 2008. 3

[13] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009. 2

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004. 2

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010. 1, 2, 5, 6, 7

[16] T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009. 2

[17] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. *CVPR*, 2009. 6

[18] B. Ommer, T. Mader, and J. M. Buhmann. Seeing the objects behind the dots: Recognition in videos from a moving camera. *Int. J. Comput. Vision*, 83:57–71, June 2009. 2

[19] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24:843–854, 1979. 2

[20] P. Remagnino and G. A. Jones. Classifying surveillance events from attributes and behaviour. In *BMVC*, 2001. 2

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004. 2

[22] V. D. Shet, D. Harwood, and L. S. Davis. Multivalued default logic for identity maintenance in visual surveillance. In *In ECCV*, pages 119–132, 2006. 2

[23] Z. W. Tu, X. R. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 63(2), 2005. 2

[24] X. Wang, X. Ma, and W. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *PAMI*, 31, 2009. 2

[25] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. In *NIPS*, 2009. 1, 3

[26] T. Xiang and S. Gong. Video behaviour profiling and abnormality detection without manual labelling. In *ICCV*, pages 1238–1245, 2005. 1, 2

[27] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, pages 819–826, 2004. 1, 2