

Spatio-temporal Video Parsing for Abnormality Detection

Borislav Antić and Björn Ommer

Abstract—Abnormality detection in video poses particular challenges due to the infinite size of the class of all irregular objects and behaviors. Thus no (or by far not enough) abnormal training samples are available and we need to find abnormalities in test data without actually knowing what they are. Nevertheless, the prevailing concept of the field is to directly search for individual abnormal local patches or image regions independent of another. To address this problem, we propose a method for joint detection of abnormalities in videos by spatio-temporal video parsing. The goal of video parsing is to find a set of indispensable normal spatio-temporal object hypotheses that *jointly* explain all the foreground of a video, while, at the same time, being supported by normal training samples. Consequently, we avoid a direct detection of abnormalities and discover them indirectly as those hypotheses which are needed for covering the foreground without finding an explanation for themselves by normal samples. Abnormalities are localized by MAP inference in a graphical model and we solve it efficiently by formulating it as a convex optimization problem. We experimentally evaluate our approach on several challenging benchmark sets, improving over the state-of-the-art on all standard benchmarks both in terms of abnormality classification and localization.

Index Terms—Abnormality Detection, Video Analysis, Surveillance, Video Retrieval, Graphical Models, MAP Inference



1 INTRODUCTION

With the rapid growth of video data, there is an increasing need not only for recognition of objects and their behavior, but in particular for detecting the rare, interesting occurrences of unusual objects or suspicious behavior in the large body of ordinary data. Finding such abnormalities in videos is crucial for applications ranging from automatic quality control to visual surveillance. Due to the large within-class variability, recognizing normal objects is already difficult. Abnormality detection in crowded scenes, however, features the additional challenge that there exist infinitely many ways for an object to appear in unusual context (irregular object instance) or to behave abnormally (unusual activity). Most of these abnormal instances are beforehand unknown, as this would for instance require predicting all the ways somebody could cheat or break a law. It is therefore simply impossible to learn a model for all that is abnormal or irregular. Consequently, recent work on abnormality detection [1] has focused on a setting where the training data contains only normal visual patterns. Thus a discriminative approach cannot be employed to directly localize irregularities in these benchmark datasets. But how can we find an abnormality *without knowing what to look for*? In spite of this fundamental problem, the main paradigm in abnormality detection is at present to independently classify individual video patches [2], [3] or regions [4].

If we want to avoid the ill-posed problem of having to decide locally and separately about the abnormality of each

image region, we need to abandon the standard approach of object detection, which aims at detecting all objects in a scene independently from one another. Since abnormality detection is typically concerned with videos from a static camera as in surveillance or industrial inspection, robust background subtraction algorithms [5] can be used for foreground/background segregation. Our goal is then to find a set of spatio-temporal object hypotheses that jointly explain all foreground pixels. This means that normal object hypotheses, which can be learned from the training data, are spread over the spatio-temporal volume of a video in order to cover foreground pixels, while protruding into the background as little as possible. These hypotheses need to explain the appearance and behavior of the underlying video regions. As objects are mutually overlapping in crowded scenes, the spatio-temporal placement of the object hypotheses can only be determined jointly. Thus, our aim is to simultaneously select those object hypotheses, which are necessary for explaining the foreground and to identify for each selected hypothesis the best matching instance from the set of all normal training samples. Abnormal objects are then those hypotheses which are required for explaining the foreground, but which themselves cannot be explained by a normal training sample. Video parsing *jointly* infers all necessary object hypotheses, so that we can *indirectly* discover all abnormal objects present in a scene without actually knowing what to look for.

Our video parsing approach consists of two stages. In the first phase, we detect a large number of object candidates in each video frame and then group them temporally into spatio-temporal object hypotheses. This shortlist of hypotheses is a superset of all candidates that might be eventually needed for parsing the video, i.e., it has a low false negative and high false positive rate. The object

• The authors are with the Institute for Scientific Computing (IWR), Heidelberg University, Germany.
E-mail: borislav.antic@iwr.uni-heidelberg.de

candidates in individual frames are obtained by running a discriminative background classifier and keeping only those patterns which are very unlikely to be background. Subsequently, object candidates in individual frames are linked temporally according to their motion cues so as to establish the shortlist of spatio-temporal object hypotheses. In the second phase of video parsing, the goal is to select hypotheses from the shortlist that can explain the foreground, and to simultaneously find normal object instances that match those hypotheses. We formulate this as an inference problem in a graphical model whose goal is to maximize the probability of the foreground explanation in a video. The inference in the graphical model is cast as a convex optimization problem where the unknown variables indicate both, the selection of hypotheses from the shortlist and their corresponding normal object prototypes learned from the training videos. Correspondences between hypotheses and normal object prototypes are based upon their shape, location as well as their appearance and behavior. The probability of abnormality of each hypothesis necessary for explaining the foreground is then calculated using the results of inference. Beside identifying abnormal objects, video parsing also computes per-pixel probability of abnormality, which effectively segments abnormalities without having any training samples for them.

We evaluate our approach on novel benchmark datasets for abnormality detection that feature highly crowded scenes. As an example, the UCSD *ped1* and *ped2* anomaly detection and localization datasets [1] contain busy walkways teeming with walking pedestrians. Abnormalities are not staged, but they occur spontaneously and correspond to unusual objects (e.g., vehicles in a pedestrian zone) or behaviors (e.g., a person cycling across walkways) in the scene. The training data features only normal patterns with large intra-class variability, whereas the test set consists of normal and abnormal instances. Due to the small resolution of videos (a person in the scene is on average only 20 pixels tall) and heavy occlusion between objects in the scene, learning models of visual patterns is difficult. We also increase the future utility of the UCSD *ped1* dataset by completing the pixel-wise ground-truth annotation for all videos in the test set that previously existed only for a small subset. The experimental results show a significant performance gain of our spatio-temporal video parsing approach in comparison to other state-of-the-art methods for abnormality detection.

2 RELATED WORK

We discuss here the previous work on abnormality detection in videos. The related problem of object recognition and tracking in crowded scenes [6], [7] aims at recognizing and tracking objects of a *known* class in a scene, whereas our goal is to detect abnormal objects, all of them being instances of an *unknown* class. Therefore, object recognition and tracking are beyond the scope of this paper and the details on these topics can be found in [8]. Majority of the work on abnormality detection relies on the extraction of

semi-local features from video [9], [10], [11], [12], [13], that are then used to train a normalcy model. Abnormalities are detected if the normalcy model does not fit the data. Some approaches [14], [4] are based on manually specifying constraints that define the condition of normalcy, whereas other methods [3], [15], [16], [17], [18], [19] learn the normalcy model directly from data in unsupervised way.

The approach of Adam et al. [20] focuses on individual activities occurring only in selected parts of a scene. Kim and Grauman [21] detect abnormalities using a spatio-temporal Markov random field that adapts to abnormal activities in videos. Loy et al. [22] use active learning methodology to integrate human feedback into the detection of abnormal events and behaviors. Unsupervised topic models are used for detection of abnormal behaviors in [23], [24]. Hospedales et al. [25] propose a semi-supervised multi-class topic model to classify and localize the subtle behavior in cluttered videos. Mahadevan et. al [1] detect unusual objects in crowded scenes by jointly modeling the dynamics and appearance with mixtures of dynamic textures. Li et al. [26] use the mixture of dynamic textures at multiple scales to detect abnormalities in a conditional random field framework.

Kratz and Nishino [27] develop a statistical model of local motion patterns in very crowded scenes to find abnormalities as local volumes with a large motion variation. Benezeth et al. [28] use low-level features to learn the co-occurrence matrix of normal behavior, and apply Markov random field to find deviating behaviors. Cong et al. [29] use sparse reconstruction cost implemented on a normal dictionary of local spatio-temporal patches to detect local and global abnormalities. Saligrama et al. [30] propose optimal decision rules for detecting local spatio-temporal abnormalities. An efficient sparse combination learning framework that achieves decent performance in the detection phase is proposed by Lu et al. [31].

Instead of independently detecting abnormal regions in video as in other approaches, abnormalities are discovered indirectly after establishing a set of spatio-temporal hypotheses that provide complete explanation of the foreground. Previous approaches related to scene parsing differ in that a parametric scene [32], [33] or object model [34], [35], [36] or a non-parametric exemplar-based representation for objects [37], [38] can be constructed. In contrast to these methods we are not provided any training samples for the abnormalities we are searching for but we can leverage a foreground/background segregation. In contrast to our previous sequential video parsing [39] that parsed video frames only spatially, one after another, the approach proposed in this paper performs a joint spatio-temporal parsing of video frames. This methodological extension is used to resolve both the spatial and temporal dependencies between objects in a scene. The new convex formulation of the inference process that improves upon the previous locally optimal inference method allows us to efficiently aggregate evidence from different frames and decide about their abnormalities in a globally optimal manner.

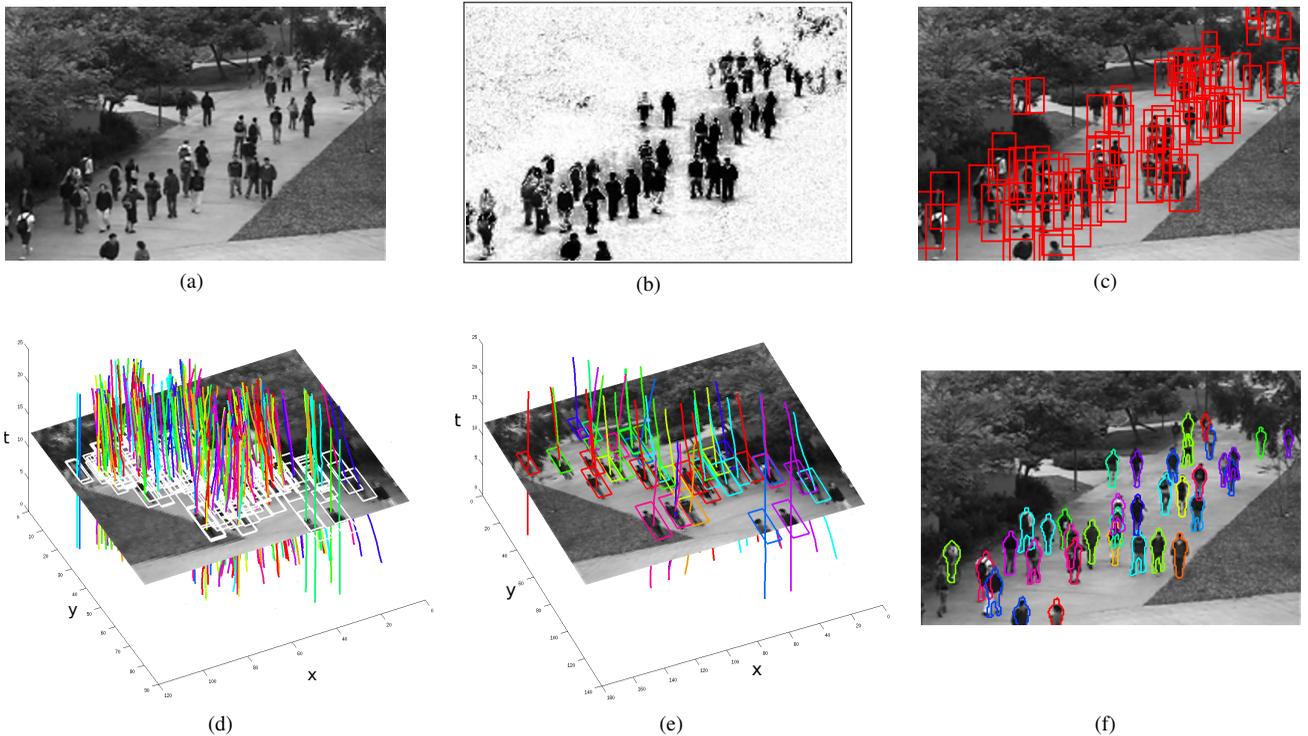


Fig. 1. Successive stages of the video parsing: (a) Source frame of a video. (b) Foreground probability map that needs to be explained by video parsing. (c) Object candidates found by inverted background detector. (d) Spatio-temporal object hypotheses found by temporal grouping serve as an input to the video parsing. (e) Subset of spatio-temporal object hypotheses that is selected by video parsing to explain the foreground pixels. (f) Normal object prototypes found by video parsing to explain the selected object hypotheses. Best viewed in color.

3 MODEL FOR SPATIO-TEMPORAL VIDEO PARSING

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 of a video using object hypotheses and to explain each hypothesis by an object model learned from the set of normal training videos. 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 any prototype from the normal object model.

Foreground segmentation. Scenarios for abnormality detection often involve the analysis of videos from static cameras. Background in such videos is constant or changes slowly over time, hence it can be learned effectively from a video. The resulting background model can then be applied to find all foreground pixels in the video. The final foreground/background segmentation is represented by a binary variable $f_j^t \in \{0, 1\}$ for all pixels j in frame t .

Background subtraction assumes that each frame I^t of a video can be expressed as the background model B^t plus a sparse vector $I^t - B^t$ whose nonzero elements are the

foreground pixels. After stacking successive video frames as columns in a matrix $I = [I^{t-\tau} \dots I^t]$, we want to find the low-rank background model B such that the sparsity inducing norm of the difference $I - B$ is the smallest possible. Following the approach of Wright et al. [5], we approximate the rank of the matrix B by a nuclear norm¹ $\|\cdot\|_*$ and use ℓ_1 as the sparsity inducing norm, so that the background subtraction becomes the following convex optimization problem,

$$B = \underset{\tilde{B}}{\operatorname{argmin}} \|\tilde{B}\|_* + \|I - \tilde{B}\|_1. \quad (1)$$

Now that we calculated the background model B , it can be used to find all foreground pixels j , $f_j^t = 1$, as those that have a large discrepancy between the observation I_j^t and the background model B_j^t . The probability that a pixel is foreground $P(f_j^t = 1)$ is obtained by the sigmoid transformation of the difference of pixel's intensity and background model,

$$P(f_j^t = 1) = 2 \left(1 + \exp(-\lambda \|I_j^t - B_j^t\|) \right)^{-1} - 1. \quad (2)$$

Pixels with foreground probability greater than 0.5 are considered as foreground, $f_j^t = 1$, and others as background, $f_j^t = 0$.

1. Nuclear norm is the sum of the singular values of the matrix and is a convex function.

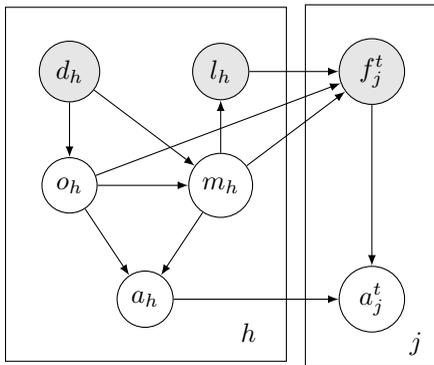


Fig. 2. Probabilistic graphical model of the spatio-temporal video parsing. The left plate contains all spatio-temporal hypotheses h with their descriptors d_h and locations l_h . The right plate comprises all pixels j with their foreground labels $f_j^t \in \{0, 1\}$. By video parsing, we infer the set of hypotheses, $o_h \in \{0, 1\}$, that are necessary for explaining the foreground, and *jointly* explain the selected hypotheses by the normal object prototypes $m_h \in \{1, \dots, K\}$. Finally, for each selected hypothesis h we decide if it is abnormal, $a_h \in \{0, 1\}$, and also mark foreground pixels that belong to abnormal objects, $a_j^t \in \{0, 1\}$.

Shortlist of Object Hypotheses. For parsing the video, we need to specify a list of spatio-temporal object hypotheses that is sufficient for explaining foreground pixels in video. An input to our video parsing algorithm consists of the most suitable object hypotheses for the task of foreground explanation. In Sect. 6 we explain the procedure for creating a shortlist of object hypotheses that has a high recall, i.e. where the majority of true-positive object hypotheses is included in the shortlist. However, as the precision rate of the proposed shortlist is low, there will be many superfluous hypotheses that are then explained away by others during video parsing.

We assume that hypotheses from the shortlist span a time window $\{t - \tau, \dots, t\}$. Each hypothesis h represents a spatio-temporal tube covering locations $l_h := (l_h^{t-\tau} \dots l_h^t)$. This is a trajectory of locations $l_h^t = (x_h^t, y_h^t, s_h^t)^\top$, which specify the center (x_h^t, y_h^t) and the scale s_h^t of a candidate object h at time t . The scale of an object represents its size relative to the size (W, H) of the object model. The *support region* of an object hypothesis h at time t is the bounding box of size $(s_h^t W, s_h^t H)$, and the set of all pixels j that belong to it is denoted by \mathcal{S}_h^t .

The goal of video parsing is then to select a subset from the shortlist of hypotheses that is both necessary and sufficient for explaining the foreground of a test video while, at same time, finding normal object prototypes that explain the hypotheses of the subset (see Fig. 1).

Spatio-temporal object descriptor. A spatio-temporal hypothesis h matches its corresponding normal object prototype both in appearance and motion. Thus, we need a spatio-temporal descriptor d_h to capture the essence of both appearance and motion of hypothesis h . We build

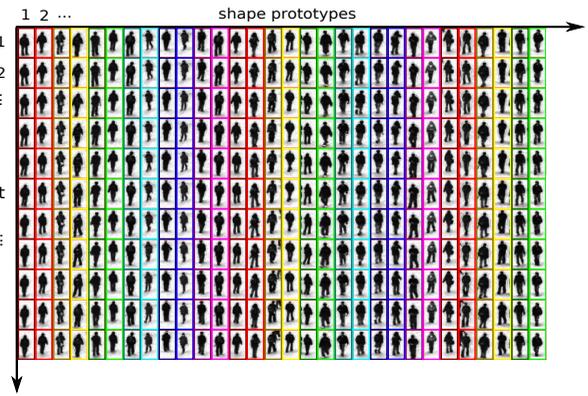


Fig. 3. The normal object model consist of a set of spatio-temporal shape prototypes, each being a sequence that captures the temporal evolution of a particular shape. Prototypes are accompanied by the appearance and motion descriptors.

a spatio-temporal descriptor $d_h := (d_h^{t-\tau} \dots d_h^t)^\top$ by concatenating frame-wise descriptors d_h^t calculated at each time t . Frame-wise object appearance is represented by the spatial derivatives of pixel's intensity in the support region \mathcal{S}_h^t of hypothesis h . Analogously, object motion is represented by the temporal derivatives of pixel's intensity. The appearance and motion representations are combined into frame-wise descriptor,

$$d_h^t := \left(\frac{\partial I_j^t}{\partial x}, \frac{\partial I_j^t}{\partial y}, \frac{\partial I_j^t}{\partial t} \right)_{j \in \mathcal{S}_h^t}. \quad (3)$$

Since the spatio-temporal descriptor d_h is long and redundant, we build its compact representation by applying PCA transformation that projects onto eigen-space such that most of the signal variation is preserved (about 95%).

Activating hypotheses needed for parsing. Not all object hypotheses from the shortlist are needed to explain foreground pixels in video. Video parsing retains only the indispensable hypotheses that cannot be explained away by other hypotheses. Therefore, we use an indicator variable $o_h \in \{0, 1\}$ for each hypothesis h to designate the hypothesis as active/inactive. To initialize parsing, a discriminative classifier is trained to distinguish background spatio-temporal patterns from anything else. This background classifier computes the probability that hypothesis h is background, $P(o_h = 0 | d_h)$, which is then inverted to obtain the foreground probability. A hypothesis with high foreground probability can still become inactive if it gets explained away by others during video parsing.

Matching with the object model. Video parsing jointly explains foreground pixels with object hypotheses, and active hypotheses $\{h : o_h = 1\}$ with normal object prototypes learned from the training data. The object model consists of K normal object prototypes that represent a diversity of normal object's shape, appearance, and motion. Video parsing then determines for each selected hypothesis h which of the K prototypes best explains it. The prototype that video parsing associates with hypothesis h is indicated

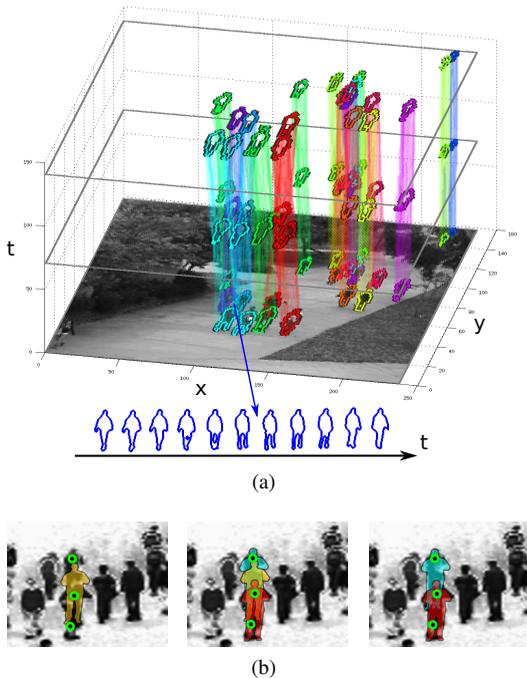


Fig. 4. (a) Spatio-temporal tubes illustrate the hypotheses selected by video parsing. Normal shape contours that explain the hypotheses are shown overlaid. (b) Superfluous hypotheses are eliminated by the statistical inference of *explaining away*. The idea is the following: Object hypothesis (yellow) is used at the beginning of video parsing to explain the foreground pixel in the middle. Other object hypotheses (red and blue) are introduced later to explain the top and bottom pixels. However, the pixel in the middle is also explained by new hypotheses, so that the original (yellow) hypothesis is not needed anymore and it can be eliminated.

by the variable $m_h \in \{1, \dots, K\}$. Sect. 5 explains in detail the learning of the normal object prototypes. For the time being, we assume that K normal object prototypes are provided as input to the parsing algorithm.

For each hypothesis h the best prototype $m_h \in \{1, \dots, K\}$ from the learned object model is sought (Fig. 3). For abnormal objects all prototypes will obviously have high matching costs. Consequently, the probability that prototype m_h is matched to a hypothesis h in a query video depends on how similar they are in both appearance and motion, $\Delta(d_h, d_{m_h})$. Here, Δ denotes a function that measures the distance of spatio-temporal descriptors in the corresponding feature space. Given the spatio-temporal descriptor d_h of hypothesis h , the probability of matching prototype m_h with the hypothesis h is the Gibbs distribution,

$$P(m_h|d_h) = \frac{1}{Z(d_h)} \exp(-\beta \Delta(d_h, d_{m_h})), \quad (4)$$

where $Z(d_h)$ is the partition function used to normalize the probability distribution.

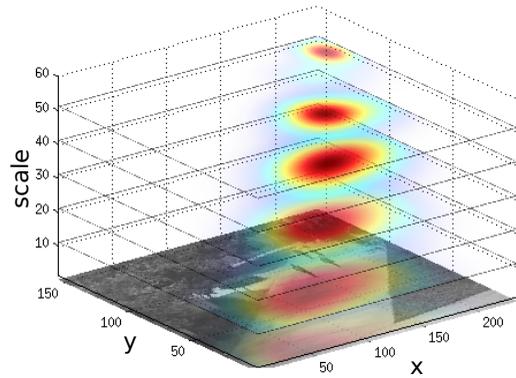


Fig. 5. The distribution of locations of normal object prototypes estimated by the Parzen windows at multiple scales (represented as horizontal slices).

Moreover, normal objects typically occupy some location in a scene more often than other, and also tend to move at a certain speed. For example, cars are more likely to drive on roads than on sidewalks, whereas pedestrians are more likely to walk on sidewalks. Consequently, the probability of observing hypothesis h that matches the prototype m_h depends on its location l_h^t and velocity $l_h^t - l_h^{t-1}$,

$$P(l_h|m_h) \propto P_{m_h}^{loc}(l_h^t) \cdot P_{m_h}^{vel}(l_h^t - l_h^{t-1}). \quad (5)$$

The normal location and velocity distributions P_{\bullet}^{loc} and P_{\bullet}^{vel} are learned for each of the K object prototypes using the Parzen window density estimator (see Fig. 5).

Therefore, the probability that hypothesis h matches to the normal object prototype m_h is

$$P(m_h|o_h, d_h, l_h) \propto o_h \cdot P(m_h|d_h) \cdot P(l_h|m_h). \quad (6)$$

Explaining foreground pixels. Video parsing selects hypotheses, $\{h : o_h = 1\}$, and finds corresponding normal object prototypes m_h to explain the foreground. The foreground probability of a pixel j depends on all hypotheses h that overlap with pixel j . Given the support regions \mathcal{S}_h^t of all hypotheses h , $\{h : j \in \mathcal{S}_h^t\}$ is the set of hypotheses that cover the pixel j . The probability that pixel j is background is equal to the product of pixel's background probabilities for each single hypothesis h that contains the pixel j . Even if all hypotheses claim that pixel j is background, $P(f_j^t = 1|o_h, m_h, l_h) = 0, \forall h$, we still allow it to be foreground with a small foreground probability $P_0 > 0$. Thus, foreground probability of pixels j given all hypotheses is

$$P(f_j^t = 1|\{o_h, m_h, l_h\}_h) = 1 - (1 - P_0) \prod_h (1 - P(f_j^t = 1|o_h, m_h, l_h)). \quad (7)$$

The foreground probability given a single hypothesis h , $P(f_j^t = 1|o_h, m_h, l_h)$, depends on the shape of the corresponding normal object prototype m_h . In the training data,

the prototype m_h covers pixels j' with some probability $P_{m_h}(f_{j'}^t = 1)$. Thus, the foreground probability of pixel j under hypothesis h is obtained by taking its corresponding object prototype m_h and “pasting” the foreground probability of m_h at the location of h . The model now needs to be brought into the reference frame of h by scaling and translating it, i.e. $l_j^t = s_h^t \cdot l_{j'}^t + (x_h^t \ y_h^t)^\top$. Then the foreground probability of pixel j given h becomes

$$P(f_j^t = 1 | o_h, m_h, l_h) = o_h \cdot \mathbf{1}[j \in S_h^t] \cdot \sum_{j'} \mathbf{1}[l_j^t = s_h^t \cdot l_{j'}^t + (x_h^t \ y_h^t)^\top] \cdot P_{m_h}(f_{j'}^t = 1). \quad (8)$$

Here $\mathbf{1}[\cdot]$ denotes the indicator function. In Eq. 8 the foreground probability of pixel j is set to zero if hypothesis h is inactive, $o_h = 0$, or the pixel j does not belong to the support region of hypothesis h , $j \notin S_h^t$.

4 INFERENCE BY FOREGROUND PARSING

The goal is now to estimate which of the hypotheses are actually needed for explaining the foreground and to find a matching normal object prototype for each hypothesis. For abnormal hypotheses Eq. 6 will yield low probabilities. If foreground $f_j^t = 1$ is observed and the pixel is covered by a hypothesis h , and no other hypothesis can be found that could explain the presence of the foreground at that pixel, then the probability of activation of hypothesis h increases. This leads to the statistical inference of *explaining away*. For an observed variable f_j^t different hypotheses h that share the same pixel j become statistically dependent so that the absence of one hypothesis can dictate the presence of another (see Fig. 4).

4.1 Joint Inference by MAP

Based on the foreground segmentation mask f_j^t and the shortlist of hypotheses h with spatio-temporal descriptors d_h and trajectories l_h , we need to jointly infer all hidden variables $\{o_h, m_h\}_h$ in our graphical model (Fig. 2). Following a maximum a posteriori (MAP) approach yields a set of hypotheses that best explain the foreground and are themselves explained by the normal object prototypes,

$$\{\bar{o}_h, \bar{m}_h\}_h = \max_{\{o_h, m_h\}_h} P(\{o_h, m_h\}_h | \{d_h, l_h\}_h, \{f_j^t\}_j) \propto \prod_j P(f_j^t | \{o_h, m_h, l_h\}_h) \prod_h P(o_h | d_h) P(m_h | o_h, d_h, l_h). \quad (9)$$

Instead of explicitly maximizing the posterior probability, we take a negative logarithm of Eq. 9 and thereby obtain the energy function $J(\cdot)$ which is then minimized. Furthermore, we decompose the energy function $J(\cdot)$ into two terms, $J_j(\cdot)$ covering the explanation of foreground pixels j , and $J_h(\cdot)$, which involves the explanation of hypotheses h by

the normal object prototypes,

$$J(\{o_h, m_h\}_h) := - \underbrace{\sum_j \log P(f_j^t | \{o_h, m_h, l_h\}_h)}_{=: J_j(\{o_h, m_h\}_h)} - \underbrace{\sum_h \left(\log P(o_h | d_h) + \log P(m_h | o_h, d_h, l_h) \right)}_{=: J_h(\{o_h, m_h\}_h)}. \quad (10)$$

To find the MAP solution, we introduce a parsing indicator $z_{h,k} \in \{0, 1\}$, that equals one if hypothesis h is active, $o_h = 1$, and their corresponding normal object prototype is $m_h = k$,

$$z_{h,k} := o_h \cdot \mathbf{1}[m_h = k], \quad \forall h, \forall k \in \{1, \dots, K\}. \quad (11)$$

To keep the notation simple, let the vector $\mathbf{z}_h := (z_{h,1}, \dots, z_{h,K})^\top$ denote the parsing indicators of hypothesis h , and the vector $\mathbf{z} := \{\mathbf{z}_h\}_h$ denote the parsing indicators of all hypotheses together. The following lemma now states that the hypotheses explanation $J_h(\cdot)$ can be expressed as a linear function of the parsing indicator \mathbf{z} .

Lemma 4.1: The hypotheses explanation term $J_h(\{o_h, m_h\}_h)$ in Eq. 10 is a linear function of the parsing indicator \mathbf{z} , i.e.

$$J_h(\{o_h, m_h\}_h) = \mathbf{b}^\top \mathbf{z} + b_0, \quad (12)$$

where the parameter vector $\mathbf{b} = \{b_{h,k}\}_{h,k}$ and scalar b_0 do not depend on the parsing indicator \mathbf{z} . The proof of Lemma 4.1 is given in Appendix A.

To express the foreground explanation term $J_j(\cdot)$ as a function of the parsing indicator \mathbf{z} , we first define a function $\Phi_{f_j^t}(\cdot)$ that is parametrized by the foreground value f_j^t of pixel j ,

$$\Phi_{f_j^t}(x) := (1 - f_j^t)x - f_j^t \cdot \log(1 - e^{-x}), \quad x > 0. \quad (13)$$

The introduced function $\Phi_{f_j^t}(\cdot)$ is convex as we show in the following lemma.

Lemma 4.2: The function $\Phi_{f_j^t}(x)$, $x > 0$ (Eq. 13) is convex for nonnegative values of the parameter f_j^t . The proof of Lemma 4.2 is given in Appendix B.

We also introduce a joint shape prototype vector $\mathbf{w} := [\mathbf{w}_1^\top \dots \mathbf{w}_K^\top]^\top$ that is obtained by concatenating all individual shape prototype vectors \mathbf{w}_k , $k \in \{1, \dots, K\}$ (c.f. Fig. 3). The component $\mathbf{w}_{k,j'}$ equals the negative logarithm of the background probability of pixel j' in the normal shape prototype \mathbf{w}_k ,

$$\mathbf{w}_{k,j'} = -\log(1 - P_k(f_{j'}^t = 1)). \quad (14)$$

The following lemma establishes a relationship between the foreground explanation term $J_j(\mathbf{z})$, the parsing indicator \mathbf{z} and the joint shape prototype vector \mathbf{w} .

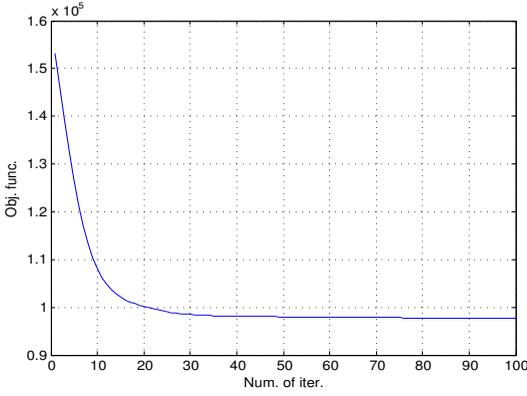


Fig. 6. Values of the objective function $J(\mathbf{z})$ (Eq. 16) that are obtained as part of the convex optimization procedure that is used to solve the proposed video parsing problem.

Lemma 4.3: The foreground explanation term $J_j(\cdot)$ is the sum over all pixels j of convex functions $\Phi_{f_j^t}(\cdot)$ whose argument is a bilinear function of the parsing indicator \mathbf{z} and the joint shape prototype \mathbf{w} ,

$$J_j(\mathbf{z}) = \sum_j \Phi_{f_j^t}(\mathbf{w}^\top \mathbf{C}_j \mathbf{z} + c_0). \quad (15)$$

The parameter matrices \mathbf{C}_j and scalar c_0 do not depend on the parsing indicator \mathbf{z} or joint shape prototype \mathbf{w} . The proof of Lemma 4.3 is given in Appendix C.

In Lemmas 4.3 and 4.1 we expressed the foreground and hypotheses explanation terms $J_j(\cdot)$ and $J_h(\cdot)$ as convex functions of the parsing indicator \mathbf{z} . Therefore, the video parsing objective function $J(\cdot) := J_j(\cdot) + J_h(\cdot)$ (Eq. 10) is a convex function of the parsing indicator \mathbf{z} . To efficiently solve the optimization problem, we relax the parsing indicator \mathbf{z} to the positive simplex, $\mathbf{z}_h \succeq 0$ and $\mathbf{1}^\top \mathbf{z}_h \leq 1$, $\forall h$. The last inequality follows from Eq. 11 and the fact that $o_h \leq 1$.

The MAP inference in our video parsing model is thus equivalent to the following constrained convex optimization problem,

$$\begin{aligned} \underset{\mathbf{z}}{\operatorname{argmin}} J(\mathbf{z}) &= \mathbf{b}^\top \mathbf{z} + b_0 + \sum_j \Phi_{f_j^t}(\mathbf{w}^\top \mathbf{C}_j \mathbf{z} + c_0), \\ \text{s.t. } \mathbf{z}_h &\succeq 0 \text{ and } \mathbf{1}^\top \mathbf{z}_h \leq 1, \forall h. \end{aligned} \quad (16)$$

After finding the optimal value of the parsing indicator \mathbf{z} , we calculate the hypothesis indicator o_h , and the matching normal object prototype m_h of hypothesis h , as

$$o_h = \sum_{k=1}^K z_{h,k}, \quad (17)$$

$$m_h = \operatorname{argmax}_k z_{h,k}. \quad (18)$$

4.2 Solving the Convex Optimization Problem

In the previous section, we showed that the joint inference of variables $\{o_h, m_h\}_h$ can be achieved by minimizing the MAP objective function $J(\mathbf{z})$ to obtain the parsing indicator

\mathbf{z} (Eq. 16), that belongs to the Cartesian product $\mathbf{Z} = \mathbf{Z}_h \times \dots \times \mathbf{Z}_h$ of positive simplexes,

$$\mathbf{Z}_h = \{\mathbf{z}_h : \mathbf{z}_h \succeq 0 \text{ and } \mathbf{1}^\top \mathbf{z}_h \leq 1\}. \quad (19)$$

The function $J(\mathbf{z})$ is convex, smooth and bounded on the set \mathbf{Z} . The *projected gradient* method [40],

$$\mathbf{z}^{n+1} = \operatorname{Proj}_{\mathbf{Z}}(\mathbf{z}^n - \alpha_n \nabla_{\mathbf{z}} J(\mathbf{z}^n)), \quad (20)$$

converges to the global optimum of the convex optimization problem in Eq. 16, because of the Lipschitz-continuity of the first derivative of function $\Phi_{f_j^t}(\cdot)$ as stated in the following lemma.

Lemma 4.4: The first derivative of the function $\Phi_{f_j^t}(x)$ is ρ -Lipschitz continuous in argument $x \geq c_0$, i.e. there is a constant ρ such that

$$|\Phi'_{f_j^t}(x_1) - \Phi'_{f_j^t}(x_2)| \leq \rho |x_1 - x_2|, \forall x_1, x_2 \geq c_0. \quad (21)$$

The proof of Lemma 4.4 is given in Appendix D.

The projection $\operatorname{Proj}_{\mathbf{Z}}(\cdot)$ requires each \mathbf{z}_h to be projected onto the positive simplex \mathbf{Z}_h . The projection onto the positive simplex is calculated by applying the method of Duchi et al. [41]. The projected gradient method finds the solution of the video parsing after only few tens of iterations (Fig. 6).

4.3 From Inference to Abnormalities

Video parsing analyses the foreground in a video and identifies objects that have atypical appearance or behave suspiciously, to label these as abnormal. Abnormalities can also be localized on the level of pixels, where it leads to a segmentation of regions in the video that contain irregular spatio-temporal patterns. Subsequently, we see how both the object-level and pixel-level abnormalities can be detected in video, based on the inference results of our video parsing approach.

Object-level abnormalities. A hypothesis h is an abnormal object, $a_h = 1$, if it is indispensable for explaining the foreground, $\bar{o}_h = 1$, but it does not have a matching normal object prototype, i.e., the best estimate \bar{m}_h of a matching prototype is unlikely to explain the hypothesis (cf. Eq. 6),

$$\begin{aligned} &P(a_h = 1 | o_h = \bar{o}_h, m_h = \bar{m}_h) \\ &\propto \bar{o}_h P(o_h = 1 | d_h) P(m_h \neq \bar{m}_h | o_h = \bar{o}_h, d_h, l_h) \quad (22) \\ &\propto \bar{o}_h P(o_h = 1 | d_h) \left(1 - P(m_h = \bar{m}_h | d_h) P(l_h | m_h = \bar{m}_h)\right). \end{aligned} \quad (23)$$

Pixel-level abnormalities. Similarly, a pixel j is part of an abnormal object, $a_j^t = 1$, if it is in the foreground, $f_j^t = 1$, and at least one of the hypotheses that extend over this pixel, $\{h : j \in \mathcal{S}_h^t\}$, is abnormal,

$$\begin{aligned} &P(a_j^t = 1 | f_j^t, \{a_h\}_{h: j \in \mathcal{S}_h^t}) \\ &\propto f_j^t \cdot P(f_j^t = 1) \cdot \max_{h: j \in \mathcal{S}_h^t} P(a_h = 1 | o_h, m_h). \end{aligned} \quad (24)$$

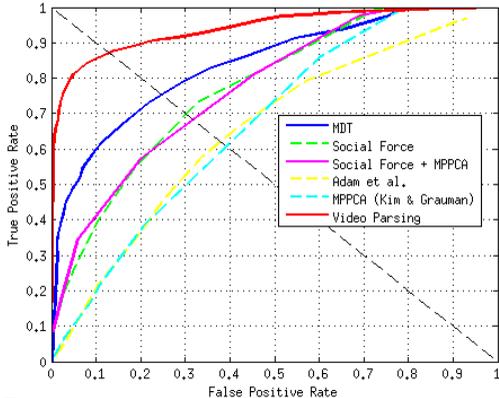


Fig. 7. Frame-wise abnormality labeling on the UCSD *ped1* dataset. Performance measures AUC and EER given in Tab. 1 are calculated from the ROC curves.

5 LEARNING AN OBJECT MODEL FOR VIDEO PARSING

Parsing query videos for abnormality detection requires an object model. We use training videos that contain a large number of normal object samples but no abnormalities to train the normal object model that consists of prototypes representing the normal object shape, appearance, and motion. As ground truth locations of objects in the training videos are not provided, we infer them by video parsing. However, for video parsing we need to know the normal object prototypes. A standard approach for solving such a problem of mutual dependencies is *expectation-maximization* (EM) [42]. Given an initial estimate of the normal object prototypes, we use them to parse the training videos, i.e. discover hypotheses that best explain the foreground and are matched to the object prototypes (E-step). Thereafter, we update the object prototypes using the matched hypotheses (M-step). We find the object model by iterating the EM steps until convergence.

The goal of learning is to estimate the normal object shape prototypes $\{\mathbf{w}_k\}_k$ (Eq. 14) and their corresponding spatio-temporal descriptors $\{d_k\}_k$, $k \in \{1, \dots, K\}$. The objective function for learning is the same as for the inference (Eq. 16), except that it is now minimized jointly in terms of shape prototypes $\{\mathbf{w}_k\}_k$, their spatio-temporal descriptors $\{d_k\}_k$, as well as the parsing indicator \mathbf{z} (Eq. 11),

$$\begin{aligned} \operatorname{argmin}_{\{\mathbf{z}, \mathbf{w}_k, d_k\}_k} J(\mathbf{z}, \{d_k, \mathbf{w}_k\}_k) &= J_h(\mathbf{z}, \{d_k\}_k) + J_j(\mathbf{z}, \{\mathbf{w}_k\}_k), \\ \text{s.t. } \mathbf{w}_k &\geq 0, \forall k, \mathbf{z}_h \geq 0 \text{ and } \mathbf{1}^\top \mathbf{z}_h \leq 1, \forall h. \end{aligned} \quad (25)$$

The hypotheses explanation term $J_h(\cdot)$ is a function of the parsing indicator \mathbf{z} and the spatio-temporal descriptors $\{d_k\}_k$,

$$J_h(\mathbf{z}, \{d_k\}_k) = \beta \sum_h \sum_k z_{h,k} \Delta(d_h, d_k) + \tilde{\mathbf{b}}^\top \mathbf{z} + b_0, \quad (26)$$

where the parameters $\tilde{\mathbf{b}}$ and b_0 do not depend on the parsing indicator \mathbf{z} or the spatio-temporal descriptors $\{d_k\}_k$ (see the proof of Lemma 4.1 in Appendix A).

From Eq. 15 we see that the foreground explanation term $J_j(\cdot)$ depends in a convex way on both the parsing indicator \mathbf{z} and the joint shape prototype vector \mathbf{w} .

Procedure for the object prototype learning. We now explain the EM algorithm used for solving the optimization problem of Eq. 25:

E-step. Given the object prototypes, we parse the training videos to infer the parsing indicator \mathbf{z} (Eq. 16) that yields the hypothesis indicator o_h for each hypothesis h , and its corresponding normal object prototype m_h (Eq. 17 and 18).

M-step. We estimate the shape prototypes $\{\mathbf{w}_k\}_k$ and their spatio-temporal descriptors $\{d_k\}_k$ from the results of video parsing. As hypotheses overlap in training videos, the corresponding shape prototypes become mutually dependent and thus need to be learned jointly. We estimate the joint shape prototype vector \mathbf{w} by the following convex optimization,

$$\mathbf{w} = \operatorname{argmin}_{\tilde{\mathbf{w}} \geq 0} J_j(\mathbf{z}, \tilde{\mathbf{w}}) = \sum_j \Phi_{f_j^t}(\tilde{\mathbf{w}}^\top \mathbf{C}_j \mathbf{z} + c_0). \quad (27)$$

The convex optimization problem of Eq. 27 can be solved efficiently by the projected gradient method that we used for solving the MAP inference problem (Eq. 20),

$$\mathbf{w}^{n+1} = \operatorname{Proj}_{\mathbb{R}_+^{|\mathbf{w}|}}(\mathbf{w}^n - \alpha_n \nabla_{\mathbf{w}} J_j(\mathbf{z}, \mathbf{w}^n)). \quad (28)$$

The spatio-temporal descriptors $\{d_k\}_k$, $k \in \{1, \dots, K\}$ are estimated separately for each normal object prototype,

$$d_k = \operatorname{argmin}_{\tilde{d}_k} \sum_h z_{h,k} \Delta(d_h, \tilde{d}_k). \quad (29)$$

In case of a squared Euclidean distance function, $\Delta(d_h, d_k) = \|d_h - d_k\|^2$, there is a closed-form solution for d_k , given as an average of spatio-temporal descriptors d_h of those hypotheses that are matched to prototype k by video parsing,

$$d_k = \frac{\sum_h z_{h,k} d_h}{\sum_h z_{h,k}}. \quad (30)$$

The EM algorithm assumes uniform location and velocity distributions (Eq. 5) for normal object prototypes. However, after the EM algorithm is converged, we estimate the prototype's location and velocity distributions from matched object hypotheses by the non-parametric *Parzen* windows.

Initialization. To start the EM algorithm, we need an initial estimate of the normal object model. After background subtraction, some foreground segments correspond to isolated normal objects that can be used to initialize our object prototypes. However, foreground/background segmentation produces also many foreground segments which correspond to interacting objects (doublets, triplets etc.). These segments are more complex and can be analyzed only by video parsing. Consequently, we need to infer which of the training foreground segments correspond to isolated normal objects and estimate object prototypes based upon them. We observe that isolated normal objects create compact clusters in the feature space. On the other hand, segments that are mixtures of two or more objects are diverse and spread out in the feature space. To detect isolated normal objects,

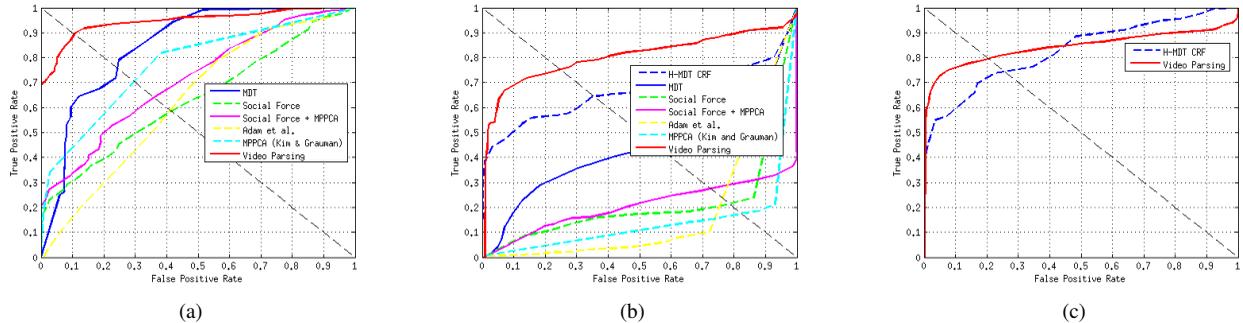


Fig. 8. (a) Frame-wise abnormality labeling for the UCSD *ped2* dataset. (b) Pixel-wise abnormality prediction evaluated by the partially annotated UCSD *ped1* dataset. (c) Pixel-wise abnormality prediction that is evaluated using the full annotation of the complete UCSD *ped1* dataset that we have assembled. In all of these cases our approach significantly improves upon the state-of-the-art, which can also be seen from the corresponding AUC and RD values provided in Tab. 1 and 2.

TABLE 1
Performance measures on the UCSD *ped1* dataset

	frame-wise		pixel-wise partial		pixel-wise full	
	AUC (%)	EER (%)	AUC (%)	RD (%)	AUC (%)	RD (%)
Social force [43]	67.5	31	19.7	21	-	-
MPPCA [21]	59	40	20.5	18	-	-
Social force + MPPCA	67	32	21.3	28	-	-
Adam [20]	65	38	13.3	24	-	-
Sparse [29]	86	19	46.1	46	-	-
LSA [30]	92.7	16	-	-	-	-
SCL [31]	91.8	15	63.8	59.1	-	-
MDT [1]	81.8	25	44.1	45	-	-
HMDT CRF [26]	-	17.8	66.2	64.8	82.7	74.5
SVP [39]	91	18	75.6	68	83.6	77
STVP	93.9	12.9	80.3	75.2	84.2	79.5

we cluster all the foreground segments and then select compact clusters in the feature space that correspond to isolated objects. We use *Ward's* method for agglomerative clustering to minimize the variance of clusters. Normal object prototypes are then computed as the centers of compact clusters.

6 CREATING INITIAL OBJECT HYPOTHESES

To initialize video parsing, we need a shortlist of spatio-temporal object hypotheses h (Sect. 3). A spatio-temporal hypothesis h consists of a sequence of object candidates in individual frames that are linked temporally. In this section we explain a method for producing per-frame object candidates and group them temporally based on their motion to obtain the shortlist of Sect. 3. Thereafter, we explain how to fill-in per-frame candidates that were missed during temporal grouping.

Temporal grouping of per-frame object candidates. To detect per-frame object candidates, we apply an inverted background detector that is trained to distinguish background patterns from everything else. The inverted background detector is trained on background and normal fore-

ground segments obtained from training videos by background subtraction. The discriminative appearance-based classifier retains in each frame the object candidates that are least likely to be background. The standard non-maximum suppression (NMS) then removes some of the candidates based on the overlap criteria. The discriminative classifier is trained using a linear SVM [44] with frame-wise descriptor of Eq. 3 extracted from background/foreground segments of training videos.

We then employ agglomerative clustering to perform a temporal grouping of candidates. This yields spatio-temporal hypotheses h , which are sequences of per-frame candidates. As usual, the clustering starts with singleton clusters (each candidate being a cluster). Then, in each round of the recursive clustering, those groups of per-frame object candidates which are most similar based on their motion and which do not share the same frames are grouped. The motion of a candidate is represented by the set of trajectories obtained by tracking the edge points inside the support region of a candidate. For tracking the feature points we use optical flow vectors that are previously computed by the method of [45]. We now define similarity of two object candidates as the ratio of the number of feature point trajectories that are shared by two candidates over the total number of trajectories in two candidates. As the result of temporal grouping, we obtain a shortlist of spatio-temporal hypotheses h .

Filling-in missing candidates by Kalman filter. The inverted background detector used for producing object candidates in each frame typically has a number of missed detections. These are the frames in which none of the object candidates is associated with a hypothesis h . We fill-in the missed object detections with the contextual help of other per-frame candidates that belong to the same hypothesis h . Therefore, the location of a missed object candidate l_h^t at time t is estimated from the available object candidate locations at times $\{t_1, t_2, \dots\}$ by a non-causal Kalman filter.

The shortlist of object hypotheses established by tem-

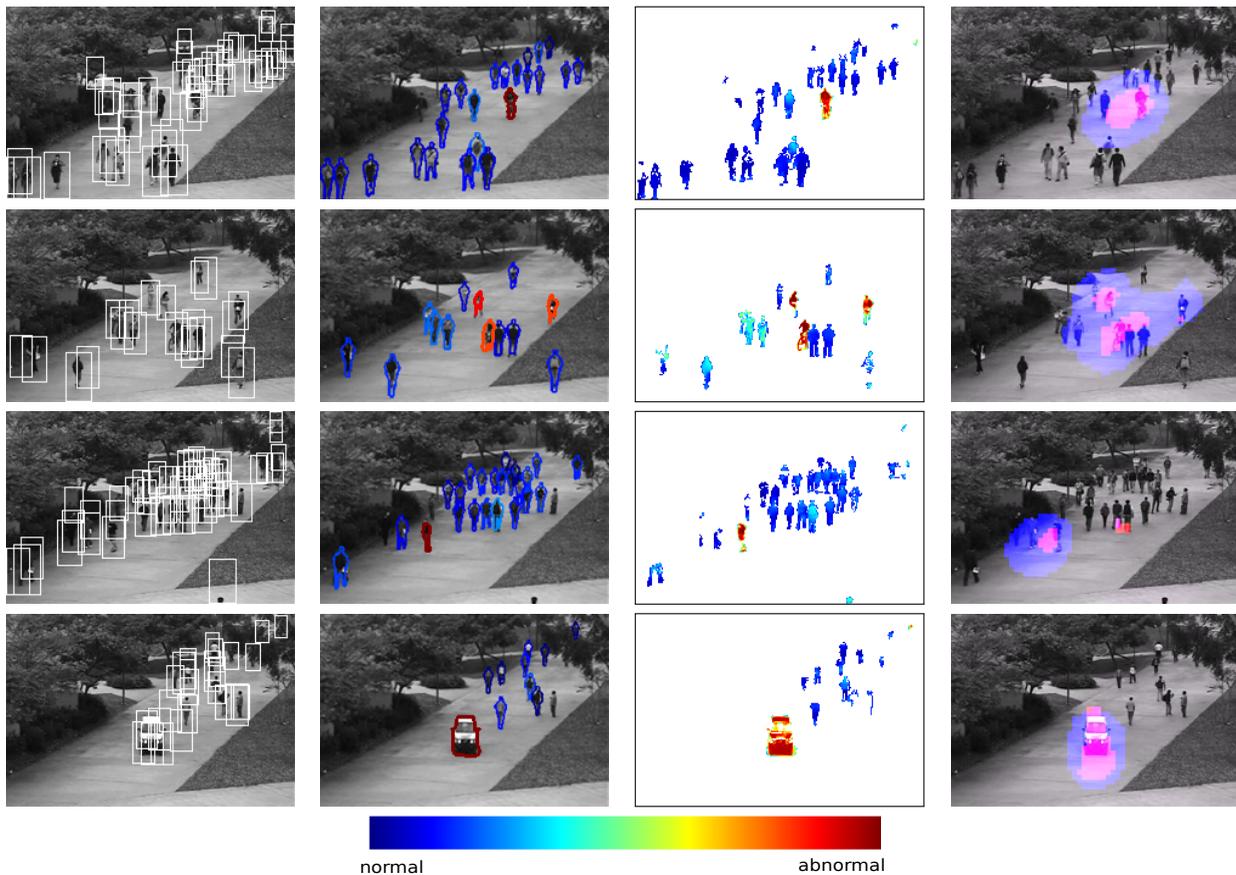


Fig. 9. Rows show results on different frames of the UCSD *ped1* benchmark. Column i) the initialization of the video parsing by a shortlist of object hypotheses, column ii) hypotheses selected by video parsing with the best matching shape prototype colored according to abnormality probability $P(a_h^t = 1)$, column iii) foreground pixel abnormality probabilities $P(a_j^t = 1)$, column iv) results by the H-MDT CRF approach [26]. Best viewed in color.

poral grouping has a high recall at the cost of low precision. By maximizing the recall, the shortlist includes all relevant hypotheses, while still maintaining a reasonable total number thereof (about one hundred). Since hypotheses are created by bottom-up grouping, there will, however, be many spurious hypotheses that can only be eliminated by video parsing.

7 EXPERIMENTAL EVALUATION

We use three standard state-of-the-art benchmark sets for evaluating our video parsing approach and comparing its performance to the other state-of-the-art methods. We first analyze the detection results of our approach on the UCSD benchmark sets *ped1* and *ped2*, then we present additional results on the UMN benchmark set. We apply the standard evaluation protocol of the datasets.

7.1 Evaluation on the UCSD Anomaly Datasets

7.1.1 Datasets Description

We use the challenging UCSD anomaly datasets *ped1* and *ped2*, that were recently proposed by Mahadevan et al. [1] for measuring the performance of abnormality detection

algorithms. Both datasets consist of videos recorded in crowded walkway scenes that also feature lots of challenging abnormal instances which are objects with unusual appearance or behavior. The UCSD *ped1* set contains 34 training and 36 test videos that are all 200 frames long. Due to the low resolution of *ped1* videos, the pedestrians who walk towards and away from the camera are only 10 – 25 pixels high. In the UCSD *ped2* dataset there are 16 training and 12 test videos that have a variable length (at most 180 frames). Pedestrians in these videos are about 30 pixels high. Videos from both benchmark sets are very crowded, so that object heavily occlude one another.

Abnormalities in the UCSD datasets are not staged but occur naturally in the scene and can be grouped into: i) objects that do not fit to the context of the scene, such as a car on a crowded walkway, or ii) objects that look normal but behave in unusual way, such as people that cycle or skateboard across the walkway or walk in the lawn. Abnormalities from the UCSD benchmark sets include also carts and wheelchairs. We emphasize that the training videos consist only of normal objects and actions, so that a model for abnormalities cannot be learned from it.

TABLE 2
Performance measures on the UCSD *ped2* dataset

	frame-wise		pixel-wise	
	AUC (%)	EER (%)	AUC (%)	RD (%)
Social force [43]	63	42	-	-
MPPCA [21]	77	30	-	-
Social force + MPPCA	71	36	-	-
Adam [20]	63	42	-	-
MDT [1]	85	25	-	-
H-MDT CRF [26]	-	18.5	-	70.1
SVP [39]	92	14	-	-
STVP	94.6	10.6	81.1	78.8

7.1.2 Evaluation Protocol

We use the standard protocol for evaluating abnormality detection results that was proposed by Mahadevan et al. [1]. The protocol consists of frame-wise and pixel-wise criteria. The frame-wise criterion labels a frame as abnormal if it contains at least one abnormal object detection. The localization accuracy of detected abnormalities is verified by the pixel-wise criterion that is more rigorous than the frame-wise criterion, since the detected abnormalities are compared to a pixel-level ground-truth mask. The pixel-wise criterion requires that at least 40% of all ground-truth abnormal pixels to be marked as abnormal in order to count a frame as true positive. By calculating the true positive rate (TPR) and false positive rate (FPR) at different detection thresholds we obtain the receiver operating characteristic (ROC).

Frame-wise and pixel-wise criteria use the area under the curve (AUC) as a performance measure calculated directly from the corresponding ROC curve. For the frame-wise criterion we calculate also the equal error rate (EER) as a value obtained when the false positive and false negative rates are equal. For pixel-wise criterion we compute the rate of detection (RD), that is equal to $1 - \text{EER}$. The pixel-wise criterion is applied on the partially labeled UCSD *ped1* dataset originally provided with the pixel-wise ground-truth annotation. Moreover, we also provide complete pixel-wise ground-truth annotations for the full datasets and evaluate thereon.

7.1.3 The Results of Evaluation

Fig. 9 compares the abnormality localization of our video parsing to the H-MDT CRF method [26] on UCSD *ped1* test videos. The first row shows a person riding a bike in a group of walking persons. In the second row there are three abnormalities in the scene: a person riding a bike, and two persons running along the walkway. The third row shows a person skateboarding along the walkway, and the fourth row shows an unusual object (car) in the scene. The columns show: (i) initial hypotheses of video parsing, (ii) hypotheses selected by video parsing, (iii) abnormality localization results of video parsing, (iv) abnormality localization results of H-MDT CRF method [26]. Due to our learned normal

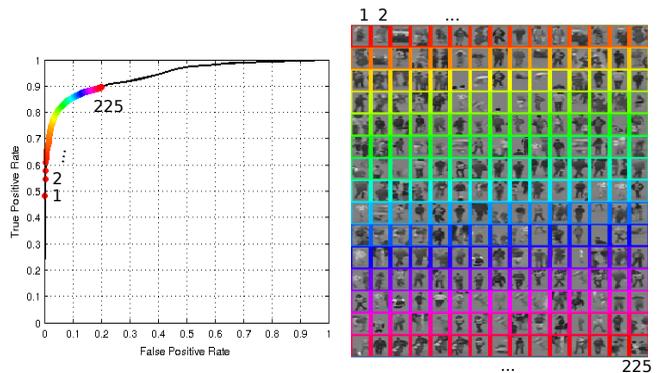


Fig. 11. Analysis of the false positive instances generated by our video parsing on the UCSD *ped1* dataset. Instances are sorted in the decreasing order of their abnormality score.

shape model used for explaining the foreground, we achieve better localization of the abnormalities in videos.

In Fig. 12 we show more examples of the video parsing on UCSD *ped1* test videos. Row 1 shows two persons skateboarding and cycling on a very crowded walkway, row 2 a skateboarder in a group of pedestrians, and row 3 two cyclists and a person walking across the walkway. By comparing the first two columns one can see that most hypotheses from the shortlist are discarded by video parsing because they get statistically explained away.

We also compare quantitatively our video parsing approach to the state-of-the-art methods on the challenging UCSD *ped1* and *ped2* benchmarks [1]. The methods used in our comparison are the mixture of dynamic textures (MDT) [1], H-MDT CRF [26], social force model (SF) [43], mixture of optical flow (MPPCA) [21], optical flow method (Adam et al.) [20], SF+MPPCA [1], sparse reconstruction (Sparse), local statistical aggregates (LSA) [30], and sparse combination learning (SCL) [31]. Our previous approach [39] which parses video frames individually, one after another, is denoted as sequential video parsing (SVP). We denote by STVP the full spatio-temporal video parsing proposed in this paper.

Our study shows that video parsing outperforms all other methods in experiments on both UCSD *ped1* and *ped2* datasets. Fig. 7 shows ROC curves for the frame-wise labeling of the UCSD *ped1* set. Tab. 1 gives the performance measures for the *ped1* dataset. We see that the inclusion of the temporal component and the improved inference enables spatio-temporal video parsing to improve upon our previous sequential video parsing by 2.9% in AUC and 5.1% in EER. From Tab. 1 we also see that our approach improves upon recently proposed powerful methods such as LSA [30] (1.2% gain in AUC and 3.1% in EER) as well as SCL [31] (2.1% gain in AUC and EER). All ROC plots for the pixel-wise labeling on *ped1* are shown in Fig. 8 b) and c). For the partial pixel-wise labeling of *ped1*, the spatio-temporal video parsing achieves an improvement of 4.7% AUC and 7.2% RD over the sequential video parsing.

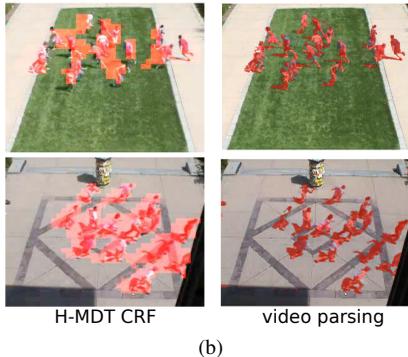
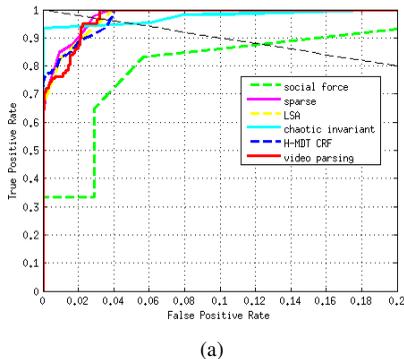


TABLE 3
Performance measures on
the UMN dataset

method	AUC (%)	EER (%)
chaotic invariants [46]	99.4	5.3
social force [43]	94.9	12.6
LSA [30]	99.5	3.4
H-MDT CRF [26]	99.5	3.7
Sparse [29] (scene1)	99.5	-
Sparse [29] (scene2)	97.5	-
Sparse [29] (scene3)	96.4	-
<i>STVP</i> (scene1)	99.5	3.2
<i>STVP</i> (scene2)	97.5	6.2
<i>STVP</i> (scene3)	99.9	1.5

Fig. 10. Abnormality detection on the UMN dataset. (a) ROC curves for frame-wise labeling. (b) Detection results of the H-MDT CRF [26] (left column) and video parsing (right column). Our approach exhibits competitive performance as can also be seen from the corresponding AUC and EER statistics in Tab. 3.

We outperform the closest competitor (HMDT CRT [26]) by 14.1% in AUC and 10.4% in RD. For the full pixel-wise labeling of *ped1*, we achieve an improvement of 2.5% in RD over the sequential video parsing. The competing HMDT CRF [26] method we outperform in this case by 1.5% in AUC and 5.0% in RD.

The ROC curves for the frame-wise labeling of UCSD *ped2* are given in Fig. 8 a). The numerical results are given in Tab. 2. We observe an improvement in performance of spatio-temporal parsing over sequential parsing by 2.6% in AUC and 3.4% in EER. The best method so far, MHDT CRF [26], we improve upon by 6.9% in EER. For the pixel-wise labeling of *ped2* dataset, we outperform the competing HMDT CRF method by 8.7% RD (AUC values for HMDT CRF are not provided in [26]). Overall we see that our spatio-temporal reasoning and the convex optimization based inference yield a significant improvement over the state-of-the-art.

Due to temporal grouping of per-frame object candidates (Sect. 6), spatio-temporal video parsing requires significantly less hypotheses (only about a hundred for the whole spatio-temporal domain) than sequential video parsing [39], which needs the same number of hypotheses for representing single frames. Since there remain fewer hypotheses to process, spatio-temporal video parsing takes less time to execute than sequential video parsing. Our non-optimized Matlab implementation on a Dual-Core 2.7GHz CPU runs at about 1 fps, whereas our previous sequential video parsing took 5-10 secs per frame. This is on par with recent H-MDT CRF [26] and Sparse [30] methods, with a notable exception of extremely fast SCL method [31].

7.1.4 Analysis of False Detections

To get a full understanding of the detection performance of proposed video parsing, we analyze the false detections on the UCSD *ped1* dataset. In Fig. 11 we see the first 225 false detections sorted in the decreasing order of their probability of abnormality. We observe several reasons for false detections: i) In many cases, false detections appear as a result of artifacts in the foreground segmentation. In

such cases, wrongly segmented pixels cannot be explained by the learned shape model and thus they are classified as abnormal. ii) Large variability of the normal human gait can sometimes be interpreted in video parsing as abnormal (e.g. running vs. fast walking). iii) Seldom errors in the provided video annotation cause that correctly detected abnormalities are sometimes considered as false (e.g. cars or running persons in Fig. 11). iv) When the true-positive hypothesis is missing from the shortlist due to a non-maximal recall, video parser can select an incorrect hypothesis as a next best fit.

7.2 Evaluation on the UMN Anomaly dataset

We additionally evaluate our video parsing on the UMN dataset that is widely used for benchmarking abnormality detection. The UMN dataset consists of three scenes in which periods of normal activity are followed by periods of emergency that are staged by people in the scene. In normal cases people are walking around alone or in groups. However, in emergency cases people start in panic to run away. For each scene several normal and abnormal events are happening one after another. In scene one, two and three there are two, six and three abnormal events, respectively. The dataset does not provide pixel-wise ground-truth abnormality maps, so we follow the standard protocol for this dataset and evaluate the detection results only in a frame-wise manner. Fig. 10 a) shows ROC curves for the frame-wise labeling. The performance measures AUC and EER are given in Tab. 3. For scene one, our performance is on par with the best competing methods in terms of AUC (99.5%) and EER(3.2%). For scene two we achieve 97.5% AUC that is equal to the best performing method (Sparse [29]). For the scene three we achieve 99.9% AUC that improves upon the best competitor (Sparse [29]) by 3.5%. A qualitative comparison of our method to HMDT CRF [26] on two frames is shown in Fig. 10 b). We see that our method achieves best localization of abnormalities that is consistent with findings from earlier experiments on UCSD *ped1* and *ped2*.

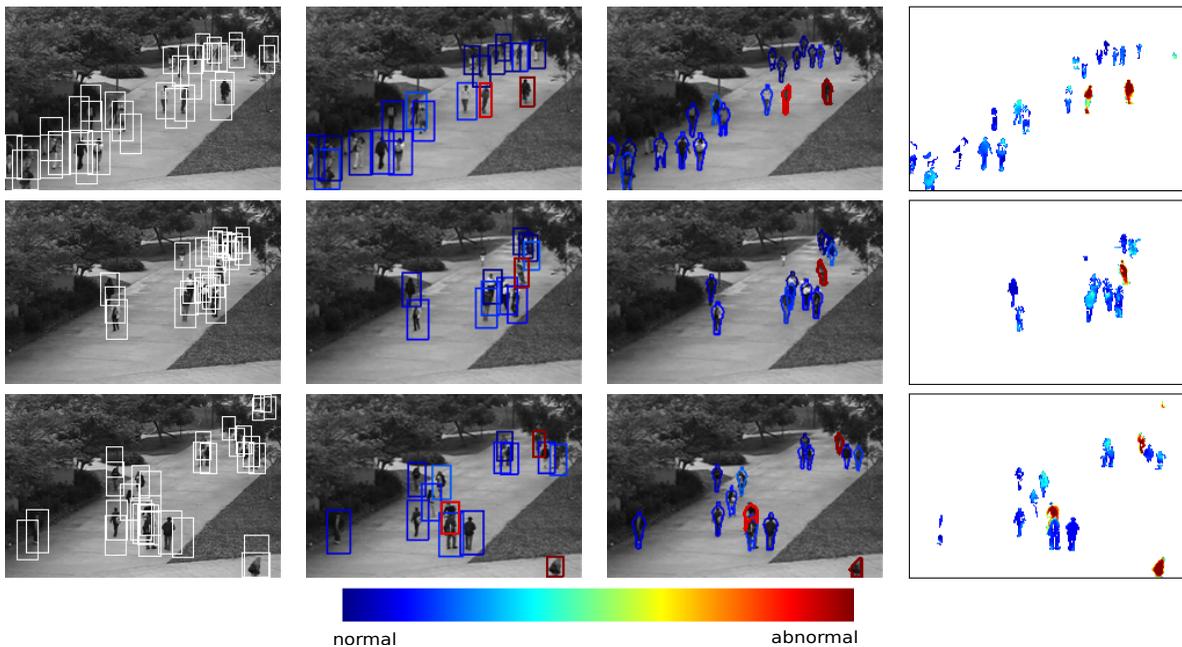


Fig. 12. Additional results of video parsing on the UCSD *ped1* dataset. Rows correspond to different examples. The first, third and fourth column correspond to the first three columns of Fig. 9. The second column shows hypotheses that are selected from the shortlist by video parsing. Other hypotheses are discarded by explaining away using the selected hypotheses. Best viewed in color.

8 CONCLUSION

In this paper we have framed abnormality detection as spatio-temporal video parsing to circumvent the ill-posed problem of directly searching for individual abnormal local image regions. We detect abnormalities by searching for a set of spatio-temporal object hypotheses that jointly explain the video foreground and which are themselves explained by normal training samples. In video parsing we do not independently detect individual hypotheses, but their joint layout that collectively describes the objects in the scene. We use MAP inference in a graphical model to effectively localize abnormalities in video and solve it as a convex optimization problem. We have evaluated our approach on several challenging datasets, which show that video parsing advances the state-of-the-art both in terms of abnormality classification and localization.

APPENDIX A PROOF OF LEMMA 4.1

Proof: The hypotheses explanation $J_h(\{o_h, m_h\}_h)$ (Eq. 10) can be written as follows,

$$\begin{aligned}
 J_h(\{o_h, m_h\}_h) &= \sum_h \left\{ -(1 - o_h) \log P(o_h = 0 | d_h) \right. \\
 &\quad - o_h \log P(o_h = 1 | d_h) + o_h \cdot \log Z(d_h) \\
 &\quad + \sum_{k=1}^M \underbrace{o_h \cdot \mathbf{1}[m_h = k]}_{=z_{h,k}} \cdot \left(\beta \Delta(d_h, d_k) \right. \\
 &\quad \left. \left. - \log P_k^{loc}(l_h^t) - \log P_k^{vel}(l_h^t - l_h^{t-1}) \right) \right\}.
 \end{aligned}$$

By replacing o_h with the sum from Eq. 17, we see that the hypotheses explanation term $J_h(\{o_h, m_h\}_h)$ can be expressed as a linear function of the parsing indicator \mathbf{z} ,

$$J_h(\mathbf{z}) = \mathbf{b}^\top \mathbf{z} + b_0,$$

where the parameter vector $\mathbf{b} = \{b_{h,k}\}_{h,k}$ and scalar b_0 are defined in the following way,

$$\begin{aligned}
 b_{h,k} &= -\log P(o_h = 1 | d_h) + \log P(o_h = 0 | d_h) \\
 &\quad + \log Z(d_h) + \beta \Delta(d_h, d_k) - \log P_k^{loc}(l_h) \\
 &\quad - \log P_k^{vel}(l_h^t - l_h^{t-1}) \\
 b_0 &= -\sum_h \log P(o_h = 0 | d_h),
 \end{aligned}$$

and they do not depend on the parsing indicator \mathbf{z} . \square

APPENDIX B PROOF OF LEMMA 4.2

Proof: The second derivative of the function $\Phi_{f_j^t}(x)$, $x > 0$ is given as follows,

$$\Phi_{f_j^t}''(x) = f_j^t \cdot \frac{e^{-x}}{(1 - e^{-x})^2}.$$

We see that the second derivative is positive, $\Phi_{f_j^t}''(x) > 0$, if the parameter f_j^t is positive, $f_j^t > 0$, so in this case the function $\Phi_{f_j^t}(x)$ is strictly convex. If the parameter f_j^t equals zero, $f_j^t = 0$, the function $\Phi_{f_j^t}(x)$ is linear, $\Phi_{f_j^t}(x) = x$, and therefore convex as well. \square

APPENDIX C PROOF OF LEMMA 4.3

Proof: The foreground explanation $J_j(\{o_h, m_h\}_h)$ depends on all hypotheses that cover pixel j ,

$$\begin{aligned} J_j(\{o_h, m_h\}_h) &= \\ &\sum_j \left\{ -(1 - f_j^t) \log P(f_j^t = 0 | \{o_h, m_h, l_h\}_h) \right. \\ &\quad \left. - f_j^t \cdot \log(1 - P(f_j^t = 0 | \{o_h, m_h, l_h\}_h)) \right\} \\ &= \sum_j \Phi_{f_j^t}(-\log P(f_j^t = 0 | \{o_h, m_h, l_h\}_h)). \end{aligned}$$

The argument of the function $\Phi_{f_j^t}(\cdot)$ in the last equation is bilinear in the parsing indicator \mathbf{z} (Eq. 11) and the joint shape prototype vector \mathbf{w} (Eq. 14),

$$\begin{aligned} &-\log P(f_j^t = 0 | \{o_h, m_h, l_h\}_h) \\ &= -\log(1 - P_0) - \sum_h \log(1 - P(f_j^t = 1 | o_h, m_h, l_h)) \\ &= -\log(1 - P_0) - \sum_h \sum_k \underbrace{o_h \cdot \mathbf{1}[m_h = k]}_{=: z_{h,k}} \cdot \mathbf{1}[j \in \mathcal{S}_h^t] \\ &\quad \cdot \sum_{j'} \mathbf{1}[l_j^t = s_h^t \cdot l_{j'}^t + (x_h^t \ y_h^t)^\top] \cdot \underbrace{\log P_k(f_{j'}^t = 0)}_{=: -\mathbf{w}_{k,j'}} \\ &= \mathbf{w}^\top \mathbf{C}_j \mathbf{z} + c_0, \end{aligned}$$

where \mathbf{C}_j is a sparse matrix with following elements,

$$\mathbf{C}_j(k, j'; h, k) = \mathbf{1}[j \in \mathcal{S}_h^t] \cdot \mathbf{1}[l_j^t = s_h^t \cdot l_{j'}^t + (x_h^t \ y_h^t)^\top],$$

and the scalar c_0 has the value $c_0 = -\log(1 - P_0)$.

Thus, the foreground explanation term $J_j(\{o_h, m_h\}_h)$ can be written as

$$J_j(\mathbf{z}, \mathbf{w}) = \sum_j \Phi_{f_j^t}(\mathbf{w}^\top \mathbf{C}_j \mathbf{z} + c_0). \quad \square$$

APPENDIX D PROOF OF LEMMA 4.4

Proof: The expression for the first derivative of the function $\Phi_{f_j^t}(x)$ is

$$\Phi'_{f_j^t}(x) = 1 - f_j^t \cdot \frac{1}{1 - e^{-x}},$$

The absolute difference of the first derivative of function $\Phi_{f_j^t}(x)$ evaluated in points $x_1, x_2 \geq c_0 = -\log(1 - P_0)$ is upper bounded in the following way,

$$\begin{aligned} |\Phi'_{f_j^t}(x_1) - \Phi'_{f_j^t}(x_2)| &= f_j^t \cdot \left| \frac{1}{1 - e^{-x_1}} - \frac{1}{1 - e^{-x_2}} \right| \\ &= f_j^t \frac{|e^{-x_1} - e^{-x_2}|}{(1 - e^{-x_1})(1 - e^{-x_2})} \leq f_j^t \frac{1}{P_0^2} |e^{-x_1} - e^{-x_2}| \\ &= f_j^t \frac{1}{P_0^2} \cdot e^{-\min\{x_1, x_2\}} \cdot (1 - e^{-|x_1 - x_2|}) \\ &\leq f_j^t \frac{1 - P_0}{P_0^2} \cdot |x_1 - x_2| = \rho |x_1 - x_2|. \end{aligned}$$

In the last line of the proof we used the inequality $1 - e^{-x} \leq x, \forall x > 0$. \square

REFERENCES

- [1] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010.
- [2] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vision*, vol. 74, no. 1, pp. 17–31, Aug. 2007.
- [3] T. Xiang and S. Gong, "Video behaviour profiling and abnormality detection without manual labelling," in *ICCV*, 2005, pp. 1238–1245.
- [4] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *CVPR*, 2004, pp. 819–826.
- [5] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *NIPS*, 2009, pp. 2080–2088.
- [6] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in *IEEE Computer Vision and Pattern Recognition*, 2006, pp. I: 594–601.
- [7] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 7, pp. 1198–1211, 2008.
- [8] 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*, vol. 83, pp. 57–71, June 2009.
- [9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
- [10] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *ICPR*, 2004.
- [11] F. Jiang, J. Yuan, S. A. Tsafaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.
- [12] M. Javan Roshkhari and M. D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," *Comput. Vis. Image Underst.*, vol. 117, no. 10, pp. 1436–1452, Oct. 2013.
- [13] R. Schuster, S. Schuster, G. Poier, M. Hirzer, J. A. Birchbauer, P. M. Roth, H. Bischof, M. Winter, and P. Schallauer, "Multi-cue learning and visualization of unusual events," in *ICCV Workshops*, 2011, pp. 1933–1940.
- [14] H. Dee and D. Hogg, "Detecting inexplicable behaviour," in *BMVC*, 2004, pp. 477–486.
- [15] T. Xiang and S. Gong, "Incremental and adaptive abnormal behaviour detection," *CVIU*, 2008.
- [16] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *CVPR*, 2008, pp. 1–8.
- [17] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical bayesian models," in *CVPR*, 2007, p. 45.
- [18] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," *J. Sel. Topics Signal Processing*, vol. 7, no. 1, pp. 91–101, 2013.
- [19] W. Yang, Y. Gao, and L. Cao, "Trasmil: A local anomaly detection framework based on trajectory segmentation and multi-instance learning," *Comput. Vis. Image Underst.*, vol. 117, no. 10, pp. 1273–1286, Oct. 2013.
- [20] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *PAMI*, vol. 30, pp. 555–560, 2008.
- [21] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," in *CVPR*, 2009, pp. 2921–2928.
- [22] C. C. Loy, T. Xiang, and S. Gong, "Stream-based active unusual event detection," in *ACCV*, 2010.
- [23] X. Wang, X. Ma, and W. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *PAMI*, vol. 31, pp. 539–555, 2009.
- [24] T. Hospedales, S. Gong, and T. Xiang, "A markov clustering topic model for mining behaviour in video," in *ICCV*, 2009.
- [25] T. M. Hospedales, J. Li, S. Gong, and T. Xiang, "Identifying rare and subtle behaviors: A weakly supervised joint topic model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2451–2464, 2011.
- [26] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, p. 1, 2013.

- [27] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1446–1453, 2009.
- [28] Y. Benezeth, P.-M. Jodoin, and V. Saligrama, "Abnormality detection using low-level co-occurring events," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 423–431, 2011.
- [29] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR*. IEEE, 2011, pp. 3449–3456.
- [30] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *CVPR*. IEEE, 2012, pp. 2112–2119.
- [31] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *International Conference on Computer Vision (ICCV)*, 2013.
- [32] Z. Tu, X. Chen, A. L. Yuille, and S.-c. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 113–140, 2005.
- [33] N. Ahuja and S. Todorovic, "Connected segmentation tree: A joint representation of region layout and hierarchy," in *CVPR*, 2008, pp. 1–8.
- [34] I. Kokkinos and A. Yuille, "Hop: Hierarchical object parsing," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 802–809, 2009.
- [35] S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 1–8, 2007.
- [36] A. Monroy and B. Ommer, "Beyond bounding-boxes: Learning object shape by model-driven grouping," in *Computer Vision–ECCV 2012*. Springer Berlin Heidelberg, 2012, pp. 580–593.
- [37] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *CVPR*, 2009, pp. 1972–1979.
- [38] T. Malisiewicz and A. A. Efros, "Beyond categories: The visual memex model for reasoning about object relationships," in *NIPS*, 2009.
- [39] B. Antic and B. Ommer, "Video parsing for abnormality detection," in *ICCV*, 2011, pp. 2415–2422.
- [40] P. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ser. Springer Optimization and Its Applications. Springer New York, 2011, pp. 185–212.
- [41] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l_1 -ball for learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 272–279.
- [42] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [43] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," *CVPR*, pp. 935–942, 2009.
- [44] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [45] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss, "Human-assisted motion annotation," in *CVPR*, 2008, pp. 1–8.
- [46] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *CVPR*, 2010, pp. 2054–2060.