Anomaly Detection in Crowded Scenes

Vijay Mahadevan Weixin Li Viral Bhalodia Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{vmahadev,wel017,vbhalodi,nuno}@ucsd.edu

Abstract

A novel framework for anomaly detection in crowded scenes is presented. Three properties are identified as important for the design of a localized video representation suitable for anomaly detection in such scenes: 1) joint modeling of appearance and dynamics of the scene, and the abilities to detect 2) temporal, and 3) spatial abnormalities. The model for normal crowd behavior is based on mixtures of dynamic textures and outliers under this model are labeled as anomalies. Temporal anomalies are equated to events of low-probability, while spatial anomalies are handled using discriminant saliency. An experimental evaluation is conducted with a new dataset of crowded scenes, composed of 100 video sequences and five well defined abnormality categories. The proposed representation is shown to outperform various state of the art anomaly detection techniques.

1. Introduction

There has recently been interest within computer vision in the analysis of densely crowded environments. Problems such as segmenting video into crowd components [3], estimating crowd size [17], determining the goal of individuals within a crowd [4] have all been subjects of research. Most of these efforts are motivated by the ubiquity of surveillance cameras, the challenges of crowd modeling, and the importance of crowd monitoring for various applications. In many of these, the goal is not so much to analyze normal crowd behavior, but to detect deviations from it. These are referred to as *anomalous* or *abnormal* events.

Anomaly detection is an active area of research on its own. Various approaches have been proposed, for both crowded and non-crowded scenes. They can be broadly categorized according to the type of scene representation adopted. One very popular category is based on trajectory modeling. It comprises tracking each object in the scene, and learning models for the resulting object tracks [5, 22, 24]. Both operations are quite difficult on densely crowded scenes, for which these approaches are not very promising.

Various authors have proposed alternative motion representations that avoid tracking. The most popular is dense optical flow, or some other form of spatio-temporal gradients [2, 16, 21]. Adam et al. [2] maintain probabilities of optical flow in local regions, using histograms. Kim and Grauman [16] model local optical flow patterns with a mixture of probabilistic PCA models, and enforce global consistency using a Markov Random Field (MRF). Mehran et al. [21] draw inspiration from classical studies of crowd behavior [13], that characterize crowd behavior using concepts such as *social force*. These concepts inspire optic flow measures of interaction within crowds, which are combined with a latent Dirichlet allocation (LDA) model for anomaly detection.

All these approaches focus uniquely on motion information, ignoring abnormality information due to variations of object appearance. This makes them impervious to abnormalities that do not involve motion outliers, e.g. a truck that crosses a bridge with weight restrictions. Furthermore, descriptors such as optical flow, pixel change histograms, or other traditional background subtraction operations, are difficult for crowded scenes, where the background is by definition dynamic, there are lots of clutter, and complicated occlusions. More complete representations, that account for both appearance and motion, have also been proposed. Boiman and Irani [6] use spatio-temporal patches and declare regions that cannot be reconstructed using data from previous frames as abnormal. Spatio-temporal gradients have been proposed in [18], where their statistics are modeled with a coupled HMM to detect abnormalities in densely crowded scenes.

Overall, there is a great diversity of approaches to abnormality detection. In general, it is quite difficult to compare two different solutions. Different representations of motion and appearance are combined with different graphical models for abnormality detection, which are typically tailored to the type of video analyzed, or a specific scene domain. Abnormalities are themselves defined in a somewhat subjective form, sometimes according to what the algorithms can detect. In some cases, different authors define different abnormalities on common datasets. Experiments are presented in datasets of very different characteristics (e.g. a traffic intersection vs a subway entrance), frequently proprietary, and with widely varying levels of crowd density.

In this work we make various contributions that address these problems. We concentrate on the issue of *representation*, namely how to design *localized video representations that enable anomaly detection in crowded scenes*. By definition, this precludes any form of global statistical inference, using MRFs, LDA, or any such models: while these can certainly improve performance, they tend to mask the limitations of the underlying visual representation. We identify three properties that the representation must have: 1) *jointly model appearance and dynamics of crowd patterns*, 2) ability to detect *temporal*, and 3) *spatial abnormalities*. We then propose the use of representations based on dynamic textures (DTs) [10]. These are joint models of appearance and dynamics, which have been shown very effective in modeling complex dynamic scenes.

As is common in the literature, we equate anomalies to *events of low-probability with respect to a model of normal crowd behavior*. We then introduce DT-based models of *normalcy* over both space and time. Temporal normalcy is modeled with a *mixture of DTs* [9] (MDT) and spatial normalcy is measured with a discriminant saliency detector [12] based on MDTs. These models *generalize* some of the approaches previously proposed to detect abnormalities in either time or space, can be easily *integrated* into a common solution, and are shown to perform well. The evaluation is based on a new dataset of crowded scenes, which is made available to the vision community. This dataset contains video of the walkways of a college campus, and crowds with naturally varying densities. It contains 100 video sequences, and a set of 5 well defined abnormality categories. These are not "synthetic", or "staged", but abnormal events that occur naturally, e.g. bicycle riders that cross pedestrian walkways. Ground truth is provided for abnormal events, as well as a protocol to evaluate detection performance. Finally the proposed abnormality detection algorithm is tested against previous approaches, establishing a set of benchmarks against which future algorithms can be compared.

2. Abnormality detection

Abnormality detection is usually formalized as an outlier detection problem. Some measurement $\mathbf{Y}$ is made, and a statistical model $P_{\mathbf{Y}}(\mathbf{y})$ is postulated for the distribution of $\mathbf{Y}$ under *normal* conditions. Abnormalities are defined as measurements whose probability is below a certain threshold under this model.

In this work, we consider the problem of *abnormality detection from localized measurements* $\mathbf{y}$ *of crowd video*. These are usually spatio-temporal patches of small dimension. A model of *normal* crowd behavior for such measurements must account for two types of normalcy, which we denote as *temporal* and *spatial*. The former reflects the intuition that normal events are *recurrent* over time. For example, cars in a highway move with a certain orientation and speed. The fact that there is no traffic at night, should not lead an anomaly detector to declare a large number of anomalies when it resumes in the morning. In this sense, a detector of temporal normalcy can be equated to a *background subtraction* algorithm in computer vision [23]. The model of normal behavior is built (and updated) over time, and all measurements that it cannot explain are denoted as *temporal abnormalities*.

Many events which would not be considered abnormal per se are abnormal *within* a crowd. This is because the crowd places constraints on individual motion, and motion patterns that would be feasible in isolation have low probability in the crowd context. For example, while there is nothing abnormal about an ambulance that rides at 50mph in a stretch of highway, the same observation within a highly *congested* highway indicates an abnormality. Note that the only indication of abnormality is the *difference* between the dynamics of the crowd and the object *at the time of the observation* and not the fact that the ambulance is moving at 50mph. Since the detection of such anomalies is mostly based on spatial processing, they are denoted as *spatial*. Their detection can be equated to the problem of *saliency detection* in computer vision [15].

While both background subtraction and saliency detection are extensively studied topics in vision, the vast majority of existing algorithms are not applicable to crowded scenes. In such scenes, where the background (or spatial surround) is highly dynamic, it is not sufficient to detect variations of image intensity, or even optical flow. Instead, the normalcy models must rely on sophisticated *joint representations of appearance and dynamics*. Even models such as the DT can be ineffective. Because crowded scenes are typically composed of distinct sub-entities - e.g. vehicles or groups of people moving in different directions - accurate detection requires the ability to model *multiple components of different appearance and dynamics*. One model that has been shown successful in this regard is the mixture of DTs of [9]. This is the representation adopted for all video analysis in this work. We next review the MDT and describe its proposed application to the design of the spatial and temporal components of an abnormality detector.

3. The mixtures of dynamic textures

The MDT [9] treats the observed video sequence $\mathbf{y}_1^\tau = \begin{bmatrix} y_1 & \cdots & y_\tau \end{bmatrix}^T$ as a sample from one of K dynamic textures. The probability of a sequence $\mathbf{y}_1^\tau$ under this model is given by

$$p(\mathbf{y}_1^\tau) = \sum_{i=1}^{K} \pi_i p(\mathbf{y}_1^\tau | z = i) \tag{1}$$

where $p(y_1^\tau | z = i)$ is the class conditional distribution of the i^{th} dynamic texture and π_i its prior probability. The generative model for the MDT is

$$\begin{aligned} x_{t+1} &= A_z x_t + v_t \\ y_t &= C_z x_t + w_t. \end{aligned} \tag{2}$$

where $z \sim$ multinomial$(\pi_1, \cdots, \pi_K)$ indexes mixture components, from which the observations are drawn. x_t is a hidden state variable, and y_t the observed video measurement. For each component, A_z, C_z are the transition and observation matrices respectively, the initial condition is given by $x_1 \sim \mathcal{N}(\mu_z, S_z)$, and the noise processes by $v_t \sim \mathcal{N}(0, Q_z)$ and $w_t \sim \mathcal{N}(0, R_z)$. The parameters of the model are learned by maximum likelihood, from a collection of spatio-temporal video patches. This is done with the expectation-maximization (EM) algorithm described in [9].

4. Temporal Abnormality Detection

Temporal abnormality detection is inspired by the popular background subtraction method of [23]. This method relies on a Gaussian mixture (GMM) at each image location, to model the local distribution of image intensities. Observations of low probability under this GMM are declared foreground. For abnormality detection, the GMM is replaced by a MDT, and the pixel-wise grid is replaced by one with a displacement of size 4. Each grid location defines the center of a video cell, from which spatio-temporal patches are extracted, and a MDT is learned during a training phase. The cell dimensions are not crucially important, in this application we use video patches of size 13×13 and cells of size 41×41. The process is illustrated in Figure 1.

After this phase, patches of low probability under the cell MDT are considered abnormalities. Given a patch $\mathbf{y}_1^\tau$, the hidden state sequence, $\mathbf{x}_1^\tau$ under this MDT model is estimated, and its log-likelihood under the mixture model $p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_1^\tau | \mathbf{y}_1^\tau)$ is computed with a Kalman smoothing filter [9]. The temporal abnormality map at location l is the negative log-likelihood of the state sequence estimated from the patch centered at l

$$\mathcal{A}_{temporal}(l) = -\log(p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_1^\tau | \mathbf{y}_1^\tau); \Theta_l). \tag{3}$$

We note that this can be seen as a generalization of the representation used by Kim and Grauman [16] which rely on a mixture of PCA models of optic flow. The matrix C_z of (2) is also a PCA basis for the patches assigned to the z^{th} mixture component. However, the PCA decomposition is

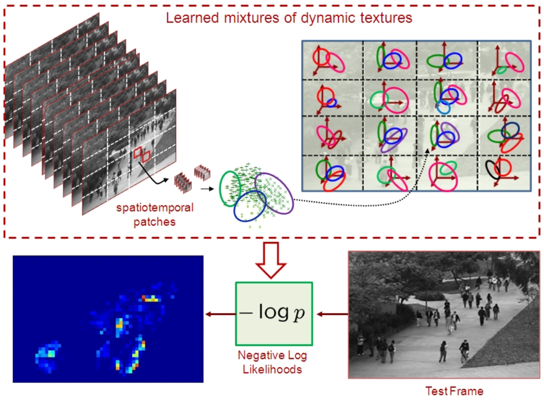


Figure 1. Learning MDTs for temporal abnormality detection. For each region of the scene, an MDT is learned during training. At test time, the negative log-likelihood of the spatiotemporal patch centered at location l is computed using the MDT whose region center is closest to l.

applied to *patch appearance*, not optic flow. The patch motion is captured by the hidden state sequence $\mathbf{x}_1^\tau$, which can be seen as a trajectory in PCA space. This implies that the representation is not *memoryless* as is the case for a mixture of optic flow. The ability to model appearance and the more sophisticated representation of dynamics, make the MDT a much more powerful representation than the mixture of PCA.

5. Spatial Abnormality Detection

Spatial abnormality detection is inspired by previous work on saliency detection in computer vision [12, 15]. Saliency is usually defined in a center-surround manner: salient locations are those with some attribute that makes them stand-out from their surround. Given an appropriate set of features, saliency provides an objective definition of *spatial anomaly*: spatially abnormal locations are those whose saliency is above some threshold. This ties the abnormality detection criteria to the criteria used to define saliency. In this work, we rely on the *discriminant saliency* criteria of [12].

Discriminant saliency formulates the saliency problem as a hypothesis test between two classes: a class of *salient stimuli*, and a *background* class, consisting of stimuli that are not salient. At each location l in the scene, two windows are defined: a *center window* $\mathcal{W}_l^1$, with label $C(l) = 1$, containing the location, and a surrounding annular window $\mathcal{W}_l^0$, with label $C(l) = 0$, containing *background*. A set of features $\mathbf{Y}$ from a predefined feature space $\mathcal{Y}$ (e.g. raw pixel values, Gabor, DCT, wavelet, or SIFT features), are computed for each of the windows $\mathcal{W}_l^c, c \in \{0, 1\}$. Given class-conditional feature densities $p_{\mathbf{Y}|C(l)}(\mathbf{y}|c)$, the *saliency* of location l, $S(l)$, is defined as the extent to which the feature $\mathbf{Y}$ can discriminate between the two classes. This is quantified by the mutual information (MI) between feature responses, $\mathbf{Y}$, and class label, C [12]

$$S(l) = \sum_{c=0}^{1} p_{C(l)}(c) KL[p_{\mathbf{Y}|C(l)}(\mathbf{y}|c) || p_{\mathbf{Y}}(\mathbf{y})] \quad (4)$$

where, $\mathrm{KL}\left(p \,\|q\right) = \int_{\mathcal{X}} p_{\mathbf{Y}}(y) \log \frac{p_{\mathbf{Y}}(y)}{q_{\mathbf{Y}}(y)} dy$ is the Kullback-Leibler (KL) divergence between the probability distributions $p_{\mathbf{Y}}(y)$ and $q_{\mathbf{Y}}(y)$ [19].

For a given feature $\mathbf{Y}$, locations of maximal saliency are those where the distinction between center and surround can be made with *highest confidence*, i.e. the MI above is maximal. Discriminant saliency can be combined with many features [11]. When $\mathbf{Y}$ consists of optical flow features, it is similar to the social force model of [21]. Under this model, saliency is defined as the difference between the optical flow at a location and the average optical flow in its neighborhood (see equation (8) of [21]). This is a simplified form of center-surround saliency, which 1) replaces the MI between features and class label by a difference to the mean background response, 2) relies on a coarse representation of dynamics based uniquely on optic flow, and 3) ignores appearance features.

Mahadevan and Vasconcelos [20] proposed the use of DTs with discriminant saliency in the context of background subtraction [20]. While this method relies on a more powerful representation of appearance and dynamics than the social force model, it is not sufficient to solve the anomaly detection problem. In the context of crowded scenes, abnormality detection requires the analysis of foreground regions, and the ability to account for diverse foregrounds. The background subtraction method of [20], which uses a single DT for both the center and surround windows, is not adequate for this purpose, producing a large number of false positives.

5.1. Center Surround Saliency with the MDT

In this work, we the adopt the MDT model of [9] as the probability distribution $p_{\mathbf{Y}|C(l)}(\mathbf{y}_1^\tau|c)$ from which spatio-temporal patches $\mathbf{y}_1^\tau$ are drawn. We start from the property that, under assumptions of Gaussian initial conditions and noise, spatio-temporal patches $\mathbf{y}_1^\tau$ drawn from a DT have a Gaussian probability distribution [8],

$$p_{\mathbf{Y}}(\mathbf{y}_1^\tau) \sim \mathcal{N}(\boldsymbol{\gamma}, \boldsymbol{\Phi}). \quad (5)$$

Assuming that the class-conditional distributions of classes $c \in \{0, 1\}$ (corresponding to center and surround) are mixtures of K_c DTs, it follows that

$$\begin{aligned} p_{\mathbf{Y}|C(l)}(\mathbf{y}_1^\tau|c) &= \sum_{i=1}^{K_c} p^i_{\mathbf{Y}|C(l)}(\mathbf{y}_1^\tau|c) \\ &= \sum_{i=1}^{K_c} \pi_c^i \mathcal{N}(\boldsymbol{\gamma}_c^i, \boldsymbol{\Phi}_c^i) \end{aligned} \quad (6)$$

for $c \in \{0, 1\}$. The marginal distribution is given by,

$$p_{\mathbf{Y}}(\mathbf{y}_1^\tau) = \sum_{i=1}^{K} p^i_{\mathbf{Y}}(\mathbf{y}_1^\tau) = \sum_{i=1}^{K} \omega^i \mathcal{N}(\boldsymbol{\gamma}^i, \boldsymbol{\Phi}^i) \quad (7)$$

Hence, evaluation of the saliency measure of (4) requires evaluation of the KL divergence between (6) and (7). This is problematic, because there is no closed form solution for the KL divergence between two MDTs. However, since the probability distribution of each MDT component is a Gaussian, it is possible to rely on common approximations to the KL divergence between two Gaussian mixtures. In this work, we adopt the variational approximation proposed in [14]

$$\mathrm{KL}\left(p_{\mathbf{Y}|C} \,\|p_{\mathbf{Y}}\right) \approx \sum_i \pi_C^i \log \frac{\sum_j \pi_C^j e^{-\mathrm{KL}\left(p^i_{\mathbf{Y}|C} \| p^j_{\mathbf{Y}|C}\right)}}{\sum_j \omega^j e^{-\mathrm{KL}\left(p^i_{\mathbf{Y}|C} \| p^j_{\mathbf{Y}}\right)}} \quad (8)$$

Each term in the exponent of (8) is KL divergence between two DTs, and can be computed in closed-form [7]. For example, the expression for the terms in the denominator is

$$\begin{aligned} &\mathrm{KL}\left(p^i_{\mathbf{Y}|C} \,\middle\|\, p^j_{\mathbf{Y}}\right) \quad (9) \\ &= \frac{1}{2}\left[\log \frac{|\boldsymbol{\Phi}|}{|\boldsymbol{\Phi}_C^i|} + \mathrm{tr}\left({\boldsymbol{\Phi}^j}^{-1} \boldsymbol{\Phi}_C^i\right) + \|\boldsymbol{\gamma}_C^i - \boldsymbol{\gamma}^j\|^2_{\boldsymbol{\Phi}^\mathbf{j}} - m\tau\right] \end{aligned}$$

where m is the number of pixels in each frame, and $||\mathbf{z}||_\mathbf{A} = \mathbf{z}^T \mathbf{A}^{-1} \mathbf{z}$. The terms in the numerator are computed similarly. These computations can be performed using efficient recursions [7].

5.2. Abnormality Detection

The spatial abnormality map is produced by computing the saliency $S(l)$ at each location l of the input video. Given a location, center surround saliency requires 1) learning MDTs from the center and surround regions, and 2) computing a weighted average of these mixtures to obtain the marginal distribution. However, learning MDTs at each scene location is computationally infeasible. To overcome this problem, we adopt the following approximation. A batch of frames around the current frame, i.e. the 3D volume $\mathcal{V}(l)$ containing l, is selected and a dense collection of overlapping spatio-temporal patches extracted from $\mathcal{V}(l)$. A single MDT with K_{global} mixture components, denoted by $(\gamma^i_{global}, \boldsymbol{\Phi}^i_{global}), i \in \{1 \ldots K_{global}\}$ is learned for the entire patch collection. Each patch in the volume is then assigned to the mixture component of largest posterior probability. This produces a segmentation of the volume into superpixel type regions, as shown in Figure 2.

At each location l, the MDTs for center and surround classes, as well as the marginal distribution, share the components of the global mixture model. Only the mixing pro-

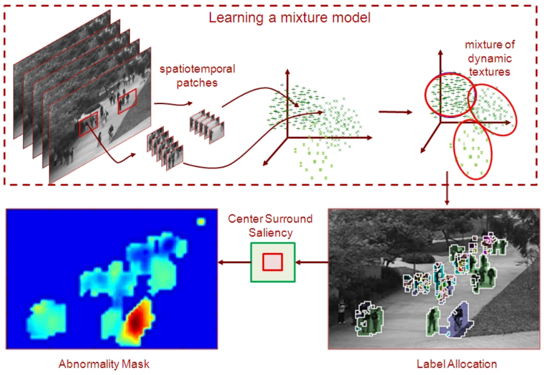


Figure 2. Illustration of spatial abnormality detection using center surround saliency with MDTs.

portions are recomputed, based on the ratio of pixels assigned to each component in the respective windows

$$p_{\mathbf{Y}|C(l)}(\mathbf{y}_1^\tau|c) = \sum_{i=1}^{K_{global}} \frac{\sum_{l\in\mathcal{W}_l^c}\mathcal{M}_{il}}{\sum_{l\in\mathcal{W}_l^c}1}\mathcal{N}(\gamma_g^i lobal, \mathbf{\Phi}_g^i lobal) \quad (10)$$

for $c \in \{0,1\}$, where $\mathcal{M}_{il} = 1$ if l is assigned to mixture component i and 0 otherwise.

The prior probabilities for center and surround, $p_C(c)$, are set according to the ratio of volumes of the center and surround windows. $S(l)$ is then computed with (4), using (8) and (9). Note that the KL divergence terms in (8) only require the computation of $\binom{K_{global}}{2}$ KL divergences between the K_{global} mixture components, and these need to be computed only once per frame. This procedure is repeated for every frame in the test video. The spatial abnormality map is then

$$\mathcal{A}_{saliency}(l) = S(l), \quad (11)$$

as illustrated in Figure 2.

The overall abnormality map is the sum of the normalized temporal and spatial abnormality maps of (3) and (11)

$$\mathcal{A}_{total}(l) = \mathcal{A}_{temporal}(l) + \mathcal{A}_{saliency}(l). \quad (12)$$

6. The crowd anomaly detection dataset

In addition to abnormality detection algorithms, this work contributes a dataset for the evaluation of abnormalities in crowded scenes. This dataset was acquired with a stationary camera mounted at an elevation, overlooking pedestrian walkways. The crowd density in the walkways was variable, ranging from sparse to very crowded. In the normal setting, the video contains only pedestrians. Abnormal events are due to either 1) the circulation of *non pedestrian* entities in the walkways, or 2) anomalous pedestrian motion patterns. Commonly occurring anomalies include bikers, skaters, small carts, and people walking across a walkway or in the grass that surrounds it. A few instances of people in wheelchair were also recorded. All abnormalities are naturally occurring, i.e. they were not staged for the purposes of assembling the dataset.

The data was split into 2 subsets, each corresponding to a different scene. The first scene (Figure 4), contains groups of people walking towards and away from the camera, and some amount of perspective distortion. The second contains scenes with pedestrian movement parallel to the camera plane. The video footage recorded from each scene was split into various clips of around 200 frames. For each clip, the groundtruth annotation includes a binary flag per frame, indicating whether an anomaly is present in that frame. In addition, a subset of 10 clips is provided with manually generated pixel-level binary masks, which identify the regions containing anomalies. This is intended to enable the evaluation of performance with respect to the ability to localize anomalies. The videos, and groundtruth annotations, are available online [1], for benchmarking of future anomaly detection algorithms.

6.1. Evaluation Procedure

Training sets - 34 clips for Peds1, 16 clips for Peds2 - are provided for learning of normalcy models. The test set (36 test clips for Peds1 and 14 test clips for Peds2) contains clips in which some of the frames have one or more anomalies present. The total number of anomalous frames (≈3400) is somewhat smaller than that of normal frames (≈5500). The task is to detect whether an anomaly is present, or not, in each frame of the test set. The evaluation has two components

- *Anomaly detection using frame level groundtruth:* Given some abnormality map, a suitable threshold is used to generate an abnormality mask. If a frame contains at least one abnormal pixel, it is considered a detection. These detections are compared to the frame level groundtruth annotation of each frame. The procedure is repeated for multiple thresholds, to determine an ROC curve. Note that this evaluation does not verify whether the detection coincides with the actual location of the anomaly. It is therefore possible for some portion true positive detections to be "lucky" co-occurrences of erroneous detections and abnormal events.
- *Anomaly localization using pixel level groundtruth:* To test localization accuracy, detections are compared to pixel level groundtruth masks, on a subset of ten clips. The procedure is similar to that described above. If at least 40% of the truly anomalous pixels are detected, the frame is considered detected correctly, and counted as a false positive otherwise. The ROC curve is based on these detection and false positive rates, for multiple threshold values.

For the anomaly detection component, the equal error rate (EER) - percentage of misclassified frames when the false positive rate is equal to the miss rate - is reported. For the anomaly localization component, detection rate at equal error is reported.

7. Experiments and Results

To evaluate the performance of the proposed anomaly representation, we compared it to three other recently proposed representations - the social force model [21] (denoted SF), the mixture of optical flow (denoted MPPCA) [16] and the optical flow monitoring method of [2]. Since code for all three was not available, the results presented reflect our own implementations. As this work addresses only the low level representation (viz. optical flow vs. dynamic textures), the Latent Dirichlet Allocation modeling of [21] and the MRF of [16] were omitted. Among the methods chosen for comparison, the social force model is a spatial anomaly detection technique, while the mixture of optical flow approach is based on temporal anomaly detection. Since the proposed MDT approach has both components, we also included the normalized combination of social force and mixture of optical flow (denoted SF-MPCCA) as a separate method in our comparisons.

7.1. Quantitative Performance Comparison

The four algorithms were run on all clips of the crowds dataset. Using a coarse threshold, abnormality maps were first thresholded to yield candidate locations. The resulting abnormality masks were passed through a spatio-temporal averaging filter to reduce noisy abnormal predictions. This can be seen as a simplified version of using a (computationally expensive) MRF model to enforce smoothness. This filtered output was subjected to the evaluation procedure of Section 6.1. The ROC curves for anomaly detection and anomaly localization are shown in Figure 3, and the EER values are tabulated in Table 1. Some examples of frames with anomalies detected by the proposed approach and the best performing competitor (SF-MPPCA) are shown in Figure 4. Video clips of the anomaly detections are available online [1].

Anomaly Detection Experiment: EER

	SF [21]	MPPCA [16]	SF-MPPCA	Adam et al. [2]	MDT
Ped1	31%	40%	32%	38%	25%
Ped2	42%	30%	36%	42%	25%
Average	37%	35%	34%	40%	25%

Anomaly Localization Experiment: Rate of Detection

Localization	21%	18%	28%	24%	45%

Table 1. Quantitative comparison of performance for the abnormality detection algorithms tested. The first two rows show the EER over the two datasets Ped1 and Ped2. The average over the two datasets is shown in the third row. The detection rate at equal error for the anomaly localization task is shown in the last row.

8. Discussion

The results show that the MDT-based anomaly detection outperforms all other approaches. The difference in performance is more pronounced in the anomaly localization task, indicating that the remaining approaches may be enjoying good detection rates in the anomaly detection task due to lucky hits. It is clear that, even when spatial and temporal detection schemes based on optical flow are combined (SF-MPPCA), they do not perform well. This suggests that optical flow representations are not powerful enough to detect anomalous occurrences in terms of joint appearance and motion. Furthermore, when there is perspective distortion, optical flow is unreliable and global comparisons of optical flow can lead to erroneous results. This is evident from the results of the social force model, shown in Figure 4 (b) and (c), where the regions of high optical flow at the near end of the camera show spurious abnormality detections. The main shortcoming of the proposed approach is the computation time. Training the mixtures of dynamic textures for videos of frame size 160×240 takes around 2hrs, while the testing time per frame is about 25secs on a standard Pentium machine with 3GHz CPU and 2GB RAM.

References

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

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *PAMI*, 30(3):555–560, March 2008.

[3] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, pages 1–6, 2007.

[4] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*, pages II: 1–14, 2008.

[5] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *CVPR*, pages 1–8, 2008.

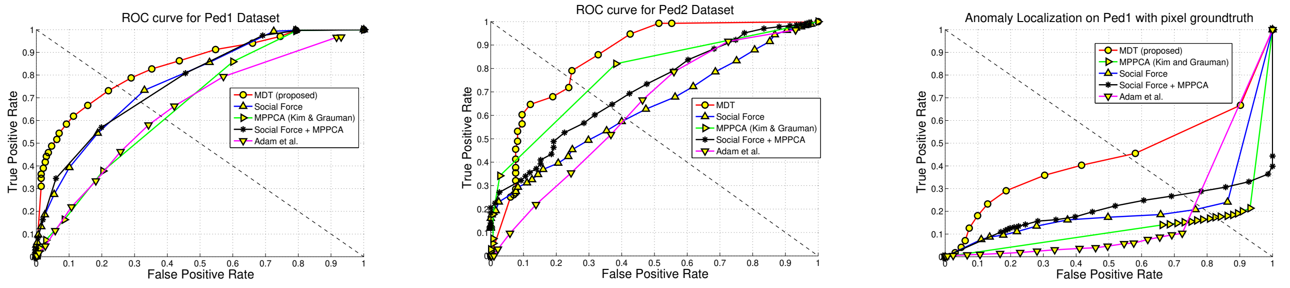
[6] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJCV*, 74(1):17–31, August 2007.

[7] A. B. Chan and N. Vasconcelos. Efficient computation of the kl divergence between dynamic textures. Technical Report SVCL-TR-2004-02, Dept. of ECE, UCSD, 2004.

[8] A. B. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *CVPR*, volume 1, pages 846–851, 2005.

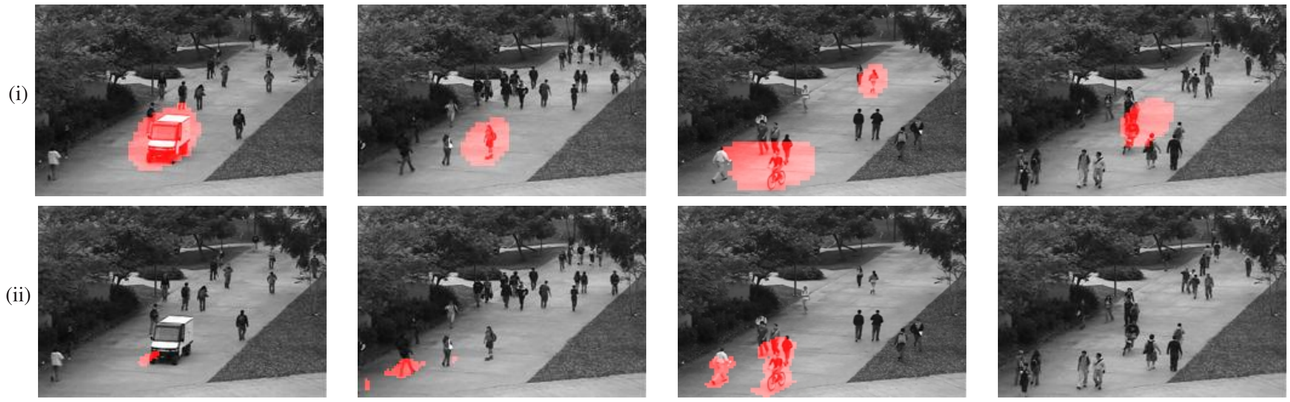
[9] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *PAMI*, 30(5):909–926, May 2008.

[10] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, 2003.



(a) (b) (c)

Figure 3. (a) and (b) Performance of the approaches tested for the anomaly detection task on the Pedestrians dataset.(c) Performance of the approaches tested on the anomaly localization with pixel level groundtruth on the Pedestrians dataset. Note that on this task, performance at chance level is *not the diagonal from* $(0, 0)$ *to* $(1, 1)$ (it is in fact close to zero)



(a) (b) (c) (d)

Figure 4. Examples of abnormal detections using (i) the MDT approach (ii) using the SF-MPPCA approach which completely misses the skater in (b), the person running in (c) and the bike in (d).

[11] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7):1–18, 6 2008.

[12] D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 21:239–271, Jan 2009.

[13] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, May 1995.

[14] J. R. Hershey and P. A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *ICASSP*, volume 4, pages IV–317–IV–320, 2007.

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998.

[16] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, pages 2921–2928, 2009.

[17] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In *BMVC*, 2005.

[18] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR09*, pages 1446–1453, 2009.

[19] S. Kullback. *Information Theory and Statistics*. Dover Publications, New York, 1968.

[20] V. Mahadevan and N. Vasconcelos. Background subtraction in highly dynamic scenes. *CVPR*, 1, 2008.

[21] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942, 2009.

[22] N. Siebel and S. Maybank. Fusion of multiple tracking algorithms for robust people tracking. In *ECCV*, page IV: 373 ff., 2002.

[23] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, volume 2, pages 2246–2252, 1999.

[24] T. Zhang, H. Lu, and S. Li. Learning semantic scene models by object classification and trajectory clustering. In *CVPR*, pages 1940–1947, 2009.