

Real-Time Anomaly Detection and Localization in Crowded Scenes

Mohammad Sabokrou¹, Mahmood Fathy², Mojtaba Hoseini¹, Reinhard Klette³

¹Malek Ashtar University of Technology, Tehran, Iran

²Iran University of Science and Technology, Tehran, Iran

³Auckland University of Technology, Auckland, New Zealand

Abstract

In this paper, we propose a method for real-time anomaly detection and localization in crowded scenes. Each video is defined as a set of non-overlapping cubic patches, and is described using two local and global descriptors. These descriptors capture the video properties from different aspects. By incorporating simple and cost-effective Gaussian classifiers, we can distinguish normal activities and anomalies in videos. The local and global features are based on structure similarity between adjacent patches and the features learned in an unsupervised way, using a sparse auto-encoder. Experimental results show that our algorithm is comparable to a state-of-the-art procedure on UCSD ped2 and UMN benchmarks, but even more time-efficient. The experiments confirm that our system can reliably detect and localize anomalies as soon as they happen in a video.

1. Introduction

The definition of an *anomaly* depends on what context is of interest. A video event is considered as being an anomaly if it is not very likely to occur in the video [6]. Describing unusual events in complex scenes is a cumbersome task, often solved by employing high-dimensional features and descriptors. Developing a reliable model to be trained with such descriptors is quite challenging and requires an enormous amount of training samples; it is also of large computational complexity. Therefore, this might face the so-called “curse of dimensionality”, in which the predictive power of the trained model reduces, as the dimensionality of the feature descriptors increases.

In recent work, one or a set of reference normal models are learned from training videos, which are then applied for detecting an anomaly in the test phase. Such methods usually consider a test video as being an anomaly if it does not resemble the learned model(s). In order to build these reference models, some specific feature descriptors should be used. In general, features usually are extracted to represent either (1) trajectories or (2) spatio-temporal changes. For

instance, [8] and [19] focus on the trajectories of objects in videos, in which each object is to be labeled as an anomaly or not, based on how they follow the learned normal trajectory. These methods could not handle the occlusion problem, and are also computationally very expensive, for the case of crowded scenes.

To overcome these weaknesses, researchers proposed methods using low-level features such as optical flow or gradients. They learn the shape and spatio-temporal relations using low-level features distributions. As an example, [13] fits a Gaussian mixture model as the features, while [1] uses an exponential distribution.

Clustering of test data using low-level features is exploited in [16]. In [2, 9, 10, 22], the normal patterns were fitted to a Markov random field, and [14, 18] apply latent Dirichlet allocations. [11] introduces a joint detector of temporal and spatial anomalies, where the authors use a mixture of dynamic textures (MDT) model.

In recent studies, sparse representations of events [6, 7, 12] in videos is being heavily explored. Notably, the proposed models in [6, 7, 11, 14, 15, 12] achieve favorable performance in anomaly detection, however they normally fail in the task of anomaly localization. All these methods, except [12], are not designed for real-time applications and commonly fail in real-world anomaly detection problems.

In this paper, we propose to represent videos from two different aspects or views, and thus two partially independent feature descriptors. Then, we introduce an approach for integrating these views in a testing step to simultaneously perform anomaly *detection* and *localization*, in real-time. Unlike previous work, instead of using low-level features, we propose to learn a set of representative features, based on auto-encoders [17].

Our detection framework identifies an anomaly in a real-time manner. Our anomaly detection method has high true-positive and low false-positive rates which make it quite reliable. We evaluate our anomaly detection and localization framework on popular datasets and report the running time for the whole procedure. The comparison with state-of-the-art methods shows the superiority of our method, both in

terms of performance and running time.

The main contributions of our work are as follows: (1) Presenting a feature learning procedure for describing videos for the task of video anomaly localization. This method is time-consuming for training, but the learned features are very discriminative to model the normal patches. (2) Introducing a descriptor-based similarity metric between adjacent patches for detecting sudden changes in spatio-temporal domains. (3) Representing video patches from two different aspects or views. Both local and global feature sets are used for each view. In the final decision, these views support each other. (4) Modeling all normal patches with Gaussian distributions. For a test video, the Mahalanobis distance is used to figure out its relevance for the normal patches. (5) Being real-time, we are able to detect and localize anomalies soon after they occur in a test video or stream.

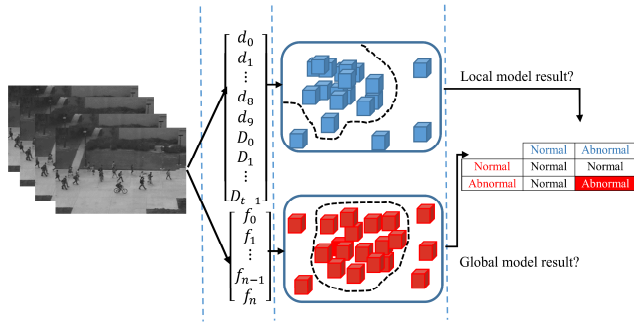


Figure 1. The scheme of our algorithm (left to right): Input frames, two views of patches (global and local), modeling the data using Gaussian distributions, and making the final decision

The overall scheme of our algorithm is shown in Fig. 1. We achieve 25 fps processing power, and with enduring some bit errors we reach up to 200 fps using a PC with 3.5 GHz CPU and 8G RAM in MATLAB 2012a.

The rest of the paper is organized as follows. The proposed approach is introduced in Section 2, where we first introduce the overall schema, and then we focus on global descriptors, local descriptors, anomaly classification scheme, and finally anomaly detection through feature learning, one after the other. Experimental results, comparisons, and analysis are presented in Section 3. Ultimately, Section 4 concludes the paper.

2. Proposed System

Overall Scheme. To represent each video, first each video is converted into a number of non-overlapping cubic patches; a sketch of this video representation is shown in Fig. 2. Generally, every video has one or a set of dominant events. Thus, one expects that *normal patches* have similar relations with their adjacent patches and a high likelihood of

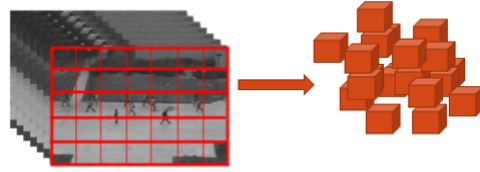


Figure 2. Video representation: Each video is represented through a number of non-overlapping cubic patches, covering the whole space-time in the video.

occurrence in the video. Therefore, these *anomaly patches* should meet three conditions:

1. The similarity between the anomaly patches and their adjacent (*i.e.*, defined by spatial changes) patches does not follow the same pattern as from normal patches to their adjacent patches.
2. It is most likely that the temporal changes of an anomaly patch would not follow the pattern in the temporal changes of normal patches.
3. It is obvious that the occurrence likelihood of an anomaly patch is less than that of normal patches.

It can be easily inferred that the above conditions 1 and 2 are characterized locally. Therefore, they can be encoded by local feature descriptors, and condition 3 is analogous to the global nature of the scene. In other words, conditions 1 and 2 consider the relation between a patch and its adjacent patches, and condition 3 describes the overall appearance of patches in the video. As a result, the first two conditions are corresponding to the spatio-temporal changes, while the latter one is different. Therefore, we model a combination of 1 and 2 through a *local* representation, and 3 by a more *global* one. On the other hand, in order to avoid the so-called “curse of dimensionality”, we model these two aspects independently.

So far, we have defined two different aspects that we approach the problem, leading to two independent models. In order to make a final decision, we aggregate the decisions from both models. If both models reject a patch it is considered to be an anomaly. This leads to a system with better performance in terms of true-positive and false-positive, since this way of combination of the two models guarantees a concrete selection of a patch as anomaly if both models agree on its being an anomaly.

In summary, the input videos are represented in two different aspects. Then, these representations are fitted to a set of Gaussian distributions and a decision boundary is calculated for each of them. Finally, based on global and local model results, a decision is reached about a patch being an anomaly or not (*detection*). The *localization* could be then easily inferred, based on which patches throughout the

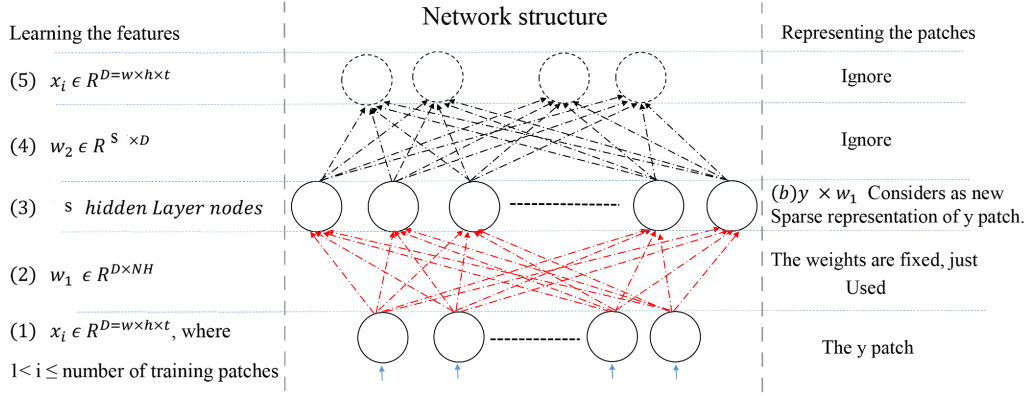


Figure 3. Summary for learning the global features using an auto-encoder. *Left*: The step for learning features uses raw normal patches; components (1), (2), (3), (4), and (5) are needed; the aim is to reconstruct the input paths with adjusting W_1 and W_2 using gradient descent. *Middle*: Auto-encoder structure. *Right*: Representing the y patch using the W_1 weights ($y \times W_1$); (1), (2), and (3) are just used; this is a multiplication of two matrixes, so it is very fast

video are classified as anomaly. In the subsequent sections, the two sets of features (global and local) are introduced.

Global descriptors. A video global descriptor is a set of features that describes the video as a whole and therefore is best able to describe the normal video patches. In [21] it is argued that classical handcrafted low-level features, such as HOG and HOF, may not be universally suitable and discriminative enough for every type of video. So, unlike previous works, that use low-level features, we use an unsupervised feature learning method based on auto-encoders. The structure of the auto-encoder is depicted in Fig. 3.

The auto-encoder learns sparse features based on gradient descent, by modeling a neural network. Suppose that we have m normal patches with the dimensions (w, h, t) , creating a data structure of $x_i \in \mathbb{R}^D$, $D = w \times h \times t$ (the raw data). The auto-encoder minimizes the objective defined in Eq. (1) by re-reconstructing the original raw data:

$$L = \frac{1}{m} \sum_{i=1}^m \|x_i - W_2 \delta(W_1 x_i + b_1) + b_2\|^2 + \sum_{i=1}^{w \cdot h \cdot t} \sum_{j=1}^s (W_{ji}^2) + \beta \sum_{j=1}^s KL(\rho \|\rho'_j) \quad (1)$$

where s is the number nodes in the auto-encoder's hidden layer, $W_1 \in \mathbb{R}^{s \times D}$ and $W_2 \in \mathbb{R}^{D \times s}$ are the weight matrices, which map the input layer nodes to hidden layer nodes, and hidden layer nodes to the output layer nodes, respectively. W_{ji} is the weight between the j^{th} hidden layer node and the i^{th} output layer node, and δ is equal to the sigmoid function. Furthermore, b_1 and b_2 , are the bias of the output layer and the hidden layer, respectively. $KL(\rho \|\rho'_j)$ is a regularization function and is set to enforce the activation of the

hidden layer to be sparse. KL is based on the similarity between a Bernoulli distribution with ρ as parameter, and the active node distribution. The parameter β is the weight of the penalty term (in the sparse auto-encoder objective). We can efficiently optimize the above objective with respect to W_1 via the stochastic gradient descent approach.

Local descriptors. To describe each video patch, we use a set of local features. The similarity between each patch and its neighboring patches are calculated. As for the neighbors, we consider nine spatial neighboring patches and one temporal neighboring patch (the one right behind the patch of interest when arranged temporally), yielding to 10 neighbors for each single patch. For temporal neighbors, we only consider the patch before the patch of interest (not the next one), as we aim to detect the anomaly soonest possible, even before the next video frames (and there-

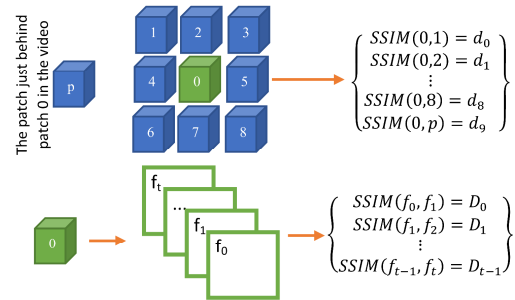


Figure 4. Illustration of our local descriptor: Similarities of each patch on interest with its neighboring patches (top), temporal inner similarities of each patch of interest (bottom).

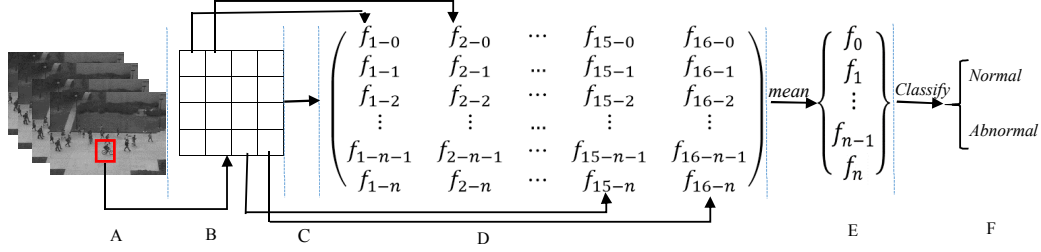


Figure 5. Large patch anomaly detection using feature learning. (A) Input video. (B) Selected test patch (e.g., $40 \times 40 \times 5$) is divided into 16 small patches. (C) $W_1 \times$ small patch. (D) Pooling all feature vectors (16 vectors). (E) Computing the mean of each feature and create one feature vector. (F) Classifying with the learned classifier using $10 \times 10 \times 5$ patches.

fore patches) in the video stream arrive. We use SSIM for computing the similarity between two patches, which is a well-known image-quality assessment tool [4]. Further, as a second type of local descriptor, we calculate the SSIM of each single frame with its subsequent frame in the patch of interest. Figure 4 illustrates our local feature assessment through the spatio-temporal neighboring. The local descriptor would be the combination of the SSIM values, i.e., $[d_0 \cdots d_9, D_0 \cdots D_{t-1}]$.

Anomaly Classifier. To model the normal activities in each video patch, we incorporate two Gaussian classifiers C_1 and C_2 . For classifying x' patches, as described, we use two partially independent feature sets (global and local), and compute the Mahalanobis distance $f(y)$. If $f(y)$ is larger than the threshold then it is considered to specify an abnormal patch, where y equals $W_1 \times x'$ in the global classifier, and $[d_0 \cdots d_9, D_0 \cdots D_3]$ for the case of the local classifier. To avoid numerical instabilities, density estimates are avoided. As a result, the C_1 and C_2 classifiers are defined as follows:

$$C_i(x) = \begin{cases} \text{Normal} & f(x) \leq \text{threshold} \\ \text{Anomaly} & \text{otherwise} \end{cases} \quad (2)$$

with

$$f(x) = (x - \mu)^T \Sigma^{-1} (x - \mu) \quad (3)$$

where μ and Σ are mean and covariance matrix, respectively. Selecting a “good” threshold is important for the performance; it can be selected based on training patches. As mentioned before, if both C_1 and C_2 classifiers label a patch as being an anomaly, it is considered to be an anomaly, but if one or neither of them considers the patch as being an anomaly, our algorithm classifies it as being a normal patch. A summary of these criteria is shown as F function in the following equation:

$$F(x) = \begin{cases} \text{Anomaly} & \text{if } C_1 = \text{Anomaly} \wedge C_2 = \text{Anomaly} \\ \text{Normal} & \text{otherwise} \end{cases} \quad (4)$$

Anomaly detection using feature learning. We learn the features from raw training data, and classify the video

patches as specified in the previous section. But based on the idea in [3], using both small patches and large patches usually leads to increased values of false-positive rate and decreased value of true-positive rate, respectively. When the patches become larger, the input dimension of the auto-encoder increases, so the number of weights in the network, which need to be learned, will also increase.

Under the condition of limited training examples, learning of features from large patches is impractical (for example $40 \times 40 \times 5$), to overcome these challenges, we learn the features from (small) $10 \times 10 \times 5$ patches. To create a model using these features, in the test phase the large patches ($40 \times 40 \times 5$) are considered. Because the learned classifier is adapted for $10 \times 10 \times 5$ patch representations, we convolve the learned feature (W_1) in $40 \times 40 \times 5$ patches, without overlapping, and pool the 16 extracted feature vectors from the $40 \times 40 \times 5$ patches. So, we use mean pooling to achieve a representation of $40 \times 40 \times 10$ patches that can be checked with the learned classifier using $10 \times 10 \times 5$ patches. This procedure is shown in Figure 5.

3. Experimental results and comparisons

We compare our algorithm with state-of-the-art methods on Ped2 UCSD¹ and UMN² benchmarks. We empirically demonstrate that our approach is suitable to be used in surveillance systems.

Experimental settings. Feature learning is done with $10 \times 10 \times 5$ patches. Training and testing phases in anomaly detection is done with $10 \times 10 \times 5$ and $40 \times 40 \times 5$ patch sizes, respectively. In anomaly detection, the size $40 \times 40 \times 5$ is exploited. Feature learning is done with an auto-encoder with 0.05 sparsity. Each $10 \times 10 \times 5$ patch is represented by a 1000-dimensional feature vector. Before feature learning, normalization is performed to set the mean and variance to 0 and 1, respectively.

UCSD datasets. This dataset includes two subsets, ped1 and ped2, that are from two different outdoor scenes. Both

¹www.svcl.ucsd.edu/projects/anomaly/dataset.html

²mha.cs.umn.edu/Movies/Crowd-Activity-All.avi

are recorded with a static camera at 10 fps, with the resolutions 158×234 and 240×360 , respectively. The dominant mobile objects in these scenes are pedestrians. Therefore, any object (*e.g.*, a car, skateboarder, wheelchair, or bicycle) is considered as being an anomaly. We evaluate our algorithm on ped2. This subset includes 12 video samples, and each sample is divided into training and test frames. To evaluate the localization, we utilize the ground truth of all test frames. We compare our results with state-of-the-art methods using receiver operating curve (ROC) and equal error rate (EER) analysis, similar to [13]. We use two evaluation measures, one at frame level and the other at pixel level. In addition to these, we define a new measure for the accuracy of anomaly localization, called *dual pixel level*. These measures are defined as follows:

Frame level measure: If one pixel detects an anomaly then it is considered as being an anomaly.

Pixel level measure: If at least 40 percent of anomaly ground truth pixels are covered by pixels detected by the algorithm, then the frame is considered to be an anomaly.

Suppose that the algorithm detects some region as being an anomaly, and just one of these regions has an overlap with anomaly ground truth; the number of false regions is not considered in the two former measures. Such a region is called a “lucky guess”. For considering the “lucky guess” regions, we introduce the dual pixel level. This measure is sensitive to a “lucky guess”.

Dual pixel level: In this measure, a frame is considered as being an anomaly if (1) it satisfies the anomaly condition at pixel level and (2) at least β percent (*i.e.*, 10%) of the pixels detected as anomaly are covered by the anomaly ground truth. If, in addition to the anomaly region, irrelevant regions are also considered as being an anomaly, then this measure does not identify the frame as being positive. Figure 6 shows an example for the different measures of anomaly detection.

Performance Evaluations. Figure 7 shows a qualita-

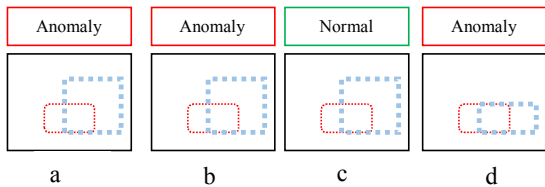


Figure 6. Measure of anomaly evaluation. The blue and red rectangles indicate the output of the algorithm and anomaly ground truth, respectively. (a) Frame-level. (b) Pixel-level evaluation: 40 percent red (ground truth) is covered with blue (detected). (c) Dual pixel-level: Evaluates that 40 percent of red is covered by blue, but at least β percent of blue is not covered by red. (d) Dual-pixel level

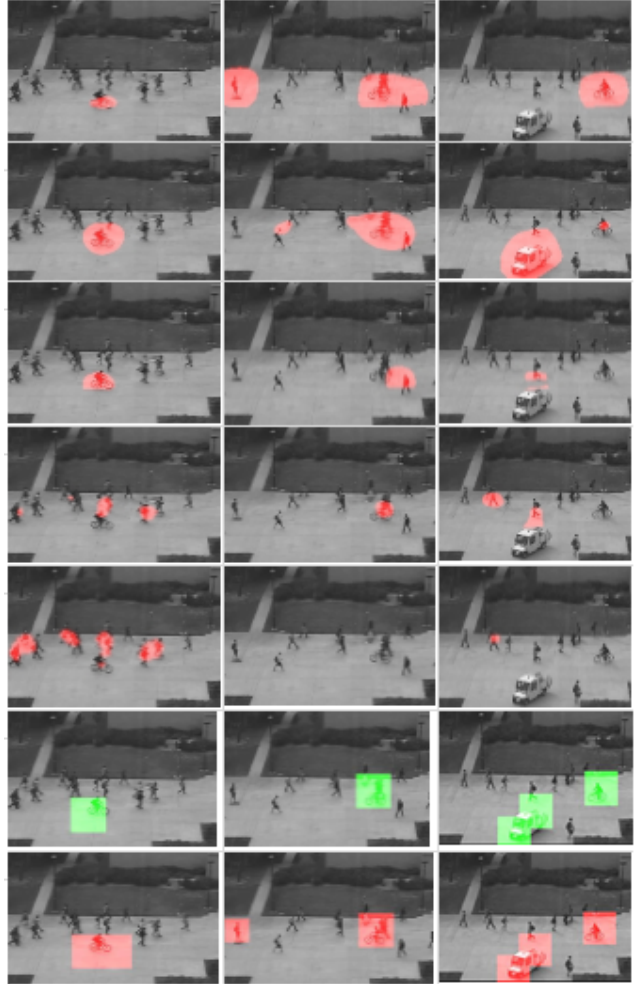


Figure 7. Example of anomaly detection from three scenes. First row to 7th row show Temporal MDT, Spatial MDT, MPPCA, Social force, Optic flow, Our method (feature learning only), an Our method (combined views)

tive comparison with other methods.³ This figure indicates that our algorithm has the best performance in comparisons with all the competing algorithms. For the run-time comparisons, see Table 1.

Method	Time (second per frame)
Xua <i>et al.</i> [20]	Offline
Li <i>et al.</i> [11]	1.38
Ours	0.04

Table 1. Run time comparison

In Figure 8 (Left), the frame-level ROC of our method is compared with other methods on the ped2 dataset. It shows that our method is comparable to other methods. For this

³ Our results are available at <http://mahfathy.iust.ac.ir/>.

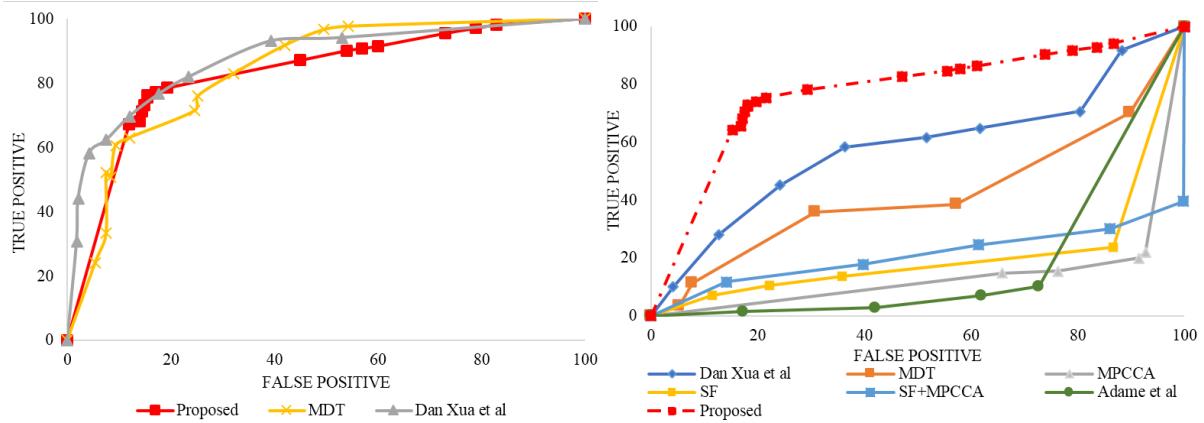


Figure 8. Comparison ROC curve (left to right): Frame-level evaluation and pixel-level evaluation

Method	Frame-level	Pixel-level
SF [14]	42	79
MPCCA [9]	30	82
MPCCA+SF [13]	36	72
Adam et.al [1]	42	76
MDT [13]	25	55
Xua et al. [20]	20	42
Li et al. [11]	18.5	29.9
Ours	19	24
Ours 0.1	—	67.5
Ours 0.05	—	27.5

Table 2. EER for frame and pixel level comparisons

measure, the EER for frame level for different methods is shown in Table 2. This confirms that our method has a good performance in comparison to others. We outperform all of the methods except the one of Li *et al.* (we are 0.5 percent below), reported in [11].

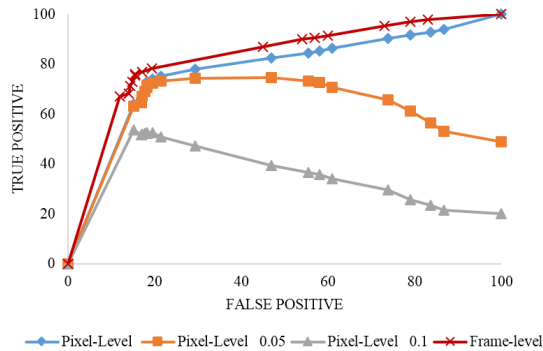


Figure 9. Comparison between dual pixel localization with β equal to 0 (pixel-level measure), 0.05, 0.10, and frame-level

Figure 8 (Right) illustrates the ROC with respect to the pixel-level measure. In Table 1, we compare the pixel level EER of our approach to that of other approaches. Our method’s EER is 24 percent where the next best result is 29.9 percent reported for the method Li *et al.* [11]. Our method is 5.9 percent better than the otherwise best result. The results show (both ROC and EER) that our algorithm outperforms the other methods for the pixel-level measure. We also use a dual-pixel level measure to analyze the accuracy of anomaly localization. Figure 9 shows the effect of the parameter β on our algorithm. The algorithm has a good performance, even better than the state-of-the-art, in pixel level with $\beta=0.05$ percent and 0 percent. Figure 9 illustrates comparisons at frame level and pixel level of our approach; in contrast to all reported algorithms, the pixel level measure is very close to frame level measure in our algorithm.



Figure 10. Examples of normal and abnormal crowded activities in scenes of the UMN dataset. Top: Normal. Bottom: Abnormal

UMN dataset. The UMN dataset has three different scenes. In each scene, a group of people are walking in an area, suddenly all people run away (escape); the escape is considered to be the anomaly. Figure 10 shows examples of normal and abnormal frames of this dataset.

Method	EER	AUC
Chaotic invariants [19]	5.3	99.4
SF [14]	12.6	94.9
Sparse [6]	2.8	99.6
Saligrama et.al [16]	3.4	99.5
Li <i>et al.</i> [11]	3.7	99.5
Ours	2.5	99.6

Table 3. Anomaly detection performance in EER and AUC

This dataset has some limitations. There are only three anomaly scenes in the dataset, and the temporal-spatial changes between normal and abnormal frames are very high. This dataset has no pixel-level ground truth. Based on this limitations, to evaluate our method, the EER and AUC in frame-level are used. The EER and AUC results are shown in Table 3. Because this dataset is simple, and anomaly localization is not important, only the global detector is used. Previous methods performed reasonably good on this dataset. The AUC of our method is comparable with the otherwise best result, and the EER of our approach is better (by 0.3 percent) than the one of the best previous method.

4. Conclusions

We presented an anomaly detection and localization method. In our method, we propose to represent a video using both global and local descriptors. Two classifiers are proposed based on these two forms of representation. Our fusion strategy on the outputs of these two classifiers achieves accurate and reliable anomaly detection and localization. However, each of the two classifiers has a good performance for anomaly detection, solely. This is especially shown on the UMN dataset where the global descriptor achieves state-of-the-art results. We introduced a new metric for region level anomaly detection for suspicious regions, as well. The performance of our approach on the UCSD dataset is better compared to recent approaches. It is also worth noting that we achieve all these good results in a much better running time than all the competing methods. Our method enjoys a low computational complexity, and can be run in real-time. This makes it quite useful for real-time surveillance applications, in which we are dealing with live streams of videos.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed location monitors. *IEEE Trans. Pattern Analysis Machine Intelligence*, 30(3):555–560, 2008. 1, 6

[2] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurrences. In *CVPR*, pages 1446–1453, 2009. 1

[3] M. Bertini, A. Del Bimbo, and L. Seidenari. Multi-scale and real-time non-parametric approach for anomaly detection and localiza-

tion. *Computer Vision Image Understanding*, 116(3):320–329, 2012. 4

[4] D. Brunet, E. R. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. *IEEE Trans. Image Processing*, 21(4):1488–1499, 2012. 4

[5] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Int. Conf. Artificial Intelligence Statistics*, pages 215–223, 2011.

[6] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, pages 3449–3456, 2011. 1, 7

[7] Y. Cong, J. Yuan, and Y. Tang. Video anomaly search in crowded scenes via spatio-temporal motion context. *IEEE Trans. Information Forensics Security*, 8(10):1590–1599, 2013. 1

[8] F. Jianga, J. Yuan, S. A. Tsafaris, and A. K. Katsaggelosa. Anomalous video event detection using spatiotemporal context. *Computer Vision Image Understanding*, 115(3):323–333, 2011. 1

[9] 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. 1, 6

[10] L. Kratz and K. Nishino. Crowded scenes using spatio-temporal motion pattern models. In *CVPR*, pages 1446–1453, 2009. 1

[11] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Analysis Machine Intelligence*, 36(1):18–32, 2014. 1, 5, 6, 7

[12] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in MATLAB. In *ICCV*, pages 2720–2727, 2013. 1

[13] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975–1981, 2010. 1, 5, 6

[14] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942, 2009. 1, 6, 7

[15] M. J. Roshtkhari and M. D. Levine. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer Vision Image Understanding*, 117(10):1436–1452, 2013. 1

[16] V. Saligrama and Z. Chen. Video anomaly detection based on local statistical aggregates. In *CVPR*, 2012. 1, 7

[17] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Int. ACM Conf. Machine Learning*, pages 1096–1103, 2008. 1

[18] X. Wang, X. Ma, and E. Grimson. Perception by hierarchical Bayesian models. In *CVPR*, pages 1–8, 2007. 1

[19] S. Wu, B. Moore, and M. Shah. Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In *CVPR*, 2010. 1, 7

[20] D. Xu, R. Song, X. Wu, N. Li, W. Feng, and H. Qian. Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. *Neurocomputing*, 143:144–152, 2014. 5, 6

[21] Y. Yang, G. Shu, and M. Shah. Semi-supervised learning of feature hierarchies for object detection in a video. In *CVPR*, pages 1650–1657, 2013. 3

[22] D. Zhang, D. Gatica-Perez, S. Bengio and I. McCowan. Semi-supervised adapted HMMS for unusual event detection. In *CVPR*, 1

2005.