

DETECTION OF ABNORMAL ACTIVITY USING ONE-CLASS CLASSIFICATION

Artem V. Gritsenko

Department of Applied Mathematics and Computer Science

Stavropol State University, Stavropol

1, Pushkina str., 355009, Russia



Abstract

By anomaly detection, we mean the detection of patterns in the video or images that do not conform to the expected behaviour. We can define anomaly as rare or infrequent behaviour compared to all other behaviours, which is also referred to as an outlier. However, detecting abnormalities in video is a challenging problem since the class of all irregular objects and behaviours is infinite and thus no (or by far not enough) abnormal training samples are available. To handle this problem we refer to the task of classification between two classes when there are no labels for an abnormal class in the training data.

We propose a framework that is based on the bag of visual words model that applied to the set of images or frames to get the robust descriptors. To learn the normal activity patterns we use the extension of support vector machine algorithm to the case of unlabeled data. An experimental evaluation of the framework is conducted with a large publicly-available dataset of crowd scenes. We demonstrate that the proposed approach outperforms state-of-the-art anomaly detection algorithms.

Key words: *anomaly in crowd scenes, contextual anomalies, crowd analysis, outlier detection, machine learning, image classification, support vector machine.*

1. INTRODUCTION

Detecting of abnormal behaviour in crowd scenes is a very challenging problem in the area of the surveillance and computer vision. Various abnormal activities have been studied, including restricted-area access detection [1], car counting [2], detection of people carrying cases [3], abandoned objects, group activity detection [5, 6], social network modelling [7], monitoring vehicles [8], and scene analysis [9].

In this paper, we focus on detecting abnormal events in crowds that involve interactions between people and carriers, e.g., moving of the vehicles at the foot-path or pedestrian crossings, which is a very important application for video surveillance. Despite the problem's practical appeal, abnormality detection remains quite challenging technically, and intellectually it can even be hard to define. The foremost challenge is that "unusual" things naturally occur with unpredictable variations, making it hard to discriminate a truly abnormal event from noisy normal observations.

The key challenges of the most approaches of detection anomalous activity are the local nature of the analyzed data and the high level of dependency on the motion information between the video frames. The main paradigm to abnormality detection is currently to classify each local image patch individually, e.g., [10, 11] or to detect abnormal image regions separately [12]. However, deciding locally and independently about the abnormality of each individual image region is an ill-posed problem. On the other hand, the common approach assumes the analysis of the motion information

that causes a lot of false alarms, e.g., the person that moves faster than others can be labelled as a vehicle.

In this work we propose a framework that analyses each image from a video sequence separately and finds the global patches that can reliably characterize the scene. For this purpose we use the bag of words representation, which represents images as histograms based on the independent features. Then we built a feature map using additive kernel. At the last stage the one-class-SVM [13] is used to discriminate between normal and abnormal frames.

The algorithm has two main advantages. First, the framework is capable to find not only local but also global anomalies. Since the framework uses the bag of word representation of video frames, it trains the model with the dictionary constructed from all images and so, considers not only several local patches, but the whole video. Second, it outperforms the state-of-the-art techniques in accuracy and creates fewer false alarms because the approach avoids tracking and calculating of spatio-temporal gradients that makes it robust to motion changes.

The evaluation is based on a novel dataset of crowded scenes of Mahadevan et al. [14], which is 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. The training data features only normal patterns with large within-class variability, whereas the test set consists of normal and abnormal instances. In the experimental evaluation, our approach has outperformed all state-of-the-art methods for abnormality detection on this dataset.

2. RELATED WORK

To detect abnormal activities, most algorithms attempt to define normal activity patterns first, and then determine how much new observations deviate. Existing techniques differ in the extent of supervision. Some approaches use both normal and abnormal activity patterns for training [14, 15, 16], while others are fully unsupervised [17, 18 19]. Broadly considered, the previous methods of anomaly detection explicitly used gradients or optical flow [17, 18, 15, 16, 19, 20, 21, 22].

Adam [17] proposed an algorithm based on the multiple integrated monitors that collect optical flow direction and magnitude. Kim and Grauman [18] associated each video patch with continual optical flow observations, and learn atomic motion patterns via MPPCA. Further they compute MRF parameters based on the learned patterns and incrementally update the model. The technique called textures of optical flow [16] based on the extension of traditional greyscale textural features to robust dense optical flow fields. The authors built a statistical model of normal motion patterns based on 3D volumes called spatio-temporal patches, and anomalies are detected as outliers from this model.

Spatio-Temporal Laplacian Eigenmap method [22] discovered the internal structure of the video using Histograms of Optical flow and produced a smooth trajectory for each video sequence. Mehran [21] was inspired by the traditional Social Force model to detect anomalies in crowd videos. The approach proposed in [20] learnt the probabilistic model from the chaotic feature set that is consists of particle trajectories. Cui and Liu [19] used optical flow features to calculate the interaction of energy potential and then SVM to find abnormalities. Tran and Yuan [23] used 3D bounding box to find the optimal spatio-temporal path for accurate event localization.

Some other techniques focused not only on motion information but considered the appearance of the objects as well [10, 24, 14]. Boiman and Irani [10] used spatio-temporal patches and declared regions that cannot be reconstructed using data from previous frames as abnormal. The approach proposed in

[24] is based on the sparse coding algorithm and labeled anomalous those segments, which reconstruction errors are significantly higher than the majority of the others. Antic and Ommer [25] got hypotheses for each object via parsing of video frames and defined them abnormal if they cannot be explained by any instance from the set of normal training samples.

Several authors [26, 14] were inspired by the dynamic texture model to represent motion. Jiang and Wu [26] used clustering to organize local patches in blobs and then k-means to detect anomalies. Mahadevan [14] modeled appearance and dynamics using a generative ARMA model for each component in the mixture of dynamic textures.

Thus, existing approaches to anomaly detection have mostly used gradients or optical flow, and these approaches are generally best suited for detecting regions of unusually high velocity. However, in many cases the objects to be detected do not necessarily travel at a substantially faster speed than regular pedestrians. For example, a cyclist or a slow-moving vehicle should be detected regardless of their velocity. Even an average pedestrian generates motion of a high velocity, due to the periodic movement of their limbs.

3. PROPOSED METHOD

We propose a new method to detect anomalies in object behaviour using ‘bag of visual words’ (BoW) representation. The framework is summarized in Fig.1. We first extract interest point using Difference of Gaussians. Second, features are represented by scale-invariant feature transform (SIFT) into numerical vectors. Third, vectors that represent patches are converted to "codewords" by the hierarchical k-means clustering. Finally, SVM is used to detect abnormal events.

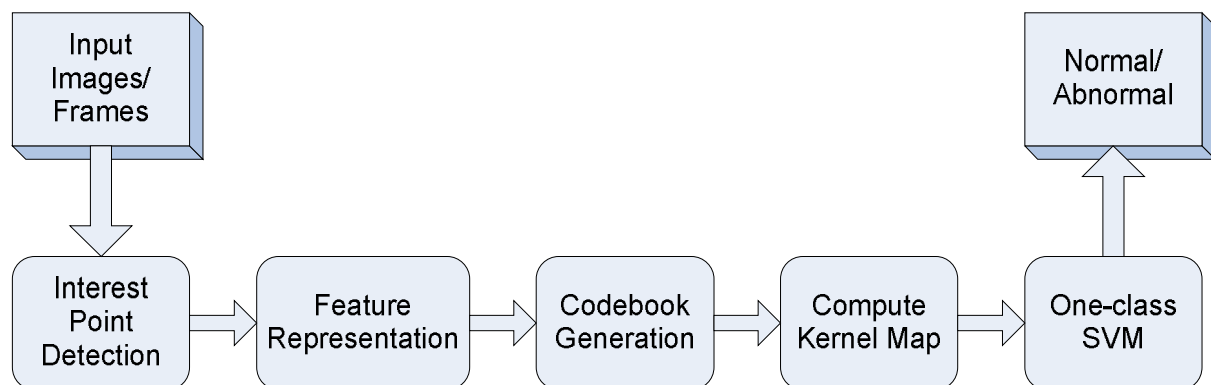


Fig.1 Summary of the proposed algorithm

3.1. Codebook Generation

We will detect keypoints using a cascade filtering approach that uses efficient algorithms to identify candidate locations that are then examined in further detail. The first stage of keypoint detection is to identify locations and scales that can be repeatedly assigned under differing views of the same object. Detecting locations that are invariant to scale change of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as scale space [27]. To efficiently detect stable keypoint locations in scale space, we use scale-space extrema in the difference-of-Gaussian function convolved with the image:

$$D(x, y, \delta) = (G(x, y, k\delta) - G(x, y, \delta)) * I(x, y) = L(x, y, k\delta) - L(x, y, \delta), \quad (1)$$

where $L(x, y, k\delta)$ – convolution of the original image $I(x, y)$ with Gaussian blur $G(x, y, k\delta)$ of radius δ . In order to detect the local maxima and minima of $D(x, y, \sigma)$, each sample point is compared to its eight neighbours in the current image and nine neighbours in the scale above and below. Afterwards some points that have low contrast (and therefore sensitive to noise) or are poorly localized along an edge are rejected.

By assigning a consistent orientation to each keypoint based on local image properties, the keypoint descriptor can be represented relative to this orientation and therefore achieve invariance to image rotation [28]. The scale of the keypoint is used to select the Gaussian smoothed image, L , with the closest scale, so that all computations are performed in a scale-invariant manner. For each image sample, $L(x, y)$, at this scale, the gradient magnitude, $m(x, y)$, and orientation, $\theta(x, y)$, is precomputed using pixel differences:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}, \quad (2)$$

$$\theta(x, y) = \tan^{-1} \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}. \quad (3)$$

An orientation histogram is formed from the gradient orientations of sample points within a region around the keypoint. Finally, we compute a descriptor for the local image region that is highly distinctive yet is as invariant as possible to remaining variations, such as change in illumination or 3D viewpoint.

The final step for the BoW model is to convert vector represented patches to "codewords" that can be considered as representatives of several similar patches. To handle this problem we utilize an efficient implementation of the Lloyd's clustering algorithm, called filtering algorithm [29]. The idea is to store the data points in a kd-tree and to maintain a subset of candidate centers, for each node of the tree. The candidates for each node are pruned as they are propagated to the node's children. Since the kd-tree is computed for the data points rather than for the centers, there is no need to update this structure with each stage of Lloyd's algorithm. Codewords are then defined as the centers of the learned clusters. Thus, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords.

3.2. Classification and anomaly detection

Now we can regard bag of visual words representation as probability distribution. To use it in the context of SVM, we should convert the data into a format suitable for SVM with the help of the feature map. The feature map transforms the data into a compact linear representation which reproduces the desired kernel to a very good level of approximation. First, we need to define an appropriate similarity function $K(x, y)$ between finite probability distributions x, y (normalized histograms). The only requirement is that K must be a positive definite function. The latter property also guarantees the existence of a feature map for additive kernels, but this is usually difficult to compute and high dimensional. To obtain compact and simple representations that are efficient in both training and testing, have excellent performance, and have a satisfactory theoretical support we will use the homogeneous kernel map [30].

For finite dimensional distributions (histograms) x, y , an additive kernel is given by:

$$K(x, y) = \sum_{b=1}^B k(x_b, y_b), \quad (4)$$

where b is the bin index. Here k is itself a positive definite kernel on the non-negative reals. After the derivation, we get an explicit form for the feature map as an infinite dimensional vector and then approximate it by a finite number of samples. The following equation defines a closed form feature map for χ^2 kernel (because it is easier to approximate), which is very simple and efficient to compute:

$$[\Psi(x)]_\lambda = e^{i\lambda \log x} \sqrt{x \sec h(\pi\lambda)}, \quad (5)$$

where $\lambda \in R$ can be thought of as the index of the feature vector $\Psi(x)$.

Finally, we utilize the algorithm that returns a function f that takes the positive value in a “small” region capturing most of the data points and negative elsewhere [13]. Our strategy is to map the data into the feature space corresponding to the kernel and to separate them from the origin with maximum margin. The decision function is defined as follows:

$$f(x) = \text{sgn}((\omega \cdot \Phi(x)) - \rho), \quad (6)$$

where, ω is a weight vector and ρ is an offset parameterizing a hyperplane in the feature space associated with the kernel. The functional form of f is given by a kernel expansion in terms of a potentially small subset of the training data; it is regularized by controlling the length of the weight vector in an associated feature space. The expansion coefficients are found by solving a quadratic programming problem, which we do by carrying out sequential optimization over pairs of input patterns. Thus, the decision function will look like:

$$f(x) = \text{sgn} \left(R^2 - \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) + 2 \sum_i a_i k(x_i, x) - k(x, x) \right), \quad (7)$$

where $\alpha_i \geq 0$ are multipliers that introduce a Lagrangian and $\sum_i \alpha_i = 1$. This function is used to compute the abnormality score for each test frame.

4. EXPERIMENTAL RESULTS

We evaluate our approach on the challenging abnormality datasets Ped1 and Ped2 that have been recently proposed by Mahadevan et al. [14]. The video sequences feature a pedestrian walkway acquired by a stationary camera with low resolution (pedestrians have a height between 10 and 30 pixels). The crowd density varies and there are numerous sequences that are very crowded and with severe occlusions. Abnormalities are not staged but are naturally occurring events such as objects that are unusual in the present surroundings (e.g. cars on walkways) or objects that behave irregularly such as people cycling across walkways or walking over the surrounding grass. Other abnormalities include skaters, small carts, and wheelchairs. The first dataset (Fig.2 (a), (b)), contains groups of people walking towards and away from the camera, and some amount of perspective distortion. The second (Fig.2 (c), (d)) contains scenes with pedestrian movement parallel to the camera plane.

The training set contains 34 short clips for learning of normal patterns, and there is a subset of 10 clips in testing set provided with pixel-level binary masks, which identify the regions containing abnormal events for Ped1 dataset and 16 clips for training and 9 for testing in the Ped2 dataset.. Each clip from Ped1 has 200 frames, with a 158×238 resolution and clips from Ped2 have 180 frames of 360×240. The dictionary size of 400 codewords was used.

To evaluate the performance of the proposed anomaly representation, we compared it to four other recently proposed representations - the social force model [21], the mixture of optical flow (denoted MPPCA) [18], the multiple fixed-location monitoring method (denoted Adam et al.) [17] and the

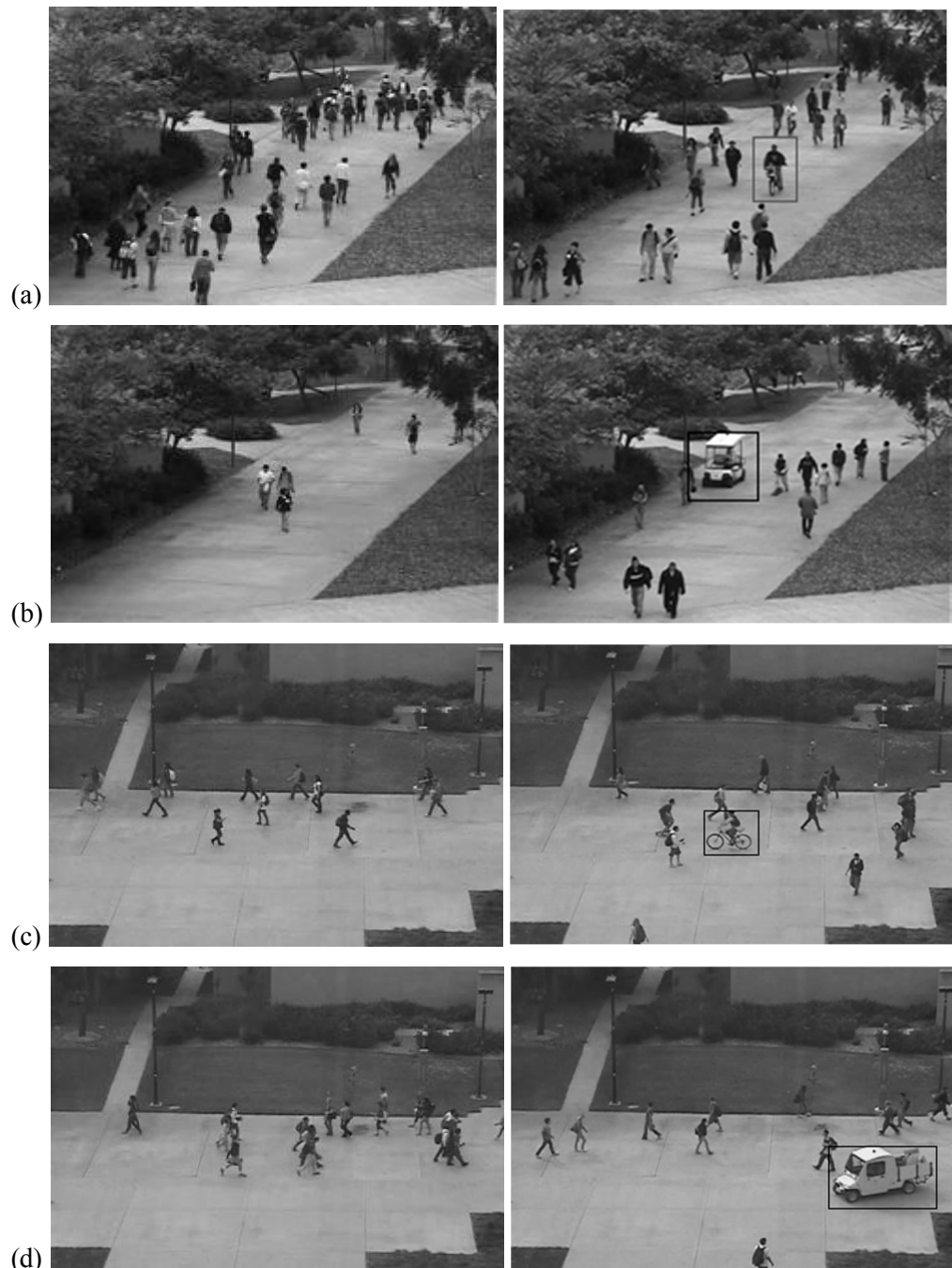


Fig. 2. Examples of normal and abnormal frames. The left row contains normal behaviour from, Peds1 dataset (a), (b) and Peds2 (c), (d). The right row consists of anomalous frames: (a) – bike, (b) –car from Peds1 and (c) – bike, (d) – car from Peds2

mixture of dynamic textures model (denoted MDT) [14]. The computer detection is compared to ground truth at each frame and the ROC curve is generated (Fig.3). 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 – and the area under ROC curve (AUC) are reported (Tab.1).

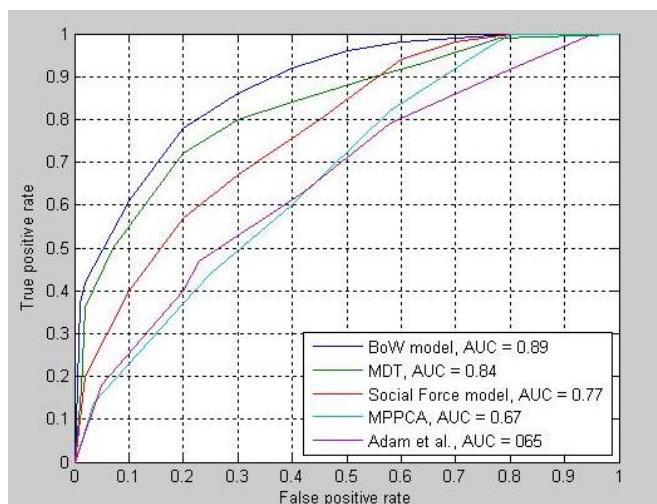


Fig. 3. The ROCs for frame-level abnormality detection

Approach	EER (equal error rate)	AUC (area under curve)
Adam [17]	40%	65%
MPPCA [18]	35%	67%
Social Force Model [21]	37%	77%
Mixture of Dynamic Textures [14]	25%	84%
BoW model	23%	89%

Tab.1. Quantitative comparison of our method

In all the experiments our method outperforms current state-of –the-art approaches. Our per-frame labeling achieves EER of 23% which is significantly better than Adam, MPPCA and Social Force approaches and slightly outperforms MDT implementation. We achieve the AUC, which is a more robust measure, as it does not depend on only a single spot on the curve, of 89% which is also a better result between compared techniques. Moreover, the proposed approach is faster (~2 seconds per frame) than other methods. The implementation on MATLAB was ran on a standard CoreDuo machine with 2.8GHz CPU and 2GB RAM

5. CONCLUSION

We have proposed a novel approach for abnormality detection in crowd scenes based on the bog of visual words representation. The approach represents each frame as a collection of visual words and then use them train SVM and make decision. We have demonstrated that the proposed approach

outperforms state-of-the-art anomaly detection algorithms on a large, publicly-available dataset. The BoW method is also faster than current techniques because it doesn't calculate high dimensional optical flow features and uses homogeneous kernel map for SVM instead.

REFERENCES

1. Konrad J. (2005) Motion detection and estimation. Handbook of Image and Video Processing, 2nd Edition.
2. N. Friedman and S. Russell. (1997) Image segmentation in video sequences: A probabilistic approach. Proc. UAI.
3. Haritaoglu I., Harwood D., Davis L. S. (2000) Realtime surveillance of people and their activities. Trans. PAMI.
4. Smith K., Quelhas P., Gatica-Perez D. (2006) Detecting abandoned luggage items in a public space. PETS.
5. Zhou, Y., Yan, S., Huang, T.S. (2008) Pair-activity classification by bi-trajectories analysis. In Proceedings of CVPR.
6. Ni B., Yan S., Kassim A. (2009) Recognizing human group activities with localized causalities. In Proceedings of CVPR.
7. Yu T., Lim S., Patwardhan K., Krahnstoever N. (2009) Monitoring, recognizing and discovering social networks. In Proceedings of CVPR.
8. Yu Q., Medioni G. (2009) Motion pattern interpretation and detection for tracking moving vehicles in airborne video. In Proceedings of CVPR.
9. Swears E., Hoogs A. (2009) Functional scene element recognition for video scene analysis. WMVC.
10. Boiman O., Irani M. (2005) Detecting irregularities in images and in video. In Proceedings of ICCV, pp. 462–469.
11. Xiang T., Gong S. (2005) Video behaviour profiling and abnormality detection without manual labelling. In Proceedings of ICCV, pp 1238–1245.
12. Zhong H., Shi J., Visontai M. (2004) Detecting unusual activity in video. In Proceedings of CVPR, pp. 819–826.
13. Scholkopf, B., Platt, J. C., Shawe-Taylor, J. S., Smola, A. J., Williamson, R. C. (2001) Estimating the Support of a High-Dimensional Distribution. Neural Computation, 13 (7). pp. 1443-1471.
14. Mahadevan V., Li W., Bhalodia V., Vasconcelos N. (2010) Anomaly detection in crowded scenes. In Proceedings of CVPR.
15. Cong Y., Yuan J., Liu J. (2011) Sparse reconstruction cost for abnormal event detection. In Proceedings of CVPR. pp. 3449-3456
16. Ryan, D., Denman, S., Fookes, C., Sridharan, S. (2011) Textures of Optical Flow for Real-Time Anomaly Detection in Crowds. AVSS, pp. 230 – 235.
17. Adam A., Rivlin E., Shimshoni I., Reinitz D. (2008) Robust real-time unusual event detection using multiple fixed-location monitors. PAMI, 30.

Journal of International Scientific Publications:
Materials, Methods & Technologies, Volume 6, Part 1

18. Kim J., Grauman K. (2009) Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In Proceedings of CVPR.
19. Cui X., Liu Q., Gao M., Metaxas D. (2011) Abnormal detection using interaction energy potentials. In Proceedings of CVPR. pp. 3161-3167.
20. Wu S., Moore B.E., Shah M. (2010) Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In Proceedings of CVPR. pp. 2054-2060.
21. Mehran R., Oyama A., Shah M. (2009) Abnormal crowd behavior detection using social force model. In Proceedings of CVPR.
22. Thida, Eng H.-L., Monekosso D. N., Remagnino P. (2010) Learning Video Manifold for Segmenting Crowd Events and Abnormality Detection. In Proceedings of ACCV (1). pp. 439-449.
23. Tran D., Yuan J. (2011) Optimal spatio-temporal path discovery for video event detection. In Proceedings of CVPR. pp. 3321-3328.
24. Zhao B., Fei-Fei L., Xing E. P. (2011) Online Detection of Unusual Events in Videos via Dynamic Sparse Coding. In Proceedings of CVPR.
25. Antic B., Ommer B. (2011) Video Parsing for Abnormality Detection. In Proceedings of ICCV. pp. 2415 – 2422.
26. Jiang F., Wu Y., Katsaggelos A.K. (2009) Detecting contextual anomalies of crowd motion in surveillance video. In Proceedings of ICIP., Cairo, Egypt, pp. 1117-1120.
27. Lowe, D.G. (1999) Object recognition from local scale-invariant features. In Proceedings of ICCV, Corfu, Greece, pp. 1150-1157.
28. Lowe, D.G. (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 2, pp. 91-110.
29. Kanungo T., Mount D.M., Netanyahu N., Piatko C., Silverman R., Wu A.Y. (2002) An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Analysis and Machine Intelligence, 24, pp. 881-892.
30. Vedaldi A., Zisserman A. (2011) Efficient Additive Kernels via Explicit Feature Maps. In Proceedings of PAMI.