

Anomaly Detection in Crowded Scenarios Using Local and Global Gaussian Mixture Models

Adrián Tomé^(✉)  and Luis Salgado 

Grupo de Tratamiento de Imágenes, E.T.S.I. de Telecomunicación,
Universidad Politécnica de Madrid, Madrid, Spain
{ata,lsa}@gti.ssr.upm.es

Abstract. This paper presents an objective comparison between two approaches for anomaly detection in surveillance scenarios. Gaussian mixture models (GMM) are used in both cases: globally, with a unique model that covers the whole scene; and locally, with one model per spatial location. The two approaches follow a “bottom-up” approach that avoids any object tracking and motion features extracted with a robust optical flow method. Furthermore, we evaluate the contribution of each feature through a statistical tool called Correlation Feature Selection in order to assure the best performance. Evaluation is done in UCSD dataset, concluding that the global model offers better results, outperforming similar anomaly detection approaches.

Keywords: Anomaly detection · Gaussian mixture model · Robust optical flow · Correlation feature selection

1 Introduction

Computer vision techniques have received much attention in the last years, mainly because they can solve tasks such as detection and recognition of interesting objects and events that could be useful for solving a varied set of issues. For instance, anomaly detection in automated surveillance [1–4, 6, 7].

In this context the main problem is that the definition of anomaly is not universal. Thus, anomalies are sometimes defined as events that differ from those considered normal [1]; other works [4] define them as unusual, uncommon or irregular events. On the other hand, authors of [2] make the definition based on the low frequency of appearance that anomalous events have compared to dominant ones.

The way to face the detection of anomalies has changed over the years. Former works needed to track the objects on scene to be able to extract high-level information such as trajectories to detect anomalous motion or speed patterns (“top-down” approach) [5]. These systems work well except in crowded scenes, where occlusions and clutter are produced. In these cases, it is convenient to extract low-level features, useful to extract information at a higher level (“bottom-up” approach) [2–4, 6]. Generally, “Bottom-up” approaches have similar feature

extraction processes. This is, they divide the sequences into spatio-temporal volumes (cuboids) from which low-level features are extracted so local events can be captured. Some popular features are gradients [2,3] or flow-based features [4,7].

We can highlight some literature methods. For instance, the one proposed by Roshtkhari and Levine [2], in which the nominal events are learned using low-level features from spatio-temporal compositions. It works without supervision, assuming that anomalous events do not occur frequently and updating the model in an online manner. On the other hand, they use multi-scale densely sampled cuboids, increasing the complexity of the system, which significantly bounds its applicability.

Alternatively, Cong et al. [4] proposed the use of a sparse representation model. They build a basis of representative samples of the training set. Later, the reconstruction cost of test samples is computed and used to select the anomalous ones. Its main disadvantage is that it requires an accurate method for selecting the representative samples through elaborated optimization algorithms.

Another remarkable work has been proposed by Zhang et al. [17]. They use optical flow and spatio-temporal gradients to detect anomalies under a scheme based on Support Vector Data Description (SVDD).

Tziakos et al. [9] used GMMs as binary classifiers applied on local detectors. Ryan et al. [7] proposed a method that builds a Gaussian mixture model with descriptors formed exclusively by optical flow features. Later, authors of [8] improved the method by incorporating more robust descriptors and combining the GMM with a Markov random field. The main drawback is its elevated computational cost.

A different approach was proposed by Mahadevan et al. [1] and later improved by Li et al. [16], in which mixture of dynamic textures are employed to characterize the appearance and dynamics of the scene. Afterwards, anomalies are detected as outliers. Although interesting, it is a significantly complex method with high computational demand. Besides, it has been outperformed by other methods.

In this work we build two normality models based on Gaussian Mixtures (global and local) and compare their performance, concluding that with the global approach we obtain the best anomaly detection rates. The features used are based on optical flow, obtained with a robust method. Furthermore, the descriptors are constructed with the best combination of features thanks to the use of a statistical tool called Correlation Feature Selection (CFS), ensuring the best performance. Thus, our approach is simple but more accurate and computationally cheaper than previous similar works, making it suitable for multiple scenarios.

2 Proposed Method

This section describes the different stages of our system, from the calculation of optical flow fields and division in spatio-temporal volumes to the extraction of features and construction of the Gaussian mixture models. Figure 1 shows the main stages of the system, in which we can find both training and test phases.

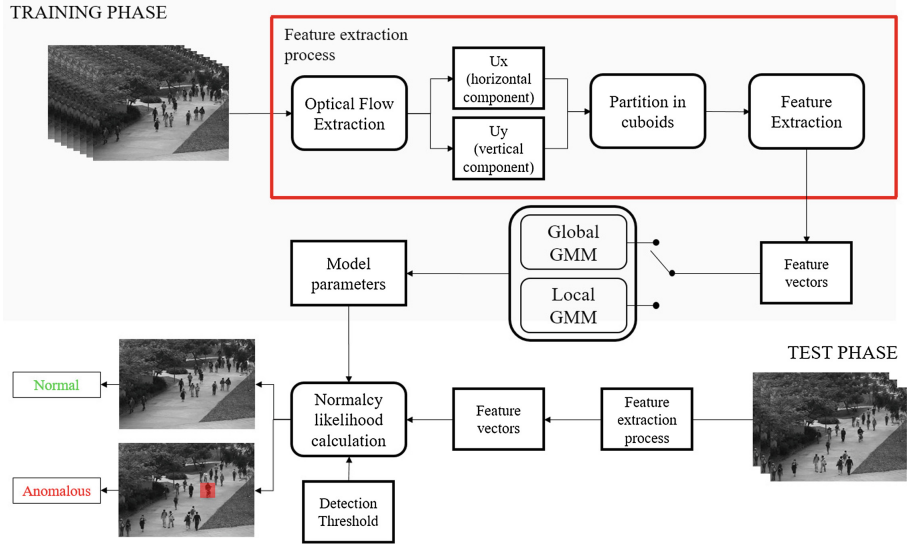


Fig. 1. Our anomaly detection system

2.1 Robust Optical Flow

The first stage corresponds to the extraction of features, which are based on optical flow. These give accurate motion information at pixel level, essential for the detection of anomalies. In this regard we can highlight the methods based on the Horn-Schunck (HS) formulation, which offers reasonable results.

Nevertheless, many recent methods based on other assumptions have better precision calculating the field. Still, authors of [10] showed that the original HS assumptions are competitive by incorporating extra stages to the process. These are: decomposition in structure and texture of the scene, the use of a specific derivative mask instead of image differences, multi-resolution pyramids, the use of weighted median filters after the warping steps to remove outliers and a graduated non-convexity (GNC) approach for the use of different penalty functions. In our contribution we have introduced two modifications. The first one is the elimination of the structure and decomposition stage, since its use does not add clear improvements in the accuracy of the field but it needs remarkable processing time. The second one is the application of only one iteration after the computation of the optical flow for each level of the pyramid, instead of using the three of the original method.

The stages that most improve the accuracy of the final flow field are the GNC approach and the application of the median filter. By using the first, we gain accuracy in the flow by combining the effect of a simple function (convex) and a more robust penalty function (non-convex). The second is useful to remove outliers in the estimation of the flow. To validate our method for the extraction of the optical flow field, we have tested it in the Middlebury dataset [13], confirming that it renders more accurate than original HS. In Fig. 2 we can see the benefits

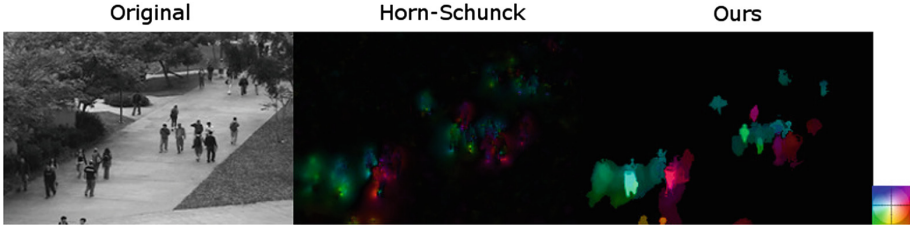


Fig. 2. Representation of optical flow fields obtained with Horn-Schunck and the proposed method. Intensity shows the magnitude of the flow vectors. Color represents the orientation. (Color figure online)

of using the proposed method. For instance, along the surface of the objects, the optical flow vectors are homogeneous and free of outliers (intensity values are normalized using the maximum magnitude value).

Once the optical flow fields are obtained, they are organized in spatio-temporal volumes (cuboids) of a specific size. These cuboids are the units from which the features are extracted and used to train the normality model. In the training phase, these cuboids are taken with no overlapping in the temporal dimension, while in the test phase new cuboids are constructed every time a new frame enters the system.

2.2 Feature Extraction

We have selected a set of significant optical flow features used in the literature. The idea is to combine them to create different descriptors and select those with the best capacity to detect anomalies. The set includes magnitude of optical flow and uniformity (also called texture) of optical flow, both proposed in [7] and the histogram of optical flow, in a similar manner to [4, 8, 14].

The magnitude of optical flow (F1) is calculated by making the summation of optical flow vectors components u_i and v_i over the total number of pixels N of the current cuboid. The second feature (F2) describes the uniformity (texture) of the optical flow vectors respect to pixels located at a distance (offset) of δ pixels. The third (F3) corresponds to the unweighted histogram of optical flow orientations, represented with n bins (each bin counts the orientations up to $(2i + 1)\pi/n$ degrees, where i is the index of the current bin).

Note that in the case of using the global GMM, the central position of the cuboid (x, y) is introduced into the descriptor, allowing to locate where the events are produced and increasing the model discriminative capabilities.

$$\mathbf{F1:} \quad \omega_x = \sum_i^N u_i, \quad \omega_y = \sum_i^N v_i \tag{1}$$

$$\mathbf{F2:} \quad \phi(\delta) = \sum_i^N u(i)u(i + \delta) + v(i)v(i + \delta) \tag{2}$$

$$\mathbf{F3:} \quad H_{OF}^n = \bigcup_{i=0}^{n-1} \{h_{\frac{(2i+1)\pi}{n}}\} \tag{3}$$

2.3 Gaussian Mixture Models: Local and Global Approaches

A GMM is a parametric model composed by K multivariate Gaussian distributions, each of them with a weight π_k , a covariance matrix Σ_k and a vector μ_k with the means of each descriptor D of size n . The likelihood of a sample given the parameters is calculated as described in (4), which is the weighted summation of the likelihood of the sample over all the distributions of the GMM (5). For detecting anomalies, we use two approaches to create the normality model. The first one uses a unique GMM distributed over the entire scene, similar to that proposed in [7, 8, 18]. The second approach sets one GMM per spatial location. Thus, only spatio-temporal volumes (cuboids) entering that specific location as time passes are used to build the local models (Tziakos et al. [9] use local GMM and Kratz and Nishino [3] use just one distribution per location).

Models are constructed using the EM algorithm and K-means++ [15] is applied for initial clustering. In the test phase, the samples are marked as anomalies if their likelihoods over the GMM exceed the threshold that indicates the boundary of normality. In the case of the global GMM, only a unique threshold is necessary, while one threshold per model is required for the implementation of the local GMM approach. Note that to avoid extra processing time we use diagonal covariance matrices instead of full matrices, since they provide good results and avoid extra processing time.

$$p(D|\Theta) = \sum_{k=1}^K \pi_k p(D|\mu_k, \Sigma_k) \quad (4)$$

$$p(D|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(D - \mu)^T \Sigma^{-1}(D - \mu)\right) \quad (5)$$

3 Results and Discussion

As in most of the related works, we have used the UCSD dataset [1] for evaluation. This dataset contains two different sets of sequences (Figs. 3 and 4) taken in a public walkway. The first set of sequences (ped1) has some perspective distortion. We have used frame-level and pixel-level criterion for the evaluation of the system. In order to perform the correlation feature selection technique and the determination of the best parameters such as the best cuboid size or the



Fig. 3. Detections on ped1 dataset



Fig. 4. Detections on ped2 dataset

appropriate number of components of the mixture, we use the global GMM approach. Right after, evaluation with local GMMs is accomplished with the best cuboid size.

3.1 Correlation Feature Selection

In order to make an objective selection of the most effective descriptor, we have performed a statistical evaluation of the possible combinations of the proposed features following the correlation feature selection (CFS) measure [12]. It works under this idea: *Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other.* With CFS it is possible to calculate the contribution that each subset of features has on the final classification by finding a score called *merit*. To do so, the average feature-feature and the average feature-classification correlations are computed under the global GMM approach.

Either detections or ground truths, signals are composed by 1s where a frame is anomalous and -1 s otherwise. For the first correlation, the detection signals under each feature are used, while for the second one, the detection signal for one feature and the ground-truth signal are used (detection signals are extracted at the point of equal error rate). The correlations are calculated by superposing the two signals and summing the element-wise product. Feature selection has been applied on ped1 dataset with an increasing number of components (from 10 to 80 in intervals of 10) and averaging the correlation values. The cuboid size is fixed to $9 \times 9 \times 7$ pixels, similarly to other methods in the literature [1, 7]. We have chosen offsets (δ) of 1, 3 and 5 pixels for the textures of optical flow (F2).

Results demonstrate that the incorporation of the optical flow histogram (with any number of bins) not only do not improve the detection rate but in many cases it significantly diminishes the performance. Therefore, the final descriptor D used to train the models is composed by features F1 and F2, using the three offsets δ proposed for F2.

$$\mathbf{D} : (x, y, \omega_x, \omega_y, \phi_{(1)}, \phi_{(3)}, \phi_{(5)}) \quad (6)$$

3.2 Global and Local GMM: Comparison

Global GMM: Varying a global threshold over the normality likelihood, true and false positive rates (TPR and FPR) are calculated and the corresponding

receiver operating curve (ROC) is constructed. The area under the curve (AUC) and equal error rates (EER) are extracted and used for comparing of results. On the other hand, pixel-level criterion is stricter, since at least 40% of the true anomalous area [1] must be covered by the area detected as anomaly. Thus, it is possible to discard detections that would be tagged as true detections with the first criterion. This occurs when the frame contains an anomaly but the detected cuboids do not coincide with the real anomalous area.

On both sets of sequences, the selected cuboid size is of $9 \times 9 \times 7$ pixels. Values of AUC and EER at frame level are calculated with a number of GMM components within the range from 10 to 500 on ped1 and up to 300 on ped2, whose event variability is lower. Additionally, to test the impact of perspective correction, results for ped1 are also obtained after applying the technique described in [11].

Including perspective correction, the average gain in terms of AUC is 3.7% on ped1, reaching a maximum of 0.8977 with 370 GMM components. On ped2, the maximum AUC is 0.9629 with 90 components. As expected, the best number of GMM components is directly related to the variability of normality behaviors in the scene: ped1 shows much more variability than ped2. Therefore, depending on the potential diversity of events in the scene under analysis, the range of GMM components to be used (or explored) can be estimated.

Besides, it is important to remark that our proposal shows significant robustness to the number of components used: for 92% of the tests modifying the number of components in ped1, the AUC deviates less than 1.5% from the maximum value; for ped2, 87% of the tests deviates less than 3.5% from the maximum AUC. Indeed, the maximum deviation from the maximum AUC value considering the whole range is 3.17% and 4.07% for ped1 and ped2 respectively.

Local GMM: One of the main advantages of the Global GMM approach, is that the probability distribution of the GMM adapts to the events that occur on the scene, in such a way that those areas with heterogeneous activities (different motion magnitudes, orientations and uniformities) will concentrate more Gaussian components, reserving less components for other areas. With the local GMM approach, the number of Gaussian components per location has to be fixed a priori. This is a problem, since it is very difficult to foresee the events that occur in each area, and consequently try to infer an appropriate number of components for the local models. In this regard, we propose a method that estimate this number of GMM components so we can obtain the maximum performance, in a similar way to the proposed by Tziakos et al. [9]. The idea is to infer how well estimated is a model based on the likelihood obtained after introducing the same training samples to the GMM distributions. Thus, a model is considered to be worse trained if its normality likelihood is low. To avoid the effect of outliers, we take the average likelihood of the training samples corresponding to their models. In Figs. 5 and 6 we can see how the value of Area Under the Curve (AUC) changes when using different Gaussian components. Using this information, we select the number of components depending on the moment in which the value of AUC is stabilized.

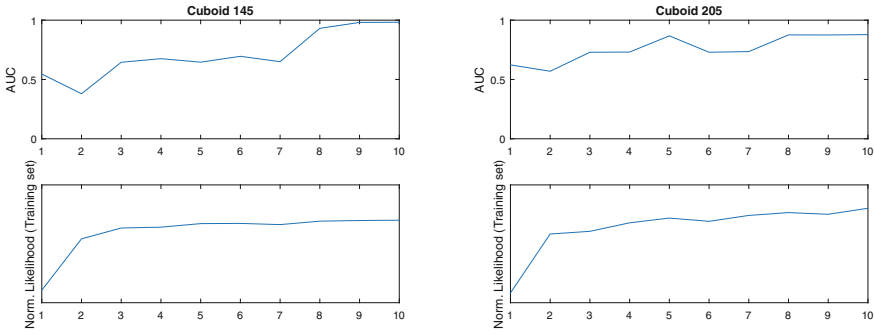


Fig. 5. Normality likelihood of the training set vs area under the curve obtained in the test phase (ped1)

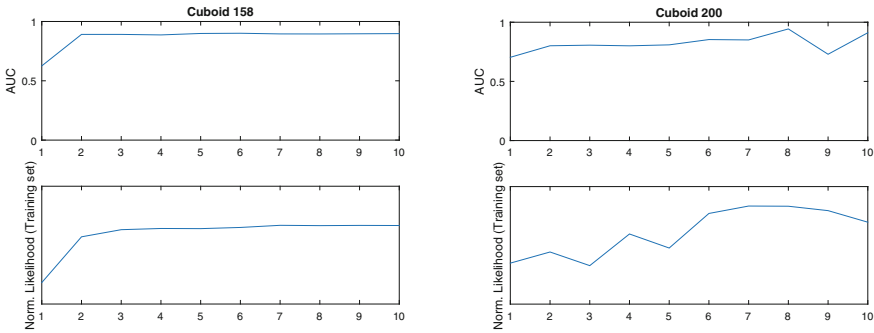


Fig. 6. Normality likelihood of the training set vs area under the curve obtained in the test phase (ped2)

To validate this idea, we have computed the ROC curves of those locations that in the test phase contain anomalies, extracting the value of area under the curve (AUC). Then, the values of AUC with different number of Gaussian components is related with the average likelihood obtained with the training samples, as shown in Figs. 5 and 6 (from 1 to 10 components on ped1 and ped2 datasets). As we can see, a better trained model (higher normality likelihood on the training set), has correlation with a better detection rate (higher AUC values). Thus, we select the number of components that maximize the average likelihood of the training samples to build the final ROC curve at frame level so we can compare the performance against global GMM approach (see Fig. 7).

The performance of the global GMM is much better in terms of AUC and EER than the local GMM approach. This is caused by the independence that local models have between each other, so when an event that covers more than one location appears, it is difficult to detect anomalies. This do not occurs when using the global model, that manages global events but at the same time their locality thanks to the addition of the cuboid position into the descriptor.

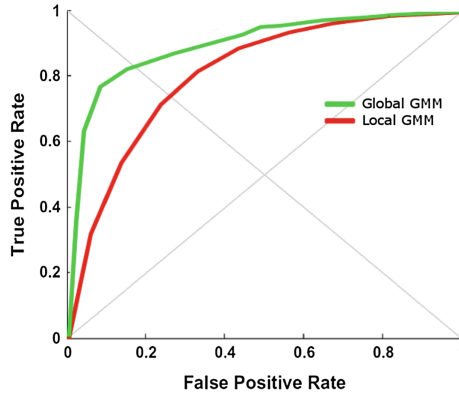


Fig. 7. Best result for global and local GMM approaches

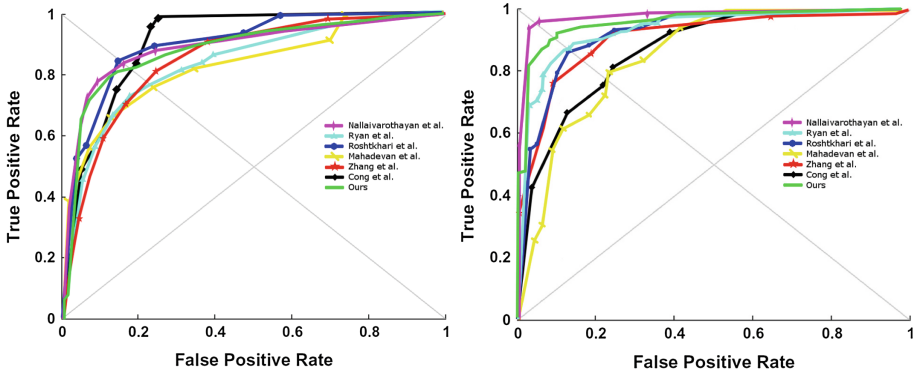


Fig. 8. ROC for ped1 and ped2 at frame level.

Figure 8 portrays how the system performs in the sequences of ped1 and ped2 datasets (with the global GMM approach from now on). Compared to other methods in the literature, we can conclude that our method is very competitive. For instance, on ped1 dataset, we obtain a similar performance to that of Nallaivarothayan et al. [8], Roshtkhari et al. [2] and Cong et al. [4] and notably better than the results obtained by Mahadevan et al. [1] and Zhang et al. [17]. Additionally, we have to highlight that we outperform the method proposed by Ryan et al. [7], where the same descriptors are used but a different optical flow method and the normalization of perspective proposed in [11] are employed, confirming their effectiveness. This is also visible on ped2, in which no perspective normalization is needed and the difference in performance respect to [7] is caused by the different optical flow algorithm used.

Our system is also quite competent for anomaly localization, as we can see in Fig. 9 (ped1 dataset). The performance of our system is similar to the proposed by Zhang et al. and Li et al. [16], being the last method the updated version of

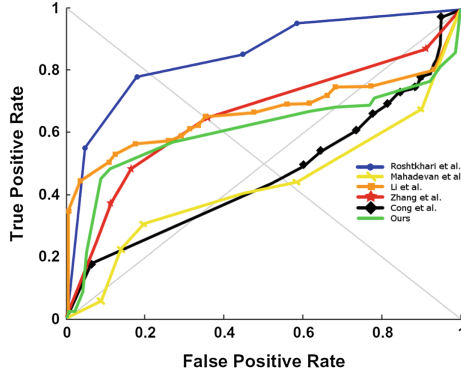


Fig. 9. ROC for ped1 dataset at pixel level.

Table 1. Performance comparison at frame and pixel level on UCSD dataset.

EER (%) frame level			EER (%) pixel level	
Method	ped1	ped2	Method	ped1
Mahadevan et al. [1]	25.00	25.00	Mahadevan et al. [1]	55.00
Roshtkhari et al. [2]	15.00	13.00	Li et al. [16]	35.00
Cong et al. [4]	19.00	21.70	Roshtkhari et al. [2]	27.00
Zhang et al. [17]	22.00	16.78	Cong et al. [4]	54.00
Ryan et al. [7]	23.10	12.70	Zhang et al. [17]	35.60
Nallaivarothayan et al. [8]	14.90	4.89	Ours	39.00
Ours	16.38	9.34		

Mahadevan et al., which together with Cong et al. have worse results than ours. Exact EER values for both ped1 and ped2 at frame level and ped1 at pixel level are given in detail in Table 1.

Note that we have used the original ground truth as proposed in [1], in which not all the sequences of ped1 dataset are labeled. The full ground truth was released later. This is important because some authors compare the results obtained with the two ground truths indistinctly, making the comparisons futile.

4 Conclusions and Future Work

The main contributions of this paper are: the construction and comparison of two GMM approaches (local and global) for the detection of anomalies; the use of a robust optical flow method for the construction of the descriptors and the application of a statistical tool based on correlation (CFS) for the selection of the most discriminative features. The global GMM approach is finally used, since is the most effective for the task of anomaly detection in surveillance scenarios. In fact, we obtain better results than similar approaches in the literature at frame and pixel level.

In the future, we intend to make the system to work in real time so it can be utilized in public surveillance scenarios. To do so, the idea is to parallelize the processes of training and test for speeding up the detection stage.

Acknowledgements. This work has been partially supported by the Ministerio de Economía, Industria y Competitividad of the Spanish Government and the European Regional Development Fund (AIE/FEDER) under projects TEC2013-48453 (MR-UHDTV), RTC-2015-3527-1 (BEGISE) and TEC2016-75981 (IVME).

References

1. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981 (2010)
2. Roshtkhari, M.J., Levine, M.D.: An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Comput. Vis. Image Underst.* **117**(10), 1436–1452 (2013)
3. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1446–1453 (2009)
4. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3449–3456 (2011)
5. Morris, B.T., Trivedi, M.M.: Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(11), 2287–2301 (2011)
6. Saligrama, V., Chen, Z.: Video anomaly detection based on local statistical aggregates. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2112–2119 (2012)
7. Ryan, D., Denman, S., Fookes, C., Sridharan, S.: Textures of optical flow for real-time anomaly detection in crowds. In: IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 230–235 (2011)
8. Nallaivarothayan, H., Fookes, C., Denman, S., Sridharan, S.: An MRF based abnormal event detection approach using motion and appearance features. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 343–348 (2014)
9. Tziakos, I., Cavallaro, A., Xu, L.Q.: Local abnormality detection in video using subspace learning. In: IEEE Conference on in Advanced Video and Signal Based Surveillance (AVSS), pp. 519–525 (2010)
10. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2432–2439 (2010)
11. Nallaivarothayan, H., Ryan, D., Denman S., Sridharan, S., Fookes, C.: An evaluation of different features and learning models for anomalous event detection. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8 (2013)
12. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML), pp. 359–366 (2000)

13. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. *Int. J. Comput. Vision* **92**(1), 1–31 (2011)
14. Yuan, Y., Fang, J., Wang, Q.: Online anomaly detection in crowd scenes via structure analysis. *IEEE Trans. Cybern.* **45**(3), 548–561 (2015)
15. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics (2007)
16. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(1), 18–32 (2014)
17. Zhang, Y., Lu, H., Zhang, L., Ruan, X.: Combining motion and appearance cues for anomaly detection. *Pattern Recogn.* **51**, 443–452 (2016)
18. Saleemi, I., Hartung, L., Shah, M.: Scene understanding by statistical modeling of motion patterns. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2069–2076 (2010)