Detection of Anomalies in Surveillance Scenarios using Mixture Models

Adrián Tomé, Luis Salgado

***Abstract*—In this paper we present a robust and simple method for the detection of anomalies in surveillance scenarios. We use a "bottom-up" approach that avoids any object tracking, making the system suitable for anomaly detection in crowds. A robust optical flow method is used for the extraction of accurate spatio-temporal motion information, which allows to get simple but discriminative descriptors that are employed to train a Gaussian mixture model. We evaluate our system in a publicly available dataset, concluding that our method outperforms similar anomaly detection approaches but with a simpler model and low-sized descriptors.**

***Index Terms*—Anomaly detection, Gaussian mixture model, robust optical flow, video-surveillance.**

I. Introduction

Detection and recognition of events in surveillance scenarios has become a popular task in the past years. Within this category, anomaly detection proposals are of special interest regarding the amount of works developed [1]–[4], [6], [7], [14]. In spite of this, it is still an open task due to the multiple problems that it generates. The most important issues are: the difficulty to define what is an anomaly; the variability of the anomalous events; and the scarcity of training samples to build models. Regarding the definition of anomaly, sometimes they are defined as irregular or unusual events [4], while in other cases these events are defined as events that differ from those considered normal [1] or the low frequency of appearance [2]. On the other hand, there are two major approaches to face the process of detection. The most used in the past is the "top-down" approach, that requires of a tracking phase in which the objects need to be detected. After this phase, trajectory patterns can be analyzed to find anomalous motion activity. The major problem with this approach is that the detection rate of anomalies decreases rapidly when applied in crowded scenarios, where occlusions and clutter appear. "Bottom-up" approaches mitigate these problems by analyzing the events firstly at pixel level and then inferring information at a higher level. This has become the most popular approach in more recent works [2]–[4], [6]. Typical low-level features are gradients [2], [3] or features based on optical flow [4], [7].

There are some state-of-the-art proposals to outline. For instance, Roshtkhari et al. [2] propose to learn a model of low-level features extracted from spatio-temporal compositions. The model is updated in an online manner and without supervision. The main drawback is that the complexity of the system is elevated.

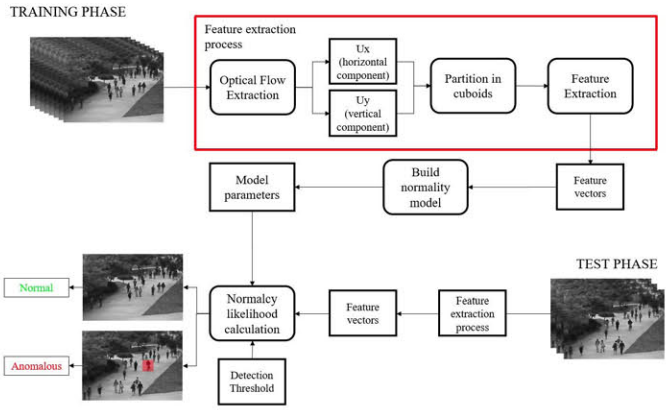

Fig. 1. Proposed anomaly detection system.

Another important approach was proposed by Cong et al. [4], method in which a set of representative samples is used to construct a basis. Its problem is that an accurate method for the selection of the samples needs to be employed.

One of the most significant approaches is detailed in [1]. They propose to use dynamic textures in a innovative manner, learning the appearance patterns of the input sequences so anomalies can be detected as outliers. The main disadvantage that it presents is related to the computational cost. Besides, it has been outperformed by other most recent works.

Ryan et al. [7] used a different approach. They implement a Gaussian mixture to create a normality model, using for that descriptors that incorporate flow-based information. This work inspired the proposal made in [9], whose method incorporates orientation histograms into the descriptors and a Markov random field to increase the detection rate of the system.

In this paper we present a simple but effective method to detect anomalies through the use of flow-based features, that feed a Gaussian mixture model that learns the normality of the scene so anomalies can be detected as samples that differ enough from the trained model. The simplicity of the system allows to apply the method to multi-camera systems, where the computational complexity is vital.

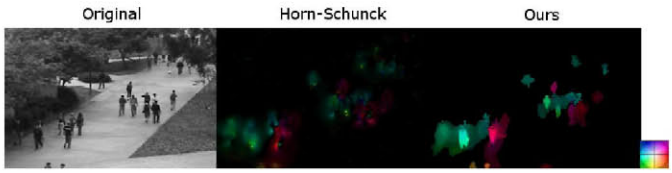

Fig. 2. Representation of optical flow fields obtained with Horn-Schunck and the proposed method. Intensity shows the magnitude of the flow vectors. Color represents the orientation.

II. PROPOSED METHOD

The main stages of our proposal are shown in Figure 1. We start with the extraction of the flow fields divided in spatio-temporal volumes (cuboids) and we continue with the construction of the Gaussian mixture model.

A. Extraction of optical flow

The popular Horn and Schunck (HS) formulation has been widely used for obtaining motion information between frames at pixel level. This method has become obsolete, since many approaches for the extraction of the optical flow with different formulations proposed in the next years get more accuracy in the final field.

In spite of this, authors of [10] claimed that the Horn-Schunck formulation can be still competent if the appropriate stages are added to the algorithm. These stages 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 on the functional to optimize. Two modifications are here proposed to simplify the approach while keeping its accuracy. The first one is the removal of the decomposition in structure and texture stage because it does not add clear improvements in the final field but it needs remarkable processing time. The second one is the application of only one iteration to recompute the optical flow before calculating it at a different resolution.

The GNC approach and the application of the median filter are the two stages that have the most remarkable effect on the final flow field. The first permits to gain accuracy 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 the original HS. In Figure 2 we can see the benefits 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).

B. Construction of the Gaussian mixture model

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 variable D of size n. The likelihood of a sample given the parameters is calculated as described in Eq. 1, which is the weighted summation of the likelihood of the sample over all the distributions of the GMM (Eq. 2). For detecting anomalies, we use a global GMM as in [7] and [9] to create a normality model, as it allows to model the events with a unique probability distribution. The GMM is constructed using the EM algorithm. K-means++ [15] is used for initial clustering. Finally, in the test phase, the samples are labeled as anomalies if their likelihoods over the GMM are below a fixed threshold. Note that we use diagonal covariance matrices instead of full matrices, since they provide good results while avoiding extra processing time.

There exist also the possibility of using local Gaussian mixture models. To do so, we need to build one GMM per spatial location of the scene. This approach permits to process the events that occur in the spatio-temporal cuboids independently and is easier to parallelize. Nevertheless, results demonstrate that the detection rate is much lower compared to the global approach, so it has been discarded, and the evaluation is performed only with the global approach.

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

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

III. FEATURE EXTRACTION

We have selected a set of features from the literature to verify what is the combination that offers the best detection rates. These features are: magnitude of optical flow and uniformity (or textures of optical flow), proposed in [7] and histogram of optical flow, similarly to [4], [9], [14].

$$\textbf{F1:} \quad \omega_x = \sum_{i}^{N} u_i, \quad \omega_y = \sum_{i}^{N} v_i$$

$$\textbf{F2:} \quad \phi(\delta) = \sum_{i}^{N} u(i)u(i+\delta) + v(i)v(i+\delta)$$

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

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 i counts the orientations up to $(2i+1)\pi/n$ degrees).

Note that to locate where the events are produced while increasing the model discriminative capabilities, the central

Fig. 3. Detections on ped1 dataset

Fig. 4. Detections on ped2 dataset

position of the cuboid (x, y) is introduced into the descriptor. We have evaluated the performance of each combination of features by verifying the detection rate that we get on UCSD and UMN datasets, trying different numbers of Gaussian components within a range between 10 and 80. The best combination of features is chosen by selecting the best areas under the curve (AUC) obtained from the receiver operating curves (ROC).

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 significantly diminish the performance. Thus, the final descriptor D used to train the model is composed by features F1 and F2. The second feature is used applying the three values of offset δ.

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

IV. RESULTS AND DISCUSSION

For evaluation, we have selected the UCSD dataset. It contains two different sets of sequences with images of a public walkway. The first set of sequences (ped1) has some perspective distortion. For measuring the performance of the system, we make use of true positive and false positive rates (TPR and FPR). They are specified in Eq. 3. The number of true positives, false negatives, false positives an true negatives are involved in the calculation.

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN} \tag{3}$$

We have used frame-level and pixel-level criterion for the evaluation of the system. With the first, for classifying a test frame as anomalous, it is enough that the likelihood of one cuboid over the GMM is lower than a fixed threshold. Varying this threshold, TPR and FPR are calculated and the corresponding receiver operating curve (ROC) is computed. The area under the curve (AUC) and equal error rates (EER) are extracted and used for result comparison. On the other hand, pixel-level criterion is stricter, since the area detected as anomaly must also cover at least the 40% of the segmented anomalous area [1]. With the final descriptor fixed, an analysis of cuboid sizes has been carried out: 9x9, 12x12 and 15x15 spatial sizes have been combined with 7, 10 and 13 frames as temporal size. For each combination, the 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].

On both datasets the best cuboid size is of 9x9x7 pixels. 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.

Figures 3 and 4 show some detections on ped1 and ped2 datasets, and Figures 5 and 6 portray the ROC curves for different methods of the state-of-the-art at frame level. The equal error rates are given in Table I, including EER values on ped1 at pixel level. At frame level, we obtain better results than all the methods except two: the method proposed by [2] only for ped1 dataset and [9] in both. In any case, the differences respect to [9] are very small, particularly if we take into account that they use a GMM but with larger descriptors and a Markov random field to model the co-occurrence of events. Additionally, the effectiveness of the robust optical flow and perspective normalization methods that we apply is proven when comparing our results with those obtained in [7], since although they use the same type of model and descriptors, their optical flow method is worse and no perspective correction is applied.

In terms of anomaly localization (EER at pixel-level in Table I), our results are above the obtained by the methods in the literature, except the proposed by [2], that has a lower equal error rate.

In terms of processing time, we need 0.4 s/frame including the calculation of the optical flow and the frame classification (we use 80 components for the GMM). It outperforms [4] (3.8 s/frame) and [1] (25 s/frame), and it is close to [2] and [14]. The method proposed in [7] operates at 0.1 s/frame (downsized images with a simpler optical flow), rendering significantly worse detection rates. In [9] no processing time is given.

EER (%) frame level		
Method	ped1	ped2
Mahadevan et al. [1]	25.00	25.00
Roshtkhari et al. [2]	15.00	13.00
Cong et al. [4]	19.00	21.70
Yuan et al. [14]	17.65	14.12
Ryan et al. [7]	23.10	12.70
Nallaivarothayan et al. [9]	14.90	4.89
Ours	16.38	9.34

EER (%) pixel level	
Method	ped1
Mahadevan et al. [1]	55.00
Roshtkhari et al. [2]	27.00
Cong et al. [4]	54.00
Yuan et al. [14]	41.00
Ours	39.00

TABLE I
PERFORMANCE COMPARISON AT FRAME AND PIXEL LEVEL.

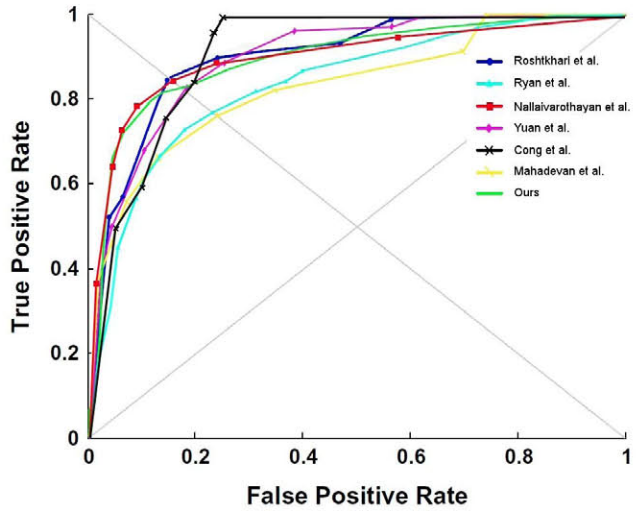

Fig. 5. ROC for ped1 dataset at frame level.

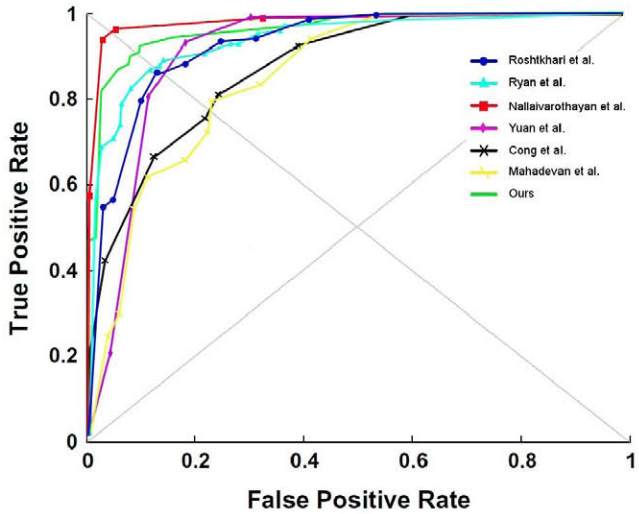

Fig. 6. ROC for ped2 dataset at frame level.

V. Conclusions and Future Work

We have presented an effective while simple method for detecting anomalies in surveillance scenarios. We can conclude that the best model is the global GMM approach, that is more effective than the local GMM approach, while needing the construction of only one probability distribution. Results confirm the goodness of our system, that obtains better results than many state-of-the-art methods on UCSD and UMN datasets. As future work, we plan to make the system to work in real time, as well as improve the descriptors and increase the range of the parameters used. We also intend to use more databases for the evaluation of the system.

Acknowledgment

This work has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government and the Fondo Europeo para el Desarrollo Regional of the European Union under the project RTC-2015-3527-1 (BEGISE). The work of Luis Salgado has been also partially supported by the project TEC2014-53176-R (HAVideo).

References

[1] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes", Computer Vision and Pattern Recognition (CVPR), pp. 1975-1981, 2010.

[2] 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 and Image Understanding, vol. 117, no. 10, pp. 1436-1452, 2013.

[3] L. Kratz, and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models", IEEE Confrence on Computer Vision and Pattern Recognition (CVPR), pp. 1446-1453, 2009.

[4] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3449-3456, 2011.

[5] B. T. Morris, and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach". IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 33, no 11, pp. 2287-2301, 2011.

[6] V. Saligrama, and Z. Chen, "Video anomaly detection based on local statistical aggregates", Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2112-2119, 2012.

[7] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Textures of optical flow for real-time anomaly detection in crowds", IEEE Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 230-235, 2011.

[8] Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly Detection and Localization in Crowded Scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no 1, 18–32 (2014)

[9] H. Nallaivarothayan, C. Fookes, S. Denman, and S. Sridharan,"An MRF based abnormal event detection approach using motion and appearance features", IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 343-348, 2014.

[10] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2432-2439, 2010.

[11] H. Nallaivarothayan, D. Ryan, S. Denman, S. Sridharan, and C. Fookes,"An evaluation of different features and learning models for anomalous event detection", Digital Image Computing: Techniques and Applications (DICTA), pp. 1-8, 2013.

[12] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning", Proc. International Conference on Machine Learning (ICML), pp. 359-366, 2000.

[13] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow", International Journal of Computer Vision (IJCV), vol. 92, no. 1, pp. 1-31, 2011.

[14] Y. Yuan, J. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis", IEEE Transactions on Cybernetics, vol. 45, no. 3, pp. 548-561, 2015.

[15] D. Arthur, and S. Vassilvitskii, "k-means++: The advantages of careful seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, pp. 1027-1035, 2007.