A VIDEO ANALYTICS FRAMEWORK FOR AMORPHOUS AND UNSTRUCTURED ANOMALY DETECTION

Martin Mueller, Peter Karasev, Ivan Kolesov, Allen Tannenbaum

Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, GA, USA

ABSTRACT

Video surveillance systems are often used to detect *anomalies*: rare events which demand a human response, such as a fire breaking out. Automated detection algorithms enable vastly more video data to be processed than would be possible otherwise. This note presents a video analytics framework for the detection of amorphous and unstructured anomalies such as fire, targets in deep turbulence, or objects behind a smoke-screen. Our approach uses an off-line supervised training phase together with an on-line Bayesian procedure: we form a prior, compute a likelihood function, and then update the posterior estimate. The prior consists of candidate image-regions generated by a weak classifier. Likelihood of a candidate region containing an object of interest at each time step is computed from the photometric observations coupled with an optimal-mass-transport optical-flow field. The posterior is sequentially updated by tracking image regions over time and space using active contours thus extracting samples from a properly aligned batch of images. The general theory is applied to the video-fire-detection problem with excellent detection performance across substantially varying scenarios which are not used for training.

Index Terms— Anomaly Detection, Video Analytics, Machine Vision, Active Contours

1. INTRODUCTION

With increasing amounts of video surveillance data, there is acute need for increased automation in *event detection*. Surveillance systems aim to detect certain events that occur rarely and exhibit unusual behavior such as a person entering a restricted area. The detection of rare and unusual events is commonly called anomaly detection.

This paper deals with digital, color video sequences. In contrast to one-dimensional signals, video signals not only possess temporal but also spatial characteristics, which greatly increases complexity. This complexity can be attributed to an event being confined to a subset of the whole frame. The type of anomaly considered in this research are amorphous, diffuse and unstructured events such as fire, smoke, objects blurred by water reflections, a crowd of animals in the distance, etc. Unlike clear-cut, structured objects that are commonly detected using shape-based methods, such as principal component analysis (PCA) [1], these diffuse events do not share distinct shape characteristics. Other methods are needed, therefore, to identify the existence and location of these events.

In the following sections, a general detection framework for amorphous and unstructured events in video sequences by means of active contours and supervised classification is presented. The framework provides consistent sampling regions (candidate regions) apart from surrounding clutter and elegantly handles the generation of ground truth data for the supervised training procedure. In comparison to unsupervised approaches [2], supervised classification is expected to produce more reliable results at the cost of having to adjust the algorithm to a specific event. Other publications related to this field address the detection of specific unstructured events such as fire [3], smoke [4] or crowds of people [5], but they do not consider general approaches.

Section 2 introduces the proposed framework, explaining the key components in detail. With minor changes to the detection framework, a tool for the convenient generation of ground truth data, which is necessary for training and testing of supervised classification, is shown in Section 3. As a demonstration, Section 4 applies the framework to a fire detection algorithm [6] showing its effectiveness in different scenarios.

2. THE FRAMEWORK

Video anomaly detection has an inherent difficulty: in the presence of simultaneous events, sampling the whole frame produces a blended picture of the events. This, consequently, diminishes the chance of detecting a particular event because its characteristics are obscured by other events occurring simultaneously. To address this difficulty, the following section proposes a three-step framework (see Fig. 1), which

1. declares a region that exhibits suspicious behavior as a candidate region,
2. tracks candidate regions over time to continuously monitor the suspicious events,
3. performs an anomaly test within each candidate region.

Before turning to aspects of implementation, the above procedure is interpreted and justified from a probabilistic point of view. Let $\Omega \in \mathcal{I}$ be the frame at time t_k in the space $\mathcal{I}$ of all possible images, and a candidate region be a subframe $\Omega_i \subset \Omega$. Bayes Theorem states that

$$\mathrm{P}(\mathrm{Event}|\Omega_i) \propto \mathrm{P}(\Omega_i|\mathrm{Event})\,\mathrm{P}(\mathrm{Event}). \tag{1}$$

The *posterior probability* $\mathrm{P}(\mathrm{Event}|\Omega_i)$ that the anomalous event occurs in the given image Ω_i is proportional to the *likelihood* $\mathrm{P}(\Omega_i|\mathrm{Event})$ and the *prior probability* of the event $\mathrm{P}(\mathrm{Event})$. The proposed framework indeed leads to a high posterior probability in the presence of an anomaly, as explained next.

Step 1 may be viewed as a restriction on the image space. Instead of propagating any image, however unlikely it is to contain

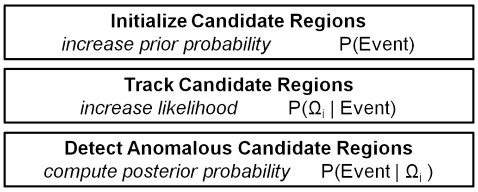


Fig. 1. Detection framework and probabilistic interpretation

an anomaly, to the classification process, step 1 filters images and passes to step 2 only those images (candidate regions) that belong to a subset $\mathcal{I}_s \subset \mathcal{I}$: the suspicious images. Thus, the probability of randomly picking an anomalous event image from $\mathcal{I}_s$ is greater than picking one from $\mathcal{I}$. In other words, by discarding from consideration images that are clearly normal, the prior probability P(Event) in Eq. (1) is increased.

Step 2 is designed to increase the likelihood term P(Ω_i|Event) in two ways. First, by segmentation of regions with similar motion statistics in one frame, a homogeneous sampling region is created. Assuming an anomalous event in Ω_i, any normal behavior included in Ω_i is considered noise, since it decreases the sampling homogeneity; the better the segmentation, the higher the likelihood P(Ω_i|Event). Secondly, tracking these candidate regions over time facilitates track-before-detect methods to remove the noise content from mis-segmentation, as will be illustrated in Section 2.3.

Step 3, in general, computes the posterior P(Event|Ω_i) in Eq. (1) via feature extraction and Bayesian classification [1] in each frame. Here, in addition to mis-segmentation, misclassification contributes to noise content. A track-before-detect method filters the classification output and draws final conclusions as to the detection output.

The following subsections present aspects of implementation, which is based on two major assumptions. First, a supervised approach is considered. In particular, it is assumed that sample videos of events of interest are available for training and testing. In some applications, where the event is very rare and hard to capture, it might be difficult to obtain these data. For many common applications, however, videos can be produced or retrieved from existing databases. Second, the anomalous event is characterized by a continued and considerable amount of motion (as opposed to events that occur very slowly or very shortly). Step 1 is realized as a weak classifier initializing active contours, which are to segment the frame in step 2 based on an optical flow motion field. A supervised classification and subsequent track-before-detect posterior filtering yield the final detection result in step 3.

2.1. Initializing Candidate Regions: The Prior

The first step provides an initial, rough estimate about suspicious regions. Due to the assumption that an anomoly in a video is moving, any background estimation technique will serve as a weak classifier. If the result is too weak, e. g. the video's content is very dynamic, the background estimation may be combined with other simple features such as color space models. Even though there exist heuristic color models for different applications, for example fire [7], we propose using a supervised classifier because the supervised framework is already assumed for the final detection step and, hence, training and testing data is available. In contrast to heuristic models that strongly depend on the parameter choice for specific scenarios, e. g. dependence on lighting conditions, supervised classifiers can learn more complicated dependencies if provided sufficient training data.

Several weak classifiers based on simple motion detection, color, and other low-level features are applied to each image. The intersection of pixels passing each test serve as a rough guess of areas containing suspicious activity in an image. These pixels are clustered using an algorithm such as k-means and thus used as the initialization for active contours, which accurately segment and track the candidate regions over time.

2.2. Tracking Candidate Regions: The Likelihood

We use active contours to segment an object in a candidate region. Details on the formulation and evolution of variational active contours can be found in [8]. Region based segmentation can be generally described by Eq. (2) where the final position of the contour is a local minimum of Eq. (2) and is the final segmentation: inside of the contour is the object and outside is the background.

$$\min_{\phi} \quad \int_{\Omega} F(I(y), \phi(y)) dy \tag{2}$$

In Eq. (2), I is the image, $\Omega \subset R^2$ is the image domain and $\phi(y)$ is the level set function that embeds the segmenting curve as its level set. A particular choice of the function F is given by Eq. (3) [9] but any region based model can be used in this framework.

$$F = \mathcal{H}(\phi)(I-u)^2 + (1-\mathcal{H}(\phi))\,(I-v)^2, \tag{3}$$

where $\mathcal{H}$ is the Heavyside step function, and

$$u = \frac{\int_\Omega \mathcal{H}(\phi) I(y) dy}{\int_\Omega \mathcal{H}(\phi) dy}, \quad v = \frac{\int_\Omega (1-\mathcal{H}(\phi)) I(y) dy}{\int_\Omega (1-\mathcal{H}(\phi)) dy}. \tag{4}$$

The segmentation is done on a feature image composed of the magnitude of the optical flow field. The motivation for using optical flow is that different events have characteristic motion statistics; thus, segmenting according to the motion field allows us to separate distinct events. The choice of optical flow computation is critical for the respective detection task at hand. While the classical optical flow formulation based on intensity constancy [10] is designed for rigid motion, [6] introduce an optimal mass transport (OMT) optical flow that accounts for diffusive motion such as that of fire or smoke. Consequently, the development of optical flow formulations for non-rigid motion of unstructured objects is of great interest for this field of research.

Above we describe the evolution of a contour in one frame. For tracking, the contour needs to be updated for arriving frames. In the simplest case, if the frame rate is high enough, one can use the contour of the previous frame as an initialization for the new frame. In the case of fast moving objects, more advanced tracking algorithms involving dynamic models need to be employed [11].

2.3. Anomaly Detection: The Posterior

Localizing suspicious activity to candidate regions is beneficial twofold. First, optical flow and other features can be computed on a smaller region, which improves computation time. Second, active contours can be initialized within the subregion, which is likely to contain a single object and background, as opposed to segmenting globally within an image, which can have multiple objects and varying background. The active contour starts at the rectangular boundary of the subregion and evolves to non-rectangular final segmentation further localizing the suspicious area.

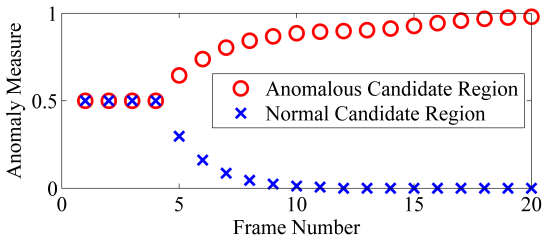


Fig. 2. Anomaly measure μ_k^i over time in the fire detection scenario Fig. 3 (Left contour - anomalous (fire). Right contour - normal (fireman)). After the initial monitoring (probability equals 0.5, $N = 5$) the track-before-detect scheme Eq. (7) and Eq. (8) pushes μ_k^i to one (zero) if the detection result is consistently positive (negative).

A supervised classifier, such as a neural network or a support vector machine, is then used for on-line classification, given samples from the candidate region. The preceding off-line training determines classifier weights such that the feature space of the training data is separated as best as possible. The major problem therein is the need for labeled training data. Section 3 addresses this topic.

The choice of features is wholly dependent on the application. Note that with the proposed framework, both pixel-based and region-based features are applicable. Using the latter would not be possible in a framework that takes the frame as a whole, since multiple events in the frame would blend into one feature.

In general, the classification provides at each time step t_k the probability $\mathrm{P}(\mathrm{Event}|f(\Omega_i))$, called P_k^i from here on, which is the probability that the event of interest is present given the features extracted from candidate region Ω_i at time t_k. Since the candidate regions are tracked, the previous probabilities are available for filtering. The filtered posterior value μ_k^i for candidate region Ω_i at time t_k with memory reaching back N time steps is called *anomaly measure* and defined as a function

$$\mu_k^i = m(\mathrm{P}_k^i, \mathrm{P}_{k-1}^i, \ldots, \mathrm{P}_{k-N}^i), \tag{5}$$

where m is an averaging function that reduces noise from mis-classifications and mis-segmentation. Moreover, m should have the tendency to converge to 0 or 1, since after some time a final decision (1 - Ω_i is anomalous, 0 - Ω_i is normal) has to be made. Fig. 2 shows an example for the behavior of the anomaly measure μ_k^i obtained from the fire results in Section 4 for two candidate regions, one that is anomalous and one that is normal.

3. GENERATION OF GROUND TRUTH DATA

This section describes an efficient procedure for labeling videos to generate ground truth data for training and testing. The idea is to transform the proposed *computer detection* approach of Section 2 into a *computer-aided human detection* method to generate ground truth data semi-automatically: the only change required is replacing the supervised classifier with a user interface in the anomaly detection step, which allows a person to label candidate regions as normal or anomalous. Having candidate regions available as contours from steps 1 and 2, the user interface displays these contours on the respective frame and the user may label a candidate region by clicking inside the contour. As opposed to freely marking regions of interest in the frame, this approach significantly speeds up the process and increases accuracy. The pixel label maps are then stored in a database for later use.

Once there is enough training data to make the detection algorithm work reasonably well, the above approach can be further simplified: instead of *replacing* the supervised classifier by a user interface, which fully transfers the decision to the human, a *correction* option may be implemented. That is, the detection algorithm runs as described in Section 2 but the user has the option to correct the algorithm's decision. The corrected data is used for continued training and improvement of the classifier. This method can be applied in the field, making it a learning detection algorithm.

4. APPLICATION TO FIRE DETECTION

The above discussion developed a general framework for amorphous and unstructured anomaly detection. Using the fire detection algorithm in [6] as an example, this section shows an application of our framework.

The fire detection algorithm in [6] performs pixel-wise classification with a neural network trained on the following features: R, G, B color channels and the magnitude of the optimal mass transport (OMT) optical flow. The OMT optical flow is based on the assumption that overall brightness is conserved between frames as opposed to the assumption that intensity is preserved as in Horn-Schunk optical flow [10]. OMT optical flow $\vec{u}$ minimizes the functional

$$\min_{\vec{u}} \alpha \int_\Omega \int_0^1 I \, \|\vec{u}\|^2 \, \mathrm{d}t \, \mathrm{d}\Omega + \|I_t + \nabla \cdot (\vec{u} I)\|^2 \tag{6}$$

where α is a weighting factor, I is the intensity and I_t is the intensity difference between two frames. The neural network's output is a probability map that is then thresholded to obtain pixels classified as anomalous.

To embed the original method in the present framework, follow the three steps of Section 2: first, initialize candidate regions with a weak classifier consisting of a neural network, trained on the RGB values, and a pixel-wise median filter, which is used as a simple motion detector. The intersection of image regions that pass the *fire test* of using an RGB-classifier and the median filter yields the final pixel map, which is then clustered via k-means for contour initialization.

Then, segmentation is performed on the magnitude of the OMT motion field with Chan-Vese [9] active contours. Overlapping contours or initializations are merged to one contour by taking their union since overlapping of regions indicates that they originate from the same population.

In the detection step, pixel-wise classification is performed as in [6] but inside each candidate region only. From the pixel probabilities, a probability P_k^i for a contour Ω_i is obtained by taking the ratio of the number of pixels classified as anomalous in Ω_i to the overall number of pixels in Ω_i. Using the moving average

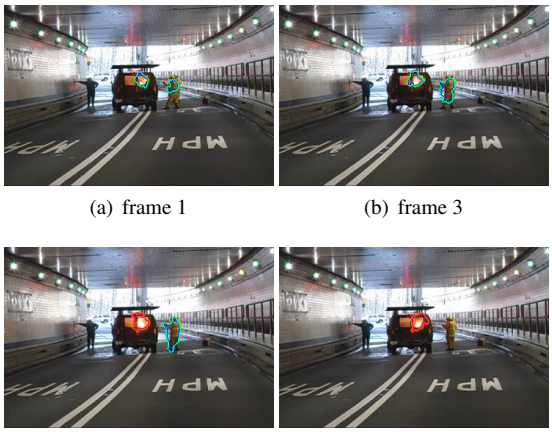
$$\bar{P}_k^i = \frac{1}{N} \sum_{k=0}^{N-1} P_{k-n}^i, \tag{7}$$

the anomaly measure μ_k^i is defined as the filter

$$\mu_k^i = \begin{cases} 0.5 \, (1 + \mu_{k-1}^i) & \text{if} \quad \bar{P}_k^i > a_{\text{thresh}} \\ 0.5 \, \mu_{k-1}^i & \text{if} \quad \bar{P}_k^i < 1 - a_{\text{thresh}} \\ \mu_{k-1}^i & \text{otherwise} \end{cases} \tag{8}$$

where $0.5 < a_{\text{thresh}} < 1$ and $\mu_0^i = 0.5$.

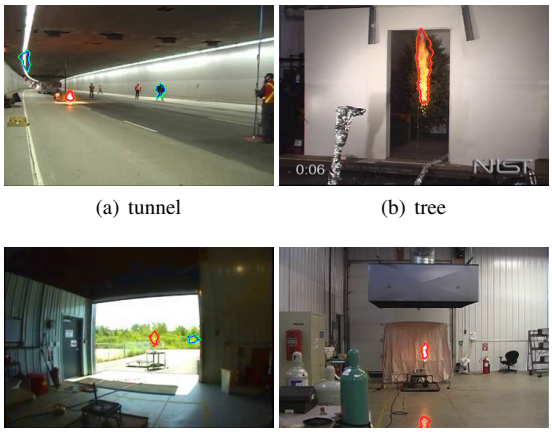
The framework is tested in five different scenarios not used in the classifier training (in contrast to [6]), and using the same set of parameters. Fig. 3 and Fig. 4 illustrate the scenarios. In all of them, the fire is detected a few frames after ignition with zero misclassifications from that point on. Also, there are no false positives (walking people, bright background) classifications in any of these scenarios. While this certainly does not imply the impossibility of false positives, it does suggest the strong robustness of this technique. Tab. 1 summarizes the detection results.



(a) frame 1 (b) frame 3

(c) frame 5 (d) frame 12

Fig. 3. Scenario 1: Van. Blue contour: tracked candidate region, red contour: detected region. (a) initial candidate regions, (b) candidate regions are tracked, (c) fire is detected, (d) yellow, moving fireman is classified as 'not fire' and contour is deleted.



(a) tunnel (b) tree

(c) outdoor (d) indoor with reflection

Fig. 4. Scenarios 2 to 5 tested from Tab. 1.

5. FINAL REMARKS

A framework for the detection of unstructured objects in videos is developed for use in video surveillance. Video fire detection is a difficult task that is chosen to demonstrate the framework's effectiveness. It is chosen primarily due to data availability, although the techniques generalize to other anomalies observed in surveillance. A particulary excellent property demonstrated in the test results is the extremely low false-positive rate. This is vital for meaningful detection when the chance of an anomaly occuring is small.

Table 1. Test results for five scenarios not present in training

1 Van	2 Tunnel	3 Tree	4 Outdoor	5 Indoor
Number of frames tested				
80	210	400	100	100
Fire first detected on frame number				
5	29	20	10	11
Number of non-fire regions correctly rejected as 'non-fire'				
7	6	1	6	0

6. ACKNOWLEDGMENT

This material is based upon work supported by the Office of Naval Research under Contract No. N00014-10-C-0204. The videos were provided by courtesy of UTRC.

7. REFERENCES

[1] C.M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893–908, 2008.

[3] B.C. Ko, K.H. Cheong, and J.Y. Nam, "Fire detection based on vision sensor and support vector machines," *Fire Safety Journal*, vol. 44, no. 3, pp. 322–329, 2009.

[4] P. Piccinini, S. Calderara, and R. Cucchiara, "Reliable smoke detection in the domains of image energy and color," in *International Conference on Image Processing (ICIP)*. IEEE, 2008, pp. 1376–1379.

[5] J. Junior, S. R. Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, 2010.

[6] I. Kolesov, P. Karasev, A. Tannenbaum, and E. Haber, "Fire and smoke detection in video with optimal mass transport based optical flow and neural networks," in *International Conference on Image Processing (ICIP)*. IEEE, 2010.

[7] T.H. Chen, P.H. Wu, and Y.C. Chiou, "An early fire-detection method based on image processing," in *International Conference on Image Processing (ICIP)*. IEEE, 2004, vol. 3, pp. 1707–1710.

[8] S. Osher and J.A. Sethian, "Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulation," *Journal of Computational Physics*, vol. 79, pp. 12–49, 1988.

[9] T.F. Chan and L.A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[10] B.K.P. Horn and B.G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[11] M. Niethammer, A. Tannenbaum, and S. Angenent, "Dynamic active contours for visual tracking," *IEEE Transactions on Automatic Control*, vol. 51, no. 4, pp. 562–579, 2006.