2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Detecting Anomalies from Video-Sequences: a Novel Descriptor

Giulia Orrù, Davide Ghiani, Maura Pintor, Gian Luca Marcialis, Fabio Roli
Department of Electrical and Electronic Engineering
University of Cagliari, Italy
Email: {giulia.orru, maura.pintor, marcialis, roli}@unica.it, davideghiani@gmail.com

***Abstract*—We present a novel descriptor for crowd behavior analysis and anomaly detection. The goal is to measure by appropriate patterns the speed of formation and disintegration of groups in the crowd. This descriptor is inspired by the concept of one-dimensional local binary patterns: in our case, such patterns depend on the number of group observed in a time window. An appropriate measurement unit, named "trit" (trinary digit), represents three possible dynamic states of groups on a certain frame. Our hypothesis is that abrupt variations of the groups' number may be due to an anomalous event that can be accordingly detected, by translating these variations on temporal trit-based sequence of strings which are significantly different from the one describing the "no-anomaly" one. Due to the peculiarity of the rationale behind this work, relying on the number of groups, three different methods of people group's extraction are compared. Experiments are carried out on the Motion-Emotion benchmark data set. Reported results point out in which cases the trit-based measurement of group dynamics allows us to detect the anomaly. Besides the promising performance of our approach, we show how it is correlated with the anomaly typology and the camera's perspective to the crowd's flow (frontal, lateral).**

I. Introduction

Anomalous event detection in crowded environments [1] by computer vision (CV) techniques is an investigation field with a strong impact in the physical security domain. In this context, the repressive action of the police could be helped if an automatic mean of fast search of anomalous events over a video-sequence would be available. Moreover, an automated system could give significant help if it would be able to predict the occurrence of such an event in real-time.

Although the notable efforts in terms of anomalous events detection and classification, especially by using deep networks [2], experimental results showed these systems cannot yet be applied in real environments without knowing their limitations.

The main motivation is that it is not easy to define in a mathematical closed-form what an anomalous event is [1]. Each application has its own specificities which may be associated with behavioural anomalies. Accordingly, behavioural biometrics [3], for example, the gestural analysis or the facial expression evaluation, are useful to assess a certain class of anomaly, such as physical or verbal violence [4].

In this paper, we focused on the assessment of how much a "state variable" denoting an anomalous event is far from the correspondent value of the "quiet" state, which denotes the absence of anomaly, over time [5]. We took into account that often video-surveillance cameras are very far from the foreground scene and give very few insights about the gestures, facial expressions and actions of individuals, although these are potentially evidence of anomalous events.

In this paper, we considered that crowds tend to assume an aggregative or dispersive behavior whose rapidity depends on the nature of the occurring event. In particular, rapid and random aggregative or dispersive behavior occurs in correspondence of an anomaly: this is evident in events causing the crowd's panic. The rapidity modelling is, therefore, the core of our approach; in other words, our goal is to assess the rapidity of the crowd's answer to a certain external stimulus.

We designed a descriptor based on accounting the number of people groups detected into the scene, and how it increases or decreases over time. A descriptor that works accordingly allows us to "explain" what is happening by evaluating how much it changes. Nevertheless, it may become less effective when a normal transition in the groups' number tends to be too similar to that supposed for an anomalous event, thus causing false alarms, even if they could be explained in retrospect. This explanation ability is still difficult to be obtained by a black-box approach like a deep network-based one [6].

Two modules compose our system. The first one is the clustering module whose aim is to detect the people groups in the scene – we implemented several state-of-the-art algorithms [7], [8], [9], [10] beside the manual counting of the groups. The second one is the original contribution of this paper: a novel temporal descriptor aimed to represent the variations of the group numbers over time.

The one-dimensional local binary pattern [11] inspired our descriptor. By a time-window sliding over a certain video-sequence's set of frames, and centred on a certain temporal instant, we compare the number of groups computed at that instant with the groups count after and before it. Three cases may happen: the number of groups increases, decreases, or remains unaltered. This leads to have, for each time-window, a set of "trinary" states (increase/decrease/unaltered), that we called string of "trits" (trit=***tri**nary digi**t***). After collecting a certain number of strings, we evaluated the histograms of the trinary states occurrences, thus obtaining a temporal description of the groups' dynamic over that set of frames. The deviation of each histogram from the "quiet" state-related histogram allows us to decide if an alarm must be output.

We adopted the Motion-Emotion data set [12] to benchmark

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

our system, and, where possible, we compared that with other state-of-the-art methods. Unfortunately, the lack of a rigorous experimental protocol shared by all research groups limits the significance of this comparison, since the same data set is used for different goals. Anyway, we computed the performance parameters usually adopted in terms of the so-called F1 score and precision/recall. Beside them, we discussed the pros and cons of the system by correlating the system's behaviour with the video-sequence. This allowed us to observe the change in the trit-based state histogram according to what happens over the video. In other words, we verified if our hypothesis is supported as well as the impact of the noise introduced by non-zero-detection-error modules (in particular, the group counting module).

Although the obtained performance is not yet ready for practical use of the system, reported experiments are in agreement with our claim, that is, we propose a novel and explainable method to anomaly detection in mass-based video-surveillance applications. This method relies on a hypothesis physically verifiable, thus it acts as full white-box whose parameters are correlated to the specific detection goal.

The paper is organized as follows. Section 2 makes the point about the state-of-the-art on modeling the crowd behavior. Section 3 describes the proposed system. Experimental results are reported in Section 4, and the concluding Section is a frank discussion about the pros and cons of this proposal.

II. Anomalous Events and Previous Works

The crowd analysis is the study of the natural movements of people and objects in the scene. Such a group of people is often named "crowd". The crowd analysis [5] can be conducted at the macroscopic level by considering the crowd global motions, or at the microscopic level that is focused on the movements of each individual. Based on the density of the crowd in the scene, crowd analysis methods can be grouped into two major categories, sparse crowd analysis and dense crowd analysis (Fig. 1) [13]. The density of a crowd is commonly an estimate of how many people are in a delimited area and it concerns the possibility of distinguishing the individuals and the background. Therefore, *sparse crowd* is a crowd where the individuals are separable and identifiable and occlusions are limited, while in the case of a *dense crowd*, people are crammed together and it is impossible to distinguish the shape of each individual due to the high level of occlusions.

The dense crowds can be divided into structured and unstructured crowds according to the motion present in the scene. If the crowd moves coherently on a macroscopic level with a common and constant direction, the crowd is structured. If the crowd is characterized by chaotic or random movements and at a microscopic level from different directions and different gestures, the crowd is called "unstructured". Crowd analysis allows to face a large number of tasks (Fig. 2) which can be schematized by macro areas as density estimation [14], crowd counting [15], behavioral analysis and prediction [16], people tracking [17] and person identification.

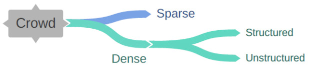


Fig. 1: Taxonomy of crowds.

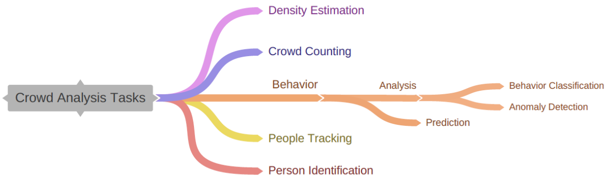


Fig. 2: Taxonomy of crowd analysis task.

Based on the task, the crowd analysis involves extremely different methods and disciplines. The analysis is performed in different levels, from raw to elaborate features.

Figure 3 shows an overview of the features used for computer vision-based crowd analysis. Usually a bottom-up approach is adopted, starting from a voluminous amount of low-level features and refining the information through processing and selection, up to high-level features [1].

At the base, a large set of **raw features** is collected. The raw features are extracted from the sensors, such as one or more videocameras. Raw features, such as the RGB level of each pixels, can be used, but the extraction of specific features is often done, such as for the optical flow-based features [18] from the frames sequence or the textural-based features [19] from the single frame.

At the middle level, **aggregate features** are obtained from the raw features, usually by reducing the dimensionality and number of features. At this level, individuals can be detected and tracked, or groups of people can be identified [20] by the extracted features. However, it is not possible to have the clear understanding of the scenario, since they are still lacking of the relationships with the whole context.

The **high-level features** allow to relate the previous features sets with their context. This step includes psychological and social knowledge in order to understand the targeted event. The features processing is strongly affected by the accuracy of the previous ones, and requires context-dependent modelling. The final interpretation of the scene as given by the third-level features is left to the humans.

In the anomaly detection task, numerous descriptors are based on raw or aggregate features, generally called low-level features, as Mixture of Dynamic Texture (MDT) [21] and the spatio-temporal descriptors Histogram of Gradient (HOG) [22] and Histogram of Optical Flow (HOF) [23]. Others combine motion information and context information, as Ref. [24] which extracts the local binary patterns for obtaining contextual information. In recent years, deep learning-based anomaly detection algorithms have become increasingly popular and allowed to achieve very accurate results.

Despite their high accuracy, the black-box nature of deep-learning approaches makes these methods poorly interpretable [25] and vulnerable to adversarial attacks [26]. In this work we propose a time descriptor of the crowd that allows to give an explanation of the detection of the anomaly.

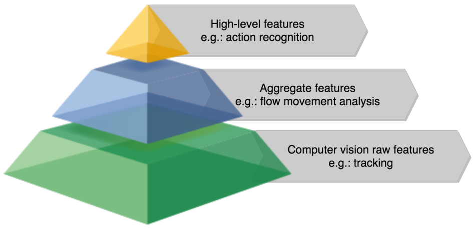


Fig. 3: Overview of features used for CV-based crowd analysis.

III. A NOVEL DESCRIPTOR FOR ANOMALOUS EVENTS DETECTION IN HIGH-DENSITY CROWDS

A. *Crowd model*

Crowd behaviour is strictly related with crowd motion patterns. In a sparse crowd scenario, these patterns are defined by individuals that move in the scene, and eventually form or break-up groups. We identify a group as a set of individuals that: 1) move together, speak together, or perform a common action; 2) are physically close; 3) are directed in a common centre of focus. Our system is based on the hypothesis that anomalous events happen when multiple *group formation events* and *group breaking-up events* suddenly appear in the scenario. It is worth noting that these events happen likewise in absence of anomaly, but with different dynamics. Figure 4 shows examples of slow and fast crowd dynamics. By counting the number of groups in the scene and monitoring its variation over time, it is possible to detect such variations and use them for the purpose of anomaly detection in crowd behaviour.

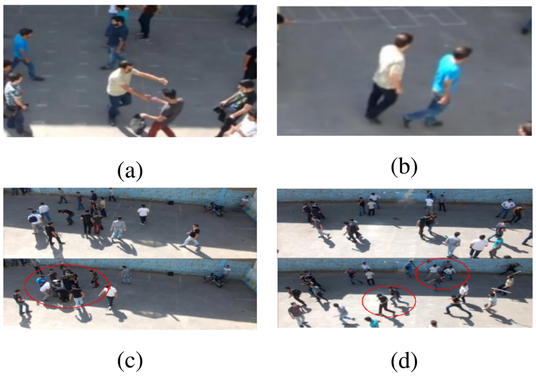


Fig. 4: Examples of crowd dynamics: (a) slow formation; (b) slow breaking-up; (c) fast formation; (d) fast breaking-up.

Our system consists of three layers, shown in Figure 5. First, we select a subset of the total number of frames for avoiding a dense analysis (1). In the second layer, we count the number of groups in the scene for each selected frame with computer vision techniques (2). This will serve as low-level features and holds little information about the anomalies. Finally, the output of this block passes through a refinement layer (3) that describes with a histogram each analysed sequence and links the low-level features with our crowd model. A threshold is applied to the central bin of the histogram, characterizing the state of quiet, and acts as a trigger for detecting the anomaly.

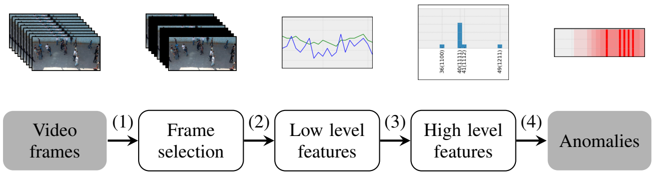


Fig. 5: Anomaly detection pipeline. (1) A subset of the total frames is selected from the whole sequence of frames; (2) Low level features are extracted for obtaining the number of groups in each scene; (3) High level features compute statistics of dynamics patterns; (4) Anomalies are obtained through thresholding a specific pattern.

B. *Low level features*

The input of our system is a video sequence of N_{tot} frames. We select one frame each F, the number of frames that we skip between two subsequent group counts, obtaining $\frac{N_{tot}}{F}$ frames. Low level features are intended to count the number of groups in the scene along time. We compare three automatic methods for counting the groups, counting each F frames and the manual counting as ground truth.

Clustering of optical flow (COF). We use dense optical flow (OF)[18] for tracking moving blocks of pixels from one (selected) frame to another. Then, a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is used for grouping into blocks the optical flow vectors that: 1) have the same angle of movement (phase of the OF); 2) have the same speed (magnitude of the OF); 3) are spatially close to each other (x,y coordinates of the pixel). We use the clustering algorithm for obtaining the number of clusters, rather than assigning each point to a cluster.

Cascade detector (CD). We estimate the number of groups by counting the number of bounding boxes resulting from the application of an off-the-shelf Multi Scale Object Detector implemented in OpenCV [10]. This method detects individual persons rather than clusters, but our hypothesis is that for each group only the front person will be completely visible. We use the number of bounding boxes as a rough estimation of the number of groups. This method is the fastest, but highly inaccurate in the absolute count.

Blob detector (BD). We implemented our own version of a bounding box detector. First, it removes the background and applies gaussian blurring to the image in order to reduce the noise. After, it binarizes the image and highlights the shapes of the blobs, i.e., the groups that we are willing to count. Finally, we apply dilation and opening for extracting the contour of the shapes, then we count the number of detected contours and this will act as our count.

To the bounding box detection is not applied any post-processing. The output of this block is a raw count of groups for each selected frame. It is important to mark that we are not interested in the accurate value of this count, rather a correct detection of the abrupt changes. This means that the group counting algorithm is supposed to capture a trend rather than exact counts.

C. *High level features*

Once the group count has been extracted, we can proceed with the high-level feature extraction (Fig. 6). We designed a descriptor that can be applied to any one-dimensional signal for anomaly detection. The output of this layer is a sequence of labels that marks where the anomaly starts in each video.

After the first layer of feature extraction, using a sliding window of size L, we obtain K arrays. Since scrolling the window on the $\frac{N_{tot}}{F}$ selected frames of one element at a time, $K = \frac{N_{tot}}{F} - L + 1$. For each array k_i, with $i = 1, ..., K$, we apply the following algorithm:

1) We scroll the array k_i with another sliding window of size $2W + 1$, being W a positive integer smaller than half of L, getting $L - 2W$ windows. For each of these windows, we compare the number of groups of the central value c with the surrounding ones. By the comparison of each element c_j of the sliding window, being $j = 0, .., 2W$ with the central value c, we obtain a trinary code by applying Eq. 1.

$$t_j = \begin{cases} 0 & if \quad c_j - c > T \\ 1 & if \quad |c - c_j| <= T \\ 2 & if \quad c - c_j > T \end{cases} \qquad (1)$$

The descriptor threshold T is set to reduce the noise influence in the count. This allows the state of quiet to evolve over time: if the groups count gradually increases or descends the sequence will not be identified as anomalous. We drop the central value c for avoiding self-comparison, and we obtain a code of $2W$ trinary digits.

2) The trinary code is transformed into a decimal number using Eq. 2.

$$d = \sum_{j=0}^{2W} (t_j \times 3^j); \qquad (2)$$

We obtain $L - 2W$ decimal numbers as the sliding window strides along the sequence of L elements.

3) The $L - 2W$ decimal values are collected into a histogram that summarises a patterns' statistic observed in the temporal sequence (without caring of the order).

At the end of this procedure, we obtain K histograms describing the subsequent statistics of patterns. The quiet state is described by a histogram with a single central bin. The lowering of the central bin means that other bins are reaching non-zero configurations. For this reason, we apply a bin threshold t^* to the central value of the histogram, that characterises the absence of anomaly. We mark as anomaly the moment in which this histogram bin changes and triggers the threshold.

TABLE I: Overview of the papers that use the ME dataset.

Refer.	Method	Application		Performance	
		Anomaly Detection	Behavior Classification	Acc[%]	Avg Error
[12]	Trajectory		✓	35.30	-
	HOG		✓	38.80	-
	HOF		✓	37.69	-
	MBH		✓	38.53	-
	Dense Trajectory		✓	38.71	-
[27]	Emotion Attribute		✓	43.6	-
	Spatio-Temporal		✓	71.70	-
[28]	Density Heatmaps/ Optical Flow		✓	90.9	-
[29]	Phase-Based Statistics	✓		-	0.118

The key observation of this process is that we model the normalcy in the scene through monitoring a trend of a specific extracted pattern over time. The histograms act as a summary of all the variations in the groups count, yet the bin that models the normalcy is only the one that represents a *static* crowd dynamics (changes in the number of groups below the threshold T).

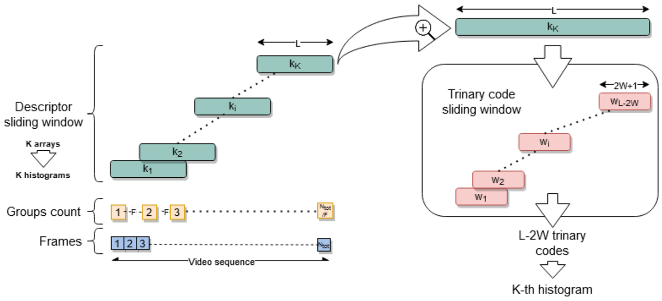


Fig. 6: High-level features extraction scheme. The group counts are divided into K arrays of size L through a first sliding windows. Each array is further divided with another sliding window in $L - 2W$ subgroups of size $2W - 1$, each of which generates a trinary code. The trinary codes transformed into decimal values are collected in a histogram.

IV. Experimental results

A. *Data set*

The algorithm has been tested using the Motion Emotion (ME) data set [12]. It contains 31 video sequences of around 44000 frames in total. Each video is recorded as 30 frames per second using a fixed video camera elevated at height, overlooking individual walkways and a video resolution of 554×235. The crowd density in the videos ranges from sparse to very crowded. The videos contain both normal and abnormal behavior, labeled frame-by-frame as 5 classes (Panic, Fight, Congestion, Obstacle and Neutral). The videos do not capture only one behavior at a time, but rather the transition from normal to abnormal situation. For each behavior, at least two videos from different point of view are recorded.

Table I shows the works that use the ME dataset in their experiments. Most of these papers focus on the behavior classification task. Even among them, it is difficult to compare the performances since the experiments do not have a common protocol. The only work on the detection anomaly task [29] analyzes the performance of a "change detector", considering not only the onset of the anomaly but also its resolution.

B. Experimental protocol and evaluation parameters

The anomaly detection performance depends on:

- the alarm synchronization with the anomaly occurrence - a maximum time delay can be taken into account in order to counteract appropriately;
- "low" false positive detection rate - false alarms reduces the effectiveness of the system; for example, too many alarms over time are an obstacle during check of true anomaly occurrences.

To evaluate the reliability of the proposed method, we consider as anomalies correctly detected all those alarms that fall within a range corresponding to about 27 seconds, centred on the actual occurrence of the anomaly. This occurrence is manually labelled, by following the perception of the human operator.

If several alarms fall in this range, only one is considered. We considered the period before the occurrence of the anomaly because some changes in the group numbers may represent the dynamics leading to the anomaly.

We divided the videos in two separate sets according to the camera position. In fact, some videos of the ME data set present a slope between wall and ground more visible than others ($> 5°$). We have included these videos in the set of "lateral camera" and those with a very small slope in the set of "frontal camera". Fig. 7 shows an example of lateral and frontal view. These views reflect two real case scenarios.

We performed the following experiments:

- Experiment 1: set of videos recorded by the camera in frontal position with respect to the scene. The set includes videos [001, 002, 005, 006, 009, 010, 015, 016, 017, 018, 019, 020, 021, 022, 023, 024, 025, 026, 028, 029, 030, 031] of the Motion Emotion Dataset.
- Experiment 2: set of videos recorded by the camera in lateral position with respect to the scene. The set includes videos [003, 004, 007, 008, 011, 012, 013, 014, 027] of the Motion Emotion Dataset.

We adopted the precision, recall and F1 score metrics to evaluate the overall system performance since anomalous detection is characterized by an uneven classes distribution: the number of frames with anomalies are considerably smaller than those without them.

We plotted the timeline description of the detection (Figs. 8-9(b)): the anomaly labels are green vertical lines and the corresponding alarms are represented by red vertical lines. The 27-seconds window centred on the anomaly is highlighted in light green. Each alarm output into that time window is evaluated as a true positive (TP). Eqs. 3-5 define the precision, recall and F1 metrics, respectively:

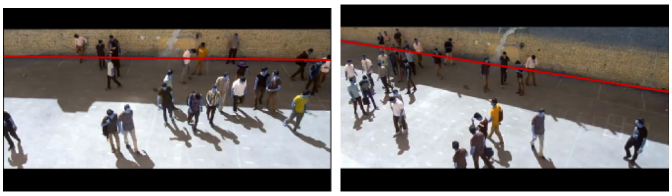


Fig. 7: Videos partition in two separate sets by the camera position: the frontal videos have the line between the wall and the ground with a zero or small slope (a), the lateral videos have a visible slope (b).

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1_{score} = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

Where TP is the number of correctly detected anomalies, FP is the number of false alarms, and FN is the number of undetected anomalies.

Table II shows the parameters shared in all the experiments. The remaining parameters have been set using a grid search, which maximize the $F1_{score}$. The ranges of the parameters set by grid search are shown in Tab. III. The grid search was done both in a supervised way, i.e. the $F1_{score}$ is maximised on all videos, and with a Leave-one-out cross validation, i.e. the parameters maximize the $F1_{score}$ on $N-1$ videos and then they are used to test the video left out. Results are reported in Table IV for each Experiment. Although supervised experiments lead to a biased estimate, they allow us to show the optimal performances obtainable with the perfect setting of the parameters.

TABLE II: Common parameters used for all the experiment.

PARAMETERS EXPERIMENT	
video frame rate	30
skipped frames (F)	20
trinary code sliding window size ($2W + 1$)	5

TABLE III: Tested parameters for tuning the algorithm.

GRID SEARCH RANGES	
descriptor threshold (T)	[2, 3, 4, 5, 6]
descriptor sliding window size (L)	[10, 15, 20, 25, 30]
bin threshold (t^*)	[0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95]

C. Overall Results

We reported the performance measurements related to Experiment 1 and 2 in Tab. V and VI. Table VII illustrates the overall performance results. In each table, the performance refers to the proposed descriptor receiving as input the number of groups obtained by the counting methods described in

TABLE IV: Parameters used for the experiments 1 and 2. For the Leave-one-out (LOO), the average and the standard deviation of the parameters used for each video were reported.

Grid search parameters		**T**		**L**		**t***	
		Sup	**LOO**	**Sup**	**LOO**	**Sup**	**LOO**
MC	**Exp 1**	3	4.25 ± 1.09	15	15.00 ± 0.00	0.85	0.76 ± 0.15
	Exp 2	2	4.00 ± 1.22	20	16.25 ± 6.49	0.65	0.68 ± 0.12
COF	**Exp 1**	4	3.50 ± 0.50	15	15.00 ± 0.00	0.50	0.68 ± 0.17
	Exp 2	6	4.25 ± 1.48	10	12.50 ± 4.33	0.85	0.76 ± 0.09
CD	**Exp 1**	4	4.25 ± 1.09	15	15.00 ± 0.00	0.85	0.76 ± 0.15
	Exp 2	4	3.75 ± 1.09	10	17.50 ± 5.59	0.85	0.69 ± 0.11
BD	**Exp 1**	3	3.75 ± 0.83	15	18.75 ± 6.50	0.50	0.74 ± 0.17
	Exp 2	5	4.25 ± 1.48	10	16.25 ± 6.50	0.70	0.76 ± 0.09

Section III-B, namely, Clustering of Optical Flow (**COF**), Cascade Detector (**CD**) and Blob Detection (**BD**).

According to Experiment 1 and 2, the different position of the camera leads to very different performances. Lateral views appear to have a deeper observation surface than frontals (Figs. 7(a-b)), and this might suggest that the descriptor has more "time" to observe changes in the scene, especially in case of anomalous events.

The group counting method affects the performance of the descriptor. The most reliable method is the Cascade Detector which on all videos achieves better detection performance than even the Manual Counting in the Leave-one-out protocol.

The difference in performance between the supervised and the leave-one-out protocol suggests that a more accurate setting of the parameters calculated by grid search would allow for more reliable detection. Another explanation is that the number of videos is too small for achieving a reliable estimation by cross-validation approaches.

D. Analysis of individual cases

We analyze some videos in detail in order to explain how the proposed descriptor behaves. In the following figures, the green lines represent the actual anomalies and the red lines the anomalies predicted by the system. If the red line lies in a light green area, the anomaly is considered correctly detected.

Firstly, we selected the video 009 to represent the case where the detector is able to correctly identify the anomaly with all the counting methods (Fig. 8). The video is characterised by an initial static flow of individuals, namely, a structured crowd, and consequently a constant number of groups. The panic event generates movement whereby the detectors are able to notice the anomaly. This is fully in agreement with the hypothesis behind our work. The video 025 (Fig. 9) presents an anomaly (protest) that is not detected except by counting with Cascade Detector. This is due to the fact that the anomaly is characterized by a single group of people who cross the scene with a controlled and slow movement. The number of groups therefore has no particular variations. This is still a structured scene, and the behaviour of the detector is in agreement with our hypothesis. The video 023 (Fig. 10) does not present anomalies. However, this case is characterised by a large number of false alarms. This is a case of unstructured and sparse crowd where we have very small changes in the number of groups. This behaviour is detected by the system and interpreted as multiple anomalies. This effect can be reduced by avoiding the bins of the trit-based histogram that represent such small variations[1].

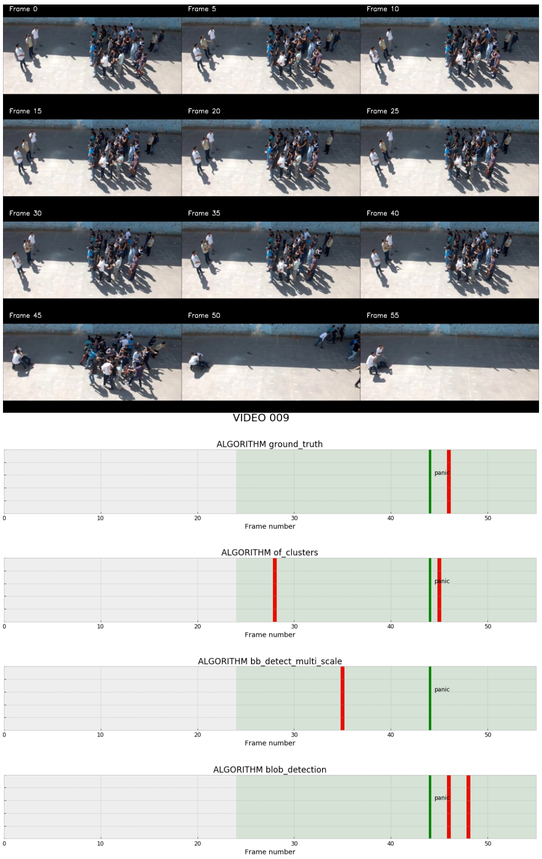


Fig. 8: Frames summary and graphical description of detection for video 009: the video presents a panic situation that has been correctly identified by all the methods.

E. Computational time analysis

The system configuration for determining the processing times of a frame was a MacBookPro with macOS Sierra, 2.7 GHz Intel Core i5, 8 GB 1867 MHz DDR3 of RAM and a Intel Iris Graphics 6100 1536 MB. The total processing time is highly influenced by the group counting method, whose times are of the order of magnitude of hundreds of milliseconds and are reported in Tab. VIII. The creation of the high-level features from the count values is in fact extremely fast, about 1 ms using a sliding window of size L=15. Considering that we analyze one frame each F = 20 frames, namely one frame per 666 ms, we can build a pipeline without delay for all the methods analyzed. However, it should be noted that the cascade detector and the blob detector, in addition to being more performing, are faster than the clustering of optical flow.

[1] The histograms that describe the frames were not reported in the paper for reasons of space but can be seen as supplementary material.

TABLE V: Performance comparison of the proposed method using different counting methods for videos recorded by the camera in frontal position.

	Supervised			Leave-one-out		
Experiment 1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MC	84.61%	91.67%	88.00%	68.42%	59.09%	63.41%
COF	62.50%	83.33%	71.43%	48.28%	58.33%	52.83%
CD	70.97%	91.67%	80.00%	70.97%	91.67%	80.00%
BD	63.33%	79.17%	70.37%	50.00%	66.67%	57.14%

TABLE VI: Performance comparison of the proposed method using different counting methods for videos recorded by the camera in lateral position.

	Supervised			Leave-one-out		
Experiment 2	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MC	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
COF	92.31%	100.00%	96.00%	63.64%	63.64 %	63.64%
CD	84.61%	91.67%	88.00%	80.00%	66.67%	72.73%
BD	85.71%	100.00%	92.31%	71.43%	90.91%	80.00%

TABLE VII: Performance comparison of the proposed method using different counting methods for all ME data set videos.

	Supervised			Leave-one-out		
All ME videos	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MC	88.89%	94.12%	91.43%	79.31%	71.87%	75.41%
COF	71.11%	88.89%	79.01%	52.50%	60.00%	56.00%
CD	75.00%	91.67%	82.50%	73.17%	83.33%	77.92%
BD	70.45%	86.11%	77.50%	56.52%	74.29%	64.20%

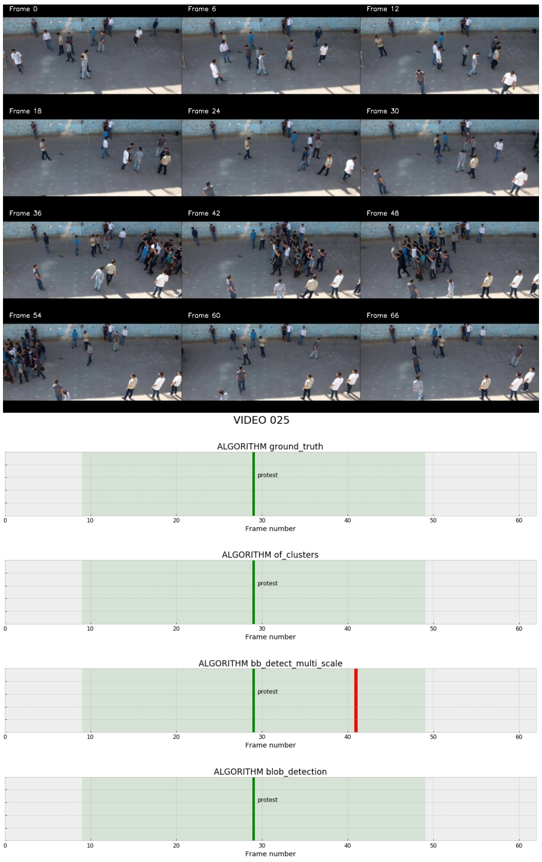


Fig. 9: Frames summary and graphical description of detection for video 025: in the video a protest crosses the scene. This anamoly is difficult to identify because crowd change is slow and controlled.

TABLE VIII: Time for the processing of a single frame with the implemented algorithms.

Algorithm	**ms per frame**
COF	∼ 650
CD	∼ 150
BD	∼ 100

V. Concluding remarks

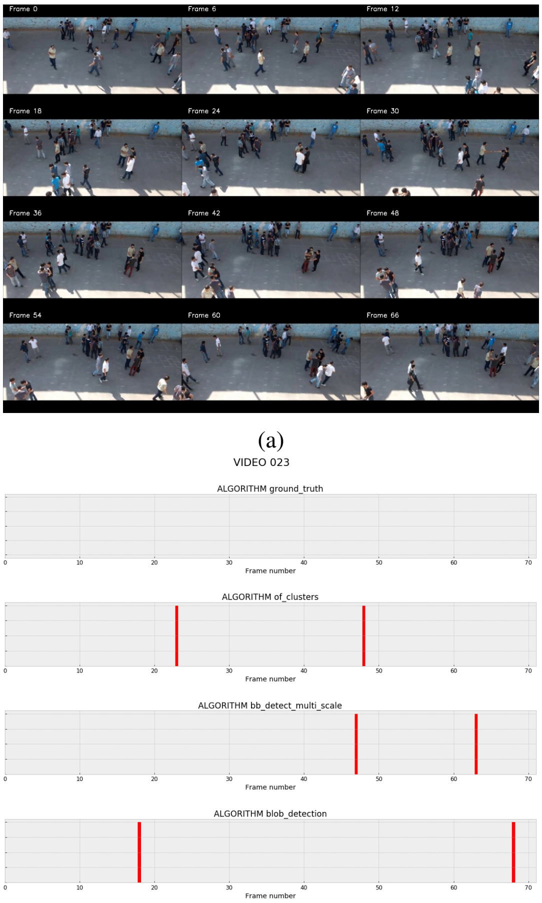
In this work, we proposed a novel temporal descriptor of small and large crowds by computer vision algorithms. We exploited the typical rapid aggregation and disruption of crowds during anomalous events and the concept of one-dimensional binary pattern. The former allowed us to focus on the number of groups of individuals over time, the latter on its variations, leading to describe them by what we called trit-based bins.

We used the Motion-Emotion data set to benchmark the proposed system and three different methods for group counting. As expected, the descriptor works well with events that clearly match our hypothesis, that is, the number of groups changes quickly and presents errors if the disruptions and aggregations are slow and controlled. On the other hand, further experiments must be aimed at analysing the trit-based bins to reduce the number of false alarms that may appear in case of sparse and unstructured crowds.

Although the performance is not good enough for real applications, the advantage of being a white box makes the descriptor easily adaptable depending on the real context and the type of anomaly to be detected.

Acknowledgment

This work is supported by the Italian Ministry of Education, University and Research (MIUR) within the PRIN2017 - BullyBuster - A framework for bullying and cyberbullying



(a)

(b)

Fig. 10: Frames summary and graphical description of detection for video 023: the video does not present anomalies but the descriptor leads to false alarms.

action detection by computer vision and artificial intelligence methods and algorithms (CUP: F74I19000370001).

References

[1] V. Murino, M. Cristani, S. Shah, and S. Savarese, Eds., *Group and Crowd Behavior for Computer Vision, 1st Edition*. Academic Press, 2017.

[2] G. Tripathi, K. Singh, and D. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: a survey," *The Visual Computer*, pp. 1–24, 03 2018.

[3] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, pp. 983–1009, 2012.

[4] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, March 2015.

[6] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, and M. A. J. van Gerven, *Explainable and interpretable models in computer vision and machine learning*, ser. The Springer Series on Challenges in Machine Learning. Springer Verlag, Jan. 2018.

[7] F. Santoro, S. Pedro, Z.-H. Tan, and T. B. Moeslund, "Crowd analysis by using optical flow and density based clustering." in *EUSIPCO*. Citeseer, 2010, pp. 269–273.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[9] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–I.

[10] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[11] N. Chatlani and J. J. Soraghan, "Local binary patterns for 1-d signal processing," in *Signal Processing Conference, 2010 18th European*. IEEE, 2010, pp. 95–99.

[12] H. Rabiee, J. Haddadnia, H. Mousavi, M. Kalantarzadeh, M. Nabi, and V. Murino, "Novel dataset for fine-grained abnormal behavior understanding in crowd," in *13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016, pp. 95–101.

[13] S. Lamba and N. Nain, "Crowd monitoring and classification: A survey," in *Advances in Computer and Computational Sciences*, S. K. Bhatia, K. K. Mishra, S. Tiwari, and V. K. Singh, Eds. Singapore: Springer Singapore, 2017, pp. 21–31.

[14] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *Robotics and Biomimetics. ROBIO'06. IEEE International Conference on*. IEEE, 2006, pp. 214–219.

[15] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[16] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Computer Vision and Pattern Recognition, 2009. IEEE Conference on*. IEEE, 2009, pp. 935–942.

[17] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1389–1396.

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

[19] Y. Hao, J. Wang, Y. Liu, Z. Xu, and J. Fan, "Extracting spatio-temporal texture signatures for crowd abnormality detection," in *Automation and Computing (ICAC), 2017 23rd International Conference on*. IEEE, 2017, pp. 1–5.

[20] H. Fradi, B. Luvison, and Q. C. Pham, "Crowd behavior analysis using local mid-level visual descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 589–602, March 2017.

[21] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1975–1981.

[22] M. J. Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2611–2618.

[23] Y. Zhao, Y. Qiao, J. Yang, and N. Kasabov, "Abnormal activity detection using spatio-temporal feature and laplacian sparse representation," in *Neural Information Processing*, S. Arik, T. Huang, W. K. Lai, and Q. Liu, Eds. Springer International Publishing, 2015, pp. 410–418.

[24] X. Zhu, X. Jin, X. Zhang, C. Li, F. He, and L. Wang, "Context-aware local abnormality detection in crowded scene," *Science China Information Sciences*, vol. 58, 05 2015.

[25] R. C. U. of Sydney, C. M. C. R. Centre, S. C. Q. C. R. Institute, and Hbku, "Deep learning for anomaly detection: A survey," 2019.

[26] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," 06 2019, pp. 49–55.

[27] E. B. Varghese and S. M. Thampi, "A deep learning approach to predict crowd behavior based on emotion," in *Smart Multimedia*, A. Basu and S. Berretti, Eds. Springer International Publishing, 2018, pp. 296–307.

[28] L. Lazaridis, A. Dimou, and P. Daras, "Abnormal behavior detection in crowded scenes using density heatmaps and optical flow," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 2060–2064.

[29] A. Briassouli, "Unknown crowd event detection from phase-based statistics," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Nov 2018, pp. 1–6.