Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos

Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz

***Abstract*— This paper presents an approach for detecting anomalous events in videos with crowds. The main goal is to recognize patterns that might lead to an anomalous event. An anomalous event might be characterized by the deviation from the normal or usual, but not necessarily in an undesirable manner, e.g., an anomalous event might just be different from normal but not a suspicious event from the surveillance point of view. One of the main challenges of detecting such events is the difficulty to create models due to their unpredictability and their dependency on the context of the scene. Based on these challenges, we present a model that uses general concepts, such as orientation, velocity, and entropy to capture anomalies. Using such a type of information, we can define models for different cases and environments. Assuming images captured from a single static camera, we propose a novel spatiotemporal feature descriptor, called *histograms of optical flow orientation and magnitude and entropy*, based on optical flow information. To determine the normality or abnormality of an event, the proposed model is composed of training and test steps. In the training, we learn the normal patterns. Then, during test, events are described and if they differ significantly from the normal patterns learned, they are considered as anomalous. The experimental results demonstrate that our model can handle different situations and is able to recognize anomalous events with success. We use the well-known UCSD and Subway data sets and introduce a new data set, namely, Badminton.**

***Index Terms*— Abnormal events, magnitude–orientation information surveillance, temporal descriptor.**

I. INTRODUCTION

IN RECENT years, video surveillance systems have become very popular due to heightened security concerns and low hardware costs. These types of systems are widely used in many applications such as nursing care institutions, law enforcement, building security, and traffic analysis. Such systems have traditionally relied on network cameras monitored by a human operator that must be aware of the actions carried out by people who are in the cameras field of view. With the recent growth in the number of cameras to be analyzed, the efficiency and accuracy of human operators have reached the limit [1]. Therefore, there is a great need for a real-time automated system that detects and locates suspicious behaviors and alerts security agents. In this way, detecting unusual or suspicious activities, uncommon behaviors, or irregular events in a scene is the primary objective of an automated video surveillance system.

Ideally, one would be interested in knowing whether suspicious activities are unfolding in the scene. However, it is extremely difficult to design activity recognition approaches without specific knowledge of the scene and the target activities [2]. In view of that, Aggarwal and Ryoo [2] have developed approaches to locate and recognize anomalous events and possibly hazardous human motions using only the knowledge regarding the normal behavior at a given location without requiring an extensive knowledge of the scene.

Jiang *et al.* [3] define anomaly detection as the identification of motion patterns that do not conform to the expected behavior in the video.[1] They also define anomaly as rare or infrequent behavior compared with all other behaviors. However, the identification of this concept requires semantic information and subjective knowledge regarding the scene and the expected behavior. Nonetheless, unknown patterns, in most cases, are very difficult to represent in automatic recognition models. Therefore, the modeling usually is built for the usual recurring patterns found in the scene, and when there is no fitting to any usual pattern, one concludes a given event as anomalous.

Since it is impossible to model every anomalous event, we must define ways of describing normal motion patterns for different regions on the scene to be able to recognize when such patterns are absent to classify them as anomalous events. With that in mind and based on common anomalous events, such as pedestrians moving with excessive speed, spatial anomaly (intruders in restricted areas or unusual locations), and the presence of nonhuman objects in unusual locations [4], we define four characteristics to be used as clues

Manuscript received December 22, 2015; revised June 6, 2016 and August 18, 2016; accepted November 1, 2016. Date of publication December 8, 2016; date of current version March 3, 2017. This work was supported in part by the Brazilian National Research Council–CNPq under Grant 487529/2013-8 and Grant 477457/2013-4, in part by the Minas Gerais Research Foundation under Grant APQ-00567-14 and Grant PPM-00025-15, and in part by the Coordination for the Improvement of Higher Education Personnel (DeepEyes Project). This paper was recommended by Associate Editor H. Yao.

The authors are with the Smart Surveillance Interest Group, Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte MG 31270-901, Brazil (e-mail: rensso@dcc.ufmg.br; carlos.caetano@dcc.ufmg.br; matheustlandrade@gmail.com; william@dcc.ufmg.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2637778

[1] The term anomalous events is sometimes referred to as abnormal events in the literature. In this paper, we opted for using the term anomalous event because abnormal might refer to a unusual event in a way that is undesirable, which is not our case since we do not have enough semantic information to know whether a given event is suspicious or just different from a normal recurring pattern, for instance.

to describe normal motion patterns in a particular region of the scene: 1) *velocity*—speed of moving objects; 2) *orientation*—common flow of the objects; 3) *appearance*—texture of the objects; and 4) *density*—the number of moving objects. We hypothesize that the use of such characteristics allows one to capture information regarding anomaly.

As an example of representation that captures one of the aforementioned desirable characteristics, we can cite the spatiotemporal feature descriptor proposed by Chaudhry *et al.* [5]. Their feature captures information based on the optical flow orientation providing the histogram of oriented optical flow (HOOF). To improve the HOOF descriptor with the raised *velocity* characteristic, in [6], we proposed a feature called *histograms of optical flow orientation and magnitude* (HOFM), which aggregates velocity information to the HOOF feature descriptor.

Based on the aforementioned characteristics and [6], we propose a novel descriptor that introduces more spatiotemporal information to describe the regions of the scene. Besides orientation and velocity, as in HOFM, our proposed descriptor, called *histograms of optical flow orientation and magnitude and entropy* (HOFME), encodes and measures the entropy of the orientation flow. We believe that the addition of the entropy will contribute mainly by capturing information regarding the appearance and density of the region, the last two characteristics that are not exploited by the HOFM [6].

Our descriptor extracts information from spatiotemporal regions sampled over several frames from nonoverlapping spatial regions. During the learning stage, with videos containing normal events presented, we extract and store the HOFME feature vectors for each spatial region, generating a set of normal patterns. Then, during the testing stage, after extracting HOFME, a nearest neighbor search[2] is performed considering only that particular region. According to the distance to the best matching pattern, the event taking place at that particular location and time might be classified as anomalous.

Besides the development of the HOFME descriptor, we propose a novel data set, called *Badminton sequence*, to evaluate anomaly detection approaches in a crowed sport event, a very common video surveillance scenario. This clip was recorded from the badminton game. We focus the recognition and ground truth on the grandstand. It is a single video data set, and the train and test are sequences extracted from it. The anomalies are people running in the corridors and jumping the chairs.

We evaluate our approach on the UCSD [7] data set, the Subway [8] data set, and the proposed Badminton sequence. According to experimental results, the proposed HOFME outperforms the traditional HOOF descriptor and our previously proposed one, and also we compared our results with those of the HOG3D [9] and motion boundary histograms (MBH) [10] descriptors.

The main differences between this paper and [6] are as follows.

1) A new spatiotemporal feature descriptor called HOFME, which adds entropy information to HOFM collecting and differentiating the orientation flow in a region.
2) A comprehensive experimental evaluation, considering long duration experiments performed in Subway video data set, the UCSD data set, and the new Badminton sequence. In addition, the proposition of a more consistent ground truth for the Subway data set.
3) A novel anomaly detection data set called *Badminton sequence* recorded in a crowded sports event.

The remainder of this paper is organized as follows. In Section II, we talk about related works on detection of anomalous events. In Section III, we introduce our approach for anomaly detection and present our proposed spatiotemporal feature descriptor. In Section IV, we analyze our experiments regarding the UCSD, Subway, and Badminton data sets. Finally, in Section V, we present the conclusion and directions to future works.

II. Related Work

Detection of anomalous events generally falls into two categories: trajectory analysis and motion analysis. While the former is based on object tracking and typically requires an uncrowded environment to operate, the latter is better suited for crowded scenes by analyzing patterns of movement rather than attempting to distinguish objects individually [4], [8]. The difficulty of the former approach increases proportionally to the number of individuals in the scene [11], [12].

Most of the existing imagery-based crowd analysis methods tend to handle a crowd as short groups of people [13]. The object-centered approaches [14], [15] require explicit detection and segmentation of individuals. These techniques in some cases are not feasible due to severe inter-object occlusion, especially in highly crowded scenes. Other models prefer to represent the crowd scenes using dense information rather than determine interest objects (groups of people). These models focus on scenes where the people movement covers the majority of the vision field.

The most common approaches to recognize anomalies are based on two steps: representation and prediction (or classification). In the former, the majority of the works employ dense feature representation, such as contextual information [3], [18], multiple information based on optical flow [4], [8], gradient-based features [16], and mixture of dynamic textures [17]. Another type of representation exploits information extracted from saliency maps, such as in [19] and in [20], where a Lagrangian particle map is used to segment the crowd. The main advantages of the methods described in this paragraph are the fixed number of features and their simplicity to set in classifiers or predictors. On the other hand, the disadvantage is that they depend much on the prior information, such as camera location.

To model and predict, the majority of techniques employ Gaussian mixture models and hidden Markov models (HMM). In [16], a multilevel HMM is used to predict anomalous events in specific regions of the crowd. In [21], [22], Markov models allow the analysis of the scene. The expectation maximization

[2] We have chosen to use a simple nearest neighbor search to be able to evaluate the quality of the feature descriptor. Therefore, we believe better results could be achieved if we employ stronger learning techniques to model the normal patterns.

algorithm has been also employed as a predictor for anomaly [15]. Another statistical model was employed in [23], where each pixel has an estimated probability to belong to foreground (there is no movement at that particular location); then using inference techniques, it determines whether a pixel is an anomaly signal. In [24], two levels of feature analysis and Gaussian regression process were employed. In [25], a robust approach uses a hierarchical mixture of dynamic textures to describe the frame. It is important to emphasize that even though the descriptors proposed in [10] and [26] share similar ideas with our proposed descriptor, their studies do not focus on anomaly detection and therefore do not capture relevant information for such an application. In the next section, we discuss the differences between our descriptors and the ones used in [10] and [26].

Despite stated in several papers that models based on the trajectory of crowds are hard to accomplish, Shao *et al.* [14] proposed a model based on group profiling that is pretty different from common models in the literature. Their model is based on group modeling, where a map of four descriptors defines the anomalies. Then, such information is quantized using a bag of features (BOF) technique and the events are classified as anomalous or not using a support vector machine. Another BOF-based technique was proposed by Saligrama and Chen [27] using a combination of codebook and Gaussian process. Deep learning technique has been also used to anomalous event detection [28]. Xu *et al.* [28] proposed a model that trains a convolutional neural network using spatiotemporal patches from optical flow images as input.

In the literature, we can also find specific works that focus on the descriptors, for example, the work of Yuan *et al.* [29] starts from a novel structure modeling of crowd behavior. They first propose an informative structural context descriptor (SCD) for describing the crowd individual. For computing the crowd SCD variation effectively, they design a multiobject tracker to associate the targets in different frames. By online spatiotemporally analyzing the SCD variation of the crowd, the abnormality is finally localized. The main contribution of Yuan *et al.* [30] lies on the motion difference between individuals, which is computed by a novel selective histogram of optical flow (SHOF). The histogram of flow (HOF) is calculated as the motion statistics where each bin of HOF represents the direction of optical flow, and the value in each bin is proportional to the magnitude of optical flow. Their SHOF represents the motion property of individuals. In [30], an interesting approach about regularity in videos is presented. This model captures the normal patters, highlighting the irregularities in the sequences. It performs two levels based on autoencoders. In the first level, local features are extracted to create the encoders; in the second level, a convolutional network reconstructs spatiotemporal regions and follows the changes in reconstructed images and the model defines the irregularities.

Approaches based on visual attention have been exploited for anomaly detection. The goal of these models is to determine the regions that capture our attention and could represent anomalies. Bruce [31] proposed a framework based on statistical density information about patches in the scene. His descriptor is based on Gabor filter patch information. Itti and Baldy [32] model a solution based on Bayesian prior information about the scene. They described the eye movement using prior information and looking for the likelihood of an eye function. Yang *et al.* [33] proposed a robust model to identify where people look. The output is a saliency map of important pixels in the image. Even though a promising direction, the majority of visual attention approaches are very complex and present difficulties to model anomalies properly.

III. Proposed Approach

In this section, we present our approach for anomaly detection. We divided the scheme in two parts: training and testing. Both schemes start with three main steps: 1) frame difference between two consecutive frames; 2) optical flow extraction; and 3) spatiotemporal description using HOFME. Our approach divides the video into nonoverlapping $n \times m \times t$ regions, referred to as cuboids, and build an orientation–magnitude representation for each cuboid. On the training step, the histograms extracted from cuboids at the same spatial location (but different temporal locations) are kept as normal patterns. During test, for a cuboid, we perform a nearest neighbor search to find similar patterns computed during training at that particular spatial region. If none is found, we consider it as abnormal. An overview of the approach is illustrated in Fig. 1.

The proposed spatiotemporal feature descriptor uses as input the optical flow. Extracting it for the whole image may be computationally expensive [4]; hence, to avoid computing optical flow for each pixel on the image, we first create a binary mask using image subtraction between the frame I_j and the frame I_{j+t}. Given a threshold d, if the resulting difference is smaller than d, the pixel is discarded; otherwise, this pixel p is set to its corresponding local cuboid C_i. Thus, each cuboid will be composed of a set of moving pixels. For each $p \in C_i^t$, we compute the optical flow using the Lucas–Kanade–Tomasi pyramidal implementation [34], where p' is the optical flow result for pixel p. The pixel p' corresponds to pixel p in C_i^t.

Our proposed feature uses optical flow information (orientation and magnitude) to build the feature vector for each cuboid. To that end, we define a cube $F_{S \times (B+1) \times E}$, where S is the number of orientation ranges, B is the number of magnitude ranges (plus one due to some magnitudes that exceed the maximum value), and E is the number of entropy ranges. We build a 3D matrix based on the orientation and magnitude information provided by the vector field resultant of optical flow (note that the magnitude of the optical flow indicates the velocity at which the pixel is moving). For the entropy information extraction, we also use the magnitude information since our main goal is to measure the variation of velocities into the cuboid. In this way, if a cuboid has different magnitudes, it is much probably that "object" or "objects" on it have a distinct magnitude appearance. Thus, given the pixels $p(x, y, t)$ and $p'(x, y, t)$ belonging to the cuboid C_i^t, the vector field $\overrightarrow{v}$ between p and p' is composed of magnitude m and orientation θ. For each cuboid at time t, we compute the cube feature F using Algorithm 1.

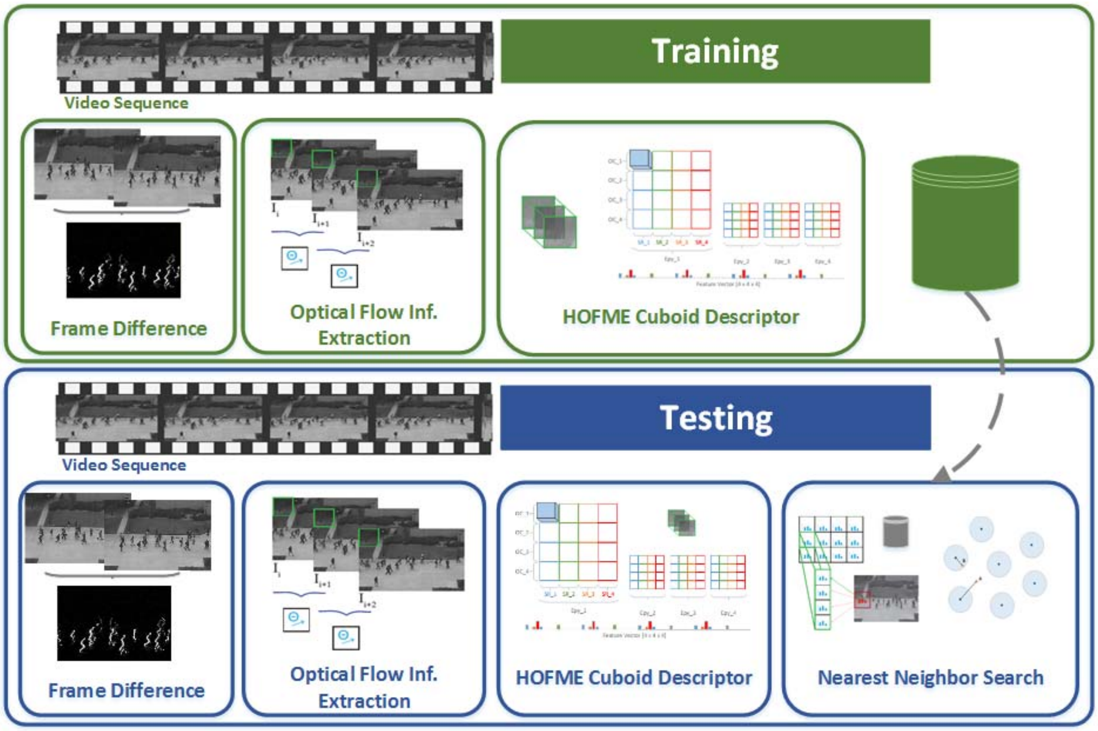


Fig. 1. Diagram illustrating the proposed approach to detect anomalous events.

Algorithm 1 HOFME Descriptor Algorithm

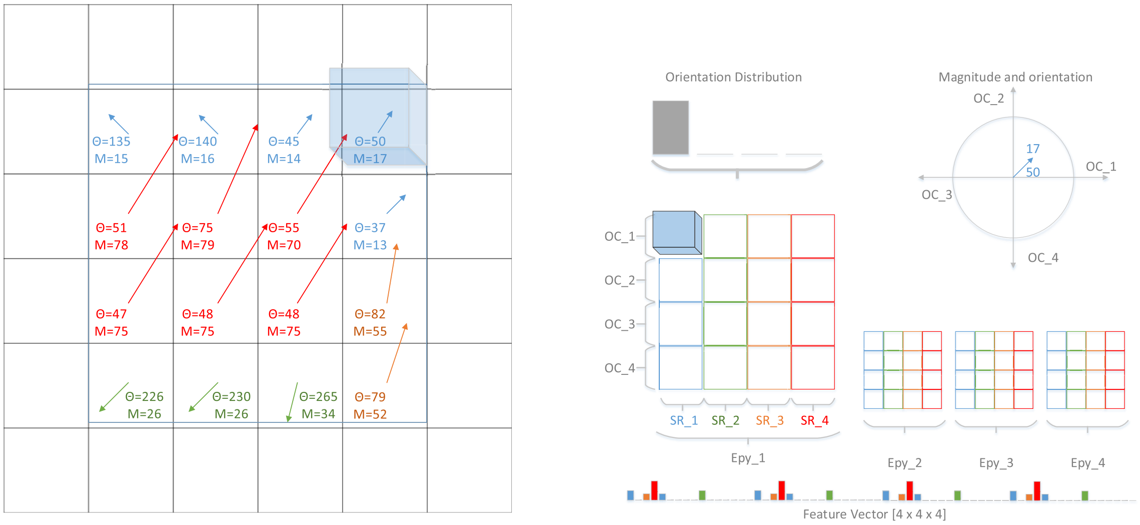
```
1: procedure HOFME($C_i^t$)
2:     $C_i^t$ is the cuboid $i$ at time $t$
3:     for Each pixel $p \in C_i^t$ do
4:         $e \leftarrow Entropy(p)/E$
5:         $s \leftarrow p_\theta/S$
6:         $b* \leftarrow p_m/(B)$
7:         $F[e, s, b*] \leftarrow Hist[e, s, b*] + 1$
8:     return $F$
```

In Algorithm 1, $s \in \{0, 1 \ldots (S-1)\}$, $b* \in \{0, 1 \ldots (B)\}$, and $e \in \{1, 2 \ldots (E-1)\}$ represent the bins in the cuboid, respectively. Mb, Bb, and Eb are the factors for the number of bins (e.g., if we use four bins for orientation, the range Mb is 90°). We increment in one dimension (bin) the magnitude axis in order to consider values that exceeds the maximum value. This is the main reason of using variable b*, in line 6 of the Algorithm 1. To compute the entropy, we use the patch around the pixel p (patch) and the neighborhood orientation information is collected using the s value of pixels (the quantized orientation value). Thus, the first step is to build the orientation distribution around the pixel p. The resultant histogram O_p is normalized to get the probability of each quantized orientation for pixel p; finally, the entropy is computed as

$$\text{Entropy}(p) = -\sum_{i=0}^{S-1} O_p(i)\log[O_p(i)]. \tag{1}$$

Fig. 2 exemplifies the HOFME computation. Fig. 2(a) illustrates the resulting matrix of optical flow from a cuboid C, where θ and m correspond to the angle and magnitude for a pixel p, respectively. Note that we use a single optical flow image and we take a cuboid of size two meaning that we have just a unity time to build a cuboid C composed of only two images. Fig. 2(b) shows a matrix presenting four magnitude and orientation ranges. Each pixel in the cuboid C increments the occurrence of a determinate bin in the cube histogram. In this way, our feature vector can be seen as a cube, where each line corresponds to a determinate orientation range, each column corresponds to the magnitude ranges, and the deep of the cube represents the entropy measure. For instance, the pixel in the example has (50, 17), orientation and magnitude values, and this pixel increments the value in $M_{1\times1}$, since 50° is in the OC_1 range and its speed is between (0, 20], corresponding to the first column. Finally, we compute the entropy orientation measure. Around the pixel $p(50, 17)$, all the adjacent pixels have the same orientation; then the histogram for orientation distribution is concentrated in just one bin and gives 0 entropy as a result. In this way, zero entropy corresponds to Epy_1. Having three coordinates, we can update the feature vector by one unity. Note that, here, we just used $t = 2$. In the case of $t > 2$, there will be more optical flow results per each image pairs, e.g., $t = 4$ yields three optical flow images. This case does not modify the main presented idea, because here we use the information of the pixel in each optical flow result, i.e., each pixel in the cuboid provides information for a determinate bin in the feature vector regarding the same cuboid.

In the previous section, we mentioned that [10], [26], similar to ours, are also based on optical flow information. However, there are some fundamental differences. First, our model instead of accumulating the magnitude value in the orientation bin represents the magnitudes in a different axis, quantizing



(a) (b)

Fig. 2. Example of feature vector extraction using the HOFME descriptor. (a) Resultant cube of the optical flow from a cuboid with (4, 4, 2) dimensions. (b) Matrix presenting four magnitude ranges: $\{(0, 20], (20, 40], (40, 60], (60, \infty)\}$, named SR_1, SR_2, SR_3, and SR_4. All the magnitudes are represented in blue, green, orange, and red, respectively, and four ranges for orientations: $\{(0, 90], (90, 180], (180, 270], (270, 360]\}$, named OC_1, OC_2, OC_3, and OC_4. The entropy can be seen as a third dimension divided into four ranges $\{(0, 1/2], (1/2, 1], (1, 3/2], (3/2, 2]\}$ labeled Epy_1, Epy_2, Epy_3, Epy_4, respectively.

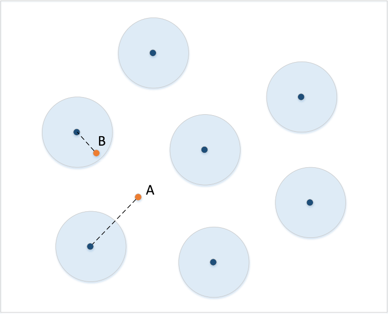


Fig. 3. Nearest neighbor search. Anomalous event pattern is represented by the case of point A and a normal event pattern by the case of point B [6].

Algorithm 2 Anomaly Detection With Nearest Neighbor Search

```
procedure NEAREST NEIGHBOR($P, C$)
    P is incoming pattern for cuboid i
    C is a set o learned patterns for cuboid i
    for w = 1 to W do                ▷ W number of learned patterns
        $d \leftarrow dist(C_w, P)$                ▷ Euclidean Distance
        if $d < \tau$ then
            return $True$
    return $False$
```

the magnitude in ranges and providing information regarding orientation and velocity separately. The second important difference is the entropy; our model also includes another axis, which represents the orientation variation, which provides some information regarding appearance and density, allowing the HOFME to capture all four desirable characteristics described earlier. Finally, the descriptors in [10] and [26] concatenate HOF, HOG, and MBH models into the histogram.

A. Detection of Anomalous Events

The main idea for the prediction step is to search for a pattern that is similar to the incoming pattern. Fig. 3 illustrates this step using blue points to represent patterns seen during training and orange points to represent incoming patterns (test samples). If an incoming pattern is similar enough to some of the known patterns, then it is considered as a normal (point B); otherwise, it will be considered as an anomalous event (point A).

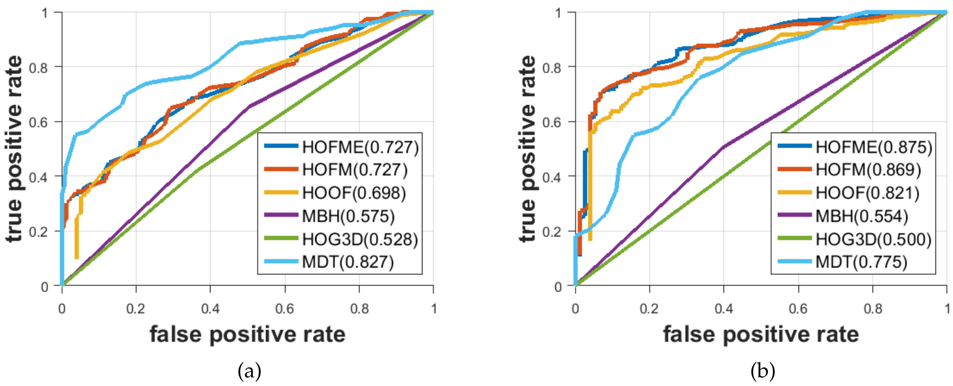
In the recognition step, for each cuboid, we use the descriptors computed during training step to classify an incoming pattern P, at the same spatial location of the cuboid, as anomalous or normal. The steps for the classification process are shown in Algorithm 2. This algorithm returns *False* when none of the patterns seen during the training step is similar to the incoming pattern P, therefore classifying P as an anomaly.

IV. Experimental Results

In this section, we present our experiments. They are presented in two parts. First, the experiments are performed on the UCSD and Subway data sets. Following, we present the experiments with our proposed video data set, Badminton. The proposed approach was developed using the Smart Surveillance Framework [35], built upon OpenCV using the C/C++ programming language.

A. UCSD Data Set

The UCSD Anomaly Detection data set [7] is an annotated publicly available data set for the evaluation of abnormal detection and localization in crowded scenarios featuring pedestrian walkways [17]. The data set was acquired with a



(a) (b)

Fig. 4. ROC curves and the AUC (in parenthesis) for the UCSD Anomaly Detection data set (curves were obtained from [6] and [25] with the addition of our results). (a) Peds1. (b) Peds2.

stationary camera with frames of 238 × 158 pixels and at a frame rate of 10 frames/s. Anomalous events are due to either the circulation of nonpedestrian entities in the walkways or anomalous pedestrian motion patterns.

The UCSD videos are divided into two scenarios: *Peds1* and *Peds2*, each captured by a camera at a different location. The videos recorded from each scenario were split into various video sequences (clips) containing around 200 frames. The number of training sequences is 27 and 16 for Peds1 and Peds2, respectively.

The criterion used to evaluate anomaly detection accuracy was based on frame level, as most of the literature works, in which the algorithm predicts which frames contain anomalous events and those predictions are compared with the ground-truth annotations.

In our experiments, we take five frames to build a cuboid. Thus, the spatiotemporal length is five. The images were scaled to twice their own size, with the aim of getting more motion information. Cuboid spatial dimension (width and height) were set to 30 × 30 pixels. We use four bins for each orientation, magnitude, and entropy range. Due to noise in the optical flow extraction, the model also filters the pixels with tiny moving magnitude by thresholding values lower than 0.5 and removing the respective pixels.

Table I shows the results considering the UCSD data set. In the Peds1 scenario, our method achieved an equal error rate (EER) of 32.0% and an AUC of 0.727, being competitive to most of the reported methods in the literature. On the other hand, on Peds2, we achieved an EER of 20% and an AUC of 0.875, outperforming all reported results. The receiver operating characteristic (ROC) curves for the two scenarios are shown in Fig. 4.

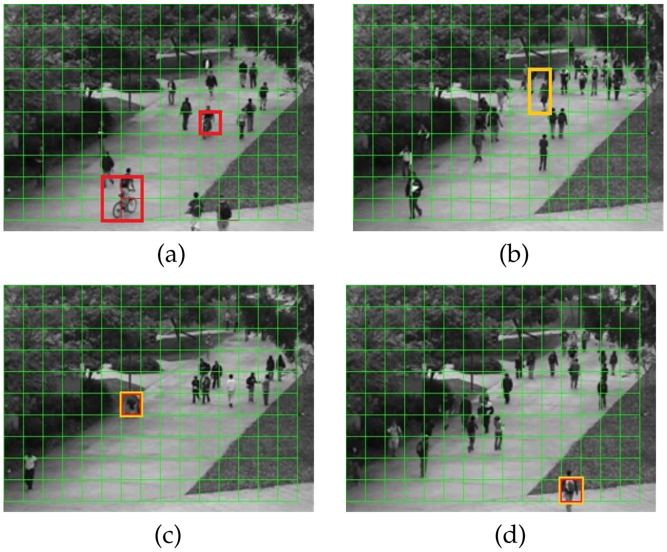
1) Discussion: Similar to our previous work (HOFM descriptor), we investigate the cases where our method failed. Most of the undetected anomalous frames correspond to very challenging cases, such as a skateboarder or a wheelchair going in an almost similar velocity of the pedestrians and with partial occlusions, as shown in Fig. 5(b) and (c). These errors occurred during sequences 21 and 12 of Peds1 and Peds2, respectively. We can see that appearance is an important criterion in the UCSD data set. In this paper, we use magnitude and orientation information. The entropy information is computed over the orientation information and it is used to filter other types of anomalies, especially in very crowd scenes.

Although our model does not include explicit appearance information, it incorporates the spatial characteristic and magnitude entropy information. Thus, higher entropy values may

TABLE I

Anomaly Detection AUC and ERR (%) Results of HOFME on the UCSD Data Set. The Results for [6] and [25] Were Obtained From Their Original Papers. The Descriptors HOG3D, MBH, and HOOF Were Implemented, and Our Anomaly Detection Approach (Switching the HOFME by the Descriptor) Was Executed Using Them

		Peds1		**Peds2**	
	Approach	**AUC**	**ERR (%)**	**AUC**	**ERR (%)**
	MDT-temporal [25]	**0.827**	**25.4**	0.775	25.9
	HOFM [6]	0.727	33.3	0.87	20.7
	HOG3D [9]	0.52	50.0	0.61	47.7
	MBH [10]	0.57	43.4	0.55	45.0
	HOOF [5]	0.69	36.4	0.82	25.9
Our results	HOFME	0.727	33.1	**0.875**	**20.0**



(a) (b) (c) (d)

Fig. 5. Examples analyzed through anomaly detection and extracted from the HOFM detector [6]. (a) True positive. (b) False negative. (c) False positive. (d) False positive.

mean that many pixels are moving with different velocities, and consequently, the object or objects in the scene does not have a regular texture or form, which captures some information regarding appearance and density of the region. For instance, some locations in the scene may not present movement patterns on the training sequences, but on the testing people or another subject might appear in those regions, which should be considered as an anomaly since such patterns were not present in training. In addition, depending on how the ground truth was labeled, these regions may be omitted or considered as a normal situation. Fig. 5(d) illustrates one of such cases.

Finally, we want to highlight the premise used by our model. We use patterns to determine the anomalous cases, i.e., patterns that do not happen during the training phase are considered as anomalous during the test phase. In this way, our model intended to be as general as possible. Another important remark is why we do not include explicitly appearance information in our descriptor. The main reason for this is that the appearance of the unknown is difficult/impossible to quantized. For instance, if we can construct a codebook, we must have the anomalous appearance descriptors; however, in an anomalous specific case, we just have the normal appearance descriptors. We may specialize our descriptor for particular scenes or environments, but this would remove some of generality to the proposed model. Here, we want to remark that even though entropy does not give appearance information directly, it introduces information regarding region texture—when something is moving in the block (see the solid rectangle in Fig. 5), the orientation of the block will be similar and entropy will be low, and in the case of many orientations, for instance, people walking against, entropy will be high. This type of information helps to differentiate some regions especially where there are different flows. Our experiments show that entropy adds some accuracy to the recognition model.

B. Subway Data Set

This data set was proposed by Adam *et al.* [8]. We evaluate our method on the two main sequences recorded from the entrance and exit subway gates. The length for the former is approximately 1 h and 36 min and 43 min for the latter. They contain typical actions performed in such places. The ground truth provided by them marks out the frames where anomalous situations are happening. Specifically, the ground truth points out two types of anomalies: walking in wrong way and jumping the ticket gate.

1) Exit Gate: This clip contains records from a subway exit. In this case, the ground truth considers only people walking in a wrong way. As in [8], we trained with the first 5 min and tested with the remaining frames. The clip contains nine anomalous situations. We counted the matches and false alarms manually. We performed various experiments and the best configuration achieved 100% of accuracy recognizing all anomalous situations. Nonetheless, our descriptor also reported 40 false alarms. Regarding the HOFME setup, we used six bins for orientations, instead of four as used in UCSD. Fig. 6 shows examples of true-positive matches.

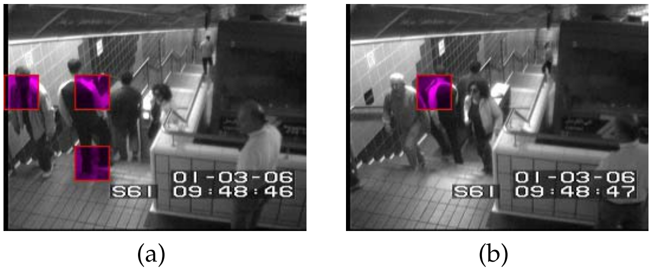


Fig. 6. Examples of true-positive matches for the exit gate clip. Figure (a) and (b) show a correct detection. However, in (a), we can notice a false positive in the leftmost region.

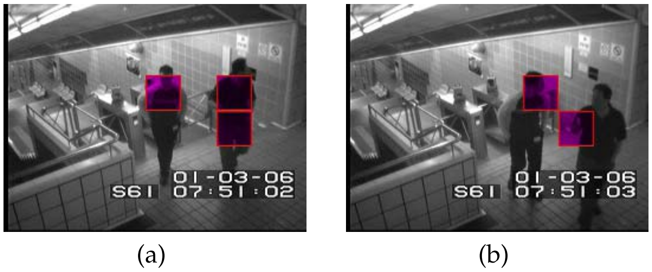


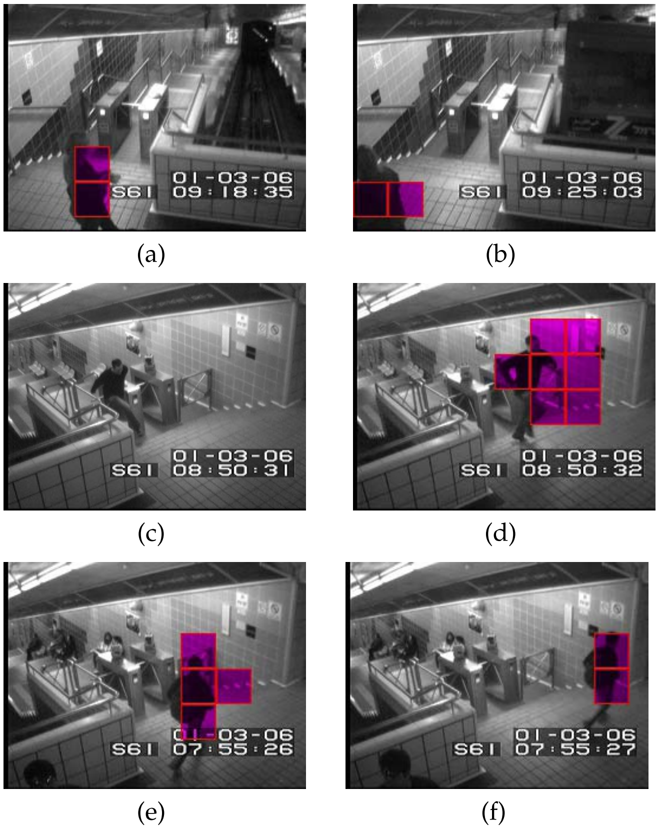
Fig. 7. Examples of true-positive matches for the entrance gate clip. Figure (a) and (b) show a correct detection in both cases.

2) Entrance Gate: This clip contains sequences for the subway entrance gate. In total, we have 31 anomalous situations. Following the protocol presented by authors, we trained with the first 20 min and tested our model using the rest of the clip, which is approximately 1 h and 16 min. We obtained 83% of accuracy. Fig. 7 presents some examples of correct matches.

3) Discussion: Although our model recognized most of the anomalous events, it also presented many false alarms (67 for the entrance gate clip and 40 for the exit gate clip), which happened for the following reason. During the training phase, people move in few directions in the entire scene. Then, in the test, our model detects anomalies where the movement direction did not appear during the training, as expected. However, the focus of the data set ground truth was only the ticket gate, information that cannot be learned from the training data, and it would need extra knowledge. Thus, the false alarms obtained by our methods are due to the creation of the ground truth assuming that anomalies could take place only near the ticket gate.

A second aspect to discuss is the anomaly based on jumping the ticket gate. In this case, the anomaly contains high semantic information. Our descriptor looks for atomic information (such as orientation, entropy, and velocity) discarding explicit modeling of appearance. When people jump the ticket gate, velocity and appearance information may be used to recognize the action. However, during the training, velocity and direction in this specific region is the same when passes through the ticket gate in much cases. In this way, we cannot recognize it as anomalous since during the training, this type of orientation and magnitude appears, and thus it is considered as normal.

In Fig. 8(a) and (b), we can see some of aforementioned situations. Most of them correspond to a person that walks by ignoring the ticket gate. Another important aspect to consider is the velocity. For instance, in Fig. 8(c) and (d), the person jumped the ticket gate, but our model did not detect this action;



(a) (b) (c) (d) (e) (f)

Fig. 8. Examples of false alarms for the subway clips. In (a) and (b) we can see a particular situation where people do not pass through the ticket gate. In (c) and (d), our model does not detect the jumping anomaly. However, it can recognize the speed when a subject starts running. (c) and (d) present a particular situation of a running person, which may represent an anomaly, however the ground truth does not warn about this situation.

however, after that the man started running and our model detected it as anomalous since nobody runs in that direction during the training phase. A very similar situation happens in Fig. 8(e) and (f), where the person appears running out the scene, and this will be considered as an abnormal behavior.

4) Subway and the New Ground Truth: Since the ground truth in subway clips focuses on events that involved the ticket gate, many other events may happen in the whole scene. For instance, the young boy running in the corridor or people walking in forbidden areas.

In this section, we propose an alternative ground truth for the subway data set.[3] The criterion used to determine the anomalies in the clip is based on the following premises: any situation that has not occurred during the training stage is reported as anomalous. We considered as "situation:" the directions, the speed, the location, and also the original subway ground truth. Therefore, for instance, if someone runs in testing and nobody ran during the training, that event will be considered as abnormal.

Table II shows our results and the results achieved using different local feature extraction approaches. In the exit scenario, our method achieved an EER of 17.8% and an AUC of 0.849, outperforming the other descriptors. On the entrance clip, we achieved an EER of 22.8% and AUC of 0.816, outperforming all reported results. The ROC curves for the two scenarios are shown in Fig. 9. Our model outperforms the HOFM, HOOF, MBH, and HOG3D descriptors.

TABLE II
ANOMALY DETECTION AUC AND ERR (%) RESULTS OF THE SUBWAY DATA SET

		Exit		**Entrance**	
	Approach	**AUC**	**ERR (%)**	**AUC**	**ERR (%)**
	HOFM [6]	0.845	18.8	0.815	23.5
	HOG3D [9]	0.524	48.6.3	0.497	50.1
	MBH [10]	0.61	43.4	0.519	48.7
	HOOF [5]	0.8	25.1	0.774	24.4
Our results	HOFME	**0.849**	**17.8**	**0.816**	**22.8**

TABLE III
ANOMALY DETECTION AUC AND ERR (%) RESULTS OF THE BADMINTON DATA SET

		Peds1	
	Approach	**AUC**	**ERR (%)**
	HOFM [6]	**0.806**	28.6
	HOG3D [9]	0.5	50.0
	MBH [10]	0.539	48.7
	HOOF [5]	0.765	**26.2**
Our results	HOFME	0.798	28.0

C. Badminton Data Set

Since there are no many anomaly data sets in the literature, we contribute in this paper by introducing a labeled video sequence.[4] This clip was recorded from a badminton game. In this video, with a total of 345 s captured at 30 frames/s and with a size of 640 × 360 pixels, the ground truth focuses on activities occurred in the grandstand. Along the video, the initial 56 s were used for training to determine what were the normal activities, such as people climbing up the stairs or walking from the right side to the left side of the camera and vice versa. Activities that occurred in the rest of the video that are different from those previously described were considered anomalies. The abnormal activities detected were people running, which occurred three times, and individuals walking down the stairs, which occurred five times.

Table III shows the AUC and EER rates for Badminton data set. According to the results, our model achieved an AUC very similar to HOFM and a smaller EER than HOOF. In Fig. 10, we can see the respective ROC curve. In this case, the HOFM descriptor achieves the better results because the entropy does not add much discriminative information, since the anomalies are activities such as running and moving in a wrong direction, which can be easily captured by HOFM.

1) Discussion: With this experiment, we intended to present a noncontrolled environment. Although the labels are simple, our model recognized accurately most of labeled situations. We focused on the region with people on the grandstand, discarding the players in game and people that cross in front of the camera. Some examples are shown in Fig. 11.

[3] This ground truth will be make publicly available.

[4] This data set will be made available publicly.

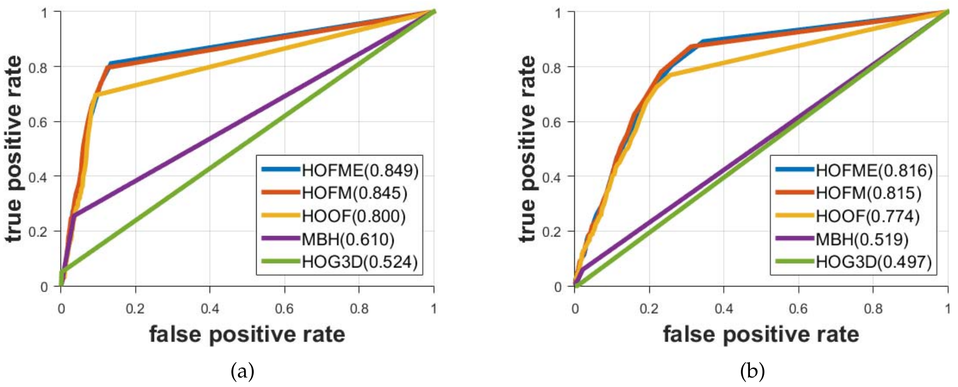


Fig. 9. ROC curve using new ground truth for the Subway data set. (a) Exit. (b) Entrance.

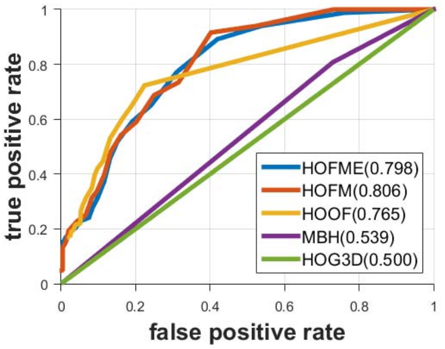


Fig. 10. ROC curves for the Badminton data set.

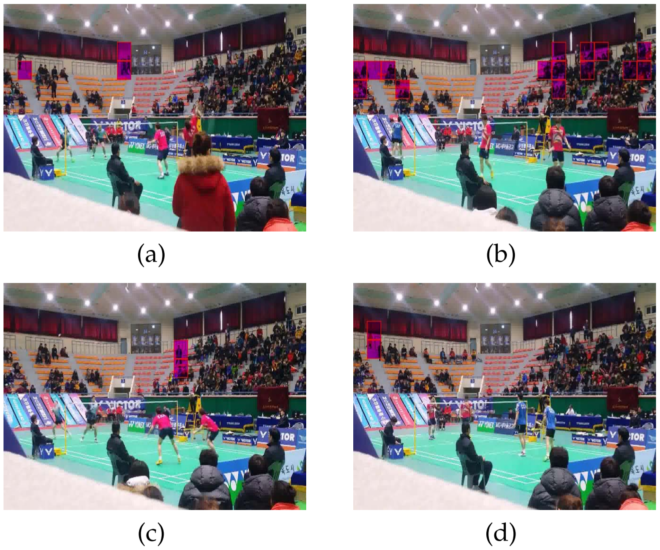


Fig. 11. Examples of anomaly detection in the Badminton dataset. (a) presents correct detections of a person running in the corridor and also a person jumping among grades. In (b) a point was scored by a team and people stand up. In (c) and (d) contain people walking in the corridor which may considered as false alarms, however, during training, nobody walks in that corridor.

V. Conclusion

This paper presented an improvement for the HOFM descriptor. The new descriptor is called HOFME. The main goal of the proposed feature is to improve anomalous event recognition tasks. It makes use of orientation and magnitude from optical flow information in order to create a feature vector for a spatiotemporal region.

We evaluated the performance of our approach compared with the other published results on the UCSD Anomaly Detection data set and Subway data set, two well-known publicly available data sets for the evaluation of anomaly detection. On the UCSD data set, we achieved state-of-the-art results in the Peds2 scenario and our model presented comparable results in the Peds1 scenario. On the Subway data set, we achieved 100% of accuracy recognizing all anomalous situations on the exit gate clip and 83% of accuracy on the entrance gate clip. Although our model recognized most of the anomalous events, it also presented many false alarms. The two main reasons are because the original ground truth focused only on the ticket gate, and during the training phase, people move in few directions. In view of that, we can state that our knowledge regarding anomalous direction is not only for the ticket gate but also for all the scenes. To cope with these situations, we also proposed a new ground truth addressing such anomalies considering all the scenes. Moreover, we introduced a new anomaly detection data set, known as Badminton, composed of a labeled video sequence recorded from a badminton game.

According to the results, HOFME achieved better results on all the performed experiments outperforming the previous HOFM descriptor and other features such as HOOF, HOG3D, and MBH.

References

[1] H. Keval, "CCTV control room collaboration and communication: Does it work?" in *Proc. Human Centred Technol. Workshop*, 2006, pp. 11–12.

[2] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, 2011, Art. no. 16.

[3] F. Jiang, Y. Wu, and A. K. Katsaggelos, "Detecting contextual anomalies of crowd motion in surveillance video," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 1117–1120.

[4] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Textures of optical flow for real-time anomaly detection in crowds," in *Proc. 8th IEEE Int. Conf. Adv. Video Signal-Based Surveill. (AVSS)*, Sep. 2011, pp. 230–235.

[5] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1932–1939.

[6] R. V. H. M. Colque, C. A. C. Júnior, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude to detect anomalous events in videos," in *Proc. 28th SIBGRAPI Conf. Graph., Patterns Images*, Salvador, Brazil, Aug. 2015, pp. 126–133.

[7] (2014). *Semiotics Visual Communication Lab UCSD Anomaly Data Set*. [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/

[8] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[9] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2008, pp. 275-1–275-10.

[10] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, 2013.

[11] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—A review," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 865–878, Nov. 2012.

[12] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *Vis. Comput.*, vol. 29, no. 10, pp. 983–1009, 2013.

[13] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 215–230.

[14] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2227–2234.

[15] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 935–942.

[16] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1446–1453.

[17] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1975–1981.

[18] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3449–3456.

[19] X. Sun, H. Yao, R. Ji, X. Liu, and P. Xu, "Unsupervised fast anomaly detection in crowds," in *Proc. 19th ACM Int. Conf. Multimedia*, Dec. 2011, pp. 1469–1472.

[20] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–6.

[21] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 175–178.

[22] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2921–2928.

[23] B. Antić and B. Ommer, "Video parsing for abnormality detection," in *Proc. ICCV*, Nov. 2011, pp. 2415–2422.

[24] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2909–2917.

[25] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[26] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[27] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2112–2119.

[28] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *Proc. BMVC*, 2015, pp. 548–561.

[29] Y. Yuan, J. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 548–561, Mar. 2015.

[30] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[31] N. D. B. Bruce, "Saliency, attention and visual search: An information theoretic approach," Ph.D. dissertation, Dept. Comput. Sci. Eng. Centre Vis Res., York Univ., Toronto, ON, Canada, 2008.

[32] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006.

[33] Y. Yang, M. Song, N. Li, J. Bu, and C. Chen, "What is the chance of happening: A new way to predict where people look," in *Proc. 11th Eur. Conf. Comput. Vis. Comput. Vis. (ECCV)*, Sep. 2010, pp. 631–643.

[34] J. Y. Bouguet, "Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm," Intel Corp., Tech. Rep. 5.1-10, 2001, p. 4.

[35] A. C. Nazare, Jr., C. E. dos Santos, R. Ferreira, and W. R. Schwartz, "Smart surveillance framework: A versatile tool for video analysis," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2014, pp. 753–760.

Rensso Victor Hugo Mora Colque received the bachelor's degree from Universidad Catolica San Pablo, Arequipa, Peru, in 2009 and the MCS degree from Universidade Federal de Ouro Preto, Ouro Preto, Brazil, in 2012. He is currently working toward the Ph.D. degree with Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil.

He is a Researcher with the Smart Surveillance Interest Group, UFMG. His research interests include computer vision and image processing.

Carlos Caetano received the B.Sc. degree in information systems and computer science from Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, Brazil, in 2011 and the M.Sc. degree in information systems and computer science from Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, in 2014, where he is currently working toward the Ph.D. degree in computer science.

He is a Researcher with the Smart Surveillance Interest Group, UFMG. His research interests include computer vision, smart surveillance and machine-learning applications, with a focus on visual pattern recognition.

Matheus Toledo Lustosa de Andrade is currently working toward the bachelor's degree in control and automation engineering with Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

His research interests include computer vision and smart surveillance, focused on anomaly detection.

William Robson Schwartz received the B.Sc. and M.Sc. degrees in computer science from Federal University of Parana, Curitiba, Brazil, and the Ph.D. degree in computer science from University of Maryland, College Park, MD, USA.

He is a Professor with the Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. His research interests include computer vision, computer forensics, biometrics, and image processing.