Rare Events Detection and Localization In Crowded Scenes Based On Flow Signature

Dieudonné Fabrice ATREVI
Univ Of Orleans, INSA Centre Val de Loire
PRISME EA 4229, F45072
Orleans, France
fabrice.atrevi@univ-orleans.fr

Damien VIVET
Univ of Toulouse
ISAE-SUPAERO, DEOS/SCAN
Toulouse, France
damien.vivet@isae.fr

Bruno EMILE
Univ Of Orleans, INSA Centre Val de Loire
PRISME EA 4229, F45072
Orleans, France
bruno.emile@univ-orleans.fr

***Abstract*—We introduce in this paper a novel method for rare events detection based on the optical flow signature. It aims to automatically highlight regions in videos where rare events are occurring. This kind of method can be used as an important step for many applications such as Closed-Circuit Television (CCTV) monitoring systems in order to reduce the cognitive effort of the operators by focusing their attention on the interesting regions. The proposed method exploits the properties of the Discrete Cosine Transform (DCT) applied to the magnitude and orientation maps of the optical flow. The output of the algorithm is a map where each pixel has a saliency score that indicates the presence of irregular motion regard to the scene. Based on the one class Support Vectors Machine (SVM) algorithm, a model of the frequent events is created and the rare events detection can be performed by using this model. The DCT is faster, easy to compute and gives interesting information to detect spatial irregular patterns in images [1]. Our method does not rely on any prior information of the scene and uses the saliency score as a feature descriptor. We demonstrate the potential of the proposed method on the publicly available videos dataset UCSD and show that it is competitive and outperforms some the state-of-the-art methods.**

***Index Terms*—Rare Event, Salient Motion, Flow signature, Crowded scenes, Discrete Cosine Transform (DCT), One class SVM.**

I. Introduction

Highlight automatically salient motions in a crowded scene is an important task for many computer vision applications such as irregular motions detection in Closed-Circuit Television (CCTV) monitoring system. For those kinds of applications, such a method can be used to help the system's operator to focus attention on irregular activities regions in the scene. As pointed out by Green et al. [2], the CCTV operators attention tends to deteriorate after a long period of monitoring. Rare events detection algorithms are then an excellent solution to reduce the cognitive efforts of operators. In opposite to salient objects detection, a rare events detection algorithm aims to highlight the regions in the scene where important and interesting activities are occurring. Setting up such a system require the need to overcome many challenges especially in machine learning and computer vision domains.

Decades ago, many approaches are proposed to detect and localize rare events in a video clip. In their review [3], Thida et al. defined three main categories of algorithms: macroscopic modeling, microscopic modeling and crowd events detection. The proposed method belongs to the crowd events detection category. This category is more challenging due to the high density of the crowds, the occlusions between individuals, emergent behaviors and self-organizing activities. In a previous work [4], we proposed to model normal event through a bag-of-words approach by using the histogram of oriented optical flow combine with the color histogram of oriented phase in order to identify global rare events sequence in videos. This method analyzes the global dynamic of the scene and identifies frames that contain rare events. The present work differs in term of features and the way to model the events. Different methods focused their effort on the localization of the event in the scene. Recently, Shi et al. in [5] proposed to model the normal events by using the multi-scale histogram of optical flow as features. For localization purpose, they used a spatiotemporal visual saliency algorithm to select the regions where the feature should be extracted. the kernel null Foley–Sammon transform is used to construct the normal event model. Lei et al. [6] proposed a new descriptor denoted "Spatial Associated Multi-scale Histogram of Optical Flow" (SA-MHOF) to perform the same task. This descriptor intends to solve the problem of the objects that are far away from the camera by taking into account the extraction position of the descriptor in the scene during the optical flow computation step. Many others papers [7]–[10] proposed different features and different ways to solve the detection and localization problems. Almost papers in the litterature are based on the histogram of the optical flow orientation combine with different appearance descriptors. Recently, a new group of papers [11]–[14] focuse on using some deep learning algorithms for the task of rare events detection. Some of them used the GAN (Generative Adversarial Networks) approach to model the optical flow information of frequent events. Since the generation of unseen datas is impossible, the rare events detection is based on the reconstruction error between the generated optical flow map of the scene that contains the rare events and the original flow

This work is part of LUMINEUX project, supported by the Regional Centre-Val de Loire (France). The authors would like to acknowledge the Conseil Regional of Centre-Val de Loire for its support.

map. it's obvious that the GAN model can't well reconstruct the optical flow of the rare events in the scene.

In this paper, we introduce a novel rare events detection method inspired by the salient object detection method [1] based on the Discrete Cosine Trasform (DCT). The proposed method is simple, fast yet effective to suppress the dominant crowd flows and model the frequent event while highlighting rares event flows in the scene. We tested the proposed method on the UCSD [15] publicly available and widely used dataset in the literature. Quantitative and qualitative results are shown below.

II. The Proposed Approach

The proposed approach can be split into four complementary steps, as shown on the Fig. 1. It takes as input two consecutive frames and gives as output the salient motion map that highlights regions that contain rare events in the scene. The different steps are described below:

- We start by computing the optical flow of two consecutive frames in order to estimate the motion field.
- The second step consists of computing the signature of both orientation and magnitude of the motion field.
- Based on the signatures of the orientation and magnitude, we estimate the salient motion map that highlights all dominant motions in the scene. The saliency map is divided into blocks and only salient blocks are chosen.
- At the last step, we construct a feature vector for each block. The feature vector is just the vectorization of the salient score of all pixels within the block. For the training phase, the one class SVM algorithm is used to construct a model of normal events. At the classification step, the algorithm gives a score for each selected block.

A. Motion Field estimation

The motion field provides the optical flow information for each pixel of the scene. It can be estimated by using many differents algorithms. The community provides, a few decades ago, robust and efficient algorithms for that purpose. Those algorithms estimate the motion information such as the direction and the velocity at each point of the scene. In this paper, we used the popular algorithm of Horn et al. [16]. Let's denote the intensity value at the pixel location (x, y) at time t by $I_{x,y,t}$. Optical flow algorithms provide coordinate space information of the flow vector at each pixel (V_x and V_y). The orientation $\varphi_{x,y,t}$ and magnitude $r_{x,y,t}$ of the associated flow vector are then expressed by the following formulation:

$$\varphi_{x,y,t} = arctan\left(\frac{V_x}{V_y}\right) \tag{1}$$

and

$$r_{x,y,t} = \sqrt{V_x^2 + V_y^2} \tag{2}$$

Since the aim of our method is to detect rare events according to the motion, we believe that both the orientation and magnitude of the flow should be used. In fact, rare events can be caused by either the orientation of the movement or its velocity. For instance, if a crowd is composed of people walking with the same average velocity and some of them have a different orientation, the method should highlight only those people with a different orientation. In that case, the salient motion is more related to the orientation than the magnitude. Let's consider the example of the salient motion of the Fig 1. In that case, the salient motion is caused by the cyclist whose velocity is more important than that of the pedestrians in the scene.

B. DCT formulation

The Discrete Cosine Transform is widely used in many image processing applications such as video compression, watermarking. It's similar to the Fourier transform with real coefficients and cosine kernel. One of the interesting property of the DCT is the energy grouping since the information of the signal is essentially supported by low-frequency coefficients. It means that the reconstruction of the image can be done from a small number of non-zero coefficients without much loss of information. Given a 1-D signal, the DCT is expressed following the equation 3.

$$F(k) = \frac{1}{2}C(k)\sum_{x=0}^{7} f(x)cos\left(\frac{(2x+1)k\pi}{16}\right) \tag{3}$$

with $F(k)$ the DCT in the frequency domain and

$$C(m) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } m = 0 \\ 1 & \text{otherwise} \end{cases}$$

For an image I represent by the intensity function $f(x, y)$, The 2-D DCT is given by the equation 4:

$$F(k,l) = \frac{1}{4}C(k)C(l)\sum_{x=0}^{7}\sum_{y=0}^{7} f(x,y).cos\left(\frac{(2x+1)k\pi}{16}\right)cos\left(\frac{(2y+1)k\pi}{16}\right) \tag{4}$$

C. Flow Signature

In our approach, we propose to use the Discrete Cosine Transform (DCT) to detect irregular patterns in the optical flow information. Let's consider each component of the original flow field map which is the mixture of regular and irregular motions according to the following structure:

$$y = r + i \quad y, r, i \in \mathbb{R}^N \tag{5}$$

r represents the regular motions and is assumed to be sparsely supported in the standard spatial basis. i represents the irregular motion and is assumed to be sparsely supported in the DCT's basis. Hou et al. [17] demonstrated in their work that it's possible to approximately isolate the support of an image by taking the sign of the mixture signal of the image in the transformed domain and then inversely-transform it back into the spatial domain. Based on their founding, we use the flow signature which is defined by (6) as the sign function of the DCT coefficients of the flow vector. The flow signature discards amplitude information across the entire frequency and is very compact with a single bit per component.

$$FlowSignature(y) = sign(DCT(y)) \tag{6}$$

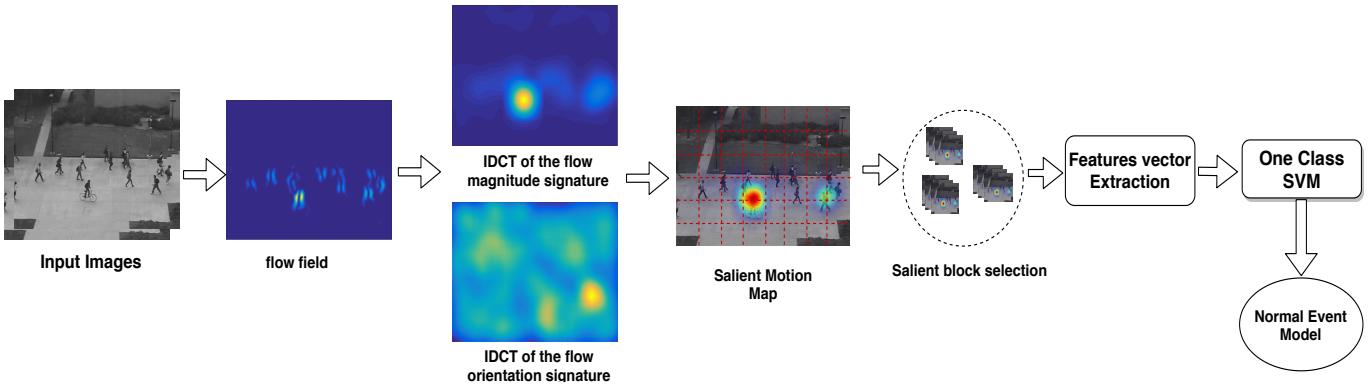

Fig. 1. Overview of the proposed method

where y can be the orientation or the magnitude information of the optical flow.

D. Motion Saliency Map Generation

Given the flow field of the scene, the computation of the signature of both the orientation and the magnitude maps is processed. Those signatures are used to reconstruct the flow field via the Inverse Discrete Cosine Transform (IDCT) which is proved to "concentrate the image energy at the locations of a spatially sparse foreground, relative to a spectrally sparse background" [17] .

$$\bar{y} = IDCT(FlowSignature(y)) \tag{7}$$

We assume that a salient motion is visually conspicuous relative to the rest of scene, then the saliency map is generated by smoothing the square of the reconstructed map defined above by the following formulation of (8) where g is a Gaussian kernel and $\circ$ is the Hadamard product operator. Such filtering is an important step as it removes the noise introduced by the sign quantization. A simple Gaussian smoothing is adequate because the support of a salient motion is usually not only spatially sparse but also localized in a contiguous region.

$$S = g * (\bar{y} \circ \bar{y}) \tag{8}$$

To take into account salient motions generated by both orientation and magnitude, the orientation saliency map S_φ and the magnitude saliency map S_r are generated following the same process. The final salient motion map is computed by weighting S_φ by S_r:

$$S_f = S_\varphi \circ S_r \tag{9}$$

E. Features Extraction and Model Construction

The salient motion map S_f is represented by a matrix corresponding to the saliency score of each pixel. The pixels with 0 as score belong to regions of the scene where no irregular motion is happening and pixels with the highest value belong to irregular motions regions. But, a salient pixel can't be automatically considered as rare event pixel. The goal of this section is to construct a model of usual events from a training set that contains only the usual events. For that purpose, each saliency map is divided into fixed size block where only blocks with a mean score above a dynamic threshold will be considered for the model construction. The value of this threshold is empirically set to $2 \times mean(saliencyMap)$. This step helps to reduce the number of blocks that are necessary for the modelization. For each block, the feature vector is obtained by vectorizing the score matrix of the block:

$$FeatureVect_i = vect(score(block[i])) \tag{10}$$

where $vect$ is the function that converts a matrix to a vector.

To build the model of usual events, we use the one class version of the SVM algorithm. The goal of this algorithm is to find the best frontier around the training set. Different kind of frontier can be used: linear, circular, etc. For identifying the rare events blocks, we use as the score the distance of the feature vector to the origin of the feature space. The higher the distance, the higher is the probability that the block contains a rare event.

III. EXPERIMENTAL RESULTS

In this section, we introduce the experimental results that show our method competitive and promising. We conducted the experiments on the publicly available dataset for abnormal events detection: the UCSD[1] dataset. The aim of the experiments is to model the usual events from a training set and localize in the scene any rare events. The UCSD dataset is composed by two separate sub-datasets: Ped 1 and Ped 2. The Ped 1 dataset contains 34 training videos and 36 test videos. The Ped 2 contains 16 training videos and 13 test videos. All training set contains people walking with various speed and orientation. The test set videos contains various rare events such as fast and zig-zag cyclists motions in the scene and vehicles passing through the scene. We start by applying a pre-processing (typically a Gaussian filter) to the image before the optical flow computation, in order to enhance the images quality. We used the Horn-Shunk algorithm [16] for the optical flow computation. The optical flow maps are resized to 64×64

[1] http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

before the DCT computation. This enable the reduction of the computation time cost. The generated saliency map is finally resized back to the original size of the image before the blocks selection step.

For quantitative evaluation, we used the AUC (Area Under the Curve) of the ROC curve, the EER (Equal Error Rate) and the RD (Rate Detection) criteria as detection performance measurement. The ROC curve is based on the false positive rate and true positive rate computed for various threshold set on the decision score. Those three criteria are commonly used in the literature to show how good is a detection method. The best approach should get the highest AUC and RD score and the lowest EER. In this paper, we conducted two levels of evaluation: the frame level and the pixel level. For the frame level evaluation, an image contains rare events if and if only at least one block gets a score above the threshold. The ground truth, in that case, contains label for each frame of the video. This evaluation doesn't give any information about the methode ability to events localization. So, for the localization, the pixel level evaluation is suitable. For this evaluation, an image is classified as containing rare events if the detected rare events cover at least 40% of the ground truth (GT). The GT is a binary mask where the rare event pixels have value 1 and the others pixels has the value 0. The EER criteria are used for frame-level evaluation and the RD criteria for the pixel level. The Ped 2 dataset is fully-annotated for both frame and pixel level while the Ped 1 dataset is partially-annotated for pixel level evaluation.

We firstly study the influence of the SVM kernel on the performance of our proposed method. Since differents kernel exist in the litterature and can influence the performance, we test three of them: the linear kernel, the polynomial kernel and the gaussian kernel.

TABLE I
SVM KERNEL INFLUENCE ANALYSIS

	Kernel		
Dataset	**Linear**	**Polynomial**	**Gaussian**
Ped 1	**0.7854**	0.7791	0.7817
Ped 2	0.7904	0.7849	**0.7977**

The table I show that the SVM kernel has less influence on the performance of our approach. Since the linear kernel is faster, we use it for next experiment. We also conduct an experiment in order to compare our descriptor based on the saliency score to the well-know Histogram of Oriented Optical Flow (HOOF) [4].

Table II and table III show that by using the saliency score as a descriptor, the detection performance is better than when the descriptor HOOF is used.

Table IV and V show the detection performances of our method compare to some of the state-of-the-art methods. From those results, one can notice that our method outperforms some prior works like the spatial version of the Mixture of Dynamic Texture (MDT), the social force approach, the MPPCA based method, Adam's method [7] and the Energy-based method [8], on both the Ped 1 and Ped 2 datasets. However, compared to Zhang et al. [18], SA-MHOF [6] and TCP [12] methods, our method is less accurate.

TABLE II
COMPARISON OF OUR DESCRIPTOR WITH HOOF ON PED 1

Evaluation level	**Frame**		**Pixel**	
Descriptor \ **Criteria**	**AUC**	**EER**	**AUC**	**RD**
HOOF	0.77	0.30	0.56	0.52
Our	**0.78**	**0.28**	**0.58**	**0.56**

TABLE III
COMPARISON OF OUR DESCRIPTOR WITH HOOF ON PED 2

Evaluation level	**Frame**		**Pixel**	
Descriptor \ **Criteria**	**AUC**	**EER**	**AUC**	**RD**
HOOF	0.75	0.33	0.69	0,63
Our	**0.79**	**0.30**	**0.77**	**0.69**

TABLE IV
COMPARISON OF OUR METHOD WITH PRIOR WORKS ON PED 1

Evaluation level	**Frame**		**Pixel**	
Method \ **Criteria**	**AUC**	**EER**	**AUC**	**RD**
MDT temporel [7]	0.82	0.23	0.57	0.59
MDT spatial [7]	0.6	0.43	0.66	0.5
Social Force [7]	0.69	0.36	0.22	0.41
MPPCA [7]	0.67	0.35	0.22	0.23
Adam (MHL) [7]	0.63	0.38	0.16	0.42
Energy-based [8]	0.7	0.35	0.49	0.67
Zhang et al. [18]	0.86	0.19	0.76	0.74
SA-MHOF [6]	0.86	0.19	0.63	0.68
Our Approach	**0.78**	**0.28**	**0.58**	**0.56**

The mains advantages of our method are its simplicity. The proposed feature based on the motion saliency score, is really simple to understand, to implement and fast to compute. We can also notice that our feature is also used for region selection and that provides an advantage in term of complexity compared to others methods which imply different algorithms for the region selection and the model construction. Unfortunately, it'll difficult to compare the computation time

TABLE V
COMPARISON OF OUR METHOD WITH PRIOR WORKS ON PED 2

Evaluation Level	Frame		Pixel	
Method \ Criteria	AUC	EER	AUC	RD
MDT temporel [7]	0.76	0.28	0.52	0.57
MDT spatial [7]	0.75	0.29	0.66	0.63
Social Force [7]	0.7	0.35	0.22	0.28
MPPCA [7]	0.71	0.36	0.22	0.22
Adam (MHL) [7]	0.58	0.46	0.16	0.22
Energy-based [8]	0.86	0.16	0.72	0.84
TCP [12]	0.88	0.18	-	-
Our Approach	**0.8**	**0.3**	**0.79**	**0.69**

of our method with others due to the difficult access to the codes.

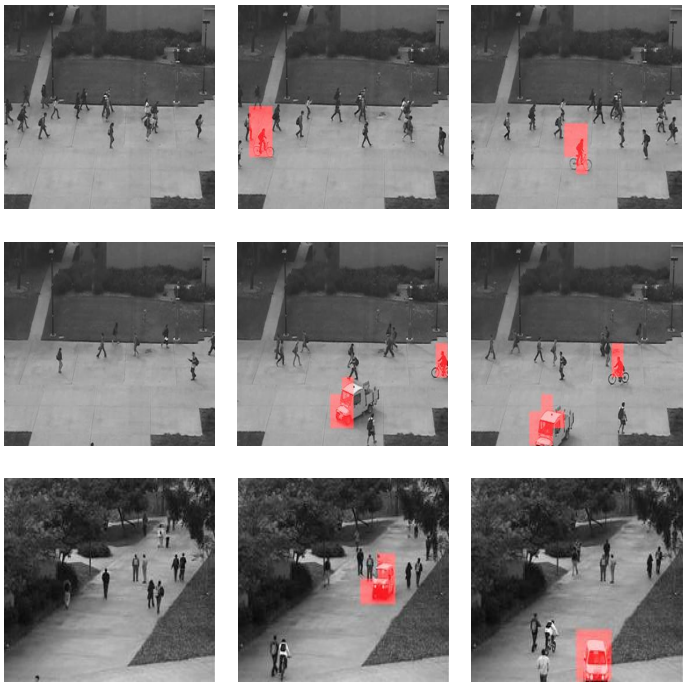

Fig. 2. Some qualitative results from Ped 1 and Ped 2 datasets

Fig. 2 shows some qualitative results extracted from three videos of the datasets. Each image is related to a scenario with one rare event that correspond to the bicycle and the car passing through the scene. The red rectangle overlay on the images shows the localization of the events. It's evidence in those images that our method accurately detects the interesting region. We can conclude that the method focuses the attention on the irregular motions that generate rare events. However, the method fails to detect some events. By analyzing the performance of the proposed method on each video of the datasets, we discover that it fails on video where events are related to the low speed of the entity that generate this event. It also fails to detect some motion where the clothe's color is mingled with the background. The former problem can be solved by improving the quality of the optical flow. It's a fact that our method highly depends on the quality of the optical flow.

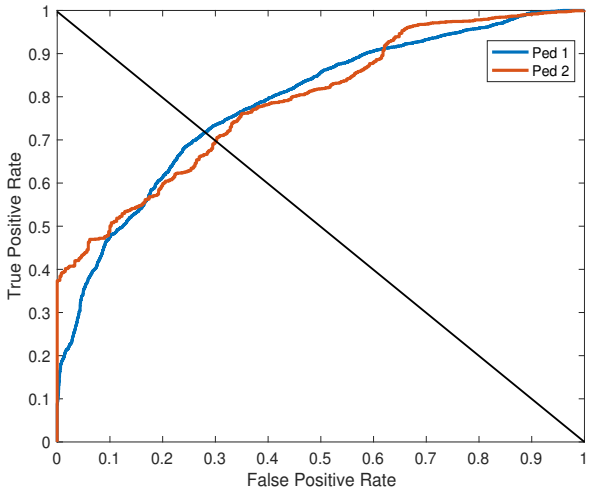

Fig. 3. ROC curve for Frame Level Evaluation

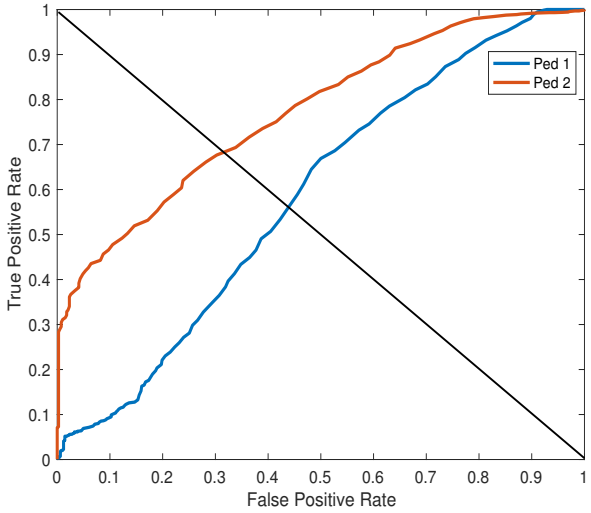

Fig. 4. ROC curve for Pixel Level Evaluation

Fig. 3 and Fig. 4 present respectively the associated ROC curve for frame level and pixel level performances of the proposed method. It's evidence that for both two levels of evaluation, the method performs better on the Ped 2 dataset especially for the localization. It's important to notice that the pixel evaluation of the Ped 1 dataset was conducted on 11 videos that have ground truth. Those 11 videos contain events with various difficulties such as the perspective deformation

which made the detection task difficult for an object which is far from the camera.

IV. CONCLUSION

In this paper, we proposed a new method that highlights rare events in videos. The proposed method uses the notion of flow signature that corresponds to the sign function of the Discrete Cosine Transform of a signal. A salient motion map is computed by using the Inverse Discrete Cosine Transform of the flow signature for both the magnitude and the orientation information. The final salient motion map is obtained by weighting the saliency map of the orientation by the saliency map of the magnitude. We proposed to use this final map to select the salient regions of the scene and use their saliency score as a descriptor. The modelization of the usual events is made with the one class SVM algorithm which helps to define a frontier in the feature space around the frequent event point. We conducted two levels evaluation based on the classification score given by the SVM. Experimental results show that the proposed method is competitive compared to prior works.

REFERENCES

[1] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):194–201, 2012.

[2] Mary W Green. The appropriate and effective use of security technologies in us schools: A guide for schools and law enforcement agencies series. *NCJ*, 1999.

[3] Myo Thida, Yoke Leng Yong, Pau Climent-Pérez, How-lung Eng, and Paolo Remagnino. A literature review on video analytics of crowded scenes. In *Intelligent Multimedia Surveillance*, pages 17–36. Springer, 2013.

[4] Dieudonne Fabrice Atrevi, Damien Vivet, and Bruno Emile. Bayesian generative model based on color histogram of oriented phase and histogram of oriented optical flow for rare event detection in crowded scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3126–3130. IEEE, 2018.

[5] Yanjiao Shi, Yunxiang Liu, Qing Zhang, Yugen Yi, and Wenju Li. Saliency-based abnormal event detection in crowded scenes. *Journal of Electronic Imaging*, 25(6):061608, 2016.

[6] Lei Hu and Fangyu Hu. Anomaly detection in crowded scenes via samhof and sparse combination. In *Computational Intelligence and Design (ISCID), 2017 10th International Symposium on*, volume 1, pages 421–424. IEEE, 2017.

[7] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2014.

[8] Hung Vu, Dinh Phung, Tu Dinh Nguyen, Anthony Trevors, and Svetha Venkatesh. Energy-based models for video anomaly detection. *arXiv preprint arXiv:1708.05211*, 2017.

[9] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu. Learning deep event models for crowd anomaly detection. *Neurocomputing*, 219:548–556, 2017.

[10] Yang Cong, Junsong Yuan, and Ji Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851–1864, 2013.

[11] Jiayu Sun, Xinzhou Wang, Naixue Xiong, and Jie Shao. Learning sparse representation with variational auto-encoder for anomaly detection. *IEEE Access*, 6:33353–33361, 2018.

[12] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698. IEEE, 2018.

[13] Tianlong Bao, Saleem Karmoshi, Chunhui Ding, and Ming Zhu. Abnormal event detection and localization in crowded scenes based on pcanet. *Multimedia Tools and Applications*, 76(22):23213–23224, 2017.

[14] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017.

[15] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE, 2010.

[16] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[17] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[18] Xinfeng Zhang, Su Yang, Xinjian Zhang, Weishan Zhang, and Jiulong Zhang. Anomaly detection and localization in crowded scenes by motion-field shape description and similarity-based statistical learning. *arXiv preprint arXiv:1805.10620*, 2018.