Atrevi, Dieudonné Fabrice and Vivet, Damien and Emile, Bruno Bayesian Generative Model Based on Color Histogram of Oriented Phase and Histogram of Oriented Optical Flow for Rare Event Detection in Crowded Scenes. (2018) In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 15 April 2018 - 20 April 2018 (Calgary, Canada).

BAYESIAN GENERATIVE MODEL BASED ON COLOR HISTOGRAM OF ORIENTED PHASE AND HISTOGRAM OF ORIENTED OPTICAL FLOW FOR RARE EVENT DETECTION IN CROWDED SCENES

Dieudonne Fabrice ATREVI[*], *Damien VIVET*[†], *Bruno EMILE*[*]

[*] University Of Orleans, INSA Centre Val de Loire, PRISME EA 4229, F45072, Orleans, France
[†] University of Toulouse, ISAE-SUPAERO, DEOS/SCAN, Toulouse, France
[*]{fabrice.atrevi, bruno.emile}@univ-orleans.fr, [†]damien.vivet@isae.fr

ABSTRACT

In this paper, we propose a new method for rare event detection in crowded scenes using a combination of Color Histogram of Oriented Phases (CHOP) and Histogram of Oriented Optical Flow (HOOF). We propose to detect and filter spatio-temporal interest points (STIP) based on the visual saliency information of the scene. Once salient STIPs are detected, the motion and appearance information of the surrounding scene are extracted. Finally, the extracted information from normal scenes are modelled by using a Bayesian generative model (Latent Dirichlet Allocation). The rare events are detected by processing the likelihood of the current scene in regard to the obtained model. The proposed method has been tested on publicly available UMN dataset and compared with different the state-of-the-art algorithms. We have shown that our method is very competitive and provides promising results.

***Index Terms*—** Rare event, Crowded scenes, CHOP, HOOF, Latent Dirichlet Allocation

1. INTRODUCTION

In order to increase the security in public areas such as airports, subways, markets, automatic scene understanding algorithms are required in order to detect rare events or dangerous behaviours. Setting up such system required to overcome many challenges specially in machine learning and computer vision domains. The objective is to propose reliable algorithms that can highlight in realtime, suspicious, abnormal and rare behavior of crowds or people. The problem of such approaches is the lack of database as we are searching for rare events, by definition, such behaviour are not included in provided database. That is why the community proposed to focus on the detection of outliers regarding a normal defined situation.

In order to solve this problem, many approaches have been proposed in the literature. In the review [1], Thida et al. defined three main categories of algorithms: macroscopic modeling, microscopic modeling and crowd events detection. The present work 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. Recently, Ang Li et al.[2] proposed an approach based on the sparse reconstruction of the histogram of maximal optical flow projection. The HMOFP feature encodes the motion of the crowd using the optical information. Prior works also proposed to model events based on the dynamic of the scene. R. Mehran et al.[3] proposed an approach based on the social force interaction between particle, moved by the optical flow on regular grid. An approach based on graph modelization was recently introduced in order to incoporate the spatial distribution of interest point detected in the scene[4]. All of these approaches focus on the modelization of the motion in the scene. However, combining motion and appearance information can improve the accuracy of the method in case when the abnormality is also related to the shape of objects present in the scene For example, in case of car or vehicle intrusion in a pedestrian area.

The present work focuses on motion and appearance feature extraction by introducing a new appearance feature based on the phase congruency combine with the histogram of oriented optical flow. In the section 2, we introduce the proposed approach and give more details about the feature description and the normality learning process. In the section 3, we present the evaluation of our method on the public abnormal dataset UMN[1].

2. METHODOLOGY

We propose a new complete framework for rare event detection in a crowd scene. The Fig 1 shows an overview of the method. Our proposition can be divided into three steps:

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.

[1] Unusual crowd activity dataset made available by the University of Minnesota http://mha.cs.umn.edu

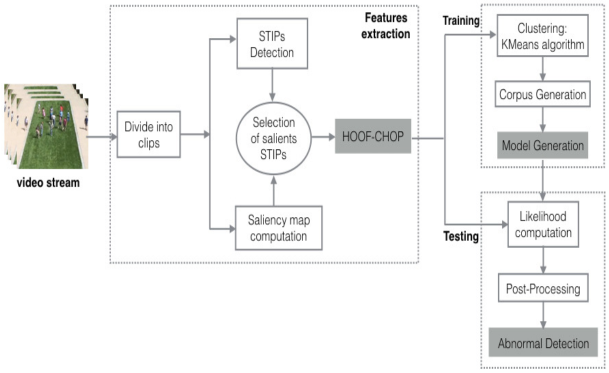

Fig. 1. Overview of the proposed approach

- spatio-temporal interest point detection based on the visual saliency of the scene,
- an appearance and motion feature description of STIP,
- a normal event modelization using the Bayesian latent dirichlet allocation algorithm [5].

2.1. Spatio-Temporal Interest Points extraction

In crowd scenes, it is a challenging task to detect people and extract information about them. To overcome this problem, in prior works, some authors modelled the scene by dividing the images in patches and then extract the global information within the patches. This can be inefficient as not all pixels are useful for event modelization. To avoid this, one can perform background subtraction or optical flow magnitude thresholding. Another solution is to use interest points in the image.

This work focuses on STIP detected in the images that have a strong visual saliency (VS) score. VS algorithms intend to highlight salient regions in images and to separate foreground and background. The salient region may contain people, cars, or other objects of interest. Our assumption is that events are generated by salient and moving object such as pedestrian, cars or cyclists. The output of such algorithm is a map of saliency score. We proposed to select the STIPs based on their saliency score. We use the well-known Harris corner detector [6] and the visual saliency algorithm proposed in [7]. STIPs are extracted in the central frame of a volume of T frames (called clips) and are ranked by the variance value of the visual saliency of the clip. The experimental result (see Fig 2) shows that the selected points are well located around moving objects in the scene while other STIPs are filtered out.

2.2. Features Extraction

In the literature, event classification models are based mainly on the motion information [2][3]. Recently, some works consider the appearance information [8][4]. The importance of each kind of information depends on the type of abnormality.

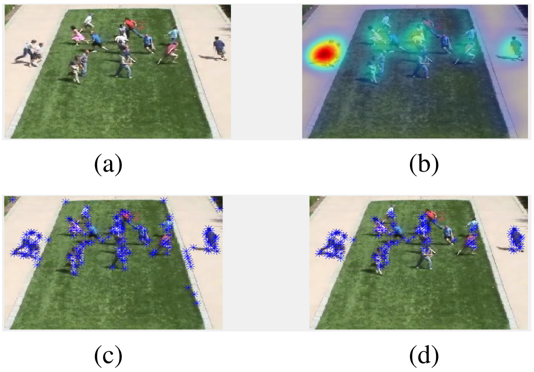

Fig. 2. Spatio-temporal Interest Points (STIPs) selection. (a) Original image, (b) Visual saliency map. (c) basic Harris feature detection (d) Saliency based STIPs filtering

In the case we have no *a priori* about the type of event, both information has to be considered. In this paper, we proposed a framework combining a modified version of HOOF[9] and the Color Histogram of Oriented Phase (CHOP). To our knowledge, it's the first time that CHOP has been used as a descriptor in event analysis context. In this section, we briefly introduce the two descriptors.

2.2.1. Histogram of Oriented Optical Flow

Inspired by the success of histogram features, Chaudhry et al.[9] introduced the histogram of oriented optical flow. The optical flow provides the motion information between two consecutive frames, such as the direction and the orientation of each pixel. In this paper, the optical flow information about the clip is computed based on the well-known Lucas-Kanade algorithm. Chaudhry et al. used the proposed descriptor (HOOF) to perform the human action recognition. The histogram building process consists of "binning the flow vector according to its primary angle from the horizontal axis and weighted according to its magnitude" [9]. Each bin of the histogram represents a set of flow orientation in the scene. In their work, they divided the orientation space in a way to be invariant to the scale and direction of the motion (θ and $\pi - \theta$ vote for the same bin), useful for human activity recognition. But, in rare event detection context, it's important to represent all motion direction. We modify the proposed algorithm of Chaudhry et al. to consider all flow orientations between $-\pi$ and π. Such descriptor captures the direction and speed of the objects in the scene, essential to differentiate normal movement from abnormal one, such as in a panic situation.

2.2.2. Color Histogram of Oriented Phase

The Color Histogram of Oriented Phase was introduced by Ragb et al.[10] as a new descriptor from humans and objects. This descriptor is based on the local phase congruency of the image. It aims to resolve the illumination and contrast varia-

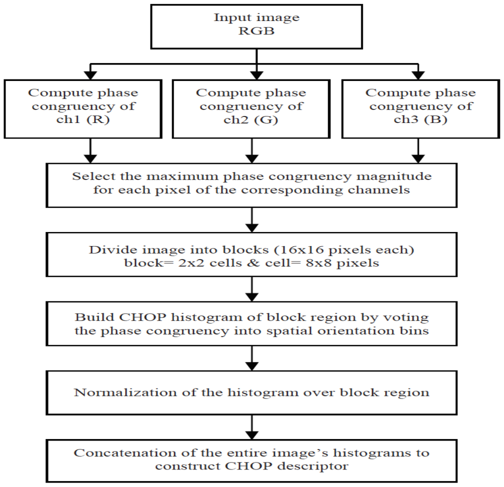
Input image
RGB
Compute phase congruency of ch1 (R)
Compute phase congruency of ch2 (G)
Compute phase congruency of ch3 (B)
Select the maximum phase congruency magnitude for each pixel of the corresponding channels
Divide image into blocks (16x16 pixels each) block= 2x2 cells & cell= 8x8 pixels
Build CHOP histogram of block region by voting the phase congruency into spatial orientation bins
Normalization of the histogram over block region
Concatenation of the entire image's histograms to construct CHOP descriptor

Fig. 3. CHOP feature vector extraction framework [10]

tion problem in object recognition. As shown in the Fig 3, the Phase Congruency (PC) is computed on each color channel of the image and the maximum values in each corresponding pixel of the three channels is selected. The resulting phase orientation map is used to construct the histogram by dividing the images into blocks and cells. The PC concept is based on the model of the local energy algorithm introduced by Morrone and Owens [11] which postulates that “the features are perceived at the points where the Fourier components of the image are maximally in phase”. According to Ragb et al. [10], following the assumption of Kovesi et al [12] who stated that “feature types include line, step edges, Mach bands, and roof edges give rise to points where the PC is high”, the PC would be maximum at the points. Given an input signal $I(x)$, the PC is defined as [10]:

$$PC(x) = \frac{E(x)}{\epsilon + \sum_n A_n} \tag{1}$$

where $E(x)$ is the local energy, A_n all the Fourier component amplitudes and ϵ a small quantity to avoid zero division. Using a filtered signal $F(x)$ of $I(x)$ from DC component and the Hilbert Transform $F(x)_H$ (90° phase shift of $F(x)$) , the local energy is defined as [10]:

$$E(x) = \sqrt{F(x)^2 + F(x)_H^2} \tag{2}$$

Ragb et al. [10] computed the PC by convolving the image with a pair of quadrature log-Gabor filter to extract the local frequencies and phase information. We use the same filter and follow the same framework in this paper. It has been shown that CHOP outperforms the popular HOG descriptor in illumination and contrast changing context. Such robustness is essential in event detection applications.

At this step, the salient elements of the scene are extracted and described in term of movement (HOOF) and in term of shape (CHOP). The final feature vector of STIPs is the concatation of the HOOF and CHOP features vectors. The normal activity learning step can be launch with this features.

2.3. Learning Process

In order to model the usual event occurring in a training dataset, we use the Latent Dirichlet Allocation (LDA) algorithm. This algorithm is well-known in the text mining domain. The LDA algorithm, introduced by Blei et al. [5], takes as input a corpus of documents and discover some topics addressed in the corpus by a generative process. Even if these topics have an intuitive interpretation in text mining, such interpretation isn't easy for event classification. We suppose that in a training set (that contain only normal scene), normal event can be identified by some repetitive behaviour, indeed patterns, in the scene. Such patterns can be seen as a group of visual word extracted from the feature vectors. Those patterns can be assimilated to a topic that can give a semantic modelisation of the normal event in the training set. The role of LDA algorithm is to extract such semantic modelisation. In order to use the LDA, we used the bag-of-word approach to represent each clip as a document. The first step consists of clustering all the feature vectors from the training data set, using the classical KMeans algorithm, to identify clusters (called vocabulary). The transformation of a video clip to a document is done by processing the histogram of the occurrence of each visual word of the vocabulary in the clip. In this, each visual word is represented by the closest cluster mean feature vector in term of Euclidian distance. The final corpus is used as an input for the LDA. As an output, LDA provides the probability density function (α and β) of the topics over the corpus. Based on these estimated parameters, we can calculate the likelihood (eq 3) of each frame inside the video clip with reference to the model .

$$l(\alpha, \beta) = \log p(D|\alpha, \beta) \tag{3}$$

where D is a video clip, α and β are the learned parameters.

3. EXPERIMENTAL RESULTS

In this section we will present the experimental results that show our method is competitive and promising. We tested our approach on the publicly available dataset of abnormal crowd activities of the university of Minnesota [13]. It contains 11 videos of panic events in three different environments (2 outdoors scenes and 1 indoor scene) (figs 4,5,6). Each video starts with people walking randomly and finishes by panic movement where people run in all directions with different speeds. Based on fixed threshold of the likelihood, video frames can be classified as containing an usual or rare event. A post-processing is applied on the likelihood of the frames belong to the same clip. This post-processing intend to correct the time lags by applying the Savitzky-Golay algorithm as filter. We consider clip size of 15 frames (about 1/2 seconds) to analyze the events. Our vocabulary is set to 40 visual words, the number of topics to 40 topics and the patches size around STIPs to 16 x 16 pixels. The feature vector size for each point in the clip is about 960-D. The training set is composed of only normal frames from the first 5 videos of scene 2 (indoor). This composition allows us to train the model on one scene and test on the remain video of different scenes. This choice intended to show that the model is not dependant on the environment but on the activity.

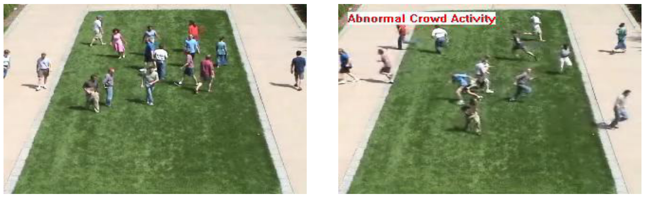

Fig. 4. Normal and abnormal frame of lawn scene

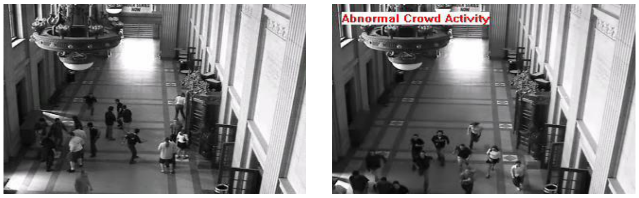

Fig. 5. Normal and abnormal frame of indoor scene

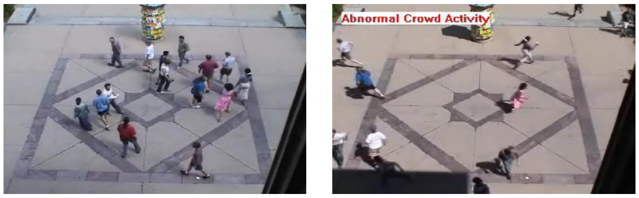

Fig. 6. Normal and abnormal frame of plaza scene

The figure 7 shows the performances of our algorithm. It outperforms the existing method based on the bag-of-words using STIP, SIFT and Dense Trajectories [4]. The table 1 summarizes the performances of different bag-of-words approaches on the UMN dataset. We also compare our method with the others algorithms not based on bag-of-words (see table 2). Our solution outperforms methods using pure optical flow [3] and is comparable to the NN [14]. However, compared to STCOG[15], HMOFP[2] and Social Force [3] scores, our results are less accurate. This can be explained by differents reasons. First, we study the entire image in globality while other technics made local analysis. Second, we choose to learn on one scene (number 2) and to test on all the outdoor scene in order to check the environment influence on the method. Moreover Scene 2 is complicated as we observe often an empty corridor or a concentration of peaple in a small area more often at the bottom of the image. Considering these parameters, we think that our approach is very promising as it provides good results that can only be improved.

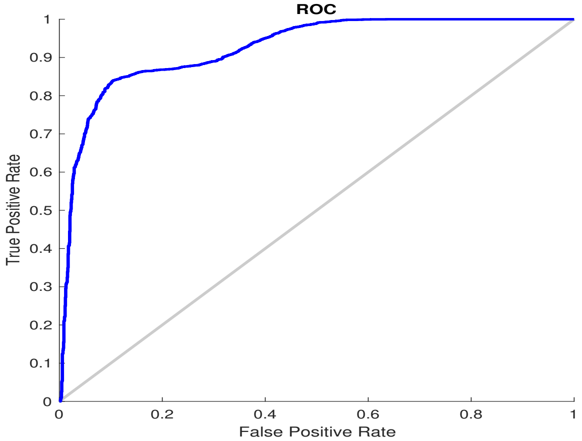

Fig. 7. ROC curve of the proposed algorithm

Table 1. Comparison of our method with bag-of-words existing methods [4]

Method	SIFT	STIP	DT	Bag of Graph	**Ours**
AUC	0.85	0.85	0.81	0.95	**0.93**

Table 2. Comparison of our method with prior works

AUC / Method	Lawn	Indoor	Plaza
Pure optical flow [3]	0.84		
Social Force [3]	0.96		
NN [14]	0.93		
STCOG [15]	0.9362	0.7759	0.9661
HMOFP [2]	0.9976	0.9570	0.9869
Ours	**0.93**		

4. CONCLUSION

In this paper, we proposed a new method that integrated the visual saliency for spatio-temporal interest point selection, a new feature descriptor based on the histogram of oriented optical flow and the color histogram of oriented phase for rare event detection in a controlled environment. The proposed approach successfully models the usual event in video and allows the detection of the rare event. We reached a competitive result with an accuracy of 93 % compared to the prior work on the abnormal UMN dataset. In future work, we'll focus on the localisation of the rare event given a scene.

5. REFERENCES

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

[2] Ang Li, Zhenjiang Miao, Yigang Cen, and Qinghua Liang. Abnormal event detection based on sparse reconstruction in crowded scenes. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 1786–1790. IEEE, 2016.

[3] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009.

[4] Dinesh Singh and C Krishna Mohan. Graph formulation of video activities for abnormal activity recognition. *Pattern Recognition*, 65:265–272, 2017.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[6] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Manchester, UK, 1988.

[7] 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.

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

[9] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *computer vision and pattern recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE, 2009.

[10] Hussin K Ragb and Vijayan K Asari. Color and local phase based descriptor for human detection. In *Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS), 2016 IEEE National*, pages 68–73. IEEE, 2016.

[11] M Concetta Morrone and Robyn A Owens. Feature detection from local energy. *Pattern recognition letters*, 6(5):303–313, 1987.

[12] Peter Kovesi. Phase congruency detects corners and edges. In *The australian pattern recognition society conference: DICTA 2003*, 2003.

[13] UMN. Unusual crowd activity dataset of university of minnesota, department of computer science and engineering. In *http://mha.cs.umn.edu/movies/crowd-activity-all.avi*, 2006.

[14] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3449–3456. IEEE, 2011.

[15] Yinghuan Shi, Yang Gao, and Ruili Wang. Real-time abnormal event detection in complicated scenes. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3653–3656. IEEE, 2010.