A Proposal on Stampede Detection in Real Environments

Antonio Carlos Cob-Parro, Cristina Losada-Gutiérrez, Marta Marrón-Romera, Alfredo Gardel-Vicente, Ignacio Bravo-Muñoz and Mohammad Ibrahim Sarker

Abstract

It is a fact that the world population has grown in recent decades, as well as the number of social and tourism events, generating situations of agglomerations where different problems may lead to generate bottlenecks stampedes or falls, that can be a risk for people. Thus, the study of the behaviour of crowds is a relevant research topic. In this context, this paper presents and approach for real-time stampede detection from images, in low and medium crowd scenarios. The proposal is based on a feature vector extracted from the optical flow entropy, and this does not require the use of thresholds. Instead of that, it includes a a Stacking classifier, based on the union of a random forest with ten estimators and an support vector classifier, that works properly in the different analyzed scenarios. The proposal has been evaluated in UMN and PETS 2009 datasets and compared to other state-of-the-art proposals in terms of accuracy and computational cost. However, since the provided ground-truth was not accurate, a new manually-labelled ground-truth has been generated and make publicly available to the scientific community. The obtained results allows validating the proposal, outperforming the state-of-the-art methods both in terms of accuracy and computational cost in all the evaluated scenarios.

1. Introduction

It is a fact that the world population has grown in recent decades. In 1950s, there was a population of 2.5 billion people, while in the year 2020, there are approximately 7.7 billion people. This fact is more shocking when the increase in the last ten years is approximately 1 billion people, suggesting that the population is increasing in a non-linear way year by year. Some population experts suggest that for the next century, the population could exceed 11 billion people. This situation creates a scenario in which it is becoming relevant to deploy surveillance systems capable of detecting individual behaviour and group behaviour. The number of social and tourism events will grow, generating situations of agglomerations where different problems may lead to generate bottlenecks stampedes, falls and a long plethora of risk scenarios. The study of crowds' behaviour is a relevant research topic, to be able to control and protect people in the face of uncontrolled events.

IPIN 2021 WiP Proceedings, November 29 – December 2, 2021, Lloret de Mar, Spain

antonio.cob@edu.uah.es (A. C. Cob-Parro); cristina.losada@uah.es (C. Losada-Gutiérrez); marta.marron@uah.es (M. Marrón-Romera); alfredo.gardel@uah.es (A. Gardel-Vicente); ignacio.bravo@uah.es (I. Bravo-Muñoz); ibrahim.sarker@uah.es (M. I. Sarker)

0000-0001-8608-7351 (A. C. Cob-Parro); 0000-0001-9545-327X (C. Losada-Gutiérrez); 0000-0002-9421-8566 (M. Marrón-Romera); 0000-0001-7887-4689 (A. Gardel-Vicente); 0000-0002-6964-0036 (I. Bravo-Muñoz); 0000-0002-9589-294X (M. I. Sarker)

Crowd analysis, in general, can be approached holistically or by object-based methods. object based methods analyse crowds as sets of objects, studying people in a particular way [1], being these objects detected and followed in a particular way. Then in this type of systems, the amount and kind of tracking performed are analysed. The problem of these object-based methods is the accuracy because, in dense crowds, the identification of people is complicated. On the other hand, holistic approaches are based on identifying the crowd as a single unit [2, 3]. These methods are based on extracting the characteristics of the crowd to deduce its behaviour. Holistic methods have a reasonable accuracy rate in detecting anomalous behaviour.

In this work, we have focused on the detection of anomalous behavior in crowds. In particular, in the behaviour related to stampedes. Detection systems usually have two differentiated phases. The first phase is based on the representation of the event by a set of characteristics. There are many methods for the extraction of features, such as the study of social force [2], which is based on the measurement of the internal motivations of individuals to perform specific actions. The use of histograms of optical flow [4] to describe movement patrons or the use of histograms of movement direction [5] to describe direction patrons. The second phase consists of momentum detection from the previously extracted features. These classification models are usually characterized by having only two classes, either a stampede or not. There are different methods such as Support Vector Machine [6], neural replicator network [7], convolutional networks [8], etc.

The latest research in the detection of abnormal behaviour in crowds uses technologies such as context location and motion-rich patio-temporal volumes [9],temporal convolutional neural network pattern [10], generative adversarial networks [11], global event influence model [12]. In this work, we have implemented a system that draws from both the most modern and the most classical approaches, Being a system based on the use of the value of the magnitude of the optical flow and from there extract the entropy to generate a series of descriptors that are used in a machine learning model for the detection of the anomaly.

Considering the previous research work, the main contribution of this paper is to deploy a robust and reliable system capable to detect stampedes in real-time. In addition, the algorithm to detect events' peak does not require a threshold. Additionally, we have manually labelled the UMN [2] and PETS 2009 [13] datasets to quantitatively evaluate the stampedes detection, and the result of the annotation has been made available to the scientific community [14].

The rest of the paper is organized as follows: section 2 describes the proposal for stampede detection, them section 3 presents the annotation procedure for the UMN and PETS-2009 datasets. Next, the main experimental results are shown in section 4.1. Finally, section 5 describes the main conclusions and future work.

2. Proposal for stampede detection

We have focused our efforts on identifying stampedes in the first two types of scenarios, the low and intermediate density crowds. As other previous approaches, such as the one described in [15], the proposal presented in this work is based on analysing the entropy obtained from the scene optical flow. It is obtained a dense optical flow by using the method of Farneback [16], instead of a punctual one as in the case of using the method of Lucas-Kanade [17].

Besides, the system works in real time. Moreover, it does not require a threshold (that must be modified for each dataset). Instead, we have extracted a feature vector from the entropy, and designed a machine learning model based on stacking classifier for stampede detection that uses a set of features generated from the entropy signal extracted from the optical flow.

The figure 1 shows a general block diagram with each of the stages of the system. Below, there are described in detail each of these stages.

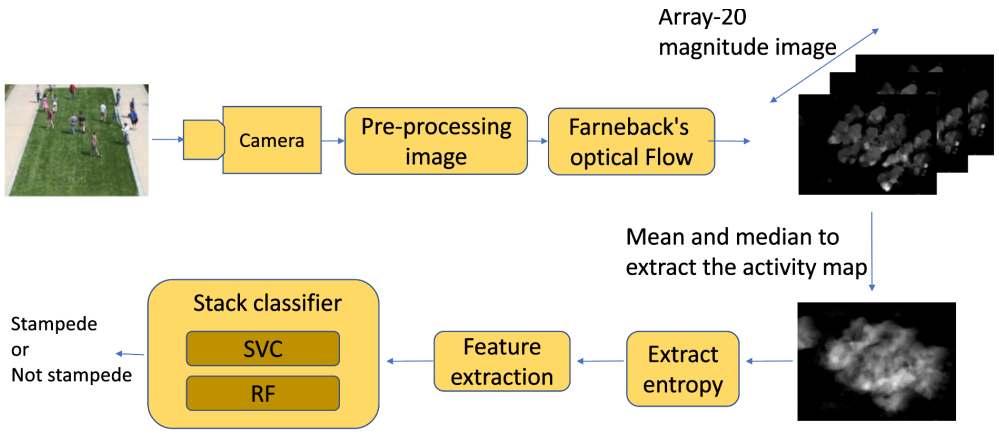


Figure 1: General block diagram of the system.

2.1. Optical flow computation

The system extracts the entropy using the optical flow of the image. For this purpose, we have used the dense optical flow of Farneback instead of the point optical flow of Lucas-Kanade. Because we wanted a system that would analyse the whole image.This system obtains the movement variations of all pixels between frames. Unlike Lucas-Kanade, which is based on the study of the variations of a specific set of pixels.

The Farneback's method is based on the estimation of the movement of the pixels between the actual and previous frames. From that movement, there are generated the displacement vectors, that are then used to study the movement variations in the image. This type of optical flow analysis presents a higher accuracy than those algorithms based on sparse optical flow.

Thus, Farneback's dense optical flow extraction is based on expanding the position of the pixel coordinates by polynomial expansion using the neighbourhood information of each pixel in the image. The original coordinates (u_0, v_0) are independent variables, and the new coordinates (u, v) are polynomials of dependent variables. The amount of motion (du, dv) of the pixel in u and v directions are determined by substituting the coordinates into them. A displacement vector is obtained for each pixel between two frames.

2.2. Feature extraction

The feature vector used for detecting stampedes is based on the entropy. To obtain the entropy, first, the image is pre-processed by blurring and grayscaling. Then the magnitude value of the optical flow is extracted using the current and previous frames These magnitude values form a matrix that correspond to the movement variation of each pixel. Then, these matrices are grouped in batches of 20 frames, and the mean of all the magnitudes is extracted for each pixel. Then the median is calculated, and the result is called activity map. From the activity map, the entropy for that frame can be extracted. To extract the entropy it is used equation 1, where x is the number of separate symbols, p_i is the frequency of the each pixel in the image and n is the actual frame.

$$Entropy(n) = \sum_{i=1}^{x} p_i log_2(p_i) \tag{1}$$

In previous works, such as [15], the authors use two thresholds to determine if there was a stampede or not. The first one was based on the entropy value, and the second one was based on the temporal occupancy variation (TOV) value between frames. For this work, the TOV has not been employed using only the entropy value, from what there is obtained a feature vector for the stacking classifier.

After obtaining the entropy, to determine if there is a stampede or not, we extract a set of features that are next classified using a stacking classifier. These features include the mean and standard deviation of the entropy. The mean is computed using a sliding window of 20 frames as shown in the equation 2, being k the size of the sliding windows . Thus the proposal requires a minimum of 20 frames for detecting a stampede, but it smooths the signal and removes the high frequency noise.

$$\mu(n) = \frac{1}{k} \sum_{i=n-k+1}^{n} Entropy(i) \tag{2}$$

The standard deviation of the entropy descriptor is used to detect the fast changes in the signal. A considerable variation generates a significant change of value in this descriptor. The equation 3 shows the mathematical definition used for the extraction of the descriptor, where x_i are the current entropy values and μ the mean entropy value.

$$\sigma(n) = \sqrt{\frac{\sum_{i=n-k+1}^{n} (Entropy(i) - \mu(i))^2}{k}} \tag{3}$$

The third feature is the distance generated as the difference between the mean plus standard deviation and the mean minus standard deviation. The standard deviation is multiplied because this value (L) is smaller than the other signal. In the equation 4 is defined mathematically.

$$Distance(n) = (\mu(n) + L * \sigma(n)) - (\mu(n) - L * \sigma(n)) \tag{4}$$

An example of the signals explained below are shown in the figure 2. The peak of the entropy (right signal) reflects the moment when the stampede happens. The features explained below

are drawn in the left picture. In blue, it is shown the entropy mean in yellow, the standard deviation and in green with shade the distance.

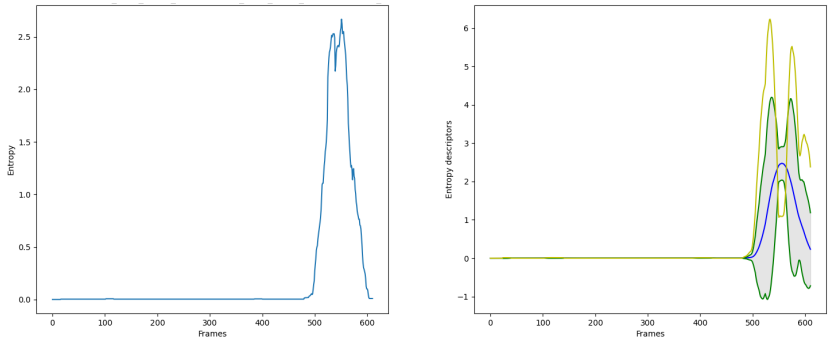


Figure 2: The graph on the right shows the entropy value, with the peak at the time of the stampede. The left plot shows the mean value in blue, the standard deviation in yellow and the distance generated in green with shading.

2.3. Classification

The stacking classifier model (figure 3) has been used, which is a combination of an support vector classifier (SVC), random forest (RF). It is an ensemble learning technique for combining multiple classification models through a meta-classifier. The individual classification models are trained based on the complete training set; then, the meta-classifier is tuned based on the results-meta-features of the individual classification models in the ensemble. The meta-classifier can be trained on the predicted class labels or the ensemble probabilities. Figure 3 shows the basic structure of a stacking classifier.

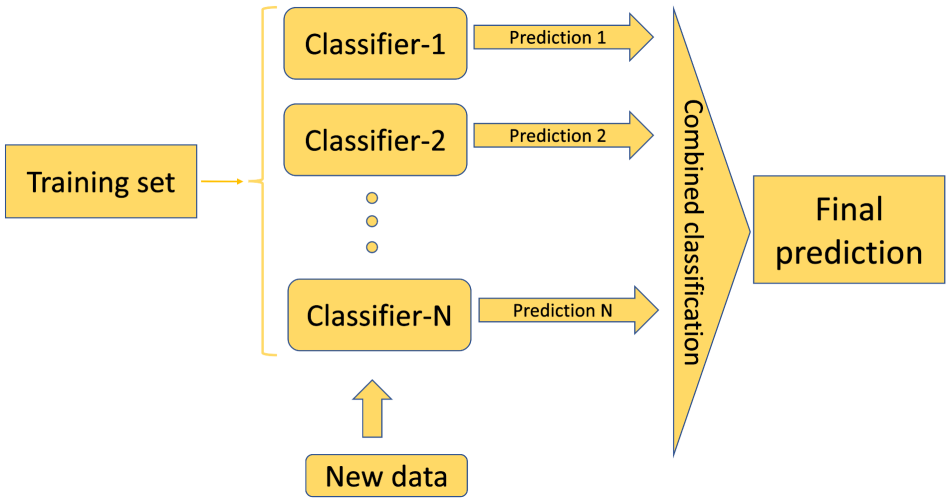


Figure 3: Example of stacking classifier structure

The stacking classifier used in this work is based on the union of a random forest with ten estimators and an SVC, that has been used to classify if there is a stampede or not. The training was performed with half of the UMN videos and half of the PETS videos, being the other half used for model testing. This 50:50 margin is used for two reasons: first, a stacking classifier does not need a large amount of information to train and second, the number of cases in which there is a stampede in the videos is much smaller than in those in which there is not, so more videos have been used for testing to achieve a more reliable result. In order to obtain the best results in terms of measuring the predictive quality of the models, a k-fold of the training process has been performed with a value of k equal to 10. The value of k is 10, because the datasets have a small number of frames and a small number of labels, so the division into 10 slots provides enough information in each slot for the training to succeed.

3. Stampede Annotation

To evaluate the system, we have used two datasets, UMN and PETS2009, that have been widely used in other works for stampede detection. These datasets include several videos with stampedes. The UMN dataset include the ground-truth, however, when analysing the videos, it can be seen that there is a delay between the beginning of the stampede and the frame in which it is labelled. Furthermore, PETS dataset does not include ground-truth information for stampede events.

The characteristics of the datasets are shown in the table 1. Both datasets have similar types of stampedes, in which people run either in the same direction or spread out. The way of recording the videos is the same for both datasets by high-angle shot. The lighting is constant in all videos but one environment of the UMN dataset. The big difference between the two datasets is the number of people in UMN is not more than 20 people in any video, while in PETS there are more than 30 people per video.

Table 1
Main characteristics of the analysed datasets

Dataset	Scenario	Resolution	Illumination	#people	#videos/frames
UMN	Lawn	240×320	Constant	15	2/1433
	Plaza			12	3/4038
	Indoor		Variable	10	6/2031
PETS 2009	Street-1	576×768	Constant	41	4/1812
	Street-2			42	4/1060

Due to the lack of an accurate ground-truth, we have analysed the two dataset, and hand-labelled the information. To label the videos, it is necessary to define two moments in a stampede, the beginning and the end of the stampede. To label the initial moment we have considered that more than four people in the image are already prepared to run or moving. The end of the stampede is defined when the people in the image stop running and start walking or when less than three people are running on the screen. By means of these guidelines, we have defined a ground-truth made by hand analysing frame by frame and indicating the moment when each stampede starts and ends. Figure 4 shows a ground-truth scheme indicating in green

the moments of calm and in red the times where it is considered that there is a stampede. In addition, in Figure 4, the frames where the stampede begins and ends and the total number of frames of each video are shown.

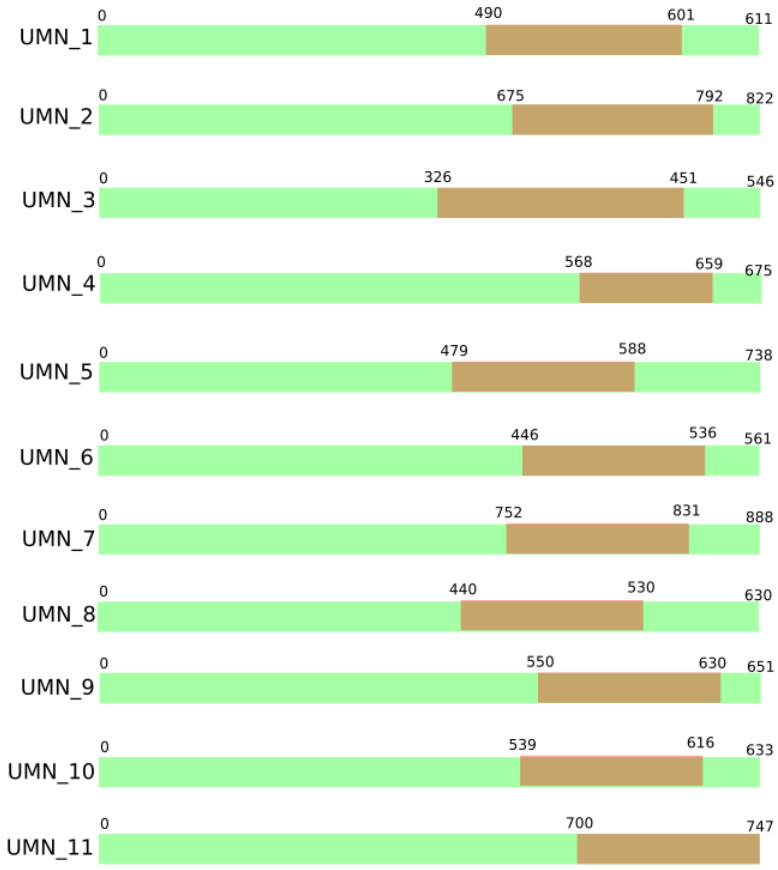


Figure 4: Ground-truth UMN

Figure 5 shows a scheme of those PETS videos that contain stampedes. Note that the PETS videos have the same video but seen from different camera points. For this reason, the beginnings and endings of the stampede are the same for all these recordings.

As mentioned in previous sections, this more accurate annotation has been made publicly available [14].

4. Experimental results

4.1. Experimental Set-up

To evaluate the system performance, we have analysed two key parameters. The first one is the speed of execution of the system, and the second one the accuracy to detect stampedes.

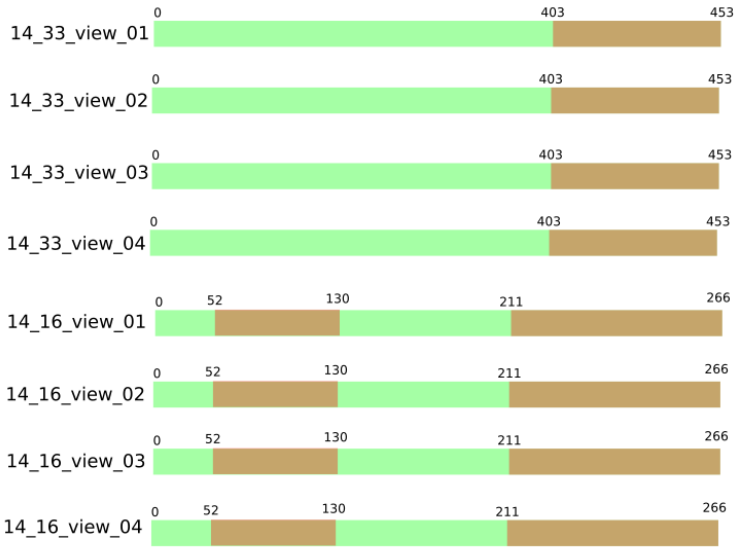


Figure 5: Ground-truth PETS

The computational cost has been evaluated for the original system executed on a CPU (Intel Core i7-9700K CPU @ 3.60GHz x 8). It is important to note that the OpenCV version is 4.5.1, and the programming language is Python version 3.7.

Regarding the accuracy, as it has been stated before, two datasets (UMN and PETS) have been used, which have been compared with the results obtained in the work [15] . The UMN dataset includes a total of 7502 frames with a resolution of 240x320 at 30 fps. The dataset has been divided into three scenarios, two outdoor and one indoor. The PETS dataset includes a single outdoor scenario with a total of 2872 frames at a resolution of 768x576 at 30 fps.

To evaluate the system we have decided to use the receiver operating characteristic (ROC) and area under the curve (AUC) metrics. A ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve represents the true positives rate (TPR) versus false positives rate (FPR) at different classification thresholds, whereas the AUC measures the entire two-dimensional area below the total ROC curve. The AUC provides an aggregate measure of performance for all possible classification thresholds.

4.2. Stampede detection

This section shows the results of stampede detection in different environments, both indoors and outdoors corresponding to the previously described UMN and PETS 2009 datasets.

As explained previously, entropy, mean, standard deviation and distance combination among entropy, mean and standard deviation are extracted of each frame. These values are analysed by a stacking classifier that determine the value of the activity.

Figure 6 shows an example of UMN where it can be clearly observed where the stampede starts, marked with an arrow.

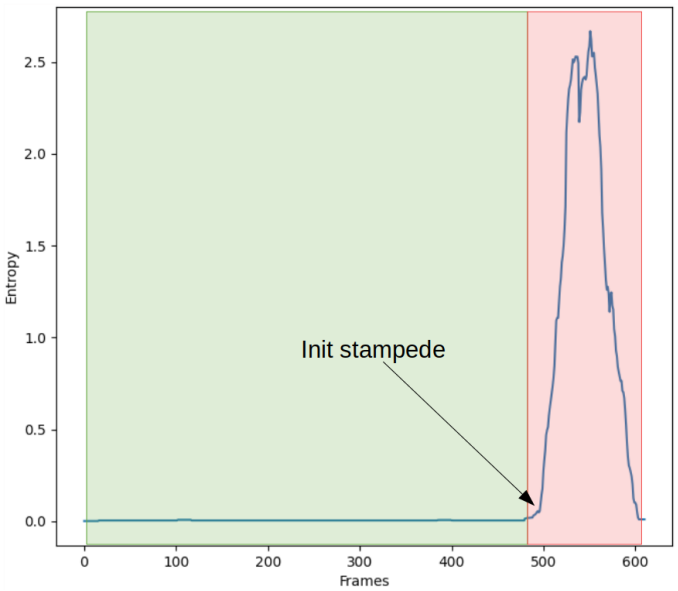


Figure 6: Green shading indicates the frames in which there is no stampede while red shading indicates the frames in which there is a stampede. The arrow indicates the moment when the stampede starts.

Due to the frame characteristics of each dataset, it is necessary to normalise the magnitude values from the optical flow. This normalisation is performed so that the entropy values are limited between two close values. This allows the dispersion between the different samples to be smaller and helps the classifier to detect better. This normalisation is adjusted to the number of pixels in the image by dividing the magnitude given by the optical flow by a constant that will be larger as the image size increases. Figure 7 shows an example in which the maximum value of the entropy is similar to the maximum value in UMN (figure 7).

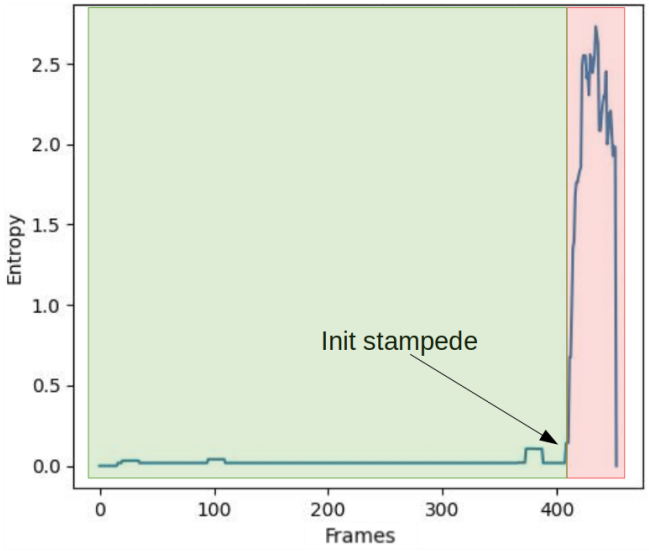
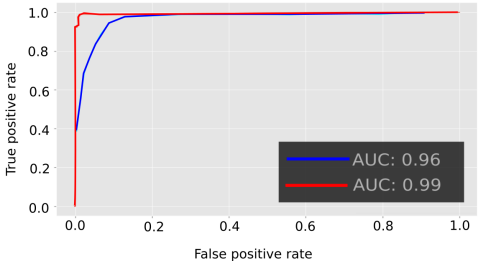


Figure 7: Green shading indicates the frames in which there is no stampede while red shading indicates the frames in which there is a stampede. The arrow indicates the moment when the stampede starts.

To observe the improvement of the system performance, we compare the ROC curves of our system with those provided in the baseline [15]. It is worth note that the ground-truth used by this paper [15] defines the beginning and the end of the stampede too late in relation to the actual times of the videos. For this reason we have manually-labelled the videos, using the above mentioned definition of stampede. Thus, our ground-truth is more accurate and realistic than the one used by [15]

The first scenario is the UMN lawn (figure 8), in which no new actors enter the scene. Figure 8a shows an example of a frame corresponding to this scenario, whereas figure 8b compares the ROCs and the AUC for our system (in red) and for the baseline [15] (in blue). Comparing the ROCs, it can be seen that the results are similar, although some improvement is obtained in our system.

(a) Lawn scenario.

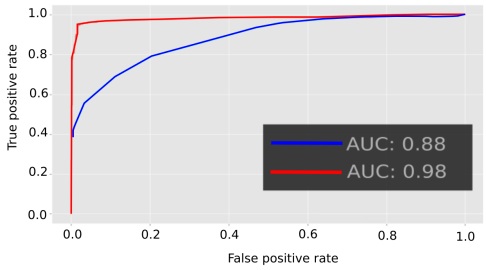


(b) ROC lawn scenario: red our system and in blue the system [15]

Figure 8: Results for Lawn UMN dataset.

The indoor scenario (figure 9) is characterized for being a recording of a hall, where unusual activity is observed, such as the entrance and exit of people, as well as much more abrupt light changes than in the other two scenarios. The indoor scenario is where the most significant improvement is observed, having important increase in the area under the curve (AUC), being the colour red our system and in the colour blue, the [15] system (figure 9b).

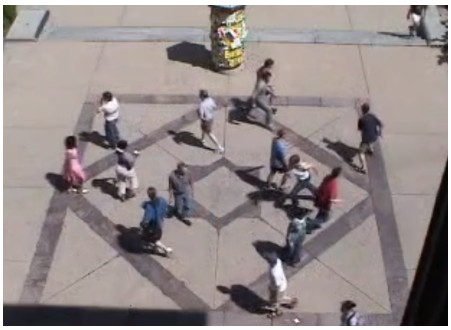
(a) Indoor scenario.



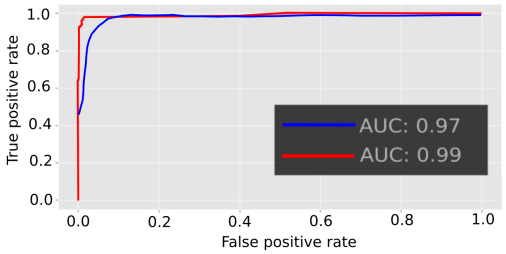
(b) ROC indoor scenario: red our system and in blue the system [15]

Figure 9: Results for Indoor UMN.

Finally, figure 10 shows the last scenario of the UMN dataset. It is recorded in a plaza, where the people is walked. This scenario (figure 10a) has the lighting, the size of the people, recording environment similar to the lawn scenario. For this reason, the values obtained in the ROC curve (figure 10b) are practically equal to the lawn scenario. As it is shown in the figure the AUC obtained by our system (colour red) and the system [15] (colour blue) is very similar, having minor improvements.



(a) Plaza scenario.



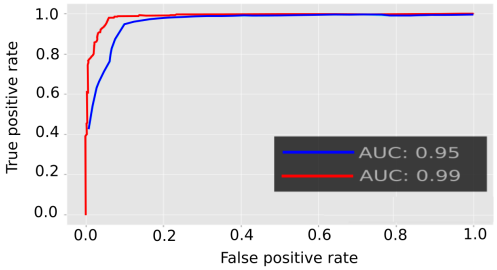
(b) ROC plaza scenario: red our system and in blue the system [15]

Figure 10: Results for Plaza UMN.

Regarding the PETS 2009 dataset (figure 11), we find a view similar to the lawn and plaza scenaios in the UMN dataset. The frames show a street crossing where a group of people run in a stampede (figure 11a). The AUC obtained with our method provides a better operation than [15].

As it has been explained before, our proposal is able to perform in real time, being faster

(a) PETS scenario.



(b) ROC PETS 2009: red our system and in blue the system [15]

Figure 11: Results for PETS 2009.

than [15]. In the table 2 is shown a comparison between our model and the [15] model. In the table there are compared both frame per second processed by each proposal and the AUC value. It can be seen that our model obtains better results in every field, both in computational cost (more than 2x speed up) and AUC (with and improvement from +3% to +10%, depending on the dataset).

Table 2
Comparison of our system with [15]

Dataset	Scenario	Our proposal		Pennisi et al. [15]	
		FPS	AUC	FPS	AUC
UMN	Lawn	54	0.99	20	0.96
	Plaza	54	0.99	20	0.97
	Indoor	53	0.98	20	0.88
PETS 2009	Street-1	22	0.99	11	0.95
	Street-2	23	0.99	11	0.96

5. Conclusions

In this work, we have proposed an approach for real-time stampede detection in low and medium density crowd scenarios. The proposal is based on a feature vector extracted from the optical flow entropy, and this does not require the use of thresholds, since there has been replaced by a stacking classifier, based on the union of a random forest with ten estimators and an SVC, that works properly in the different analysed scenarios.

We have selected and extracted features for stampede detection. These features are extracted through the entropy information given by the optical flow. These features are suitable descriptors for the detection of the stampede beginning.

The proposal has been evaluated exhaustively in UMN and PETS 2009 datasets, that has been widely used in stampede detection works, and compared to other state-of-the-art proposals in

terms of accuracy and computational cost. For this evaluation, the UMN and PETS (stampedes) datasets have been hand labelled, adjusting in a more precise way the start and end frames of the stampedes. For this reason, our ground-truth is tighter and more precise than the existing ones.

As conclusion, the system offers an improvement of the results respect to previous state-of-the-art works both in precision and computational cost. In addition the system does not depends on a threshold to run correctly and it is able to adapt to different environments. This has been possible by replacing thresholds with machine learning models that are able to generalise a solution independently of the environment.

As future lines of work, we plan to develop this system but running on dedicated vision processing hardware such as GPU or VPU. This will allow a faster response in the processing and add another layer of parallelisation to the system. Regarding the classification method, it is proposed to use a dense neural network instead of a classical machine learning model and compare how the system behaves in deep learning models. In addition, it is intended to test the system on more datasets in order to evaluate more environments. Evaluating in a more extensive way the generality of the designed system.

Acknowledgments

The authors would like to thank the GEINTRA research group (geintra-uah.org) for their support and background work. This research has received funding from the European Union's Horizon 2020 Research and Innovation Programme under PALAEMON project (Grant Agreement nº 814962), and by the Spanish Ministry of Economy and Competitiveness under project HEIMDAL-UAH (TIN2016-75982-C2-1-R).

References

[1] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, T. Yu, Unified crowd segmentation, in: European conference on computer vision, Springer, 2008, p. 691–704.

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

[3] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, p. 1–6.

[4] A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors, IEEE transactions on pattern analysis and machine intelligence 30 (2008) 555–560.

[5] H. M. Dee, A. Caplier, Crowd behaviour analysis using histograms of motion direction, in: 2010 IEEE International Conference on Image Processing, IEEE, 2010, p. 1545–1548.

[6] C. Direkoglu, M. Sah, N. E. O'Connor, Abnormal crowd behavior detection using novel optical flow-based features, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2017, p. 1–6.

[7] S. Hawkins, H. He, G. Williams, R. Baxter, Outlier detection using replicator neural networks, in: International Conference on Data Warehousing and Knowledge Discovery, Springer, 2002, p. 170–180.
[8] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, Z. Zhang, Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes, Signal Processing: Image Communication 47 (2016) 358–368.
[9] N. Patil, P. K. Biswas, Global abnormal events detection in crowded scenes using context location and motion-rich spatio-temporal volumes, IET Image Processing 12 (2018) 596–604.
[10] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, N. 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), IEEE, 2018, p. 1689–1698.
[11] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, Abnormal event detection in videos using generative adversarial nets, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, p. 1577–1581.
[12] L. Pan, H. Zhou, Y. Liu, M. Wang, Global event influence model: integrating crowd motion and social psychology for global anomaly detection in dense crowds, Journal of Electronic Imaging 28 (2019) 023033.
[13] J. Ferryman, A. Shahrokni, Pets2009: Dataset and challenge, in: 2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance, IEEE, 2009, p. 1–6.
[14] A. C. Cob-Parro, Umn repository with the updated grountruth, 2021. Https://github.com/CarlosCobParro/UMN-groundtruth-update.
[15] A. Pennisi, D. D. Bloisi, L. Iocchi, Online real-time crowd behavior detection in video sequences, Computer Vision and Image Understanding 144 (2016) 166–176.
[16] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Scandinavian conference on Image analysis, Springer, 2003, p. 363–370.
[17] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'81, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1981, p. 674–679.