Deep learning-based anomalous object detection system for panoramic cameras managed by a Jetson TX2 board

Jesús Benito-Picazo, Enrique Domínguez, Esteban J. Palomo, Gonzalo Ramos-Jiménez, Ezequiel López-Rubio
Department of Computer Languages and Computer Science
University of Málaga
Málaga, Spain
{jpicazo, enriqued, ejpalomo, ramos, ezeqlr}@lcc.uma.es

Abstract—**Social conflicts appearing in the media are increasing public awareness about security issues, resulting in a higher demand of more exhaustive environment monitoring methods. Automatic video surveillance systems are a powerful assistance to public and private security agents. Since the arrival of deep learning, object detection and classification systems have experienced a large improvement in both accuracy and versatility. However, deep learning-based object detection and classification systems often require expensive GPU-based hardware to work properly. This paper presents a novel deep learning-based foreground anomalous object detection system for video streams supplied by panoramic cameras, specially designed to build power efficient video surveillance systems. The system optimises the process of searching for anomalous objects through a new potential detection generator managed by three different multivariant homoscedastic distributions. Experimental results obtained after its deployment in a Jetson TX2 board attest the good performance of the system, postulating it as a solvent approach to power saving video surveillance systems.**

Index Terms—**Deep learning, video surveillance, panoramic cameras, power saving**

I. INTRODUCTION

Impelled by the abundance of social conflicts appearing in the media, citizens of modern societies demand higher security rates requiring effective security systems. One of the multiple uses of computer vision is the construction of automatic video surveillance systems. The different challenges faced by the researchers when building these kinds of systems are addressed in [1], where the author provides a survey on vision-based human action recognition. W^4 system detailed in [2] presents a real-time video surveillance system capable of performing the detection and tracking of people in video streams coming from certain cameras placed in outdoor environments. The system employs a combination of shape analysis and tracking to locate people and the different parts of their anatomy in order to create models of people's appearance so their activities can be monitored through different types of interactions such as occlusions. These video surveillance systems usually are equipped with person identification modules such as the one presented in [3], where the authors propose a person re-identification system based on an effective feature representation called Local Maximal Occurrence (LOMO), and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA). Along the same line goes the work illustrated in [4], describing a face identification system using eigenface recognisers and Intel's Haar cascades, intended to monitor the identity of the people appearing in video streams filmed by cameras installed at the entry points of certain facilities.

Sometimes, people are not the only objective of a video surveillance system. In fact, sometimes is important that video surveillance systems are prepared to alert from the presence of certain object in the scene that is under vigilance. As it is reflected in the works referred to above, in order to have a correct understanding of the scene that is being monitored, the system must be able to identify the different objects swarming in the scene, and for that, a background modelling algorithm is needed. A good example of these algorithms can be found in [5] and [6]. These algorithms continued their evolution until nowadays as it is attested in [7], where deep learning-based background subtraction model based is presented for flexible foreground segmentation.

Deep learning is a machine learning technique, which stands out in terms of accuracy and performance. This technique uses Deep Neural Networks (DNN) to learn a series of features from the input data, and is specially useful in the field of computer vision, where it has several applications, as it is illustrated in the work presented in [8]. In this work we can find a review of various recent uses of convolutional neural networks (CNNs) to solve inverse problems in imaging such as deconvolution, denoising, medical image reconstruction and superresolution. For instance, in [9], the authors present a deep neural network-based denoising filter and a practical method of deep neural network training with noisy patterns to improve its performance against noisy test patterns. All of these techniques have proven their effectiveness in computer vision-based automated video surveillance systems. A good example of can be found in [10] and [11], where several types of deep convolutional networks are proposed to be used in the construction of crack detection systems on asphalt pavement surfaces. [12] Also presents a civil engineering deep learning-based inspection system. In this case, the system proposes a post-disaster inspection of the reinforced concrete bridge structures.

The choice of an appropriate camera is critical in the design of any video surveillance system at this device will be in charge of supplying the image input of the cited system. Pan-tilt-zoom (PTZ) cameras are powerful, yet affordable devices with a high acceptance in the construction of video surveillance systems because of their versatility and motion capabilities. For example, in [13] the authors present the development of a novel salient motion detection method mainly for non-stationary videos captured by PTZ cameras. In [14] we can find a background subtraction algorithm designed for PTZ cameras that works without the need for explicitly registering images. Video surveillance systems often operate using PTZ camera networks as it is documented in [15] and [16].

Due to their restricted field of view, PTZ cameras present some limitations that can be overcome by using 360° panoramic image capture devices. These devices supply different sorts of 360° spherical images offering the possibility of covering most of the monitored area on each video frame as it is described in works like [17] and [18]. This sort of systems have been in continuous improvement by scientists and engineers, reaching high performance levels in object detection and tracking as it is described in [19].

Finally, computer vision in general and more specifically deep learning-based video surveillance systems, usually have heavy computational requirements that in many occasions only can be supplied by expensive, high power consuming devices, severely limiting their autonomy and versatility. Thus the necessity for low power automatic video surveillance systems with an acceptable performance. Along this line, some works have been developed in the last years. In [20] the authors describe the design of a computationally efficient and low power demanding system for detecting moving objects which can be deployed into unmanned aerial vehicles (UAV). The work presented by the authors in [21] propose a tracking pipeline for fixed smart cameras that reaches real-time processing on a low-cost embedded smart camera composed of a Raspberry-Pi board and a RaspiCam camera. Finally, the work described in [22] describes a deep learning-based automatic video surveillance system for panoramic cameras, specifically designed to be deployed on cheap and power efficient hardware devices, such as a Raspberry-Pi microcomputer.

In this work we describe an improved deep learning-based automatic video surveillance system for panoramic cameras, specifically optimized to be deployed on a Jetson TX2 board. This system relies on a novel potential detection generator based on three multivariate homoscedastic distributions and a MobileNet [23] Deep Convolutional network.

The rest of paper is organized as follows. Section II presents the mathematical model of our proposal. Section III presents the architecture of the system. In Section IV, our experimental results are provided. Finally, Section V concludes this paper.

II. METHODOLOGY

The environments that are considered in this work contain anomalous objects, which means that they do not belong to the most frequently found classes in the scene. An alarm must be activated in case that an anomalous object is found.

In this section, the proposed methodology to detect anomalous objects is detailed. This method is an extension of our previous work in [22].

The basis of our approach is the analysis of the most recent detections. A set is maintained with such detections, i.e. the active detections. The set comprises the objects that have been detected recently by the surveillance camera. We define a detection as a vector of four real numbers (π_i, x_1, x_2, x_3) where:

- π_i is the *a priori* likelihood of the object.
- (x_1, x_2) are the coordinates of the object (vertical and horizontal), which refer to the panoramic coordinate system defined by the video camera.
- x_3 is the number of pixels that comprise the length of the bounding box that surrounds the detected object.

Also, we define a forgetting rate α that is employed to update the a priori probability π_i. If a detection is lost from the sight, then the associated detection goes non active.

If we note $\mathbf{x} = (x_1, x_2, x_3)$, then we can express the possible range for $\mathbf{x}$ this way:

$$\mathcal{V} = [1, N_{rows}] \times [1, N_{cols}] \times [S_{min}, S_{max}] \subset \mathbb{R}^3 \qquad (1)$$

where $N_{rows} \times N_{cols}$ is the size in pixels of the acquired video frame, so that the possible sizes of the bounding boxes are limited by S_{min} and S_{max}, respectively.

Then the following probabilistic model is employed to estimate the possible positions of the objects:

$$p(\mathbf{y}) = qU_{\mathcal{V}}(\mathbf{y}) + (1-q)\frac{1}{M}\sum_{i=1}^{M} \pi_i K(\mathbf{y}, \mathbf{x}_i, \sigma) \qquad (2)$$

where $U_{\mathcal{V}}(\mathbf{y})$ is the uniform probability distribution on $\mathcal{V}$, $K(\mathbf{y}, \boldsymbol{\mu}, \sigma)$ stands for a multivariate distribution having a mean vector $\boldsymbol{\mu}$ and a constant spread parameter σ, M is the number of detections which are active, $q \in (0,1)$ is a tunable mixing parameter and σ is the spread parameter of the multivariate distribution.

The main novelty introduced in the mathematical model presented in this work with respect to the mathematical model in [22], is that the a priori probability π_i of the i-th detection is proportional to the probability that the i-th detection belongs to the most likely class associated to that detection.

We have selected three multivariate distributions for our probabilistic model. Such distributions are: Gaussian, Student-t and triangular. Their equations are shown in Table I, so that $\|\cdot\|$ denotes the Euclidean norm of a vector, while ν stands for the degrees of freedom of the Student-t distribution. It must be highlighted that the Gaussian and Student-t options include

$K_{Gaussian}\left(\mathbf{y},\boldsymbol{\mu},\sigma\right)=\left(2\pi\right)^{-\frac{3}{2}}\sigma^{-3}\exp\left(-\frac{1}{2\sigma^{2}}\left\Vert \mathbf{y}-\boldsymbol{\mu}\right\Vert ^{2}\right)$	(3)
$K_{Student}\left(\mathbf{y},\boldsymbol{\mu},\sigma\right)=\frac{\Gamma\left(\frac{\nu+3}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{3}{2}}\pi^{\frac{3}{2}}\sigma^{3}}\left(1+\frac{1}{\nu\sigma^{2}}\left\Vert \mathbf{y}-\boldsymbol{\mu}\right\Vert ^{2}\right)^{-\frac{\nu+3}{2}}$	(4)
$K_{Triangular}\left(\mathbf{y},\boldsymbol{\mu},\sigma\right)=\prod_{j=1}^{3}k_{Triangular,j}\left(y_{j},\mu_{j},\sigma\right)$	(5)
$k_{Triangular,j}\left(y_{j},\mu_{j},\sigma\right)=\begin{cases}0 & \text{for } y_{j}<\mu_{j}-\sigma\\ \frac{y_{j}-\mu_{j}+\sigma}{\sigma^{2}} & \text{for } \mu_{j}-\sigma\leq y_{j}<\mu_{j}\\ \frac{1}{\sigma} & \text{for } y_{j}=\mu_{j}\\ \frac{\mu_{j}+\sigma-y_{j}}{\sigma^{2}} & \text{for } \mu_{j}<y_{j}\leq\mu_{j}+\sigma\\ 0 & \text{for } y_{j}>\mu_{j}+\sigma \end{cases}$	(6)

TABLE I
GAUSSIAN, STUDENT-T, AND TRIANGULAR MULTIVARIATE DISTRIBUTIONS.

a spread parameter σ which coincides with the standard deviation.

The aim of our proposal is to focus the search on the regions of the incoming video where objects have been detected recently. This corresponds to the multivariate distribution of the probabilistic model. Nevertheless, the rest of the video frame must also be queried. This is accommodated by the uniform distribution of the probabilistic model.

The proposed algorithm to find objects of an anomalous class within a panoramic video recording is given next:

1) Setup the set of current detections $\mathcal{A}$ to the empty set.
2) Acquire a new video frame from the surveillance hardware.
3) Update the a priori probabilities π_i of the set of active detections by application of the forgetting rate α. Any objects that have been lost of sight because they are outside $\mathcal{V}$ are erased since they have become inactive.
4) Choose at random a set of M samples from the multivariate distribution (2). Find the bounding box corresponding to every sample, and change its size so that it matches the required size for the deep neural network. After that, pass the bounding box to the deep network. If the output of the deep network reveals that there has been a detection, then the associated sample is inserted into $\mathcal{A}$, and the sample is annotated with the probability that the object is actually there, i.e. the reliability of the detection.
5) Go to step 2.

Next, a concrete framework to implement the above detailed procedure is given, so that a cheap microcontroller based surveillance system can be obtained.

III. SYSTEM ARCHITECTURE

As it was mentioned in Section I, the main piece of a deep learning-based video surveillance system consists of a powerful foreground object detection and identification engine. However, object detection and classification from images filmed by a high resolution camera usually requires high amounts of computing power, specially if it incorporates deep learning techniques. But at the same time, it looks reasonable that a video surveillance system must have a fast response in order to be effective. Consequently, these systems are often supported by high performance GPU-based hardware that has also high electric power demands.

Nevertheless, there are occasions where is very difficult or just impossible to supply a general power connection to those systems, specially in the case of systems with a high degree of autonomy, either because they are constantly in motion or because they are intended to be installed in natural environments where no general power connections are available.

This fact motivated the authors of this work to tackle the design and implementation of a deep learning video surveillance system capable of detecting anomalous foreground objects but at a small fraction of the electrical power needed by a conventional deep learning-based system. Traditional deep learning-based object detection and classification models, such as Faster-RCNN, often rely on performing a massive number of inference passes to all the surface of the frame, but this is unfeasible to be done when designing a system that is meant to be deployed in a low profile hardware device. Thus, we decided to design a system that presents an architecture that features a potential detection generator that will test just a limited number of areas whose position and size will be designated by a mixture of a random distribution and three multivariate homoscedastic distributions. These areas of the frame will be fed to a convolutional neural network who is in charge of identifying the possible anomalous objects enclosed in those areas.

Hence, the architecture of the object detection and identification system detailed in this work will consist of two well differentiated parts: A software architecture that will implement the object detection and classification algorithm, and a hardware architecture integrated by a panoramic camera and a Jetson TX2 board where the algorithm is meant to be deployed.

A. Software architecture

The software architecture of the system is shown in Figure 1. In this figure it can be observed that it consists of two different modules: The first one is the potential detection generator, a program developed in Python that acquires one frame from the video stream supplied by a panoramic spherical camera and scatters certain number of random windows, namely potential detections, in the frame that will be delimited and situated by one of the three mixtures of random and homoscedastic multivariant distributions presented in Section II. This module also is in charge of updating the position and size of the existing detections according to the equations of the model proposed in Section II.

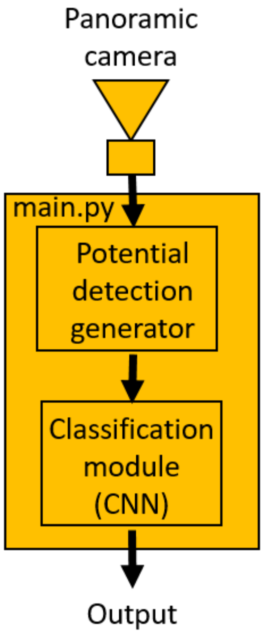

Fig. 1. Schematic diagram of the software architecture

The second module will try to identify the object enclosed in each potential detection by feeding them to a convolutional neural network.

Convolutional neural networks often require large amounts of computing power to work in a timely manner. Thus, the choice of the convolutional neural network to be used is critical in a low power consuming system with limited computing power. Therefore, the chosen CNN must present a balance between accuracy and speed. These reasons led the authors of this work to select the *Pytorch* framework implementation of the MobileNet [23] network properly trained using the *Pascal VOC2012* dataset [24].

B. Hardware architecture

Hardware choice is also a critical matter when designing a power efficient video surveillance system insofar as it must have a limited power consumption and at the same time it must be capable of working as fast as possible, because these kinds of systems will be more useful as their operating speed approaches to real-time. Therefore, we decided to use a Jetson TX2 board to deploy our system as it features a cuda-capable 256 core NVIDIA Pascal GPU, specially useful for deep learning tasks, with a power consumption of 7.5 watts.

IV. EXPERIMENTAL RESULTS

In order to properly test the anomalous object detection system for panoramic cameras presented in this work we have developed a benchmark program including the two modules referred to in Section III. The program uses a panoramic video from a well-known 360° videos dataset hosted by the *Virtual Human Interaction Lab* at the University of Stanford [25] to simulate the video stream provided by a 360° camera. This is a very convenient configuration in order to make more reproducible experiments. Thus, the program works as follows (Figure 2):

First, a panoramic video frame is supplied to the program. Second, the potential detection generator module will use one of the homoscedastic multivariate distributions to generate the frame coordinates for a certain number of windows enclosing the areas that will be examined in search for anomalous objects. These will be our "potential detections". After the potential detection generation phase has been completed, the program feeds the areas enclosed by these potential detections to the identification module, where a convolutional neural network will identify the possible anomalous objects present in that area of the frame. Next, every potential detection containing any object will be included in the detections set referred as $\mathcal{A}$ in the model described in Section II. Finally every new detection is compared to the list of anomalous objects, and if the program finds any coincidence, it raises an anomalous object detection alert informing about the coordinates of the upper left corner and side sizes of the bounding box that surrounds the anomalous object detected. It also informs about its category and the accuracy of the detection.

With the objective of exhaustively testing the system presented in this work, a series of experiments has been performed. These experiments included the separate utilisation, one by one, of the three probability distributions in which the potential detection generator relies and that were described in Section II. In order to create the same conditions for all the experiments we have used an unaltered 360° video from the public dataset hosted by the Human Interaction Lab of the University of Stanford. In this video we have localised and tagged manually all the appearances of objects from four different classes of the well-known and widely used Pascal VOC 2012 dataset, that we have considered as anomalous for that scene. These classes are "person", "dog", "car" and "motorcycle". As it was mentioned before, given the limited computing power of the software our system is going to be deployed in, it is very important to choose a classifier that is well balanced in terms of accuracy and performance. Thus, as the basis of the object location and classification module, we have used the Pytorch framework implementation of the also

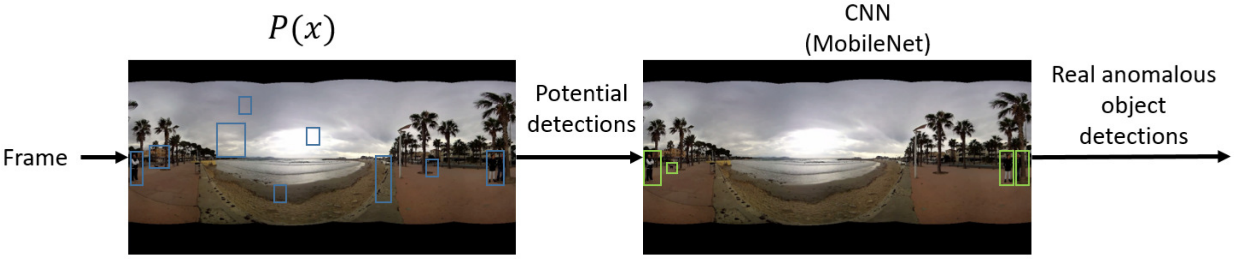

Fig. 2. Illustration of the system's operating with real frames from the [25] video dataset

well-known MobileNet convolutional neural network trained using the cited Pascal VOC 2012 dataset.

Basically, experiments consisted of feeding the 100 first frames of the 360° manually tagged video to the system and counting the total number of anomalous objects detected after processing the 100 frames. This process has been performed for a number of potential detections that goes from 1 to 10 and for each of the three multivariate homoscedastic distributions featured by the potential detection generator, namely a mixture of uniform and gaussian distribution, a mixture of a uniform distribution and a triangular distribution and a mixture of a uniform distribution and a Student-t distribution. In order to calculate the amount of anomalous objects detected on each frame, the system compares the position and identity of the detected objects with the tags appearing on the .xml file associated to that frame and calculates how many of the detected objects have been detected in their correct place with an adjustable margin of error. Despite the relative simplicity of the system's operating, some parameters had to be properly tuned in order to find the proper configuration that allowed to extract the best performance of the entire system. Thus, after the parameter optimisation process, we realised that the optimal parameter values were different for each one of the three distributions. Therefore we have fixed them to the values appearing in Tables II - IV.

Parameter	Value
q	0.4
σ	0.3
detection size (% of frame height)	10
α	0.1

TABLE II
SELECTED PARAMETER VALUES FOR THE UNIFORM-GAUSSIAN MIXTURE DISTRIBUTION.

Parameter	Value
q	0.2
σ	0.3
Max detection size (% of frame height)	10
α	0.1

TABLE III
SELECTED PARAMETER VALUES FOR THE UNIFORM-TRIANGULAR MIXTURE DISTRIBUTION.

Parameter	Value
q	0.7
σ	0.3
Max detection size (% of frame height)	10
α	0.1

TABLE IV
SELECTED PARAMETER VALUES FOR THE UNIFORM-STUDENT-T MIXTURE DISTRIBUTION.

The results obtained are shown in Figure 3 and they represent the number of anomalous objects correctly identified for a number of potential detections spanning from 1 to 10 and for each of the three considered distributions supporting the potential detection generator.

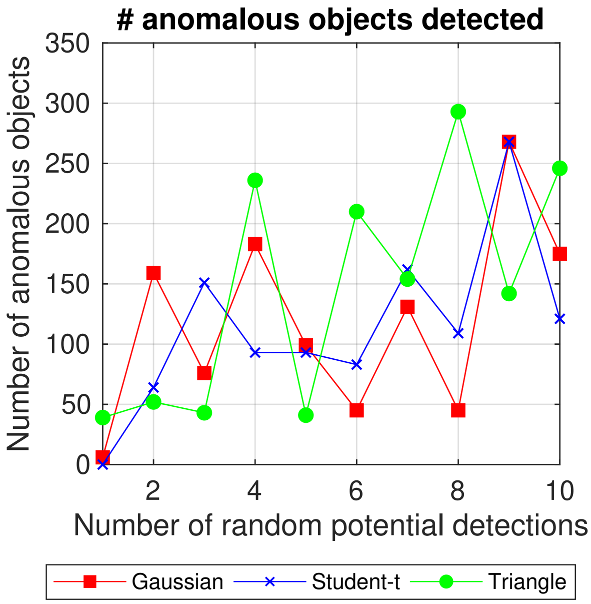

Fig. 3. Mean number of anomalous object detections for each of the three multivariate homoscedastic distributions considering a number of potential detections between 1 and 10.

In general, the plot reveals that the number of anomalous objects detected increases as the number of potential detections does so. However, it can also be observed that the number of detected objects presents several oscillations for all three multivariate distributions utilised by the potential detection generator.

# Windows	1	2	3	4	5	6	7	8	9	10
Gaussian Mixture (fps)	**5.4645**	4.7182	3.8535	**3.3569**	2.8472	**2.5620**	2.1691	**2.0162**	1.7665	**1.7558**
Triangular Mixture (fps)	5.0401	**4.8257**	**3.9403**	3.2813	2.7894	2.4990	**2.2304**	2.0058	1.7528	1.6959
Student-t mixture (fps)	5.2514	4.7676	3.8529	3.3008	**2.8830**	2.4920	2.2292	1.9963	**1.8504**	1.6981

TABLE V

SYSTEM PERFORMANCE EXPRESSED IN MEAN FPS. VS NUMBER OF POTENTIAL DETECTION GENERATIONS FOR THE THREE MIXTURE MODELS.

Individually, Figure 3 reveals that the potential detection generator managed by the uniform-triangular mixture detects more objects than the potential detection generator managed by the other two distributions. However, the uniform-triangular mixture version presents deeper oscillations whilst the uniform-Student-t mixture presents a more stable behaviour despite the fact that it detects a lower amount of objects. The generator managed by the uniform-gaussian distribution seems to present the most balanced behaviour as, although it detects less objects than the generator powered by the uniform-triangular mixture, it also looks more stable in terms of oscillations and it detects more objects than the uniform-Student-t mixture.

Speed is very important issue when designing any automatic video surveillance system. Indeed, if the system is not fast enough, its usefulness may be very limited. But is very important to take into account that we are dealing with a low power consuming system with limited computational power. Therefore, it is critical to know the capabilities of the systems in terms of speed performance when deployed in a Jetson TX2 board. Thus, the system speed experiments consisted of calculating the mean system speed in frames per second (fps) when running a test program that processes 612 frames of the above referenced 360° video for each of the three distributions used by the potential detection generator and for a number of potential detections between 1 and 10. The obtained results are illustrated in Table V.

Considering that the time spent by the MobileNet neural network in processing one potential detection is constant, experiments reveal a similar behaviour for the three homoscedastic distributions in terms of processing speed. However, there are some differences that must be analysed. The maximum processing speed achieved by the entire system is 5.4 frames per second when using a potential detection generator supported by the uniform-gaussian mixture and the number of potential detections is set to 1. The minimum processing speed is also achieved by the system powered by the uniform-gaussian mixture for a number of 10 potential detections. In general, is difficult to establish a clear winner between the three distributions used to implement the potential detection generator, in terms of speed. But according to the values shown in Table V, the uniform-gaussian distribution achieved the highest processing speed for 1, 4, 6, 8 and 10 windows. This means that the uniform-gaussian distribution is faster than the others in 50% of the executions, whilst the uniform-triangular distribution is faster in the 30% of the times, and the uniform-Student-t distribution stands out from its competitors 20% of the times. Therefore, even though they have similar execution speeds, uniform-gaussian distribution should be recommended in terms of time consumption.

V. CONCLUSION

In this paper it is presented the design and implementation of a deep learning-based video surveillance system, managed by low power hardware devices, capable of detecting anomalous objects in a video stream shot by a panoramic camera. The system presents a potential detection generator, relying on a new mathematical model based on three multivariate homoscedastic distributions, in charge of deciding what parts of the scene are more likely to contain an anomalous object; and a MobileNet convolutional neural network that will perform the classification of the detected objects. Parameter optimisation is such a heavy task in terms of computational requirements and this fact conditioned the dimension for the parameter optimisation process performed insofar as more time is needed to perform a deeper parameter optimisation process. This circumstance leads us to think that there is still some margin for improvements in this direction allowing us to be optimistic with respect to future performance of our model. Nevertheless, performance tests at this moment reveal that the system is capable of detecting anomalous objects with an acceptable accuracy at a speed of up to 5.4 fps, positioning our approach as a good foundation for the construction of energy saving automatic video surveillance systems.

ACKNOWLEDGMENT

This work is partially supported by the Ministry of Economy and Competitiveness of Spain under grants TIN2016-75097-P and PPIT.UMA.B1.2017. It is also partially supported by the Ministry of Science, Innovation and Universities of Spain under grant RTI2018-094645-B-I00, project name Automated detection with low-cost hardware of unusual activities in video sequences. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project UMA18-FEDERJA-084, project name Detection of anomalous behavior agents by deep learning in low-cost video surveillance intelligent systems. All of them include funds from the European Regional Development Fund (ERDF). The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X GPUs used for this research. The authors acknowledge the funding from the Universidad de Málaga.

References

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010, cited By :1484. [Online]. Available: www.scopus.com

[2] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000, cited By :2192. [Online]. Available: www.scopus.com

[3] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, 2015, pp. 2197–2206, cited By :1279. [Online]. Available: www.scopus.com

[4] D. Dalwadi, Y. Mehta, and N. Macwan, *Face recognition-based attendance system using real-time computer vision algorithms*, ser. Advances in Intelligent Systems and Computing, 2021, vol. 1141. [Online]. Available: www.scopus.com

[5] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, 2003, pp. 2–10, cited By :330. [Online]. Available: www.scopus.com

[6] L. Li, W. Huang, I. Y. . Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1459–1472, 2004, cited By :781. [Online]. Available: www.scopus.com

[7] M. Vijayan and R. Mohan, "A universal foreground segmentation technique using deep-neural network," *Multimedia Tools and Applications*, vol. 79, no. 47-48, pp. 34 835–34 850, 2020. [Online]. Available: www.scopus.com

[8] M. T. McCann, K. H. Jin, and M. Unser, "Convolutional neural networks for inverse problems in imaging: A review," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 85–95, 2017.

[9] M. Koziarski and B. Cyganek, "Image recognition with deep neural networks in presence of noise - dealing with and taking advantage of distortions," *Integrated Computer-Aided Engineering*, vol. 24, pp. 337–349, 2017.

[10] A. Zhang, K. C. P. Wang, Y. Fei, Y. Liu, C. Chen, G. Yang, J. Q. Li, E. Yang, and S. Qiu, "Automated pixel-level pavement crack detection on 3d asphalt surfaces with a recurrent neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 3, pp. 213–229, 2019, cited By :28. [Online]. Available: www.scopus.com

[11] S. Bang, S. Park, H. Kim, and H. Kim, "Encoder–decoder network for pixel-level road crack detection in black-box images," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 8, pp. 713–727, 2019.

[12] X. Liang, "Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with bayesian optimization," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 5, pp. 415–430, 2019.

[13] C. Chen, S. Li, H. Qin, and A. Hao, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *Pattern Recognition*, vol. 52, pp. 410 – 432, 2016.

[14] H. Sajid, S.-C. S. Cheung, and N. Jacobs, "Appearance based background subtraction for PTZ cameras," *Signal Processing: Image Communication*, vol. 47, pp. 417 – 425, 2016.

[15] C. Micheloni, B. Rinner, and G. Foresti, "Video analysis in pan-tilt-zoom camera networks," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 78–90, 2010.

[16] C. Ding, B. Song, A. Morye, J. Farrell, and A. Roy-Chowdhury, "Collaborative sensing in a distributed PTZ camera network," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3282–3295, 2012.

[17] G. Scotti, L. Marcenaro, C. Coelho, F. Selvaggi, and C. S. Regazzoni, "Dual camera intelligent sensor for high definition 360 degrees surveillance," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 250–257, 2005.

[18] Y. Sato, K. Hashimoto, and Y. Shibata, "A new networked surveillance video system by combination of omni-directional and network controlled cameras," in *Network-Based Information Systems*, M. Takizawa, L. Barolli, and T. Enokido, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 313–322.

[19] Q. Fan and Y. Xu, "A robust target recognition and tracking panoramic surveillance system based on deep learning," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 11342, 2019. [Online]. Available: www.scopus.com

[20] P. Angelov, P. Sadeghi-Tehran, and C. Clarke, "AURORA: Autonomous real-time on-board video analytics," *Neural Comput. Appl.*, vol. 28, no. 5, pp. 855–865, 2017.

[21] A. Dziri, M. Duranton, and R. Chapuis, "Real-time multiple objects tracking on raspberry-pi-based smart embedded camera," *Journal of Electronic Imaging*, vol. 25, p. 041005, 2016.

[22] J. Benito-Picazo, E. Domínguez, E. J. Palomo, and E. López-Rubio, "Deep learning-based video surveillance system managed by low cost hardware and panoramic cameras," *Integrated Computer-Aided Engineering*, vol. 27, no. 4, pp. 373–387, 2020. [Online]. Available: www.scopus.com

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[24] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, p. 303–338, Jun. 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4

[25] V. H. I. Lab, "360 video database," accessed: 2021-02-09. [Online]. Available: https://vhil.stanford.edu/