Object Detection and Multi-class Classification from Videos: An Application of AI-enabled Surveillance Camera

P.Anil Jawalkar[1], Akshaya[2], B. Swathi[2], B. Sraveena[2],B. Satya Sahithi[2]

[1] Assistant Professor, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India, anil.jawalkar022@gmail.com.

[2] Student, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India.

Abstract

Surveillance cameras play a crucial role in upholding public safety, safeguarding properties, and acting as a deterrent to criminal activities. However, the conventional surveillance systems heavily rely on basic image processing techniques and rule-based algorithms for tasks like object detection, tracking, and recognition. Unfortunately, these methods have their limitations when dealing with complex scenarios such as occlusions, changes in lighting conditions, and variations in object appearances. Therefore, false alarms or missed detections can occur, ultimately diminishing the overall effectiveness of the surveillance system. To tackle these challenges head-on, researchers and engineers have been actively exploring the integration of advanced computer vision techniques, particularly Convolutional Neural Networks (CNNs), to augment the capabilities of smart surveillance cameras. The aim is to improve the efficiency and accuracy of these systems by leveraging cutting-edge CNN modifications. The need for enhancing the performance of smart surveillance cameras arises from the ever-growing demand for more intelligent and reliable surveillance solutions. As the number of surveillance cameras deployed in public spaces, commercial areas, and private premises continues to increase, it becomes imperative to develop advanced algorithms capable of accurately and efficiently analyzing vast amounts of video data. By elevating the performance of surveillance cameras, it becomes feasible to provide real-time and accurate insights to security personnel, law enforcement agencies, and other stakeholders. Hence, the core objective of this project revolves around enhancing the performance of smart surveillance cameras using advanced CNN modifications. A custom CNN architecture is proposed to optimize object detection, tracking, and recognition tasks. This approach undergoes meticulous experimentation with real-world surveillance data to evaluate the effectiveness of the modifications. The results manifest significant improvements in the system's accuracy and efficiency, paving the way for more intelligent and reliable surveillance solutions across various applications. The outcomes of this research hold the potential to contribute significantly to the fields of computer vision and smart surveillance technology, enhancing public safety and security overall.

1. Introduction

Smart surveillance cameras play a pivotal role in contemporary security and monitoring systems. They serve a multitude of purposes, from safeguarding public spaces and managing traffic to protecting valuable assets [1]. These cameras, capable of capturing copious visual data, necessitate real-time processing and analysis. The application of CNNs in computer vision and image analysis traces its roots to the 1990s. Over time, CNNs have evolved with more intricate architectures and refined training techniques [2]. The adoption of CNNs in surveillance cameras burgeoned in tandem with advancements in hardware and software capabilities, enabling the real-time scrutiny of video streams [3, 4]. The journey of enhancing smart surveillance camera performance is intricately entwined with the evolution of deep learning and CNNs in the domain of computer vision. Image classification, as a classical research topic in recent years, is one of the core issues of computer vision and the basis of various fields of visual recognition. The improvement of classification network performance tends to significantly improve its application level, for example to object-detection, segmentation, human pose estimation, video classification, object tracking, and super-resolution technology. Improving image classification technology is an important part of promoting the development of computer vision. Its main process includes image data preprocessing, feature extraction and representation, and classifier design. The focus of image classification research has always been image feature extraction, which is the basis of image classification. Traditional image feature extraction algorithms focus more on manually setting specific image features. This method has poor generalization ability and portability. So, letting a computer have the ability to process images similar to biological vision is what researchers dream of.

In the last decade, modern video surveillance systems have attracted increasing interest with several studies focusing on automated video surveillance systems, which involve a network of surveillance cameras with sensors that can monitor human and nonhuman objects in a specific environment. Pattern recognition can be used to find specific arrangements of features or data, which usually yield details regarding a presented system or data set. In a technical context, a pattern can involve repeating sequences of data with time, and patterns can be utilized to predict trends and specific featural configurations in images to recognize objects. Many recognition approaches involving the use of the support vector machine (SVM) [5], artificial neural network (ANN) [6], deep learning [7], and other rule-based classification systems have been developed. Performing classification using an ANN is a supervised practical strategy that has achieved satisfactory results in many classification tasks. The SVM requires fewer computational requirements than the ANN; however, the SVM provides lower recognition accuracy than the ANN. In recent years, networks have played a significant role in a wide range of applications, and they have been applied to surveillance systems. In recent years, as the amounts of unstructured and structured data have increased to big data levels, researchers have developed deep learning systems that are basically neural networks with several layers. Deep learning allows one to capture and mine larger amounts of data, including unstructured data. This approach can be used to model complicated relationships between inputs and outputs or to find patterns. However, the associated accuracy and classification efficiency are generally low [8]. Many strategies have been developed to increase the recognition accuracy. In this work, we discuss the accuracy gains from adopting certain saliency methods to improve the recognition and detection of an object and isolate it from a scene.

2. Literature survey

The performance efficiency of existing surveillance systems is highly dependent on the activity of human operators who are responsible for monitoring the camera footage. In general, most medium and large surveillance systems involve numerous screens (approximately 50 or more) that display the streams captured by numerous cameras. As the number of simultaneous video streams to be viewed increases, the work of surveillance operators becomes considerably challenging and fatiguing. Practically, after twenty minutes of continuous work, the attention of the operators is expected to degrade considerably. In general, the operators check for the absence or presence of objects (i.e., people and vehicles) in surveillance areas and ensure that the maximum capacity of a place remains intact, such as by ensuring that no unauthorized people are present in restricted areas and no objects are present in unexpected places. The failures of such systems in alarming authorities can be attributed to the limitations of manual processing. Generally, most traditional methods used to obtain evidence depend heavily on the records of the security camera systems in or near accident sites. Practically, when an incident occurs in a vast space or considerable time has elapsed since its occurrence, it is difficult to find any valuable evidence pertaining to the perpetrators from the large number of surveillance videos, which hinders the resolution of the cases. Thus, to minimize the mental burden of the operators and enhance their attention spans, it is desirable that an automated system that can reliably alert an operator of the presence of target objects (e.g., a human) or the occurrence of an anomalous event be developed.

Pattern recognition, which is widely used in many recognition applications, can be performed to find arrangements of features or data, and this technique can be applied in the surveillance domain. Several recognition approaches involving the support vector machine, artificial neural networks, decision trees, and other rule-based classification systems have been proposed. Machine learning typically uses two types of approaches, namely, supervised and unsupervised learning. Using these approaches, especially supervised learning, we can train a model with known input and output data to ensure that it can estimate any future output. Moreover, in some existing systems, an artificial immune system (AIS)-inspired framework, where the AIS is a computational paradigm that is a part of the computational intelligence family and is inspired by the biological immune system that can reliably identify unknown patterns within sequences of input images, has been utilized to achieve real-time vision analysis designed for surveillance applications.

The field of video surveillance is very wide. Active research is ongoing in subjects, such as automatic thread detection and alarms, large-scale video surveillance systems, face recognition and license plate recognition systems, and human behavior analysis. Intelligent video surveillance is of significant interest in industry applications because of the increased requirement to decrease the time it takes to analyze large-scale video data. Relating to the terminology, Elliott [9] recently described an intelligent video system (termed IVS) as “any kind of video surveillance method that makes use of technology to automatically manipulate process and/or achieved actions, detection, alarming and stored video images without human intervention.” Academic and industry studies are focused on developing key technologies for designing powerful intelligent surveillance systems along with low-cost computing hardware; and the applications include object tracking [10], pedestrian detection, gait analysis, vehicle recognition, privacy protection, face and iris recognition, video summarization, and crowd counting. Nguyen [11] described the implementation and design of an intelligent low-cost monitoring system using a Raspberry Pi and a motion detection algorithm programmed in Python as a traditional programming environment. Additionally, the system utilizes the motion detection algorithm to considerably reduce storage usage and save expense costs. The motion detection algorithm is executed on a Raspberry Pi that enables live streaming cameras together with motion detection. The real-time video camera can be viewed from almost any web browser, even by mobile devices. Sabri et al. [12] present a real-time intruder monitoring system based on a Raspberry Pi to deploy a surveillance system that is effective in remote and scattered places, such as universities. The system hardware consists of a Raspberry Pi, long-distance sensors, cameras, a wireless module and alert circuitry; and the detection algorithm is designed in Python. This system is a novel cost-effective solution with good flexibility and improvement needed for monitoring pervasive remote locations. The results show that the system has high reliability for smooth working while using web applications; in addition, it is cost-effective. Therefore, it can be integrated as several units to catch and concisely monitor remote and scattered areas. Their system can also be controlled by a remote user geographically or sparsely far from any networked workstation. The recognition results prove that the system efficiently recognized intruders and provided alerts when detecting intruders at distances between one to three meters from the system camera. The recognition accuracy is between 83% and 95% and the reliable warning alert is in the range of 86–97%. Turchini et al. [13] proposes an object tracking system that was merged with their lately developed abnormality detection system to provide protection and intelligence for critical regions. In recent years, many studies have focused on using artificial intelligence for intelligence surveillance systems. These techniques involve different approaches, such as the SVM, the ANN, and the latest developed types based on deep learning techniques. However, deep neural networks are computationally challenging and memory hungry; therefore, it is difficult to run these models in low computational systems, such as single board computers [14]. Several approaches have been utilized to address this problem. Many approaches have reduced the size of neural networks and maintained the accuracy, such as MobileNet, while other approaches minimize the number of parameters or the size [15].

3. Proposed System Model

The class diagram is used to refine the use case diagram and define a detailed design of the system. The class diagram classifies the actors defined in the use case diagram into a set of interrelated classes. The relationship or association between the classes can be either an “is-a” or “has-a” relationship. Each class in the class diagram may be capable of providing certain functionalities. These functionalities provided by the class are termed “methods” of the class. Apart from this, each class may have certain “attributes” that uniquely identify the class as shown in Figure 1. According to the facts, training and testing of proposed model involves in allowing every source image via a succession of convolution layers by a kernel or filter, rectified linear unit (ReLU), max pooling, fully connected layer and utilize SoftMax layer with classification layer to categorize the objects with probabilistic values ranging from [0,1]. Convolution layer as is the primary layer to extract the features from a source image and maintains the relationship between pixels by learning the features of image by employing tiny blocks of source data. It’s a mathematical function which considers two inputs like source image $I(x,y,d)$ where x and y denotes the spatial coordinates i.e., number of rows and columns. d is denoted as dimension of an image (here $d = 3$, since the source image is RGB) and a filter or kernel with similar size of input image and can be denoted as $F(k_x, k_y, d)$.

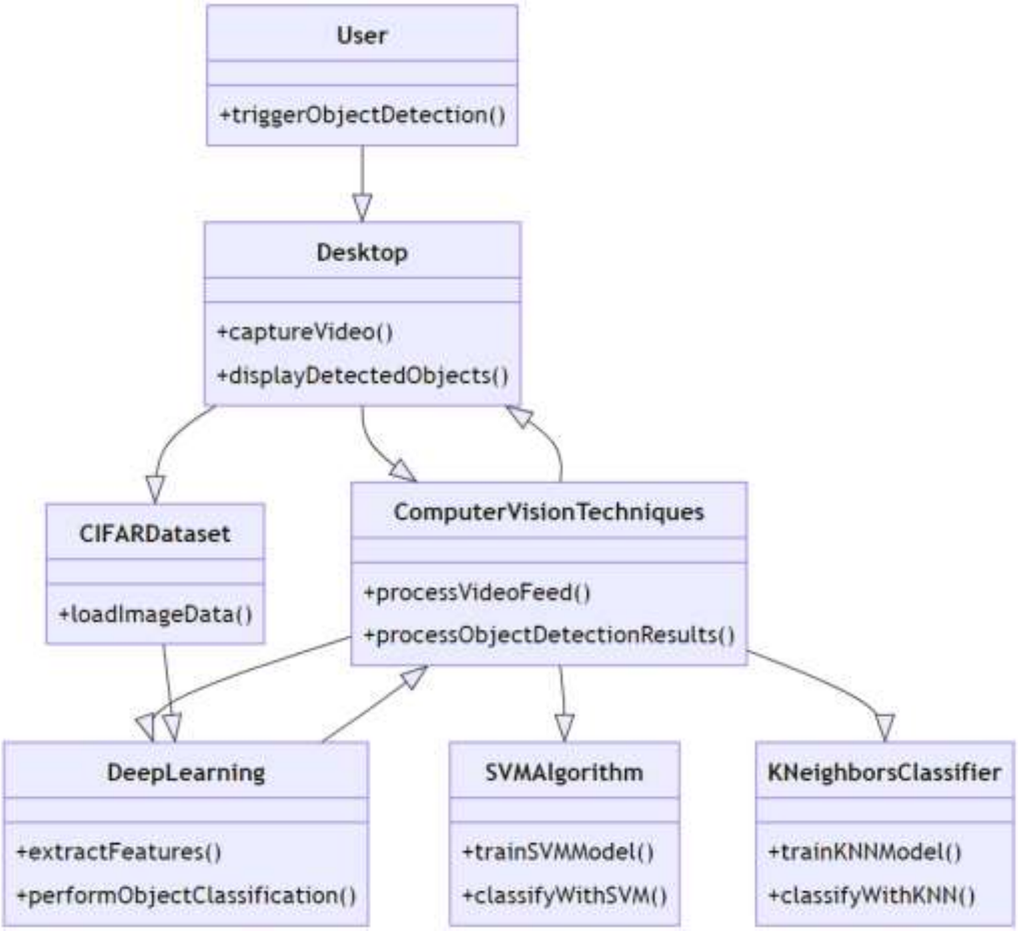


Figure 1. Proposed system model.

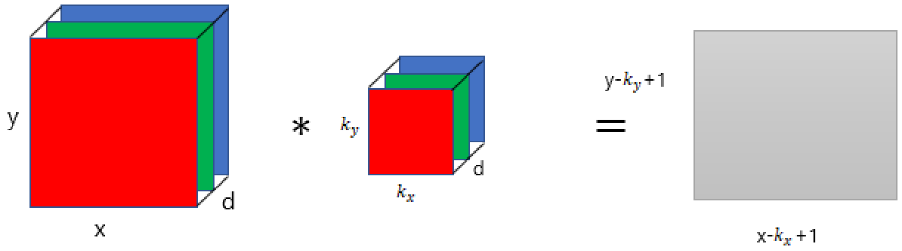


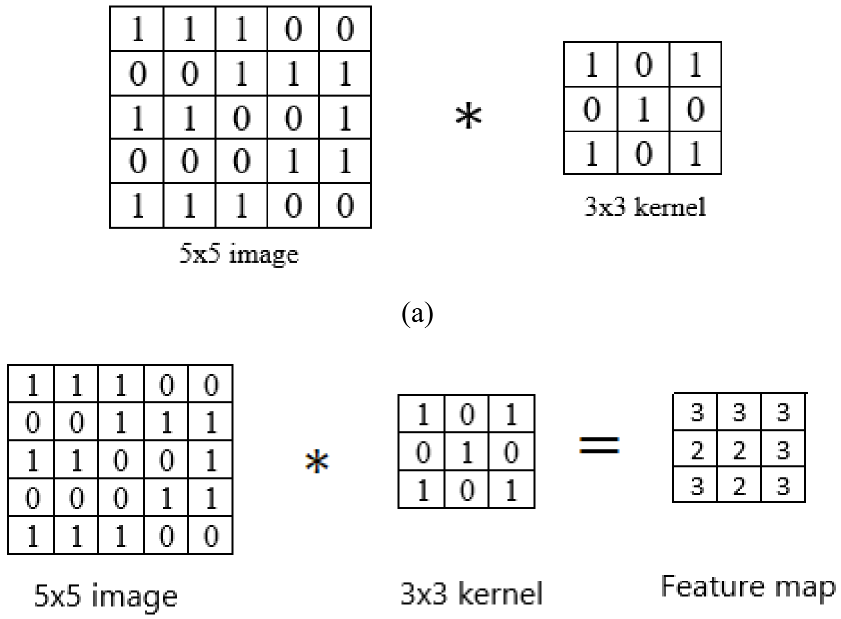
Fig. 2: Representation of convolution layer process.

The output obtained from convolution process of input image and filter has a size of $C\left((x - k_x + 1), (y - k_y + 1), 1\right)$, which is referred as feature map. Let us assume an input image with a size of 5×5 and the filter having the size of 3×3. The feature map of input image is obtained by multiplying the input image values with the filter values.

ReLU layer: Networks those utilizes the rectifier operation for the hidden layers are cited as rectified linear unit (ReLU). This ReLU function $\mathcal{G}(\cdot)$ is a simple computation that returns the value given as input directly if the value of input is greater than zero else returns zero. This can be represented as mathematically using the function $max(\cdot)$ over the set of 0 and the input x as follows:

$$\mathcal{G}(x) = \max\{0, x\}$$

Max pooing layer: This layer mitigates the number of parameters when there are larger size images. This can be called as subsampling or down sampling that mitigates the dimensionality of every feature map by preserving the important information. Max pooling considers the maximum element form the rectified feature map.



1	1	1	0	0
0	0	1	1	1
1	1	0	0	1
0	0	0	1	1
1	1	1	0	0

5x5 image

$*$

1	0	1
0	1	0
1	0	1

3x3 kernel

(a)

1	1	1	0	0
0	0	1	1	1
1	1	0	0	1
0	0	0	1	1
1	1	1	0	0

5x5 image

$*$

1	0	1
0	1	0
1	0	1

3x3 kernel

$=$

3	3	3
2	2	3
3	2	3

Feature map

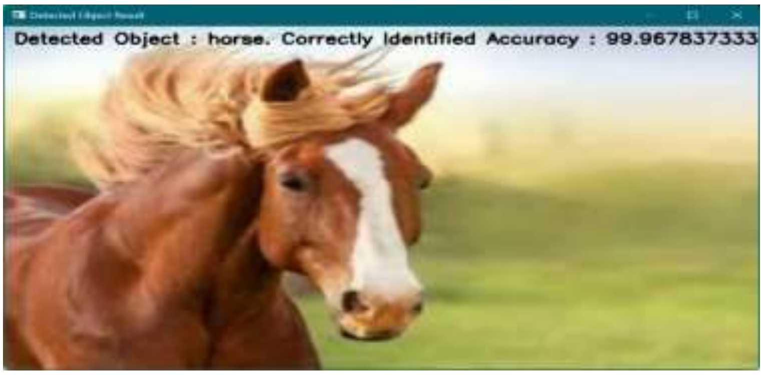
(b)

Fig. 3: Example of convolution layer process (a) an image with size $\mathbf{5} \times \mathbf{5}$ is convolving with $\mathbf{3} \times \mathbf{3}$ kernel (b) Convolved feature map

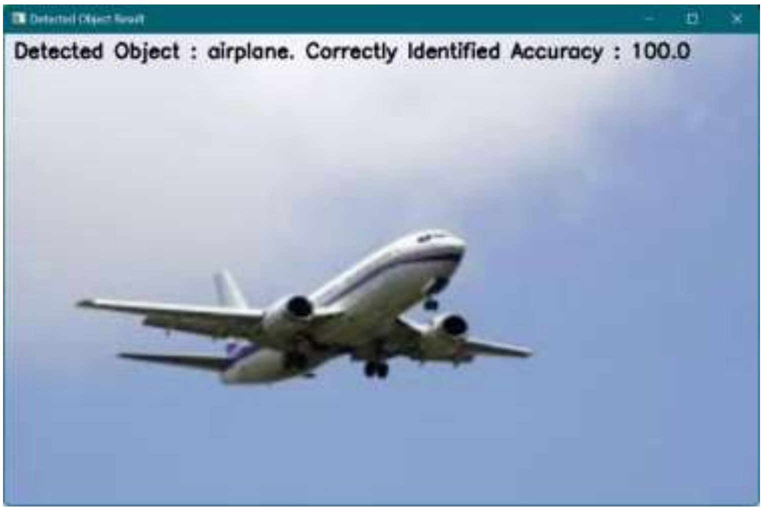
Softmax classifier: Generally, as seen in the above picture softmax function is added at the end of the output since it is the place where the nodes are meet finally and thus, they can be classified. Here, X is the input of all the models and the layers between X and Y are the hidden layers and the data is passed from X to all the layers and Received by Y. Suppose, we have 10 classes, and we predict for which class the given input belongs to. So, for this what we do is allot each class with a particular predicted output. Which means that we have 10 outputs corresponding to 10 different class and predict the class by the highest probability it has.

4. Results description

Figure 4 shows the initial interface of the application. It displays the graphical user interface (GUI) with various buttons, labels, and widgets for different functionalities related to image classification and object detection for enhancing the smart surveillance camera performance.



Detected Object : horse. Correctly Identified Accuracy : 99.967837333



Detected Object : airplane. Correctly Identified Accuracy : 100.0

Figure 4: Sample predicted objects from given test video.

5. Conclusion

The project on "Object Detection and Classification for Enhancing Surveillance Camera Performance Using CNN Model" has successfully demonstrated the potential for improving the capabilities of smart surveillance systems. The use of CNNs and advanced computer vision techniques has led to significant advancements in object detection and classification, which are crucial for surveillance applications. Key accomplishments and findings of the project includes improved object detection, where the CNN model employed in the project has shown remarkable accuracy in detecting objects within surveillance video feeds. This enhances the system's ability to identify and track objects of interest effectively. Precise object classification, where the CNN model not only detects objects but also classifies them into predefined categories. This classification capability provides valuable insights and context to surveillance personnel. Integration of advanced techniques, where the project leveraged advanced techniques such as Global Color Histogram, Local Binary Pattern (LBP), Histogram of Oriented Gradients (HOG), and Support Vector Machines (SVM) for further enhancing object recognition and classification. Finally, this proposed system has processed video streams in real-time, enabling immediate responses to detected objects. This is critical for security and surveillance applications.

References

[1] Selvi E, Adimoolam M, Karthi G, Thinakaran K, Balamurugan NM, Kannadasan R, Wechtaisong C, Khan AA. Suspicious Actions Detection System Using Enhanced CNN and Surveillance Video. Electronics. 2022; 11(24):4210.

[2] Yu Wan, Zhaohong Liao, Jia Liu, Weiwei Song, Hong Ji, Zhi Gao, Small object detection leveraging density□aware scale adaptation, The Photogrammetric Record, 10.1111/phor.12446, 38, 182, (160-175), (2023).

[3] Xu, J. A deep learning approach to building an intelligent video surveillance system. Multimed Tools Appl 80, 5495–5515 (2021). https://doi.org/10.1007/s11042-020-09964-6

[4] Sun Y, Zhi X, Han H, Jiang S, Shi T, Gong J, Zhang W. Enhancing UAV Detection in Surveillance Camera Videos through Spatiotemporal Information and Optical Flow. Sensors. 2023; 23(13):6037.

[5] Junoh et al. (2012) Junoh AK, Mansor MN, Abu SA, Ahmad WZW. Procedia engineering, vol. 38. Elsevier Ltd; 2012. SVM classifier for automatic surveillance system; pp. 1806–1810.

[6] Petrosino & Maddalena (2012) Petrosino A, Maddalena L. Handbook on soft computing for video surveillance. Chapman & Hall/CRC; 2012. Neural networks in video surveillance: a perspective view; pp. 59–78.

[7] Wang et al. (2019) Wang W, Shen J, Xie J, Cheng MM, Ling H, Borji A. Revisiting video saliency prediction in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019;43:220–237. doi: 10.1109/TPAMI.2019.2924417.

[8] Liu & An (2020) Liu JE, An FP. Scientific programming, vol. 2020. Hindawi: 2020. Image classification algorithm based on deep learning-kernel function; p. 7607612.

[9] Elliott (2010) Elliott D. Intelligent video solution: a definition. Security. 2010;47(6):46–48.

[10] Khan & Gu (2010) Khan Z, Gu I. Joint feature correspondences and appearance similarity for robust visual object tracking. IEEE Transactions on Information Forensics and Security. 2010;5(3):591–606. doi: 10.1109/TIFS.2010.2050312.

[11] Nguyen et al. (2015) Nguyen HQ, Loan TTK, Mao BD, Huh EN. Low-cost real-time system monitoring using Raspberry Pi. 2015. Seventh international conference on ubiquitous and future networks; Sapporo, Japan. 2015. pp. 857–859

[12] Sabri et al. (2018) Sabri N, Salim MS, Fouad S, Aljunid SA, AL-Dhief FT, Rashidi CBM. Design and implementation of an embedded smart intruder surveillance system. MATEC web of conferences 150 Malaysia technical universities conference on engineering and technology (MUCET 2017). Vol. 150. 2018:1–6.

[13] Turchini et al. (2018) Turchini F, Seidenar L, Uricchio T, Bimbo A. Deep learning-based surveillance system for open critical areas. Inventions. 2018;3(4):69-1–69-13. doi: 10.3390/inventions3040069.

[14] Verhelst & Moons (2018) Verhelst M, Moons B. Embedded deep neural network processing: algorithmic and processor techniques bring deep learning to IoT and edge devices. IEEE Solid-State Circuits Magazine. 2018;9(4):55–65. doi: 10.1109/MSSC.2017.2745818.

[15] Véstias (2019) Véstias MP. A survey of convolutional neural networks on edge with reconfigurable computing. Algorithms, MDPI. 2019;12(8):154. doi: 10.3390/a12080154.