Post-processing of video surveillance systems alarm signals using the YOLOv8 neural network

Olga Pavlova[1,*,†], Ivan Rudyk[1†] and Houda EL Bouhissi[2†]

[1] *Khmelnytskyi National University, Instytuts'ka str., 11, Khmelnytskyi, 29016, Ukraine*
[2] *LIMED Laboratory, Faculty of Exact Sciences,University of Bejaia, 06000, Bejaia, Algeria*

Abstract

The method of solving the security problem of a warehouse equipped with external surveillance cameras by processing the video stream using artificial neural networks is considered. Experiments were conducted to test already existing pre-trained models and the model that gave the highest recognition quality - YOLOv8 was determined. A dataset of images taken from outdoor surveillance cameras was also compiled for training, validation and experiments. It was proven that the YOLOv8 neural network model currently coped best with the given task, so it will be used for further experiments. However, upon manual verification, it was observed that the proportion of objects identified with errors by the newly developed model decreased to 10.7%. The obtained metrics reflect the success of the training process, as evidenced by improvements in parameters such as Train Box Loss (reduced to 0.5135) and mAP50 (increased to 0.98367) with each successive epoch. However, the relatively stable Validation Box Loss value (0.69291) from epoch 84 onward suggests inherent performance fluctuations possibly due to the limited size of the validation sample (6% of the training sample).

Keywords
video-image processing, pattern recognition, objects detection, neural networks, YOLOv8

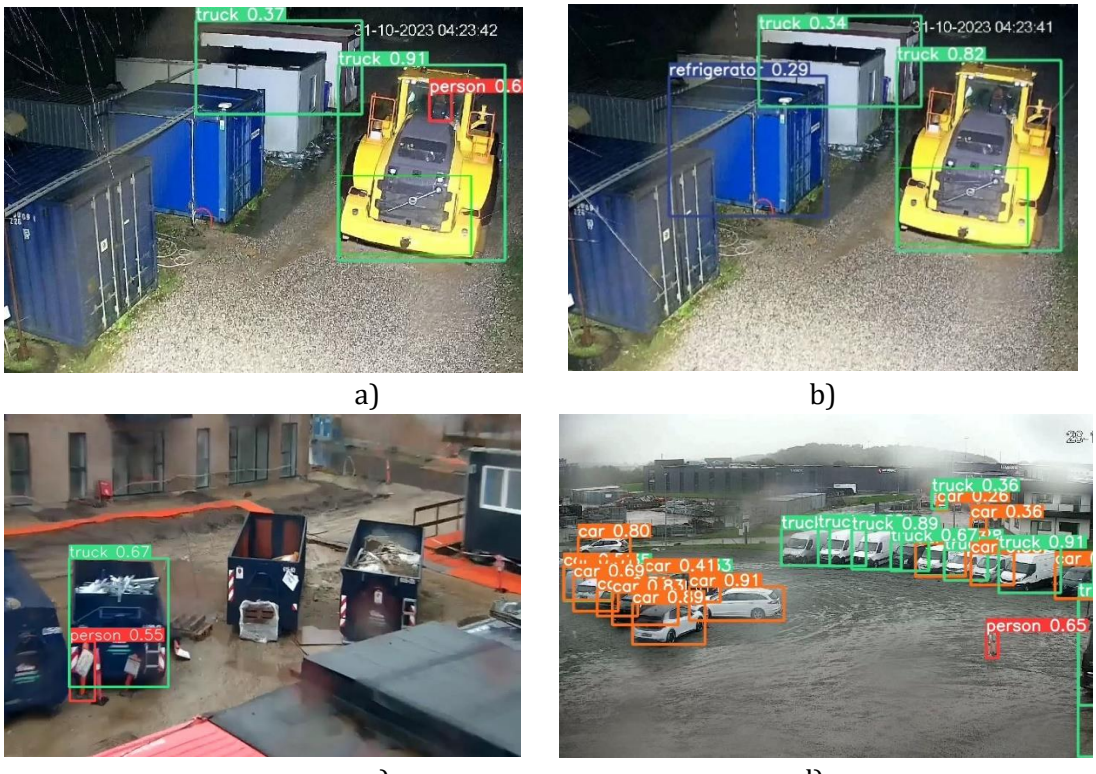
1. Introduction

Companies that specialize in providing comprehensive video surveillance services often face the problem of pattern recognition quality. Every day, their operators receive signals about violations from more than 400 cameras. The average daily number of signals is almost 3,000 (more than 86,000 for the past month). It should be noted that the majority of cameras are located outside and are significantly affected by various factors that cause false alarms. These factors can be divided into categories:

IntelITSIS'2024: 5th International Workshop on Intelligent Information Technologies and Systems of Information Security, March 28, 2024, Khmelnytskyi, Ukraine
* Corresponding author.
† These authors contributed equally.
pavlovao@khmnu.edu.ua (O. Pavlova); noveimya@gmail.com (I. Rudyk); houda.elbouhissi@gmail.com (H. El Bouhissi)
0000-0003-2905-0215 (O. Pavlova); 0009-0009-8355-3471(I. Rudyk); 0000-0003-3239-8255 (H. El Bouhissi)

1) precipitation (rain, snow);
2) wind (shakes objects in the field of view of the camera, or the camera itself);
3) light effects (flickering of light on the territory or outside it, car headlights at night, glare from the sun, or just a shadow from a cloud during the day).

All the mentioned factors, and more often – their combinations – make up a significant part of all camera activations. So, for the mentioned last month, the number of false alarms reached 44 thousand, that is, more than 51% of all alarm signals. Considering that each alarm must be verified by a person - false alarms cause significant losses to the owner, forcing to keep an increased number of operators on the job. Figure 1 shows examples of errors in pattern recognition on images from outdoor surveillance cameras (a) false recognition of a truck and a person driving a bulldozer; b) false recognition of a refrigerator and a truck at the place of cargo containers; c) false identification of a truck and a person at the place of the cargo container and demarcation column; d) false recognition of trucks in the place of minibuses and a person in the place of a limiting partition).



a) b)

c) d)

Figure 1: Examples of errors in pattern recognition on images from outdoor surveillance cameras.

Since external surveillance cameras are intended mainly for enterprise security purposes, namely, to prevent illegal entry into the territory of outsiders and vehicles, *the problem of high-quality recognition of patterns on the video stream from the cameras is quite important* for the automation of the operator's work.

2. Related works

In the course of the study, an analysis of the latest scientific publications in the field of pattern recognition using artificial neural networks was carried out. Scientific publications devoted to the application of models based on artificial neural networks such as Google Cloud Vision API, Pytorch Faster R-CNN, OpenCV+CNN and YOLO libraries for various fields such as biology, medicine, smart cities and cyber-physical systems, recognition of gestures and facial expressions were reviewed. The results of scientific publications analysis are presented in Table 1.

Table 1
Analysis of ready existing computer vision approaches for pattern recognition

Reference	Year	Algorithm / Model	Scope of application	Brief Description of the approach
1	2	3	4	5
Pavlova O. et al. [1]	2021	CNN	Smart parking system	The research in general is aimed at image recognition for camera-based smart parking using convolutional neural network (CNN).
Zeng Y. et al. [2]	2020	Google Cloud AutoML Vision	Medicine. Early diagnosis of carcinoma.	The paper assesses the feasibility of an AutoML approach for the identification of invasive ductal carcinoma (IDC) in whole slide images (WSI). An experimental IDC identification model is built with Google Cloud AutoML Vision.
Lu Y. et al. [3]	2020	Google Cloud Vision (GCV) APIs	Cybersecurity	The transferability of adversarial examples across a wide range of real-world computer vision tasks, including image classification, object detection, semantic segmentation, explicit content detection, and text detection were investigated.
Radiuk P. et al. [4]	2022	An Ensemble of Fine-tuned NNs	statistical analysis	The classification task of emotional expressions was performed according to several machine learning algorithms: raw random forest, gradient boosting random forest, support vector machine, multilayer perceptron,

				recurrent neural network, and convolutional neural network
Sahoo J. et al. [5]	2022	pre-trained CNN model with score-level fusion technique	Hand gesture recognition	A real-time American sign language (ASL) recognition system is developed and tested using the proposed technique.
Gazda M. et al. [6]	2021	multiple-fine-tuned CNNs	Medicine. Parkinson's Disease Diagnosis	Multiple-fine-tuned convolutional neural networks for Parkinson's disease diagnosis from offline handwriting.
Radiuk P. et al. [7]	2022	Google Cloud Vision OpenCV	Parking slots detection	Google Cloud Vision technology as parking slots detector and a pre-trained convolutional neural network as a feature extractor and a classifier were selected to develop a cyber-physical system for smart parking.
Bussa S.et al. [8]	2020	OpenCV	Face recognition	Smart attendance system using OpenCV based on facial recognition.
Safran M. et al. [17]	2024	YOLOv8 and CNN	License plate recognition	Efficient Multistage License Plate Detection and Recognition Using YOLOv8 and CNN for Smart Parking Systems
Melnychenko O. et al. [20]	2023	YOLOv5-v1	Apples recognition	A novel architecture to identify apples in images with occlusions was created.

As can be seen from Table 1, a deep learning approach, particularly CNNs, has been most frequently used over the past five years and has shown the most robust recognition of individual objects, parking slots, faces, emotions, gestures, medical patterns etc. Therefore, considering the abovementioned analysis, it was decided to apply the pre-trained model for objects captured by the surveillance cameras recognition.

3. Dataset preparation and model selection for the implementation

The search for a solution has been ongoing for almost 10 years. The number of cameras is increasing, the latest streaming video analysis systems are being put into operation. However, the quality of automated image recognition that could satisfy the company's security parameters and eliminate the human operator from the process of monitoring the cameras was not achieved.

Based on the above analysis of existing solutions, it was decided to collect a dataset of 12 test images and conduct a comparative analysis on the quality of pattern recognition. Google Cloud Vision API [22], Pytorch FasterR-CNN [23] and YOLOv8 neural network [27] were

chosen for testing. The dataset that was collected and prepared for models testing is presented in Figure 2. It contains images in infrared light, in shades of gray, on which there are images of various objects on a construction site, cars, people.

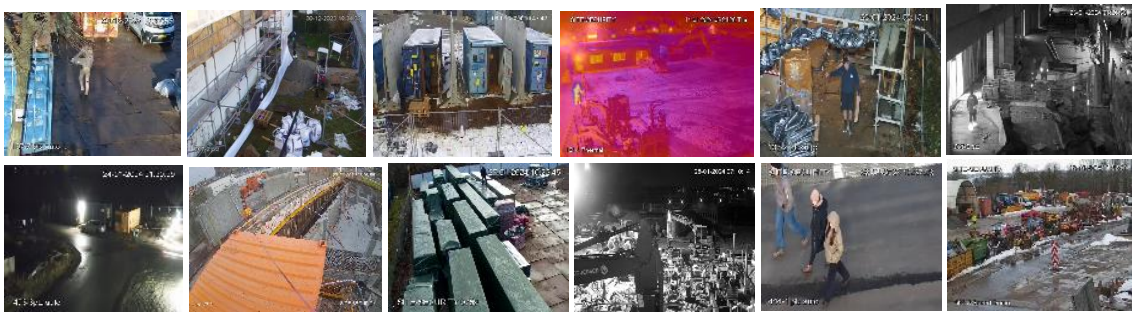


Figure 2: A dataset of target images for models testing.

Since the target objects for recognition in the images are people, the focus of the experiment was on recognizing the images of people and establishing the presence of people. According to the results of testing the models [22, 23, 27], a table was compiled with the effectiveness of recognizing people's images and with the available advantages and disadvantages of each of the models. The test results are presented in Table 3.

Table 3
The results of the neural networks models testing using the pre-prepared dataset

Criterion	Google Cloud Vision API	YoloV8	PyTorch Faster R-CNN
Cases of count people matched	1	8	8
Found bigger count of people than were presented really	2	1	2
Found fewer count of people than were presented really	9	3	2
% of matched	8,3	66,7	66,7
Advantages	Good integration with other Google services	User-friendly API, high accuracy	High accuracy
Disadvantages	Low accuracy	-	Complicated API (compared to others)

As can be seen from Table 3, the result of the analysis showed that the YoloV8 model coped best with pattern recognition on the test dataset, so it was decided to build further work on its basis.

4. Image recognition using the YOLOv8 neural network

Below is the experience of object recognition in video clips sent by surveillance cameras after they were triggered. It should be noted that cameras were chosen for analysis, where, according to expectations, there should be no people at the time of observation.

By conducting the experiment, the minimum value of fps = 12 in the video stream was selected, at which the model based on the artificial network clearly captured the objects captured by the thermal cameras. On thermal cameras, it is especially difficult to recognize silhouettes of people, because the range on cameras is from 1 to 100 fps, and videos are processed frame by frame. That is, for a 10-second video at fps=100 and the processing time of one frame t=170-175 ms, the processing time of the video stream increases to 175 seconds. Therefore, by the formula (1) parameter N was calculated. N - interval in source video between frames that should be copied to target video, to get desired fps. Means, for example: if fps in source video is 100, and desired fps is 25 - in this case each 4-th (100/25) frame from source video to target video - 4th, 8th,12th..., should be taken, others frames should be skipped.

$$N = round\ (fps(source)/fps(target)), \quad (1)$$

A Python program was created that uses the YOLOv8 neural network and the yolov8x.pt model (the model was pre-trained on the COCO set by the network manufacturer) [24]. The classes of objects of interest, their correspondence to the classes of the COCO model, and the priority of their detection are shown in Table 4.

Table 4
The correspondence of objects of interest to the classes of the COCO model

Priority	Custom class	Classes of COCO model
1	Person	'person'
2	Vehicle	'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train', 'truck', 'boat'
3	Animal	'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe'
4	Light	'traffic light'
5	NIL	none of the above classes were detected

The operation of the algorithm in the part of detecting objects of the customer's classes in the clip can be reduced to several points:

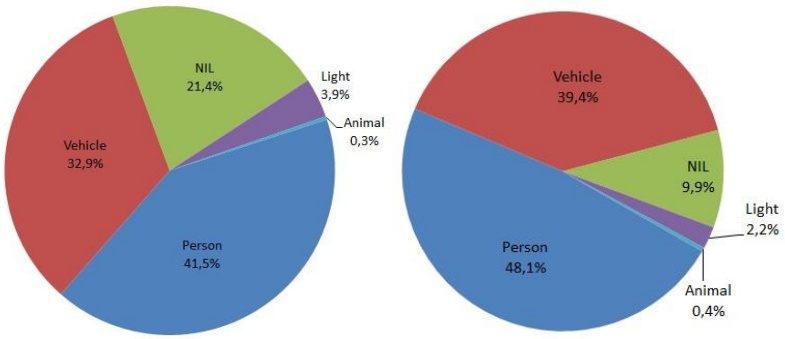
1. Objects were searched for in each frame of the video clip. Search results for each clip were saved and statistics were recorded in a table. A sample of training statistics is presented in Table 5.

2. Assigning a clip to a certain class was carried out from the highest -1 to the lowest -5 priority. That is, for example, if at least one object of priority 1 (Person) was detected in the clip, the clip belonged to this class regardless of the number of objects of lower classes. Similarly, if no objects of priority 1 were detected, the presence of objects of priority 2 was considered, etc.

Table 5
A sample of training statistics

Frame Number	Object COCO code	Object Name	Confidence
1	0	person	0.5774214267730713
1	13	bench	0.4358726441860199
2	0	person	0.5948778986930847
2	13	bench	0.484416606760025024
3	0	person	0.5487728118896484
3	13	bench	0.4547703266143799
4	13	bench	0.5063151121139526
4	0	person	0.4895791709423065
4	0	person	0.25702717900276184
5	13	bench	0.49720075726909094

The model was trained according to the classes listed in Table 1. The actual distribution of clips by classes on the example of a daily (a) and weekly (b) selection is shown in the diagrams in Figure 3.



a) Daily selection (1440 clips) b) Weekly selection (11247 clips)

Figure 3: The actual distribution of clips by classes on the example of a daily (a) and weekly (b) selection.

In order to improve the quality of recognition, it was decided to train the model for the "person" class. To create a dataset with the "human" class, the authors selected video clips from 153 cameras. Object recognition in video clips was used to mark objects in images, where the presence of people was determined by a human operator beforehand. The largest of the ready-made models from the YOLOv8 neural network developer of the Ultralitics company was used as a model: yolo8x.pt [26,27]. During the preparation of the dataset, 4908 images containing 6121 objects of the "person" class were obtained, all of which were manually checked for possible neural network errors before the start of training. The number of background images (which did not contain objects of the "person" class) - 92.

All received images were reviewed, when recognition errors were detected, label files were created for such images manually (on a local computer, using the Labelimg program). For the training dataset, 4908 images containing 6121 objects of the "person" class were obtained. The number of background images (which did not contain any objects) is 92. The percentage of detected neural network errors is 19.4%. The number of prepared images for verification is 300. Training was carried out on Google Collab data was downloaded from Google Drive. 120 learning epochs were launched. Training was started with empty weights for the yolo8s model. Tesla V-100 GPU was used, each cycle lasted about 2 min. Considering that the version "from the box" was trained on 640px images, and the size of the video from the cameras can also be 704, 1920 - the decision was made in the training parameters to use a size of 1280, to provide better detail.

5. Experiments and Results

Since the main purpose of the video surveillance conducted by the client company is the protection of objects, at this stage it was decided to additionally verify the results where people were detected (Person class). Verification was carried out by human operators. The results of videos verification in which the neural network detected people (a sample of 4837 clips over 6 days) are shown in Figure 4 in the form of a diagram. Only 0.7% of the total number of clips identified by the program as Person were confirmed by the operators. That is, the error of the second kind was 99.3%. The number of Person objects detected by the neural network in each video clip varies from 1 to 373 (it should be noted that the total number of frames in the clips is from 8 to 900). The frequency with which objects of the Person class occur in files is plotted in Figure 5.

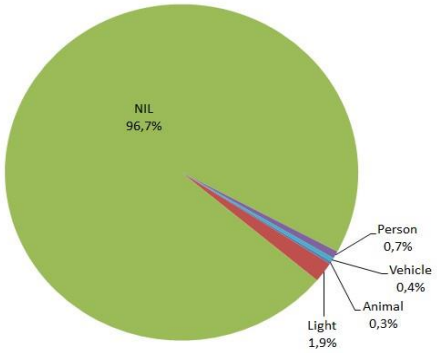


Figure 4: The results of videos verification in which the neural network detected people.

Count of Person objects detected in clips	Count of the files (video-clips)
1	1248
2	572
3	414
4	278
5	208
6	193
7	169
8	149
9	137
10	101
...	...
230	1
285	1
288	1
372	1
373	1

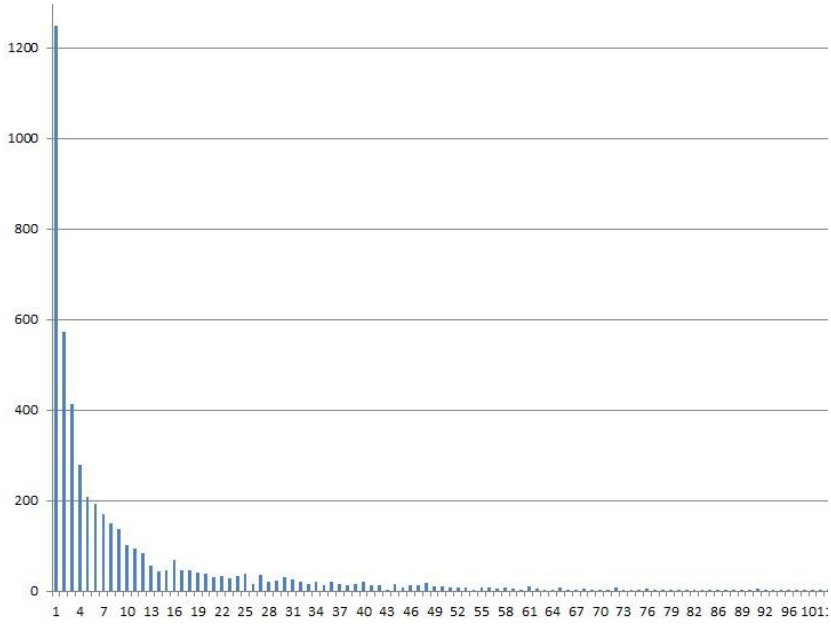


Figure 5: The frequency with which objects of the Person class occur in files.

Evaluation of training results was carried out in 2 ways: using the obtained metrics (Figures 6-8) and manually (saving data from the video using the newly created model, manually searching for erroneous results).

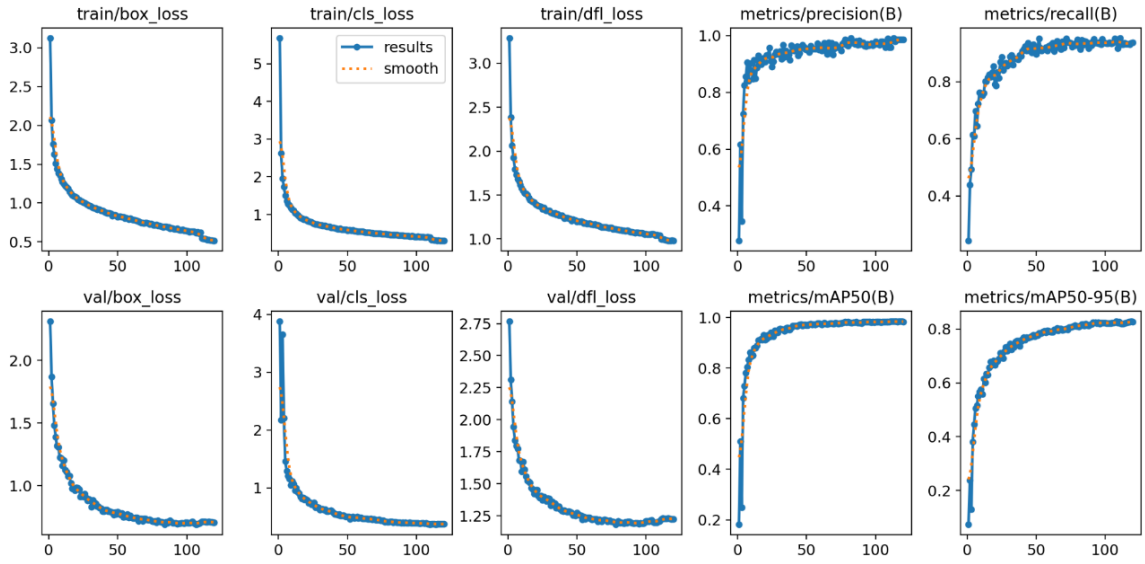


Figure 6: Metrics for training results evaluation.

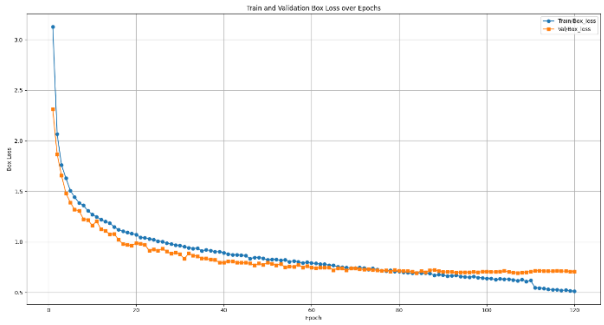


Figure 7: Train and Validation loss over epochs.

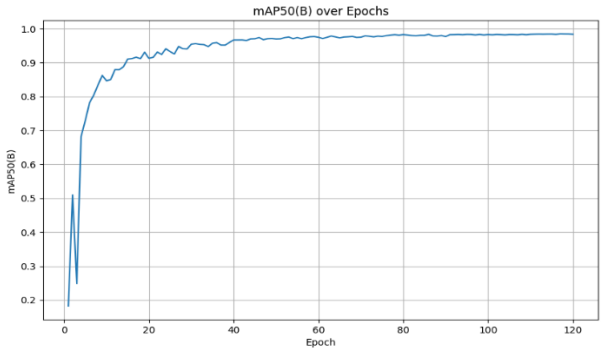


Figure 8: Mean average precision calculated at an intersection over union (IoU) threshold of 0.50.

Manual verification showed a decrease in the proportion of objects identified by the newly created model with errors to 10.7%.

As for the obtained metrics, they indicate, on the one hand, the success of training: there is an improvement in the values of the parameters Train Box Loss (up to 0.5135), mAP50 (up to 0.98367) with an increase in epochs. On the other hand, the relatively stable value of Validation Box Loss (0.69291) starting from epoch 84 suggests natural performance fluctuations due to too small a validation sample (6% of the training sample).

Further efforts will be aimed at increasing the number of objects for training to the recommended value of 10,000, the validation sample to 10-20% (1000-2000 images), the number of epochs - 300. Authors also believe that the dataset should include images from a larger number of cameras.

6. Conclusions

Therefore, in the course of work on the problem of increasing the security of the site equipped with external surveillance cameras, it was decided to develop a model for post-processing of alarm signals using a neural network. For this, an analysis of scientific publications over the past 5 years was conducted and the most effective pre-trained models that provide the highest recognition results of various objects were investigated. An experiment was also conducted by testing three models on a custom dataset of 12 images,

during which it was determined that the YOLOv8 neural network model currently gives the best result. For the security of the site, it was determined that it is most important to recognize a person in the images from the cameras. Therefore, the YOLOv8 neural network was pre-trained on the COCO dataset. The results of experiments on the recognition of people in the video stream made it possible to achieve higher efficiency.

Also, manual verification revealed a notable decrease in identification errors to 10.7% with the newly developed model. Training metrics, including Train Box Loss and mAP50, demonstrated improvement, while Validation Box Loss remained relatively stable.

Future efforts will focus on augmenting the training dataset to 10,000 objects, expanding the validation sample to 10-20%, increasing epochs to 300, and incorporating images from a broader range of cameras to enhance dataset diversity.

References

[1] O. Pavlova, V. Kovalenko, T. Hovorushchenko, Neural network-based image recognition method for smart parking. Comput. Syst. Inf. Technol. J. 1 (2021) 49–55.

[2] Y. Zeng, J. Zhang, A machine learning model for detecting invasive ductal carcinoma with Google Cloud AutoML Vision. Computers in biology and medicine. J. 122 (2020)

[3] Y. Lu, Y. Jia, J. Wang, B. Li, W. Chai, L. Carin, Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. CVF conference on Computer Vision and Pattern Recognition. 2020. pp. 940-949.

[4] P. Radiuk, O. Pavlova, N. Hrypynska, An Ensemble Machine Learning Approach for Twitter Sentiment Analysis. Proceedings of the 3d International Workshop, volume 3171, 2022, pp. 387–397.

[5] J. P. Sahoo, A. J. Prakash, P. Pławiak, S. Samantray, Real-time hand gesture recognition using fine-tuned convolutional neural network. Sensors. J. 22 (2020).

[6] M. Gazda, M. Hireš, P. Drotár, Multiple-fine-tuned convolutional neural networks for Parkinson's disease diagnosis from offline handwriting. IEEE Transactions on Systems, Man, and Cybernetics. Systems. J. 52(1), (2020) 78-89.

[7] P. Radiuk, O. Pavlova, H. El Bouhissi, V. Avsiyevych, V. Kovalenko, Convolutional Neural Network for Parking Slots Detection. CEUR Workshop Proceedings, 3156, 2022, pp. 284–293.

[8] S. Bussa, A. Mani, S. Bharuka, S. Kaushik, Smart attendance system using OpenCV based on facial recognition. Int. J. Eng. Res. Technol, 9 (2021), 54-59.

[9] A. P. Ismail, F. A. Abd Aziz, N. M. Kasim, K. Daud, Hand gesture recognition on python and OPENCV. IOP Conference Series: Materials Science and Engineering, Vol. 1045, No. 1, IOP Publishing, 2020, p. 012043.

[10] M. Kushal, M. Pappa, ID Card Detection with Facial Recognition using Tensorflow and OpenCV. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 742-746.

[11] A. Sharma, K. Shah, S. Verma, Face recognition using Haar cascade and local binary pattern histogram in OpenCV. In 2021 Sixth International Conference on Image Information Processing (ICIIP), Vol. 6, pp. 298-303.

[12] Tu-Liang Lin, Hong-Yi CHANG, Kai-Hong CHEN, The pest and disease identification in the growth of sweet peppers using faster R-CNN and mask R-CNN. Journal of Internet Technology, 21(2020) 605-614.
[13] T. Mahmood, M. Arsalan, M. Owais, M. Lee, K. Park, Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. Journal of clinical medicine, 9 (2020) 749-759.
[14] N. Palanivel, Automatic number plate detection in vehicles using faster R-CNN. 2020 International conference on system, computation, automation and networking (ICSCAN), 2020. p. 1-6.
[15] S. WAN, S. GOUDOS, Faster R-CNN for multi-class fruit detection using a robotic vision system. Computer Networks. J, 168 (2020)
[16] D. Al-Obidi, S. Kacmaz. Facial Features Recognition Based on Their Shape and Color Using YOLOv8. 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2023.
[17] M. Safran, A. Alajmi, S. Alfarhood, Efficient Multistage License Plate Detection and Recognition Using YOLOv8 and CNN for Smart Parking Systems. Journal of Sensors, 2024.
[18] B. Xiao, Y. Nguyen, W. Yan, Fruit ripeness identification using YOLOv8 model. Multimedia Tools and Applications. J. 2023, 1-18.
[19] A. Chabi, E. Mahama, A. Gouton, P. Tossa, Automatic localization of five relevant Dermoscopic structures based on YOLOv8 for diagnosis improvement. Journal of Imaging, 9 (2023) 148-159.
[20] O. Melnychenko, O. Savenko, P. Radiuk, Apple Detection with Occlusions Using Modified YOLOv5-v1, 2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Dortmund, Germany, 2023, pp. 107-112, doi: 10.1109/IDAACS58523.2023.10348779.
[21] O. Melnychenko, O. Savenko. A Self-Organized Automated System to Control Unmanned Aerial Vehicles for Object Detection. The 4th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS-2023): CEUR-Workshop Proceedings, volume 3373, 2023, pp.589-600
[22] Vision AI. Google Cloud. URL: https://cloud.google.com/vision
[23] PyTorch Faster R-CNN. URL: https://pytorch.org/vision/stable/models/faster_rcnn.html
[24] COCO. Common Objects in Context. URL: https://cocodataset.org
[25] Tips for Best Training Result URL: https://docs.ultralytics.com/yolov5/tutorials/tips_for_best_training_results/
[26] Model Training with Ultralytics YOLO. URL:https://docs.ultralytics.com/modes/train/
[27] Train YOLOv8 on Custom Dataset – A Complete Tutorial. URL:https://learnopencv.com/train-yolov8-on-custom-dataset/
[28] Adding background images to a classification dataset. URL: https://github.com/ultralytics/ultralytics/issues/4479 (last accessed February 1, 2024)