UDC 004.032.2 https://doi.org/10.15407/vidbir2022.50.069

APPLICATION OF YOLOX DEEP LEARNING MODEL FOR AUTOMATED OBJECT DETECTION ON THERMOGRAMS

I. O. Skladchykov, A. S. Momot, R. M. Galagan, H. A. Bohdan, K. M. Trotsiuk

Igor Sikorsky Kyiv Polytechnic Institute, Kyiv

E-mail: skladchykov.vania@gmail.com, drewmomot@gmail.com

A method of automating the data analysis of thermal imaging systems in the field of safety control is proposed. It has been established that today video surveillance technologies have a number of disadvantages that can be eliminated by using thermal imaging cameras. Analysis of infrared images can be automated in order to reduce percentage of false positives and increase the efficiency of thermal imaging video surveillance systems. A high level of interference, unclear object contours and low image resolution are real problems in automating the object detecting process on thermographic images. The traditional and promising methods of thermograms analysis and approaches that can lead to creating the automated thermal video surveillance systems are discussed. It is proposed to use deep learning, which in recent years has proven itself as an effective way of image analysis. The study is based on review of existing works, as methods of automating the object detection process on thermograms. It is proposed to use YOLOX as a deep learning model, which has one of the best quality indicators and speed processing input parameters on standard datasets. FLIR's Thermal Starter annotated set of thermal images is used to train the model, which value of mAP at the level of 55% is obtained according the results of model training for recognizing 4 classes of objects on thermograms. Different advantages and disadvantages of this development are analyzed. Ways of further improvement of the neural network method of automation of thermal imaging safety control systems have been determined.

Keywords: *thermal monitoring, deep learning, object detection.*

ЗАСТОСУВАННЯ МОДЕЛІ ГЛИБИННОГО НАВЧАННЯ YOLOX ДЛЯ АВТОМАТИЗОВАНОГО ДЕТЕКТУВАННЯ ОБ'ЄКТІВ НА ТЕРМОГРАМАХ

І. О. Складчиков, А. С. Момот, Р. М. Галаган, Г. А. Богдан, К. М. Троцюк

Київський політехнічний інститут ім. Ігоря Сікорського, Київ

Запропоновано метод автоматизації аналізу даних тепловізійних систем у галузі контролю безпеки. Встановлено, що на сьогодні технології відеоспостереження мають низку недоліків, яких можна позбутись, використовуючи тепловізійні камери. Для зниження відсотків хибних спрацювань та підвищення ефективності тепловізійних систем відеонагляду аналіз інфрачервоних зображень можна автоматизувати. Недоліком в автоматизації детектування об'єктів на термографічних зображеннях є високий рівень завад, нечіткі контури об'єктів, низька роздільна здатність зображень. Розглянуто традиційні та перспективні методи аналізу термограм та підходи до створення автоматизованих систем теплового відеонагляду. На основі огляду існуючих праць як метод автоматизації детектування об'єктів на термограмах запропоновано використовувати глибинне навчання, яке за останні роки зарекомендувало себе як ефективний засіб аналізу зображень. За модель глибинного навчання запропоновано вживати YOLOX, яка має одні з найкращих показників якості та швидкості оброблення вхідних параметрів на стандартних наборах даних. Для навчання моделі використано анотований набір теплових зображень Thermal Starter від компанії FLIR. За результатами навчання моделі для розпізнавання чотирьох класів об'єктів на термограмах отримано значення mAP на рівні 55%. Проаналізовано переваги та недоліки цієї розробки. Визначено шляхи подальшого вдосконалення нейромережевого методу автоматизації тепловізійних систем контролю безпеки.

Ключові слова: *тепловізійний відеонагляд, глибинне навчання, детектування об'єктів.*

Introduction. Nowadays video surveillance technologies are used almost everywhere in different ways. They are especially widely used in the field of security control.

In recent years, it has already become increasingly common practice to rely on thermal imaging equipment to achieve the best results for perimeter protection, object identification, and other similar security tasks. Thermal imagers give a “different” view of the world. Interpreting thermal images is not easy. As a result, human operator becomes an even more critical element in the threat detection chain. A human in a thermal surveillance system imposes limitations that must be taken into account. The human eye is not perfect and has limitations on the amount of information it can perceive [1, 2]. In addition, it is difficult for a human to provide high-quality continuous supervision of the territory. Therefore, the task of automating data analysis of thermal imaging video surveillance systems is becoming more and more relevant.

Analyzing thermal images for meaningful information is a challenging task. Therefore, digital processing of thermograms is considered as a very important area of research for their automated analysis and interpretation. There are many traditional approaches to perform image processing, but artificial intelligence plays an important role in performing automation [3].

In recent decades artificial intelligence methods have firmly established themselves as possible alternative math tools. They are widely used in many information systems. Special attention should be paid to deep learning methods, which are most effective in signal processing, automatic control and image analysis That is why the use of artificial intelligence is making great interest as a means of automating object detection on thermal images. In particular, the usage of deep learning will reduce the frequency of false alarms and increase the overall reliability and efficiency of thermal imaging safety control systems, minimizing human role in decision-making process [4].

Review of previous studies. Thermal security cameras work reliably in places with insufficient lighting and poor visibility. Fig. 1 shows a comparison of visible and infrared images. Image courtesy by FLIR company. Looking at the image we can see that no important details are visible in visible spectrum. In particular, it is impossible to recognize a person due to insufficient lighting. This defect is not present on the thermogram.

Fig. 1. Comparison of the image in visible (*a*) and infrared (*b*) spectrum.

Another advantage of thermal surveillance is the ability to ignore most of the visual camouflage. For example, thick layer leaves can often be found near offices and warehouses. In addition, thermal surveillance cameras can be equipped with intelligent sensors and advanced analytics technology, which will help to reduce the number of false alarms. Finally, thermal imaging-based systems are often cheaper to install and operate in the long term [5].

In study [6], the authors considered the history of development and latest achievements in the field diagnostics of malfunctions of power equipment based on infrared

thermography. This work indicates that the rapid increase in the amount of equipment in electrical networks requires the replacement of human labor with automatic and intelligent technologies. With appearance of deep networks, intelligent identification of equipment defects on thermal images has become increasingly popular recently. This method, using a training data set, provides fully automatic diagnostic detection features, without human intervention. The small resolution of infrared images is a main disadvantage of such a system. This prevents widespread use of intelligent diagnostics of energy equipment using thermal imagers. It is proposed to create an open infrared images database to solve this problem. It will lead to an improvement in the efficiency algorithms for processing thermographic images using artificial intelligence methods.

The authors of the work [7] conducted an experiment with simulation of an active thermal field. Infrared images of the object with a high level of interference were obtained. This work considered the usage of various methods for processing thermograms, such as: wavelet analysis, principal component analysis, and artificial neural networks. It has been established that the methods of digital thermograms processing allow us to improve the quality of an image compared to optimal thermogram, to increase the signal-to-noise ratio and, as a result, the reliability of testing. Modeling showed that the main problem of most methods is low noise-resistance. The use of neural networks for thermal field data analysis showed higher efficiency compared to the principal component analysis method. The effectiveness of deep learning has been experimentally proven, which is confirmed by quantitative characteristics.

Work [8] describes a cost-effective solution for using a wireless infrared sensor device that can be used in intelligent systems for protection of private areas. This approach uses a new high-resolution infrared sensor and implements the concept Internet of Things (IoT) architecture, which is a goal of Industry 4.0. Authors of this development propose to create a network of IoT devices to monitor physical parameters in a smart house and control the security of the territory. It is proposed to develop appropriate automated notification and response programs based on deep learning. This approach has proven itself as an effective way to detect objects in images in automatic mode. In particular, this development can be useful when it is necessary to detect presence of person by thermal imaging systems.

The research [9] examines the prospects of using thermal imaging systems in safety control tasks. Ways of developing and increasing efficiency of thermal imaging systems are analyzed. Along with the improvement of technical parameters of heat engineering equipment, automation of thermal data analysis is a promising direction. Due to a number of advantages, the usage of convolutional neural networks is proposed as a method of automating the thermal images processing. It was decided to artificially increase the data volume by applying augmentation in order to improve the performance system. The approach described by the authors made it possible to automatically detect and recognize the object class on thermal images with an accuracy of 97.92%. The disadvantage of this system is a detection of only one object in the image, which is impractical for use in thermal video surveillance systems, because more than one important object for recognition can be located on secure territory at the same time.

As it is known, there is a large number of deep learning models [10]. The ways to optimize the neural network architecture are limitless. It leads to the task of analyzing the possibility of using different types of deep learning models in thermal imaging security control systems. The speed and quality of a deep learning model is becoming increasingly important for computer vision. The paper [11] presents various variants of neural network architecture for object detection called EfficientNet. Several key architecture optimizations are proposed to improve the efficiency of the models. First, a profound bidirectional feature pyramid network (BiFPN) is proposed, which allows easy and fast multi-scale feature fusion. Next a comprehensive scaling method is created. It

uniformly scales the resolution, depth, and width of the model. Increasing the scale of the model in any one dimension (width, depth, resolution) can improve accuracy, but when the model becomes too large, the improvement in accuracy is not obvious.

The latest modification, named EfficientDet, is designed for object detection in images and consists of EfficientNet as a base, to which a BiFPN functional pyramidal feature detection unit is added. This network shows a mean Average Precision (mAP) of up to 46% at a data processing rate of about 35 frames per second on the MS COCO dataset. Such characteristics are a good combination of speed and accuracy of object recognition. The disadvantage of this modification is a large number of functional blocks. Accordingly, they lead to an increase in the number parameters. This model requires significant resources to process the input parameters, increasing the requirements for computing power and data transfer speed in the thermal video surveillance system.

Objectives of research. Based on the studies mentioned above, it can be assumed that the detection of objects in the infrared range will be more informative, qualitative and practical than in the visible spectrum. It can be concluded that thermal imaging cameras provide powerful new evaluation capabilities for video surveillance systems. Operating individually or in combination with video surveillance cameras, thermal imagers give security operators much more data to identify and track intruders in a protected area. However, today there is a problem in choosing an effective method of automating the process of thermal image analysis, and especially – object detection.

To develop an automated object detection system in order to solve this problem the use of deep learning is proposed. This should lead to the increase in the informativeness and reliability of the operation of thermal video surveillance systems, as well as to the reduction of the influence of the system operator on decision-making. The purpose of this work is to implement a deep learning model for detecting objects on thermal images. Taking into account the proven researches, it is planned to create a software for automated analysis data from infrared cameras used for security control. In order to reduce the percentage of false positives, such a system should detect objects in images with high reliability, have significant noise-resistance and high operation speed. This requires choosing the optimal deep learning model and using a representative dataset.

Description of the training dataset. To train the deep learning model, it is suggested to use Thermal Starter dataset provided by the FLIR company. This is a ready-made annotated set of thermal images for training and validating neural networks for object detection. The image was obtained using a thermal imaging camera installed on a car. The dataset contains a total 14,452 annotated thermal images, including 10,228 images taken from short videos and 4,224 images from a continuous 144-second video. The resolution of thermal images is 640×640 pixels. The videos were shot under the normal clear sky conditions, both during a day and a night. Human annotators marked four categories of objects, namely: person, bicycle, cars and dogs. The MS COCO label vector was used for numbering the classes [12].

Annotators made bounding boxes around objects as tight as possible. Tight bounding boxes that let through small parts of the subject, such as limbs, were preferred over wider boxes. Personal accessories were not included in the restrictive bounding boxes on people. Heads and shoulders were a higher priority for boxing than other body parts of humans and dogs. Minimal license plate blur was applied to all images to make them illegible.

Obtained dataset of experimental images was divided into two subsets, namely training (8862 samples) and validation (1366 samples). Training set is actually used for training the network; validation set serves to select the hyperparameters of the network in the learning process.

Implementation of deep learning model. Python programming language is used for effective work with neural networks. The advantage of this language is primarily

the existence of a large number of libraries that have a wide set of tools for creating artificial intelligence models and data analysis. Using this language, it is much easier to cope with the tasks of image analysis and visualize analyzed data. The PyTorch framework was chosen to implement the deep learning model. This framework includes many different tools for creating neural networks. Compared to other tools, models based on PyTorch have a higher operating speed.

An urgent and important task is to determine architecture of deep network, which will allow us to detect the object with the greatest reliability. Since modern infrared cameras do not have a high frame rate, the calculation speed of the deep learning model takes a back seat. However, for real-time monitoring, data processing frequency should not be lower than 20–25 frames per second. This fact must be taken into account when choosing a suitable model for object recognition. In addition, the task of object detection in infrared images is complicated by the low detail thermal imprints of these objects, which is related to physical features of nature infrared radiation. Therefore, the deep learning model should be generally generalizable and have high adaptability.

To assess the quality of object detection models, mAP metric is used – an indicator of the average correctness of recognition of various object classes. This indicator calculates the average probability of correct answers in the range from 0 to 1 for all object classes that the model can recognize. For each class, average accuracy of recognition (Average Precision) is defined as the area under Precision-Recall curve. Higher the value of this indicator, fewer false recognitions model performs on the test data.

Today, YOLO (You Only Live Once) is considered as one of the most promising models for object detection. Its main feature is object detection in one data pass. There are no explicit loops in the YOLO architecture, which makes network fast. YOLO uses a grid of predefined windows – areas in which objects are classified. On the MS COCO dataset, modern YOLO modifications show up to 51.2% mAP at a data rate of up to 60 frames per second. Networks with this architecture are among the fastest in the object detection task, which makes them promising for use as part of real-time thermographic systems.

This model has many different modifications, such as YOLOv5, YOLOX and others. The work [13] describes in detail the difference between them, their advantages, disadvantages and main parameters. Having analyzed all aspects, YOLOX-M model was chosen, which is considered balanced in terms of the average accuracy of object recognition, speed and a number of model parameters. YOLOX object detector is a very interesting addition to YOLO family.

YOLOX detector, released in July 2021, switched to an anchor-less approach, which differs from previous YOLO networks. This model includes 25.3 million parameters. On the test dataset, mAP indicator of various object classes from the MS COCO set is 46.4%. YOLOX is prospective for several years ahead, as it allows object recognition on video with a frequency of 81 frames per second. This is even an excess – the frame rate in the most modern thermal imaging systems does not exceed 40 frames per second, and human eye does not need more than 60 frames per second for comfort. Fig. 2 shows a simplified general architecture of YOLOX-M neural network head.

Authors of [14] have made prototypes of standard YOLOX models freely available and presented small aspects for working with it. Analyzing architecture, it can be seen that a fully connected convolution is first used to reduce the feature extraction pyramid (FPN) based channel width to 256, and then two parallel branches with two fully connected convolutions each are added to solve classification tasks (Cls.) and regression (Reg.) tasks in the form of bounding box prediction around objects, respectively. IoU branch is added to regression and used to estimate presence of object in the predicted bounding box. IoU (Intersection-Over-Union) parameter is a metric used to assess reliability of bounding box detection.

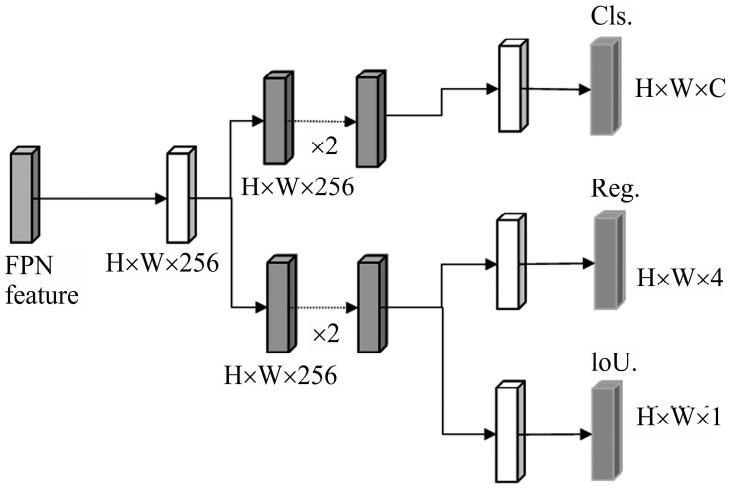

Fig. 2. Architecture of the head YOLOX-M network.

During training, the loss functions cls_loss (measures correct classification of each predicted bounding box: each box can contain an object class or “background”) and total_loss (total loss of the model) [15] were used. The loss functions on training set are presented in Fig. 3.

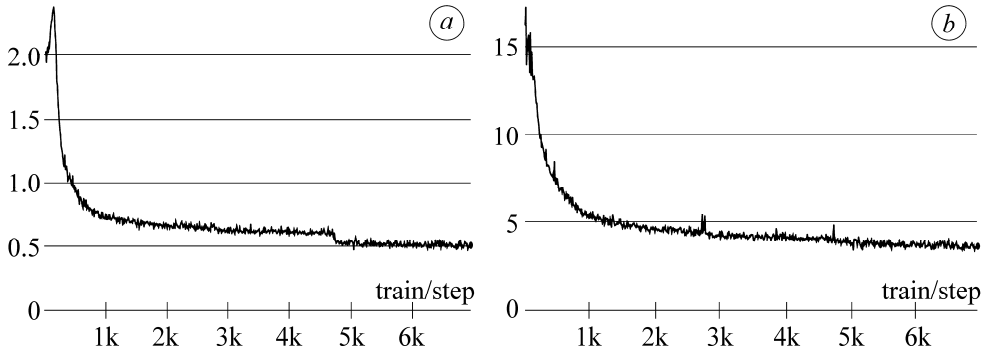

Fig. 3. Loss function by classes (*a*) and overall loss function (*b*) on the training data set.

For this model, values of width 0.75 and depth 0.67 were experimentally selected, which achieved most optimal results in terms of general quality criterion. The total number of training objects presented in one batch was equal to 64, and training took place on one core of system for 50 epochs. The “SiLU” activation function was used in this development.

The YOLOX-M model uses data augmentation techniques to improve training outcomes. With help of augmentation, each training image was randomly modified in order to increase the representativeness of dataset [16]. The following augmentation parameters are selected: flip probability = 0.5; degress = 10.0; translate = 0.1; mosaic scale = (0.1, 2); mixup scale = (0.5, 1.5); shear = 2.0.

A graph of training results on the validation dataset is shown in Fig. 4. As can be seen, a larger number of epochs would lead to overtraining of the model. The mAP value on validation sample reaches a maximum of 0.55 at IoU of 50%. The processing speed of one image was 22 ms, which corresponds to a frequency of about 45 frames per second.

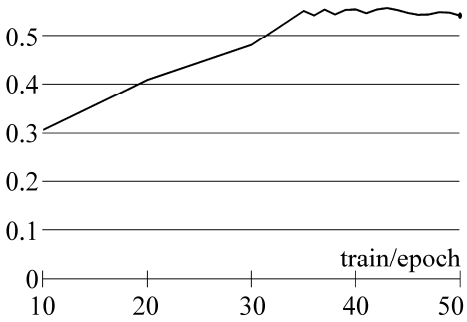

Fig. 4. Value of mAP on the validation sample.

Results. An example of the model operation is shown in Fig. 5. It can be seen that the system detects existing objects in the image quite well. In particular, positions and boundaries of cars and people present in the frame are correctly defined. At the same time, one false positive can be noticed, since there are actually only 3 people in the image, not 4. However, this can be filtered by removing all frames with objects whose detection accuracy is below some threshold.

It is worth noting that for the naked human eye it is difficult to immediately see details in thermal image. According to subjective estimates, time to recognize all objects on a given frame would be more than 22 ms, for which the model solved this task. This once again confirms the importance to use an automated data analysis for increasing the effectiveness of object detection.

Fig. 5. An example of the program operation on test data.

After training the model, quality of object detection was assessed using various metrics on the validation set. Table 1 shows the values of the obtained metrics. It can be concluded that the best results according to Average Precision indicator are achieved with an IoU value of 0.50. Mean of Average Precision in the range of IoU from 0.50 to 0.95 is lower, because of increasing requirements for reliability in bounding boxes detection around objects leads to missing data. In a security control task, missing an object has more negative consequences than false detection. This should be taken into account when setting IoU threshold during model setup.

Table 2 shows the Average Precision and Average Recall values for each object class. After analyzing data, we can pay attention to relatively small values of metrics for the dogs and bicycle classes. This drawback is associated with a small number of thermal images of these classes in the training data set. That is, classes were not properly balanced by the authors of the dataset. This can be explained by the fact that dogs

and bicycles are less common on city streets than cars and pedestrians. When using a larger volume of representative data for model training, the quality of the indicators system can be significantly improved. This may be an important task for further developments.

Table 1. Average Precision and Average Recall of the model

Metric	IoU	Result
Average Precision	0.50	0.539
	0.75	0.276
	0.50:0.95	0.430
Average Recall	0.50:0.95	0.533

Table 2. Average Precision and Average Recall by classes

Class	Average Precision, %	Average Recall, %
person	55.11	59.40
bicycle	27.80	46.54
cars	65.95	70.01
dogs	12.30	31.29

This model showed a better result of object class recognition than YOLOX-M presented in [13]. In the mentioned study, a set of images of visible spectrum was used. For comparison, the mAP value of the mentioned work is 0.464 at IoU = 0.5, while for thermal images this indicator is 0.539. As a conclusion, the use of thermal images to detect objects on city streets is more effective. This is explained by greater contrast of thermal objects and independence of result on the illumination level. Object detection in the visible spectrum is faster than in infrared (11 ms vs. 22 ms, respectively). However, for modern thermal imaging systems, such a difference is not significant, especially considering obtained advantage in detection reliability.

Disadvantage of this system is long training or retraining of the model for determination of the specific objects classes, which requires large computing resources. The large weight of model does not allow it to be used directly on the basis of thermal imager itself. This leads to the need to use of cloud computing or transfer data to a separate device for further processing.

A small training database can lead to low-quality training of the network and erroneous operation system, as well as omission of some object classes in the image. To solve this issue, it is promising to create a single database of annotated thermal images in MS COCO format.

CONCLUSIONS

The advantages of using thermal imaging systems over video cameras visible spectrum in video surveillance technologies were investigated. The current situation of thermal video surveillance systems was analyzed and directions for their improvement were determined. Taking into account the technological development of thermal imaging devices, automation of object detection on thermal images is a promising direction. For this, a deep learning model was developed for the automated object detection in thermal images on training dataset. To improve the efficiency of object detection among existing deep learning models, the latest YOLOX-M architecture was chosen. This model has a high speed (up to 45 frames per second) and mAP value of 0.539 at IoU = 0.5 based on experimental data. The proposed software system showed a possi-

bility of using thermal images to increase reliability of object detection in comparison with similar models that work with images in the visible spectrum.

In the nearest future it is possible to use a larger database training data to obtain better system of quality indicators. It has been confirmed that the use of intelligent data analysis systems in thermal imaging security control systems allows us to improve speed of threat recognition, reduce proportion of false positives due to subjective factor. It will be also possible to avoid the need for round-the-clock monitoring of the territory by a human. Optimization of existing and development of new deep learning models for solving the tasks of automated object detection on thermal images is a promising direction for development of security systems in the coming years.

1. Wong, L.; Wai, K. An effective surveillance system using thermal camera. In IEEE 2009 International Conf. on Signal Acquisition and Processing, Kuala Lumpur, Maylaysia, 3–5 April 2009, pp. 13–17. https://doi.org/10.1109/ICSAP.2009.12
2. Alokhina, O.V.; Ivchenko, D.V.; Pits, N.A. Thermal remote sensing data analysis in monitoring of natural objects. *Information Extraction and Processing*, **2020**, 48(124), 61–71. https://doi.org/10.15407/vidbir2020.48.061
3. Muraviov, O.V.; Petryk, V.F.; Lysenko, I.I.; Bohdan, H.A.; Nakonechna, A.V. Automatization of thermographic diagnostic method of human body pathologies. *Taurida Scientific Herald. Series: Technical Sciences*, **2022**, 1, 47–53. https://doi.org/10.32851/tnv-tech.2022.1.5 (in Ukrainian)
4. Kosarevych, R.Ya.; Alokhina, O.V.; Rusyn, B.P.; Lutsyk, O.A.; Pits, N.A.; Ivchenko, D.V. Analysis of remote sensing images by methods of convolutional neural networks and marked random point fields. *Information Extraction and Processing*, **2021**, 49(125), 45–51. https://doi.org/10.15407/vidbir2021.49.045
5. Zhivkovic, A.; Muravyov, A. Modern technologies of non-contact temperature measurement. Nauka i studia, 2020. (in Russian)
6. Xia, C.; Ren, M. Infrared thermography based diagnostics on power equipment: State of the art. *High Voltage*, **2020**, 6, 387–407. https://doi.org/10.1049/hve2.12023
7. Galagan, R.; Momot, A. Analysis of application of neural networks to improve the reliability of active thermal NDT. *KPI Science News*, **2019**, 1, 7–14. https://doi.org/10.20535/kpi-sn.2019.1.157374
8. Skladchikov, I.; Momot, A. Using the MLX90640 sensor as part of a smart thermographic camera. In *Efficiency and automation of engineering solutions in instrument construction*, **2020**, 16, 240–242. (in Ukrainian)
9. Momot, A.; Skladchykov, I. Deep learning automated data analysis of security infrared cameras. *Slovak International Scientific Journal*, **2021**, 52, 13–16.
10. Trask, A. Grokking Deep Learning. Manning Publications, 2019.
11. Mingxing, T.; Ruoming, P. EfficientDet: Scalable and Efficient Object Detection. *IEEE Xplore*, **2021**, 10781–10790. https://doi.org/10.48550/arXiv.1911.09070
12. FLIR Thermal Images Dataset. 2021. https://www.kaggle.com/datasets/albertofv/flir-thermal-images-dataset-reduced (accessed 2022-07-08)
13. Zheng, G.; Songtao, L. Exceeding yolo series in 2021. arXiv preprint arXiv, 2021, # 2107.08430. https://doi.org/10.48550/arXiv.2107.08430
14. Zheng, G.; Songtao, L. Exceeding yolo series in 2021. 2021a. https://github.com/Megvii-BaseDetection/YOLOX. (accessed 2022-07-08)
15. Storozhyk, D.V.; Muraviov, O.V.; Protasov, A.G.; Bazhenov, V.G.; Bohdan, G.A. Multispectral image combining as a method of information content increasing at binary segmentation. *KPI Science News*, **2020**, 82–87. https://doi.org/10.20535/kpi-sn.2020.2.197955 (in Ukrainian)
16. Cao, J.; Wu, W.; Wang, R.; Kwong, S. No-reference image quality assessment by using convolutional neural networks via object detection. *International Journal of Machine Learning and Cybernetics*, **2022**, 1–12. https://doi.org/10.1109/ICDSP.2016.7868646

Received 06.09.2022