https://doi.org/10.31891/csit-2024-4-19
UDC 004.93.1

Xia GUANXIANG
Vinnytsia National Technical University
Hunan Mass Media Vocational and Technical College
Viacheslav KOVTUN
Vinnytsia National Technical University

THE MODEL OF THE SYSTEM FOR OBJECTS RECOGNITION IN THE REAL-TIME VIDEO STREAM

This study addresses the development of a robust object recognition system tailored for real-time video streams. With the increasing integration of mobile devices in diverse applications, the research focuses on leveraging temporal and spatial data inherent in video streams to mitigate challenges such as environmental noise, preprocessing defects, and algorithmic errors. The proposed system incorporates dynamic models and convolutional neural networks (CNNs) to enhance recognition accuracy. Experimental evaluations using various datasets demonstrate the efficacy of combining classifier outputs and applying integration strategies suited for mobile platforms. The findings have practical implications for automated document processing, security systems, and mobile technology advancements, contributing to the broader field of computer vision..

Keywords: real-time video stream, classifier integration, mobile platforms, computer vision, environmental noise, dynamic models, automated systems.

Ся ГУАНСЯНГ
Вінницький національний технічний університет
Хунанський коледж масових медіа
В’ячеслав КОВТУН
Вінницький національний технічний університет

МОДЕЛЬ СИСТЕМИ РОЗПІЗНАВАННЯ ОБ’ЄКТІВ У ВІДЕО ПОТОЦІ РЕАЛЬНОГО ЧАСУ

У статті досліджується процес розробки надійної системи розпізнавання об'єктів, адаптованої для роботи з відео потоками реального часу. З огляду на зростаючу інтеграцію мобільних пристроїв у різні сфери людської діяльності, дослідження зосереджено на використанні часових і просторових даних, притаманних відео потокам, для подолання таких викликів, як шум навколишнього середовища, дефекти попередньої обробки та алгоритмічні помилки. Запропонована система інтегрує динамічні моделі та згорткові нейронні мережі для підвищення точності розпізнавання. Експериментальна оцінка з використанням різних наборів даних демонструє ефективність комбінування результатів класифікаторів та застосування інтеграційних стратегій, оптимізованих для мобільних платформ. Результати мають практичне значення для автоматизованої обробки документів, систем безпеки та вдосконалення мобільних технологій, сприяючи розвитку галузі комп'ютерного зору.

Ключові слова: відео потік реальному часу, інтеграція класифікаторів, мобільні платформи, комп'ютерний зір, шум навколишнього середовища, динамічні моделі, автоматизовані системи.

THE PROBLEM STATEMENT IN GENERAL FORM AND ITS CONNECTION WITH IMPORTANT SCIENTIFIC OR PRACTICAL TASKS

The rapid integration of mobile devices and technologies into technological, social, and commercial processes has fundamentally transformed the landscape of real-time object recognition systems [1]. These systems are increasingly leveraging the capabilities of mobile devices, such as smartphones and tablets, to replace traditional stationary recognition systems. A significant driver of this transformation is the need to develop technical vision systems that can operate under the hardware constraints of mobile platforms, making the development of such technologies a critical and timely challenge. Traditional object recognition systems rely on individual photographs or scanned images to analyze and recognize objects [2]. While effective in controlled environments, these systems face substantial limitations when applied to real-world scenarios involving variable lighting, object motion, and environmental noise. The use of video streams as a digital representation of objects offers an innovative alternative, providing access to a sequence of observations that capture richer visual information compared to static images. One pressing challenge in this domain involves addressing recognition errors caused by algorithmic imperfections, preprocessing defects, and environmental noise [3]. These errors are magnified in single-frame recognition approaches, where even minor distortions or changes in input can lead to incorrect classification. Conversely, video streams inherently mitigate some of these limitations by capturing multiple frames, allowing for the aggregation of visual information across observations. This approach can enhance robustness against noise and preprocessing defects, enabling more accurate recognition of complex objects such as holographic security elements or documents with reflective surfaces.

The development of object recognition systems capable of leveraging video streams requires innovative strategies for combining the results of frame-by-frame recognition. Existing research primarily focuses on merging outputs from heterogeneous classifiers, leaving a gap in strategies tailored to sequential observations in video

streams [4]. Addressing this gap involves designing algorithms that integrate temporal and spatial information to optimize recognition accuracy while minimizing computational overhead – a crucial consideration for mobile device implementation. Considering the advancements in convolutional neural networks (CNNs), which have demonstrated human-level performance in certain recognition tasks, integrating these technologies with video-based object representation holds immense potential [5]. However, CNNs remain susceptible to instability caused by minimal input variations, emphasizing the need for systems that capitalize on the redundancy and variability inherent in video streams.

The proposed research addresses the development of a robust system for object recognition in real-time video streams, focusing on reducing errors associated with noise, preprocessing defects, and algorithmic limitations. This effort has significant implications for practical applications, including automated document input, identity verification, and security systems, as well as broader scientific contributions to the fields of computer vision and mobile technology. By advancing the capabilities of technical vision systems, this research aims to enable seamless integration of intelligent recognition technologies into everyday life, enhancing efficiency and accessibility across diverse domains.

THE PROBLEM STATEMENT AND REVIEW OF RECENT RESEARCH

Consider the model of a system for recognizing a single object ρ. Let there be a set containing N classes $K = \{k_1, k_2, \ldots, k_N\}$. For example, when considering the task of recognizing individual characters in the "Surname" field of a Ukraine citizen's passport, the set of classes consists of the Ukrainian alphabet with the addition of space and hyphen characters. In the task of document page typization in an image after localizing its boundaries and projective correction, the set of classes can represent a collection of document page types available for further processing. It is worth mentioning separately that sometimes in object and phenomenon recognition tasks, the presence of an "empty class" is allowed, which should be the system's response to an input image of an object unknown to the system or to an image that does not contain an object.

Let an image of an object $P(\rho)$ from a certain set of all possible images P be given, and within the framework of the interaction model between the recognition system and the user/operator (or with other system components), there exists a class $k^*(\rho) \in K$ to which the object ρ belongs. The task of recognizing a single object image consists of determining this class. The result of the recognition system's work can generally be represented as a well-defined mapping from the set of classes K to the set of membership scores $\hat{f}: K \to \mathbb{R}$. Given that the set of classes K contains exactly N elements

$$\hat{f}(P(\rho)) = \{(k_1, \varphi_1), (k_2, \varphi_2), \ldots, (k_N, \varphi_N)\}, \quad (1)$$

where $\varphi_i \in \mathbb{R}$, $i \in \{1, \ldots, N\}$, represents the real-valued membership scores of object ρ to class $k_i \in K$ under the condition that the image of object $P(\rho)$ is observed. The final classification decision is made for class $k^*(\mathrm{P}(\rho)) = \arg\max \hat{f}(\mathrm{P}(\rho))$. The trivial scheme of an object recognition system within the described model is illustrated in Fig. 1.

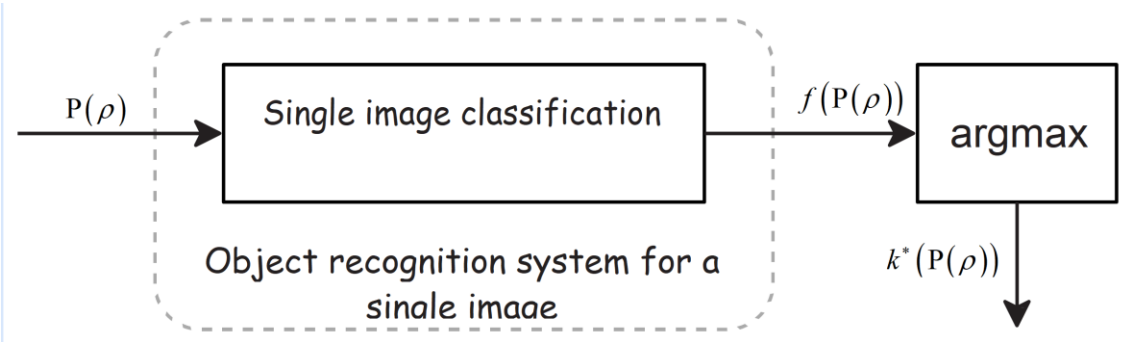

Fig. 1. Trivial Scheme of a Single Object Recognition System

If we exclude the validation process of recognition results and the process of training the parameters of the recognition system (in cases where machine learning methods, such as artificial neural networks, are used for classification), and focus directly on the recognition process, such a recognition system is static and does not involve feedback.

Let's consider the task of recognizing object ρ in a video stream. The video stream is generated by a capturing device that provides a sequence of frames, each of which is an independent image of object ρ. Under the condition of a fixed number of frames, the task of recognizing an object in a video stream can be treated as a static system, similar to the one presented in Fig. 1, but with a more complex input model. Thus, the sequence of M

frames can be considered as a set of images of object ρ as $\mathrm{P}(\rho)=\{\mathrm{P}_1(\rho),\mathrm{P}_2(\rho),\ldots,\mathrm{P}_M(\rho)\}\subset\mathrm{P}$. Meanwhile, the output model of the system remains unchanged. Implementations of such a system may differ in their approaches to data integration. A trivial approach might consider the classification process as a "black box" that processes multiple images simultaneously. Other options partially or fully use methods for recognizing individual object images and perform integration either at the input image level or at the recognition results level of each individual image.

The development of object recognition systems for real-time video streams has garnered increasing attention due to advancements in computer vision and mobile technology. Traditional systems typically rely on single-frame analysis, which limits their robustness under variable conditions such as noise, environmental distortions, and preprocessing defects. Recent studies emphasize the need to move beyond single-image recognition to leverage the temporal and spatial information available in video streams.

Research on video-based recognition systems often intersects with studies on combining outputs from multiple classifiers. Strategies such as weighted significance levels of classifiers [6], training combination rules based on statistical features [6, 7], and multiset approaches for group classification [8] provide a foundation for integrating multiple observations. However, these approaches are primarily designed for heterogeneous classifiers and lack direct applicability to sequential data in video streams.

Incorporating convolutional neural networks (CNNs) into recognition systems has significantly advanced their accuracy and scalability [9]. CNNs have achieved human-level performance in specific tasks [10] but remain sensitive to minor input variations, which can result in unstable recognition outcomes [9]. This instability underscores the importance of using video streams to mitigate the effects of noise and preprocessing defects by aggregating information across multiple frames.

Another area of focus is the use of "super-resolution" techniques to enhance object recognition by generating high-resolution images from multiple low-resolution frames [11]. While effective for image quality improvement, these methods do not address inherent algorithmic errors in recognition systems. Recent research has also explored the challenges of recognizing complex objects, such as holographic security elements, that are indistinguishable in single frames but can be identified using temporal data from video streams.

Despite these advancements, there is limited exploration of strategies specifically tailored to combining frame-by-frame recognition results in video streams. This gap highlights the need for novel algorithms that can effectively utilize temporal data while operating within the computational constraints of mobile devices. Addressing these challenges is critical for developing robust, real-time object recognition systems that meet the demands of modern applications in mobile technology, security, and automated document processing.

FORMULATION OF THE ARTICLE'S OBJECTIVES

The goal of this research is to develop a robust object recognition system for real-time video streams, leveraging dynamic modeling and integration strategies to enhance recognition accuracy under the constraints of mobile platforms and varying environmental conditions.

Research objectives are:

1. Design a dynamic model for real-time video stream processing that effectively utilizes temporal and spatial data to mitigate the impact of environmental noise and preprocessing defects.
2. Explore and evaluate methods for combining the outputs of individual frame classifiers to improve recognition accuracy, considering the limitations and computational constraints of mobile devices.
3. Conduct experimental evaluations on diverse datasets to compare the effectiveness of the proposed recognition system against existing approaches, particularly under conditions involving noise, segmentation errors, and varying input quality.

PRESENTATION OF THE MAIN MATERIAL

The previously mentioned static models of the object recognition system in a video stream do not fully reflect the recognition scenario using a mobile device, as these models assume a set of frames as input, without ordering, and do not account for changes in the system's state during the capture process. Additionally, given the hardware limitations of mobile devices, storing and processing multiple images may be impractical or impossible. To more accurately correspond to the object recognition process in a video stream on a mobile device, it is proposed to consider a dynamic model with discrete time.

For the purposes of formalization, let us represent the video stream as a sequence of images of the object generated over time. Thus, discrete time is given as $t=0,1,2,\ldots$, and the video stream contains images of the observed object $\mathrm{P}_t(\rho)\in\mathrm{P}$. This discrete model of the video stream corresponds to the principles of representing a coded video stream in software systems [12].

To define the object recognition system in a video stream, which is generated independently, it is necessary to define a service model that acts as an intermediate layer between the video stream and the direct flow of images processed by the recognition system. The most trivial service model is one where images generated during the

processing of the previous image by the recognition system are discarded. If storing a collection of images is possible, an alternative service model is one with a buffer that allows incoming images to be accumulated and delivered to the system on request at any arbitrary time, without limitations associated with the discretization of image generation by the source. From the perspective of the recognition system of the image sequence, the set of recognition methods and algorithms, as well as result integration, do not depend on the service model. Therefore, in this work, it will be assumed that at any time t, the "current" image $\mathrm{P}_t(\rho)$ can be captured, and during system loading periods, images may be discarded. The recognition system maintains some internal state $v_t \in V$, which changes over time. The time Δ required to obtain an updated result after inputting the next image $\mathrm{P}_t(\rho)$ is, in general, a function of the image and the system's internal state: $\Delta = f\left(\mathrm{P}_t(\rho), v_t\right)$, which may be computationally infeasible at time t. The recognition result, which takes into account the information contained in the image captured at time t, may only be available at time $D(t) = t + \Delta$.

At the initial moment in time: $t = 0$, the system's internal state is initialized to v_0. Let image $\mathrm{P}_t(\rho)$ be captured at time t, which is then fed into the recognition module $\hat{f}$. The recognition result $\hat{f}\left(\mathrm{P}_t(\rho)\right)$ becomes available at time $t' \geq t$ and is stored in the system's memory module (i.e., it becomes part of the state $v_{t'}$). After that, the results of object image recognition accumulated up to that point are combined, and at time $D(t) \geq t'$, the recognition result $Y_{D(t)}$ is output. After the result is output, the next image $\mathrm{P}_{D(t)}(\rho)$ is captured, and the process continues. Thus, the result $Y_{D(t)}$ takes into account the information contained in the images with indices $0, D^1(0), D^2(0), \ldots, t$ (the superscript sign in the function $D(t)$ indicates multiple composition of the function, not exponentiation). The quality of the result is characterized by how close result $Y_{D(t)}$ is to the true value $w(\rho)$ of object ρ, according to a certain metric. The diagram of the described recognition system is presented in Fig. 2.

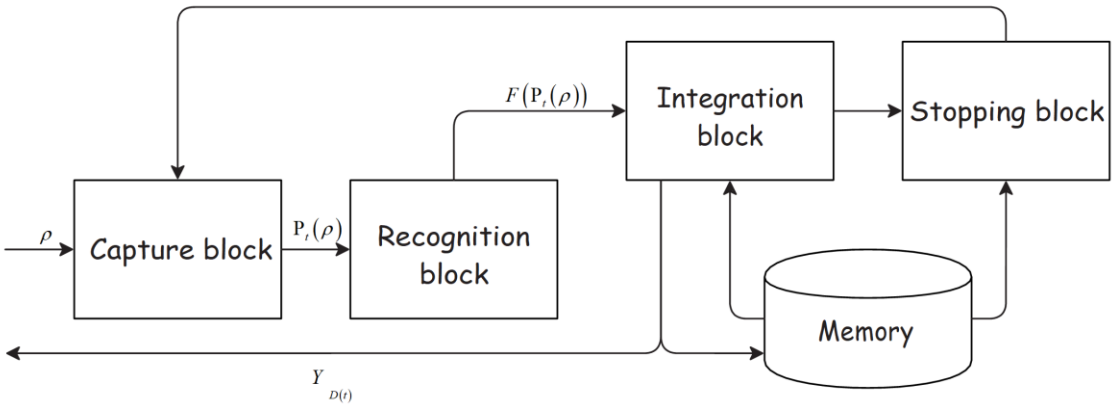

Fig. 2. Diagram of the object recognition system in a video stream with pause

The feature extraction and object classification methods applicable in static systems (see Fig. 2.5) are also applicable in the dynamic model. However, the dynamic model of the object recognition system in a video stream has a number of specific properties. First and foremost, it is important to highlight the increased influence of the performance of single-image recognition algorithms on the system's output. Indeed, reducing the time Δ required to recognize a single image $\mathrm{P}_t(\rho)$ allows more information about the object ρ to be processed in the same absolute time (i.e., within the same time frame from the perspective of the user/operator). In addition, within such a system, tasks arise that are atypical for traditional object recognition systems on images. The first such task is obtaining the result $Y_{D(t)}$ – the task of combining (integrating) the recognition results of the same object in different images into a single result. The second task is stopping the recognition process – since image capture may not be naturally limited, at time $D(t)$, the task arises of deciding whether the image capture process should be terminated and the accumulated result up to that point should be accepted as final.

As the performance measure of the system at the stop moment $t = t_{fin}$, it is proposed to consider a linear combination:

$$\alpha l\left(Y_{t_{fin}}, w(\rho)\right) + \beta\Pi\left(t_{fin}\right), \tag{2}$$

where α, β are constants, $l\left(Y_{t_{fin}}, w(\rho)\right)$ is the distance from the integrated result Y_t to the true value $w(\rho)$, characterizing the quality of the result, and $\Pi(t)$ is a penalty function based on time. A special case of the penalty function $\Pi(t)$ is the number of processed images:

$$\Pi(t) = \max\left\{i \middle| D^i(0) \le t\right\}. \tag{3}$$

The primary task of traditional object recognition systems is to maximize recognition accuracy (i.e., to maximize the proportion of "correct" object classifications). The task of integrating object recognition results is to maximize the accuracy of recognizing a set of various images of the same object, given the recognition results of individual images. To formalize the problem of integration from the perspective of the object recognition system model in a video stream, let us assume that a set of objects $\mathrm{P} = \{\rho_1, \rho_2, \ldots, \rho_J\}$ of cardinality I and a set of video sequences

$$X = \left\{\mathrm{P}_1\left(\rho_{x_1}\right), \mathrm{P}_2\left(\rho_{x_2}\right), \ldots, \mathrm{P}_J\left(\rho_{x_J}\right)\right\} \tag{4}$$

of cardinality J are given, where x_j is the index of an object from the set X for each $j \in \{1, 2, \ldots, J\}$, and each video sequence $\mathrm{P}_j\left(\rho_{x_j}\right) = \left\{\mathrm{P}_{j1}\left(\rho_{x_j}\right), \mathrm{P}_{j2}\left(\rho_{x_j}\right), \ldots, \mathrm{P}_{jMj}\left(\rho_{x_j}\right)\right\}$ is a sequence of images of object $\rho_{x_j} \in \mathrm{P}$, which may be affected by environmental noise and preprocessing defects (see Section 2.1). Additionally, a set of classes $K = \{k_1, k_2, \ldots, k_N\}$ and information about the ideal belonging of each object to the corresponding class w are provided: $\mathrm{P} \to K$.

The task of object recognition in a video stream can be formulated as finding a classification function F: $\mathrm{P}^* \to K$, that maximizes recognition accuracy [13]:

$$W_F(X) = \frac{1}{J}\sum_{j=1}\left[F\left(\mathrm{P}_j\left(\rho_{x_j}\right)\right) = w\left(\rho_{x_j}\right)\right] \to \max_F. \tag{5}$$

A more specific task of integrating the results of single-object recognition involves an integration function $Y: \left(\mathbb{R}^K\right)^* \to \mathbb{R}^K$ that transforms a sequence of recognition results for individual images into a unified recognition result for video sequences (here, $\mathbb{R}^K$ is the set of all possible mappings from the set of classes K to the set of scores $\mathbb{R}$, i.e., the set of all possible classification outcomes). Since the final recognition result for a video sequence is the class corresponding to the highest score in the recognition output $F(\mathrm{P}) = \arg\max Y\left(\hat{f}\left(\mathrm{P}\right)\right)$ (see Section 2.2), the formulation of the integration task is based on (5) and takes the form of

$$W_F(X) = \frac{1}{J}\sum_{j=1}\left[\arg\max Y\left(\hat{f}\left(\mathrm{P}_j\left(\rho_{x_j}\right)\right)\right) = w\left(\rho_{x_j}\right)\right] \to \max_Y. \tag{6}$$

In the ideal case, the classification function F or the result integration function Y should have the ability to filter out outliers that appear in the input data stream due to environmental noise or preprocessing defects, and should also be capable of filtering classifier noise, mitigating random internal errors.

It is easy to notice that the approach to integration as the task of constructing a classification function F can be reduced to the task of building a result integration function Y by applying the existing method of classifying individual object images. Alternative approaches include, for example, super-resolution techniques [11], which perform pixel matching of a set of input images $\mathrm{P}(\rho)$ and create a unified (ideal) image of the object, which is then classified. However, it should be noted that due to the specific characteristics of the most accurate existing method for image classification (convolutional neural networks), namely, their instability to random pixel distortions, this work will focus on methods for constructing the integration function Y for the results of single-image recognition.

In the literature, the task of combining single-object classification results is usually discussed in the context of methods for obtaining more accurate classification by combining the results of several different classifiers [14].

Depending on the model used for the classification result of an object and the interpretation of the classifier's scores, various combination methods are considered.

The task of combining object classification results can be considered as a collective decision-making problem. Let us introduce the concept of a predictor of classifier result reliability as a real-valued function $r\left(\mathrm{P}(\rho), \hat{f}\right)$, reflecting the degree of confidence that the classification result of image $\mathrm{P}(\rho)$ by the function $\hat{f}$ will be correct. It makes sense to use computable image characteristics as predictors, which are known to influence classification accuracy [9], such as blur and focus level assessments [5], noise level estimation, digitization artifacts [10], and so on. (Such predictors can be considered a priori, as they directly rely on the characteristics of the input images). Another class of predictors is conditioned by the values of classification scores (posterior predictors), which are related to the concept of the recognition result reliability score [5]. An example of a widely used posterior predictor of reliability is the value of the first (maximum) alternative score [15]:

$$r\left(\mathrm{P}(\rho), \hat{f}\right) = \max \hat{f}\left(\mathrm{P}(\rho)\right). \tag{7}$$

Let a certain reliability predictor be given. Then, the task of integrating the recognition results of a sequence of images $\mathrm{P}(\rho) = \left\{\mathrm{P}_1(\rho), \mathrm{P}_2(\rho), \ldots, \mathrm{P}_M(\rho)\right\}$ of cardinality M can be considered as a collective decision-making problem with experts, whose competence levels are functions of the values of the reliability predictor. It is worth noting that the experts' competence levels in this model reflect the input data, as the characteristics of individual observations (i.e., individual images $\mathrm{P}_1(\rho), \ldots, \mathrm{P}_M(\rho)$) are necessary for evaluating the significance of the experts.

To answer this question, an experimental study is proposed. Four datasets were prepared, the characteristics of which are provided in Table 1. The MRZ-MSEGM and MRZ-CLEAN datasets contain video sequences of recognition results for characters in the machine-readable zone of international documents [16]. The ICN-MSEGM and ICN-CLEAN datasets contain video sequences of recognition results for the "Number" field characters of payment bank cards, captured using indent printing. The images of characters in the considered test datasets are subject to a wide range of distortions: uneven or insufficient lighting, digital noise, defocusing and "blurring" due to the optical sensor's movement relative to the medium, glare from an external light source, and interference created by the holographic security layer of the document, among others. The recognition result for each individual character image was obtained using convolutional neural networks, separately trained for the machine-readable zone characters and for the "Number" field characters of payment bank cards, on separate training image sets using data augmentation techniques [15]. The MRZ-MSEGM and ICN-MSEGM datasets contain errors caused by the incorrect or insufficiently accurate operation of document localization algorithms and text line segmentation algorithms. The MRZ-CLEAN and ICN-CLEAN datasets are subsets of the corresponding MRZ-MSEGM and ICN-MSEGM datasets, which do not contain such errors. Thus, in the MRZ-CLEAN and ICN-CLEAN datasets, each video sequence contains images of the exact same character, without any segmentation defects.

Table 1.

Characteristics of the test datasets MRZ-MSEGM, MRZ-CLEAN, ICN-MSEGM, and ICN-CLEAN

Dataset description	MRZ-CLEAN	MRZ-MSEGM
Cardinality of the set of classes K	37	
Total number of symbol images	631530	637874
ecognition accuracy of individual images, %	96.8994	96.7357
Number of video sequences	7508	7581
Minimum length $P(\rho)$	3	
Maximun length $P(\rho)$	223	
Mean length $P(\rho)$	21	
Dataset description	ICN-CLEAN	ICN-MSEGM
Cardinality of the set of classes K	10	
Total number of symbol images	29166	31580
ecognition accuracy of individual images, %	96.8936	90.9816
Number of video sequences	1748	1898
Minimum length $P(\rho)$	3	
Maximun length $P(\rho)$	25	
Mean length $P(\rho)$	12	

The comparison of basic classifier combination strategies presented in the overview chapter was conducted on the provided test datasets: the product rule (8), sum rule (9), minimum rule (10), maximum rule (11), and median rule (12):

$$MUL\left(\mathrm{P}\right)(\phi) = P\left(\phi|\mathrm{P}\right) = \frac{1}{P(\phi)^{N-1}} \prod_{i=1}^{N} C_{\Sigma}\left(\mathrm{P}_i\right)(\phi); \tag{8}$$

$$SUM\left(\mathrm{P}\right)\left(\phi\right)=\frac{1}{N}\sum_{i=1}^{N}C_{\Sigma}\left(\mathrm{P}_{i}\right)\left(\phi\right); \tag{9}$$

$$MIN\left(\mathrm{P}\right)\left(\phi\right)=\left(\min_{i=1}^{N}C_{\Sigma}\left(\rho_{i}\right)\left(\phi\right)\right)\left(\sum_{k=1}^{K}\min_{i=1}^{N}C_{\Sigma}\left(\rho_{i}\right)\left(\phi_{k}\right)\right); \tag{10}$$

$$MAX\left(\mathrm{P}\right)\left(\phi\right)=\left(\max_{i=1}^{N}C_{\Sigma}\left(\rho_{i}\right)\left(\phi\right)\right)\left(\sum_{k=1}^{K}\max_{i=1}^{N}C_{\Sigma}\left(\rho_{i}\right)\left(\phi_{k}\right)\right); \tag{11}$$

$$AVG\left(\mathrm{P}\right)\left(\phi\right)=\left(\operatorname*{avg}_{i=1}^{N}C_{\Sigma}\left(\rho_{i}\right)\left(\phi\right)\right)\left(\sum_{k=1}^{K}\operatorname*{avg}_{i=1}^{N}C_{\Sigma}\left(\rho_{i}\right)\left(\phi_{k}\right)\right). \tag{12}$$

The recognition accuracy of a video sequence of characters is the relative proportion of video sequences for which the ideal answer matches the class that received the highest score according to a given combination rule. Additionally, a comparison was made between the basic combination rules and the voting method:

$$\begin{aligned} &Poll\left(\sigma\right)\left(\hat{f}\left(\mathrm{P}\left(\rho\right)\right)\right)\left(k\right)= \\ &=\sigma\frac{1}{M}\sum_{i=1}^{M}1_{\mathrm{P}_{k}\left(\rho\right)}\left(\mathrm{P}_{i}\left(\rho\right)\right)+\left(1-\sigma\right)\max_{i=1}^{M}\left(1_{\mathrm{P}_{k}\left(\rho\right)}\left(\mathrm{P}_{i}\left(\rho\right)\right)r\left(\mathrm{P}_{i}\left(\rho\right),\hat{f}\right)\right), \end{aligned} \tag{13}$$

where $\mathrm{P}_{k}\left(\rho\right)=\left\{\mathrm{P}\left(\rho\right)\in\mathrm{P}\left(\rho\right)\middle|\, f\left(\mathrm{P}\left(\rho\right)\right)=k\right\}$ is a subset of video sequence elements for which the classifier choice corresponds to class k, $1_{\mathrm{P}_{k}\left(\rho\right)}\left(\mathrm{P}\left(\rho\right)\right)$ is the indicator function for the membership of image $\mathrm{P}\left(\rho\right)$ in subset $\mathrm{P}_{k}\left(\rho\right)$, and $r\left(\mathrm{P}_{i}\left(\rho\right),\hat{f}\right)$ is the credibility predictor. The posterior predictor "first alternative rule" (7) was used as the credibility predictor.

Figure 3 shows the comparative values of recognition accuracy for video sequences using combination rules (8), (9), (10), (11), (12), and (13) on the test datasets MRZ-MSEGM, MRZ-CLEAN, ICN-MSEGM, and ICN-CLEAN. The horizontal axis of the graphs corresponds to the values of parameter a for the combination rule (13). The recognition accuracy using the other combination rules is represented by horizontal lines.

Fig. 3 demonstrates a significant difference in the optimal choice of combination strategy depending on the input data model: on test sets where errors in character localization and segmentation occur, higher recognition accuracy for video sequences is achieved by the product rule (8), voting (13), and the sum rule (9) (Fig. 3a, 3c). However, on test sets where such errors were excluded (Fig. 3b, 3d), higher recognition accuracy is provided by the maximum rule (11). In other words, when considering this task as a collective decision-making problem, in the case of a stricter input data model (with no errors in character localization and segmentation), it is more advantageous to trust a single competent expert rather than the collective opinion of several experts.

In the presence of errors in character localization and segmentation, the stability of the credibility predictors decreases, which, in turn, increases the difference between the competency ratings of the experts (constructed based on the values of the predictors) and the actual competency values of the experts (which correspond to the posterior probabilities of making the correct decision). In such cases, selecting the expert with the highest competency level is more often incorrect, and thus the difference between the actual competency level of the chosen expert and the competency levels of the other experts decreases. Therefore, the optimal choice, in terms of the reliable predictor of video sequence credibility, is absent when segmentation and localization errors are present, which aligns with a broader interpretation of collective decision-making theory [2, 6]. According to this theory, a violation of the first part of Condorcet's assertion (that with an increasing number of experts, the probability of collective correct decision-making increases if each expert has a higher probability of making a correct individual decision than a wrong one) occurs when the difference between the competency levels of the most competent expert and the others increases.

The results of the experiment allow the conclusion that when building an object recognition system in a video stream, the choice of a combination strategy for the results should be guided not only by the model of the object recognition results but also by the noise model of the input data. In this case, with a fixed noise model for the input data, the results of studies aimed at combining different classifiers to maximize the recognition accuracy of a single object can be used for integrating classification results of individual objects. However, direct application of the considered combination rules is not possible if the object recognition result model is more complex than a simple classification result (1). An example of such an object is a text string, for which classification is performed independently for each character. The task of integrating such objects will be addressed in Chapter 3.

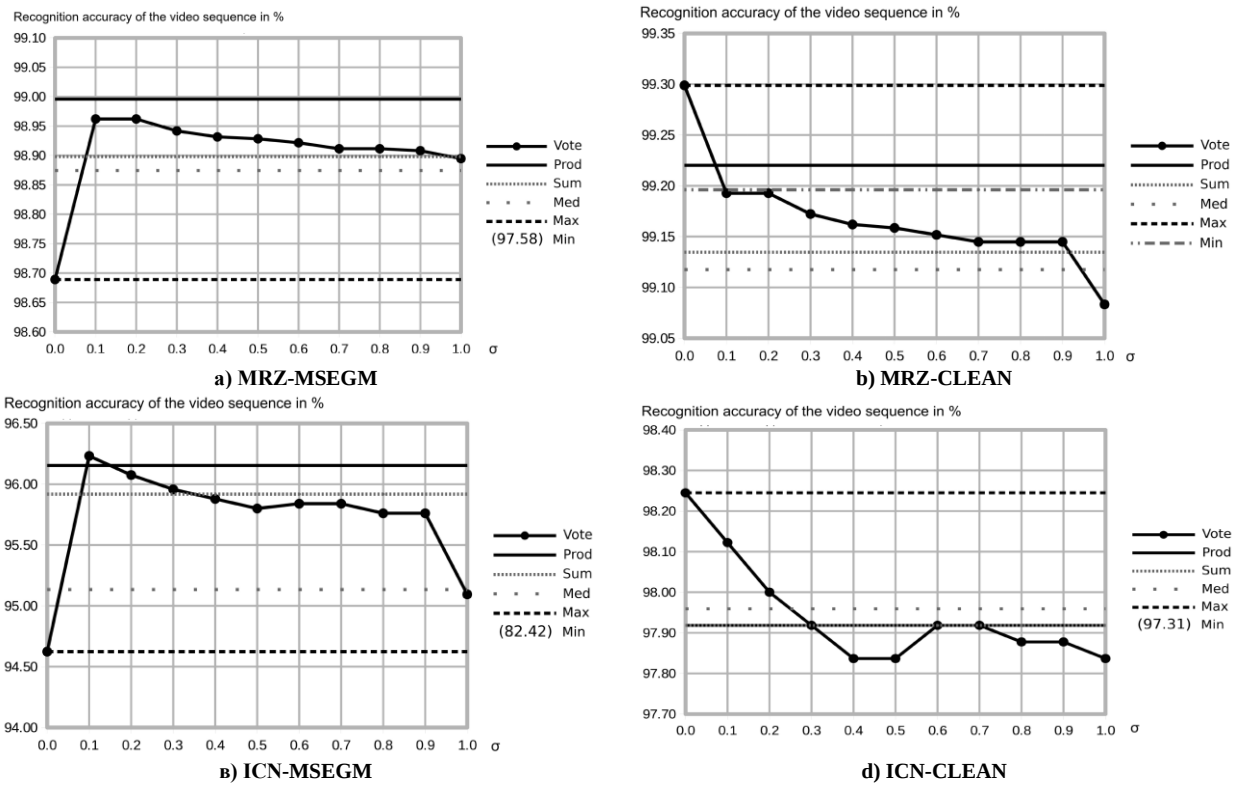

Fig. 3. Comparison of recognition accuracy for video sequences of characters using basic combination strategies

CONCLUSIONS FROM THE PRESENT STUDY AND PROSPECTS FOR FURTHER RESEARCH IN THIS AREA

This study has proposed a robust system for real-time object recognition in video streams, addressing key challenges such as environmental noise, preprocessing defects, and algorithmic limitations. By incorporating dynamic modeling and leveraging the capabilities of convolutional neural networks (CNNs), the system effectively utilizes temporal and spatial information to improve recognition accuracy. Experimental results have demonstrated the benefits of integrating classifier outputs through optimized combination strategies, highlighting the system's adaptability to varying noise models and input conditions.

The research findings have practical implications for a wide range of applications, including automated document processing, identity verification, and security systems. They also contribute to advancing mobile and embedded vision technologies, offering solutions for hardware-constrained environments.

Future research in this domain should focus on the following areas:

1. Developing advanced algorithms for real-time processing that further minimize computational overhead, enabling deployment on low-power mobile devices.

2. Exploring the integration of novel deep learning architectures and attention mechanisms to enhance recognition accuracy for complex and dynamic input scenarios.

3. Investigating methods for adaptive noise modelling and real-time data augmentation to improve system robustness in diverse and unpredictable environments.

These advancements will pave the way for more efficient and versatile object recognition systems, fostering innovation in both academic research and practical applications.

References

1. Verhoef P. C., Broekhuizen T., Bart Y., Bhattacharya A., Qi Dong J., Fabian N., Haenlein M. Digital transformation: A multidisciplinary reflection and research agenda // Journal of Business Research. – 2021. – Vol. 122. – P. 889–901. – Elsevier BV. DOI: https://doi.org/10.1016/j.jbusres.2019.09.022.

2. Sun Y., Sun Z., Chen W. The evolution of object detection methods // Engineering Applications of Artificial Intelligence. – 2024. – Vol. 133. – Article ID: 108458. – Elsevier BV. DOI: https://doi.org/10.1016/j.engappai.2024.108458.

3. Verhoef P. C., Broekhuizen T., Bart Y., Bhattacharya A., Qi Dong J., Fabian N., Haenlein M. Digital transformation: A multidisciplinary reflection and research agenda // Journal of Business Research. – 2021. – Vol. 122. – P. 889–901. – Elsevier BV. DOI: https://doi.org/10.1016/j.jbusres.2019.09.022.

4. Xiong B., Kalantidis Y., Ghadiyaram D., Grauman K. Less Is More: Learning Highlight Detection From Video Duration // Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2019. – P. 1–35. – IEEE. DOI: https://doi.org/10.1109/cvpr.2019.00135.

5. Kumari S., Prabha C., Karim A., Hassan Md. M., Azam S. A Comprehensive Investigation of Anomaly Detection Methods in Deep Learning and Machine Learning: 2019–2023 // IET Information Security. – 2024. – Vol. 2024. – Issue 1. – Institution of Engineering and Technology (IET). DOI: https://doi.org/10.1049/2024/8821891.

6. Suresh V., Mohan C. K., Kumaraswamy R., Yegnanarayana B. Combining multiple evidence for video classification //

Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing. – 2005. – P. 187–192. – IEEE. DOI: https://doi.org/10.1109/icisip.2005.1529446.

7. Alqahtani A. F., Ilyas M. An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying // Machine Learning and Knowledge Extraction. – 2024. – Vol. 6. – Issue 1. – P. 156–170. – MDPI AG. DOI: https://doi.org/10.3390/make6010009.

8. Chen R.-C., Dewi C., Huang S.-W., Caraka R. E. Selecting critical features for data classification based on machine learning methods // Journal of Big Data. – 2020. – Vol. 7. – Issue 1. – Springer Science and Business Media LLC. DOI: https://doi.org/10.1186/s40537-020-00327-4.

9. Alzubaidi L., Zhang J., Humaidi A. J., Al-Dujaili A., Duan Y., Al-Shamma O., Santamaría J., Fadhel M. A., Al-Amidie M., Farhan L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions // Journal of Big Data. – 2021. – Vol. 8. – Issue 1. – Springer Science and Business Media LLC. DOI: https://doi.org/10.1186/s40537-021-00444-8.

10. Rakhmatulin I., Dao M.-S., Nassibi A., Mandic D. Exploring Convolutional Neural Network Architectures for EEG Feature Extraction // Sensors. – 2024. – Vol. 24. – Issue 3. – Article ID: 877. – MDPI AG. DOI: https://doi.org/10.3390/s24030877.

11. Umirzakova S., Ahmad S., Khan L. U., Whangbo T. Medical image super-resolution for smart healthcare applications: A comprehensive survey // Information Fusion. – 2024. – Vol. 103. – Article ID: 102075. – Elsevier BV. DOI: https://doi.org/10.1016/j.inffus.2023.102075.

12. Peña-Ancavil E., Estevez C., Sanhueza A., Orchard M. Adaptive Scalable Video Streaming (ASViS): An Advanced ABR Transmission Protocol for Optimal Video Quality // Electronics. – 2023. – Vol. 12. – Issue 21. – Article ID: 4542. – MDPI AG. DOI: https://doi.org/10.3390/electronics12214542.

13. Vaidya G. A. Object Detection In Video Streaming Using Machine Learning And Cnn Techniques // Journal of Advanced Zoology. – 2024. – P. 186–196. – Green Publication. DOI: https://doi.org/10.53555/jaz.v45is4.4183.

14. Arco J. E., Ortiz A., Ramírez J., Martínez-Murcia F. J., Zhang Y.-D., Górriz J. M. Uncertainty-driven ensembles of multi-scale deep architectures for image classification // Information Fusion. – 2023. – Vol. 89. – P. 53–65. – Elsevier BV. DOI: https://doi.org/10.1016/j.inffus.2022.08.010.

15. Johnson M. S., Sinharay S. The Reliability of the Posterior Probability of Skill Attainment in Diagnostic Classification Models // Journal of Educational and Behavioral Statistics. – 2019. – Vol. 45. – Issue 1. – P. 5–31. – American Educational Research Association (AERA). DOI: https://doi.org/10.3102/1076998619864550.

16. Liu Y., Joren H., Gupta O., Raviv D. MRZ code extraction from visa and passport documents using convolutional neural networks // International Journal on Document Analysis and Recognition (IJDAR). – 2021. – Vol. 25. – Issue 1. – P. 29–39. – Springer Science and Business Media LLC. DOI: https://doi.org/10.1007/s10032-021-00384-2.

Xia Guanxiang **Ся Гуансянг**	Graduate student, Vinnytsia National Technical University, Vinnytsia, Ukraine, Teacher, Hunan Mass Media Vocational and Technical College, Changsha, China, ORCID N/A e-mail: xiaguanxiang@163.com	Аспірант, Вінницький національний технічний університет, Вінниця, Україна, Викладач, Хунанський коледж масових медіа, Чанше, Китай.
Viacheslav Kovtun **Вячеслав Ковтун**	Doctor of Technical Sciences, Professor, Head of the Computer Control Systems Department, Vinnytsia National Technical University, Vinnytsia, Ukraine, https://orcid.org/0000-0002-7624-7072 e-mail: kovtun_v_v@vntu.edu.ua Scopus Author ID: 57195679681 ResearcherID: M-9043-2019	доктор технічних наук, професор, завідувач кафедри комп'ютерних систем управління, Вінницький Національний Технічний Університет, Вінниця, Україна.