Augmented reality audiostream creation using CNN: boosting inclusion and safety for visually impaired people

Olexander Mazurets[1], Olena Sobko[1,*], Rostyslav Dydo[1], Olha Zalutska[1] and Maryna Molchanova[1]

[1] *Khmelnytskyi National University, 11, Instytuts'ka str., Khmelnytskyi, 29016, Ukraine*

Abstract

In paper proposed approach for boost the inclusion and safety of visually impaired people by creating method of augmented reality audiostream creation which contains objects and named persons identified and classified from a video stream using CNN. Developed method differs from existing ones by two advantages: ability to classify named individuals in the video stream for boosting the inclusion of visually impaired people; and ability to consider of prioritizes dangerous objects when generating audiostream for boosting the safety of visually impaired people. By this, method is contributed to achieving UNDP SDG No. 3, No. 10 and No. 11. For research the method effectiveness, test software was developed. Obtained the applied comparison results of neural network models YOLOv5n, YOLOv5x, YOLOv8n, YOLOv8x, YOLOv11n and YOLOv11x revealed that the YOLOv11x model is better suited for objects identifying and classification in environment; obtained sufficient values of macrometrics for 80 objects classes: Accuracy 74.1%, Precision 73.74%, Recall 65.85%, mAP@0.5 71.34%, and mAP@0.5:0.95 54.85%. Obtained the applied results from CNN neural network model macrometrics indicate high-quality classification of named persons: Accuracy 97%, Precision 95%, Recall 96%, and F1-score 95%.

Keywords

augmented reality, visually impaired people, audiostream, object detection, YOLO, CNN

1. Introduction

According to the World Health Organization, there are about 2.2 billion people with visual impairments in the world, and their number is increasing every year due to population aging and insufficient access to health services in developing countries [1]. In Ukraine, there are no official data on the number of people with visual impairments, but, according to various estimates, their number varies from 70 to 300 thousand people [2]. The situation is complicated by military aggression, which leads to an increase in the number of people with eye injuries. People with visual impairments face a number of serious problems, especially with regard to orientation in the environment. The main problems are difficulties with mobility, insufficient accessibility of the urban environment, limitations in the perception of information and difficulties in social interaction [3, 4]. One of the main problems is the difficulty in identifying physical obstacles. Due to the lack of visual information, people with visual impairments are forced to rely on auditory cues, tactile markers, and assistive devices such as canes or guide dogs [5, 6].

There are also challenges with the use of current technologies. For example, while mobile applications based on artificial intelligence help with navigation, they often have limited

MoMLeT-2025: 7th International Workshop on Modern Machine Learning Technologies, June, 14, 2025, Lviv-Shatsk, Ukraine

[*] Corresponding author.

exe.chong@gmail.com (O. Mazurets); olenasobko.ua@gmail.com (O. Sobko); rostyslav0805@gmail.com (R. Dydo); zalutska.olha@gmail.com (O. Zalutska); m.o.molchanova@gmail.com (M. Molchanova)

0000-0002-8900-0650 (O. Mazurets); 0000-0001-5371-5788 (O. Sobko); 0009-0005-4377-7607 (R. Dydo); 0000-0003-1242-3548 (O. Zalutska); 0000-0001-9810-936X (M. Molchanova)

functionality or low accuracy [7]. At the same time, the implementation of such solutions in everyday life can significantly improve the ability to be autonomous, providing timely information about objects and obstacles in real time for people with visual impairments, for example, based on portable devices such as GoogleGlass, smart canes, etc. [8, 9].

Other problems include psychological barriers [10], such as fear of new routes or public transport, and a lack of specialists who can provide quality spatial orientation training. The latter is especially relevant in the context of the growing number of people with visual impairments due to injuries or diseases [11], and has been complicated by the impact of quarantine restrictions [12, 13] and propaganda influence [14, 15], which has given special importance to the issue of inclusion of people with visual impairments. Although measures are being taken at the state level to increase inclusion for people with visual impairments, in particular, equipping the urban environment with tactile elements, ensuring access to education, implementing programs to support employment and legal protection, these efforts are still insufficient [16]. The essence of the problem lies in the lack of accessibility of information technologies, the lack of modern equipment for the integration of people with visual impairments into everyday life, as well as the uneven implementation of inclusive solutions in different regions [17].

The use of such technologies will contribute to increasing inclusion and safety for people with visual impairments and creating an environment that takes into account the needs of people with disabilities, and the created software will help achieve the UN Sustainable Development Goals (SDGs), in particular SDG No. 3 «Promote healthy lives and well-being for all ages», SDG No. 10 «Reduce inequalities within and among countries» and SDG No. 11 «Ensure accessible, safe, sustainable and environmentally sustainable cities and human settlements» [18]. In particular, achieving SDG No. 3 through the development of information technologies that enhance the mobility and autonomy of people with visual impairments will occur by improving their psycho-emotional state and physical health [19]. The achievement of SDG No. 10 is justified by the fact that the integration of information technologies for people with visual impairments will reduce inequalities in access to education, employment and social life [20]. The use of audio prompts and automated navigation systems will help ensure safe movement in urban environments to achieve SDG No. 11 [21].

Research goal is to boost the inclusion and safety of visually impaired people by creating an audio stream of objects and named persons identified and classified from a video stream using deep convolutional neural networks.

The main paper contribution is the proposed method of augmented reality audiostream creation using CNN, which differs from existing ones in that when classifying objects in a video stream, it is able to recognize named persons among them, which contributes to increasing the level of inclusion. Also, thanks to the introduced system of listing priority objects by danger, the proposed method can be used to increase the safety of people with visual impairments.

2. Related Works

The relevance of using information technology to improve inclusion and safety for people with visual impairments is undeniable. Some researchers focus on integrating multisensory data, combining images, sounds, and information from different sensors, creating a more complete picture of the environment [22]. Such systems provide adaptive audio feedback according to the users needs and conditions. Other researchers are working on developing deep learning-based systems that use computer vision to recognize objects and create their audio descriptions [23]. The following is a review of current research on the use of artificial intelligence tools and methods for recognizing objects in the environment to ensure inclusion and safety for people with visual impairments, as well as identifying named individuals in a video stream.

A walking assistance system for people with visual impairments using XR glasses for safe movement on the street is proposed in the study [24]. The mobile-optimized YOLOv8n model was used for implementation, and three specialized models were developed for pedestrian paths, transport infrastructure and obstacle recognition, which includes 9 classes in total. The average metrics are Precision 87.05%, Recall 80.8%, F1-Score 84%.

The authors of study [25] developed a mobile application that helps people with visual impairments in their daily lives, increasing their autonomy, mobility and independence. The authors used YOLOv5 for real-time object detection and Google ML Kit for text-to-speech conversion. The YOLOv5 model is able to classify 7 classes of objects with an accuracy of 96%.

In [26], an assistance system for visually impaired people is described, which combines an obstacle detection algorithm with the use of sensors. The proposed system provides distance measurement to objects and provides tactile feedback. An additional camera provides contextual information about the environment using audio instructions. The system is able to recognize 9 objects, for which an average Accuracy of 91.7% was obtained.

A voice application for smartphones was created in [27], which helps blind people in object recognition and navigation in the environment. This system integrates text, people and object recognition functions, and also uses sensors for obstacle detection, applying the K-nearest neighbor algorithm for image processing.

In paper [28], an automated object identification system for the blind is described, which uses RFCN and Mask RCNN models for object detection. The RFCN model showed the best results with an average exact match (mAP) of 0.825. The system uses a Raspberry Pi with a connected camera and an ultrasonic sensor to determine the distance, and the information is transmitted to the user via audio feedback.

The authors of the article [29] presented an image-to-speech system for the blind, combining OpenCV and YOLO to identify objects and create audio descriptions. The system achieves an Accuracy of 96.60%, but the work only considers the recognition of a limited number of environmental objects.

In [30] presents a system for real-time object recognition using the YOLO algorithm optimized for mobile devices. This system allows blind users to identify objects around them through text-audio notifications, promoting their autonomy and reducing the need for help from other people or special devices. The study [31] aims to develop and evaluate a prototype face recognition system for use in the educational process. For implementation, an Android application was created that classifies recognized faces. During the research, the authors obtained a face recognition accuracy rate of 78.57%.

In the paper [32], a new approach to face recognition in video surveillance systems for university laboratories is proposed, aimed at solving the problems of low resolution and partial face overlap. For this purpose, Retinaface-Resnet was reconstructed and combined with Quality-Adaptive Margin (adaface). Experiments showed an accuracy of 96.12% on the WiderFace set and 84.36% in real laboratory conditions.

The authors of the study [33] proposed a new approach to face detection based on deep learning, aimed at solving problems related to facial expressions, overlap and lighting changes. The optimized ResNet-50 architecture with hyperparameter selection using the Gray Wolf algorithm was used. 94% accuracy was achieved on both training and test data.

A method combining LBPH and CNN for image preprocessing using equalized histograms was proposed in the study [34]. LBPH is used to extract and combine histogram values into a single vector, which reduces training loss and increases accuracy to 96.5%.

Analysis of recent studies allows us to conclude that there are unresolved problems on the way to increasing the level of inclusion. In particular, the identification of named persons, which can promote social communication and a sense of security, has not been implemented. Meanwhile, existing solutions do not take into account priorities for identified environmental objects, although the found objects in different situations may have diametric priorities.

In this regard, the study aims to increase the level of inclusion by identifying named persons, and makes an attempt to increase the level of safety of people with visual impairments by introducing priorities for identified environmental objects.

3. Method

Artificial intelligence tools and methods allow us to create systems to support people with special needs. One important direction is the creation of an augmented reality audio stream that can increase inclusion and safety for people with visual impairments by analyzing the environment through a video stream. Figure 1 proposes a general approach to converting input information in the form of a video stream and a prioritized list of dangerous objects for people with visual impairments into output information in the form of an audio stream.

In accordance with the presented approach in Figure 1, a diagram of the method of augmented reality audiostream creation using CNN is shown in Figure 2. As part of the conducted research, the use of YOLO models for identification and classification of objects from the video stream and CNN for classification of named individuals is proposed.

The input data of the method are a video stream, a neural network model for object identification, a priority list of dangerous objects for people with visual impairments, a delay in voicing objects, as well as a confidence threshold for identifying objects in the video stream and a neural network model for classifying named persons. The video stream provides dynamic visual information about the environment in which a person with visual impairments is located, and neural network models process this video stream to identify and classify objects according to a set confidence threshold from which an audio stream is formed, which is relevant for people with visual impairments.

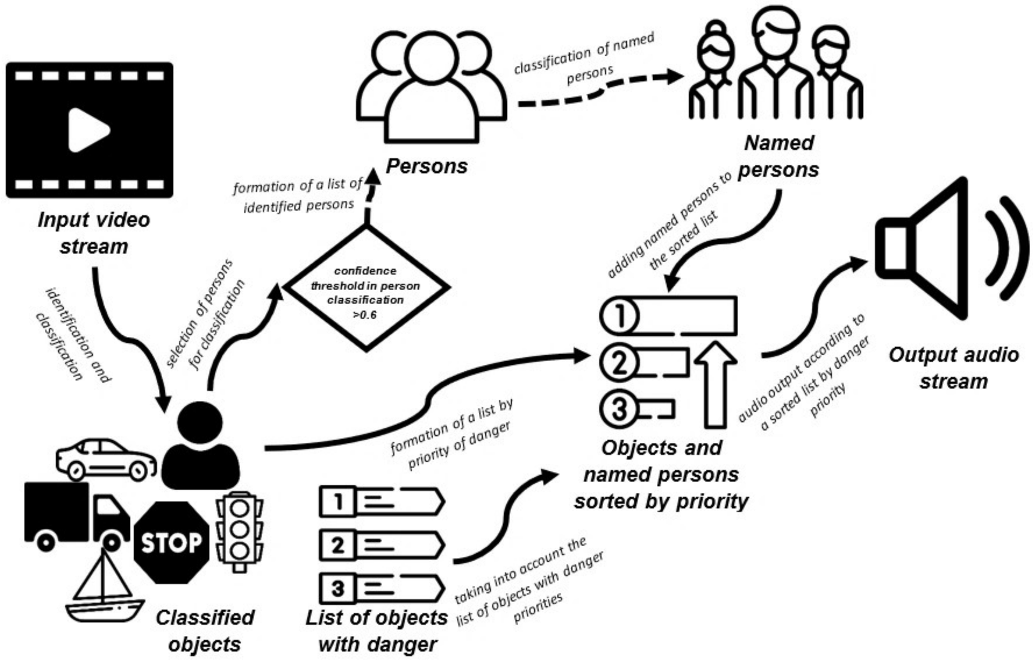

Persons
classification of named persons
Named persons
Input video stream
formation of a list of identified persons
confidence threshold in person classification >0.6
identification and classification
selection of persons for classification
adding named persons to the sorted list
Output audio stream
audio output according to a sorted list by danger priority
formation of a list by priority of danger
Objects and named persons sorted by priority
STOP
taking into account the list of objects with danger priorities
Classified objects
List of objects with danger

Figure 1: Scheme of data conversion for audio stream generation.

The first stage of the method is the pre-processing of the video stream. Preprocessing for the YOLO neural network model [35] includes preparing data for processing, in particular scaling frames to a single size that varies depending on the versions, but for most it is 640x640, normalizing pixel values to reduce variability and converting images to the appropriate RGB color space. After that, the data is converted into tensors compatible with the requirements of the neural network model.

The second stage includes determining the boundaries of objects on the received and pre-processed in stage 1 frames of the video stream and estimating the probability of their belonging to a certain class. YOLO performs classification of objects in the frame by one-stage detection, dividing the image into an S×S grid. The grid size depends on the YOLO versions and the detection level and can be: 80×80, 40×40, 20×20 cells, where each cell is responsible for predicting bounding boxes and class probabilities for objects located within its boundaries. The neural network architecture simultaneously calculates the coordinates of the frames, the degree of confidence in the presence of the object and its class affiliation, using a forward pass. After that, confidence thresholds and the Non-Maximum Suppression algorithm are applied to filter out bounding boxes and leave only the most relevant ones. Using the confidence threshold set by the user (from 0.0 to 1.0), filtering of objects occurs: only those whose probability exceeds this threshold will be selected for output to the audio stream in the following stages. If an object is detected in the frame that is identified as a person with an additional threshold value above 0.6, an additional determination of this person is performed at stage 3.

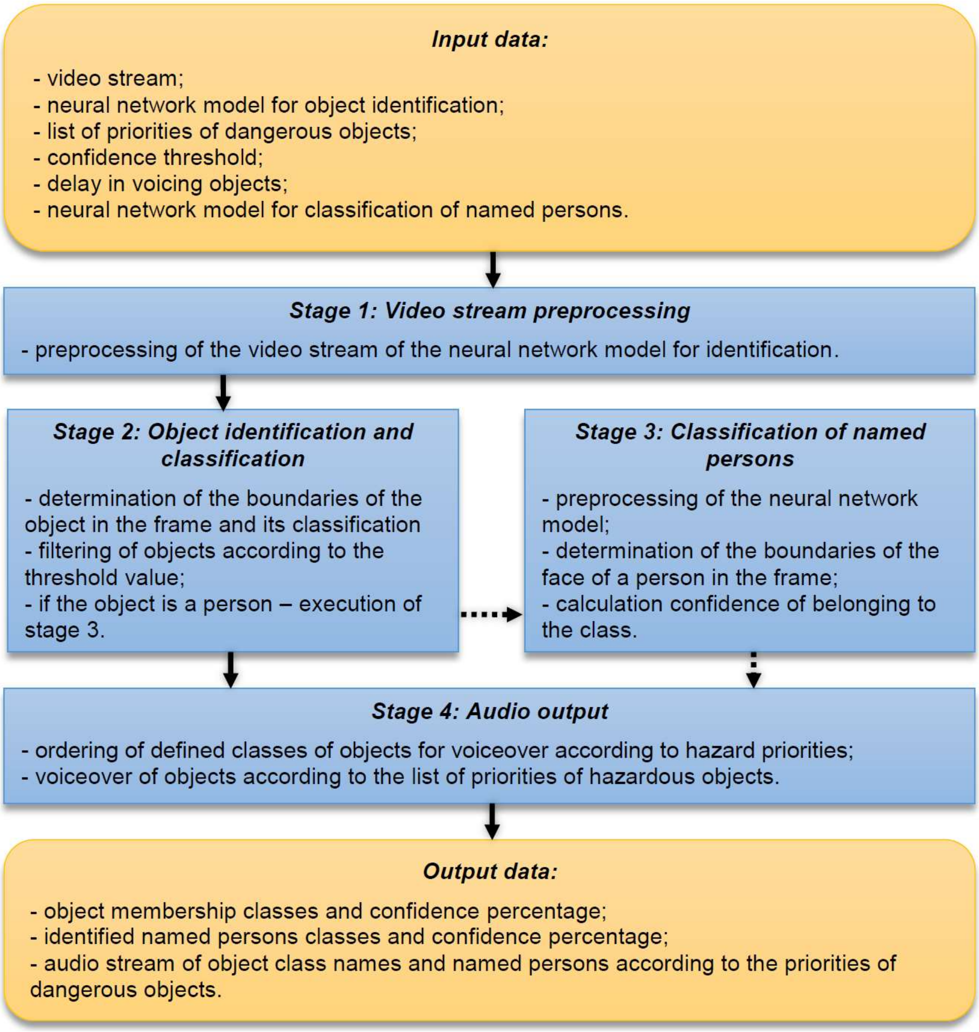

Input data:
- video stream;
- neural network model for object identification;
- list of priorities of dangerous objects;
- confidence threshold;
- delay in voicing objects;
- neural network model for classification of named persons.
Stage 1: Video stream preprocessing
- preprocessing of the video stream of the neural network model for identification.
Stage 2: Object identification and classification
- determination of the boundaries of the object in the frame and its classification
- filtering of objects according to the threshold value;
- if the object is a person – execution of stage 3.
Stage 3: Classification of named persons
- preprocessing of the neural network model;
- determination of the boundaries of the face of a person in the frame;
- calculation confidence of belonging to the class.
Stage 4: Audio output
- ordering of defined classes of objects for voiceover according to hazard priorities;
- voiceover of objects according to the list of priorities of hazardous objects.
Output data:
- object membership classes and confidence percentage;
- identified named persons classes and confidence percentage;
- audio stream of object class names and named persons according to the priorities of dangerous objects.

Figure 2: Scheme of method of augmented reality audiostream creation using CNN.

In the third stage, the named persons are classified. First, the boundary of the identified person in the frame is determined, after which the probability with which this person belongs to a specific class is calculated. The class of the named person in the next stage will be announced as "person is"+name of person. If the person could not be identified, his name will not be announced in stage 4, but "unknown person" will be announced instead. Using YOLO to identify named persons is not advisable, since this approach is optimized for fast detection of object classes, and not for accurate recognition of specific persons. YOLO performs well in identifying people in the frame, but its final layers are not designed to extract unique features necessary for identifying and classifying named persons. For the identification and classification of named persons, a CNN neural network model is used in the study. The steps involved in face identification using a cascade classifier based on Haar functions and classification of a named person using it using CNN are shown in Figure 3.

To classify a person detected in a frame from a video stream, a cascade classifier based on Haar functions is used to determine the boundaries of the face [36]. As input parameters, the cascade of classifiers receives the coordinates of the bounding box of the person that was classified by YOLO. This allows localizing the area containing the persons face, which is necessary for further classification of named persons. If the cascade classifier does not determine that there is a face in the bounding box of the person (for example, the person in the frame has his back turned, or the face is at an angle that does not allow him to be identified), then the classification of named persons does not occur, and the class name «person» is provided in the audio output. After isolating the area with the face, a CNN is used to identify the person by classifying the resulting image. The CNN generates a vector representation of the face and compares it to the named persons trained by the classifier to determine the most likely class match. This is done by using a softmax function that converts the model's output values into probabilities of belonging to each class. The maximum value of the softmax function will correspond to the identified class of the named person, which generates audio output indicating the name of the recognized person.

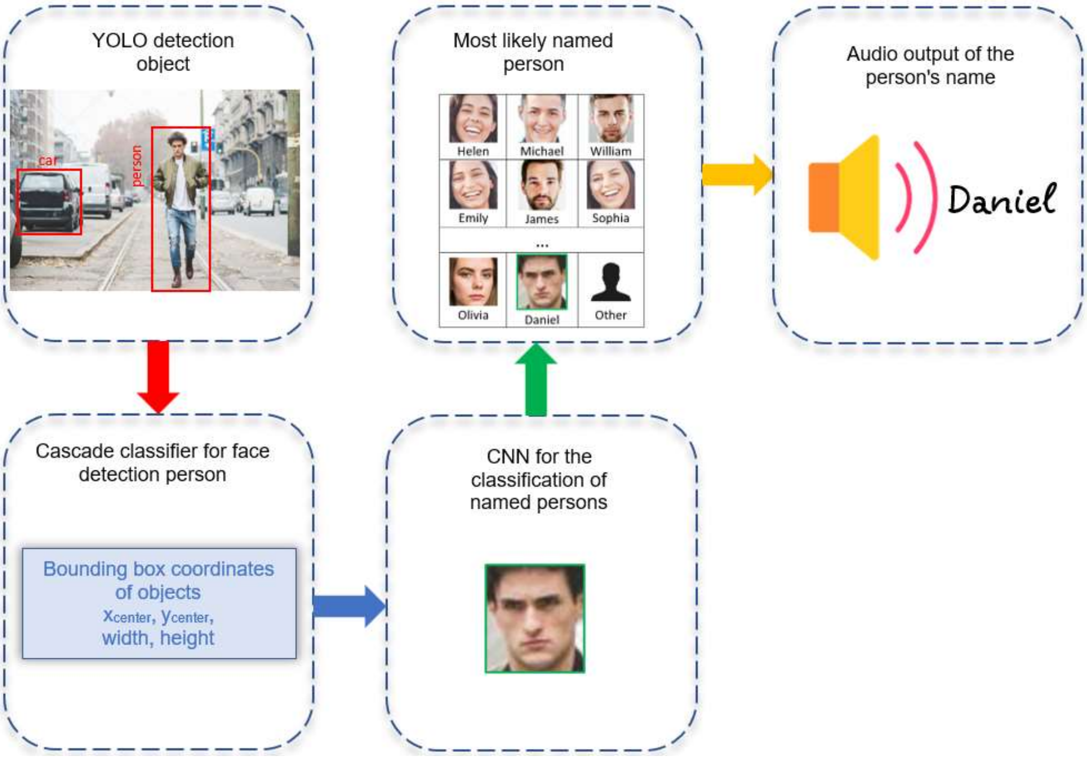

YOLO detection object
car
person
Most likely named person
Helen
Michael
William
Emily
James
Sophia
...
Olivia
Daniel
Other
Audio output of the person's name
Daniel
Cascade classifier for face detection person
Bounding box coordinates of objects
x_{center}, y_{center},
width, height
CNN for the classification of named persons

Figure 3: Steps scheme for detecting and identifying of named persons faces.

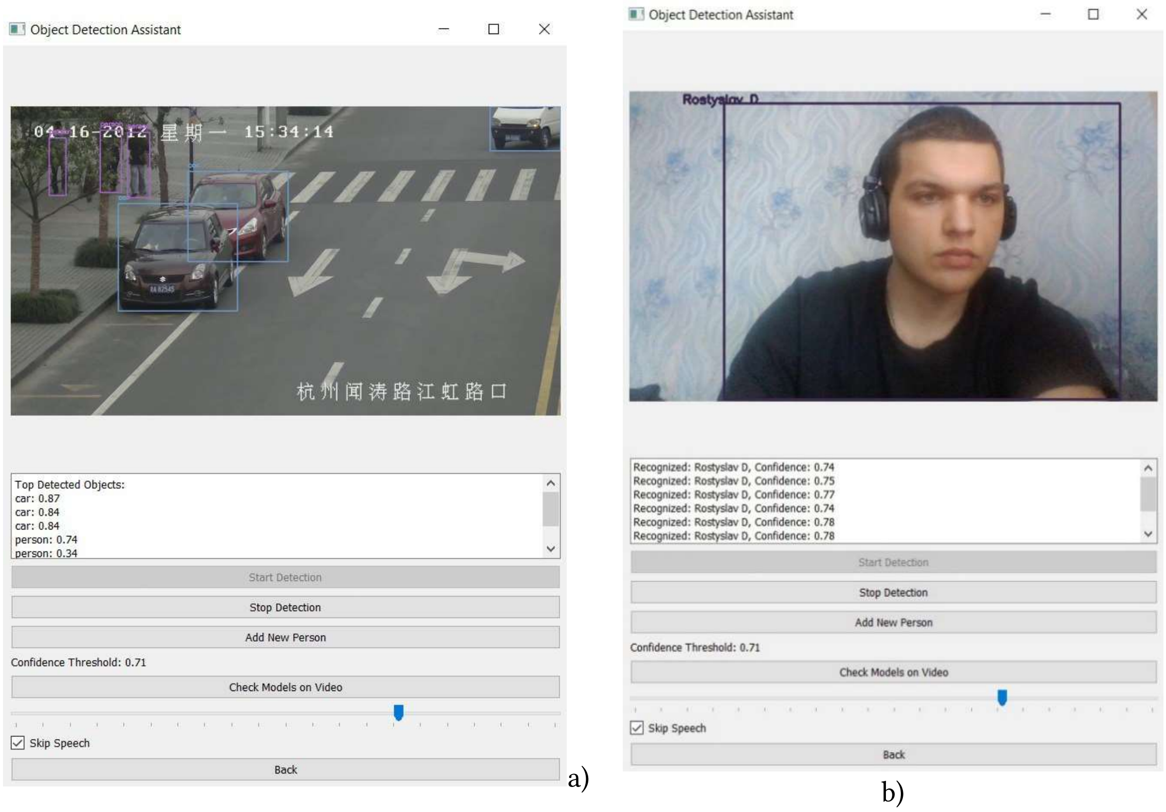
The fourth stage involves the formation of an audio stream. The classes of objects from the video stream are sorted for voiceover according to the defined priorities of danger for visually impaired people, as well as with identified named persons. It is important to note that if several objects with the same priority are classified in the frame, they will be voiced in the order in which YOLO classifies them. If the user has set a priority for the "person" class, then all persons (including named persons) will be voiced according to this priority. If there are several named persons in the frame, they will be voiced in the order of their classification. If the priority for the "person" class is not set, named persons will be voiced after all objects with a higher priority, in the order in which they were classified. The voiceover update frequency is set by the user to prevent excessive load on the perception of information. Since this study is experimental in nature, the audio stream is updated within a 5-second interval, which can be changed by the user if necessary, and the output of classified objects to the application log has no delays. As part of the study, the following list of priority objects for visually impaired people was determined: bicycle, car, motorcycle, bus, train, truck, traffic_light, stop_sign. The specified list is purely experimental and can be changed or supplemented if necessary.

The initial data of the method are classes of objects with an assessment of the confidence of their belonging to the corresponding classes, classes of classified named persons, if they have been identified, as well as an audio stream formed on the basis of the priority of dangerous objects for visually impaired people, and containing the specified classes of objects and named persons, if such are classified.

Thus, the presented method of augmented reality audiostream creation using CNN can contribute to increasing inclusion and safety for people with visual impairments.

4. Experiment

To test the presented method, it is necessary to study the existing pre-trained YOLO neural network models, as well as to investigate the influence of the batch size and the number of epochs when training the CNN model on the accuracy of the classification of named persons by test software (Figure 4).



```
Top Detected Objects:
car: 0.87
car: 0.84
car: 0.84
person: 0.74
person: 0.34
```

Confidence Threshold: 0.71

☑ Skip Speech

a)

```
Recognized: Rostyslav D, Confidence: 0.74
Recognized: Rostyslav D, Confidence: 0.75
Recognized: Rostyslav D, Confidence: 0.77
Recognized: Rostyslav D, Confidence: 0.74
Recognized: Rostyslav D, Confidence: 0.78
Recognized: Rostyslav D, Confidence: 0.78
```

Confidence Threshold: 0.71

☑ Skip Speech

b)

Figure 4: Test software interface: a) object detection and classification;
b) identification of named persons (author of paper Rostyslav Dydo is identificated).

In the process of experimental research, test software was created for creating an augmented reality audio stream in Python (Figure 4), which allowed testing the proposed method of augmented reality audiostream creation using CNN.To implement this software prototype, the following Python libraries were used: «tensorflow» [37], «pandas» [38], «PyQt5» [39], «numpy» [40], «ultralytics» [41]. All experiments were performed on a computer with following configuration: Intel Core i5-11400H processor, 16 GB DDR4 (3200MT/s) RAM and NVIDIA GeForce RTX 3050 Laptop graphics processor.

To study the models, the COCO dataset was used, which contains 330,000 images (of which over 200,000 are annotated), 80 object classes [42], which was tested for balance and representativeness [43]. The results of experiments using developed test software are presented in the next section.

5. Results and Discussion

To study of object classification at stage 2 of developed method, the YOLOv5n, YOLOv5x, YOLOv8n, YOLOv8x, YOLOv11n, and YOLOv11x models were compared, which are pre-trained to classify 80 environmental objects, including, for example: car, motorcycle, traffic light, bench, dog, backpack, tennis racket, apple, chair, toothbrush, etc. The models were compared using the metrics Accuracy, Precision, Recall, etc. [44, 45]. Pre-trained objects can be divided into the following categories: people, transport, infrastructure, animals, personal belongings, sports equipment, dishes and kitchen items, food, furniture, household appliances, stationery. Figure 5 shows the results of the macrometrics Accuracy, Precision, Recall, mAP@0.5, mAP@0.5:0.95 for each of the tested YOLO models.

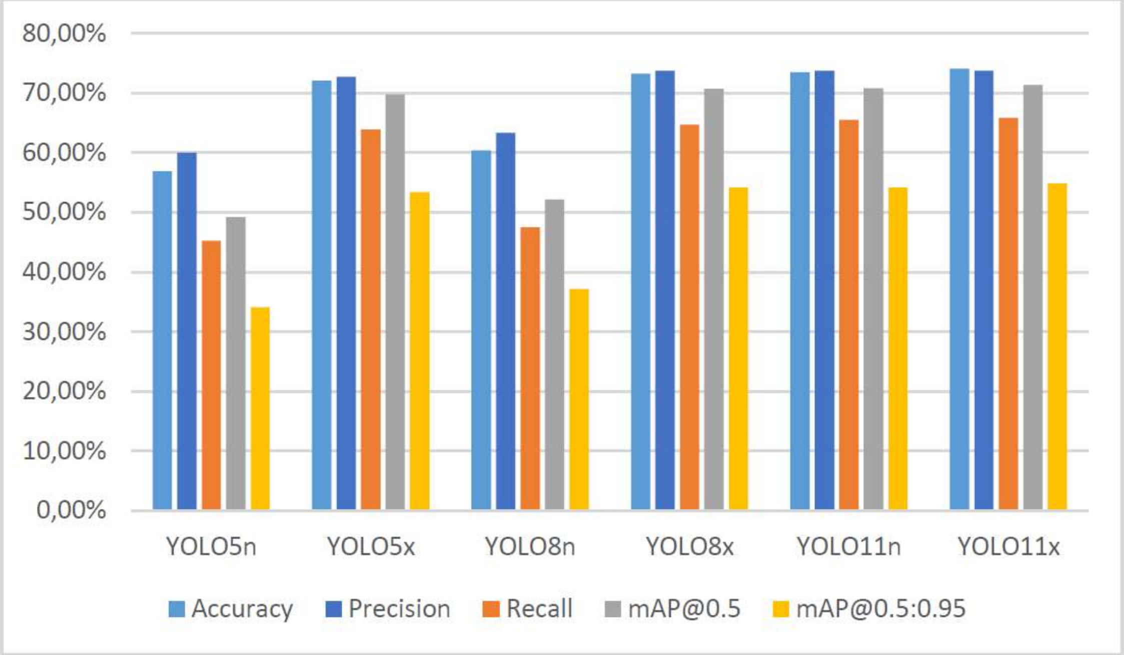


Figure 5: Histogram comparison of YOLO versions macrometrics for object classification.

The study found that YOLOv11X has the best metrics for classifying objects in a video stream. In particular, Accuracy 74.1%, Precision 73.74% and Recall at 65.85% indicate that the model can correctly identify objects. The obtained mAP@0.5 indicator (average accuracy at the IoU threshold of 0.5) is 71.34%, and mAP@0.5:0.95 has a value of 54.85%. The obtained results of these indicators indicate that the model finds objects well even with a small error in positioning. In modern computer vision systems aimed at helping people with visual impairments, one of the key indicators is the speed of real-time image processing. In particular, the frame rate per second (FPS) determines how quickly the model can analyze the environment and provide the user with relevant information. High FPS is important for ensuring continuous and accurate object recognition, especially in changing environmental conditions. As part of the study, FPS was calculated for the YOLOv5x, YOLOv8x, YOLOv11x models, which allows us to assess their suitability for real-time application. The models were tested on a video [46] containing scenes with people and cars. The results of FPS calculations are presented in Table 1.

Table 1
FPS calculation results of YOLOv5x, YOLOv8x, YOLOv11x models

Model	Frames processed	Average FPS	Total time (sec.)
YOLOv5x.pt	750	143.47	5.23
YOLOv8x.pt		155.32	4.83
YOLOv11x.pt		182.66	4.11

The YOLO model test results show a noticeable increase in FPS when moving to newer versions. In particular, the YOLOv5x model has the lowest average FPS (143.47), which means a longer frame processing time, while YOLOv8x improves this figure to 155.32 FPS, reducing the total processing time by 0.4 seconds. The best results are shown by YOLOv11x, which reaches 182.66 FPS, which is 27% faster than YOLOv5x and almost 18% faster than YOLOv8x.

The reduction in the total frame processing time from 5.23 seconds (YOLOv5x) to 4.11 seconds (YOLOv11x) indicates a significant increase in efficiency, which is critical for real-time object recognition, especially for systems aimed at assisting the visually impaired, where even small delays can affect safety and usability.

The FPS indicators given in Table 1 were calculated during the processing of the video stream by YOLO models without displaying the image on the screen and without visualizing the superimposed bounded boxes. At the same time, for the correct display of the video in the software prototype together with the visualization of the superimposed bounded boxes of the detected objects, the optimal frame rate of 30 FPS was empirically determined, which can be changed in the settings depending on the user's requirements and the computing resources of the system.

Considering the conducted studies, YOLOv11x is the best of the tested models, providing the highest indicators of Accuracy, Precision, Recall, mAP@0.5, mAP@0.5:0.95, FPS and the lowest total frame processing time, therefore, the micrometric indicators of the YOLOv11x model were further investigated in more detail. Figure 6 shows the micrometric indicators for each class of only the priority objects in terms of danger for people with visual impairments, which were listed in section 3.

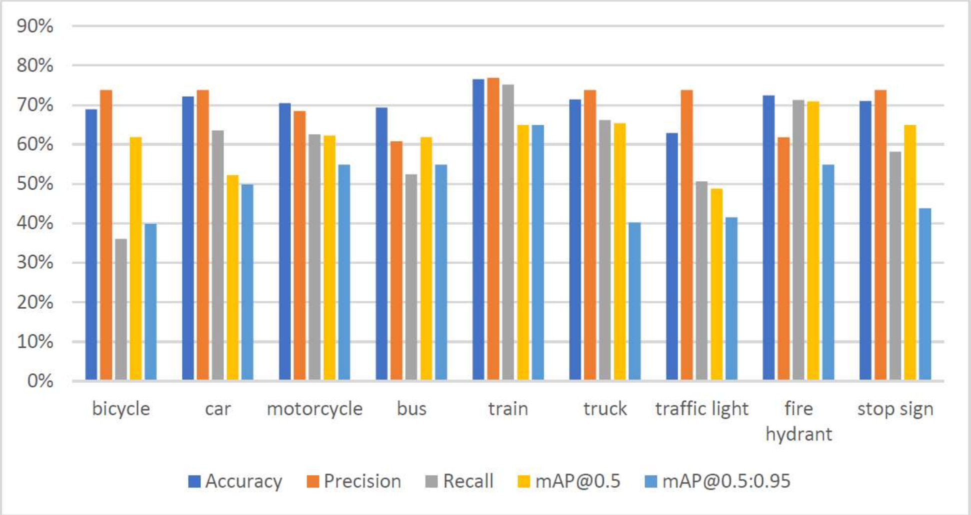


Figure 6: Micrometrics of YOLOv11X model for classes of objects prioritized by danger.

The evaluation of the YOLOv11X model performance indicates its balanced performance in classifying objects of different classes. The Precision value is stable for most classes, approaching 73.74%, which indicates the overall high ability of the model to correctly classify objects without a significant number of false positives. However, the Recall indicator shows significant variability, which may indicate the heterogeneous ability of the model to find all objects of certain categories in the input images.

The lowest Recall level is observed for the «bicycle» class (36.06%), which indicates the difficulty of model in detecting this type of object due to the peculiarities of their shape, size, or the presence of partial overlap with other objects. At the same time, the «train» class (75.13%) shows the highest Recall indicator, which indicates a high level of detection of this object in the test set.

As can be seen from Figure 6, the YOLOv11X neural network model demonstrates above-average prediction accuracy for each class, but for some classes, such as «bicycle» and «traffic light», it demonstrates lower values than for other classes, which is associated with the complexity of their detection. However, if we evaluate the model in the context of 80-class classification, the calculated indicators are good.

For the classification of named persons, the CNN neural network model was trained to identify 19 named persons and a separate class was «other», in which photos of different persons were mixed, which were taken in the amount of 50 photos from the dataset [47]. To train the neural network model, 50 face frames were taken from the video stream for each person in different positions using the software prototype interface. The total number of samples is 1000 images and all classes were used equally in the training process.

The influence of batch size and number of epochs when training the CNN model on the accuracy of classification of named persons is given in Table 2. The following macrometric indicators were obtained: Accuracy 97%, Precision 95%, Recall 96%, F1-score 95%.

Micrometrics Accuracy, Precision, Recall, F1-score are given in Table 3.

Table 2
Results of the influence of batch size and number of epochs on classification accuracy

Macrometrics	Epochs	Batch size		
		3	5	10
Validation Loss	16	11%	2%	3%
	32	24%	10%	3%
	64	41%	28%	1%
Validation Accuracy	16	94%	96%	96%
	32	89%	93%	97%
	64	79%	88%	96%

Table 3
CNN micrometric indicators Accuracy, Precision, Recall, F_1-score for recognizing named persons

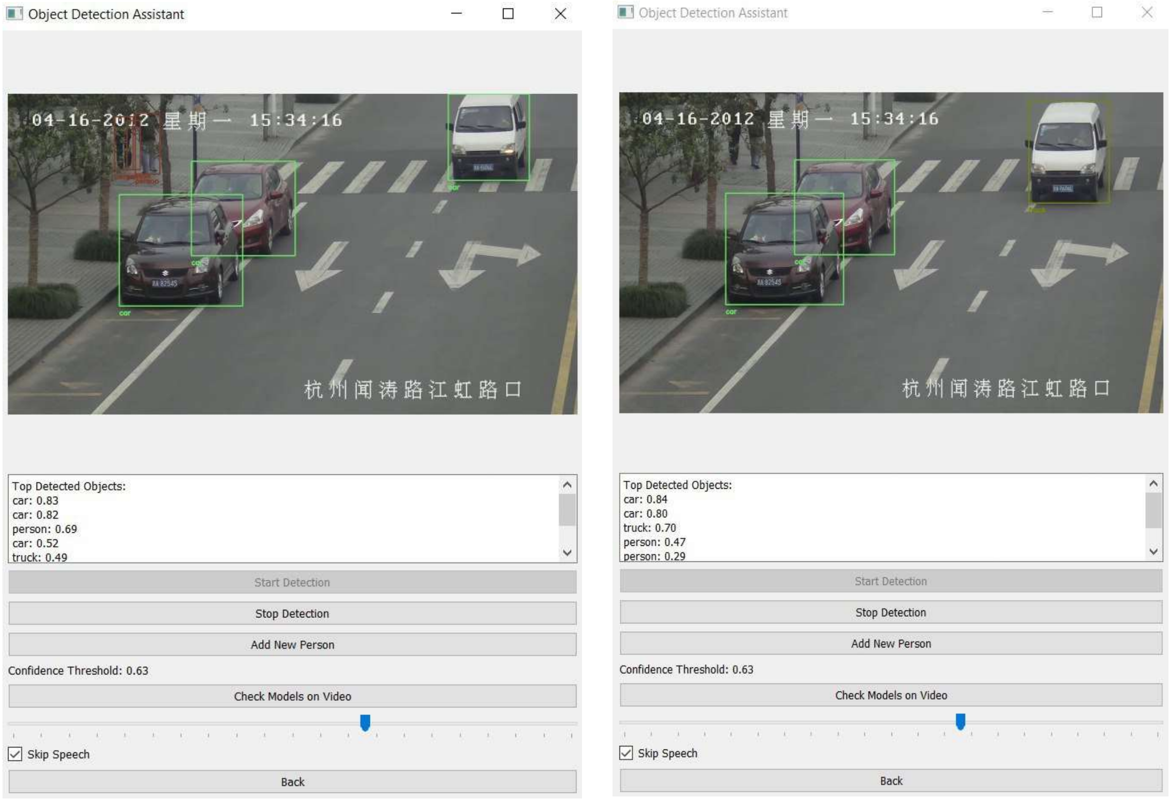
Class	Accuracy	Precision	Recall	F1-score
Person 1	98%	97%	98%	97%
Person 2	95%	96%	97%	96%
Person 3	96%	95%	96%	96%
Person 4	97%	95%	95%	95%
Person 5	96%	96%	97%	96%
Person 6	96%	95%	95%	95%
Person 7	97%	94%	95%	94%
Person 8	97%	96%	97%	96%
Person 9	97%	94%	95%	94%
Person 10	96%	94%	95%	94%
Person 11	95%	95%	96%	95%
Person 12	97%	94%	95%	94%
Person 13	96%	95%	95%	95%
Person 14	96%	96%	97%	96%
Person 15	97%	94%	95%	94%
Person 16	97%	95%	96%	95%
Person 17	97%	96%	96%	95%
Person 18	98%	96%	96%	95%
Person 19	97%	95%	95%	95%
Other	96%	96%	95%	95%

According to the results presented in Table 3, the optimal values of batch size are 32 and the number of epochs is 16, since with these parameters the highest classification accuracy of 97% is achieved.

The results of the CNN neural network model for the classification of named persons demonstrate high accuracy. In particular, the micrometrics Accuracy, Precision, Recall and F1-score indicate sufficient accuracy and completeness of recognition for each class. The obtained neural network model has higher accuracy rates, compared to [32], where the Accuracy rate reaches 96.12%, as well as compared to [33] with an Accuracy rate of 94% and [34] with an Accuracy rate of 96.5%.

The limitations of the used CNN model are that it demonstrates a decrease in the level of confidence in the classification of named persons with a significant head rotation, in particular, if the face is in a 3/4 angle or in profile. In such cases, the model cannot identify a person due to the lack of a sufficient number of recognition features. Also, glasses and masks reduce the classification confidence score, but headgear does not affect the classification accuracy.

Another major limitation is the lack of an algorithm for measuring the distance to objects in the frame. Since the model analyzes two-dimensional images, there is a problem with correctly identifying objects that are either too close or too far from the camera. Also, due to the distortion of the perspective of objects in the frame and the loss of details, such objects may be incorrectly classified or even missed. For example, a car that is at a considerable distance may be recognized as a different class due to its reduced size and distortion of proportions. An example of incorrect classification of a car is shown in Figure 7. This is also especially critical for named persons, since the accuracy of their classification directly depends on the quality and size of the face in the frame.



Top Detected Objects:
car: 0.83
car: 0.82
person: 0.69
car: 0.52
truck: 0.49

Confidence Threshold: 0.63

☑ Skip Speech

a)

Top Detected Objects:
car: 0.84
car: 0.80
truck: 0.70
person: 0.47
person: 0.29

Confidence Threshold: 0.63

☑ Skip Speech

b)

Figure 7: Classification example of the same object by YOLOv11x model at different distances: a) correct; b) incorrect.

Another limitation is the frame refresh rate and the creation of the current audio stream. The delay between frame processing can cause situations where the object has already left the field of view, but its sound is still being played. This can create problems for users, since the audio stream will not be synchronized with the current situation in the environment.

The prospect of further research is to study the method in real conditions to ensure the optimal delay of the audio stream output. As well as the use of sensors to estimate the distance to objects, this will compensate for the scaling problem to avoid false classifications for objects that are too close or far from the device camera.

6. Conclusions

Method of augmented reality audiostream creation using CNN is proposed, which can increase inclusion and safety for people with visual impairments due to additional classification of named persons and priority audio output of dangerous objects into the audio stream.

To identify objects in the video stream, the YOLOv11x neural network model was used, which allows in real time by analyzing the video stream received by the camera, to identify and classify objects from the environment, both outdoors and indoors. To classify named persons identified by the YOLOv11x model, a cascade classifier for face identification and a CNN neural network model trained for multi-class classification of named persons were used.

For applied research of the method proposed in paper, test software was developed and comparison of neural network models YOLO5n, YOLOv5x, YOLOv8n, YOLOv8x, YOLOv11n, and YOLOv11x was performed. By analyzing the obtained results, namely the macrometrics Precision, Recall, mAP@0.5, mAP@0.5:0.95 and FPS, it was determined that the YOLOv11x model is better suited for identifying the classification of objects in the environment, as it received higher indicators. In particular, the Accuracy value is 74.1%, Precision is 73.74%, which indicates the high accuracy of the model in assigning the detected objects to the correct classes. Recall, which is 65.85%, reflects the ability of the model to find a significant proportion of all objects present, although some of them may remain undetected. The mAP@0.5 value, which is 71.34%, indicates the overall quality of the model at a relatively low overlap threshold, and mAP@0.5:0.95, which is 54.85%, demonstrates its accuracy at different overlap thresholds. The calculated FPS of the YOLOv5x, YOLOv8x, YOLOv11x models also indicates that YOLOv11x, compared to others, is able to process the same number of frames in a shorter period of time.

The obtained results of macrometrics and micrometrics Accuracy, Precision, Recall and F1-score of the CNN neural network model for the task of classification of named persons demonstrate high classification accuracy. Macrometrics (Accuracy 97%, Precision 95%, Recall 96%, F1-score 95%) indicate a general high accuracy of classification of named persons. Minor variations in the indicators for individual classes indicate a minimum number of errors, which does not significantly affect the overall accuracy of classification of named persons.

Proposed method has limitation in that there is no algorithm for measuring the distance to objects, which leads to a decrease in the accuracy of their identification and classification, especially for distant objects. In addition, limited frame refresh rate can cause delay in the audio stream, due to which objects are voiced with delay or after they disappear from the view field.

Therefore, the method presented in the paper can increase the inclusion and safety of people with visual impairments. Developed method differs from existing ones by able to classify named individuals in the video stream for boosting the inclusion of visually impaired people; also, method prioritizes dangerous objects when generating audiostream for boosting the safety of visually impaired people. Considering the above, the method of augmented reality audiostream creation using CNN will contribute to achieving UNDP SDG No. 3, No. 10 and No. 11.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] WHO, Blindness and Visual Impairment, 2025. URL: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] 300, Useful Information, 2025. URL: https://0300.com.ua/cikave/useful-info.

[3] S. Cai, A. Ram, Z. Gou, M. A. W. Shaikh, Y.-A. Chen, Y. Wan, K. Hara, S. Zhao, D. Hsu, Navigating Real-World Challenges: A Quadruped Robot Guiding System for Visually Impaired People in Diverse Environments, In: CHI '24: CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2024. doi:10.1145/3613904.3642227.

[4] S. Dembitska, I. Sivert, Digital accessibility in education: challenges and prospects, Health Saf. Pedagog. 9.2 (2024) 57–63. doi:10.31649/2524-1079-2024-9-2-057-063.

[5] A. M. R. Nielsen, B. L. Due, L. Lüchow, The eye at hand: when visually impaired people distribute seeing with sensing AI, Vis. Commun. (2024). doi:10.1177/14703572241227517.

[6] L. R. d. Souza, R. Francisco, J. E. d. Rosa Tavares, J. L. V. Barbosa, Intelligent environments and assistive technologies for assisting visually impaired people: a systematic literature review, Univers. Access Inf. Soc. (2024). doi:10.1007/s10209-024-01117-y.

[7] A. Lavric, C. Beguni, E. Zadobrischi, A.-M. Cailean, S.-A. Avatamanitei, A Comprehensive Survey on Emerging Assistive Technologies for Visually Impaired Persons: Lighting the Path with Visible Light Communications and Artificial Intelligence Innovations, Sensors 24.15 (2024) 4834. doi:10.3390/s24154834.

[8] V. Moram, S. Zahruddin, S. Kumar, Multifunctional Assistive Smart Glasses for Visually Impaired, SN Comput. Sci. 6.2 (2025). doi:10.1007/s42979-025-03701-2.

[9] I. Patel, M. Kulkarni, N. Mehendale, Review of sensor-driven assistive device technologies for enhancing navigation for the visually impaired, Multimedia Tools Appl. (2023). doi:10.1007/s11042-023-17552-7.

[10] Z. J. Muhsin, R. Qahwaji, F. Ghanchi, M. Al-Taee, Review of substitutive assistive tools and technologies for people with visual impairments: recent advancements and prospects, J. Multimodal User Interfaces (2023). doi:10.1007/s12193-023-00427-4.

[11] I. Krak, O. Sobko, M. Molchanova, I. Tymofiiev, O. Mazurets, O. Barmak, Method for neural network cyberbullying detection in text content with visual analytic, CEUR Workshop Proceedings, 3917 (2025) 298-309. URL: https://ceur-ws.org/Vol-3917/paper57.pdf .

[12] I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, R. Bahrii, O. Sobko and O. Barmak, Abusive speech detection method for Ukrainian language used recurrent neural network, in: CEUR Workshop Proceedings, 3688 (2024) pp.16–28. doi: 10.31110/COLINS/2024-3/002.

[13] O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina, Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets, Lecture Notes on Data Engineering and Communications Technologies, 149 (2023) 591–607. doi:10.1007/978-3-031-16203-9_33.

[14] I. Krak, V. Didur, M. Molchanova, O. Mazurets, O. Sobko, O. Zalutska and O. Barmak, Method for political propaganda detection in internet content using recurrent neural network models ensemble, CEUR Workshop Proceedings, 3806 (2024) 312–324. URL: https://ceur-ws.org/Vol-3806/S_36_Krak.pdf.

[15] I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, E. Manziuk and O. Barmak, Method for neural network detecting propaganda techniques by markers with visual analytic, CEUR Workshop Proceedings, 3790 (2024) 158-170. URL: https://ceur-ws.org/Vol-3790/paper14.pdf.

[16] R. Szekely, C. Holloway, M. Bandukda, Understanding the Psychosocial Impact of Assistive Technologies for Blind and Partially Sighted People: Protocol for a Scoping Review (Preprint), JMIR Res. Protoc. (2024). doi:10.2196/65056.

[17] C. G. Richardson, The Role of Rehabilitation Medicine in the Psychological Etiology of Blindness or Visual Impairment: A Critical Synthesis, in: Disease and Health Research: New Insights Vol. 6, BP International, 2024, c. 62-117. doi:10.9734/bpi/dhrni/v6/859.

[18] UNDP, Ukraine, 2025. URL: https://www.undp.org/uk/ukraine.

[19] N. Radwan, M. Farouk, The Growth of Internet of Things (IoT) In The Management of Healthcare Issues and Healthcare Policy Development, Int. J. Technol., Innov. Manag. (IJTIM) 1.1 (2021) 69–84. doi:10.54489/ijtim.v1i1.8.

[20] A. Sharifi, Z. Allam, S. E. Bibri, A. R. Khavarian-Garmsir, Smart cities and sustainable development goals (SDGs): A systematic literature review of co-benefits and trade-offs, Cities 146 (2024) 104659. doi:10.1016/j.cities.2023.104659.

[21] M. Mishra, S. Desul, C. A. G. Santos, S. K. Mishra, A. H. M. Kamal, S. Goswami, A. M. Kalumba, R. Biswal, R. M. da Silva, C. A. C. dos Santos, A bibliometric analysis of sustainable development goals (SDGs): a review of progress, challenges, and opportunities, Environ., Dev. Sustain. (2023). doi:10.1007/s10668-023-03225-w.

[22] K. Ghafoor, T. Ahmad, M. Aslam, S. Wahla, Improving social interaction of the visually impaired individuals through conversational assistive technology, Int. J. Intell. Comput. Cybern. (2023). doi:10.1108/ijicc-06-2023-0147.

[23] S. Chaple, V. Raut, J. C. Patni, A. Banode, S. Ninawe, N. Shelke, Artificial Intelligence on Visually Impaired People: A Comprehensive Review, in: 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, 2024. doi:10.1109/icicv62344.2024.00052.

[24] I. Jeong, K. Kim, J. Jung, J. Cho, YOLOv8-Based XR Smart Glasses Mobility Assistive System for Aiding Outdoor Walking of Visually Impaired Individuals in South Korea, Electronics 14.3 (2025) 425. doi:10.3390/electronics14030425.

[25] M. A. Kamran, A. Orakzai, U. Noor, Y. S. Afridi, M. Sher, Visually: Assisting the Visually Impaired People Through AI-Assisted Mobility, (2025).

[26] G. I. Okolo, T. Althobaiti, N. Ramzan, Smart Assistive Navigation System for Visually Impaired People, J. Disabil. Res. 4.1 (2025). doi:10.57197/jdr-2024-0086.

[27] J. Hemavathy, A. Sabarika Shree, S. Priyanka, K. Subhashree, AI Based Voice Assisted Object Recognition for Visually Impaired Society, In: 2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), IEEE, 2023. doi:10.1109/icdsaai59313.2023.10452456.

[28] S. Sameer, P. Madan, S. Kannan, V. J. Upadhye, H. Patil, S. Rajkumar, AI-based Object Detection for Assisting the Visually Impaired People, in: 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), IEEE, 2024. doi:10.1109/icmcsi61536.2024.00080.

[29] A. G. Hagargund, S. V. Thota, M. Bera, E. F. Shaik, Image to speech conversion for visually impaired, Int. J. Latest Res. Eng. Technol. 3.06 (2017) 09–15. URL: https://www.academia.edu/download/82052515/2_B2017160.pdf

[30] D. Das, S. Roy, Object Detection with voice output for visually impaired, in: 2024 International Conference on Communication, Computing and Internet of Things (IC3IoT), IEEE, 2024. doi:10.1109/ic3iot60841.2024.10550247.

[31] Haryanto, N. Kholis, S. Hadi, E. Supriyadi, M. Ikhsani, The development of face recognition application system (SAFR) as an adaptive teaching support, in: The 6th International Conference of Ice-Elinvo 2023: Digital Solutions for Sustainable and Green Development, AIP Publishing, 2025, 020007. doi:10.1063/5.0261215.

[32] Z. Ren, X. Liu, J. Xu, Y. Zhang, M. Fang, LittleFaceNet: A Small-Sized Face Recognition Method Based on RetinaFace and AdaFace, J. Imaging 11.1 (2025) 24. doi:10.3390/jimaging11010024.

[33] N. Sabah Abbod, J. B. Mohasefi, Designing Face Detection Systems with Gray Wolf Optimization, Iraqi J. Electr. Electron. Eng. 21.2 (2025) 64–75. doi:10.37917/ijeee.21.2.7.

[34] R. K. Shukla, A. K. Tiwari, A. Ranjan Mishra, Face Recognition Using LBPH and CNN, Recent Adv. Comput. Sci. Commun. 17 (2024). doi:10.2174/0126662558282684240213062932.

[35] A. Vijayakumar, S. Vairavasundaram, YOLO-based Object Detection Models: A Review and its Applications, Multimedia Tools Appl. (2024). doi:10.1007/s11042-024-18872-y.

[36] OpenCV, Cascade Classifier Tutorial, 2025. URL: https://docs.opencv.org/4.x/db/d28/tutorial_cascade_classifier.html.

[37] TensorFlow, TensorFlow Official Website, 2025. URL: https://www.tensorflow.org/.

[38] Pandas, Pandas Official Website, 2025. URL: https://pandas.pydata.org/.

[39] PyQt5, PyQt5 Official Website, 2025. URL: https://pypi.org/project/PyQt5/.

[40] NumPy, NumPy Official Website, 2025. URL: https://numpy.org/.

[41] Ultralytics, Ultralytics Official Website, 2025. URL: https://ultralytics.com/.

[42] COCO Dataset, 2025. URL: https://cocodataset.org/.

[43] O. Sobko, O. Mazurets, M. Molchanova, I. Krak, O. Barmak, Method for analysis and formation of representative text datasets, CEUR Workshop Proceedings, 3899 (2024) 84-98. URL: https://ceur-ws.org/Vol-3899/paper9.pdf.

[44] Y. Krak, O. Barmak, O. Mazurets, The Practice Investigation of the Information Technology Efficiency for Automated Definition of Terms in the Semantic Content of Educational Materials, CEUR Workshop Proceedings, 163 (2016) 237–245. URL: doi:10.15407/pp2016.02-03.237.

[45] O. Zalutska, M. Molchanova, O. Sobko, O. Mazurets, O. Pasichnyk, O. Barmak, I. Krak, Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network, CEUR Workshop Proceedings, 3387 (2023) 344-356. URL: https://ceur-ws.org/Vol-3387/paper26.pdf.

[46] TechChannel00001, YouTube Video, 2025. URL: https://www.youtube.com/watch?v=Gr0HpDM8Ki8.

[47] Kaggle, LFW People Dataset, 2025. URL: https://www.kaggle.com/datasets/atulanandjha/lfwpeople.