An intelligent approach for creation of control systems for road sections with pedestrian crossings

V S Tynchenko¹², A S Bondarev¹, S V Apanasenko¹, D V Rogova¹, A S Yadrov² and S A Moskvin²³

¹Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russia
²Siberian Federal University, 79, Svobodny pr., Krasnoyarsk, 660041, Russia
³State Institution - Krasnoyarsk Regional Branch of the Social Insurance Fund of the Russian Federation, 117, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660010, Russia

E-mail: vadimond@mail.ru

Abstract. The article discusses the development of a system for recognizing people at a pedestrian crossing. Such recognition system includes a trained classifier and two sets of images taken from an open database containing images of city streets from outdoor cameras. Some shortcomings that can be easily solved with the help of the developed recognition system are analyzed. We also studied common methods of detection, for example, the Viola-Jones method, and their advantages in the operation of the system. A comparison of all the considered methods is carried out and the most suitable one for the set goal is selected and the method of constructing a classifier is described separately. This system is becoming more and more relevant with the development of the technological capabilities of urban life and improves its quality due to its smooth operation and good performance indicators.

1. Introduction
In large cities, in practice, centralized traffic light control systems are used, taking into account the developed cycles and phases of automobile and pedestrian traffic, for the development of which empirically collected information on flows and density of each type of traffic is used [1 - 3].

Taking into account the development of machine learning, which provides a wide range of tools for working with data in real time, this direction is a promising direction for implementation into existing systems or their expansion, allowing one to optimize or automate existing solutions [4 - 6].

Next, the existing solution to the task of adjusting the pedestrian phase will be considered, and the requirements for the system being developed will be described.

2. Recognition system for pedestrian crossing people
At the moment, there is a solution to the problem of adjusting the pedestrian phase in the form of pedestrian crossing sections (PCS): in the immediate vicinity of the pedestrian crossing. A button is installed with a connection to the controller, when pressed, the pedestrian traffic light switches to a green signal, a road traffic light – to a red signal. After a given amount of time, the traffic lights switch back. An example of such a PCS is shown in Figure 1.
Despite the simplicity of this solution, it has significant drawbacks, which are successfully solved by the developed recognition system.

First, the button design is unreliable. The buttons themselves are firm enough to press. Moreover, due to frequent use, vandalism and dampness, the PCS begins to work intermittently, leading either to the impossibility of switching when pressed, or to the constant inclusion of the pedestrian phase.

The system does not require direct pedestrian intervention, so the service life of the equipment will be higher than the corresponding PCS indicator. This also applies to acts of vandalism: the surveillance camera is out of reach for a person without special technical means, so the likelihood of equipment breakdown in this way will be much lower.

Second, PCSs do not allow maintaining the cycle length and the same number of phases, since the controllers do not support the phase selection mode according to the principle “either one or the other”.

Likewise, PCS is incompatible with traffic lights in coordinated control, when they are controlled remotely and turn on simultaneously or with a shift according to the “green wave” principle. When working in remote mode, the technical control device intercepts control of the controller installed directly on the road, including phases according to the specified personal rules. This relationship is one-way, describing a top-down relationship. At the same time, in order to be included in the nearest window, communication must be two-way, which is currently not technically implemented.

The developed system has the prospect of expanding to a network of interconnected controllers transmitting information to each other, or implementation into existing coordination systems. This will allow you to individually and centrally adjust the adjustment of the pedestrian phases for each traffic situation, both for the local controller system and for each traffic situation, as well as in accordance with the existing algorithms of the coordination systems with their expansion.

3. Description of system requirements
The system being developed is faced with the task of recognizing a human figure in a video fragment in the context of an urban environment, that is, it must be able to distinguish people from all other objects of the environment [7 - 9].

It is understood that the system can both solve the problem of detecting a human figure and fix the fact of the presence of a pedestrian in the frame for a certain amount of time to transmit information either to a traffic light or to a local or central traffic controller [10].

The recognition system requires a video surveillance camera or other video recorder above the pedestrian crossing area [11]. The device should be installed at a distance that allows the human figure to be clearly distinguished.

In addition to the reader, hardware is required to operate the system itself. It is possible to implement the developed solution into the existing coordination systems.
An important initial step is the selection of the best solution for the classification problem. The chosen method must meet the specified requirements, that is, the final decision based on the analysis results must be based on the specifics of the subject area.

This section provides an overview of common detection methods and indicates the selected development tools. Moreover, simple indicators of the efficiency of the classifier are described for further evaluation of the results obtained at the end of training.

Object recognition is a scientific discipline that aims to classify objects into several categories or classes. The recognition task is to classify the considered new, unexplored object to any given class [12]. Thus, in this context, object recognition is one of the varieties of classification of identification; that is, assigning an unambiguous name to the object in question. In other words, obtaining characteristic points on the image will allow further classification of the object under consideration.

If we consider the problem of finding objects on a digital video fragment, then it can be divided into three subtasks:

- Detection: selection of a fragment in the image that may contain the desired object.
- Recognition: determination of the found object to a certain class.
- Tracking: The found object is fixed and its position is tracked.

These tasks are reduced to the construction of an image classifier.

For the system being developed, the first two subtasks are considered taking into account the tasks assigned to the system.

Two of the most common object recognition methods are discussed below: convolutional neural networks and the Viola-Jones method.

The convolutional neural network is a special neural network architecture proposed in 1988 and aimed at efficient recognition of objects and patterns. The architecture of such neural networks models some of the features of the visual cortex of the brain: there are some simple cells that respond to straight lines at different angles, and complex ones, the work of which is associated with the launch of a specific group of simple cells.

This architecture received its name in connection with the use of the convolution operation, the essence of which is element-wise multiplication of each image fragment by the convolution kernel and then writing the result to the corresponding position of the output image (Figure 2).

![Figure 2. Convolutional neural network architecture.](image-url)

The method under consideration uses three types of layers: convolution, downsampling (pulling), and a fully connected layer.

The convolution operator forms the basis of the convolutional layer of the network. This layer consists of three levels:

- Creation of a set of linear activations by performing one or more parallel convolutions.
- Detection, that is, the application of a non-linear activation function to all linear activations.
- Spatial aggregation to modify the output for transmission to the next network layer.
A layer consists of a set of cores, and its task is to calculate the convolution of the output image from the previous layer using this set and add an offset corresponding to the core at each iteration. The essence of these processes is the addition and scaling of the input image parameters (pixels).

The pulling layer is a non-linear compaction of the feature map. The layer receives at the output separate fragments of the image and combines them into one value. There are various ways of aggregation, usually the maximum function is used. This operation can significantly reduce the spatial volume of the image and avoid overfitting.

After several passes of image convolution and compaction using the pooling operation, the system rebuilds from a specific pixel grid to more abstract feature maps, as a rule, on each next layer, the number of channels increases and the image dimension in each channel decreases. Thus, by the end, there is a large set of channels storing a small amount of data, which are interpreted as the most abstract concepts identified from the original image.

The simplest and most common learning method is supervised learning - back propagation. Convolutional neural networks have the following advantages:

- Solving problems with unknown patterns.
- Relative resistance to rotation and shift of the recognized image.
- Resistance to noise in the input data.
- Potentially fast work.
- Fault tolerance with hardware implementation of a neural network.

The Viola-Jones method was developed and presented in 2001, but it is still effective for finding objects in images and video sequences in real time. Initially, its task was to recognize faces, but the scope of possible application of the algorithm is quite wide.

A simplified diagram of the algorithm is shown in Figure 3. At the output of the algorithm, a set of found objects at different scales are given. As mentioned earlier, the classifier is trained with a teacher, i.e., there is a training sample and information indicating which of the training images contain the desired object.

The fundamental idea of the algorithm is the extraction of image features and the subsequent training of the algorithm on them. As features for the recognition algorithm, the authors proposed Haar features based on Haar wavelets.

The Haar trait is a set of adjacent rectangular regions, which are divided into two types. There are a lot of possible Haar features (various combinations of regions of different widths and heights with different positions in the image). In the standard Viola-Jones method, the features presented in Figure 4, called Haar primitives, are used.

To obtain the value of a certain feature for the image under consideration, it is necessary to add the brightness of the pixels in both groups of rectangular regions separately, and then subtract the second from the first resulting sum. The resulting difference is the value of a particular Haar feature for the image under consideration.
Calculating the value of each Haar feature for an image requires many addition operations. For quick calculation of the total brightness, the calculation operations are performed using integral representations of images.

In the integral form of the image, the value of each pixel is the sum of the brightness of that pixel and all the pixels that are above and to the left of it (in case the pixel with unit coordinates is located in the upper left corner of the image). By translating the image into an integral form, you can calculate the values of Haar features for it, without performing the summation of all the required brightness values each time anew. It is enough to calculate the sum of the brightness for each rectangular area using the properties of the integral shape of the image.

Boosting is a set of methods that improve the accuracy of analytical models. Boosting literally means "strengthening" of "weak" models - it is a procedure for sequentially building a composition of machine learning algorithms, when each next algorithm seeks to compensate for the shortcomings of the composition of all previous algorithms. The idea of boosting was proposed by Robert Shapir in the late 90s, when it was necessary to find a solution to the problem of obtaining one good one, having many slightly different learning algorithms.

Thus, upon completion of training, the cascade shown in figure 5 will be obtained, consisting of a chain of classifiers (layers) trained using the boosting procedure.

Figure 3. Generalized recognition scheme in the Viola-Jones algorithm.

Figure 4. Haar primitives.
Figure 5. Cascade structure.

The Viola-Jones method has the following advantages:
- The ability to detect a large number of objects in the original image.
- Good speed of work.
- The possibility of training the classifier on the location of any object.
- Checking a feature at a certain position in the image takes linear time due to the use of the integral representation of the image.
- A large number of implementations in various open-source libraries and in various programming languages.
- A small number of false positives when searching for an object in the image.

Based on the results of the analysis of the considered methods, we can say that both methods are suitable for solving the problem of classifying pedestrians in a video stream.

Both algorithms are fast, with high detection rates. Both the convolutional neural network and the Viola-Jones method require a lot of time and resources in the learning process, so this factor, like the speed of work, cannot be considered as decisive.

However, one of the advantages of the Viola-Jones method is the low rate of false positives, which is a critical feature in the work being done. It is assumed that a security camera or other video recording device will have a static field of view. Thus, in case of false classification of a third-party relatively static object (plant, road sign, etc.), the resulting classifier will have no practical application: a signal will be transmitted on an ongoing basis to switch the pedestrian green traffic light in the absence of a pedestrian, which corresponds to normal operation controllers at the moment without implementing the system.

For the above reason, the Viola-Jones method was chosen for the developed system.

4. Conclusion
As a result of this work, a review of development tools for control systems for road sections with pedestrian crossings was carried out. To implement the classifier, the methods of object detection were considered. The Viola-Jones method was selected for the most suitable for the specifics of the task at hand, and a method for constructing a classifier based on the AdaBoost algorithm was described. The considered classifier is offered to systems for regulating traffic sections with pedestrian crossings.

References
[1] Datola G, Bottero M and De Angelis E 2020 Addressing Social Inclusion Within Urban Resilience: A System Dynamics Approach In: INTERNATIONAL SYMPOSIUM: New Metropolitan Perspectives pp 510-9
[2] Dong C Z, Bas S and Catbas F N 2020 A portable monitoring approach using cameras and computer vision for bridge load rating in smart cities Journal of Civil Structural Health Monitoring 10(5) 1001-21
[3] Oc T and Tiesdell S 1998 City centre management and safer city centres: approaches in Coventry and Nottingham Cities 15(2) 85-103
[4] Obedin A V, Soroka E O, Kukartsev V V, Mikhailov A S, Tynchenko V S, Semenova E I and Bashmur K A 2019 The developing program system of social monitoring of road improvement and urban infrastructure *Journal of Physics: Conference Series* 1399(5) 055021

[5] Ocot T and Tiesdell S 1999 The fortress, the panoptic, the regulatory and the animated: Planning and urban design approaches to safer city centres *Landscape Research* 24(3) 265-86

[6] Ballesteros J, Rahman M, Carbunar B and Rishe N 2012 Safe cities. A participatory sensing approach *In: 37th Annual IEEE Conference on Local Computer Networks* pp 626-34

[7] Tubaishat A and Elsayed R 2017 Safe city: Introducing a flawless model and implementation guidelines to reduce crime rate and response time *In: Proceedings of the 29th International Business Information Management Association Conference - Education Excellence and Innovation Management through Vision 2020: From Regional Development Sustainability to Global Economic Growth* pp 474-85

[8] Li Y H, Zhang X D, Feng Z C and Luan Z H 2014 Research on the Key Equipment Performance of Stage Video Surveillance and Scheduling System Based on IP Network *Applied Mechanics and Materials* 644 957-60

[9] Idimechev I A, Mikhailov A S, Goponenko A A, Stavenko S S, Kukartsev V V and Tynchenko V S 2020 Comparative analysis of background subtraction algorithms in the task of recognizing and identifying people in a video stream *Journal of Physics: Conference Series* 1661(1) 012041

[10] Zeng X, Guo H and Hu W 2019 Design and Implementation of Shipping Video Surveillance Equipment Based on Raspberry Pi *In: 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)* pp 66-70

[11] Pan X, Hu H, Xu J and Li M 2020 Research on Video Surveillance Robot Based on SSH Reverse Tunnel Technology *In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* pp 298-302

[12] Kukartsev V V, Mikhailov A S, Tarasevich A V, Tynchenko V S, Ogol A R and Khramkov V V 2019 Selection of appropriate architecture and parameters of neural network for images recognition and classification *Journal of Physics: Conference Series* 1399(3) 033105