ANALYSIS OF THE SITUATION ON THE OBSERVED SCENE CONTAINING MANY MOVING OBJECTS

D. S. Girenko [1] *, N. V. Kim [2]

[1]PAWLIN Technologies Ltd, Dubna, Russian Federation - dima@webiceberg.com
[2]Moscow Aviation Institute, National Research University, Russian Federation - nkim2011@list.ru

Commission II, WG II/8

KEY WORDS: Artificial Vision Systems (AVS), Automatic situation control, Deep learning, Scene understanding, Transformer

ABSTRACT:

In many processes, control is based on the classification of the current states of objects of interest or situations and the selection of appropriate control actions. In the presented work, the task of classifying dynamic situations based on the analysis of the received video sequence is considered, for example, when monitoring traffic, observing the behavior of a crowd of people or animals, etc. abnormal, dangerous, etc.). At the same time, the classification of situations by separate static images of similar observed scenes can be difficult, physically unrealizable or impractical. The aim of the study is to increase the efficiency of process management by using some features (signs) of dynamic situations for classification. Automatic classification of current dynamic situations in such processes will allow timely organization of measures to correct undesirable development of situations and/or reduce possible predicted losses. A technique for classifying dynamic situations based on the analysis of motion parameters identified by the transformer network and subsequent classification implemented by the perceptron is presented. As a demonstration example, the classification of street situations determined by people's behavior is investigated. Examples of distinguished classes of situations are given, confirming the possibility of implementing the proposed methodology.

1. Introduction

Processes are considered, the control of which is based on the classification of current situations, determined by the state of objects on the observed scenes, and the choice of appropriate control actions.

The aim of the study is to increase the efficiency of managing processes in which the classification of situations by individual frames (static images) can be difficult, physically unrealizable or impractical. We assume that such processes include dynamic processes. The observed scenes of dynamic processes contain many objects, the nature of the movement of which is determined by the class of the current situation (regular, abnormal, dangerous, etc.).

The classification of situations necessary to control the processes under study is generally based on situation analysis technologies (Abrosimov et al., 2010 & Endsley and Jones, 2016 & Fridman and Kulik, 2019). The choice of methodology for conducting a situation analysis requires determining the components of each situation. As shown in the article (Abrosimov et al, 2010), situational awareness includes awareness of processes in the environment, in order to understand how information, events and one's own actions will affect goals and objectives in the current and immediate moment. Insufficient or incorrect awareness of the situation is one of the factors associated, in particular, with accidents that are caused by the "human factor".

In many cases, the analysis of situations conducted as part of situational awareness (Endsley and Jones, 2016 & Fridman and Kulik, 2019) is based on the study of static events with a small number of moving objects. Thus, visual analysis can use images of observed scenes in the post-processing mode of occurred events (Kim and Chervonenkis, 2015). Such an analysis and further forecast of the development of situations may turn out to be too late for effective management or may give erroneous results, since it does not take into account the dynamics of previous events.

The widespread introduction of modern stationary and mobile surveillance systems, including aviation monitoring systems, significantly expands the range of constantly observed processes and in many cases makes it possible to obtain video materials of the development of many different ones, incl. dynamic situations in real time. Thus, in the event of a road traffic accident (RTA), depending on the class of emergency, ambulances, tow trucks or fire trucks can be used to eliminate the consequences of an accident (Kim and Chervonenkis, 2015). Automatic classification of accidents will reduce the time of calling the relevant services and improve the efficiency of assistance to victims.

Situation analysis is used in a wide range of tasks, incl. in video analytics tasks that require the study of dynamic situations, in particular, the tasks of determining the state of mobile objects, behavior analysis, and others. Classification based on the analysis of dynamic situations makes it possible to evaluate the predicted developments of events, taking into account various options for implemented controls.

According to (Fridman and Kulik, 2019), there are three main elements in the concept of “situational awareness”: a) the formation of information about the surrounding situation in time and space; b) understanding the meaning of the situation and c) predicting the development of the situation, their own actions and the actions of other participants.

We accept that the process under study is the situation on the street, from the point of view of the behavior (actions) of pedestrians. Traffic accidents are not considered.

When undesirable situations arise, management (to participate in them) can be carried out by sending medical workers or law enforcement officers.

2. Methodology for classifying dynamic situations

The considered problem of classification of situations refers to the tasks solved by recognition systems. In traditional methodologies associated with the development of recognition systems (Forssyth and Ponce, 2003 & Kim, 2001 & Visilter et al., 2010), the following stages of development are discussed:

1. Definition of classes of situations;
2. Formation of a dictionary of features;
3. Description of classes of situations in the language of features;
4. Splitting the a priori feature space into areas corresponding to the a priori alphabet of classes;
5. Choice of recognition algorithms that allow to attribute the recognized object to the appropriate class.

The implementation of stages 2, 3 of this technique in a number of practical cases is difficult due to the complexity of identifying individual features and their instability under the influence of various destabilizing factors.

The complexity of determining the position of decisive boundaries (stage 4), in particular, when using statistical methods (Kim, 2001), is associated with the need for a preliminary assessment of a priori probabilities of possible outcomes and detailed descriptions of the features used. The development of modern neural network technologies allows solving these problems based on the use of appropriate training samples.

In the general case, the 1st stage of the methodology (determining the classes of situations) is based on the target task being solved by the decision maker. Various approaches are possible in the formation of classes included in the working classification. So, one of the options is to classify situations into three intuitive classes, divided according to the degree of danger to humans: safe - a regular situation or the absence of a special situation of other classes; dangerous - situations and processes leading to the potential for injury; catastrophic - situations and processes with a relatively high probability of human casualties.

In specific cases, when forming alphabets of classes of situations, it is necessary to take into account: possible options for the development of situations and available management resources. In this paper, as a demonstration example, situations on the street related to the behavior of pedestrians are considered. In accordance with the available management resources, we define the following classes of situations:

1) Class 1 - regular situation. Control actions: no.
2) Class 2 - emergency situation "accident". Control actions: call "Rescue Service".
3) Class 3 - emergency situation "fight". Control actions: call the "Police".
4) Class 4 - non-standard "indeterminate". The situation class is not defined. Control actions: participation of a human operator in an additional expert assessment of the situation.

The main features that characterize the state of objects of interest in dynamic scenes are the parameters of the movement of objects (value and direction of speed) and their position in space.

Feature extraction on video sequences can be implemented by various methods. Thus, algorithms based on the determination of the motion vectors of the optical flow (Vaswani et al., 2017) can be used with further classification of these vectors. The complexity of this approach lies in determining the required tolerances for the values of features for each class of situation. These tolerances determine the position of the decisive boundaries that divide the feature space into separate classes.

In statistical methods, the calculation of decisive boundaries requires the formation of appropriate process models, which in many cases is quite complicated. An approach based on the use of neural network technologies seems to be more effective.

In this paper, it is proposed to use a transformer architecture network that is able to find relationships in image sequences. An observation scenario is considered as an example, when the parameters of the video camera from which the shooting is performed are known, the video camera is stationary.

The following algorithm for solving the situation classification problem is proposed, which ensures the implementation of stages 3, 4, 5 (Fig.1)

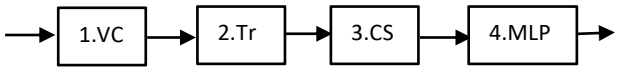

Figure 1. Situation classification algorithm

1. Obtaining a video recording of the observed situation (VC - video camera).
2. The transformer (Tr) selects and tracks objects on the received video sequence; at the output of the transformer, a list of objects and their corresponding trajectories are provided. At the preparation stage, the transformer was trained to identify objects of interest - people.
3. The resulting trajectories (in the image coordinate system) are recalculated (CS block) into motion parameters (direction and velocity for each scene object) in a flat metric coordinate system related to the position of the receiving video camera. The recalculation is made taking into account the parameters of the video camera and the conditions of observation.
4. The recalculated motion parameters and the position of each object are fed into the multilayer perceptron (MLP), which binds the observed situation to one of the previously defined classes.

At the output of the algorithm, information about the class of the observed situation is displayed.

3. EXPERIMENTS AND RESULTS

3.1. Data for testing the approach

For the class of situations “regular situation”, the MOT17 sample was used, which represents the flow of people on the street.

For the class of “accident” situations, video recordings containing people falling to the floor were used (Eraso et al., 2022). Also used were several videos obtained from city surveillance cameras, with situations when a person faints.

For the class of situations “fight”, a sample was used (Soliman et al., 2019) containing many examples of violence. It is worth

noting that a fight means exactly the active influence of the objects of observation on each other, and not “beating”.

3.2. Transformer

To obtain the trajectories of objects, it was decided to use a kind of transformer network that solves the problem of tracking people. This type of networks occupies one of the leading positions, such as (Yihong et al., 2021).

In this work, the TrackFormer network (Meinhardt et al., 2021) is used. This network was launched on Windows. The weights posted by the author were used, trained on the MOT17 sample using pre-training on CrowdHuman.

3.3. Obtaining motion parameters

The following parameters are used as attributes:

V – object speed (meter/second)
o - object orientation in space
υ - The number of frames on which the direction of movement of the object continuously changes

During operation, TrackFormer returns object trajectories in the form of a text table with the following content: Image number, object number, coordinates of the circumscribing rectangle.

Classification is carried out on the basis of an assessment of the movement of an object in space and an assessment of the orientation of the object in the image according to the following formula:

$$o = \frac{h}{w}, \quad (1)$$

где: h - the height of the object in the image, w - the width of the object in the image

However, for cases where the angle between the beam directed from the camera to the object and the vertical is less than 45 degrees, the orientation is calculated by the formula:

$$o = \frac{w}{h}, \quad (2)$$

3.4. Accounting for observation parameters

Obviously, the features of the object in the image depend on the parameters of the surveillance system, in particular, the position of the video camera relative to the observed scene. The resulting image distortions make it impossible to correctly compare the parameters obtained under different shooting conditions. Let's estimate the speed of objects in a flat metric coordinate system, the origin of which is at the camera's projection point on the plane: the y axis is directed along the projection of the camera's optical axis, the x axis is directed perpendicular to the right.

To calculate the required data, the following are used: the installation height of the video camera, the angle of inclination of the optical axis relative to the horizon plane, the image resolution, the viewing angles of the video camera, and the shooting frequency.

Consider the projection of an object in a previously defined flat coordinate system. Since the TrackFormer for each object returns a rectangle (describing the given object), we will assume that the bottom side of the rectangle is always on the plane. Then the position of the object on the plane is determined by the following coordinates:

$$\begin{cases} x = x_{rect} + \frac{w}{2} \\ y = y_{rect} + h \end{cases}, \quad (3)$$

where: x, y - coordinates of a point indicating the position of the object on the plane; x_{rect}, y_{rect} - circumscribing rectangle coordinates; w - the width of the circumscribing rectangle; h - the height of the bounding box.

Then the speed of approach (removal) of the object relative to the camera:

$$V_y = \frac{D_2 - D_1}{\Delta} = \frac{\frac{H}{atan(\gamma_1)} - \frac{H}{atan(\gamma_2)}}{\Delta}, \quad (4)$$

where: V_y - displacement of the object during the observation period, D_1 - distance from the camera to the object in the horizontal plane at the initial time of observation, D_2 - distance from the camera to the object in the horizontal plane at the end time, Δ - the number of frames elapsed between the start and end time of observation, γ_1 - the angle of inclination between the beam directed from the camera to the plane through the object at the initial moment of time and the projection of the beam onto the horizontal plane, γ_2 - the angle of inclination between the beam directed from the camera to the plane through the object at the final moment of time and the projection of the beam onto the horizontal plane.

Angle γ can be estimated from the following equation:

$$\gamma = \gamma_{уст} + \frac{y - cy}{res_y} * FOV_y, \quad (5)$$

where: $\gamma_{уст}$ - camera installation angle, cy - image center coordinate, res_y - vertical image resolution FOV_y - vertical field of view of the camera.

The lateral velocity of the object in the plane, observed in the image, has the following dependence:

$$V_x = \frac{D_1}{\Delta} * \sin(\alpha), \quad (6)$$

where: V_x - lateral speed of an object, α - the angle in the horizontal plane between the ranges at the start and end times.
Thus, at $\Delta = 1$ we get the absolute displacement of the object (in metric coordinates) for the time elapsed between adjacent frames:

$$V_{frame} = \sqrt{Vx^2 + Vy^2}, \quad (7)$$

where: V_{frame} – is the absolute displacement of the object.

At the final stage of calculating the speed of an object, it is necessary to take into account the frequency of saving frames per second (fps):

$$V = V_{frame} * fps. \quad (8)$$

3.5. Situation classification

As shown, to classify dynamic situations, it is necessary to select objects in images and determine their features: motion

parameters and orientation in space. The classification is made on the basis of a comparison of the obtained values of features with the areas of acceptable values in the feature space.
In the presented examples, the boundary values of the regions were determined on the basis of expert assessments for each attribute.
Consider the following examples of situation classes:

A characteristic sign of a regular situation is the stability of the direction of movement of objects, with low values of speed. The examples below show data from the MOT17 sample (Fig 2, 5) and extracted feature values (Fig 3, 4, 6, 7).

Figure 2. Test 1, network detections

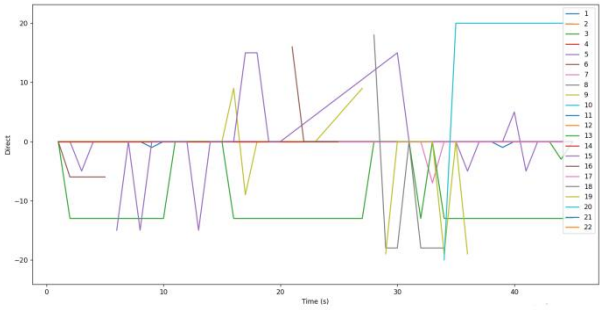

Figure 3. Test 1, the dependence of the direction of movement on time

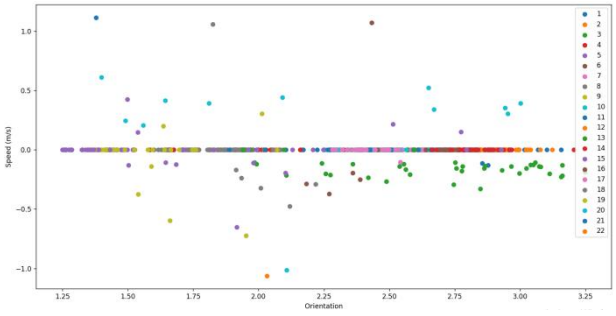

Figure 4. Test 1, the dependence of the speed by orientation

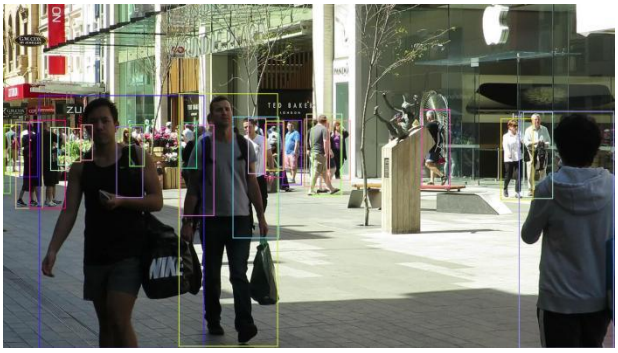

Figure 5. Test 2, network detections

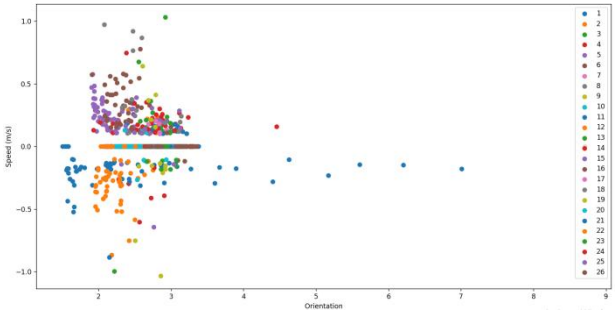

Figure 6. Test 2, the dependence of the direction of movement on time

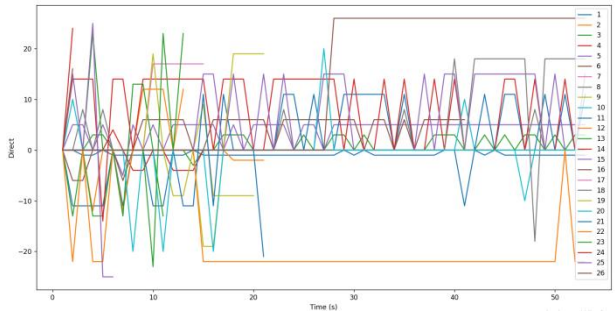

Figure 7. Test 2, the dependence of the speed by orientation

We keep the boundary values for this class of situation, not taking into account isolated cases of exceeding the speed limit:

$$\begin{cases} o >= 1 \\ abs(V) < 1 \end{cases} \quad (9)$$

A characteristic feature of the “accident” situation is a change in the orientation of the object, with low velocities after the change. The first example is from a street camera (Fig 8), the second is taken from the previously specified sample (Fig 11). The selected values of features are shown in Fig 9, 10 and Fig 12, 13, respectively.

Figure 8. Test 3, network detections

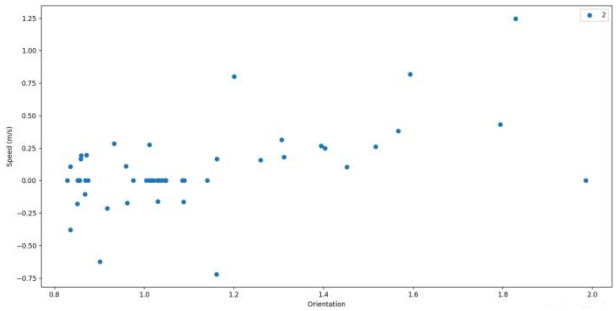

Figure 9. Test 3, the dependence of the direction of movement on time

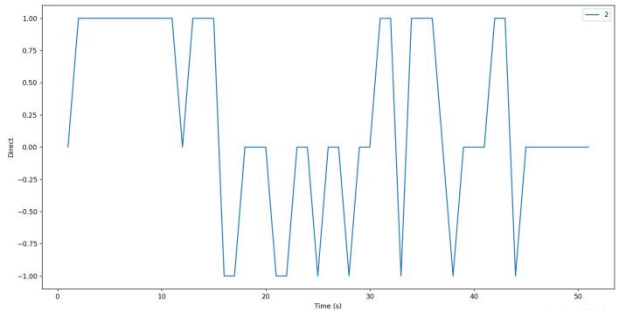

Figure 10. Test 3, the dependence of the speed by orientation

Figure 11. Test 4, network detections

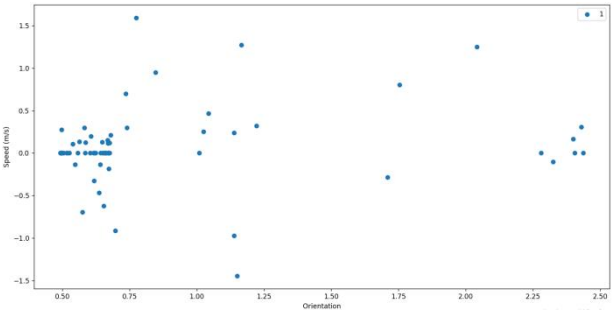

Figure 12. Test 4, the dependence of the direction of movement on time

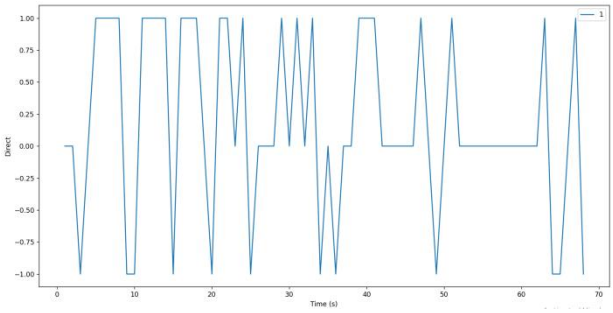

Figure 13. Test 4, the dependence of the speed by orientation

We determine the boundary values for this class of situation, namely, the satisfaction of the following condition of the object parameters for at least half of the observed time:

$$\begin{cases} o < 1 \\ abs(V) < 0.5 \end{cases} \quad (10)$$

A characteristic feature of the “fight” situation is a frequent change in the direction of movement of objects with high velocities. The figures below show data from the previously mentioned sample containing examples of violence (Fig 14, 17) and highlighted feature values (Fig 15, 16, 18, 19).

Figure 14. Test 5, network detections

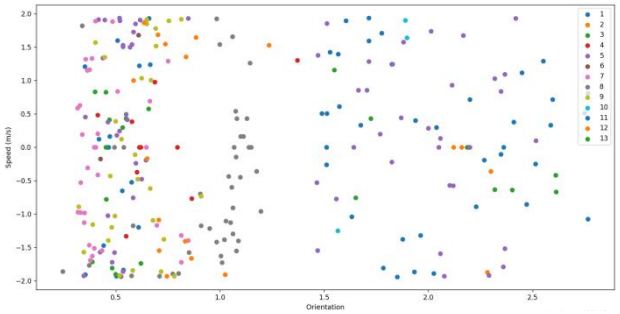

Figure 15. Test 5, the dependence of the direction of movement on time

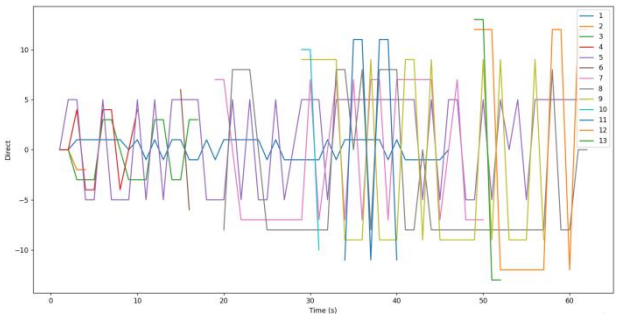

Figure 16. Test 5, the dependence of the speed by orientation

Figure 17. Test 6, network detections

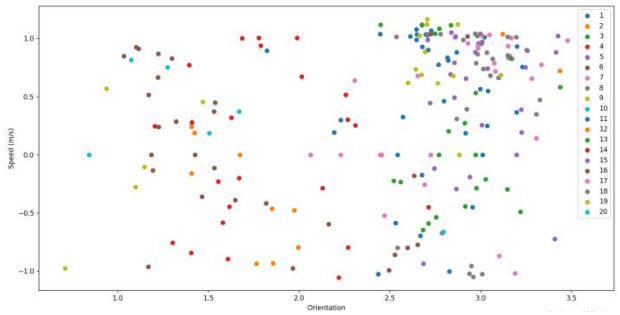

Figure 18. Test 6, the dependence of the direction of movement on time

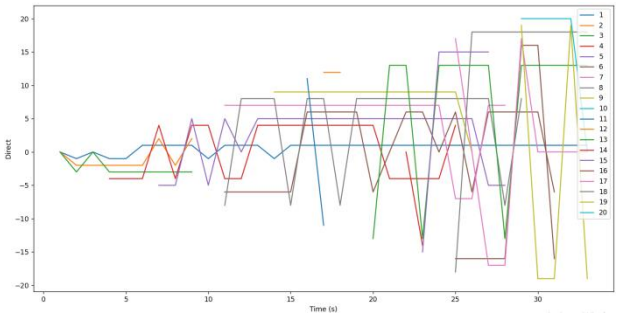

Figure 19. Test 6, the dependence of the speed by orientation

We define the boundary values for this class of situation that are inherent in at least one object:

$$\begin{cases} \upsilon > 5 \\ abs(V) >= 1 \end{cases} \quad (11)$$

To determine specific boundaries between situations, it is proposed to feed the existing data set into a multilayer perceptron. In this case, uncertain situations will be considered those that are characterized by overestimated or underestimated values of features.

4. CONCLUSIONS

1. In the work, a method was proposed for classifying dynamic situations on observed scenes containing many moving objects.
2. As an example, the dynamic situations of the street situation, observed by a fixed video camera, are considered. A classification of situations, the objects of which are people, has been carried out. 4 classes of situations are distinguished.
3. Based on the results of the work trained on the MOT17 sample of the TrackFormer network and the algorithm for converting the position of an object from the image coordinate system to a flat metric coordinate system, a classification of situations on test data was carried out, as a result of which logical rules were obtained that determine the positions of the boundaries of the class areas in the feature space.
4. Conducted experiments confirm the efficiency of the developed algorithms and programs.
5. As part of improving the efficiency of the proposed approach, it is planned to improve the architecture of the algorithm, as well as expand the database for training the neural network of the transformer and perceptron.

5. REFERENCES

1. Abrosimov V., Gaydin M., 2010, SIMULATION MODEL OF THE SITUATIONAL AWARENESS FORMATION IN A GROUP OF AUTONOMOUS ROBOTS IN THE POTENTIAL THREATS, IZVESTIYA SFedU. ENGINEERING SCIENCES № 3 (104), pp. 14-20
2. Bichen W., Chenfeng X., Xiaoliang D., Alvin W., Peizhao Z., Masayoshi T., Kurt K., Vajda P., 2020, Visual Transformers: Token-based Image Representation and Processing for Computer Vision, https://doi.org/10.48550/arXiv.2006.03677
3. Eraso, Jose C., Muñoz E.; Muñoz M., Pinto J., 2022, “Dataset CAUCAFall”, Mendeley Data, V4, doi: 10.17632/7w7fccy7ky.4
4. Endsley M., Jones D., 2016, Designing for Situation Awareness. CRC Press.

5. Fridman A, Kulik B., 2019., QUANTITATIVE EVALUATION OF SITUATIONAL AWARENESS IN CONCEPTUAL MODELLING SYSTEM., System Analysis in Engineering and Control № 3, pp. 449-460
6. Forssyth D., Ponce J., 2003, Computer Vision: a Modern Approach, Prentice Hall.
7. Kim N., 2001, Image processing and analysis in technical vision systems: Textbook, Moscow, Russian Federation.
8. Kim N., Chervonenkis M., 2015, Situation Control of Unmanned Aerial Vehicles for Road Traffic Monitoring. Modern Applied Science. 9(5), http://dx.doi.org/10.5539/mas.v9n5p1
9. Meinhardt T., Kirillov A., Leal-Taixé, L., & Feichtenhofer C., 2021, TrackFormer: Multi-Object Tracking with Transformers, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8834-8844.
10. Soliman M., Kamal M., Nashed M., Mostafa Y., Chawky B., Khattab D., 2019, “ Violence Recognition from Videos using Deep Learning Techniques”, Proc. 9th International Conference on Intelligent Computing and Information Systems (ICICIS'19), Cairo, pp. 79-84.
11. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Llion J., Aidan G., Lukasz K., Polosukhin I., 2017, Attention Is All You Need, https://doi.org/10.48550/arXiv.1706.03762
12. Visilter Yu., Zheltov S., Bondarenko A., Ososkov M., Morzhin A, 2010, Image processing and analysis in machine vision problems, Moscow, Russian Federation.
13. Yihong X., Yutong B., Guillaume D., Chuang G., Daniela R., Xavier A., 2021, TransCenter: Transformers with Dense Queries for Multiple-Object Tracking, IEEE Transactions on Pattern Analysis & Machine Intelligence, № 1, pp. 1-16, doi: 10.1109/TPAMI.2022.3225078