Automatic Detection and Tracking of Objects in Video Data with Global Motion

Nataliia A. Obukhova, Alexandr A. Motyko, Alexandr A. Pozdeev, Konstantin A. Smirnov
Saint Petersburg Electrotechnical University "LETI"
Saint Petersburg, Russia
{natalia172419, motyko.alexandr}@yandex.ru, puches4@gmail.com, KonstantinAndSmi@yandex.ru

***Abstract*—The automatic method of non-point objects of interest detection and tracking in video data obtained by a video camera mounted on a mobile carrier is proposed. Additional features of the problem are a non-uniform background, the presence of objects overlapping with the background and each other, significant and rapid changes in the object of interest size of. The automatic detection is based on a convolutional neural network with YOLO architecture. Due to limitations on computing resources, object tracking is implemented without neural network solutions. To ensure stable tracking, several detectors are used simultaneously with subsequent analysis of the obtained data. The tracking stage is based on a detector based on histograms of oriented gradients (HOG), supplemented by a detector based on correlation filtering and motion trajectory prediction based on the Kalman filter.**

The proposed method allows detecting and successfully tracking objects at the distance of 1500 meters with an object projection size on the frame 5 x 5 pixels in conditions of global movement, non-uniform background and significant dynamics of object of interest properties. At the automatic detection stage TPR averaged over all video files participating in the experiments corresponds to 0.81, FPR corresponds to 0.10. At the tracking stage, the failure rate (tracking failures) is $6*10^{-5}$

I. Introduction

In recent years, increased attention has been paid to television systems and optoelectronic complexes for automatic detection and tracking of moving objects. This is due to the emergence of high-quality video sensors and new hardware solutions with high computational power and high performance. The new capabilities made it possible to implement complex image processing methods in real-time, which increased tactical and technical characteristics of television systems. At the same time, the problem of automatic object detection and tracking in video data remains one of the most challenging and not fully solved.

This article proposes a method for automatically detection and tracking objects of interest in complex background conditions, where objects may overlap with the background and each other, and with insufficient lighting.

The significant difficulties of task are connected with the observation conditions and include:

- The placement of the video sensor on a moving carrier. The movement of the carrier has a complex trajectory with sharp changes of the movement direction.
- Significant and rapid changes in the size of the object of interest, ranging from 5*5 pixels to 200*200 pixels (at a resolution of 1920*1080).

An additional requirement during the tracking stage is a limitation on computational resources.

II. Automatic object detection

By the result of automatic object detection, we understand the object selecting using a strobe, the center of which corresponds to the center of mass of the pixels in the object's projection plane. The borders of the strobe represent a bounding rectangle around the object on the image.

Object detection can be realized using methods based on discriminant features (brightness, color, texture, and motion), as well as deep learning methods.

When detecting objects based on brightness features, the presence of a uniform background is assumed [1], [2]. Detecting objects of interest on a non-uniform background poses significant challenges and is not very effective.

Texture features and the algorithms developed based on them [3], [4], [5] enable solving tasks of automatic detection and robust tracking, the method based on histogram of oriented gradients [6] is the most efficient approach.

In television (TV) systems for object detection and tracking, information about the object is represented as a set of frames. This allows to track the dynamics of the object of interest.

The most common motion feature is evaluated using the absolute difference of TV signals, known as motion energy.

The absolute difference is a scalar estimation. However, it is not possible to separate images of objects that are in close proximity to each other or to resolve situations where objects overlap (occlusion). The presence of global motion assumed in the solving task poses a significant problem.

The alternative way to estimate the motion feature is through motion vectors. After projecting the real three-dimensional object motion onto the frame plane, it is reflected as two-dimensional motion, which can be estimated by discrete displacements of image fragments - an optical flow field or a field of motion vectors. Having information about the direction and magnitude of the displacement allows:

- To detect objects on a non-uniform background that are in close proximity to each other.

- To resolve occlusion situations during tracking by identifying the foreground object.

The main problem with using optical flow vectors is the high probability of obtaining anomalous vectors that do not reflect the true motion, which occur on weakly-textured image fragments.

Currently, the best performance in general object detection tasks is achieved by methods based on convolutional neural networks.

A convolutional neural network can be represented as two sequentially executed blocks: an automatic feature extractor from images and a decision-making block that performs the target analysis (detection, classification, identification, etc.). An illustration for a typical network architecture is shown in Fig. 1.

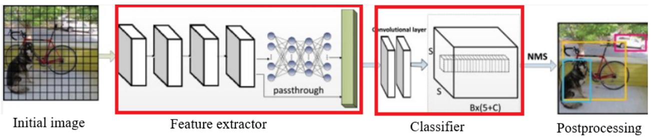

Fig. 1. Example of convolutional neural network architecture (YOLO)

Using convolutions with multiple filter kernels, features are extracted, which are then fed into a classical fully connected neural network (in general, the decision-making block can be built on algorithms of various types).

To automatic evaluation the parameters of the feature extractor filters and the decision-making block, a training dataset is required. This dataset consists of marked video data, where the target responses are known for a series of input images. It is worth noting that for training modern convolutional neural networks, the size of the training dataset needs to be sufficiently large: solving a typical industrial problem requires a minimum of 50.000 to 100.000 marked images.

Among the neural network architectures designed for detection, the YOLO-type architectures stand out (the latest version being 8, with versions 5-8 actively used). This architecture provides the best tradeoff between detection accuracy and speed [7]. Fig. 2 illustrates the advantage of YOLO version 1 (2017) over contemporary competitors of this model. The current versions of the model for 2023 are even more significantly superior to competitors.

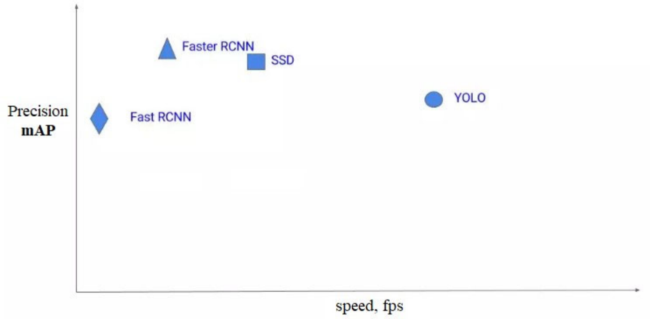

Fig. 2. Comparative advantage of YOLO architecture

Based on the above, the most efficient approach within the scope of the problem is to use convolutional neural networks, the texture feature with an estimation based on the histogram of oriented gradients, and the motion feature estimated based on optical flow.

Within the framework of the problem to be solved, a comparative study of these approaches was carried out. Based on the results obtained, a neural network solution - Yolo8 architecture - was selected for automatic detection of the object of interest.

The resolution of the frames received from the video camera corresponds to the FHD standard. Using large-size images for model training is inefficient: there is a problem with memory shortage, the procedure is slow, and the resulting model is «heavy». Image scaling, which is traditionally used in such cases, is impossible, as it results in the loss of small-sized objects of interest and the model will not be able to detect them later.

In this regard, the detection model was trained on images of standard size for Yolo v8 architecture - 320 by 320 pixels. These images were obtained by slicing the original images into fragments. The fragmentation algorithm contained special logic to prevent “cutting” of target objects. Thus, the fragments were formed around the marked objects of interest existing in the training sample. In addition, to ensure the balance of the sample, negative fragments were prepared - containing backgrounds of various kinds in the absence of objects of interest.

The initial set contained more than 120000 images, including data from sensors of different types. After the decorrelation procedure the data volume of training sample amounted to 43000 images, validation sample 10000.

After fragmentation to 320*320 size, the training samples set was 55000 and the validation samples set was 15000 images.

After automated sample analysis and cleaning of data with incorrect or inaccurate markup, the output was 27000 images for model training and 8000 for validation. After adding negative examples, the final output was: 87000 for training, 20000 for validation.

The following results were obtained by estimating the accuracy parameters of the models from the validation samples.

For the model that performs detection of objects of interest in the image, Fig. 3-5 show graphs illustrating the obtained characteristics.

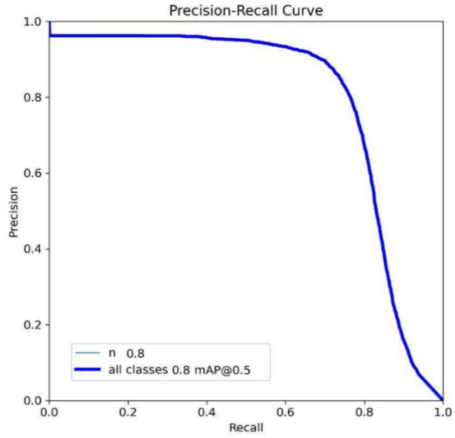

Fig. 3. Precision-recall curve for the model performing detection

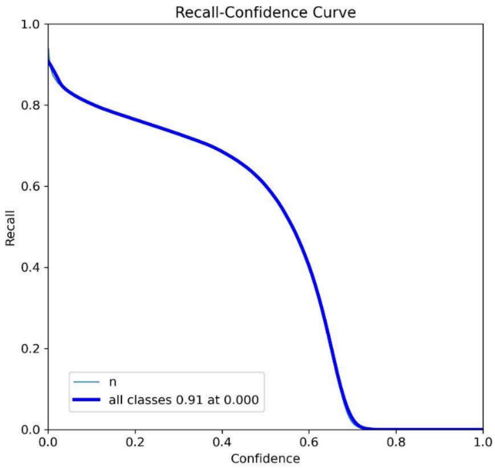

Fig. 4. Recall-confidence curve for the model performing detection

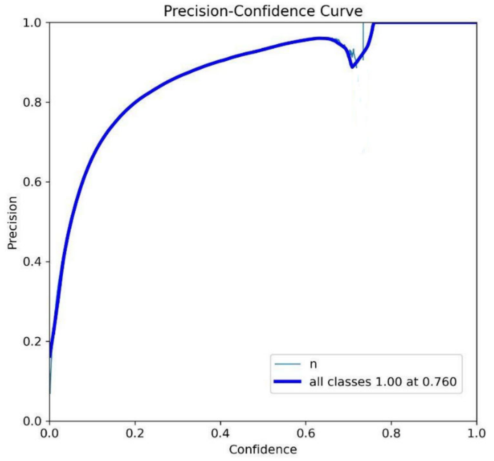

Fig. 5. Precision-confidence curve for the model performing detection

For the detection model, the mAP metric (average accuracy when the confidence parameter value is 0.5) reached a value of 0.8.

The obtained precision characteristics are high enough taking into account that the detection result is not “final”. Based on the results of the neural network detector, a strobe is formed for further automatic tracking of the object of interest.

III. Tracking

A. Basic principles of the method of automatic tracking of objects of interest on a complex background

Analysis of work in the field of video analytics highlight two main approaches to the task of tracking: algorithms based on convolutional neural networks and non-neural network "classical" algorithms. Algorithms based on neural networks show good results, however, their advantage compared to analogues is not as significant as it is for detection and classification tasks. At the same time, convolutional neural networks require significant computing resources, which is a problem when deploying such solutions in real systems with resource constraints. Therefore, approaches that do not use neural networks were used to solve the formulated problem with the limits of computational complexity.

Among such approaches, approaches based on correlation filtering and based on histograms of oriented gradients can be distinguished [8].

In correlation-based approach, the tracking is performed by comparing the current image with a reference image. The reference image is recorded when the operator (or an automatic detection algorithm) forms a region of interest for the object. It contains video information not only about the object but also about the background within the region of interest. The deviation of the current position of the object from the previous position and the error signal are determined by comparing the reference and current images using correlation filtering (correlation decision function).

In the process of object tracking, it is essential to not only detect the object but also create a model of the changes in its parameters. This model allows for predicting how the object's parameters will evolve over time and enables tracking even when the object is temporarily obscured. The most effective tool for implementing object tracking is the Kalman filter. The widespread use of the Kalman filter in tracking tasks can be attributed to its ability to extract valuable information from noisy data. Mathematically, this model involves two stages of computation. The first stage, the prediction stage, involves predicting the value of the object position based on previous values. The second stage, the correction stage, involves refining the predicted value and estimating the error. The refinement is based on the measurement of current data about the object of interest.

For the problem being solved in a two-dimensional space, the state vectors of the system (the object of interest) and the control vector have the following form:

$$\vec{x} = \begin{bmatrix} x \\ y \\ \dot{x} \\ \dot{y} \end{bmatrix}, \tag{1}$$

$$\vec{u}_k = \begin{bmatrix} \ddot{x} \\ \ddot{y} \end{bmatrix}, \tag{2}$$

where x , y are coordinates of the object of interest center, $\dot{x}$, $\dot{y}$ are the object of interest velocities of, $\ddot{x}$, $\ddot{y}$ are the object of interest accelerations .

The expressions for the prediction stage are given by:

$$\hat{\vec{x}}_{k|k-1} = \mathbf{F}_k \vec{x}_{k-1|k-1} + \mathbf{B}_k \vec{u}_k , \tag{3}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}_k , \tag{4}$$

where $\mathbf{P}_k$ is the covariance matrix that describes the relationship between the object's position and velocity, $\mathbf{Q}_k$ is the covariance matrix of the noise.

The system evolution matrix $\mathbf{F}_k$ and control matrices $\mathbf{B}_k$:

$$\mathbf{F}_k = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (5)$$

$$\mathbf{B}_k = \begin{bmatrix} \frac{\Delta t^2}{2} & 0 \\ 0 & \frac{\Delta t^2}{2} \\ \Delta t & 0 \\ 0 & \Delta t \end{bmatrix}. \quad (6)$$

At the stage of measuring and correcting the prediction:

$$\hat{\vec{x}}_{k|k} = \hat{\vec{x}}_{k|k-1} + \mathbf{K}_k \tilde{\vec{y}}_k, \quad (7)$$

$$\mathbf{P}_{k|k} = [1 - \mathbf{K}_k \mathbf{H}_k] \mathbf{P}_{k|k-1}, \quad (8)$$

$$\tilde{\vec{y}}_k = \vec{z}_k - \mathbf{H}_k \hat{\vec{x}}_{k|k-1}, \quad (9)$$

where $\vec{z}_k$ is the obtained position of the object of interest as a result of detection, $\mathbf{H}_k$ is a matrix that characterizes the obtaining of information from different sensors.

Within the framework of the task being solved:

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \quad (10)$$

The Kalman Gain:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}_k^T \mathbf{S}_k^{-1}, \quad (11)$$

$$\mathbf{S}_k = \mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k. \quad (12)$$

The matrix $\mathbf{R}$ is initialized with the squares of deviations along the corresponding coordinate:

$$\mathbf{R} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}. \quad (13)$$

The peculiarities of the solved problem are non-uniform background, significant dynamics of the object of interest properties, global motion in the frame caused by the placement of the video camera on a moving carrier, at the same time there are limitations on computational resources.

Subject to limitations on computational resources, it is advisable to abandon the neural network detector, and to use several detectors simultaneously to ensure stable tracking. A histogram oriented gradient (HOG) based detector is used as the basis for the tracking phase, it is complemented by a correlation filtering based detector and Kalman filter based motion trajectory prediction. The need to supplement the HOG-based detector with a correlation detector is due to the unstable operation of the HOG detector when the object of interest size o is sharply dynamic.

Thus, the tracking method at each frame has three data sets to analyze: the correlation matching result, the histogram-based oriented gradient extraction result (HOG detector), and the predicted parameter estimates based on the Kalman filter. Each dataset includes the position of the object of interest center and its size.

The comparison of detection and prediction results from the Kalman filter-based model is performed according to the Intersection Over Union (IOU) measure:

$$IOU = \frac{|A \cap B|}{|A| \cup |B|}, \quad (14)$$

where A and B are rectangular regions representing the detection or prediction result; $|\cdot|$ is the power of the corresponding set (area of the rectangular region); $\cap$, $\cup$ is the intersection and union of sets, respectively.

To ensure stable tracking in conditions of sharp changes in the object of interest size and dynamic maneuvering of the carrier on which the video sensor is installed, the following algorithm for analyzing the received data is proposed.

If the data of both detectors and the estimates predicted by the Kalman filter coincide, the correlation detector is reinitialized. By reinitialization we will understand updating the content of the reference image based on the current video data. If the data obtained by the correlation detector and the HOG detector coincide, but significantly diverge from the estimates predicted by the Kalman filter, it means the presence of a significant global motion (sharp carrier maneuver) and requires reinitialization of the Kalman filter parameters. If the data from the Kalman filter and the correlation detector match, but the HOG and correlation detector results have a discrepancy, it means that the object of interest is approaching with a significant increase in its size. In this case, to prevent the loss of the object of interest, the tracking is implemented only on the basis of the correlation filter and reinitialize the HOG detector. Transition to correlation tracking is realized by threshold limitation:

$$area\_thr \cdot S_c > S_{tr}, \quad (15)$$

where S_c is the response area of the correlation detector, S_{tr} is the response area of the HOG detector, $area\_thr$ is the threshold limit.

The $area\_thr$ value is determined based on the following considerations. A high value of $area\_thr$ threshold allows to increase the quality of tracking when the target is approaching, but significantly reduces the quality of tracking on a non-uniform background. In this case there is often switching to the correlation tracker and reinitialization of the CSRT-tracker. Therefore, it is recommended to make the threshold value equal to 0.6-0.7 to provide a balance between the quality of tracking when the object is approaching quickly and against a non-uniform background.

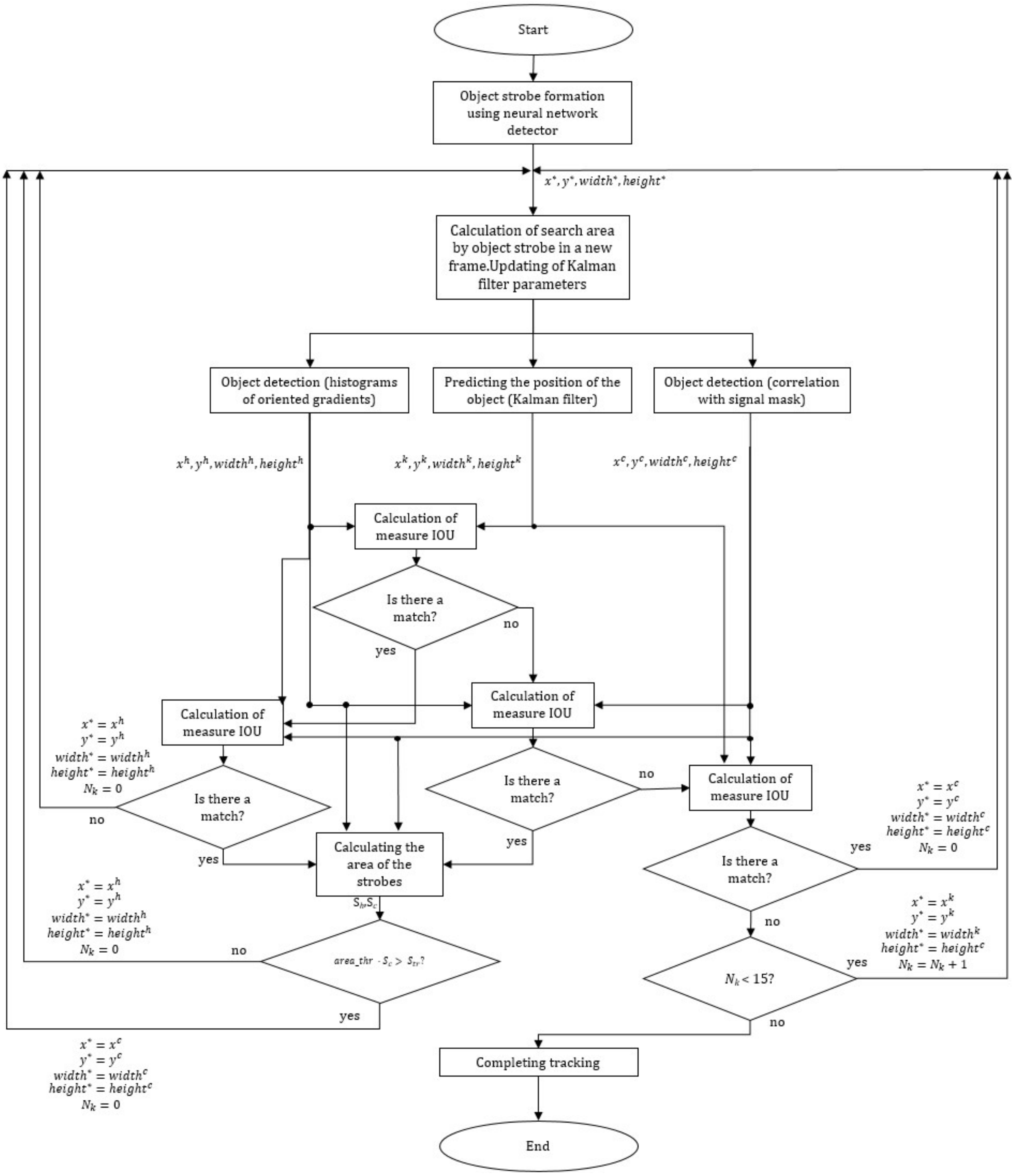
Start
Object strobe formation using neural network detector
$x^*, y^*, width^*, height^*$
Calculation of search area by object strobe in a new frame.Updating of Kalman filter parameters
Object detection (histograms of oriented gradients)
Predicting the position of the object (Kalman filter)
Object detection (correlation with signal mask)
$x^h, y^h, width^h, height^h$
$x^k, y^k, width^k, height^k$
$x^c, y^c, width^c, height^c$
Calculation of measure IOU
Is there a match?
no
yes
Calculation of measure IOU
Calculation of measure IOU
$x^* = x^h$
$y^* = y^h$
$width^* = width^h$
$height^* = height^h$
$N_k = 0$
Is there a match?
no
yes
Is there a match?
no
yes
Calculation of measure IOU
Calculating the area of the strobes
S_h, S_c
$x^* = x^c$
$y^* = y^c$
$width^* = width^c$
$height^* = height^c$
$N_k = 0$
yes
Is there a match?
no
$x^* = x^h$
$y^* = y^h$
$width^* = width^h$
$height^* = height^h$
$N_k = 0$
no
$area\_thr \cdot S_c > S_{tr}?$
yes
$x^* = x^k$
$y^* = y^k$
$width^* = width^k$
$height^* = height^c$
$N_k = N_k + 1$
yes
$N_k < 15?$
no
$x^* = x^c$
$y^* = y^c$
$width^* = width^c$
$height^* = height^c$
$N_k = 0$
Completing tracking
End

Fig. 6. Block diagram of the proposed tracking algorithm

The condition for tracking termination is the absence of coincidence of data sets from detectors and Kalman filter in more than *N* frames. To fix this situation, a special counter N_k is provided. The counter value N_k increases if all three data sets for analysis (the result of HOG and correlation detector and Kalman filter) do not coincide and tracking is performed on the basis of Kalman filter (tracking "by memory"). If a match is detected and the result of one of the detectors is returned for further tracking, the N_k counter is reset to zero.

The block diagram of the proposed algorithm is shown in Fig.6.

B. Features of software implementation of the proposed tracking method

In the software solution of the proposed algorithm, Channel and Spatial Reliability Tracker (CSRT) [9] is used to implement the detector based on histogram of oriented gradients (HOG). CSRT uses Color Names attributes in addition to HOG to describe the detected object. HOG reflects information about texture and shapes, while Color Names reflect information about color, which helps to use a more complete representation of the object. The use of Color Names, instead of the classical three-component description allows to achieve greater stability to changes in the background color or lighting conditions.

The main idea of CSRT lies in the calculation of spatial and channel reliability. By channels in this case we mean HOG and Color Names feature channels [10]. The authors of [9] use multiple HOG channels to represent texture and edge features of an object more accurately. Different channels may include different gradient orientations and different mesh sizes. For each channel, a weighting factor is calculated to characterize its degree of reliability for object tracking.

The spatial reliability map reflects the “confidence” that certain areas of an object are reliable for tracking. This map helps the algorithm to focus on those areas of the object that better match the original features and ignore areas that may be distorted or hidden, and better track non-rectangular shaped objects.

KCF (Kernelized Correlation Filters) [11] was used as a correlation detector in the implementation of the proposed method. KCF operation is based on the application of correlation filters, which are trained on the basis of object examples in different positions and scales. Unlike the classical correlation method, KCF is highly computationally efficient by utilizing the properties of cyclic matrix and fast Fourier transform (FFT) [11]. One of the key features of KCF is the use of the “kernel trick”, which allows transforming the input data into a higher dimensional space where objects are more easily distinguishable.

The experiments have shown that the best quality of tracking with increasing target size is provided by KCF tracker, and tracking of an object on a complex background is provided by CSRT tracker. High reliability of object tracking on a complex background is due to the built-in learning algorithm in CSRT tracker. This approach in its leads to the fact that reinitialization of CSRT tracker leads to accuracy decrease on the first few frames, so this procedure should be performed only if it is strongly necessary.

To improve the reliability of tracking after reinitialization of the CSRT tracker, the following procedure is proposed. The result of the correlation tracker without reinitialization of the CSRT tracker is taken as the main one. If the KCF and CSRT results coincide during the processing of subsequent frames, the CSRT tracker becomes the main tracker again. Otherwise the CSRT tracker is reinitialized again.

IV. Experimental research

The main objective of the experimental study of the automatic detection method is to evaluate the probability of correct detection at a given distance. The probability of correct detection was evaluated for a distance of 1500m.

Video data was prepared with the following characteristics:

- The size of the object of interest was 30cm * 30cm * 30cm.
- The area of the object of interest ranged from 5*5 to 200*200 pixels for a 1920*1080 pixel, with a signal-to-noise ratio of at least 40 dB and a brightness contrast of the object-background at least 15%.
- The meteorological visibility distance was at least 1500m.
- The illumination level was at least 500 lux.

A test sensor with a resolution 1920*1080, variable focal length and field of view angles from 2.3 to 63 degrees was used to capture video data. The object of interest used in the experiments had size 20 cm * 18 cm * 5.5 cm. The distance from the object of interest in the experiments was 100 meters.

When video data capturing, the focal length of the camera lens was changed so that the field of view angle corresponded to the value from Table I. This ensured that the object of interest image size obtained when the object of interest was at the 100 meters distance was the same as for an object of interest with size no larger than 30 cm x 30 cm x 30 cm at the distance of 1500 meters when using a camera with a narrow viewing angle.

Video data is captured for the task of the maximum detection distance at two different distances (Table I). Examples of frames from test videos are shown in the Fig. 7.

TABLE I. Parameters of range and object size for the experimental study

Parameter	Detection Camera Parameters	Parameters of the test camera in detection mode
Field of view angle	6 degrees	27.5 degrees
Distance No.1	1500 meters	100 meters
The size of the object at distance No.1	7.3 pixels	7.35 pixels
Distance No.2	2230 meters	150 meters
The size of the object at distance No.2	4.9 pixels	4.9 pixels

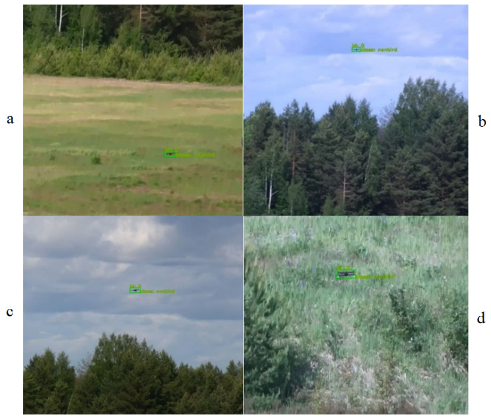

Fig. 7. Examples of frames from test videos where: a – object on a complex background (field), b, c – object on a simple background (sky), there is a brief flight of the object over the horizon line to a complex background, d – object on a complex background (field), there are moments when the object merges with the background

For each video, the object of interest was automatically detected in each frame. According to the results of the frame analysis, the following dates were recorded: the frame number and the presence/absence of detection. When an object of interest is detected, its coordinates are recorded.

Based on the obtained data, an estimation of the correct detection probability TPR is calculated, an estimate of the probability of a false alarm FPR at a given detection range:

$$TPR = TP / QP, \tag{16}$$

$$FPR = FP / Q, \tag{17}$$

where TP – detected objects of interest, FP – the number of false alarms of the detector, QP – the total number of objects of interest, Q – the total number of detection.

The results are shown in Table II.

TABLE II. TEST DATA TO ASSESS THE PROBABILITY OF CORRECT DETECTION AT A MAXIMUM RANGE OF 1500 M

Video	Object	TP	FP	QP	Q	TPR	FPR
Daytime, object against a cloudy sky	1	1149	20	1151	1169	0.99	0.02
Twilight, object against cloudy sky and forest background	1	2506	278	2784	2784	0.90	0.10
Twilight, object against vegetation	1	46	2	74	48	0.62	0.04
Twilight, 2 objects on dynamic background (wind, leaves)	1	49	36	56	85	0.88	0.42
	2	76	9	88	85	0.86	0.11
Daytime, object on low-detail background, haze	1	339	6	699	345	0.48	0.02
Daytime, 2 objects on highly detailed background, haze	1	1601	114	1722	1715	0.93	0.07
	2	1388	79	1722	1567	0.81	0.05
Daytime, the object crosses the horizon line (sky and forest)	1	234	21	298	261	0.78	0.08

During the tracking stage, the main metric considered was the number of tracking failures. In this research, a tracking failure is defined as the loss of the object, requiring subsequent automatic recapture using the neural network-based detector. The newest algorithm based on deep learning technology, SAM [12], was used as a competing solution for the comparative study. This neural network algorithm is currently one of the leaders in terms of accuracy and is popular among researchers.

The results of the experimental study conducted on a set of 10 videos are summarized in Table III.

As can be seen from the description of the video clips, they contain situations that occur in real conditions and complicate the tracking process.

Examples of frames with a significant increase in the size of the object of interest and frames with a sharp change of direction due to the maneuver of the carrier are shown in Fig. 8 and 9.

TABLE III. RESULTS OF THE EXPERIMENTAL STUDY

Video, No	Description	Number of frames	Number of tracking failures (proposed algorithm)	Number of tracking failures (SAM)
1	Object on a plain background (sky)	2987	0	0
2	Object on a simple background (sky), there is a brief flight of the object over the horizon line to a complex background (Fig. 7)	9108	0	0
3	Object on a complex background (forest, house)	1906	1	1
4	Object on a complex background (forest, field, road) (Fig. 7)	633	0	0
5	Object on a simple background (sky), significant change in the size of the object (Fig. 8)	1160	0	0
6	Object on a complex background (forest, field), there are moments when the object merges with the background (Fig. 7)	1237	0	1
7	Object on a complex background (sky, forest, road), there are moments with sharp global movement, change in the size of the object	1110	4	3
8	Object on a complex background (forest, sky), there are moments when the object merges with the background	926	2	1
9	Object on a complex background (forest, field, road), there is global movement, significant change of object size, rotations of the object of interest and the camera carrier (Fig. 9)	702	2	2
10	Object on a simple background (sky), low SNR value, interference, appearance of “shadows” of objects as a result of channel mismatch, global motion, significant change of object dimensions	675	4	2

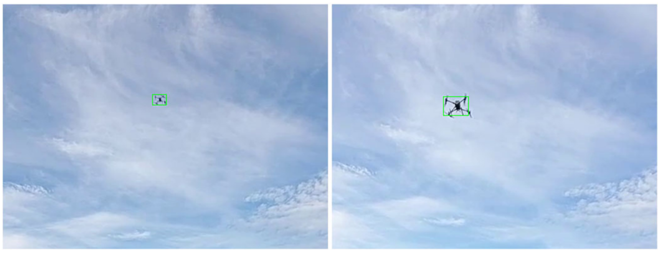

Fig. 8. Example of significant increase of object size. The object is on a simple background (sky) and has increased in size from 15*15 pixels to 40*30 pixels in 63 frames

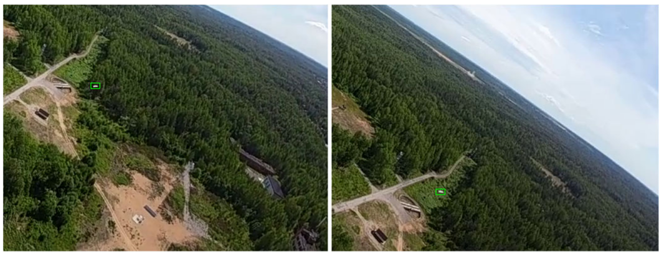

Fig. 9. Example of global motion due to the maneuver of the sensor carrier. The object is on a complex background (forest, field, road) and for 10 frames the sensor carrier has rotated by 12 degrees

The data obtained in Table III allow us to calculate the failure rate (tracking failures) [13]:

$$F = F_t / N, \tag{18}$$

where F_t is the number of tracking failures, N is the total number of frames of the video sequence.

Accordingly, according to the experimental data, the proposed method has a failure rate of $6*10^{-4}$. At the same time, the similar parameter of the competing solution SAM was $5*10^{-4}$. As can be seen from the obtained values and Table III, both algorithms show similar results. SAM shows relatively better quality of maintenance (fewer failures), but from the practical point of view, the characteristics of the two algorithms (achieved values of failure rate) are comparable.

V. Conclusion

The proposed method of automatic detection and tracking allows to detect objects of interest at the distance of 1500 meters with a minimum object size 5x5 pixels. The averaged value over all video files of the correct detection probability TPR equals 0.81, the false alarm probability FPR corresponds to the value 0.10

The experimental study showed that the tracking algorithm shows good results on uniform and non-uniform backgrounds. Accompaniment failures mainly occur in cases of the object of interest significant rotations with its size significant changing, at the boundaries of transition from one background to another. Failures can take place in the presence of several complex situations at the same time, which are not critical separately, for example, a sharp global movement and interference. The proposed tracking algorithm has comparable accuracy characteristics from the practical point of view with the advanced modern analog based on deep learning technologies. At the same time, the proposed algorithm is much easier to deploy on the target hardware: it does not require graphics gas pedals, is characterized by high performance, simplicity and flexibility of customization for the specifics and characteristics of the system.

The achieved parameters are due to the use of multi-feature detection with subsequent processing, allowing to detect the sensor carrier maneuvers and other complex situations. The advantages of the proposed method also include low computational requirements, as the neural network detector is only present during the automatic detection stage.

Acknowledgment

The work was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation within the framework of realization of the complex project on creation of high-tech production on the theme "Multimodal complex of airport airspace control" (Agreement on granting a subsidy from the federal budget for the development of cooperation between a state scientific institution and an organization belonging to real sector of the economy for the purpose of realization of the complex project on creation of high-tech manufacturing No. 075-11-2023-007 dated 10.02.2023) and within the framework of the Resolution of the Government of the Russian Federation No. 218 dated April 9, 2010 on the basis of the head executor: federal state autonomous educational institution of higher education the Saint Petersburg Electrotechnical University “LETI” (SPb ETU “LETI”).

References

[1] D. Koller, K. Daniilidis, H.-H Nagel, “Model-based object tracking in monocular image sequences of road traffic scenes”, *International Journal of Computer Vision*, v. 10, №3, 1993. P. 257-281.

[2] H. Mobahi, S. Rao, A. Yang, S. Sastry, Y. Ma, “Segmentation of Natural Images by Texture and Boundary Compression”, *International Journal of Computer Vision*, 95, 2011, pp. 86–98.

[3] S. Rao, H. Mobahi, et al, “Natural Image Segmentation with Adaptive Texture and Boundary Encoding”, *Proceedings of the Asian Conference on Computer Vision (ACCV)*, H. Zha, R.-i. Taniguchi, and S. Maybank (Eds.), Part I, LNCS 5994, Springer, 2009, pp. 135-146.

[4] M. Gao, H. Chen, S. Zheng, “Texture image segmentation using fused features and active contour”, *23rd International Conference on Pattern Recognition*, April, Cancun, Mexico, 2016, pp. 520-526.

[5] G. L. Foresti, C. S. Regazzoni, “Coding of noisy binary images by using statistical morphological skeleton”, *IEEE Workshop Nonlinear Signal Processing*, Cyprus, Greece, 1995, pp. 354-359.

[6] G. L. Foresti, “A change detection method for multiple object localization in real scenes”, *IEEE Conf*, Indust, Electron, Bologna, Italy, 1994, pp.984-987.

[7] M.R.M Jenkin, A.D. Jepson, J. K. Tsotsos, “Techniques for disparity measurement”, *CVGIP*, 53(1), 1991, pp 14-30.

[8] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection”, *Computer Vision and Pattern Recognition, 2005, CVPR 2005, IEEE Computer Society Conference on*, volume 1, IEEE, 2005, pages 886–893.

[9] A. Lukezic, T. Voj'ir, L. C. Zajc, J. Matas, M. Kristan, “Discriminative correlation filter tracker with channel and spatial reliability”, *International Journal of Computer Vision*, 2018.

[10] J. van de Weijer, C. Schmid, J. Verbeek, D. Larlus, “Learning color names for real-world applications”, *IEEE Trans. Image Proc.*, July 2009, 18(7):1512–1523.

[11] J. F. Henriques, R. Caseirio, P. Martins, J. Batista, “High-Speed Tracking with Kernelized Correlation Filters”, *IEEE Trans on PAMI*, 2015, 37(3):583-596.

[12] A. Kirillov, et al, “Segment Anything”, *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 3992-4003.

[13] A.E. Shchelkunov, V.V. Kovalev, K.I. Morev, I.V. Sidko, “The metrics for tracking algorithms evaluation”, *Izvestiya YuFU, Engineering Sciences*, 2020, №1, pp. 233-245.