UDC 004.05 DOI: 10.18698/0236-3933-2019-2-40-55

# QUALITY ASSESSMENT OF OBJECTS DETECTION AND LOCALIZATION IN A VIDEO STREAM

**D.A. Gavrilov** gavrilov.da@mipt.ru

**Moscow Institute of Physics and Technology (State University),
Dolgoprudny, Moscow Region, Russian Federation**

**Abstract**

The paper presents a software and hardware system for quality assessment of detection and localization of objects of interest (in most cases, the functional systems of this type coincides with the functionality of television machines), depending on the content of the input video. The development of a generator for recording the input video sequence along the trajectories of objects, the trajectory of the camera and image distortion in the frame is performed. The description of the main modules of the complex is presented. The description of the simulation process in a distributed testing system is given. Client applications can be run on the same computer as the server, and on remote computers

**Keywords**

*Detection and localization, image processing, automatic target tracking*

Received 07.09.2018

**Introduction.** Target acquisition (TA) is the mode of operation of systems for analyzing the environment in which a change in the orientation of the axis or direction of movement of certain elements of the complex after changes in the trajectories and speeds of the observed objects is provided without the operator's participation, but only under his control. Detection and "lock-in n" of a target precede the task of target acquisition [1]. In the TA process, current coordinates and advance angle are calculated.

Natural or artificial noises may cause a necessity to change from the TA mode to manual tracking, which leads to a decrease in tracking stability.

The TA mode can be based on radar, photooptical, infrared, quantum-optical, acoustic, magnetic and other principles of radiation and reception of signals processed by a computing device according to a given program [2–4].

Testing and assessment the performance of the detection and localization algorithms during the development process go a long way. A comparison of the error probabilities of the algorithm is a cross functional way to test it. Assessments of the error probabilities can be obtained empirically or by testing on some scope of examples [5]. There are known methods for studying the stability of algorithms for detecting and localizing objects to various distortions, based on the calculation of the error probability estimation using the Monte-Carlo method [6]. Using the superposition of noise on the image, the changes in the quality of the algorithm depending on the level of noise can be calculated. Besides, noise signal [7] or a multi-purpose scene [8] simulators to assess the quality of algorithms can be used. The disadvantages of the known simulators are the imitation of work only of the radar equipment and the inability to test and control systems operating in the visible and/or infrared wavelength ranges, and / or based on quantum-optical, acoustic, magnetic and other principles of radiation, as well as the impossibility of simulating an artificial video signal, allowing to reproduce the generated file in the system of automatic and / or semi-automatic detection of the location of the target, tracking the target [9].

Software allowing to automate the process of checking the operation of the systems for detecting and localizing objects of interest is presented in this paper. The software mentioned provides a certain features, which in most cases has often a television automat or teleautomat supplied as part of environmental analysis systems. Software features:

– generating a video file with specified parameters of camera and targets movement, size and types of interferences;

– simulation of an artificial video signal, which allows to play the generated video file in the teleautomat;

– teleautomatics logging mode;

– automated analysis of the teleautomats log and obtaining statistics on various criteria for the “quality” of the teleautomats operation.

The result is achieved in that it provides the ability to assess the correspondence between the parameters of the input video and the numerical coefficients for assessment the quality of object of interest tracking, as well as the possibility of modifying the video signal to create unique parameters of the original video, which makes it possible to improve the quality of assessment of target detection and tracking systems.

**Problem statement. The aim of the work:**

– simulation of input signals for teleautomat;

– visual and numerical monitoring of teleautomat operational results.

To achieve this goal, it is necessary to develop a complex for assessing the quality of teleautomats’ work, depending on the content of the input video.

To solve the problem, it is necessary to develop a generator, with which an incoming video sequence is recorded for a set of certain parameters. The trajectories of motion of objects and the camera can be as the parameters, as well as the distortion of the image of the frame.

The complex for assessing the quality of teleautomats operation consists of the following independent modules.

*Simulator* — generates incoming video for a teleautomat, and also records the initial data about camera and targets movement (in foreign literature — Ground Truth) — information about finding marks on each frame, that is, the positions of their centers and masks.

*Visualizer* — forms a video with labels' markers, which was found by the teleautomat.

*Evaluator* — provides quantitative indicators of the quality of the teleautomat operation (takes Ground Truth to the input and the result of the teleautomat operation, then checks how close the original values to the teleautomat outputs are).

Teleautomats testing system architecture is presented in Fig. 1.

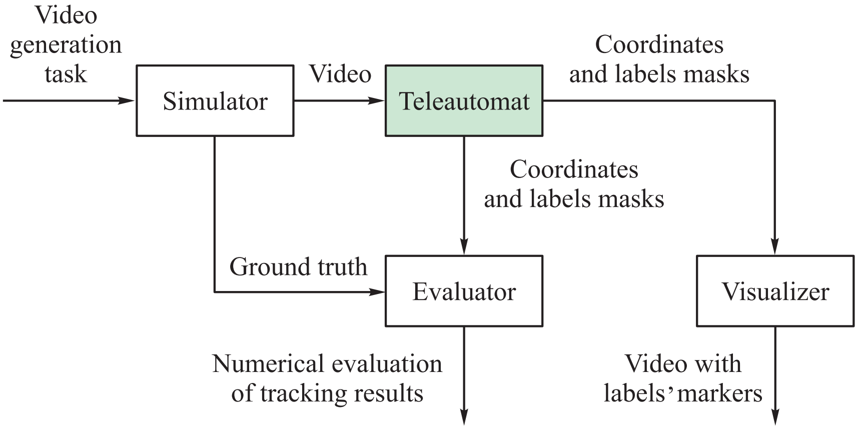
Video generation task
Simulator
Video
Teleautomat
Coordinates and labels masks
Coordinates and labels masks
Ground truth
Evaluator
Visualizer
Numerical evaluation of tracking results
Video with labels' markers

**Fig. 1.** Teleautomats testing system architecture

The result of the system operation is the correspondence between the parameters of the input video and the numerical coefficients of assessing the quality of tracking the object of interest (tracking).

Thus, the formation of a set of objective parameters that affect the quality of tracking is one of the most important tasks of the study. These developed and formulated parameters are calculated at the stage of formation of the original video.

**Simulator description.** The input video simulator (Fig. 2) consists of three separate modules: a trajectory interpolator; dynamic parameter interpolator; video generator based on a 2.5-dimensional scene model.

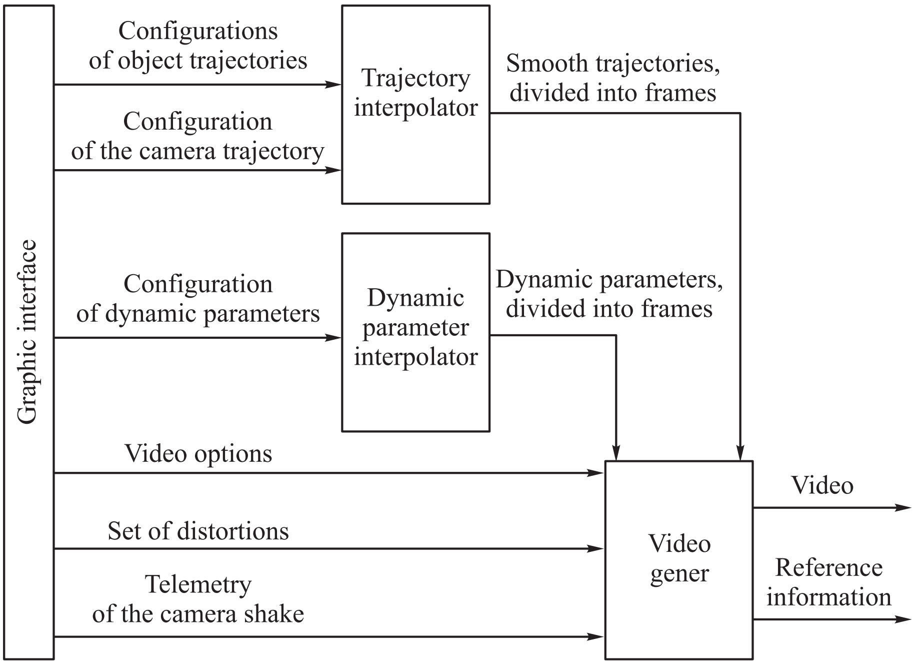

**Fig. 2.** High-level simulator diagram

The trajectories interpolator transforms the input set of key points of the movement trajectories, as well as the target and camera speeds into smooth trajectories divided into frames with a given video frequency.

Dynamic parameters interpolator transforms a set of key points of a time diagram of dynamic parameters into parameter sets for each frame. Interpolation between key points is linear.

The video generator from discrete trajectories set produces the formation of an image for each frame and recording of complete information about the scene, which is later used for tracking quality assessment.

A 2.5-dimensional model of a scene will be called a scene composed of two-dimensional layers ordered by distance from the observer. Most of the elements of the scene are sprite oriented, i.e., they are not three-dimensional models, but animated two-dimensional images that undergo projective transformations. In general, the elements of the scene can be replaced depending on the position (angle) of the object relative to the observer. Consider the implementation of a 2.5-dimensional model of the scene.

The model can be described by the following characteristics:

– the camera is fixed at one point, the trajectory of the camera is a set of rotations of the direction of its optical axis;

– the background image is “attached” to an infinitely distant sphere;

– the images of the targets do not change the angle; the sprites of the targets can only be rotated around the camera optical axis, move in and out, while removing the target from the observer can be replaced by the sprite scale. We will call this degree of freedom a 2.5-dimensional space.

The video generator is a module implemented in the MATLAB environment that generates images of the frame for each point in time.

**Generator main features.** 1. The definition of the background area getting in frame. To implement this function, the image is scaled according to the distance of the object from the camera and the focal distance of the lens, then the translation vector and rotation are applied to it.

2. For each target trajectory, the following actions are performed:

– the image of the sprite target is scaled and rotated in accordance with the specified values of scale and angle;

– the image of the sprite is superimposed on the background image of the current frame at the given coordinates.

In the future, the implementation of a 2.5-dimensional model of the scene will be called a 2.5D-model.

**Trajectories calculator.** A schematic representation of the trajectory calculator is shown in Fig. 3

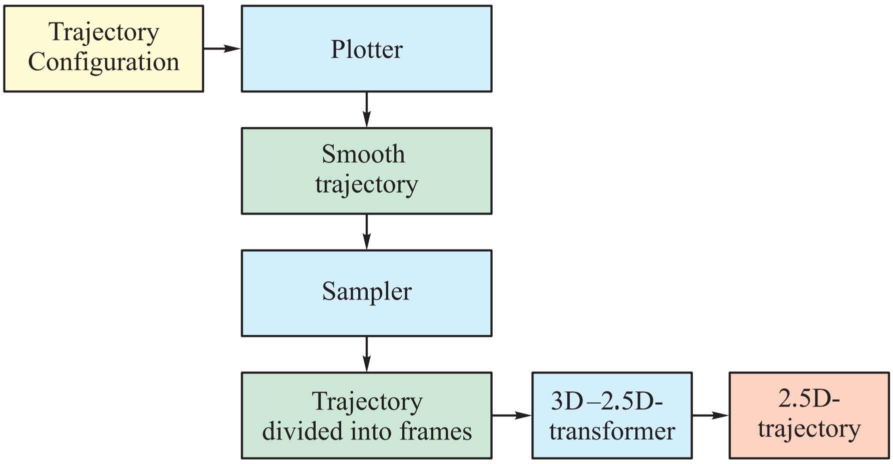

**Fig. 3.** Trajectories calculator

The input data for the simulator are prepared by a specially developed utility that solves the problem of creating a smooth target motion path.

A spline-approximated curve, as well as the properties of the sprite target are constructed by a set of key points at the input. The key points of the target trajectory are given in the earth axes system. For definiteness, it is assumed that the origin of coordinates coincides with the position of the camera, the axis *Ox* is directed to the north, the axis *Oy* — to the west, the axis *Oz* — along the outer normal to the earth's surface.

To calculate the image seen by the camera, the orientation of the camera is required. To set the camera direction, the azimuth and zenith angles are used. Thus, spherical coordinates are used in intermediate calculations.

In the spherical coordinate system (SCS), three coordinates ($r$, θ, φ) are used to specify the position of a point in three dimensions, where $r$ is the distance to the origin and θ and φ are the zenith and azimuthal angles of the camera direction (given in radians). In Fig. 4 shows an example of constructing a smooth trajectory along anchor points.

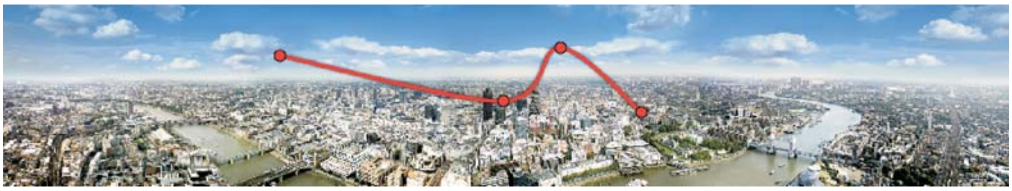

**Fig. 4.** A smooth trajectory along anchor points constructing

The configuration also indicates the type of trajectory — the camera or target, as well as the video frame rate. Cartesian coordinates and speeds are indicated in meters and meters per second.

Interpolation is implemented for the following dynamic parameters: brightness and scale — for the object; background brightness and noise dispersion — for the scene.

**Video generator interface.** The main task of the video generator is to create video sequence based on motion trajectories of the objects and the camera. The objects are represented as sprites and they move on the panoramic image's background. In this case the camera motion trajectory describes the relationship between the frame's system of coordinates and the real reference system.

Besides, the video generator adds distortions in the form of noise in each frame. The parameters of the noise are specified in the incoming task.

A panorama is an image of infinitely remote objects in the equidistant cylindrical projection.

The transformation of coordinates in the panorama raster is given by:

$$\begin{aligned} x &= (\varphi - \varphi_0)\sin\theta, \\ y &= \left(\frac{\pi}{2} - \theta\right). \end{aligned} \qquad (1)$$

A "binding" is used in order to relate the panorama's points with the camera's direction in the direction of the polar and azimuthal angles. The "binding" is based on the image stitching program during panorama's creation.

Rotation matrices around the axis of the cartesian system through angle α in the 3-dimensional space are the following.

Rotation around *X* axis (from the resulting coordinate system to the original one)

$$\mathrm{M}_x(\alpha)=\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\alpha & -\sin\alpha\\ 0 & \sin\alpha & \cos\alpha\end{pmatrix}. \tag{2}$$

Rotation around *Y* axis (from the resulting coordinate system to the original one)

$$M_y(\alpha)=\begin{pmatrix}\cos\alpha & 0 & \sin\alpha\\ 0 & 1 & 0\\ \sin\alpha & 0 & \cos\alpha\end{pmatrix}. \tag{3}$$

Rotation around *Z* axis (from the resulting coordinate system to the original one)

$$M_z(\alpha)=\begin{pmatrix}\cos\alpha & -\sin\alpha & 0\\ \sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{pmatrix}, \tag{4}$$

$$\begin{gathered}P=\begin{pmatrix}0 & 0 & 1\\ 1 & 0 & 0\\ 0 & 1 & 0\end{pmatrix},\\ R=M_z(\omega)M_x\left(\frac{\pi}{2}-\theta\right)M_y(\varphi)P,\\ K=diag\left(f,f,1\right),\\ T=K^{-1}R.\end{gathered} \tag{5}$$

In order to form a new frame the generator the coordinate from the earth-bound system of coordinates (in meters) to the system of coordinates of the image plane (pixels).

**Camera vibrations.** Vibrations loaded from telemetry files are added to the camera motion trajectory. The telemetry file contains data about car's motion referenced to earth, as well as about the camera's vibrations references to the car. Thus, the total vibration of camera referenced to earth is calculated and added to the trajectory.

**Output data.** The simulator output data contains scene images compiled into a video sequence. The images are broken down into frames.

The monochrome image of the current video frame will be denoted by an $N \times M$ matrix $A$.

Besides, the simulator output also contains: the set of parameters; Ground Truth.

The process of forming the video sequence and the ground truth is presented in Fig. 5.

**Noises and effects.** Each video sequence is modified using various effects or noises that indirectly influence the tracking quality. Noise level variation allows creating unique combinations of the calculated parameters of the original video and the results of the tracking quality evaluation.

Right now we have developed the following effects.

*Fog (reduced image contrast).* This effect creates a half-tone image by opacification of the original half-tone image. The brightness values in the range [0, ..., 1] are transformed into brightness values in the range [*bottom* ... *top*]. The brightness values less than *bottom* are set equal to *bottom*, the brightness values greater than *top* are set equal to *top*.

*Exposure error.* This effects adds exposure correction to the intensity value of each point of the original image. The correction can range in [–1, ..., 1] (1 and –1 correspond an absolutely black and absolutely white image).

*Non-uniform (e.g. gaussian) light exposure.* This effect simulates a point source of light located in the upper left corner of the image. The relationship between light intensity and distance is given by

$$\begin{aligned} &I = Ae^{-\frac{x^2}{2\sigma^2}}; \\ &\sigma = Max\left(width, height\right) * scale, \end{aligned} \tag{6}$$

where *scale* is exposure scale coefficient, in range [0, ..., 1] (the value of 1 corresponds to full exposure of the frame); $A$ is exposure amplitude ranging in [0, ..., 1]; *width, height* is frame image dimensions.

*Focusing error.* The focusing error is simulated by blurring object boundaries. This effect is achieved by applying 2-dimensional circular averaging filter. The parameter *diam* (in pixels) sets the diameter of the averaging disc.

*Motion blur*. This effect is a blurring of the image caused by camera motion. It is simulated by applying averaging two-dimensional filter parametrized by length (amplitude) and the inclination angle. The amplitude is measured in pixels and the inclination angle in degrees in a clockwise direction.

Some noise types can lead to improvement of the values of the randomly chosen parameters, e.g. create an additional texture. It is necessary to evaluate

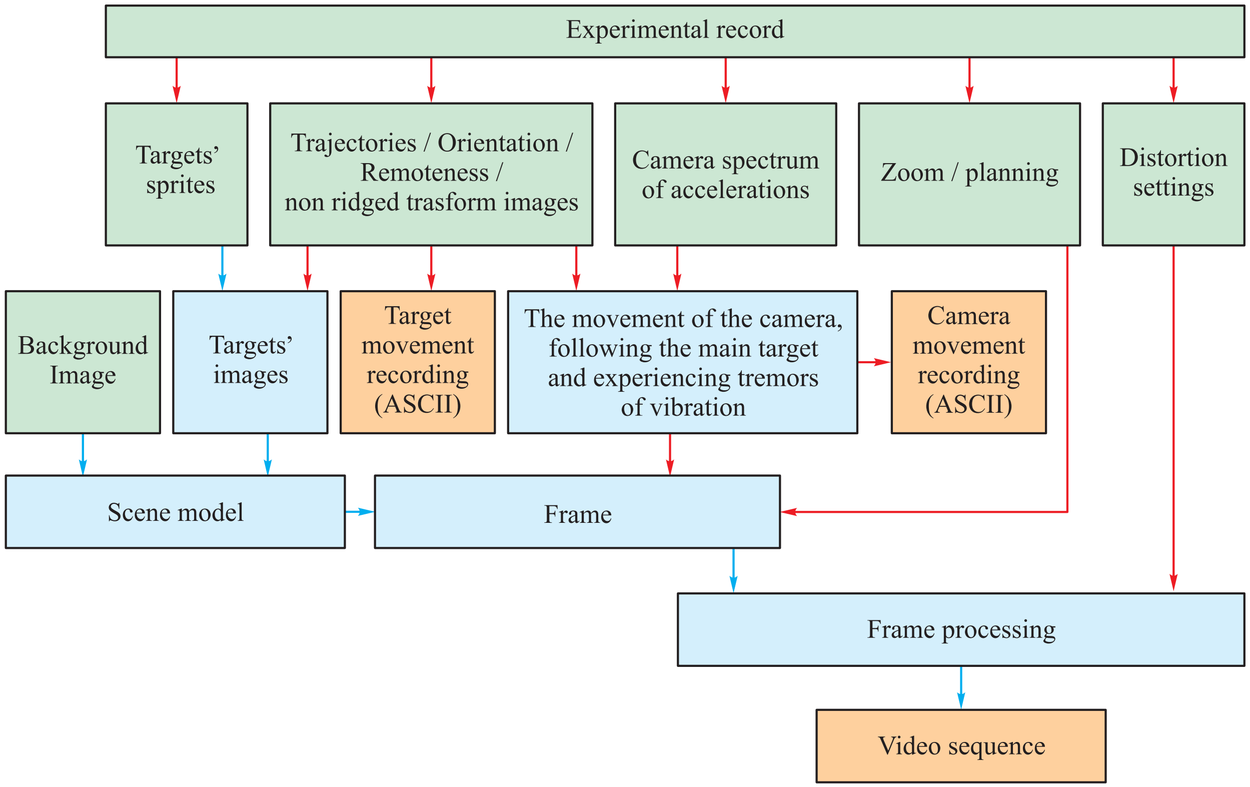
Experimental record
Targets' sprites
Trajectories / Orientation / Remoteness / non ridged trasform images
Camera spectrum of accelerations
Zoom / planning
Distortion settings
Background Image
Targets' images
Target movement recording (ASCII)
The movement of the camera, following the main target and experiencing tremors of vibration
Camera movement recording (ASCII)
Scene model
Frame
Frame processing
Video sequence

**Fig. 5.** Video sequence and ground truth generation

each parameter before and after applying distortions in order to highlight singularities created by distortions. The parameter data difference norm is treated as component of the distortion value. Thus the "artefact" keypoints are taken into account as a noise component.

**Tracking quality assessment.** The tracking results are represented by foreground mask. Going further this term is going to denote areas of images occupied by markers. In this case the task of the teleautomat is forming foreground mask segmented by markers for each frame of the input video.

The main components used to assess the algorithm output are: segmentation of objects; continuity of the segmentation labels; deviation of the center's position; object detection probability.

The general flowchart of the teleautomat operation quality assessment system is presented Fig. 6.

***Segmentation.*** The segmentation results in assigning labels of belonging to this or another marker. The number of a marker is the label of the majority of the visible pixels for the given marker.

The following assessment characteristics are calculated: pixel false positive count; pixel false negative count.

Let $\mathbf{F}_S$ by an $N \times M$ matrix, containing the original foreground binary mask obtained from simulation, $\mathbf{F}_t$ is $N \times M$ matrix containing binary foreground binary mask from tracking results; $A_S$ is number of points $\mathbf{F}_S$ equal to 1; $A_t$ is number of points $F_t$ equal to 1.

Then the coefficients of segmentation assessment can be numerically described in the following way:
coefficient of truly detected pixels of markers

$$TP = \frac{\mathbf{F}_S \cap \mathbf{F}_t}{A_S};$$

coefficient of false positive detected pixels of marks

$$FP = \frac{\mathbf{F}_t - \mathbf{F}_S \cap \mathbf{F}_t}{A_S};$$

coefficient of false negative detected pixels of marks

$$FN = \frac{-\mathbf{F}_t \cap \mathbf{F}_S}{A_S}.$$

***Reacquisition.*** Reacquisition is the situation of marker number change. The metric is going to be the average number of reacquisitions per one frame.

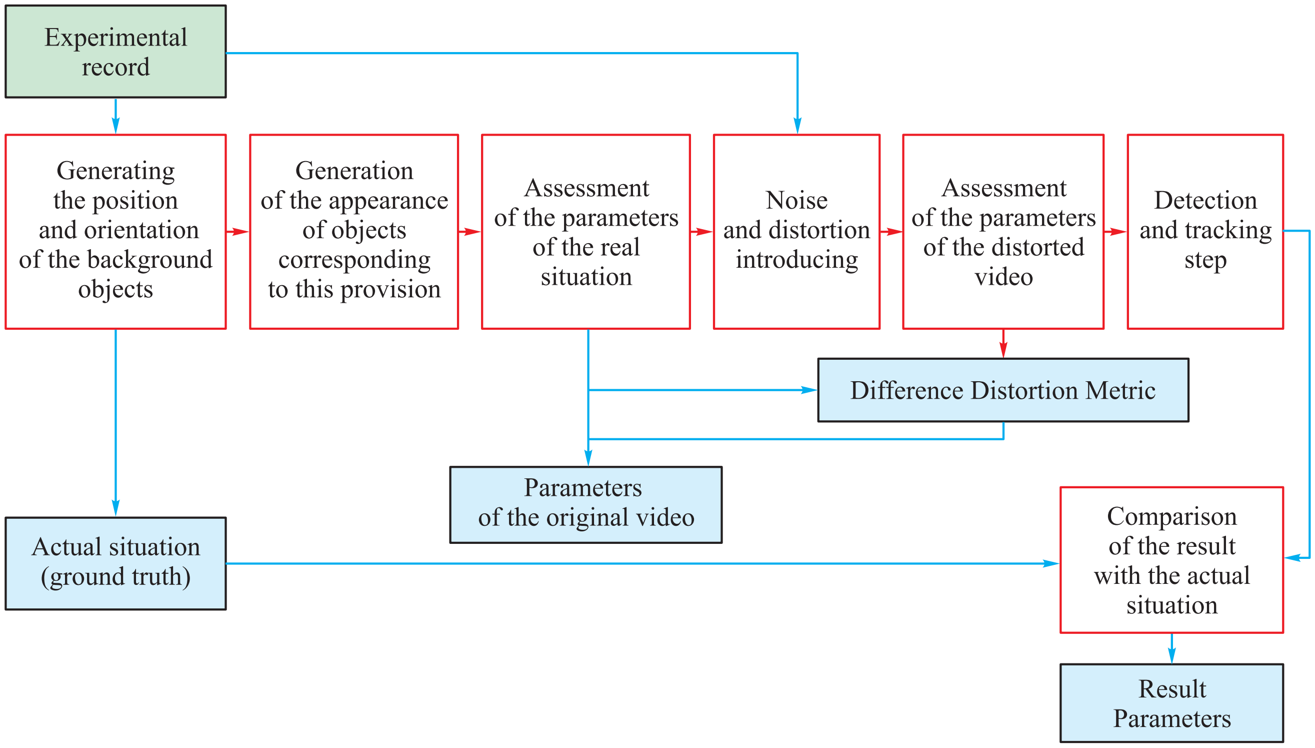
Experimental record
Generating the position and orientation of the background objects
Generation of the appearance of objects corresponding to this provision
Assessment of the parameters of the real situation
Noise and distortion introducing
Assessment of the parameters of the distorted video
Detection and tracking step
Difference Distortion Metric
Parameters of the original video
Actual situation (ground truth)
Comparison of the result with the actual situation
Result Parameters

**Fig. 6.** Flowchart of the teleautomat operation quality assessment system

***Deviation of the center position.*** The distance between the true center of a marker and the center calculated by the tracker is calculated for each marker. Then an averaged characteristic of the number of markers and frames is calculated. The deviation of the center point is shown in Fig. 7.

***Object detection probability.*** The relationship between the change $q$ of detecting an object and the size of the neighborhood ε is calculated. A successful detection is an event where for a true marker there is at least one detected object the center of which lies within the radius of the given neighborhood (Fig. 8).

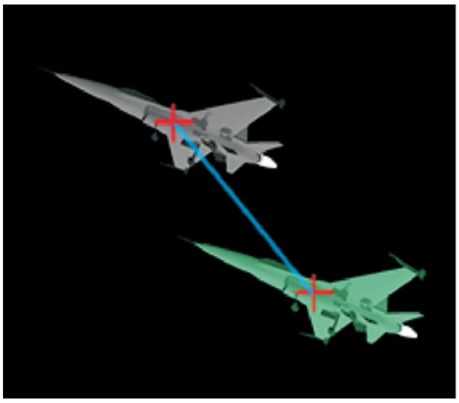

**Fig. 7.** Deviation of the center position

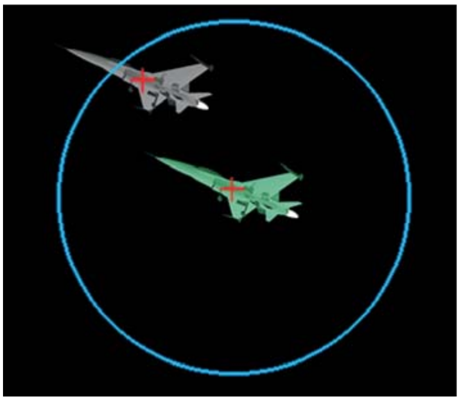

**Fig. 8.** Neighborhood of the object detection probability

Thus, the probability of detecting an object in the neighborhood ε is given by

$$p_\varepsilon = \frac{\sum\limits_n^N \sum\limits_m^M \frac{q_{n,m,\varepsilon}}{M}}{N}, \tag{7}$$

where $N$ is number of frames in the video; $M$ is number of true frames; $q_{n,m,\varepsilon}$ is probability of detecting $m$-th marker on the $n$-th frame in the neighborhood ε.

**Matching video parameters and the result of quality assessment.** Quality is a function of *groundTruth* and frame image $A$

$$q = F\,(groundTruth, A). \tag{8}$$

Expression (8) is true for parameters of the original video

$$p = F\,(groundTruth, A). \tag{9}$$

Thus each set of parameters is corresponding to the result of the quality assessment, i.e., quality is a function of the video parameters.

The quality evaluator can work in the following modes in order to support teleautomats depending on teleautomat’s functionality:

- enables transfer of the objects' masks;
- sends only the center and the bounding box of a marker;
- sends only the center of a marker.

All assessment criteria apply for the first mode. The quality assessment for the second mode is similar to the first mode, except that the object's mask is represented by a rectangular area corresponding to the bounding box of a marker.

Segmentation criteria are not calculated in the third mode as they cannot be applied because of shortage of input data.

**Distributed complex of teleautomats' testing** is created with the aim of reducing the time of generating video sequence and calculating parameters. The time is reduced by executing processes for each dataset in parallel. The client applications can be launched on one computer with a server as well as on remote computers. The sequence diagram of the simulation process is shown in Fig. 9.

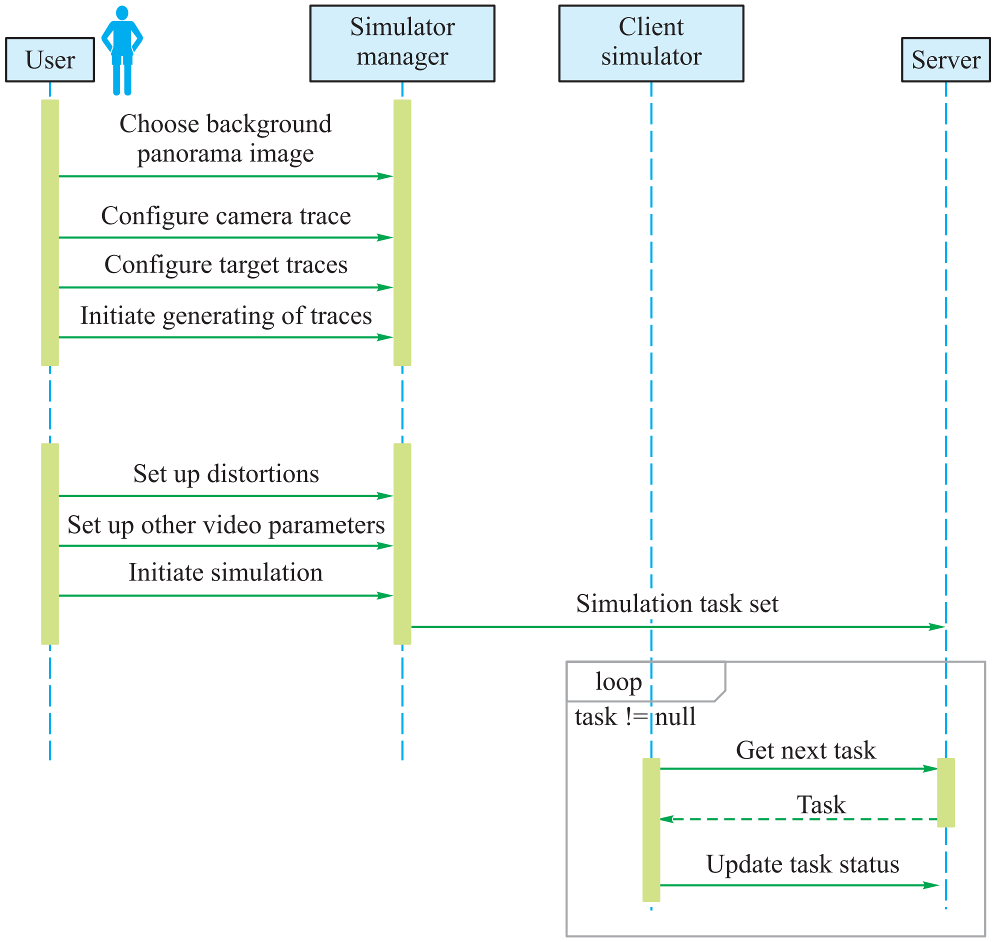

**Fig. 9.** Sequence diagram of the simulation process

The Simulator Manager is used to enable users to configure trajectories of the camera and the targets.

After constructing the trajectories and the targets the user sets up the distortion parameters, video properties and launches the simulation. A server is being sent a query containing the simulation-task-set.

After the server receives the simulation-task-set it sends the tasks to the client applications. A client is sending a query for receiving the next task, in return it gets the simulation task. Having achieved the task, the client sends a notification on completing the task.

After receiving a new task the client generates a new video sequence calculates the parameters and the input data and writes them to a database. The recorded data gets a unique video identifier in the database — Video ID.

After completing all the task the server launches the tracking and the tracking quality assessment procedures for each video, a teleautomat is writing the tracking results in the database, the evaluator records the assessment results.

**Experimental studies of the detection and localization algorithms.** Video signal primary processing algorithms were tested using the developed hardware and software suite.

Unique fragment extraction algorithm (algorithm No. 1).

Algorithm of follow-up binding of the object's fragments moving together (algorithm No. 2).

Pyramid algorithm of extracting and calculating features of the objects with intensity different from the average local background intensity (algorithm No. 3).

Algorithm of calculating features of the objects differing from the local background the by the gray scale (algorithm No. 4).

The main metrics for assessing the algorithms are target detection probabilities and the false positive probability. Over the course of the experiment the algorithms were tested using video signal generator on synthetic video and real videos of the target environment. Throughput capacity of the video signal was studied as well. The results of each of the algorithms were visualized and quantitatively assessed. The testing results in the form of average indicative data are shown in the table.

**Results of testing the algorithms**

| Primary processing method | Average detection probability | Average false positive probability |
|---|---|---|
| Algorithm No. 1 | 0.72 | 0.31 |
| Algorithm No. 2 | 0.79 | 0.28 |
| Algorithm No. 3 | 0.71 | 0.35 |
| Algorithm No. 4 | 0.68 | 0.38 |

The quality of the algorithm's work is visually controlled by an operator using video file generated by the "Visualizer" module. The quantitative characteristic of the algorithm's work quality, statistical data is calculated using "Evaluator" module and then recorded (table).

**Conclusion.** The following results were obtained. 1. The software that provides verification of the operation of teleautomats with different parameters of the external environment was developed.

2. The generator of the registration of the incoming video sequence, the parameters of which are the trajectories of objects and camera movement and camera image distortion was developed.

3. The main modules of the teleautomats quality assessment complex are described.

4. A description of the effects and noise used to modify the video and allowing to create unique combinations of parameters of the original video to improve the assessment of the quality of tracking are presented.

5. The basic principles and criteria for assessing the quality of the teleautomats operation are presented.

6. The simulation process is described.

7. Experimental studies of algorithms for the primary processing of a video signal have been performed, as a result of which visualization and quantitative evaluation of the results of each of the tested algorithms has been carried out.

Translated by U. Gordeeva

## REFERENCES

[1] Boykov V.A., Kolyuchkin V.Ya. Algorithm for object image automatic tracking. *Vestn. Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Priborostr*. [Herald of the Bauman Moscow State Tech. Univ., Instrum. Eng.], 2017, no. 5, pp. 4–13 (in Russ.). DOI: 10.18698/0236-3933-2017-5-4-13

[2] Voroshilina E.P., Tislenko V.I. Analysis of autotracking methods by range. *Izvestiya Tomskogo politekhnicheskogo universiteta. Inzhiniring georesursov* [Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering], 2006, no. 8, pp. 67–71 (in Russ.).

[3] Korzunov O.V., Luzhinskiy A.I. Analysis of detection and coordinates' measurement algorithms in optoelectronic systems. *Izvestiya TulGU. Tekhnicheskie nauki* [News of the Tula State University. Technical sciences], 2017, no. 12-3, pp. 164–172 (in Russ.).

[4] Mazo A.M., Markova E.I., Lapteva R.R. Guidance system of tracking radar. *Sovremennye problemy proektirovaniya, proizvodstva i ekspluatatsii radiotekhnicheskikh sistem*, 2016, no. 1, pp. 105–107 (in Russ.).

[5] Courtney P., Thacker N.A. Performance characterisation in computer vision: the role of statistics in testing and design. *Imaging and vision systems*. Nova Science Publishers, 2001, pp. 109–128.

[6] Perevalov D.S. Analysis of an object detection and localization algorithms in images with structural noise. *Vychislitel'nye tekhnologii* [Computational Technologies], 2009, vol. 14, no. 1, pp. 94–106 (in Russ.).

[7] Bekirbaev T.O., Bondarenko I.O., Starikova T.A., et al. Tsifrovoy imitator bortovykh radiolokatsionnykh system [Digital simulator of onboard radar system]. Patent 75058 RF. Appl. 24.03.2008, publ. 20.07.2008 (in Russ.).

[8] Sirotin A.I., Mukhin V.V., Nesterov Yu.G., et al. Imitator radiolokatsionnoy tseli [Radar target simulator]. Patent 2412449 RF. Appl. 26.12.2008, publ. 20.02.2011 (in Russ.).

[9] Bakhrushin N.I., Dogadkin V.A., Meshcheryakov V.V., et al. Gruppovoy imitator-trenazher dlya sistemy upravleniya ob"ektami samonavedeniya [Group simulator-simulator of homing object control system]. Patent 39964 RF. Appl. 26.04.2004, publ. 20.08.2004 (in Russ.).

**Gavrilov D.A.** — Cand. Sc. (Eng.), Head of the Laboratory of Digital Special Purpose Systems, Assoc. Professor, Department of Radio Electronics and Applied Informatics, Moscow Institute of Physics and Technology (State University) (Institutskiy per. 9, Dolgoprudny, Moscow Region, 141701 Russian Federation).

**Please cite this article as:**

Gavrilov D.A. Quality assessment of objects detection and localization in a video stream. *Herald of the Bauman Moscow State Technical University, Series Instrument Engineering*, 2019, no. 2, pp. 40–55. DOI: 10.18698/0236-3933-2019-2-40-55