The Principles of Organizing the Search for an Object in an Image, Tracking an Object and the Selection of Informative Features Based on the Visual Perception of a Person

Vitaliy Boyun(✉)

National Academy of Sciences of Ukraine, Kyiv, Ukraine
vboyun@gmail.com

Abstract. A significant expansion of the scope of computer vision, in particular in real-time systems, places very high demands on them in terms of productivity and efficiency of information processing, and in feedback systems, it also requires information lag in it. Such requirements are not ensured by traditional approaches. The way out of the situation may be to use as a prototype the principles of organization of the human visual system, which has a very high selectivity of perception of video information. The paper presents a generalized dynamic model of the organization of these principles. It is proposed to use them to organize the search for an object in a coarse image of a scene, to track an object and, if necessary, to carry out its classification or recognition at a more detailed level.

Keywords: Retinal neural network · Receptive fields · Image preprocessing · Object search · Informative signs · Local (ring) organization of neurons · Adaptation mechanisms · Intelligent video systems

1 Introduction

Intelligent video cameras and real-time video systems play a large role in automation systems for production processes, visual quality control of products, robotics, security and military systems, automation systems for scientific and biomedical research, etc. Moreover, the range of their application, the requirements for them are constantly expanding. This is especially true for video systems with feedback, where the results of real-time information processing are used to control the process or for other actions. Such systems put forward increased demands not only on the performance of computing facilities, but also on the lag of information in the feedback loop, which are not provided within the framework of traditional approaches [7,9]. On the other hand, the human visual system has improved over millions of years and has reached an extremely

S. Babichev et al. (Eds.): DSMP 2020, CCIS 1158, pp. 22–44, 2020.
https://doi.org/10.1007/978-3-030-61656-4_2

high level of organization. Therefore, the phenomenon of vision provides an extremely many diverse solutions for computer vision systems. Despite the enormous amount of information in the image, and especially in the video sequence, the human visual-analyzing system very effectively and efficiently copes with these problems due to its extremely high selectivity [4,8,16,18,25].

There is a significant semantic gap in how a person perceives and describes an image and how the image is perceived by the video system. A person identifies the semantics of the image, and the video system represents the visual content of the image in the form of its low-level characteristics such as color, texture, orientation, shape, the presence of movement, etc. [22].

The second section briefly discusses the principles of organization of the human visual system, the third - the state of the problem, in the fourth - a generalized model of human perception in dynamics. The fifth section is devoted to the principles of organizing the search for an object in an image, tracking it and extracting informative features based on a person's visual perception.

2 The Principles of Organization of the Human Visual System

The following is a brief description of the structural features of the organization of the retina, the mechanisms and processes associated with the processing of video information in it [8,15,18,22].

The scene image projected on the retina is inverted along two axes (top-bottom and left-right) and is a set of individual points obtained from rods and cones. These points represent, respectively, different light intensities in shades of gray and different wavelengths, perceived by the brain as three basic colors (red, blue and green with their shades). Moreover, dots in shades of gray are mainly located on the periphery of the retina with a low spatial resolution, and colored dots, more densely packed, are located in the central fossa (fovea zone). This representation of the environment is extremely redundant and not informative for the brain; it simply cannot cope with such huge flows of information. In the process of evolution, selective mechanisms have been developed in the human visual system to extract the most informative features for the perception of the environment, reducing the redundancy of scene images by many orders of magnitude. These mechanisms use the presence in the images of edges, borders of objects that differ in color, texture, shape, orientation, movement in space, which allows using different methods to distinguish between objects. The basic information in this case is the contrast, which, together with the retinal neurons (on- and off-centers) organized by the ring principle and lateral inhibition mechanisms, allows us to distinguish these primary signs of images of objects.

Although the rods and cones of the retina perceive absolute values of brightness and color, however, information is analyzed on the values of the differences between adjacent pixels in space (to detect the edges or borders of objects) or in time (between adjacent frames of a video sequence to detect movement or other

changes in the scene). Therefore, color or brightness differences are a very important characteristic for extracting many informative features. There is evidence of the existence of spatial frequency detectors in the visual system. which are tuned to a limited frequency range. Spatial frequency analysis is a simple and reliable way to describe and generalize the structural details of visual objects. Such information is used by the brain to evaluate fragments of the image and to control the further process of perception of the image by the retina depending on the goal (search for a specific object in the scene, tracking it or recognizing it). In the case of contemplation of scenes without a goal (for example, from a car window), perception is carried out reflexively without the intense involvement of the brain, we only see an enlarged picture of flickering objects.

Similarly, perception in tracking mode of a fast-moving object works. In this case, shorter eye movements, with a previously determined scale of the object, carry out tracking. Visual information, before getting from the eye to the brain, passes through many layers of specialized retinal neurons on the retina. The retina is a complex network of photoreceptors (rods and cones) and nerve cells (neurons). The rods provide a perception of the brightness of achromatic light (in shades of gray), and cones - electromagnetic waves of different lengths, which are perceived by the human visual system as three basic colors (red, blue, green with their shades). The peripheral zone of the retina, on which the rods and partly cones are predominantly located, provides a wide field of view with a low spatial resolution. The dense placement of cones in the central fossa (fovea zone) provides clear color vision.

The peripheral retina and central fossa are organized according to the ring principle (the so-called on- and off-centers). In the on-center, the excitation zone occupies the central part, and around it there is an inhibition zone. In off-centers, the zones of excitation and inhibition are interchanged. Such centers of the peripheral retina are organized using horizontal and diffuse bipolar cells and extract larger spatial features. The centers of excitation-inhibition of the central fossa are organized using horizontal and small bipolar cells and extract more subtle features of the image. The sizes of receptive fields are controlled, respectively, by interplexiform and amacrine cells using horizontal cells. Such an organization allows you to extract spots or points that differ from the background in brightness, color, orientation, texture, etc.

The well-known phrase “the eye looks and the brain sees” quite accurately describes the process of perception, that is, thinking about stimulating our sensory receptors. In this case, ascending and descending processes are involved. Ascending processes are also called processes of transmitting information data from the receptors. And the descending processes, also called the processes of conceptualizing information data, are based on previously acquired knowledge, previous experience, understanding and expectations. To increase sensitivity in conditions of insufficient illumination, the rods of the peripheral retina are combined into groups with the sum of signals from larger receptive fields, i.e. there is an exchange of spatial resolution for increasing sensitivity in brightness.

At the level of the retina, a number of adaptation processes operate. In particular, it is the brightness adaptation to the level of illumination (10–12 orders of magnitude: from the threshold of sensitivity of night vision, which is ensured by the excitation of more sensitive rods of the retina, to the threshold of blinding brightness, which is limited by the level of perception of cones of color vision). The subjective brightness, which is perceived by the human visual analyzer, is a logarithmic function of the physical brightness of the light that has entered the eye. However, the range of brightness levels, which is simultaneously perceived by the eye, is about 3 orders of magnitude.

Excitation of retinal neurons in the presence of objects (spots) in the focused image on it, which differ from the background in brightness or color values, spatial frequencies, orientation, and the presence of movement, is used to quickly search for objects in the scene image. It consists in the fact that the place in the scene that caused the excitement, with the help of saccades, is refluxed (focused) into the central fossa for a more detailed analysis [1,4,21].

In this case, when the plasticity property of neurons is used, the formation of receptive fields of various sizes, the change in their shape, and the possibility of extracting informative features from an object that display various physical properties of the object [1]. Special (interlexiform and amacrine) cells control these processes by changing the sizes of receptive fields, the sizes of the excitation and inhibition zones of on- and off-centers, as well as neuron activation thresholds. In accordance with the theory of integration of the features of the object, the perception of the object is carried out in two stages [28]:

- previous attention - the quick extraction of simple visible features of an object (perceptual primitives), is carried out in parallel throughout the retina automatically without conscious effort and concentration;
- focused attention - requires the observer's efforts and close examination of the object, is carried out sequentially with the help of the central fossa.

Information from diffuse bipolar cells of the peripheral retina and small bipolar cells of the central fossa is transmitted using G_M- and G_P-type ganglion cells, respectively, to the magno- and parvocellular regions of the lateral geniculate body (LGB). In the parvocellular region, color information from the central fossa is also added. These areas, like the retina, are organized according to the concentric ring principle, but operate with larger receptive fields, thus increasing the level of abstractness of the presentation of information. LGB, in addition, performs the function of switching the received and processed information into the corresponding layers of the primary visual cortex of the brain.

Thus, a significant part of the preliminary processing of information is carried out already at the level of the retina. The peripheral zone of the retina is oriented toward a coarse spatial perception of the scene, with a high sensitivity to perception of brightness. The central zone of the retina is focused on clear vision both in spatial resolution and in color. Using the principles of organizing the retina of the eye allows us to more effectively organize the search for an object in the scene image and preliminary preparation of video data for recognition by extracting low-level informative features of the image.

3 Problem Statement

Quite often, the retina is considered as a structure with a logarithmic-polar organization [3], in the center of which is the central fossa (fovea zone), and the peripheral retina is located around. This provides a wide peripheral inspection and the possibility of a detailed perception of information in the central fossa. Attempts have been made to repeat this foveal organization of visual perception in technical and algorithmic models. One of the directions is the creation of foveal sensors with radial and hierarchical (pyramidal) organization of the receptive field [24]. The disadvantage of radial organization is the need to implement the management of the "look", i.e. direction of the optical axis of the sensor, which requires the use of a high-speed drive and its control system. In addition, the implementation of logarithmic-polar sensors and the subsequent processing of information on them also present significant difficulties.

To speed up the process of searching for an object in the scene, the most commonly used method is the pyramidal organization of the process of taking and processing information [14]. It consists in the fact that at first the image with the maximum resolution is read, and then, by smoothing and thinning, for example, using the Gauss or Laplace pyramids, the next layer of the pyramid is formed, whose dimensions are 4 times smaller than the original. This procedure is repeated several times until the desired level of coarsening is achieved. It is easier to find an object on a coarse image, and then, by applying the reverse procedure, it is possible to restore the original resolution of the image of the object for its recognition, measurement, etc. However, this approach, which was used to reduce the amount of information when transmitting images under conditions of limited channel capacity, does not fully meet the principles of organizing the human visual system and the requirements of real-time systems.

An important characteristic for texture analysis is spatial frequencies, for the evaluation of which the Fourier analysis methods are traditionally used, however, they require large computational costs [4]. In [23], models of the organization of the retina and central fossa are presented, the principles of the organization of connections on layers of neurons of different specializations, adaptive mechanisms for the perception of lighting and contrast, and the path of signals from receptors to ganglion cells are considered. In the process of modeling, the reliability of these models was confirmed to a large extent; therefore, in some parts of our work, we focus on their use. Close approaches to the principles of processing information on the retina, in terms of extracting informative features, are used in convolution neural networks, which are a further development of the cognitron and neocognitron [19]. They serve as the first layers of recognizing artificial neural networks (ANN). However, convolution neural networks realize only a small part of the arsenal of techniques developed in the process of evolution and embedded in the human visual analyzer. In particular, are not used:

- mechanisms of "attention" and techniques for quick search for an object on the coarse peripheral retina, followed by detailed analysis in the central fossa;
- adaptive mechanisms for controlling the dynamic range of perception of brightness, contrast and sensitivity in low light conditions.

This leads to a significant increase in the database of objects and scenes, as well as the training time of the ANN to improve the quality of classification and recognition. In addition, convolution neural networks are currently not designed to connect to real cameras and to work in real time with video sequences.

Neurophysiologists, neuropsychologists, gestalt psychologists, cognitive psychologists, cybernetics consider the human visual analyzer from different angles, from their fields of knowledge, but there is no generalized representation of the process of perceiving the environment, and especially in dynamics. The situation is aggravated by the use in different publications of different terminology of the same processes and the conflicting interpretation of the results, which makes it difficult to understand the complete process of perception.

4 Dynamic Model of the Processes of Searching for an Object in a Scene, Tracking It and Extracting Informative Features

The generalized model of the human visual system is multifunctional and consists of hundreds of local models that describe a number of structural, physical, biochemical, psychophysical mechanisms and processes. The process of perception of visual information by a person is dynamic, with many parameters that change in the process of perception, with many feedbacks. This is evidenced by a huge number of articles devoted to studies of specific properties, features, mechanisms, processes etc. in a human visual system [2,17,20,25–27,29]. However, there is practically no general idea and understanding of the process of perception by the visual system in dynamics for various modes.

In this work, a generalized model of the organization and functioning of the retina, as well as approaches to the implementation of its most important and fundamental principles, which could help increase the intelligence of computer vision systems, are considered from a general perspective. In accordance with this model, enlarged fields of receptors (rods) are created on the retina using horizontal cells, which create a center of excitation using diffuse bipolar cells. With the help of other horizontal cells from receptors adjacent to the receptors of the excitation zone, larger circular inhibition zones are formed, i.e. on- and off-centers are organized. This process is carried out along the entire periphery of the retina in parallel. The results of the work of these centers through Gm-type ganglion cells in parallel and topologically (i.e. with reference to the location on the retina) are transmitted to the LGB. At the same time, the dimensions of the center and surround gradually decrease, i.e. such centers work, as it were, on different scales of perception. In each center, a search is made for the maximum response of the on-center with the specification of the size of the region.

Amacrine cells control the outputs of on- and off-centers and control the change in the size of receptive fields, helping to search for the maximum response of these filters. It should also be noted that the search on the peripheral retina is carried out in parallel on many features: shades of gray, texture, orientation, movement, etc. The flows of this enlarged information through G_M-type ganglion

cells enter the region of LGB magnocellular cells organized according to the same center-surround ring principle with local connections between neurons, but operating with larger receptive fields. Magnocellular cells work with fuzzy images; they are more sensitive to the difference in surface illumination (gray on gray), which contributes to the perception of the depth of the scene. In addition, they quickly respond to stimuli, but the reaction quickly dies away, which is well suited for detecting movement on the scene. The results of the generalization of information are transmitted from the LGB to the corresponding layers of the visual cortex. The visual cortex is organized on a linear basis, reacts to lines, stripes, rectangular segments. It contains simple cells, complex and hypercomplex, i.e. it also, according to the hierarchical principle, increases the level of abstractness of the features of an object for recognition [25].

If the results of the visual cortex using a coarse image of a fragment with additional fine information from the central fossa do not satisfy the search target, then the next image fragment detected on the retina is transferred to the central fossa to check if the searches target matches. This process is repeated for all identified fragments until a final positive or negative result is obtained. If a more detailed analysis is needed, similar to the above, a fragment is analyzed with additional color information for classification, recognition of an object, determination of its geometric dimensions, etc. Based on these coarse informative features, the visual cortex searches for their correspondence to some image of an object (model) from the ones in memory.

In the case of detecting an object when contemplating the environment or tracking a fast-moving object, this process is repeated many times. If it is necessary to search for an object in the scene for its classification, more detailed examination or recognition, a finer analysis of image fragments that have fallen on the retina is performed. It is possible that the visual cortex of the brain, controlling the response maxima throughout the retina, sets priorities for their sequential "transmission" to the central fossa (fovea zone). The process of "transmission" is carried out by saccades (rapid eye movements) by controlling the focusing system. In the central fossa, organized by the same principle of excitation-inhibition, due to the denser packing of cones, a more detailed analysis, including color, of the obtained image fragment is carried out in order to extract informative features from it. The results of the analysis in the central fossa through ganglion cells of the G_P type enter the region of parvocellular cells of the LGB, which is also organized on the principle of center-surround. These results are also parallel and topologically transmitted to the corresponding cortical layers of the visual cortex. In particular, layer V1 is responsible for frequency and orientation analysis, layers V2 and V3 process information about the shape and position in space, layer V4 is responsible for the perception of color information, layer V5 is for movement.

The presented work is not intended to accurately copy the functions of the retina and the processes occurring in it. The aim of the work is to understand the general structure and organization of the human visual system and use these principles to build specialized and problem-oriented video systems for various

purposes, taking into account the state and capabilities of modern microelectronics. And, perhaps, at the same time to put before it those problems that will make it possible to more effectively solve the problems of finding an object in an image, tracking it, classifying, recognizing objects, scenes and situations in real time.

The advantage of this approach is the use of a number of adaptive mechanisms to increase sensitivity in low light conditions, increase contrast, attention mechanisms and coarse-accurate presentation of the scene, which will significantly reduce the amount of databases and time for training ANN, link them to the processes of shooting video sequences and use in computer vision systems.

4.1 Coarsening (Assumptions) Accepted in the Model to Ensure the Possibility of Its Technical Implementation at the Current Level of Microelectronics Development

Instead of a logarithmic-polar retina representation system, a Cartesian image representation system with a uniform grid is adopted. Accordingly, the ring organization of on- and off-centers has turned into a square-ring. Parallel processes of extracting all possible (or necessary for a given perception mode) features of an object at the level of the peripheral retina and central fossa, as well as the processes of their transmission to the higher layers of the brain, have been replaced by sequential or parallel-sequential ones.

The paper does not examine in detail the principles of organization of the primary visual cortex, since they are already actively used in the ANN. The results of the study can be used as the first layers of the ANN for preliminary preparation of information for recognition on them. This is an important point, since the analysis of large scenes is computationally significantly more expensive than recognizing the objects themselves, providing a reduction in the amount of information for recognition by many orders of magnitude, although the recognition process is more complicated in the logical sense.

5 Implementation of the Principles of Human Visual Perception for Building Computer Vision Systems and Their Interaction with the Outside World

The purpose of this section is to highlight the fundamental principles of the organization of the human retina and their implementation, taking into account the capabilities of the modern level of microelectronics and computer vision systems (CVS) requirements. These fundamental principles, which, in our opinion, will help create more advanced computer vision systems, are [5,6]:

- the ability to change the resolution when working with the video sensor;
- implementation of methods for coarse-accurate search for an object in an image;
- the ability to expand the dynamic range of perception of brightness;

- the ability to implement various modes of perception of video information;
- the ability to use the differential presentation of information for image analysis;
- the ability to control in dynamics the parameters for reading information;
- organization of ascending and descending processes;
- ensuring the principle of ring organization of receptive fields;
- implementation of adaptation mechanisms to the level of lighting and contrast;
- the use of spatial frequencies, as an important principle of texture analysis;
- phased increase in the level of abstractness (generalization) of information.

The basis for the construction of CVSs can be modern CMOS video sensors, which are very complex devices and contain, in addition to the sensor matrix, about 10 specialized processors for preliminary (technological) image preparation organized in a conveyor, and several hundred registers for detailed programming of processes removal and preprocessing of the image. Thanks to the pipelined organization of the computing process, they provide the necessary rate of data output. In particular, the following operations are performed in the video sensor and conveyor:

- automatic exposure time control;
- the perception of light and converting it into analog signals of brightness or color;
- primary analog processing;
- automatic control of analog gain;
- conversion of analog signals into digital codes;
- determination of the level of "black";
- establishing a balance of "white" to adapt the sensor to different types of lighting;
- correction of the heterogeneity of the sensitivity of the elements of the sensor matrix;
- adaptation to the level of lighting;
- automatic focusing when using a lens with a servo drive;
- color correction due to imperfect characteristics of Bayer filters sprayed on each element of the sensor matrix;
- correction of defective pixels;
- the formation of all color components in each pixel (demosaicing or Color Filter Array (CFA) interpolation);
- gamma correction and saturation correction;
- conversion of presentation formats of the output RGB image to different formats (YCbCr, YUV, JPEG, MPEG, etc.).

Some of the above list of operations at the request of the user may be excluded. This organization of the video sensor is due to the task facing the video camera - providing high quality presentation of information for humans and the convenience of working with the camera. At the same time, the information captured by the camera is either stored in its memory or transmitted for

storage and use by some information receivers. Given the enormous amount of information obtained from the video sequence, and the need for its storage and transmission, much attention was paid to the implementation of various image compression methods (JPEG, MPEG, etc.) when organizing the pipeline.

In modern video cameras, the principles of organizing the human visual system are already used to a large extent. In particular, these are:

- focusing the image by moving the objective lens (changing the shape of the lens);
- control of the perception of light by controlling the diaphragm and exposure (change in pupil diameter);
- expansion of the dynamic range due to the nonlinear conversion of signals from the sensor matrix (logarithmic characteristic of the perception of light by the rods and cones of the retina);
- the use of 3 Bayer demosaic filters, etc., sprayed onto the corresponding pixels (using the property of selective sensitivity of cones in 3 ranges of wavelengths of the spectrum interpreted by the brain as three basic colors);
- panning, tracking, etc. modes used in smart video cameras (different types of eye movements when searching for an object, tracking it, etc.).

Computer vision systems are mainly real-time systems and, in addition to registering a video sequence and its preliminary processing, should perform its functions of image segmentation, search for an object in an image, and selection of informative features for its classification, recognition, tracking, and other actions at the rate of information. Traditional video cameras and their approaches to the collection and processing of information, image compression do not always meet the high requirements of real-time systems and, in addition, require the use of additional computing tools to implement the above functions.

The main requirement for real-time systems is to ensure the rate of information retrieval and processing, and for real-time systems with feedback, it is also necessary to minimize the delay in the feedback loop. This is especially important for computer vision systems working with fast processes and high-speed objects. Detailed studies of video sensors and the organization of the pipeline for preliminary processing of information have shown that they have large reserves for increasing productivity in CVSs, but these reserves are practically not used in practice. Let us consider how the video sensor pipeline can be modified and what needs to be added to it in order to implement a computer vision system with the fundamental principles of the human visual analyzer, discussed at the beginning of this section.

The possibility of changing the resolution of the video sensor will allow, on the one hand, to provide a wide field of view, and on the other hand, the possibility of a detailed analysis of the selected object. Image coarsing is an effective method of reducing the amount of information when searching for an object in an image, tracking it, as well as in the mode of simple contemplation of the situation and panning of a scene. Unfortunately, the modern technology for the production of CMOS video sensors does not fully ensure such capabilities. An exception is the possibility of thinning the image into rows and columns when

reading it, which allows 4 times to coarsen the representation of the image. In addition, it is possible to increase the resolution of a color image by 4 or more times compared to the original physical resolution due to linear or quadratic interpolation between adjacent image pixels. It is proposed to provide coarsening of the image by summing and then averaging the signals of neighboring pixels of the receptive fields, as it is the case in the peripheral retina of the human eye. Such summation can not only reduce image size, but also increase sensitivity in low light conditions. This process can be controlled (by analogy with amacrine cells) depending on the average or local brightness of the image by changing the sizes of the summed receptive fields and the gain of the on- and off-centers.

Summing and averaging the brightness values of the sensor matrix across rows and columns can be determine local levels of illumination. An analysis of these values allows us to identify local areas that differ significantly in the degree of illumination, and take measures to expand the dynamic range of perception of brightness.

The ability to change the resolution when working with a video sensor will allow, on the one hand, to provide a wide field of view, and on the other hand, the possibility of a detailed analysis of the selected object. Image coarsening is an effective method of reducing the amount of information when searching for an object in an image, tracking it, and also in the mode of simple contemplation of the situation.

The ability to change the resolution of the video sensor allows us to implement the method of coarse-accurate search for an object in the image, in which the search is performed on the coarse image for informative features of higher order, which significantly reduces the amount of processed information [6]. These features are transmitted in ascending channels to the central processor or to the recognition layers of the ANN, performing the functions of the primary visual cortex of the brain, where they are compared with the existing models of objects accumulated as a result of previous experience. If the comparison of the features of the object and the model was unsuccessful, then, in accordance with the priorities, the coarse features of the next object selected in the scene are compared. In the case of a successful comparison of features, depending on the purpose of the search, an object can be accompanied on a coarse image or, if necessary, a given image fragment with a higher resolution can be read for a more detailed examination, classification or recognition of an object. The center of the selected object and its overall dimensions has already been obtained when searching for the object. The central processor in descending channels controls these processes.

The expansion of the dynamic range of brightness perception can be provided by a nonlinear, for example, a logarithmic, scan of the reference voltage during parallel (throughout the matrix) analog-to-digital conversion (scanning method by parameter) [7,10]. If it is necessary to obtain a binarized image, it is also possible to use a nonlinear sweep of the reference voltage, but a threshold element and a trigger must be integrated in each element of the sensor matrix, which records the achievement of the threshold signal level by this matrix element [11].

The human visual system has in its arsenal a number of types of eye movements (saccades, mini-saccades, tremors, microtremors, tracking eye movements, vestibular-ocular movements to stabilize the image on the central fossa, vergent movements to reduce and dilate the axes of the eyes when they focus on the selected object). They contribute to a more efficient perception of video information in various modes (contemplation, panning when turning the head or observing a person, searching for an object in an image, detailed examination of an object for its classification/recognition).

When you turn the observer's head, for example, to the right on the retina, a part of the scene image is added on the right side and the part of the image on the left disappears. This feature can be effectively used to pan the scene, while eliminating the repeating part of previous frames and using only new information that arose when the camera was rotated. The transition from a dynamic image to a panoramic one can significantly reduce the amount of information that depends on the speed of head rotation or the movement of the observer. At the same time, it remains possible to restore a dynamic video sequence at any tempo in time and along an arbitrary trajectory of moving the gaze.

The amount of information in a video sequence can be estimated based on the amplitude-spatial and temporal resolution according to the formula

$$C_{v.seq} = \frac{X}{\Delta x} \cdot \frac{Y}{\Delta y} \cdot \log_2\left(\frac{Z}{\delta z} + 1\right) \cdot \frac{1}{\Delta t} \tag{1}$$

where X and Y are the sizes of the image field; Z is the brightness coordinate of the image; Δx, Δy, δz, Δt are discreteness of presentation of the corresponding image coordinates; $\frac{X}{\Delta x} \cdot \frac{Y}{\Delta y}$ is the number of pixels in the frame of the video sequence; $log_2(\frac{Z}{\delta z} + 1)$ is the bit depth representation of pixels; $\frac{1}{\Delta t}$ is the frame rate.

The values of X, Y and Z in a formula are usually fixed (maximum). The values Δx, Δy, δz, Δt are also fixed, therefore, this approach gives a (very high!) upper estimate of the amount of information and does not indicate ways to reduce the redundancy of the digital representation of images.

Providing the ability to dynamically change the parameters in the above formula (in accordance with the types of eye movements) will allow you to adapt the camcorder to the requirements and features of the tasks, as well as significantly reduce the redundancy of the representation of images and video sequences. In accordance with this approach, dynamic models of the processes of panning and editing of video sequences, automatic search for changes in the systems of circular and sector view, search and selection of objects or changes in the scene (movement, color, size) and tracking them were developed and studied [4,8,9].

Searching for an object by the visual system is carried out by identifying on the peripheral retina (coarse image) using special mechanisms of informative features of "suspicious" image fragments corresponding to the rough model of the object. Using saccades, moving and focusing in the central fossa of the image fragment with the supposed object is carried out. According to the principle of

a hierarchical coarse-grained representation of information [4], a coarse image of the scene (thinned out and then further coarsened) is first read out, in which all fragments that differ from the background are highlighted. In accordance with priority, these fragments are analyzed, during which crude informative features characterizing the object are identified and compared with the model of the desired object. The process continues until information is received about the presence of an object in the scene or its absence. Such an approach when searching for "suspicious" fragments due to image coarsenining allows reducing the amount of information by a factor equal to the degree of roughening for each fragment. Image coarsening, i.e. increase in the sampling step over the space Δx and Δy in n times:

$$\Delta x' = n \cdot \Delta x, \quad \Delta y' = n \cdot \Delta y$$

$$C_{image} = \frac{X}{\Delta x'} \cdot \frac{Y}{\Delta y'} \cdot \log_2(\frac{Z}{\delta z} + 1) = \frac{1}{n^2} \cdot \frac{XY}{\Delta x \Delta y} \cdot \log_2(\frac{Z}{\delta z} + 1) \qquad (2)$$

reduces the amount of information by n^2 times.

Reading an arbitrary rectangle from an image ranging from X_1 to X_2 and from Y_1 to Y_2

$$C_{image} = \frac{\Delta X}{\Delta x} \cdot \frac{\Delta Y}{\Delta y} \cdot \log_2(\frac{Z}{\delta z} + 1) \qquad (3)$$

reduces the amount of information in the image representation in $\dfrac{XY}{\Delta x \Delta y}$ times, where

$$\Delta X = X_2 - X_1, \quad \Delta Y = Y_2 - Y_1$$

Tracking of an object can be carried out both by its rough features, and by more subtle ones. In this case, the trajectory of the object's movement, changes in its shape, color are tracked and the rest of the scene is ignored. When implementing this approach, the video camera, having information about the overall dimensions of the object and its location, reads only a fragment of the image with dimensions slightly exceeding (taking into account the speed of the object) the overall dimensions of the object. Moreover, the reduction in the amount of information is determined in accordance with formula (3). A coarse presentation of information may concern not only the resolution of the spatial representation of the image, but also the accuracy of the representation of brightness (color), i.e. bit depth. The brightness (color) picture of the image is not constant, but varies from pixel to pixel and from frame to frame. Therefore, the value Z in the formula (1) is not a constant, but a function of the image coordinates $Z_{ij} = f(x_i, y_j)$.

In [8], the concept of the entropy of a random variable value was

$$H_N = \sum_{i=1}^{k} p_i \log_2(N_i + 1)$$

introduced, which is a measure of the uncertainty of the random variable value itself and represents the average number of digits per one random variable value. Considering the matrix (m × n) of pixel brightness values as random values, we obtain an image entropy estimate [4]

$$H_{image} = \sum_{j=1}^{n} \sum_{i=1}^{m} p_{ij} \log_2 \left(\frac{z_{ij}}{\delta z} + 1 \right) \tag{4}$$

As for the Shannon statistical entropy, the logarithm base determines the unit of measurement of the entropy of a random variable. At the logarithm base of two, the unit of entropy of the value is a bit. Normalizing the H_{image} by the value of $\log_2(\frac{z_{max}}{\delta z} + 1)$, we obtain the reduced entropy of the image brightness characteristic,

$$h_z = \frac{H_{image}}{\log_2(\frac{z_{max}}{\delta z} + 1)} \tag{5}$$

The reduced entropy of the brightness characteristic of the image characterizes the spread in the bitness of the representation of the brightness of the image pixels and varies in the range (0 1). This characteristic will allow us to evaluate the effectiveness of using various methods for representing the variable bit depth of the brightness (color) parameter of the image. A decrease in bit depth leads, respectively, to a decrease in the amount of information in the image, however, in most cases it is rather difficult to use this decrease in the amount of information in full.

Conditionally, we can assume that the amount of information received from a color video camera increases by 3 times. However, in real-time systems, in a number of applications, it is possible to use only individual *R*-, *G*-, and *B*-components, or, going to the HSB model, for example, use only the color or brightness characteristics, which to a greater extent carry useful information for this task. Therefore, to increase the selectivity, it is advisable to provide the ability to read only the necessary information in this task. Change the frequency of video recording, i.e. an increase or decrease in the sampling step in time Δt, leads to a proportional decrease in the amount of information in the video sequence. Video frequency control can be organized taking into account the dynamism of the scene, i.e. the speed of ongoing processes or the speed of moving objects in the scene.

Thus, using the considered dynamic models for selecting the parameters for reading information from the video sensor and the principles of controlling its pipeline for processing this information, it is possible to significantly increase the performance of the camera by adapting it to the conditions of the current task. Further, we consider those computing operations that cannot be implemented on the pipelines of existing video sensors and require the use of additional computing tools (central processor, group of graphic processors, FPGAs, etc.) for their implementation.

It is known that the retina operates when analyzing images not with absolute values of brightness or color of pixels, but with their differences between adjacent

pixels in a row, column or between frames. The differential representation of information on the retina is obtained due to special eye movements (tremor and microtremor). As studies have shown, the use of difference representations, which actually represent the contrast or changes between frames, is an effective way to highlight a number of useful properties in the image and video sequence. Taking into account the presence of correlation between neighboring elements in the row and column of the matrix, the image can be represented as a matrix of differences between neighboring elements (growth matrix). In this case, the entropy of the image is determined in this way:

$$H_{image} = \sum_{j=1}^{n} \sum_{i=1}^{m} p_{ij} \cdot log_2 \left(\left| \frac{\Delta z_{ij}}{\delta z} \right| + 1 \right) \tag{6}$$

where the differences Δz_{ij} can be determined both by rows ($\Delta z'_{ij}$), and columns (Δz_{ij}) of the matrix:

$$\Delta z'_{ij} = z_{i+1,j} - z_{ij}, \quad \Delta z''_{ij} = z_{i,j+1} - z_{ij}$$

Typically, fewer bits are required to encode differences in brightness, which will also reduce the amount of information in the image. Given the correlation between the corresponding pixels of two adjacent frames, we obtain a difference image matrix [4], the entropy of which is defined as

$$H_{dif} = \sum_{j=1}^{n} \sum_{i=1}^{m} p_{ij} \cdot \log_2 \left(\left| \frac{\Delta z_{ij}^{k}}{\delta z} \right| + 1 \right) \tag{7}$$

where $\Delta z_{ij}^{k} = \Delta z_{ij}^{k+1} - \Delta z_{ij}^{k}$, k is the frame number of the video sequence.

In the difference image, similarly to the previous one, useful information can also be distinguished, i.e. moving object or scene changes between frames. In this case, similarly to the previous one, a reduction in the amount of information in the image is also provided here. In contrast to the statistical measure of information by C. Shannon, developed to compress information during transmission and storage, a dynamic measure of information is proposed - δ-entropy. It is defined as the average value of the modulus of the video derivative and is a measure of the uncertainty of a random process change [25]:

$$H_{\delta} = \frac{\Delta t}{\delta_z} \cdot M[|f'(t)|] \tag{8}$$

δ-entropy allows you to extract useful (dynamic) information from signals, images, video sequences, iterative processes, etc., significantly reducing its redundancy, and can be the basis for evaluating the processes of conversion and processing of information in real-time systems.

In a discrete form, the estimate of the δ-entropy of the image is defined as

$$H_{\delta} = \sum_{j=1}^{n} \sum_{i=1}^{m} p_{ij} \cdot \left| \frac{\Delta z_{ij}}{\delta z} \right| \tag{9}$$

where Δz are the differences between the brightness of pixels in rows or columns; δ-entropy is a measure of the dynamism of an image, characterizes its contrast and can be used to assess the information content of an image, the amount of useful (dynamic) information in it, spatial frequencies, search and evaluate the characteristics of textures and objects, as well as to control the level of contrast.

The known method of averaging transverse slices of the image brightness profile over rows and columns [16] allows you to determine the location of an object in the image that differs from the background, but does not allow you to select objects of the texture type. To determine the location of such objects, it is proposed to average (modulo) the transverse sections of the image according to dynamic signs, i.e. by differences in brightness between adjacent pixels in rows and columns [20]:

$$H_{\delta_j} = \frac{1}{M} \cdot \sum_{i=1}^{M} \left| \frac{\Delta z_{ij}}{\delta z} \right|, \quad H_{\delta_i} = \frac{1}{N} \cdot \sum_{j=1}^{N} \left| \frac{\Delta z_{ij}}{\delta z} \right| \tag{10}$$

An important characteristic for texture analysis and a number of other applications are spatial frequencies, for the analysis of which the Fourier analysis is most often used. In this paper, instead of the Fourier analysis, it is proposed to use δ-entropy, which is calculated much easier. Such a dynamic measure of information can be effectively used for a number of applications. In particular, for:

- image segmentation into high and low dynamic sections;
- segmentation of text information in the image;
- search and classification of textures;
- search for barcodes, DMX codes, fingerprints, car numbers, etc.;
- extracting movement, changes in brightness or color;
- control the frequency of shooting a video camera, etc.

Estimating the contrast level of the scene image by rows and columns, we can select areas with significantly different values and take measures to correct it. In addition, you can classify textures as irregular and regular (Fig. 1), determine the step of the texture, its contrast, identify changes in the texture (defect), etc.

According to the graphs of δ-entropy (Fig. 1), we can distinguish between rows and columns: a) - irregular texture; b), c), d) - regular textures.

From the distances between the maxima or minima, you can determine the step of regular textures, by the magnitude of the signal amplitude - the value of the contrast of the texture; if the regularity of the texture is violated, a defect can be detected - c). Contrast control can be carried out by changing the size of the masks (filters on-and off-centers) and their gains. For example, a 3×3 mask provides maximum contrast enhancement of 4 or 8 times, a 5×5 mask - 16 or 32 times, etc. The gain is selected taking into account contrast by choosing a mask so that the result of its application is, for example, in the second half of the brightness representation scale (128–255 for an 8-bit scale). Masks are desirable to adjust so that their coefficients are multiples of the power of two in order to exclude the operation of multiplying by coefficients. This will replace

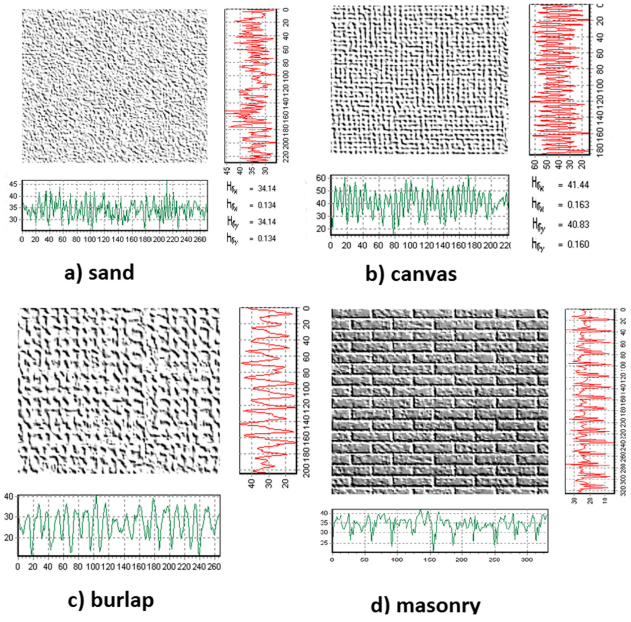

Fig. 1. Assessment of texture parameters

the operations of multiplication by the operations of shift by the corresponding number of bits, which is especially effective in the hardware implementation of such procedures.

A very important point is the providing of the ***principle of ring organization of receptive fields***, since it equally effectively works both on the model of the peripheral retina and on the model of the central fossa. In addition, on the basis of the *center-surround* principle, one can extract many informative features of a higher level on the retina model (spots that differ from the background in brightness, color, orientation, dynamic characteristics - spatial frequencies, the presence of motion in the video sequence, etc.) that provide the search for objects in the scene. The principle of *center-surround* in accordance with the three-component color theory is used by the visual system not only to extract individual colors, but also at a higher level of the visual system when extracting colors in accordance with the theory of opponent processes.

It is problematic to organize parallel processing of all signs and sequential on different scales, as is done on the retina, on modern video sensors. However, they can be implemented in parallel-serial or serial-parallel manner on a large number of GPU or on the FPGA. Although with large coarsening on the peripheral retina, it becomes necessary to repeatedly calculate the sums of square or rectangular fragments, this can be effectively solved using an integrated matrix [2]. The same principle can be applied to calculate the *center-surround* operator (with the same inhibition zone coefficients).

In this case, the total amount of the entire *center-surround* fragment is calculated from 4 points, and then the center area is determined from 4 points, i.e. it takes only 7 operations of addition/subtraction and one shift operation for several digits (using masks with a central coefficient multiple of degree 2). In the central fossa, according to the principle of *center-surround*, more detailed informative features are extracted (brighter/darker or different in color or orientation points that characterize the edges of areas or contours of objects, spatial frequencies of textures, etc. that stand out from the background). The *center-surround* principle is implemented in computer vision using appropriate masks (Laplace, Roberts, Sobel, Prewitt, etc.), which are already widely used in image processing techniques to extract the edges and boundaries of areas, object contours, orientation features, etc.

With an insufficient level of illumination, by analogy with the grouping of rods on the peripheral retina, the summation of signals from rods for larger receptive fields is performed, which can significantly increase the sensitivity to light perception, i.e. exchange spatial resolution for increased sensitivity.

In the case of insufficient light in the arsenal of the human visual system, there are techniques that can significantly increase the sensitivity to light perception in exchange for a decrease in spatial resolution. They consist in the possibility of combining and summing the signals of neighboring rods of the peripheral retina using horizontal cells, as well as the ring concentric organization of neurons, which allows to increase the contrast in the perception of the scene.

The sensitivity of light perception is controlled by changing the number of summed signals of the rods and by changing the number of concentric rings around the central excitation zone of on- and off-centers. Thus, the coefficient of sensitivity increase is directly proportional to the number of summed signals from neighboring retinal rods and the gain due to the ring organization of neurons. Examples of using this approach and the results of increasing the sensitivity are shown in Fig. 2.

However, with large degrees of coarsening, i.e. combining a large number of rods, the gain of sensitivity is very large (in the range from 1 to n^2) and they need to be adjusted. At the same time, large degrees of roughening can be effectively used to quickly search for "interesting" places in the scene image in accordance with the dynamic model of the visual analyzer. In this case, combining and summing the signals from the rods can be carried out with the averaging operation, i.e. with a coefficient of gain equal to the central element of the mask.

An increase in contrast can be achieved by building up the brake rings around the exciting center, i.e. by using masks of large sizes (for example, 5×5, 7×7, etc.). The principle of ring concentric organization, as noted above, can be used to extract a large number of coarse and subtle informative features not only in monochrome, but also in color images. Let's consider this issue in more detail on the example of the on-center. As you know, in the Bayer color representation

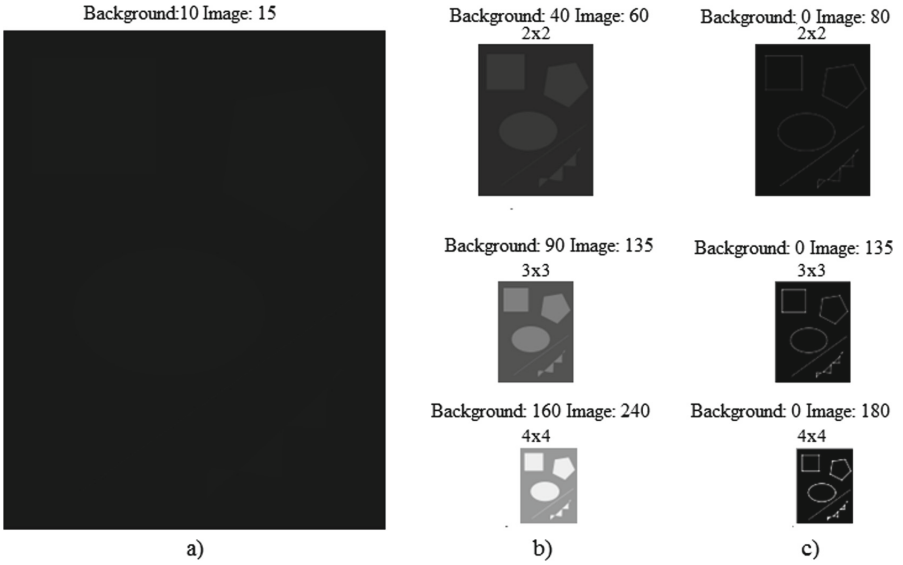

Fig. 2. Increasing the sensitivity of the peripheral retina by summing the signals from the rods and their ring organization: a) input image; b) the result of amplification due to the summation of the signals of fields of size 2×2, 3×3 and 4×4; c) the result of amplification due to the ring organization with a Laplace 3×3 mask

model, three basic colors are placed in the sensor matrix as follows:

$$\begin{array}{c} r\ g\ r\ g\ r\ g\ r\ g\ r\ g \\ g\ b\ g\ b\ g\ b\ g\ b\ g\ b \\ r\ g\ r\ g\ r\ g\ r\ g\ r\ g \\ g\ b\ g\ b\ g\ b\ g\ b\ g\ b \\ r\ g\ r\ g\ r\ g\ r\ g\ r\ g \\ g\ b\ g\ b\ g\ b\ g\ b\ g\ b \end{array}$$

If we consider 3×3 masks, we can see that to extract blue (exciting) color, you need to use an 8-connected mask h_1. Moreover, for color B, the colors r and g are inhibitory (opposing), and colors b and g are inhibitory for color R. For color G, the colors r and b are inhibitory, but this uses a 4-connected Laplace mask h_2.

$$h_1 = \begin{vmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{vmatrix}, \quad h_2 = \begin{vmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{vmatrix}$$

For the off-center, the signs in the masks are reversed, and with their help it is possible to identify color pixels that are smaller in magnitude of the corresponding opponent colors in the mask. This fully corresponds to the three-component theory of color vision [11]. In accordance with the theory of opponent processes

in the human visual system, in the second stage, which takes place at higher levels (possibly in the LGB), according to the same principle, center-surround, differences in colors are revealed in larger receptive fields. As a result of the implementation of the three-component color theory, 3 color planes are obtained: R, G and B. Additionally, the Y plane is formed in accordance with the formula

$$Y = (r + g)/2 - |r - g|/2 - b$$

Further, negative values are zeroed, after that pairs of opponent colors RG and BY are considered, for example: R is the center, G is the surround; B is the center, and Y is the surround, or vice versa. In this case, the selection of colors can be performed at different scales of the presentation of information. To reduce the amount of information when roughening the color layer or implementing the *center-surround*, as for monochrome images, an integrating matrix can be used.

An important point in order to reduce the delay in the feedback circuit of the CVS is the combination of the processes of reading and processing information from the sensor matrix. This combination is implemented using 2–3 channels of direct access to memory. In this case, the first channel controls the input of information into memory, the second - reads information for processing, and the third - controls the recording of results in a new memory array. This approach reduces the processing processor's time spent working with several sources and receivers of information and allows obtaining information for controlling the reading parameters of the next frame of the video sequence with minimal time after reading data. Typically, the frequency of reading information in modern video sensors is about one and a half to two orders of magnitude lower than the frequency of the processing processor (for example, a digital signal processor or graphic processors), which allows up to hundreds of image processing operations to be performed during the reading of one data from a video sensor. Since the time diagram of the video sensor also contains additional losses of time for switching to reading the next line and for completing a frame, the delay in information in the CVS feedback circuit is reduced by approximately the processing time of one frame in comparison with the traditional approach.

Such approaches were used to create the first in Ukraine IVK-1 family of intelligent video cameras (2000) and a number of real-time video systems based on them. Among them:

- product quality control system for color, size, shape;
- control system for static and dynamic parameters of physical, chemical and biological objects;
- digital optical capillaroscope for non-invasive control of static and dynamic parameters of human blood microcirculation;
- MacroMicroPotok hemodynamic laboratory based on an advanced Dopplerograph and a digital optical capillaroscope for monitoring the human cardiovascular system at macro and micro levels;
- tracking device for the selected object, etc.

6 Conclusions

Using an arsenal of methods and mechanisms for processing information on the retina of a human visual analyzer made it possible to propose a number of original principles for organizing the search for an object in an image, tracking it and extracting informative features for recognition. In particular:

- the modernization of the video sensor conveyor based on dynamic models and the control of information reading parameters can significantly (by several orders of magnitude) reduce the amount of processed information. This is achieved due to the possibility of changing the resolution of the video sensor, coarse-accurate representation of the scene (coarse - when searching for an object or tracking it, and accurate representation of only fragments of the image of the object - during recognition) [9]. At the same time, the same video sensor is used for coarse and detailed processing, and the combination of image input and processing processes can significantly increase the efficiency of processing video information;
- the use of the proposed principles of the ring concentric organization of on- and off-centers and an integrating matrix for their implementation, adaptation mechanisms to the level of illumination and contrast, the method of determining spatial frequencies based on δ-entropy, the differential representation of images for their analysis can significantly simplify the implementation of these principles and accelerate the process of training ANN in the subsequent stages of classification/recognition of objects. The selection of informative features directly in CVS allows reducing the amount of information transmitted to higher levels of information processing (for example, in the ANN) by 3–4 orders of magnitude [9];
- the introduction of a processing processor into the CVS (DSP processor, a group of graphic processors, FPGAs, ANNs of convolution type, etc., depending on the characteristics of the task and performance requirements) to implement the above principles will significantly increase the productivity, efficiency and effectiveness of the CVS. A significant increase in productivity can be achieved by constructing a multilayer matrix that combines parallel reading, analog-to-digital conversion and information processing (suitable for specialized applications) [6,10–13].

Acknowledgments. This work was carried out in the framework of fundamental competitive topics (VFK 200.15 and VFK 200.19), which were funded by the Presidium of the National Academy of Sciences (NAS) of Ukraine.

I express my sincere gratitude to the employees of the Department of Intelligent Real-Time Video Systems of the Institute of Cybernetics named after V.M. Glushkov NAS of Ukraine, which took part in the creation of tools, modeling and verification of theoretical provisions.

References

1. Anderson, D.: Cognitive Psychology, 5th edn. Piter, St. Petersburg, Russia (2002). (Russian translation)

2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. **110**(3), 346–359 (2008). https://doi.org/10.1016/j.cviu.2007.09.014
3. Benoit, A., Caplier, A., Durette, B., Herault, J.: Using human visual system modeling for bio-inspired low level image processing. Comput. Vis. Image Underst. **114**(7), 758–773 (2010). https://doi.org/10.1016/j.cviu.2010.01.011
4. Boyun, V.: Intelligent selective perception of visual information in vision systems. In: Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Application, IDAACS 2011, Czech Republic, Prague, vol. 1, pp. 412–416 (2011)
5. Boyun, V.: Directions of development of intelligent real time video systems. Appl. Theor. Comput. Technol. [S.l.] **2**(3), 48–66 (2017)
6. Boyun, V.P., Voznenko, L.O., Malkush, I.F.: Principles of organization of the human eye retina and their use in computer vision systems. Cybern. Syst. Anal. **55**(5), 701–713 (2019). https://doi.org/10.1007/s10559-019-00181-0
7. Boyun, V.: The dynamic theory of information. Fundamentals and applications. Institute of Cybernetics of NASU, Kyiv, Ukraine (2001)
8. Boyun, V.: A human visual analyzer as a prototype for construction of the set of dedicated systems of machine vision. In: Proceedings of the International Science and Technology Conference on "Artificial Intelligence", Intelligent Systems II-2010, vol. 1, pp. 21–26 (2010)
9. Boyun, V.: Intelligent selective perception of visual information: informational aspects. Artif. Intell. **3**, 16–24 (2011). (in Ukrainian)
10. Boyun, V.: Device for determining the location and parameters of image objects, UA patent no. 76597, BI no. 6 (2013)
11. Boyun, V.: Sensor device for determination of location and center of gravity of an object, UA patent no. 106292, BI no. 12 (2014)
12. Boyun, V.: Sensor device for determining the location and moments of inertia of an object in an image, UA patent no. 106301, BI no. 15 (2014)
13. Boyun, V.: Sensor matrix with image processing, UA patent no. 109335, BI no. 6 (2015)
14. Burt, P.: Smart sensing within a pyramid vision machine. Proc. IEEE **76**(8), 175–185 (1988). https://doi.org/10.1109/5.5971
15. Gollisch, T., Meister, M.: Eye smarter than scientists believed: neural computations in circuits of the retina. Neuron **65**(2), 150–164 (2009). https://doi.org/10.1016/j.neuron.2009.12.009
16. Gonsales, R., Woods, R.: Digital Image Processing. Technosphere, Moscow, Russia (2005)
17. Hubel, D.H.: Eye, Brain and Vision. Scienceific American, New York (1988)
18. Kolb, H.: How the retina works: much of the construction of an image takes place in the retina itself through the use of specialized neural circuits. Am. Sci. **91**(1), 28–35 (2003). https://doi.org/10.1511/2003.1.28
19. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: NIPS Proceedings. Advances in Neural Information Processing Systems, vol. 25 (2012). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-network
20. Marr, D.: Computational Investigation into Human Representation and Processing of Visual Information. W.H. Freeman and Company, New York (1987)
21. Podvigin, N., Makarov, F., Shelepin, Y.: Elements of Structural and Functional Organization of Visual Oculomotor System. Nauka, Leningrad, USSR (1986). (in Russian)

22. Schiffmann, H.: Sensation and Perception: An Integrated Approach. Piter, St. Peterburg (2003). (Russian translation)
23. Shah, S., Levine, M.: Visual information processing in primate cone pathway - part i: a model, part ii: experiments. IEEE Trans. Syst. Man Cybern. Syst. Part b Cybern. **26**(2), 259–289 (1996). https://doi.org/10.1109/3477.485837
24. Shelepin, Y., Bondarko, V., Danilova, M.: Foveola construction and visual system pyramidal organization model. Sens. Syst. **9**(1), 87–97 (1995). (in Russian)
25. Shevelev, I.: Neurons of Visual Cortex. Adaptability and Dynamics of Receptive Fields. Nauka, Moscow (1984). (in Russian)
26. Siagian, C., Itti, L.: Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Trans. Pattern Anal. Mach. Intell. **29**(2), 300–312 (2007)
27. Supin, A.: Neuron Mechanisms of Visual Analysis. Nauka, Moscow, USSR (1974). (in Russian)
28. Tagare, H., Toyama, K., Wang, J.: A maximum-likelihood strategy for directing attention during visual search. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 490–500 (2001)
29. Yamasaki, H., Shibata, T.: A real-time image-feature-extraction and vector-generation VLSI employing arrayed-shift-register architecture. IEEE J. Solid-State Circ. **42**(9), 2046–2053 (2007)