Theoretico-informational approach to the introduction of feedback into multilevel machine-vision systems

A. S. Potapov

NPK S. I. Vavilov State Optical Institute, St. Petersburg

(Submitted April 4, 2007)
Opticheskiĭ Zhurnal **74**, 59–65 (October 2007)

This paper discusses the problem of introducing feedback in multilevel machine-vision systems. Based on a theoretical-informational analysis, it is shown that feedback is needed in such systems because the individual components of the quality criterion common to the entire system are optimized at different levels. As a result, the decisions made at the earlier stages of the analysis use only part of the components of this criterion, and this makes these decisions less than optimal. An approach is proposed to introduce feedback as a method of achieving the global optimum of the informational quality criterion by iterative improvement of the decision. Based on these results, it is shown that the information contained in the contours is inadequate for a robust construction of the structural elements. © *2007 Optical Society of America*.

I. INTRODUCTION

The importance of the concept of feedback is well known. The presence of feedback in guidance from the environment in biological systems largely determines their capability for survival. Adaptive and trainable robotic systems are also based on feedback, which appears as the result of machine-perception subsystems. However, feedback can be established not only from the environment to a cybernetic system, but also between the subsystems of the system itself, where it plays just as great a role.

This article discusses machine-vision systems, which in most cases are currently being developed as multilevel systems.[1] In multilevel systems, a signal arrives at the input—the lower level—and then propagates to higher levels, being subjected to ever more complex analysis. However, it is well known[2] that in biological perception systems there are also numerous feedbacks from the upper to the lower levels. The upper levels in this case, by invoking wider contextual information, control the analysis process carried out at the lower levels.

Machine vision involves methods in which an information flux from the upper to the lower levels is introduced. These methods predominantly relate to methods based on knowledge and are ordinarily used when the contents of the input signal at the upper level are fairly well known *a priori* and require only to be refined (see, for example, Ref. 3).

When one is dealing with feedback from the environment, the meaning of this connection is fairly clear. For example, in automatic control systems with feedback, the problem to be solved usually involves maintaining a given value of an adjustable quantity that describes the state of the controllable object (the environment) on the basis of information contained in the input signal (See Ref. 4, page 26). The information arriving at the feedback is generated by the environment itself.

However, when feedback is introduced into multilevel machine-vision systems, it becomes necessary to determine not only the reaction to the signals that arrive via these connections but also the character itself of the connections; i.e., one needs to determine just what information should be transmitted “downward” and with what purpose. Naturally, the introduced feedback must make it easier to solve the problem of the inductive conclusion itself. We treat this problem as a problem of inductive logic, and we investigate it in terms of the theoretical-informational approach.[5] It is shown in this paper how the theoretical-informational approach can be used to explain the purpose of feedbacks in multilevel systems of perception and to develop techniques for using them.

II. EXISTING APPROACHES

A. Adaptive resonance theory

One of the first detailed investigations of the role of feedback in perception led to the development of adaptive resonance theory (ART) at the end of the 1970s. The concept itself of adaptive resonance is most obviously expressed in a neural network of the ART type, proposed by S. Grossberg.[6,7]

Such a network consists of several (in the simplest case, two) layers. The first layer is the input layer. Each of its neurons represents a certain attribute of an object. In the second layer, the neurons correspond to classes of objects. Each neuron of the second layer acquires information on the neurons of the first layer and is activated when the attributes correspond to the class of objects associated with the given neuron. This is similar to a direct-propagation network, but connections also go from the neurons of the second layer to the first layer. If any neuron of the second layer is active it encourages the neurons of the first layer whose activity pattern corresponds to the attributes of the reference object of the given class. Current information is thus transmitted from bottom to top, while expectation is transmitted from top to bottom.

An ART network is recursive; i.e., variation of the activity of the neurons of the first layer affects the activity of the neurons of the second layer, which in turn affects the activity of the neurons of the first layer. If any first-layer neurons do not obtain reinforcement from the second layer, their activity gradually damps out. Conversely, the activity of neurons that have received an initially weak sensory signal can be amplified if they receive strong reinforcement from above (if the initial activity of the second-layer neurons corresponds to the *a priori* expectations).

Only resonance states are maintained in such a network—i.e., activity patterns of the neurons at two levels that support each other. The result of perception is thus built up from expectations and new information. The activity of the lower-level neurons is thus corrected, and this makes it possible to restore distorted and omitted information.

The time to establish resonance can serve as the explanation of why perception is a slower process than direct transmission of signals from the sense organs to the brain.

There are a number of illusions of perception, which are explained by this theory.[2] The reinforcement that comes from the upper levels of an ART neural network also makes it possible to model the control of attention: The attributes of those objects in the perceptive field to which attention is directed are reinforced. However, the most important thing is the fact that this model has disclosed the mechanisms of perception under conditions of noise and indeterminacy of the sensory signals, and this is extremely critical for machine-perception systems.

Unfortunately, the concept of adaptive resonance has been developed predominantly in terms of the neural-network approach,[7] even though its significance is substantially broader. The ART neural network itself (like its refinement) possesses certain disadvantages in practice. The architecture of such a network is convenient for recognizing a single object with a fixed number of attributes. However, this architecture is not suitable for analyzing a complex visual scene or verbal statements of indeterminate length.

The architecture itself of a network with feedback does not tell how these connections should be adjusted to optimize its functioning. Although an ART network makes it possible to elucidate a number of questions in perception and implements a number of fundamental concepts, the construction of the optimum technical system of perception lies beyond the neural-network paradigm.

B. The Bayesian approach

Let us consider the problem of introducing feedback in terms of the Bayesian approach, which is well regarded and widely used in interpreting images.[8]

Let D be the initial data, while H is some hypothesis concerning their contents. Then, in terms of the Bayesian approach, the best hypothesis will be that which maximizes the *a posteriori* probability $P(H|D)$, expressed, according to Bayes's rule, in terms of the probability of the data, $P(D|H)$, and the *a priori* probability of the hypothesis, $P(H)$:

$$p(H|D)=\frac{P(D|H)P(H)}{P(D)}, \tag{1}$$

where the *a priori* probability of the data, $P(D)$, is independent of the hypothesis and can be neglected in the decision process.

We shall consider multilevel systems of inference, and therefore we need to write

$$\begin{aligned}
&P(H_1|D)=P(D|H_1)P(H_1)/P(D),\\
&P(H_2|H_1)=P(H_1|H_2)P(H_2)/P(H_1),\\
&\dots\\
&P(H_N|H_{N-1})=P(H_{N-1}|H_N)P(H_N)/P(H_{N-1}).
\end{aligned} \tag{2}$$

Here H_i is an ith-level hypothesis, for the selection of which an $(i-1)$st-level hypothesis is used as the initial data.

In the case of a multilevel system, the joint *a posteriori* probability $P(H_1,H_2,\dots,H_N|D)$ needs to be maximized, where $(H_1,H_2,\dots,H_N)$ is the general hypothesis concerning the contents of the sensory signal, which is represented in the form of a collection of partial hypotheses of different levels.

Let us consider two levels. Then $P(H_1,H_2|D)=P(H_2|H_1)P(H_1|D)$. We now assume that at the first level a certain best hypothesis H_1^* is chosen for which the maximum $P(H_1|D)$ is achieved. After this, it is possible to maximize the value of

$$P(H_1,H_2|D)=P(H_2|H_1^*)P(H_1^*|D), \tag{3}$$

where $P(H_1^*|D)=\text{const}$.

We now point out the following: The maximization of $P(H_1|D)$ does not necessarily maximize $P(H_1,H_2|D)$, since, for a chosen H_1^*, the other factor $[P(H_2|H_1^*)]$ may be small.

However, an exhaustive search for all possible pairs (H_1,H_2) in order to find the maximum $P(H_1,H_2|D)$ is inefficient from a computational viewpoint. As the amount of initial data increases (for example, as the image size increases), the hypothesis space increases exponentially, and the problem of combinatorial explosion arises. Thus, it is necessary on one hand to avoid an exhaustive search for all possible combinations of hypotheses, and, on the other hand, it is necessary to ensure that the maximum is found.

In order to do this, one usually passes from the lower to the upper levels not one best hypothesis but several alternatives. That hypothesis H_1^* is left that does not simply give the maximum $P(H_1|D)$ but also for which such a hypothesis H_2^* is found that the factor $P(H_2^*|H_1^*)$ will be large enough [so that $P(H_1^*,H_2^*|D)$ will be greater than for the alternatives]. Too many of such alternatives, however, cannot be considered, since, for each lower-level hypothesis, hypotheses will need to be constructed on the next level. Consequently, the most distinct alternatives are selected. From hypotheses that are similar in content, the best is chosen, but this hypothesis may also be nonoptimal at the next level.

In this connection, it is possible to introduce the following operation, which, it is true, is usually not implemented in the Bayesian approach. After the best hypothesis H_2^* is found for some hypothesis H_1^*, it is possible to find such a hypothesis H_1' that maximizes the second factor $P(H_2^*|H_1)$, after which H_1^* can be approximated to H_1'. It is possible to conditionally write $H_1^*=(1-\varepsilon)H_1^*+\varepsilon H_1'$, where ε is some factor. The expected hypothesis tells what the hypothesis of the preceding level must be. Iterative mutual correction of H_1 and H_2 in order to maximize the product gives the optimum values of H_1 and H_2.

Thus, in the Bayesian approach, as in the approach based on adaptive resonance, several hypotheses can be activated at each level, while the effect of adaptive resonance itself corresponds to the iterative optimization of the product $P(H_2|H_1)P(H_1|D)$ due to both cofactors, and this replaces an exhaustive search in hypothesis space.

The Bayesian approach is mathematically correct, and it can be used to track fairly closely the causes of the appearance and the meaning of feedback; however, it can be hard to use because of the need to compute the probabilities for hypotheses that possess a complex structure. For example, it is extremely problematic to evaluate $P(H_1|D)$, where D is an image, while H_1 is a collection of contours distinguished on it. The use of the Bayesian approach in multilevel systems causes no difficulties until one reaches the middle levels (phonemes or letters) when recognizing oral or written speech or when recognizing a narrow class of objects from attribute vectors of a fixed size.

It can also be nontrivial to assign *a priori* probabilities that determine expectations. If there is no absolutely clear expectation of the contents of the input signal, it is problematic to specify the *a priori* probabilities, especially at the lower levels of a system of perception. For example, how can one specify *a priori* probabilities for various kinds of contours on an image when there is no expectation of a specific contour at a specific site? The problem of *a priori* probabilities is very deep[9] and will not be considered here.

III. THE THEORETICAL-INFORMATIONAL APPROACH TO THE INTRODUCTION OF FEEDBACK

We propose a solution of the feedback problem in terms of a theoretical-informational approach based on the principle of minimum description length. In a verbal formulation, this principle states:[9] Among all the models that explain data, one should choose the one that minimizes the sum of the description length of data by the model and the description length of the model itself. The description length, as a rule, is computed as the length of the lines that code the data by means of the model and that code the model itself. The coding scheme is determined by the representation of the information that corresponds to the metamodel of the given object region.

The hypothesis is thus chosen by minimizing the description length

$$L(D,H)=L(D|H)+L(H), \tag{4}$$

and this is the information analog of Bayes's rule. However, unlike the Bayesian approach, the theoretical-informational approach makes it possible to evaluate the description length of hypotheses with a complex structure.

Let us consider the theoretical-informational approach to the introduction of feedback in image-interpretation systems. We shall not analyze the question itself of the possible representations of information in these systems but shall use the representations described in Ref. 5.

We shall assume that image $f(x,y)$, given in region G, i.e., $f:G\rightarrow R$, is described as a collection of nonintersecting regions $G=\bigcup_i G_i$. The contents of each region (the narrowing of image f onto region G_i, i.e., $f|_{G_i}$) is described by an individual stochastic model. In the simplest case, the intensities of the pixels in region G_i are described as statistically independent and identically distributed (the probability distribution function is different for different regions). In a more complex case, the spatial dependences of the pixel intensities can be taken into account. Such a representation is caused by the fact that observed scenes consist of visible surfaces, each of which possesses an individual reflectivity pattern.

Besides the pixel intensities inside each region, it is also necessary to describe the position of these regions. This can be done by describing their boundaries ∂G_i.

The description length of an image in terms of some hypothesis concerning how this image is broken up into regions is then represented in the form of a sum:

$$DL=\sum_i \left(L(f|_{G_i})+L(\partial G_i)\right). \tag{5}$$

The question next arises of the coding of the contours (the boundaries of the regions). In the representation that we considered, the contours are described by a collection of structural elements, each of which corresponds to some segment of the contour. The description must indicate the boundaries of the segments (this corresponds to the corners and intersections on the contour) and the shape of the line that describes the segment itself (for example, the parameters of a straight line or arc of a circle or ellipse). Since the contour itself is described by structural elements with certain errors, these errors also need to be coded, in order to compute the correct description length.

Let $s_j^{(i)}$ be the jth structural element on the ith contour ∂G_i. Then

$$L(\partial G_i)=\sum_j \left(L(\partial\, G_i|_{s_j^{(i)}})+L(s_j^{(i)})\right). \tag{6}$$

Next, the structural elements $s_j^{(i)}$ can be described by grouping them according to the similarity and regularity of the spatial position (Ref. 5, p. 316), but we shall restrict ourselves to the first three levels.

Thus, to compute the exact value of the quality criterion of Eq. (5) for each hypothesis concerning how the image breaks up into regions, it is necessary to compute the value of Eq. (6), for which the structural elements should be constructed on the basis of the boundary of these regions [i.e., one must find the $s_j^{(i)}$ values that minimize $L(\partial G_i)$ for each collection of regions G_i]. This is equivalent to an exhaustive search for all possible pairs of hypotheses H_1 and H_2 in the Bayesian approach.

As was said above, this is not efficient from a computational viewpoint. On the other hand, the minimization of the

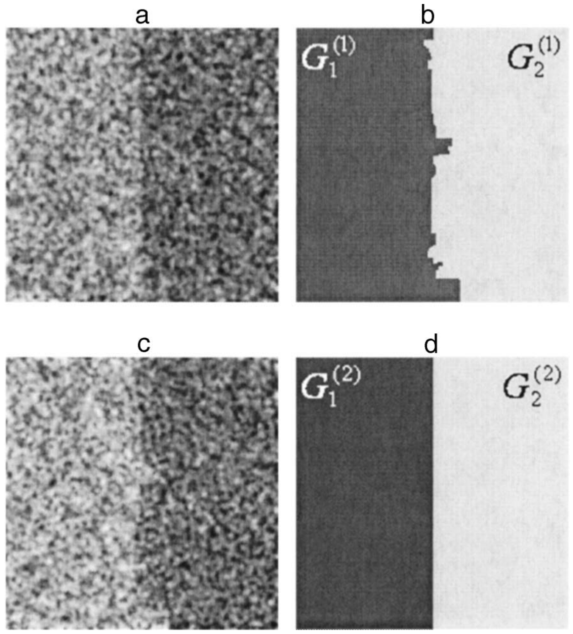

FIG. 1. Separating an image into regions using feedback. (a) Synthesized image consisting of two regions separated by a straight (vertical) boundary, (b) regions found by minimizing the quality criterion of Eq. (7), (c) image synthesized using the regions of (b), (d) regions found for image (a) as a result of correcting regions (b) by means of feedback.

term $L(\partial G_i|_{s_j^{(i)}})$ neglecting the term $L(s_j^{(i)})$ results in the construction of a large number of regions with strongly irregular boundaries, and therefore some simple estimate of the description length $L(\partial G_i)$ should be used for which it is not required to construct a next-level hypothesis. In other words, instead of refined coding of the contours via the structural elements, it is possible to use somewhat less efficient but far faster coding of the contours—for example, chain coding, which indicates the first point of a contour and the direction from each point to the next.

Thus, the quality criterion

$$DL_0 = \sum_i (L(f|_{G_i}) + L_0(\partial G_i)) \tag{7}$$

can be used to construct the regions, where $L_0(\partial G_i)$ is a quick estimate of the value of $L(\partial G_i)$, with $L_0(\partial G_i) \geqslant L(\partial G_i)$.

Next, the construction of the structural elements with minimization of the quality criterion of Eq. (6) can be accomplished with fixed contours.

Obviously, the hypothesis concerning the regions on the image found by minimizing the criterion of Eq. (7) may not minimize the general criterion of Eq. (5). The question of drawing a boundary between any two regions may be ambiguous when the description length at that level is almost invariant. Such cases can be fairly frequent, especially on noisy images. This ambiguity can be resolved at the next level, if one bears in mind the description length of the boundaries of the regions. Thus, in the given approach, several alternative hypotheses should also be transferred to the subsequent levels.

At the subsequent level, the description of the contours by structural elements is inaccurate; i.e., the contours that have been found deviate from the ideal sections and segments. Here two methods of describing these deviations are possible. First, the deviations themselves can be described; i.e., the deviation of each contour point from the points given by the structural elements is described, and this is expressed by the term $L(\partial G_i|_{s_j^{(i)}})$. Second, the contour itself can be corrected. In this case, the boundaries of the regions change, and hence the term $L(f|_{G_i})$ changes. If, because of a smooth transition or large noise, the boundary of the region is determined inaccurately, the displacement of this boundary has no great effect on the description length of the contents of the regions but can strongly influence the description length of the boundaries.

Feedback proper corresponds to correction of the boundaries of the regions. The structural elements specify the expectation of the position of the contours, and a contour can be attracted to this expectation, provided that the total description length decreases, i.e., if the increase of the description length of the contents of the region is compensated by a reduction of the description length of the contours.

IV. EXPERIMENTAL VERIFICATION

Let us check whether simultaneously taking into account both terms $L(f|_{G_i})$ and $L(\partial G_i)$ in Eq. (5), implemented by introducing feedback from the structural-element level to the contour level, actually makes it possible to obtain an adequate result. We shall use $2\|\partial G_i\|$ as an estimate of $L_0(f|_{G_i})$, where $\|\partial G_i\|$ is the number of points in a contour (to determine the position of each point on the contour, it is required to know the direction from the preceding point, which can be coded by two bits for a four-connected contour.

Figures 1a and 1c show two synthesized images, for each of which the regions shown in Fig. 1b are found by optimizing the criterion of Eq. (7). The separation shown in Fig. 1d would be legitimate for image (a), but the resulting boundary is displaced relative to the true boundary because of the local brightness fluctuations. Let us denote images (a) and (c) as f_1 and f_2, respectively. Tables I and II show the

TABLE I. Description length of contained image f_1, compressed into regions $G_1^{(1)}$, $G_2^{(1)}$, $G_1^{(2)}$, and $G_2^{(2)}$.

Version of separation into regions	G_1		G_2		$L\left(f_1\|_{G_1}\right)+L\left(f_1\|_{G_2}\right)$
	Entropy	$L\left(f_1\|_{G_1}\right)$, bit	Entropy	$L\left(f_1\|_{G_2}\right)$, bit	
$G^{(1)}$	6.362	129826	6.285	123140	252966
$G^{(2)}$	6.361	127219	6.295	125906	253125

TABLE II. Description length of contained image f_2, compressed into regions $G_1^{(1)}$, $G_2^{(1)}$, $G_1^{(2)}$, and $G_2^{(2)}$.

Version of separation into regions	G_1		G_2		$L\left(f_2\|_{G_1}\right)+L\left(f_2\|_{G_2}\right)$
	Entropy	$L\left(f_2\|_{G_1}\right)$, bit	Entropy	$L\left(f_2\|_{G_2}\right)$, bit	
$G^{(1)}$	6.328	129142	6.236	122178	251320
$G^{(2)}$	6.343	126859	6.313	126242	253101

description lengths of images f_1 and f_2 as functions of the regions that were distinguished.

The difference between the estimates of the description lengths of the boundaries in the two cases equals about 72 bits. At the same time, when more accurate results are obtained as a result of constructing structural elements, the difference of the description lengths $L(\partial G^{(1)})-L(\partial G^{(2)})$ increases to 232 bits. Using the data shown in the tables, we get

$$\left[L\left(f_1|_{G_1^{(1)}}\right)+L\left(f_1|_{G_2^{(1)}}\right)+L_0\left(\partial G^{(1)}\right)\right]$$
$$-\left[L\left(f_1|_{G_1^{(2)}}\right)+L\left(f_1|_{G_2^{(2)}}\right)+L_0\left(\partial G^{(2)}\right)\right]=-97 \text{ bit},$$
$$\left[L\left(f_2|_{G_1^{(1)}}\right)+L\left(f_2|_{G_2^{(1)}}\right)+L_0\left(\partial G^{(1)}\right)\right]$$
$$-\left[L\left(f_2|_{G_1^{(2)}}\right)+L\left(f_2|_{G_2^{(2)}}\right)+L_0\left(\partial G^{(2)}\right)\right]=-1709 \text{ bit},$$
$$\left[L\left(f_1|_{G_1^{(1)}}\right)+L\left(f_1|_{G_2^{(1)}}\right)+L\left(\partial G^{(1)}\right)\right]$$
$$-\left[L\left(f_1|_{G_1^{(2)}}\right)+L\left(f_1|_{G_2^{(2)}}\right)+L\left(\partial G^{(2)}\right)\right]=63 \text{ bit},$$
$$\left[L\left(f_2|_{G_1^{(1)}}\right)+L\left(f_2|_{G_2^{(1)}}\right)+L\left(\partial G^{(1)}\right)\right]$$
$$-\left[L\left(f_2|_{G_1^{(2)}}\right)+L\left(f_2|_{G_2^{(2)}}\right)+L\left(\partial G^{(2)}\right)\right]=-1549 \text{ bit}.$$

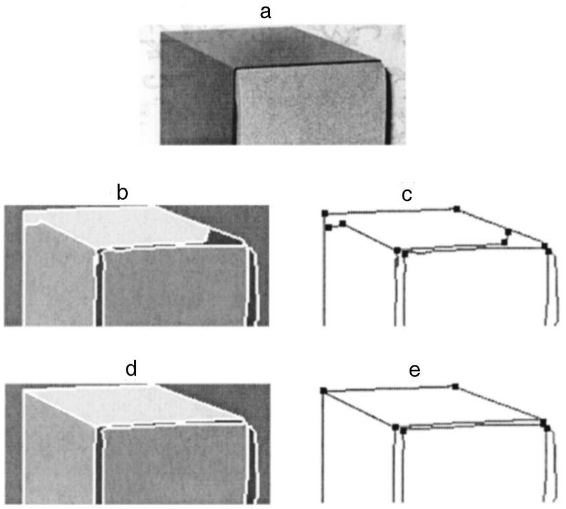

FIG. 2. Separating the regions and structural elements on an image. (a) Fragment of an initial image, (b) result of separating out regions by minimizing the target function of Eq. (7), (c) result of constructing structural elements on the basis of the boundaries of the regions of (b), (d) result of improving the regions of (b) by feedback from the structural-element level, (e) structural elements constructed as a result of iterative optimization of the quality criterion of Eq. (5) along with the regions of (d).

Thus, a more accurate estimate of the description length of the boundaries, obtained as a result of constructing the structural elements, makes it possible to choose separation (d) for f_1 while maintaining the choice of separation (b) for f_2.

It should be pointed out here that, when an image is separated into regions, the structural elements are not yet constructed, and therefore the separation shown in Fig. 1b is constructed as the initial hypothesis in both cases. The resulting contours are ordinarily used for further analysis without returning to the original intensities of the pixels. However, in our example, the contours were constructed to be identical for both images f_1 and f_2; i.e., these images would then be interpreted as equivalent at the contour level. The advantage of feedback can also be seen when working with actual images (see Fig. 2).

The essence of feedback is that, after the structural elements were constructed, new hypotheses were advanced concerning the position of the contours, and, to determine the quality of these hypotheses, it is necessary to return to the preceding level to obtain estimates of $L(f|_{G_i})$. In other words, the contours do not contain all the information needed for robust construction of the structural elements. This conclusion can also be extended to the subsequent levels of visual perception. Moreover, as the levels increase in hierarchical systems in the absence of feedback, the nonoptimality of the decisions made at the lower levels increases at the subsequent levels, and this can produce a catastrophic falloff of the efficiency of the system as a whole.

CONCLUSION

This paper has discussed a theoretical-informational approach to the introduction of feedback in machine-vision systems. This approach is more rigorous than that based on adaptive resonance theory and makes it possible to determine the optimum form of the feedback for a given representation of the information. This approach is also preferable in practice to the Bayesian approach, since it does not require explicit specification of the *a priori* probabilities of the hypotheses (the form of which, as a rule, is not known) and since it makes it possible to utilize more complex descriptions of the images, which are hard to use in the Bayesian approach because it is impossible to specify the corresponding stochastic models.

This investigation shows that the given approach makes it possible to enhance the robustness of the methods for constructing hierarchical structural descriptions of the images, with the problem of introducing feedback becoming central when the degree of *a priori* indeterminacy of the contents of the images is increased.

To develop the given approach further, it is necessary to develop specific mechanisms of iterative optimization of a global quality criterion that connects the individual levels of the visual system.

[1] A. Rares, M. J. T. Reinders, and E. A. Hendriks, “Image Interpretation Systems,” Technical Report (MCCWS 2.1.1.3.C), TU Delft, 1999.

[2] A. Gove, S. Grossberg, and E. Mingolla, “Brightness perception, illusory contours, and corticogeniculate feedback,” Visual Neurosci. **12**, 1027 (1995).

[3] C.-E. Liedtke, O. Grau, and S. Growe, “Use of explicit knowledge for the reconstruction of 3-D object geometry,” in *International Conference on Computer Analysis of Images and Patterns, 1995*, pp. 580–587.

[4] R. M. Kozlov, *Adaptation and Training in Robotics* (Nauka, Moscow, 1990).

[5] A. S. Potapov, *Pattern Recognition and Machine Perception. General Approach Based on the Principle of Minimum Description Length* (Politekhnika, St. Petersburg, 2007).

[6] S. Grossberg, “Adaptive pattern classification and universal recoding. I: Parallel development and coding of neural feature detectors,” Biol. Cybern. 23, 121 (1976) S. Grossberg, and “II: Feedback, expectation, olfaction, and illusions,” Biol. Cybern. 23, 187 (1976).

[7] S. Grossberg, “Adaptive resonance theory,” Technical Report CAS/CNS TR-2000-024, in *The Encyclopedia of Cognitive Science* (Macmillan, London, 2003).

[8] S. K. Kopparapu and U. B. Desai, “Bayesian Approach to Image Interpretation,” Springer International Series in Engineering and Computer Science, 2001, vol. 616.

[9] M. Li and P. Vitanyi, “Philosophical Issues in Kolmogorov complexity,” Proc. ICALP92 (invited lecture), 1992, pp. 1–15.