Principle of Representational Minimum Description Length in Image Analysis and Pattern Recognition[1]

A. S. Potapov

Vavilov State Optical Institute, 12 Birzhevaya line, St. Petersburg, 199034 Russia

e-mail: pas.aicv@gmail.com

Abstract—Problems of decision criteria in tasks of image analysis and pattern recognition are considered. Overlearning as a practical consequence of fundamental paradoxes in inductive inference is illustrated by examples. Theoretical (based on algorithmic complexity) and practical formulations of the minimum description length (MDL) principle are given. A decrease in the overlearning effect is shown on examples of modern recognition, grouping, and segmentation methods modified by the MDL principle. The representational MDL principle is introduced as an extension of the MDL principle, which makes it possible to take into account the dependence of the optimality criterion of the model from prior information given in data representation, as well as to perform optimization of representations. Novel possibilities of constructing learnable image analysis algorithms by optimizing the representation based on the extended MDL principle are described.

Keywords: image analysis, pattern recognition, overlearning, inductive inference, algorithmic complexity.

DOI: 10.1134/S1054661812010294

1. INTRODUCTION

Tasks of image analysis and pattern recognition are rather varied. At the same time, they possess certain similarities, since they contain induction as an essential part. Inductive inference consists of searching for regularities in observation data and in the construction of models of the data source. Methods of inductive inferences contain general components, such as model space, decision criterion, optimization, or search algorithm. Methods of image analysis and pattern recognition always include these components in explicit or implicit form.

Studies in the aforementioned fields frequently invent particular ad hoc optimality criteria and data description models based on heuristic considerations. The areas of application of these methods appear to be greatly restricted, due to the narrow model space (detectable regularities) and inaccuracy of optimality criteria. As a result, effects of overlearning arise in the field of pattern recognition and problems of constructing nontrivially learnable algorithms arise in the field of image analysis. These problems are inherent, not only to purely heuristic methods of recognition and analysis, but also to the application of seemingly correct mathematical approaches, such as the Bayes' rule for the most probable model selection.

In this work, the well-known principle of minimum description length (MDL) is described, which usage helps to introduce correct decision criterion for solving overlearning problem. A novel representational MDL (RMDL) principle has been introduced as an extension of the MDL principle that makes it possible to take into account the dependence of the model optimality criterion on the prior information given in data representation. Moreover, the RMDL principle makes it possible to construct image analysis methods with strong learnability via the automatic optimization of representations.

2. BAYES' CRITERION

One of the most widely used mathematical criteria in inductive inference is based on the following Bayes' rule:

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

where $P(H|D)$ is the posterior probability of model H with the given data D, $P(H)$ and $P(D)$ are prior probabilities, $P(D|H)$ is likelihood of data D with the given model H.

For example, the Bayes' rule can be directly applied to the classification problem. Let D be one pattern and H be one of the classes. If one has probability density distributions $P(D|H)$ of patterns within each class and unconditional probabilities $P(H)$, he can easily select the most probable class for the given pattern maximizing $P(H|D)$.

Learning in statistical pattern recognition consists of inducing probability distributions based on training

[1] The article was translated by the authors.

Received October 16, 2011

set $\{d_i, h_i\}$, where d_i is the ith pattern and h_i is its class. Here, prior probabilities $P(H)$ can be estimated from the frequencies of each class in the training set. Distribution $P(D|H)$ should be represented as an element of some family $P(D|H, \mathbf{w})$, where $\mathbf{w}$ is an indicator (e.g. parameter vector) of specific distribution. Using Bayes' rule and supposing independence of patterns one can obtain the following:

$$P(\mathbf{w}|D) = \frac{P(\mathbf{w})\prod_i P(d_i|h_i, \mathbf{w})}{P(D)}. \qquad (2)$$

Values of $P(d_i|h_i, \mathbf{w})$ can be explicitly calculated for the specific distribution defined with $\mathbf{w}$. However, there is a problem with evaluation of prior probabilities $P(\mathbf{w})$. In order to correctly specify these probabilities, one must have a large number of training sets and the true probability for each of them should be known. This is impossible because these true probabilities are unknown, even for human experts constructing a training set.

Many researchers prefer to ignore prior probabilities and to use the maximum likelihood (ML) approach. The same result will be obtained if one supposes that prior probabilities are equal. This supposition is evidently incorrect because, in the case of infinite model spaces, prior distributions become nonnormalized. In practice, this leads to the overlearning problem. For example, consider mixture Gaussian models. The likelihood of data will be maximized for the maximum number of components (equal to the number of training patterns) in a mixture leading to degenerate distribution. This is the so-called overlearning (or overfitting) effect.

This effect also appears in the task of regression. For example, if one tries to find a polynomial that fits the given points with minimum error (maximum likelihood), he will obtain a polynomial with the maximum degree that follows all errors in the data and possesses no generalization or extrapolation capabilities. An oversegmentation effect of the same origin appears in various image segmentation tasks [1]; models with more segments will be more precise.

As has been pointed out in studies [2, 3], the problem of prior probabilities is fundamental. This is connected with some paradoxes in inductive inference, such as Goodman's Grue emerald paradox (grue emeralds are green before some date and blue after it). The paradox lies in the fact that the observational data show the same evidence for emeralds to be green or grue.

Many criteria with an heuristically introduced penalty exist for model complexity. Still, new criteria are being invented for particular tasks of information processing. This is surprising because correct general criterion was proposed 50 years ago and it is well known; however, its importance is still underestimated.

3. ALGORITHMIC PROBABILITY

Consider the following notion of (prefix) algorithmic complexity of a binary string β introduced by A. N. Kolmogorov:

$$K_U(\beta) = \min_{\alpha}[l(\alpha)|U(\alpha) = \beta], \qquad (3)$$

where U is a universal Turing machine (UTM), α is an arbitrary algorithm (program for UTM), and $l(\alpha)$ is its length. In accordance with this notion, the amount of information contained in the data (string) equals the length of the shortest program that can produce this data.

Unlike the Shannon theory, this notion of information quantity relies not on probability, but on pure combinatorial assumptions. R. Solomonoff proposed to derive the probability from the algorithmic complexity and to use it in induction. Indeed, if it is possible to derive optimal codes from probabilities, one can invert this task and find probabilities from optimal codes.

The probability of a program α is connected with its length $l(\alpha)$ as follows:

$$P(\alpha) = 2^{-l(\alpha)}. \qquad (4)$$

Arbitrary string β can be generated by a number of programs α_i, so its algorithmic probability can be calculated using the equation

$$P(\beta) = \sum_i 2^{-l(\alpha_i)}. \qquad (5)$$

This is the so-called universal distribution.

The algorithmic probability can be called the theoretical information formalization of Occam's razor and solves the problem of prior probabilities [3]. Apparently, if some string has any regularity, it can be generated by some shorter (than string) program, so its probability is higher. If one supposes that there are no nonalgorithmic regularities, the set of algorithms will give universal model space. With a correct-decision criterion, this will yield a universal inductive inference method. Unfortunately, algorithmic probability is incalculable.

4. MINIMUM DESCRIPTION LENGTH PRINCIPLE

More practical schemes, such as the principles of Wallace's minimum message length (MML) and Rissanen's minimum description length (MDL) avoid the incalculability problem by considering restricted subsets of models. Li and Vitanyi's ideal MDL principle also utilizes computable models. In general, these principles can be obtained from algorithmic complexity if one divides program for UTM $\alpha = \mu\delta$ into the

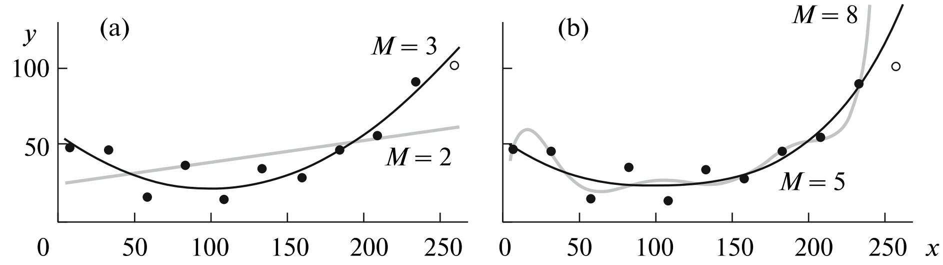

Fig. 1. Example of polynomial approximation.

algorithm itself (regular component of the model) μ and input data (random component) δ as follows:

$$\begin{aligned} K(\beta|\mu) &= \min[l(\delta)|U(\mu\delta) = \beta], \\ K(\beta) &= \min[l(\mu) + K(\beta|\mu)]. \end{aligned} \quad (6)$$

Consequently, the equation

$$\mu^* = \underset{\mu}{\operatorname{argmin}}[l(\mu) + K(\beta|\mu)] \quad (7)$$

gives the best model via minimization of model complexity $l(\mu)$ and model precision $K(\beta|\mu) = l(\delta)$, where δ describes deviations of data β from model μ. Equation (7) is similar to Bayes' rule if one assumes $l(\mu) = \log_2 P(\mu)$ and $K(\beta|\mu) = \log_2 P(\beta|\mu)$. Here, prior probabilities can be calculated.

If one considers the calculable complexity, the best model can be calculated for any string. However, calculability is not sufficient for practical use because inductive inference problem appears to be *NP*-hard in these settings.

Universal inductive inference method turned out to be practically impossible even with correct decision criterion, because of essential search problem. Nevertheless, the MDL principle (or its equivalents) in the following verbal form [2] appears to be very fruitful in analysis and pattern recognition. The best model of the given data source is the one that minimizes the sum of the length (in bits) of the model description and the length (in bits) of data encoded with the use of the model.

In order to apply this principle in practice, Heuristic coding schemes are introduced for each particular task. For example, when considering some parametric subfamily of algorithms, one must describe only the values of parameters that specify the concrete model. Searching within this subfamily can be easy, but only an a priori restricted set of regularities can be captured.

5. PRACTICAL USE OF THE MDL PRINCIPLE

Consider the task of polynomial approximation. Some polynomial can be interpreted as an algorithm that produces a string $y_1y_2...y_n$ from the given input string $x_1x_2...x_n$, where (x_i, y_i) are points to be fit (probably, with some residuals) by polynomial with *M* dimensional parameter vector **w**

$$f(x|\mathbf{w}) = \sum_{k=0}^{M-1} w_k x^k. \quad (8)$$

The model description contains information about parameters **w**. If the polynomial precisely fits the given data, the data description length will be zero. Otherwise, one must also describe deviations of the data from the model. The data encoded using the model include deviations $e_i = y_i - f(x_i|\mathbf{w})$, which description length can be estimated based on entropy $L_e = nH(\{e_i\})$ in assumption of their independence. If deviations have a Gaussian distribution, the entropy will be proportional to the logarithm of dispersion; i.e., the optimization of L_e will lead to the least-squares method (LSM), which is a type of ML method.

In order to avoid overfitting, one must take into account the description length of model parameters $L_\mathbf{w}$, which depends on the precision (number of bits per parameter) with which they are described. Each parameter can be rounded with a different precision (this will result in changes in both L_e and $L_\mathbf{w}$) in order to find the optimum. We have the following rough estimation for *M* parameters and *n* data elements [4]: $L_\mathbf{w} = 0.5M\log_2 n$. The resulting equation for the MDL criterion is as follows:

$$L(\mathbf{w}) = nH(\{y_i - f(x|\mathbf{w})\}) + \frac{M}{2}\log_2 n. \quad (9)$$

Figure 1 shows an example of fitting different polynomials to some set of noisy points.

One point was not included into the training set; it is used to check the extrapolation quality. Table 1 shows the average error in the point of training set, the average error in some wider interval, and the description length for polynomials of different degrees ($M - 1$).

It can be seen that the average error at points from the training set decreases with the degree of the polynomial; it does not correspond to true extrapolation precision. The MDL criterion helps to choose optimal model complexity. Of course, similar results can be achieved using the cross-validation technique. The latter, however, has several drawbacks, i.e., it does not give understanding of overlearning origin, the model is constructed using only the portion of data that results in a loss of precision, and cross-validation is difficult to apply in unsupervised learning or image analysis tasks.

The same results can be achieved in the task of pattern recognition, when this task is reduced to the approximation problem, and nonlinear (e.g. polynomial) discrimination functions may have different complexity [5]. Similarly, the MDL criterion helps to choose the optimal complexity of support vector models [6] and the number of components in mixture models [7]. There are also many applications of the MDL principle in image analysis tasks, such as texture segmentation, feature extraction, structural description, object recognition, spatial transformation estimation, optical flow estimation, change detection, and many others. Figures 2 and 3 shows some initial images and results of segmentation based on the MDL criterion (Fig. 3 also shows results with least squares criterion for comparison).

6. REPRESENTATIONAL MDL PRINCIPLE

As can be seen, the MDL principle helps to partially solve problems, such as overfitting (in the task of approximation), overlearning (in the task of pattern recognition), and oversegmentation (i.e., to automatically select the number of regions in the image), in practice. However, it can also be seen that coding schemes for estimating the description length are introduced heuristically in the MDL-based methods. Ungrounded coding schemes are non-optimal and non-adaptive (independent of the given data). These schemes define algorithmically incomplete model spaces that cause corresponding methods of image analysis and pattern recognition to be fundamentally restricted. Thus, there is a large gap between the theoretical MDL principle with universal model space and prior probability distribution and its practical applications.

In order to overcome this gap, tasks of inductive inference should be considered as mass problems. Indeed, image analysis and pattern recognition methods are usually executed independently for each image or pattern. It is easy to show that algorithmic complexity of concatenation of some data strings $\beta_1\beta_2...\beta_n$ is smaller than the sum of their individual algorithmic complexities as follows:

$$K_U(\beta_1\beta_2\ldots\beta_n) \leqslant \sum_{i=1}^{n} K_U(\beta_i). \tag{10}$$

Moreover, a universal prior probability distribution appears to be dependent on the choice of universal Turing machine (UTM), which can be considered an additional theoretical difficulty. Usually, this difficulty is assumed to be nonessential, since a constant string ν exists for any two UTMs, U and V, such that $(\forall\alpha)U(\nu\alpha) = V(\alpha)$. In other words, $(\forall\beta)K_U(\beta) \le K_V(\beta) + C$; i.e., the algorithmic complexities of any data on two different UTMs differ by a constant that does not depend on the given data. Consequently, the influence of the difference between UTMs will decrease with an increase in the volume of data and equivalent models will be selected.

Table 1. Results of polynomial approximation

M	$\bar{e}$	$\bar{e}$ [−100, 350]	L
1	20.8	64.5	45.5
2	18.0	62.8	45.1
3	8.4	6.0	35.6
4	8.1	27.0	36.9
5	8.0	70.9	38.2
6	7.6	326.8	39.3
7	7.5	590.9	40.8
8	6.6	8332	40.6
9	6.0	34912	40.9

However, in practice, constant C may be very large. Moreover, difference in algorithmic complexities will be unbounded in mass problems of automatic data processing, because only the following inequality will be held

$$\sum_{i=1}^{n} K_U(\beta_i) \le \sum_{i=1}^{n} K_V(\beta_i) + nC. \tag{11}$$

It can now be seen why heuristic coding schemes are used in practical applications of the MDL principle in the tasks of image analysis and pattern recognition, rather than a universal model space defined by some UTM. Not only does the search for an algorithmically complete space leads to computational problems, but the selection of a specific coding scheme exerts a strong influence on the model-quality crite-

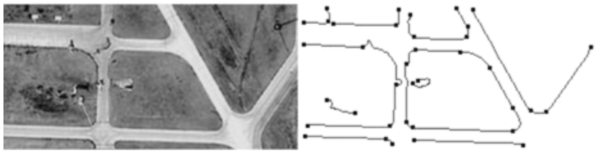

Fig. 2. Example of contour segmentation.

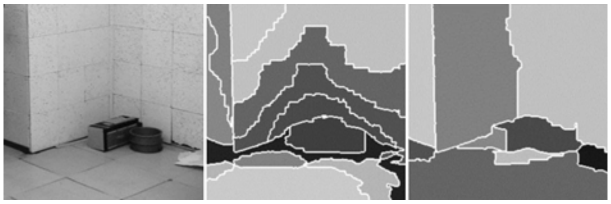

Fig. 3. Example of image segmentation.

rion and, consequently, on the efficiency of the corresponding method.

These difficulties are the most crucial for image analysis and pattern recognition as mass problems; therefore, the difference between left and right parts of Eq. (10) and Eq. (11) should be minimized. The sum of algorithmic complexities of data strings (sum of lengths of their independent descriptions) is much larger than the algorithmic complexity of their concatenation (length of their joint description) because these sets contain mutual information. This mutual information should be removed from descriptions of individual data strings, and should be considered as prior information in corresponding methods. This implies the use of conditional algorithmic complexity. Indeed,

$$K_U(\beta_1\beta_2\ldots\beta_n) \approx \min_S\left(\sum_{i=1}^{n} K_U(\beta_i|S) + l(S)\right), \quad (12)$$

where conditional algorithmic complexity can be calculated as $K_U(\beta_i|S) = \min_{\mu}(l(\mu)|U(S\mu) = \beta_i)$, S is some string, and $l(S)$ is its length [8].

It can be shown that $(\forall U, V, S)(\exists S')(\forall\beta)K_U(\beta|S') = K_V(\beta|S)$. Let $S' = \nu S$, where ν is an interpreter of V on U, then $(\forall\alpha)U(S'\alpha) = U(\nu S\alpha) = V(S\alpha)$. Consequently,

$$\sum_{i=1}^{n} K_U(\beta_i|S') = \sum_{i=1}^{n} K_V(\beta_i|S). \quad (13)$$

Since S can be interpreted as an algorithm (some program for UTM) that produces any given data string from its description, algorithm S precisely fits the verbal notion of representation formulated by David Marr [9]. Therefore, the following string definition can be given.

Definition. Program S for UTM U will be called the "representation" of the collection of data strings (images, patterns) $\mathbf{B} = \{\beta_1, \ldots, \beta_n\}$, if $(\forall\beta \in \mathrm{B})(\exists\mu, \delta \in \{0, 1\}^*)U(S\mu\delta) = \beta$. String β will be called the "description" of μ within representation S. This description consists of regular μ and random δ components.

If data description is carried out within some representation, then the mentioned above difficulties will be eliminated. In particular, because of Eq. (13), the choice of the UTM will not influence the selection of the model for specific data string and we will omit the indication of the specific UTM and write $K_S(\beta)$ instead of $K_U(\beta|S)$. It should be noted that some representation S usually specifies the algorithmically incomplete model space and complexity $K_S(\beta)$ turns out to be calculable in practice (in contrast to complexity $K_U(\beta)$).

The formal notion of representation can be used to extend the MDL principle on mass problems giving representational MDL principle that consists of two parts [8].

1. The best model μ of data β *within given representation S* is the model, for which the sum of the following components is minimized:

—the length of the model $l(\mu)$;

—the length of data described with the use of model $K_S(\beta|\mu)$.

Selection criterion and the best model can be calculated as

$$\begin{aligned} L_S(\beta, \mu) &= K_S(\beta|\mu) + l(\mu) \\ \text{and } \mu^* &= \underset{\mu}{\arg\min} L_S(\beta, \mu). \end{aligned} \quad (14)$$

2. The best representation S for the collection of data strings $\mathrm{B} = \{\beta_1, \ldots, \beta_n\}$ is the representation for which the sum of the following components is minimized:

—the length of representation $l(S)$;

—the sum of lengths of data strings described within the representation $\sum_{i=1}^{n} K(\beta_i|S)$.

Selection criterion and the best representation can be calculated as

$$\begin{aligned} L(\mathrm{B}, S) &= l(S) + \sum_{i=1}^{n} K_S(\beta_i) \\ \text{and } S^* &= \underset{S}{\arg\min} L(\mathrm{B}, S). \end{aligned} \quad (15)$$

The RMDL principle specifies dependence of model quality criterion from used representation (description language), and also gives a criterion for optimizing the representation itself. Thus, theoretical grounds for optimizing the representation depending on the problem domain are obtained instead of the heuristic selection of coding schemes that occurs during the practical implementation of the MDL criterion.

Of course, the RMDL principle does not give a complete solution of the problem of automatic representation optimization; rather, it only gives a criterion for their comparison that can only be practically used with an efficient representation search (or generation) procedures. Nevertheless, this principle can be used for an objective comparison of hand-crafted representations (question of optimality and bounds of applicability of heuristic coding schemes was not even stated), as well as for the automatic optimization of representations within their simple families.

7. SYNTHETIC METHODS OF PATTERN RECOGNITION

The RMDL principle shows that the existing pattern recognition methods use particular representations that specify algorithmically incomplete model spaces. Universal recognition systems built using this approach are doomed to fail. Furthermore, proofs of the universality of artificial neural networks based on the fact that they can be used for arbitrarily precise functional approximations are incorrect. For example, the polynomial approximation of exponential function results in the classical effect of overfitting because, in this representation, the correct model will have infinite complexity (and its reconstruction will require an infinite training set).

It is natural that algorithmically complete spaces are not used in practical recognition methods, but it leads to a restricted (and varying for different methods) set of detectable regularities. Growing popularity of composition of classifiers (e.g. [10]) is not surprising because each classifier uses its own particular model space, and their composition extends the set of representable regularities. However, the use of different voting schemes cannot be considered optimal.

The RMDL principle helps to build a synthetic pattern-recognition systems [11], in which the choice of the best particular classifier is carried out based on the description length criterion. As an example, consider the extension of family of representations in the Gaussian mixture method. Gaussian mixture is represented in the form

$$p(\mathbf{x}|\mathbf{w}) = \sum_{i=1}^{d} P_i p(\mathbf{x}|\mathbf{C}_i, y_i), \qquad (16)$$

where $\mathbf{x}$ is the pattern (feature vector), P_i is the weight of ith component of the mixture, $p(\mathbf{x}|\mathbf{C}_i, \mathbf{y}_i)$ is the normal distribution with the covariance matrix $\mathbf{C}_i$ and mean vector $\mathbf{y}_i$, and $\mathbf{w}$ is the combined vector of parameters of the Gaussian mixture.

In this case, pattern recognition task is reduced to estimation of mixture parameters $\mathbf{w}$ based on training set $\mathrm{B} = \{\mathbf{x}_i\}_{i=1}^{n}$. As it was noted, the MDL principle helps to solve the problem of selection of number d of mixture components, but it does not show the possibility to extend this representation. The RMDL criterion will have the following form for this representation

$$L(\mathrm{B}, S\mathbf{w}) = l(S) + l(\mathbf{w}) - \sum_{i=1}^{n} \log_2 p(\mathbf{x}_i|\mathbf{w}). \qquad (17)$$

Selection between different representations (different forms of probability density function p) can be carried out based on this criterion.

Consider simpler representations as alternatives, which can appear to be more efficient. Gaussian mixtures will be specified by diagonal covariance matrix $C_{i,i} = \sigma_i^2$ in one representation, and unity matrix multiplied by dispersion $\mathbf{C} = \sigma^2\mathbf{I}$ in another representation. Seemingly, these two representations define models that comprise a subset defined by representation based on full Gaussian mixtures; thus, their inclusion should be useless. However, identical distributions in different representations correspond to different models in inductive inference because their complexity (number of parameters) is different.

Table 2. Description lengths (in bits) of training set (n = 24) within different mixture models

Type	$M = 1$	$M = 2$	$M = 3$	$M = 4$
C	834	855	855	–
$C_{i,j} = \sigma_i^2$	856	838	817	826
$C = \sigma^2 I$	859	857	826	823

Consider the example in Fig. 4.

Here, a five-dimensional space of features is used (the subset for two features is shown in the figure), and the patterns are distributed in three clusters (these clusters are separable in the given five-dimensional space). The results of clustering using three different types of representations and different numbers of components in mixtures are shown.

The corresponding description lengths (in bits) are given in Table 2. As can be seen from the table, the minimum description length in different representations is achieved for different numbers of components in the mixtures. The reason is that Gaussian distribution is defined by 20 parameters in five-dimensional space, and there is not enough information in the used training set even to estimate each parameter of mixture with three components. At the same time, the simplest representation appears also to be less efficient despite models have smaller number of parameters within it.

As it can be seen, in order to select correct number of clusters one needs to take different representations into account and to make choice between them. Even small difference between representations can be intrinsic. Of course, automatic selection (based on the RMDL criterion) between more diverse representations will significantly increase capabilities of pattern recognition and clustering methods.

8. COMPARISON OF IMAGE REPRESENTATIONS USING THE RMDL CRITERION

The development of the RMDL principle was primarily motivated by the problems of image analysis. Indeed, the notion of representation is one of the most crucial elements in image analysis methods. The construction of a model (description) of an image is always carried out by some representation, and there are very different classes of image representations. Particular criteria for optimal model selection are usually deduced for each representation, but the efficiency of representations themselves is rarely investigated. In this context, the RMDL principle, which shows an explicit connection between the selection of model

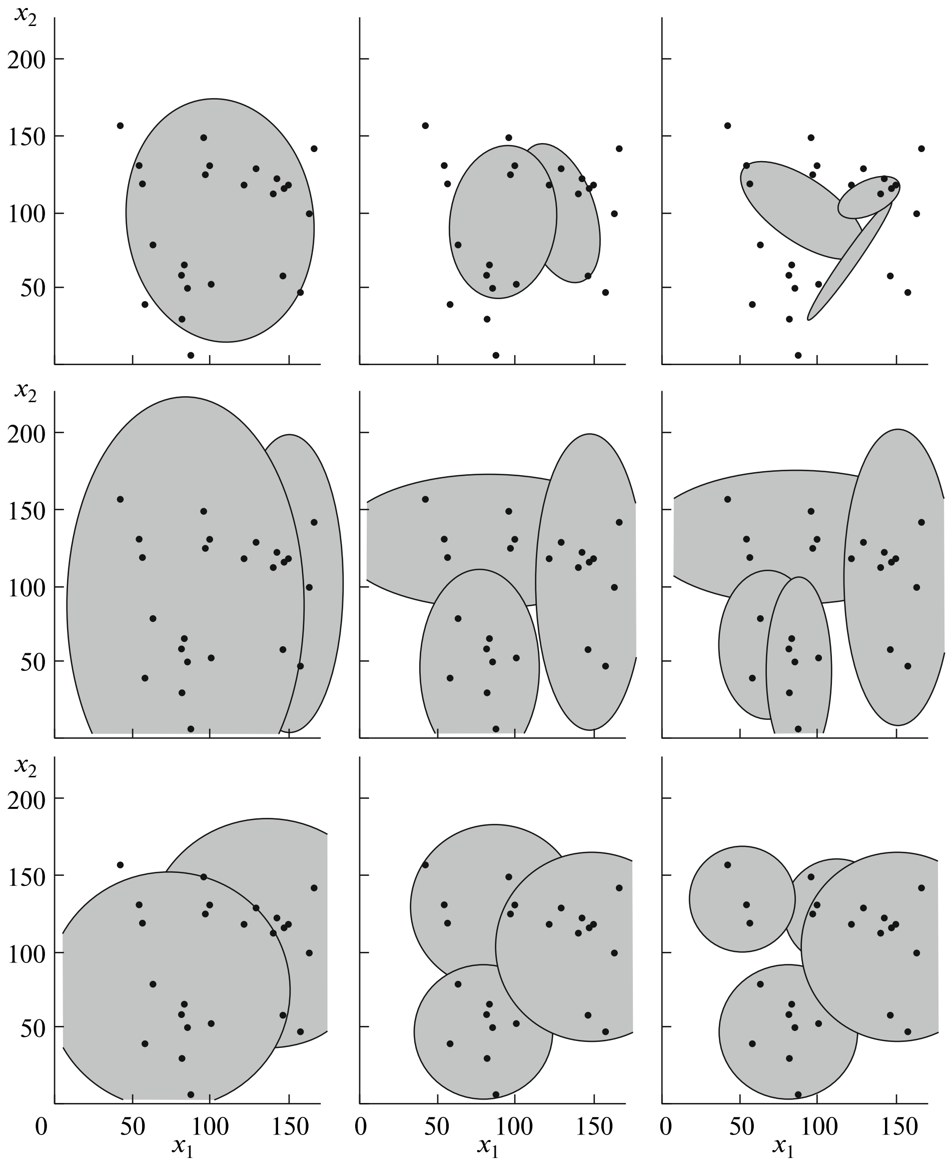

Fig. 4. Mixtures with different restrictions on covariance matrix and different numbers of components.

criteria and data representations and also allows defining optimality criterion of representations, can become very useful element of image analysis methodology.

Examples of a comparison of the efficiency of image representations in segmentation algorithms and representations of contours (borders of regions) in algorithms of structural analysis are given in the work [12]. The simplest representation, which reflects some image properties, is the representation of an image by an array of brightness values as independent outcomes of some random variable. In the simplest representation $S_0^{(1)}$, the description length of an image $f(x, y)$: $G \longrightarrow R$ can be estimated as follows:

$$L_{S_0}(f) = \|G\|H(f) + N_{\text{int}}\log_2 N_{\text{int}},$$
$$H(f) = -\sum_{f=0}^{N_{\text{int}}-1} p(f)\log_2 p(f), \tag{18}$$

where $\|G\|$ is the area of the region G (number of pixels in the image f), $H(f)$ is entropy of brightness values in supposition of their statistical independence, and N_{int} is the number of different brightness levels. The first summand is the description length of pixel brightness values encoded with Huffman code. In order to decode it, one must use a code (frequency) table, which should also be stored in the image description; its length can be roughly estimated as N_{int} bits.

Some image is divided into regions in the task of segmentation. In other words, the image is represented as a set of regions, each of which is described independently. The simplest method consists of describing pixel brightness in regions independently of the outcomes of some random variable; however, probability distributions will be unique for each region. The image is assumed to be divided into the set of regions $G_1, ..., G_d$, in which the representation $S_1^{(1)}$. Here, borders of the region G_i should also be described in addition to pixel brightness. That is, the image description length within this representation can be estimated as

$$L_{S_1^{(1)}}(f) = \sum_i (\|G_i\| H(f_i) + N_{\text{int}} \log_2 N_{\text{int}} + \|\delta G_i\| \log_2 N_{\text{dir}}), \tag{19}$$

where $f_i(x, y) = f(x, y)|_{G_i}$ is the image f in the region G_i with border length $\|\delta G_i\|$, N_{dir} is the number of possible directions from the current border point to the next point (e.g., $N_{\text{dir}} = 8$).

Image segmentation based on the criterion (19) leads to detection of regions with minimal entropy. However, the increase in the number of regions is also penalized, i.e., a compromise between the number of regions and their entropy is sought.

The further complication of the representation supposes the description of the dependence between the brightness of pixels inside of regions, which can be represented in general form as an approximation of the image in each region by some family of functions.

Let $g_i(x, y, \mathbf{w}_i)$ be some function defined by parameters $\mathbf{w}_i$ that approximate the image in the region G_i, and $r_i(x, y) = [f_i(x, y) - g_i(x, y, \mathbf{w}_i)]$ be residuals of approximation. The description length in representation $S_2^{(1)}$ can be estimated as

$$L_{S_2^{(1)}}(f) = \sum_{i=1}^{d} (\|G_i\| H(r_i) + N_{\text{int}} \log_2 N_{\text{int}} + \|\delta G_i\| \log_2 N_{\text{dir}} + l(\mathbf{w}_i)),$$

where $l(\mathbf{w}_i) = \frac{m_i}{2} \log_2 \|G_i\|$ is the length of description of parameters $\mathbf{w}_i$.

Quadratic functions for the representation $S_2^{(1)}$ are considered as functions $g_i(x, y, \mathbf{w}_i)$ in paper [12]. The representation $S_3^{(1)}$ is also considered, in which approximation is performed using a system of Gabor functions.

Table 3 shows the results of a comparison of the description lengths within different representations for three sets, i.e., F_1, F_2, and F_3,which contain aerospace photographs, SAR images, and indoor images, correspondingly. It should be pointed out that ratio of average description lengths is shown here without taking complexity of representations themselves into account. Difference in the latter complexity can be tens kilobytes that will result in slight changes in total value of the RMDL criterion (and the ratio $L(F_2, S_2^{(1)})/L(F_2, S_1^{(1)})$ will be larger than 1).

Table 3. Comparison of quality of representations $S_n^{(1)}$

	Image set		
Description length ratio	F_1	F_2	F_3
$\overline{L_{S_1^{(1)}}(f)}/\overline{L_{S_0^{(1)}}(f)}$	0.837	0.968	0.713
$\overline{L_{S_2^{(1)}}(f)}/\overline{L_{S_1^{(1)}}(f)}$	0.985	0.999	0.921
$\overline{L_{S_3^{(1)}}(f)}/\overline{L_{S_1^{(1)}}(f)}$	0.946	0.988	0.996

It can be seen that the efficiency of different representations varies in different image samples. However, the separation of images into regions appears to be efficient in terms of the RMDL criterion on all samples that can be used to objectively ground representations of this type; however, its efficiency is not guaranteed for image samples from different problem domains. The higher efficiency of the representation using of the polynomial approximation of brightness distribution in regions for indoor image samples is rather interesting. This result corresponds with the presence of smooth changes in brightness on these images. At the same time, the approximation with Gabor functions appears to be more efficient for aerospace images of landscapes with natural textures.

Borders of the region obtained as a result of segmentation or contours detected by local operators are commonly used as the basis for image representations at the next level of abstraction. The description of contours in the form of their structural elements inevitably leads to the problem of selecting an alphabet for these elements. Typical structural elements are line segments and arcs of circles and ellipses. One can consider a contour representation in the form of chain code ($S_0^{(2)}$) without dividing the contours into structural elements, as well as representations that include approximation contour segments by line segments only ($S_1^{(2)}$), by line segments and arcs of second order curves ($S_2^{(2)}$), and by curves of the first, second, and third order ($S_3^{(2)}$). The description length criterion can be estimated for each of these representations of contours. Table 4 shows estimations of average description lengths of contours detected on images of the same samples.

Table 4. Comparison of quality of representations $S_n^{(2)}$

	Image set		
Description length ratio	F_1	F_2	F_3
$\overline{L_{S_1^{(2)}}(\delta G)}/\overline{L_{S_0^{(2)}}(\delta G)}$	0.809	0.812	0.679
$\overline{L_{S_2^{(2)}}(\delta G)}/\overline{L_{S_1^{(2)}}(\delta G)}$	0.831	0.845	0.791
$\overline{L_{S_3^{(2)}}(\delta G)}/\overline{L_{S_2^{(2)}}(\delta G)}$	1.007	1.007	1.006

The efficiency of representations differs, but here, in contrast to the previous example, the selection of the most efficient representation does not depend on the sample. This is the representation that includes line segments and arcs of circles and ellipses. The additional inclusion of third-order curves into the alphabet of structural elements does not increase the efficiency of the representation. Of course, this conclusion cannot be extended to images of other problem domains without corresponding experimental validation.

The selection of image representations based on the RMDL criterion can be performed automatically, which greatly increases the adaptive capabilities of image analysis methods while solving problems with prior uncertainty. For example, this selection can be performed in methods of the structural matching of arbitrary images. The possibility of optimizing feature-based representations is theoretically and empirically grounded in [13] for image analysis systems that function in the specific visual environment, e.g., for the case when a serial model of a robot is used in a priori unknown apartments. The optimization of representation should be considered the essential or “strong” learning of image-analysis systems, whereas the accumulation of information in existing representations is superficial or “weak” learning.

Thus, the optimal selection of the model within some representation is the main element in tasks of perception and optimal representation selection is the main element both types of problems, which makes this principle a powerful tool in image analysis and pattern recognition.

CONCLUSIONS

One of the general problems in inductive inference consists of specifying a decision criterion, without which correct data interpretation is impossible. Even widely used Bayes’ criterion frequently causes an a priori probability problem, which (when ignored) results in overlearning effect.

Algorithmic complexity provides a reliable theoretical basis for solving these problems. However, universal prior distribution over algorithmically complete model space does not allow practically realizable search. Instead, restricted families of models and heuristic coding schemes are applied, which result in a practical MDL principle, the use of which has shown many positive results in solving tasks of pattern recognition and image analysis. However, these results brought limited progress in these fields because the fundamental problems remain unsolved.

One of the current directions of research consists in filling the gap between theoretical and practical versions of the MDL principle. In order to do this, the notion of representation should be incorporated into this principle, which leads to the representational MDL principle, based on which model selection criteria are constructed that explicitly depend on the given representation and criteria for the automatic optimization of the representation are obtained, rather than using static heuristic coding schemes. In particular, the application of this approach yields essentially learnable image analysis algorithms.

The primary reason for the non-universality of existent methods of image analysis and pattern recognition consists of using algorithmically incomplete solution spaces caused by intractable search problem.

ACKNOWLEDGMENTS

Work is carried out with support of Ministry of Education and Science and Russian Federation President’s grant Council (MD-2040.2010.9).

REFERENCES

1. T. Lee, “A Minimum Description Length Based Image Segmentation Procedure, and Its Comparison with a Cross–Validation Based Segmentation Procedure,” J. Am. Stat. Assoc. **95**, 259–270 (2000).
2. M. Li and P. Vitanyi, “Philosophical Issues in Kolmogorov Complexity,” in *Proc. ICALP’92* (Vienna, 1992), pp. 1–15.
3. R. Solomonoff, “Does Algorithmic Probability Solve the Problem of Induction?,” in *Proc. Conf. on Information, Statistics and Induction in Science (ISIS)* (World Sci., Melbourne, 1996).
4. J. J. Rissanen, “Modeling by the Shortest Data Description,” Automatica **14**, 465–471 (1978).
5. M. Sato, M. Kudo, J. Toyama, and M. Shimbo, “Construction of a Nonlinear Discrimination Function Based on the MDL Criterion,” in *Proc. 1st Int. Workshop on Statistical Techniques in Pattern Recognition* (Prague, 1997), pp. 141–146.
6. U. von Luxburg, O. Bousquet, and B. Schölkopf, “A Compression Approach to Support Vector Model Selection,” Mach. Learn. Res. **5**, 293–323 (2004).

7. H. Tenmoto, M. Kudo, and M. Shimbo, “MDL—Based Selection of the Number of Components in Mixture Models for Pattern Classification,” in *Lecture Notes Computer Science* (Springer-Verlag, London, 1998), Vol. 1451, pp. 831–836.

8. A. S. Potapov, “How to Choose Image Presentation on the Base of Minimization of Representation Length of Image Description,” Izv. Vyssh. Uchebn. Zaved. Priborostroen. **51** (7), 3–7 (2008).

9. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, 1982; Radio i svyaz’, Moscow, 1987).

10. D. Ruta and B. Gabrys, “Classifier Selection for Majority Voting,” Inf. Fusion **6** (1), 63–81 (2005).

11. A. S. Potapov, “Synthetic Pattern Recognition Methods Based on the Representational Minimum Description Length Principle,” in *Proc. 2nd Int. Topical Meeting on Optical Sensing and Artificial Vision OSAV’2008* (St. Petersburg, 2008), pp. 354–362.

12. A. S. Potapov, “Comparative Analysis of Structural Images Presentation by Using the Representation Principle of Minimal Description Length,” Opt. Zh. **75** (11), 35–41 (2008).

13. A. S. Potapov, I. A. Malyshev, A. E. Puysha, and A. N. Averkin, “New Paradigm of Learnable Computer Vision Algorithms Based on the Representational MDL Principle,” Proc. SPIE 7**696**, 769606(2010).

Alexey Potapov graduated from the Department of Mathematics and Mechanics of the St.Petersburg State University, Russia, in 2002. In 2005, he received the Ph.D. degree for the thesis in the field of automatic image analysis at the Vavilov State Optical Institute, St.Petersburg, Russia, where he is currently Head of research laboratory. From 2006 till now, he is also with National Research University of Information Technology, Mechanics, and Optics, St. Petersburg, Russia, where he received the Dr. Sc. degree in 2008, and currently he is Professor in the Department of Computer Photonics and Videoinformatics. He has more than 70 papers in the fields of image analysis, pattern recognition, and machine learning, including the monograph titled “Pattern recognition and machine perception: general approach on the base of the minimum description length principle” (in Russian).