SEPARATION OF HOMOGENEITY FIELDS IN SPACE-MADE PHOTOGRAPHS BY A NONPARAMETRIC SEGMENTATION ALGORITHM IN SPACES OF INFORMATIVE FEATURES

K.T. Protasov

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences, Tomsk
Received April 16, 1998

A new combined nonparametric algorithm based on a four-step procedure is developed for segmentation of multi-spectral space-made photographs of the Earth's underlying surface and cloudiness. At the first step, fragment-by-fragment local clustering of video data is performed by use of the Bhattacharyya distance or Kullback divergence. At the second step, adjacent obtained classes are unified by use of the empirical risk functional. At the third step, enlarged classes serve as a learning material for a nonparametric algorithm of pattern recognition. Finally, at the fourth step, the pattern recognition algorithm performs segmentation of the whole image. This approach makes it possible to solve the problem of compromise between awkwardness of the initial data and necessity to use adequate models of images under recognition based on nonparametric estimates of unknown conditional probability distributions. Besides, the problem of studying complexes of features for information content is being solved in the sense of the minimum of the empirical risk criterion.

INTRODUCTION

Multichannel space-made photographs of the Earth's underlying surface and cloudiness are main sources of fast information in solving problems of use of nature resources, climate-ecological monitoring, and evaluation of states of natural complexes. Since recording of images is performed under conditions of broken cloudiness in the overwhelming majority of cases, there arises a problem of automatic revealing of cloudiness fields by use of algorithms of video data segmentation.[1–4] The presence of cloudiness is a disturbing factor, so these parts of an image must be highlighted before solving the problem of classification of textural homogeneity of video data concerning the Earth's underlying surface.

The algorithm of automatic classification proposed below has good local properties and can work with large arrays of video data. It is a four-step procedure. At the first step, cluster analysis of small fragments of multispectral images is performed. The analysis is based on seeking for modes of mixing distributions followed by enlargement of classes by use of the Bhattacharyya's distances or the Kullback's information criterion. At the second step, the obtained classes of all fragments are united into larger blocks by use of empirical risk. At the third step, the algorithm of pattern recognition learns to distinguish classes obtained at the second step of data aggregation, and complexes of features are studied for information content. Finally, at the fourth step, the learned decision rule recognizes components of the whole image. Since a small amount of fragments statistically equivalent to the whole image is sufficient to serve as a learning material, this approach leads to sharp decrease of computation expenses with saving the accuracy characteristics of the decision rule.

The salient feature of the proposed algorithm is the fact that a nonparametric estimation of the risk functional or nonparametric estimations of the boundaries of this functional serve as a measure of closeness or distinguishability of distinguished classes, and probability models of classes are recovered by use of nonparametric approximations of unknown conditional probability density functions.

MATHEMATICAL FOUNDATIONS FOR SYNTHESIS OF PATTERN RECOGNITION ALGORITHMS

To begin, let us consider the problems connected with construction of automatic classification algorithms and present principle mathematical relations of synthesis of pattern recognition algorithms used below.[1–5,8]

Let the result of observation be a set of numbered fields of video data given in several spectral ranges so that every pixel of an image of the Earth's underlying surface and cloudiness recorded by the recording system

0235-6880/98/07 678-07 $02.00

is characterized by a random vector $\mathbf{X} = (X^1, \ldots, X^n)^{\mathrm{T}}$, where T is the transposition sign; $\mathbf{X} \in R^n \equiv \chi$, and R^n is the n-dimensional space of observations. The components X^i, $i = 1, \ldots, n$ of the observation vector $\mathbf{X}$ characterize the reflecting (radio brightness) properties of landscapes and cloudiness in every corresponding spectral range. We assume that the joint distribution of the vector $\mathbf{X}$ components in the space of observations can be presented as the following mixing probability density function:

$$f(\mathbf{x}) = \sum_{\nu \in \mathrm{L}} P(\nu)\, f_\nu(\mathbf{x}; \theta_\nu),\ \nu \in \mathrm{L} \equiv \{1, \ldots, L\}, \qquad (1)$$

where L is the space of classes; L is the number of classes; $f_\nu(\mathbf{x}; \theta_\nu)$ is the conditional unimodular parametric (with the parameter vector $\theta_\nu \in R^{m\nu}$, m_ν is the dimension of the space of parameters) probability density function of class ν; and $P(\nu)$ is the weight of the probability density function $f_\nu(\mathbf{x}; \theta_\nu)$ in the mixture having a meaning of *a priori* probability of appearance of the class ν; $\sum_{\nu \in \mathrm{L}} P(\nu) = 1$, unknown. The problem is to identify all the components of the mixture (1) $\{L, P(\nu), f_\nu(\mathbf{x}; \theta_\nu), \nu \in \mathrm{L}\}$ from the available non-classified sample of observations $\mathbf{X}_1, \ldots, \mathbf{X}_N$ of size N.

It should be noted that the problem of recovering of mixture (1) components has a solution only if it can be identified.[1–3] This condition is difficult to verify in practice. From the geometrical point of view, it means that $f(\mathbf{x})$ must have "well pronounced" local modes generated by cluster-forming subsamples of a mixed sample; besides, the behavior of $f(\mathbf{x})$ in a vicinity of the mode must permit recovering of the parametric functions $f_\nu(\mathbf{x}; \theta_\nu)$, just which are the models of the sought classes, with accuracy sufficient for practice. In such a general formulation, the problem of seeking of unknown parameters θ_ν and other components of Eq. (1), for instance, by the maximal likelihood method, is desperately awkward. If nonparametric estimates of unknown distributions are taken as $f(\mathbf{x})$, the problem becomes even more complicated.

Let us suppose that decomposition of the mixture (1) is performed in a certain way and the corresponding parametric probability measures $f(\mathbf{x}/\nu)$, $\nu \in \mathrm{L}$, are recovered. Then the problem of construction of decision rules for pattern recognition, with estimate of their quality, can be formulated in the following way. Let probability measures with *a priori* distributions of situations $P(\nu)$, conditional probability density functions $f(\mathbf{x}/\nu)$, random vector of observations $\mathbf{X} \in R^n \equiv \chi$, $\nu \in \mathrm{L}$, and a simple loss matrix $1 - \delta_{\nu\mu}$, where $\delta_{\nu\mu}$ is the Kronecker delta, $\nu, \mu \in \mathrm{L}$, be defined in the Euclidean n-dimensional space of observations R^n and the space of hypotheses $\mathrm{L} \equiv \{1, \ldots, L\}$. Then the quality of supposed classification can be estimated by the minimum of mean losses or minimum of mean recognition errors (minimum of the risk functional)

$$r = \sum_{\nu \in \mathrm{L}} \int_{R^n} P(\nu)\, f(\mathbf{x}/\nu) \left[1 - \prod_{\nu \in \mathrm{L}} E\{l_{\nu\mu}(\mathbf{x})\}\right] \mathrm{d}\mathbf{x}, \qquad (2)$$

where

$$E\{t\} = \begin{cases} 1, & t \geq 0, \\ 0, & t < 0; \end{cases}$$

$l_{\nu\mu}(\mathbf{x}) = P(\nu)\, f(\mathbf{x}/\nu) - P(\mu)\, f(\mathbf{x}/\mu) \geq 0;\ \forall \mu \in \mathrm{L},\ \mu \neq \nu, \mathrm{L}$ is the space of classes, $\mathrm{L} \equiv \{1, \ldots, L\}$, L is the number of classes enumerated by the natural scale. In this case, for the simple loss matrix it is the averaged error probability carried by the Bayes decision rule $l_{\nu\mu}(\mathbf{x})$. Decomposition in Eq. (1) can be performed in other nonparametric way, namely, by indication of classes containing some sampled values of a mixed sample represented by separate classes $\mathbf{X}^\nu_1, \ldots, \mathbf{X}^\nu_{N_\nu}$, $\nu \in \mathrm{L}$. In this case, with $f(\mathbf{x}/\mu)$ given, it is naturally to estimate the mean risk (2) by the empirical risk, namely,

$$\hat{r} = \sum_{\nu \in \mathrm{L}} \frac{1}{N_\nu} \sum_{j=1}^{N_\nu} P(\nu)\, I\{\nu = \arg\max_{\mu \in \mathrm{L}} P(\mu)\, f(\mathbf{X}^\nu_j/\mu)\}, \qquad (3)$$

where $I\{\text{"true"}\} = 0$, $I\{\text{"false"}\} = 1$ is the characteristic function; N_ν is the size of the sample of class $\nu \in \mathrm{L}$, and

$$u(\mathbf{x}) = \arg\max_{\mu \in \mathrm{L}} P(\mu)\, f(\mathbf{x}/\mu) \qquad (4)$$

is the Bayes decision rule written in another, identical form; $u(\mathbf{x})$ is the taken decision (in the simplest case, elements of L can be decisions), $u \in \mathrm{L}$.

If the nonparametric estimates $\hat{f}(\mathbf{x}/\mu)$ by learning sequences $\mathbf{X}^\mu_1, \ldots, \mathbf{X}^\mu_{N_\mu}$, $\mu \in \mathrm{L}$, are taken as unknown conditional probability density functions $f(\mathbf{x}/\mu)$, $\mu \in \mathrm{L}$, the empirical risk (3) is calculated by the method of "running" testing in the following way. When $\hat{f}(\mathbf{x}/\nu = \mu)$ is calculated in Eq. (3) for $\nu = \mu$ at the point $\mathbf{x} = \mathbf{X}^\nu_j$, the latter is excluded from the sampling values, by which, properly speaking, $\hat{f}(\mathbf{x}/\nu)$ is estimated. The following two estimates differing in the kernel type will be used as nonparametric estimates for unknown probability density functions. For instance, the estimate with a Gaussian kernel has the following expression:

$$\hat{f}(\mathbf{x}/\nu) = \frac{1}{N_\nu} \sum_{j=1}^{N_\nu} (2\pi)^{-n/2} |\hat{R}_\nu|^{-1/2}\, h_\nu^{-n} \times$$

$$\times \exp\left\{-\frac{1}{2h_\nu^2} (\mathbf{x} - \mathbf{X}^\nu_j)^{\mathrm{T}}\, \hat{R}_\nu^{-1}\, (\mathbf{x} - \mathbf{X}^\nu_j)\right\}, \qquad (5)$$

where $\hat{R}_\nu$ is the covariation matrix (sampled estimate); T is the transposition sign; h_ν is the smoothing parameter whose properties guarantee asymptotic

convergence of $\hat{f}(\mathbf{x}/\nu)$ to the corresponding probability density function.[1–3] In other case, to make the calculations shorter, we use the Epanechnikov kernel with an "internal" coordinate system providing turn of the spread ellipse in correspondence with the spread of sampled data[6,7]

$$\hat{f}(\mathbf{x}/\nu) = \frac{1}{N_\nu} \sum_{j=1}^{N_\nu} \prod_{i=1}^{n} \times$$

$$\times \left\{ \frac{1}{\lambda_\nu^{i1/2} h_\nu} \left[a - b \frac{(\mathbf{g}_i^{\mathrm{T}} (\mathbf{x} - \mathbf{X}_j^\nu))^2}{\lambda_\nu^i h_\nu^2} \right] \right\}, \qquad (6)$$

where the following auxiliary coordinate system is introduced:

$$\mathbf{u} = G\mathbf{x}, \quad M[\mathbf{U}\mathbf{U}^{\mathrm{T}}] = GM\left[\overset{\circ}{\mathbf{X}}\overset{\circ}{\mathbf{X}}^{\mathrm{T}}\right]G^{\mathrm{T}} =$$

$$= G\hat{R}_\nu G^{\mathrm{T}} = \Lambda, \quad \Lambda = \begin{pmatrix} \lambda^1 & & 0 \\ & \ddots & \\ 0 & & \lambda^n \end{pmatrix}, \quad G = \begin{pmatrix} \mathbf{g}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{g}_n^{\mathrm{T}} \end{pmatrix},$$

G is the matrix of the decorrelating orthogonal transform; $M[.]$ is the operator of mathematical expectation; $\overset{\circ}{\mathbf{X}}$ are centered observations; Λ is the diagonal matrix of eigenvalues; $a = 3/4\sqrt{5}$; $b = a/5$. The salient feature of the modified Epanechnikov kernel is that the common smoothing parameter for all dimensions of the space of observations is scaled by coordinates with eigenvalues λ_i, $i = 1, \ldots, n$ of the estimation correlation matrix $\hat{R}_\nu$.

When using nonparametric estimates of unknown distributions (5) and (6), the smoothing parameter h_ν, $\nu \in \mathrm{L}$, remains poorly defined. So there arises a possibility of additional adaptation of probability image models to concrete conditions of observation and learning samples. The most natural, although rather awkward for calculations, approach to determination of smoothing parameters is the way based on minimization of the risk functional (empirical risk) by a set of smoothing parameters with allowance made for the fact that the functional has several extremes and is not differentiable. In this connection, let us consider the following two-step procedure of the search for the global extreme of the functional (3). At the first step, a point is "thrown" randomly with uniform distribution into the search domain which is a multi-dimensional square

$$\prod_{\nu=1}^{L} [h_{\min}^\nu, h_{\max}^\nu], \ \nu \in \mathrm{L},$$

where $h_{\min}$ and $h_{\max}$ are the lower and upper estimated boundaries of the smoothing parameter, respectively. Then the gradient descent from this point is performed; here we use seeking methods of adaptation.[8] For this purpose, the quality functional (3) is varied with respect to the smoothing parameters in the following way. Values of functional increments

$$r_+[\mathbf{h}, a] = (r[\mathbf{h} + a\mathbf{e}_1], \ldots, r[\mathbf{h} + a\mathbf{e}_L]),$$

$$r_-[\mathbf{h}, a] = (r[\mathbf{h} - a\mathbf{e}_1], \ldots, r[\mathbf{h} - a\mathbf{e}_L])$$

are calculated, where L is the number of parameters h corresponding to the number of classes and collected into the parameter vector $\mathbf{h} = (h^1, \ldots, h^L)^{\mathrm{T}}$; a is the scalar parameter defining the value of the search step; $\mathbf{e}_i = \left(\underbrace{0, \ldots, 1}_{i}, \ldots, 0\right)^{\mathrm{T}}$ $i = 1, \ldots, L$ are basis vectors of search directions.

The estimated value of the gradient is calculated in the following way:

$$\frac{r_+[\mathbf{h}, a] - r_-[\mathbf{h}, a]}{2a} = \nabla_{h^\pm} r[\mathbf{h}, a],$$

where $\nabla_{h^\pm}$ is the gradient sign. The recurrent form of the search adaptation algorithm is as follows:

$$\mathbf{h}[j] = \mathbf{h}[j-1] - \gamma[j] \nabla_{h^\pm} r[\mathbf{h}[j-1], a[j]], \qquad (7)$$

the choice of the search $a[.]$ and operating $\gamma[.]$ steps is considered in Ref. 8 (here $\gamma[.] < a[.]$).

It should be noted that the risk functional is the only functional adequate to the problem of quality estimation for decision rules of pattern recognition. However, it is rather awkward for analytical and numerical methods of synthesis of optimum decision rules. In this connection, let us consider simpler criteria of quality estimation for recognizing systems.

Risk for recognizing two patterns $\mu, \nu \in \mathrm{L}$, when a simple loss matrix is given, coincides with the averaged recognition error and is limited from above by the following value ε, which is referred to as the Chernov boundary[2]:

$$r \le [P(\mu)\, P(\nu)]^{1/2} \exp\left\{-\beta_h\left(\frac{1}{2}\right)\right\} = \varepsilon, \qquad (8)$$

where $\beta_h\left(\frac{1}{2}\right) = -\ln \int_\chi [f(\mathbf{x}/\mu)\, f(\mathbf{x}/\nu)]^{1/2}\, \mathrm{d}\mathbf{x}$ is the Bhattacharyya distance.

Probability of error can be represented through the variational Kolmogorov distance

$$2r = 1 - \int_\chi |P(\mu)\, f(\mathbf{x}/\mu) - P(\nu)\, f(\mathbf{x}/\nu)|\, \mathrm{d}\mathbf{x}.$$

Using the Schwarz inequality, we can simultaneously obtain the lower boundary for error probability r:

$$\frac{1}{2} - \frac{1}{2}(1 - 4\varepsilon^2)^{1/2} \le r \le \varepsilon; \qquad (9)$$

if ε is small, $\varepsilon^2 \le r \le \varepsilon$. The Chernov boundaries and Bhattacharyya distance have all the necessary properties for distinguishability criteria of conditional probability distributions that are models of images.

Let us consider the following version for estimation of the Bhattacharyya distance. It is based on the technique of functional integration over empirical distributions.[9] The integral in the expression for the Bhattacharyya distance can be written in the symmetrical form

$$2\int_{\chi}[f(\mathbf{x}/\mu)\,f(\mathbf{x}/\nu)]^{1/2}\,d\mathbf{x} \cong \int_{\chi}\left[\frac{f(\mathbf{x}/\mu)}{f(\mathbf{x}/\nu)}\right]^{1/2} dF_N(\mathbf{x}/\nu) + \int_{\chi}\left[\frac{f(\mathbf{x}/\nu)}{f(\mathbf{x}/\mu)}\right]^{1/2} dF_N(\mathbf{x}/\mu) \cong$$

$$\cong \frac{1}{N_\nu}\sum_{j=1}^{N_\nu}\left\{\frac{\frac{1}{N_\mu}\sum_{i=1}^{N_\mu}(2\pi)^{-n/2}|\hat{R}_\mu|^{-1/2}\,h_\mu^{-n}\exp\left\{-\frac{1}{2h_\mu^2}\rho_\mu(\mathbf{X}_j^\nu,\mathbf{X}_i^\mu)\right\}}{\frac{1}{N_\nu-1}\sum\limits_{\substack{i=1\\ i\ne j}}^{N_\mu}(2\pi)^{-n/2}|\hat{R}_\nu|^{-1/2}\,h_\nu^{-n}\exp\left\{-\frac{1}{2h_\nu^2}\rho_\nu(\mathbf{X}_j^\nu,\mathbf{X}_i^\nu)\right\}}\right\}^{1/2} +$$

$$+\frac{1}{N_\mu}\sum_{j=1}^{N_\mu}\left\{\frac{\frac{1}{N_\nu}\sum_{i=1}^{N_\nu}(2\pi)^{-n/2}|\hat{R}_\nu|^{-1/2}\,h_\nu^{-n}\exp\left\{-\frac{1}{2h_\nu^2}\rho_\nu(\mathbf{X}_j^\mu,\mathbf{X}_i^\nu)\right\}}{\frac{1}{N_\mu-1}\sum\limits_{\substack{i=1\\ i\ne j}}^{N_\mu}(2\pi)^{-n/2}|\hat{R}_\mu|^{-1/2}\,h_\mu^{-n}\exp\left\{-\frac{1}{2h_\mu^2}\rho_\mu(\mathbf{X}_j^\mu,\mathbf{X}_i^\mu)\right\}}\right\}^{1/2}, \qquad (10)$$

where $F_N(\mathbf{x}/\nu)$ is the empirical distribution function.[6,9] The distance function

$$\rho_\mu(\mathbf{X}_j^\nu, \mathbf{X}_i^\mu) = (\mathbf{X}_j^\nu - \mathbf{X}_i^\mu)^{\mathrm{T}}\,\hat{R}_\mu^{-1}\,(\mathbf{X}_j^\nu - \mathbf{X}_i^\mu) \qquad (11)$$

is introduced in Ref. 10.

Thus, nonparametric estimate of the Bhattacharyya distance can serve as a measure of image distinguishability, just as the risk or, more exactly, empirical risk.

Now let us consider another measure of closeness of probability models of images. The measure is similar to the Bhattacharyya distance. When constructing decision rules for pattern recognition, the fundamental part belongs to likelihood relation or monotone transforms of this relation, e. g., $\ln\{f(\mathbf{x}/\mu)/f(\mathbf{x}/\nu)\}$. In this connection, the Kullback divergence[10] having all the necessary properties of distances is an efficient distinguishability measure of classes. Using the technique of integration over empirical distributions[9] and substituting nonparametric estimates of unknown probability density functions by samples into the expression for divergence, we obtain

$$D = \int_{\chi}\ln\frac{f(\mathbf{x}/\mu)}{f(\mathbf{x}/\nu)}\,dF_N(\mathbf{x}/\mu) -$$

$$-\int_{\chi}\ln\frac{f(\mathbf{x}/\mu)}{f(\mathbf{x}/\nu)}\,dF_N(\mathbf{x}/\nu) \cong$$

$$\cong \frac{1}{N_\mu}\sum_{j=1}^{N_\mu}\ln\frac{\frac{1}{N_\mu-1}\sum\limits_{\substack{i=1\\ i\ne j}}^{N_\mu}(2\pi)^{-n/2}|\hat{R}_\mu|^{-1/2}\,h_\mu^{-n}\to}{\frac{1}{N_\nu}\sum\limits_{\substack{i=1\\ i\ne j}}^{N_\nu}(2\pi)^{-n/2}|\hat{R}_\nu|^{-1/2}\,h_\nu^{-n}\to}$$

$$\frac{\to\exp\left\{-\frac{1}{2h_\mu^2}\rho_\mu(\mathbf{X}_j^\mu,\mathbf{X}_i^\mu)\right\}}{\to\exp\left\{-\frac{1}{2h_\nu^2}\rho_\nu(\mathbf{X}_j^\mu,\mathbf{X}_i^\nu)\right\}} -$$

$$-\frac{1}{N_\nu}\sum_{j=1}^{N_\nu}\ln\frac{\frac{1}{N_\mu}\sum\limits_{\substack{i=1\\ i\ne j}}^{N_\mu}(2\pi)^{-n/2}|\hat{R}_\mu|^{-1/2}\,h_\mu^{-n}\to}{\frac{1}{N_\nu-1}\sum\limits_{\substack{i=1\\ i\ne j}}^{N_\nu}(2\pi)^{-n/2}|\hat{R}_\nu|^{-1/2}\,h_\nu^{-n}\to}$$

$$\frac{\to\exp\left\{-\frac{1}{2h_\mu^2}\rho_\mu(\mathbf{X}_j^\nu,\mathbf{X}_i^\mu)\right\}}{\to\exp\left\{-\frac{1}{2h_\nu^2}\rho_\nu(\mathbf{X}_j^\nu,\mathbf{X}_i^\nu)\right\}}. \qquad (12)$$

Note that nonparametric estimates of distributions with the Epanechnikov kernel (6) can be similarly used in Eqs. (10) and (12).

Thus, estimates of simpler criteria (10) and (12) with the "smoothness" property are obtained together with the direct distinguishability criterion for recognized classes (3). This simplifies the solution of some optimization problems.

Now let us turn to the steps of construction of the segmentation algorithm. First of all, we dwell on the problem of choice of the learning material.

CHOICE OF LOCAL FRAGMENTS AND FORMATION OF LEARNING MATERIAL

Segmentation of large arrays of video data makes it necessary to select fragments for learning of the algorithm of pattern recognition. Quality of classification of the whole image depends on the quality of learning to a large extent. In this connection, the set of fragments for learning material must be set in such a way as to reflect the variety of the whole field of video data. In other words, a little set of fragments must reflect statistical properties of the whole general set of data. In one of versions of algorithm operation, the choice of fragments in a large image can be left to operator's intuition or given by a random mechanism.

Let us dwell on the possibility to choose learning material "purified" from mixed pixels to some extent. Note that one of the causes of large variety of classes to be distinguished is, on one hand, the variety of states of natural formations and, on the other hand, appearance of a large number of mixed pixels generated by bad resolution of scanning systems. For instance, resolution of the NOAA satellite and AVHRR device in nadir is only 1.1 × 1.1 km and, as facies is the smallest observable unity in aerial photography of landscapes,[13] we can suppose that a large number and variety of elementary landscape forms is concentrated on an area of 1.1 × 1.1 km. Just this leads to appearance of mixed pixels, which are integral characteristics of real situations. In this connection, when forming the learning material, it is expedient to try to separate some parts with stationary behavior of radio brightness, rather than choose data fragments. We supposedly can formulate a hypothesis that a stationary part of radio brightness corresponds also to a stationary landscape formation, whose portrait we want to separate. In this connection it is expedient to enter these quasi-stationary parts of the image in the learning material. This problem can be solved by spatial differentiation of video data followed by separation of parts with small and close to zero gradient.

SEARCH FOR LOCAL MODES OF THE MIXING DISTRIBUTION (FIRST STEP OF THE ALGORITHM)

Let us again turn to Eq. (1). We accept the hypothesis that the problem of mixture identification has a solution meaning that the mixing distribution is multimodal. Every mode pretends to form its own class, which will be called a subclass or a local class. Thus, there arises a primary problem of seeking local modes of the mixing distribution (1). In the language of a sample consisting of representatives of all the classes, this means seeking local aggregations, blobs of sampling values in the general bulk of data of the learning material $\mathbf{X}_1, \ldots, \mathbf{X}_N$ obtained from a fragment of video data. The main formation idea of the algorithm for seeking local modes is that in the vicinity of each local mode, by definition, the number of sampling values is larger as compared with adjacent domains. Let an elementary volume described, for instance, by a spread ellipsoid be given and transported, with fixed number of points inside the volume, along the sampled space. Then the number of sampled vectors in the ellipsoid is largest in the case when the center of the seeking ellipsoid is close to a local mode of the mixing distribution. The problems of convergence for such an algorithm of mode seeking are considered in Ref. 11. It should be noted that the number of directions, in which the seeking ellipsoid may move, sharply increases with increasing dimensionality of the observation space. As a simplified variant, it is natural to estimate every local mode by the nearest sampled vector. Thus, the center of the seeking ellipsoid should be placed only at the points of a mixed sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$, and the number of nearest neighbors falling within the ellipsoid should be estimated. This significantly simplifies the problem. To form the function of distance (11) between sampled observations, let us estimate the covariation matrix with use of the whole mixed sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$:

$$\hat{R} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{X}_i - \hat{\mu})\, (\mathbf{X}_i - \hat{\mu})^{\mathrm{T}},$$

where $\hat{\mu}$ is the estimated mathematical expectation by the same sample. Let us define the generalized function of distance in the following way:

$$\rho(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^{\mathrm{T}}\, \hat{R}^{-1}\, (\mathbf{x} - \mathbf{y}), \qquad (13)$$

where $\mathbf{x}$ and $\mathbf{y}$ are replaced with observations from the sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$.

For definiteness sake, let the Gaussian kernel be taken as a kernel of a recovered nonparametric estimate of the unknown probability density function. Let us set the boundaries of smoothing parameters obtained from considerations connected with maximization of the likelihood functional or, more precisely, empirical estimate of the entropy functional[12]:

$$h_{\min} = \sqrt{\frac{\sum_{i=1, i \neq j}^{N} \min_{\{j\}} \rho\,(\mathbf{X}_i, \mathbf{X}_j)}{N\, n}},$$

$$h_{\max} = \sqrt{\frac{\sum_{i=1}^{N} \max_{\{j\}} \rho\,(\mathbf{X}_i, \mathbf{X}_j)}{N\,n}}, \qquad (14)$$

where $h_{\min}$ and $h_{\max}$ are estimated boundaries of the smoothing parameter. They are written without the index of belonging to a class. We will place the "center" with the point $\mathbf{Y}$ of the kernel function

$$f(\mathbf{x}) = (2\pi)^{-n/2}\,|\hat{R}|^{-1/2}\,h_{\min}^{-n} \times$$

$$\times \exp\left\{-\frac{1}{2\,h_{\min}^{2}}\,\rho(\mathbf{X}, \mathbf{Y})\right\} \qquad (15)$$

sequentially at every of the sampling points $\mathbf{X}_1, \ldots, \mathbf{X}_N$, $\mathbf{Y} \in \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and, for every such position, estimate the number of sampled values falling within the vicinity of the central point $\mathbf{Y}$. Here the vicinity is defined by a certain level Δ, so that $\{\mathbf{X}: f(\mathbf{X}_j) \geq \Delta\}$, where Δ is significance level of the field of influence.

Let us consider the set of maximal size of N_1 sampled values $\mathbf{X}_1^1, \ldots, \mathbf{X}_{N_1}^1$ falling within the Δ-vicinity (the set was separated at the first step), and remove the values from the general sample. Let the remained part be renumbered as $\mathbf{X}_1, \ldots, \mathbf{X}_N$ by a sequential index, where the new value N is the preceding value without N_1. Repeating the process, we separate some number of mode-forming subsamples. Let this number be L^1 for the first fragment of the learning set. Some sampled points (vectors) may appear single and we set them aside for a while. The similar procedure is performed for each elementary fragment chosen as a learning one. So, in the general set, we obtain a sufficiently large number of classes $\sum_k L^k$, $k = 1, \ldots, K$, where K is the number of the chosen fragments.

AGGREGATION OF LOCAL CLASSES USING THE BHATTACHARYYA OR KULLBACK MEASURES OF CLOSENESS

The aim of this step is to separate the closest subclasses among the whole variety and unite them. The subclasses are closest in the sense of measures (10) and (12). For this purpose, we study distances between pairs of local classes. Suppose that a certain minimal expected number of classes $L_{\min}$ and their maximal number $L_{\max}$ be known *a priori*, so it is expedient to consider a decomposition of the whole image into some number L classes and $L_{\min} \leq L \leq L_{\max}$. The iteration process is exhaustion of all the possible pairs of classes with calculation of nonparametric estimates for the Bhattacharyya distance (10) or the Kullback's measure (12). In the process of the first iteration, united is only the pair of classes with minimal distance in the sense of the chosen measures. At this step, we use the generalized metric based on $\rho(\mathbf{x}, \mathbf{y})$, as the formed classes are not sufficient to estimate the proper metrics of the type (11). Thus, continuing the iteration process of uniting pairs of local classes, we reduce their number to the number closest to $L_{\max}$.

UNIFICATION OF CLASSES BY THE CRITERION OF MAXIMUM EMPIRICAL RISK (SECOND STEP OF THE ALGORITHM)

At this stage, it is expedient to perform more correct unification into classes by use of, first, the proper metrics of classes with individual measures (11) and, second, the criterion of empirical risk. The first unification of class pairs is performed for the largest risk values $r \cong 0.5$ (it means that the classes are fully indistinguishable). When recovering the conditional distributions, calculated are covariation matrices of classes $\hat{R}_\nu$ with smoothing parameter expressed through observations[12]:

$$h_\nu \cong \sqrt{\sum_i \sum_{j \neq i} \rho_\nu(\mathbf{X}_i, \mathbf{X}_j) / nN_\nu\,(N_\nu - 1)},$$

where $\nu = 1, \ldots, L$, L is a certain number of classes synthesized at the given step. To over the quality of classes distinguishability, the generalized risk (3) is calculated and its value for a given set of classes is stored. Continuing the hierarchic process of unification and enlargement of classes from $L_{\max}$ to $L_{\min}$ and comparing values of the generalized risk, we stop at its minimal value which results from the set of classes $L = L_{\text{opt}}$. Below we use this set of classes in order to optimize the Bayes decision rules (4) by smoothing parameters and to estimate the quality of feature subspaces.

SELECTION OF INFORMATIVE FEATURES BY THE CRITERION OF MINIMUM EMPIRICAL RISK IN RECOGNITION OF MULTIZONAL IMAGE COMPONENTS (THIRD STEP OF THE ALGORITHM)

Two main points should be highlighted in the problem of choice of informative features, namely, it is necessary to define the functional of information content of the feature subsystem, as well as the formation technology of sequences of feature subspaces that are studied for information content.

First of all, let us note that only the mean risk is adequate to the problem of estimation of quality (information content) for complexes of features. In place of the mean risk, the empirical estimate of the latter by the learning sample, i.e., the same criterion, by minimization of which the optimal (Bayes) rule of pattern recognition was obtained, may be also used. As to the ways for the choice of feature subspaces, the

variety of methods applied in practice is not large. Note that the solution of the formulated problem is known and trivial: to obtain the optimal system consisting of k features chosen among n initial components of the observation vector, one needs only to compare the values of the information content criterion that were calculated at different k-dimensional subspaces and to fix the set of k features, at which the chosen criterion reaches its optimum. The number of these calculations of the optimum criterion equals the number $\binom{n}{k}$ of combinations of n features taken k at a time, what is astronomically computationally expensive even for comparatively small k and n. That is why truncated exhaustion is widely applied in practice. For instance, the algorithm which is conventionally denoted as "A" performs truncated exhaustion reducing the system of features by sequential elimination of features with low information content. In another algorithm "B", the system of informative features is constructed sequentially by inclusion of features with high information content. We used the combined algorithm for choice of informative subspaces of k features. It is a modified version of truncated exhaustion. This algorithm resides in such procedure that full exhaustion by i features of n is used until the number of combinations $\binom{n}{i}$ determining versions of full exhaustion of systems by i features is small and acceptable in the sense of computational expenses for calculation of the functional of information content of i-dimensional subspaces. Thus, at the first step, we select i ($i \ll k$) informative features. At the second step, fixing $(n - i)$ remaining features which pretend to complement the informative set, we select, again by full exhaustion, a system of j features such that the set of $i + j$ ($i + j \ll k$) features is optimal with respect to the information content criterion, and so on. The system of features is extended by blocks until the informative set of $i + j + \ldots + l$ features reaches the sought value k. In a particular case, taking $i = j = \ldots = l = 1$, we obtain the algorithm "B". A similar generalization is possible for the algorithm "A" of truncated exhaustion of subspaces, in which the initial dimensionality is decreased also by blocks in the mode of conditional full exhaustion. So, the proposed algorithm permits us to consider additional versions of feature spaces and test them for information content.

Thus, all the components of the information content criterion (3) are defined completely, and r can be used for estimation of information content of feature sets what is just performed at this step.

NONPARAMETRIC CLASSIFICATION OF THE WHOLE IMAGE BY THE PATTERN RECOGNITION ALGORITHM IN THE SPACE OF INFORMATIVE FEATURES (FOURTH STEP OF THE ALGORITHM)

Finally, after the probability models of the classes are recovered and the optimal complex of informative features is separated, the estimating Bayes decision rule (4) relates the unknown re-observed vector $\mathbf{X} \in R^k$, where R^k is the k-dimensional space of informative features (the vector is sequentially selected from the whole field of the analyzed multilayer image), to one of the given classes. So the whole image is segmented by a pattern recognition algorithm; as a result, texturally homogeneous fields of video data become resolved.

REFERENCES

1. S.A. Aivazyan, V.M. Bukhshtaber, I.S. Enyukov, and L.D. Meshalkin, *Applied Statistics: Classification and Decrease of Dimensionality* (Finansy i Statistika, Moscow, 1989), 607 pp.
2. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press, New York, 1972).
3. J.T. Tou and R.C. Gonzalez, *Pattern Recognition Principles* (Addison-Wesley Publishing Company, Reading Mass., 1974).
4. K.T. Protasov, Atmos. Oceanic Opt. **11**, No. 1, 74–79 (1998).
5. K.T. Protasov, Izv. Vyssh. Uchebn. Zaved., Ser. Fiz. **38**, No. 9, 59–64 (1995).
6. C.R. Rao, *Linear Statistical Inference and Its Applications* (John Wiley & Sons Inc., New York–London–Sydney).
7. V.A. Epanechnikov, Teor. Ver. Primen. **14**, No. 1, 156–161 (1969).
8. Ya.Z. Tsypkin, *Adaptation and Teaching in Automatic Systems* (Nauka, Moscow, 1968), 400 pp.
9. F.P. Tarasenko, *Nonparametric Statistics* (State University, Tomsk, 1976), 294 pp.
10. S. Kullback, *Information Theory and Statistics* (John Wiley & Sons Inc., New York, Chapman & Hall Limited, London, 1959).
11. A.A. Zhiglyavskii, *Mathematical Theory of Global Random Seeking* (State University, Leningrad, 1985), 296 pp.
12. N.V. Ivanova and K.T. Protasov, in: *Mathematical Statistics and Its Applications* (State University, Tomsk, 1982) issue 8, pp. 50–65.
13. Yu.S. Tolchel'nikov, *Optical Properties of a Landscape* (Nauka, Leningrad, 1974).