ISOLATION OF CLOUD FIELDS IN SATELLITE PICTURES USING THE SEGMENTATION ALGORITHM BASED ON CLASSIFICATION AND PATTERN RECOGNITION

K.T. Protasov

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences, Tomsk
Received August 18, 1997

We have developed a combined algorithm involving four-step procedure for making segmentation of multispectral pictures of clouds and underlying surface taken from a satellite. The procedure steps include (1) fragment-by-fragment, clusterization of video data into classes; (2) then the closest classes identified are united using Bhattacharia spacing; (3) thus formed ensembles of classes are used to teach the pattern recognition algorithm used; (4) and finally, this algorithm is used to make fragmentation of the entire image. This approach enables one to compromise between the necessity of using certain models of the patterns to be recognized, for instance, those based on the Johnson approximations we use here, and bulky and cumbersome arrays of the initial data. The algorithm performance is illustrated by an example of discriminating cloud fields from the patterns recorded with an Advanced Very High Resolution Radiometer (AVHRR) onboard NOAA satellite.

The major source of operational information for solving problems of nature management and climate and ecological monitoring are multichannel satellite pictures of clouds and underlying surface. Because satellite imagery data are typically acquired under conditions of broken clouds an additional problem arises on distinguishing cloud fields routinely by means of algorithms of segmenting the imagery data. The problem of cloud field discrimination itself is important in cloud water content assessment, microstructure sampling, cloud classification, and thunderstorm prediction. A specific feature of this class of problems is the necessity of handling large-scale fields (for instance, a NOAA satellite imagery field consists of 2048 by 5000 readouts for each of the five AVHRR spectral intervals), which makes it difficult to use standard approaches.

Let us now consider a combined algorithm of segmenting multichannel satellite images which is based upon cluster analysis of local data segments whose results are input to pattern recognition techniques.

The image segmentation is normally understood as an automated image splitting into the interpretable areas. These areas may be associated with the objects to be detected or recognized each with its own brightness, geometrical, and texture characteristics. As such, the image segmentation is the initial step in constructing a formal description of the scene to be analyzed. This formal description then is used to detect, identify, and recognize objects and phenomena.

The simplest segmentation algorithms are often constructed based on the concepts of boundaries and constant radiance levels. A detailed review of the results from these algorithms can be found elsewhere.[1,2] More elaborate segmentation algorithms use marking of the region points based on the homogeneity concept. In this case the image segmentation is related to the problem on data clustering which is solved by means of cluster analysis and techniques of automatic classification and taxonometry.[3–6,8]

The segmentation algorithms use the classification methods in the following way. For each point (pixel) of a given image, certain set of characteristics is fixed which are called features. In particular, these may be digitized values of the recorded brightness fields in all spectral intervals. In the space of these features, certain groups (clusters) of points are isolated. In the initial image, points of each cluster make up connected regions separated out by a segmenting algorithm. Performance of a segmenting algorithm will depend on the system of features formed and on the strategy of selecting closest (in some sense) points.

Two approaches may conventionally be pointed out to formation of features. In the first one the features characterize some portion of a cloud or surface scene viewed by a device and map this data (within the sensor resolution) as an element (point, pixel) of the image analyzed, so that a set of features is created for every image pixel. These may be brightness characteristics of the image element from a series of spectral intervals or functional transforms of those

0235-6880/98/01 74-06 $02.00

brightness characteristics. In this way, the dimensionality of the feature space is defined by the number of sensor spectral channels in use.

According to the second approach the features are defined within a vicinity of the image element. In this case one may use statistical properties of the spectral brightnesses, brightnesses averaged over an image fragment, covariation or texture characteristics of the digitized fields. In this case the set of features for the pixel analyzed depends on the nearby elements within the vicinity chosen.

When constructing an image segmentation algorithm one must (a) form suitable set of features for an imagery element, either pixel or a fragment; (b) construct a criterion of point clustering in the feature space by introducing a proximity measure; and (c) outline the strategy of detecting some (desired) number of clusters once their number is not prescribed *a priori*.

Despite of vast literature on synthesizing algorithms for an automated classification (for many examples, see Refs. 3–6), this problem is not yet properly formalized and the list of algorithms is permanently extended with new ones. Classification algorithms often use heuristic tricks that allow for a variety of problems on the video data analysis. It is also worth noting that the analyzed satellite pictures of the earth's surface are generally over 1024×1024 readouts size being set by the series of spectral intervals in use. There number constantly grows, so that hyperspectral imagery data including hundreds spectral intervals have recently emerged, making cluster analysis techniques computationally expensive as well.

The Isodata-Iterative Self-Organizing Data Analysis Techniques (IZODATA) are used, in simple cases, as part of the ERDAS and ENVI software packages. It is based on formalizing empirical experience where the image points are grouped into a number of clusters, by minimizing the sum of rms deviations of points from the cluster centers and maximizing the dispersal of centers themselves.[4,5] Large-format images are clustered block-by-block, and the local clustering results are then matched together.

However, among the approaches to constructing the clustering algorithms there is one theoretically most well grounded that is based on describing classes in the feature space using combinations of the probability distributions. In this case, extraction of individual clusters is connected with solution of the problem of splitting joint distribution into elemental conditional unimodal probability functions that, in fact, are the models of the classes sought. Unfortunately, one needs for identifying poorly pronounced local extrema of a function in order to solve this problem and this is a computationally too a cumbersome task even for a small bulks of data.[3–5]

Below we present an algorithm of automatic classification which preserves good local properties and, yet, is capable of working with large video data arrays. It is a four-stage procedure. First, small fragments of multispectral imagery data are clusterized using a threshold decision making rule applied to all variants possible. The second step is to normalize the data and unite local classes of all fragments into larger blocks with the help of intergroup proximity measure. Third, the pattern recognition algorithm is taught to distinguish between classes produced at the data aggregation stage. The final stage is to recognize the image as a whole by using the decision making rule chosen. Considering that material chosen as teaching can be few fragments which are statistically equivalent to the entire image, this approach offers considerable reduction of computer time yet preserving the accuracy of the decision making rule.

It is specific feature of the algorithm proposed that the proximity or distinguishability of the classes is quantified in terms of a risk functional or the upper or lower boundaries of the class, while the probabilistic models of classes are retrieved using Johnson approximation.

Let us now assume that a small image fragment locally is clustered using an available technique which is chosen to be computationally cheap yet rigorously founded, and concentrate on the second and third phases of the algorithm construction.

First, consider in a more detail construction of the classification algorithm, followed by pattern recognition. Let a set of the digitized video data fields be available from observations at several spectral intervals, so that each pixel of the surface and cloud image, as viewed by an imager, is characterized by the random vector $\mathbf{X} = (X^1, \ldots, X^n)^{\mathrm{T}}$, where T stands for transposing, $\mathbf{X} \in R^n$, while R^n is the n-dimensional space of observations. The components X^i, $i = 1, \ldots, n$ of the observation vector $\mathbf{X}$ characterize reflective (radiobrightness) properties of land areas and clouds at each spectral interval, respectively. We assume that the joint distribution of the components of $\mathbf{X}$ vector in the observation space can be expressed by the probability density function as

$$f(\mathbf{x}) = \sum_{j=1}^{Q} P_j \, f(\mathbf{x}; \theta_j), \qquad (1)$$

where Q is the number of classes, $f(\mathbf{x}; \theta_j)$ is the conditional unimodal parameterical (with the parameter vector $\theta_j \in R^m$ and m being the dimensionality of the parameter space) probability density function for the class j; P_j is the weight of the probability density function $f(\mathbf{x}; \theta_j)$ in a mixture that has the meaning of *a priori* probability that the class j occurs; $\sum_{j=1}^{Q} P_j = 1$; and θ_j are the parameters of the probability density function. All parameters defined above are unknown. The task is to retrieve, from the available non-classified sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ of N observations, all the components $\{Q, P_j f_j(\mathbf{x}; \theta_j), \; j = 1, \ldots, Q\}$ of the mixture (1).

It should be noted that the problem of reconstructing components of the mixture (1) is only solvable if the mixture is identifiable.[3,6] This condition can hardly be checked in practice and, geometrically, that means that $f(\mathbf{x})$ must have well pronounced local modes produced by cluster-generating subsamples of a mixed sample. Moreover, the behavior of $f(\mathbf{x})$ in the vicinity of a mode must enable the reconstruction of the parametric functions $f_j(\mathbf{x}; \theta_j)$ sufficiently accurate. The latter functions are the models of the classes sought.

Now, let us consider the task of choosing approximating distributions $f(\mathbf{x}; \theta_j)$ for unknown cluster models once a data sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ of this class with the volume N is available. Here we consider the reconstruction of parametric distributions using Johnson approximations.[7,10] In so doing, we will allow for the consistency of individual feature distributions with the true, one-dimensional distributions, and describe, to a certain degree of correctness, statistical correlations between the vector components being observed. The reconstruction of parametric probability density functions can be represented as a two-step procedure. First, an appropriate Johnson transformation is selected for each feature to sufficiently closely match the unknown true distribution. At the second stage interfeature correlation coefficients are evaluated to describe correlation between the transformed components of the observed vector. Then, joint distribution of the components of the observed vector is written as (with class index not shown)

$$f(\mathbf{x}; \theta_j) = \frac{\prod_{i=1}^{n} \delta^i \tau'_x (x^i; \varepsilon^i, \lambda^i)}{(2\pi)^{n/2} |R|^{n/2}} \exp \{-\frac{1}{2} (\gamma + \delta\tau(\mathbf{x}; \varepsilon, \lambda))^{\mathrm{T}} R^{-1} (\gamma + \delta\tau(\mathbf{x}; \varepsilon, \lambda))\}, \quad (2)$$

where vector $\gamma + \delta\tau(\mathbf{x}; \varepsilon, \lambda) = \begin{bmatrix} \gamma^1 + \delta^1\tau(x^1; \varepsilon^1, \lambda^1) \\ \cdot \\ \cdot \\ \cdot \\ \gamma^n + \delta^n\tau(x^n; \varepsilon^n, \lambda^n) \end{bmatrix}$;

$$R = (\tau_{ij}); \quad \tau_{ij} = \frac{1}{N} \sum_{k=1}^{N} [\gamma^i + \delta^i\tau(x_k^i; \varepsilon^i, \lambda^i)] \times [\gamma^j + \delta^j\tau(x_k^j; \varepsilon^j, \lambda^j)];$$

$\tau(x; \varepsilon, \lambda)$ is one of the normalizing Johnson transformations; $\theta = (\gamma^1, \ldots, \gamma^n; \delta^1, \ldots, \delta^n; \varepsilon^1, \ldots, \varepsilon^n; \lambda^1, \ldots, \lambda^n; (\tau_{ij})_{n\times n})^{\mathrm{T}}$, and

$$\tau'_x(x; \varepsilon, \lambda) = \left| \frac{\mathrm{d}\tau(x; \varepsilon, \lambda)}{\mathrm{d}x} \right| .$$

The parameters ε and λ are the lower boundary and the span of a sample of each component of the observation vector. These parameters can be evaluated either from physical considerations or directly from a sample. When ε and λ are known, both γ and δ are readily determined by the maximum likelihood method.[7,10]

After the probabilistic models of classes are reconstructed, Bayes decision making rule refers the newly observed vector $\mathbf{x}$ to one of the classes available[4–6]

$$u = \arg\max_{i=1,\ldots,Q} P_i f_i(\mathbf{x}; \hat{\theta}_i), \quad (3)$$

where P_i, $i = 1, \ldots, Q$ are *a priori* probabilities of occurrence of the classes or their estimates, and u is the decision made (the number of the class recognized). This decision making rule minimizes the probability of wrong decisions (averaged recognition error). The criterion of a minimal classification error is a particular case of the criterion of the type of risk. These criteria are the only ones that are adequate to hypothesis recognition and testing.

Since the probability of classification error is difficult to estimate[4] since the integration of the weighted probability density functions is required over multidimensional feature spaces, it is worth using the boundaries of the error probability. Really, given two classes ($Q = 2$), the average probability of wrong decisions ε is expressed through the Kolmogorov variational distance as[4]

$$\varepsilon = \frac{1}{2} \left[1 - \int_X |P_1 f(\mathbf{x}/1) - P_2 f(\mathbf{x}/2)| \mathrm{d}\mathbf{x} \right], \quad 0 \le \varepsilon \le 1/2, \quad (4)$$

where $P_1 f(\mathbf{x}/1)$ and $P_1 f(\mathbf{x}/2)$ are the class 1 and class 2 conditional probability density functions, respectively, and X is the integration domain. It appears that

$$(1/2) - (1/2)(1 - 4\varepsilon_n^2)^{1/2} \le \varepsilon \le \varepsilon_n,$$

where $\varepsilon_n = [P_1 P_2]^{1/2} \exp\{-\mu(1/2)\}$, so that the quantity $\mu(1/2) = -\ln \int_X [f(\mathbf{x}/1) \times f(\mathbf{x}/2)]^{1/2} \mathrm{d}\mathbf{x}$, which is called the Bhattacharia distance, can be used as a simplified criterion of the class separability.

Consider now a simple variant of the pattern recognition or automatic classification with the Gaussian models of the class description. For Gaussian model, the Bhattacharia distance is[4]

$$\mu_N\left(\frac{1}{2}\right) = \frac{1}{8} (m_1 - m_2)^{\mathrm{T}} \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (m_1 - m_2) + \frac{1}{2} \ln \frac{\frac{1}{2}(\Sigma_1 + \Sigma_2)}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}}, \quad (5)$$

where m_i and Σ_i are the mathematical expectation and correlation matrix for the ith class, $i = 1, 2$. If $\Sigma_1 \cong \Sigma_2$, which is true for close classes, then

$$\mu_N\left(\frac{1}{2}\right) = \frac{1}{8}(m_1 - m_2)^{\mathrm{T}}\,\Sigma^{-1}\,(m_1 - m_2) +$$

$$+\frac{1}{2}\ln\frac{|\Sigma|}{|\Sigma|} \cong 0, \qquad (6)$$

and the Bhattacharia distance can serve as a criterion for making concatenation of two distributions or, more exactly, two sampled ensembles of observations (for which μ_N is small) into a single group thus making up a homogeneous class.

The class separability criterion (5) can also be used for distributions constructed with the help of Johnson normalizing transformations as follows. First, note that Chernov boundary and Bhattacharia distance[4] are invariant relative to one–to–one transformations. It is readily shown that, for multidimensional Johnson distributions, it is possible to construct a one–to–one transformation that converts distributions (2) to the normal probability density functions, for which distributions, the Bhattacharia distance is calculated and has the form of Eq. (5). Actually, for normalizing transformation of S_L family with the interface ε and common parameter λ, we obtain the following normalizing nondegenerate transformation (written for one component of the vector, for brevity):

$$\xi = \gamma + \delta \ln (x - \varepsilon)/\lambda \qquad (7)$$

and the transformation inverse to it

$$x = \varepsilon + \lambda \exp\{(\xi - \gamma)/\delta\}. \qquad (8)$$

By, substituting ξ into the probability density function (2) we obtain, in the new space, that

$$f_j(\xi) = \frac{1}{(\sqrt{2\pi})^n\,|G_j|^{1/2}} \times$$

$$\times \exp\left\{-\frac{1}{2}\left(\xi - \left(\gamma - \frac{\delta}{\delta_j}\gamma^j\right)\right)^{\mathrm{T}} G_j^{-1}\left(\xi - \left(\gamma - \frac{\delta}{\delta_j}\gamma^j\right)\right)\right\}, \qquad (9)$$

where $G_j = M\left(\xi - \left(\gamma - \frac{\delta}{\delta_j}\gamma^j\right)\right)\left(\xi - \left(\gamma - \frac{\delta}{\delta_j}\gamma^j\right)\right)^{\mathrm{T}}, j = 0, \ldots, Q;$ M is the operator of mathematical expectation; γ and δ are the parameters of the transformation function, and ε and λ are the same quantities as in Eq. (2). Similar expressions may be obtained for the other Johnson distributions.

The program for implementing this algorithm begins with choosing fragments from the image analyzed that are to be used as teaching material. These fragments can be selected manually by an operator instructed to follow fragmentation objectives; namely, by increasing representativeness of some class or another. It is possible, following this way, to increase the accuracy of its description and selection. If necessary, the arrangement of fragments within the image analyzed can be made using some random mechanism, for instance, it may be set uniform.

The idea of a strategy of fragment selection should naturally give teaching material based on the properties of statistical homogeneity and representativeness characteristic of the entire ensemble of the video data on the field to be analyzed.

As the teaching fragment sampling is formed next step is to extract local clusters within each fragment. Since the clusters are small, any cluster analysis procedure can be used to isolate clusters, including the procedures of running over all interpoint distances capable of identifying point clusters by the nearest-neighbor technique.

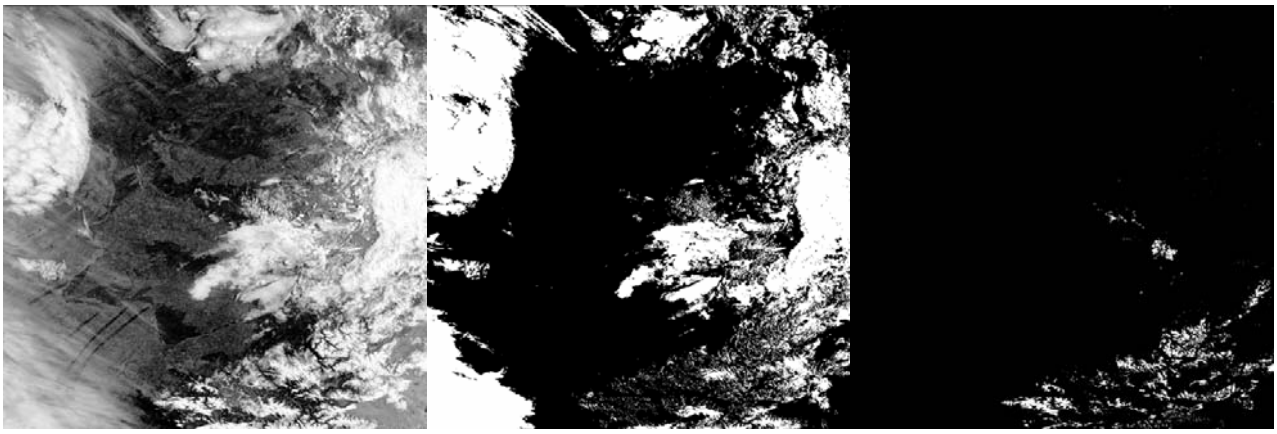
In the case, being considered here we used the method of extracting left mode in the interpoint distance histogram,[9] which provides for obtaining the statistically mean measure of the point clusterization. The main idea of this method is as follows. For each vector with the components being the spectral brightness values, one calculates distances, in the Euclidean metrics, to all other vectors of the fragment analyzed and constructs histograms of these distances.

Then, from all the histograms formed, one selects the histogram that has the mode positioned to the left of modes of other histograms. The adaptive threshold cuts off the group of points making up this histogram. Then these points form a cluster and are excluded from the subsequent consideration. The process is repeated many times until only few vectors remain that do not belong to any class established. These are anomalous vectors useless for teaching, and their classification is performed at the stage of recognition.

At the second step, results of local clusterizations are aggregated by using Bhattacharia distance (6) as a measure of class closeness in the space previously normalized with the transformation (8). All possible pairs of classes of all fragments are checked for closeness, in the sense of the minimum of this distance, and the closest ones are concatenated. As a result, we obtain a few number of classes preset by the operator.

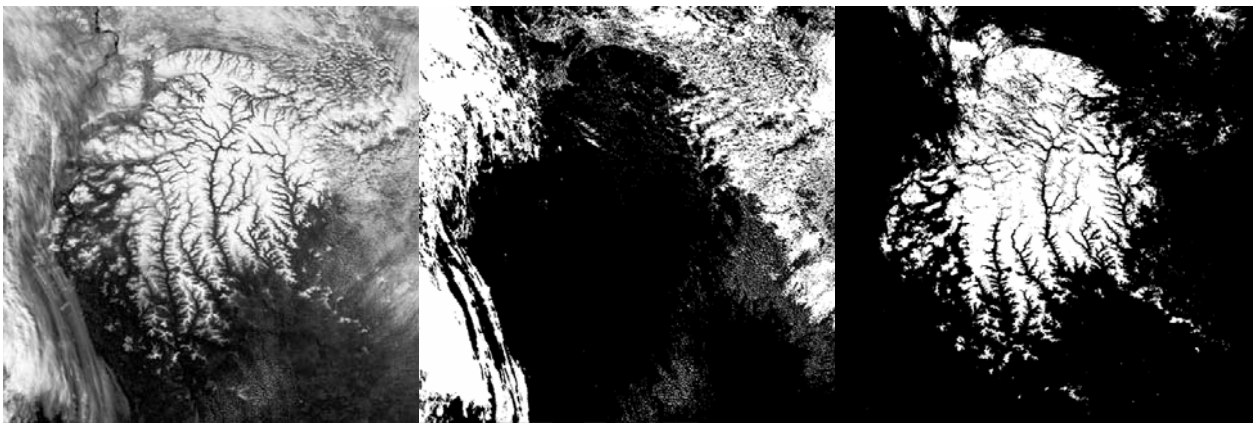
At the third stage, the classes, thus obtained, serve as teaching samples for a pattern recognition algorithm. In so doing, parameters of the decision making rule (3) are estimated using Eq. (2). At the fourth stage, the decision making rule obtained is used to classify the entire image by a standard technique.

To illustrate the algorithm performance, 3 images were taken with the size 1024×1024 readouts, acquired with the AVHRR from a NOAA satellite in 5 spectral intervals (channels), 0.58–0.68 μm, 0.725–1.1 μm, 3.55–3.93 μm, 10.3–11.3 μm, and 11.5–12.5 μm, with the spatial (footprint) resolution of 1×1 km^2 per pixel.



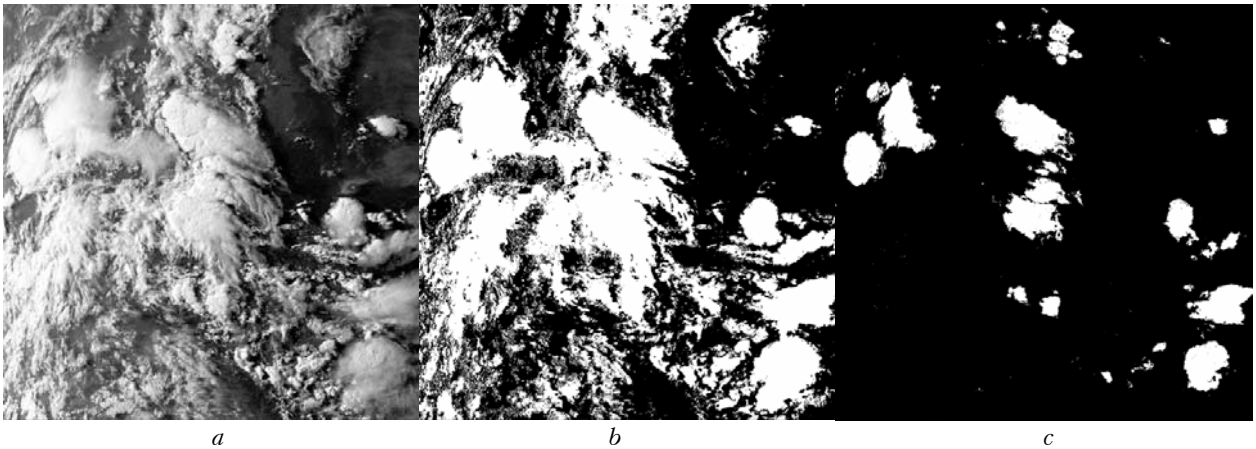
a *b* *c*

FIG. 1. (a) Image acquired with the channel No. 1 of NOAA-12 AVHRR; (b) cloud fields identified and (c) snow-cover in mountains.



a *b* *c*

FIG. 2. (a) Image acquired with the channel No. 1 of NOAA-14 AVHRR; (b) cloud fields identified and (c) snow-cover on an elevated plateau.



a *b* *c*

FIG. 3. (a) Image acquired with the channel No. 1 of NOAA-12 AVHRR; (b) cloud fields identified and (c) thunderstorm clouds.

The first one was taken over West Siberia and Altai from NOAA-12 satellite, from the orbit with apogee at 81°83′, at 09:34 LT on May 14, 1997, and shows the presence of clouds and snow cover over Altai mountains. Eighteen fragments of 64 by 64 readouts were chosen to teach the segmenting algorithm. These fragments were uniformly spaced within the field of the initial, 1024 by 1024 readouts image. The results of processing were 233 classes, finally aggregated to give 50 classes. Then the pattern recognition algorithm was used to analyze the entire image. Some results obtained using this algorithm are shown in Fig. 1. Figure 1*a* presents the initial image in the radiometer channel number one. The result of extracting the class "cloud" is presented in Fig 1*b* and that of the class "snow" in Fig. 1*c*. Based on the data presented in the figures we conclude that the classes are resolved pretty well in a five-dimensional feature space, while being indistinguishable in individual spectral channels.

The second image was recorded over Putoran plateau (medium-mountain elevation of the Central Siberian plateau with the height marks of 900 to 1701 m) by NOAA-14 satellite, orbiting with the apogee of 28°00′, on May 29, 1997, at 14:30 LT), and shows the presence of clouds and snow cover. The fragments of 64×64 pixel size uniformly spaced over the entire image were isolated to teach the algorithm. A total of 223 classes were identified that finally were grouped to give 50 classes. Then the pattern recognition algorithm was applied to analysis of the entire image. Some results thus obtained are depicted in Fig. 2. Figure 2*a* shows the initial image from the radiometer channel number one. The result of extracting the class "cloud" is presented in Fig 2*b*, and the class "snow" (snow-covered flat tops of ridges) in Fig. 2*c*. Such a clear separation between the classes could hardly be feasible if only single channel data were used.

The third image was recorded over West Siberia and Altai Mountains by NOAA–12 satellite, orbiting with the apogee of 76°00′, on August 8, 1997, at 19:37 LT, and shows the presence of clouds including thunderstorm ones. Twenty five fragments of 64 by 64 pixel size were chosen to teach the segmenting algorithm. These fragments were spaced uniformly over the entire image of 1024 by 1024 readouts. As a result 221 classes were identified that finally were combined into 50 classes. Then the entire image was analyzed using the pattern recognition algorithm. Some results of the analysis are shown in Fig. 3. Figure 3*a* shows the initial image from the radiometer channel number one. The result of extracting the class "cloud" is presented in Fig 3*b*, and the class "thunderstorm clouds" in Fig. 3*c*. From the figures it follows that the classes are again well resolved in the five-dimensional feature space.

REFERENCES

1. P.A. Bakut, G.S. Kolmogorov, and I.E. Vornovitskii, Zarub. Radioelektronika, No. 10, 6–24 (1987).
2. P.A. Bakut and G.S. Kolmogorov, Zarub. Radioelektronika, No. 10, 25–46 (1987).
3. S.A. Aivazyan, V.M. Bukhshtaber, I.S. Enyukov, and L.D. Meshalkin, *Applied Statistics: Classification and Reduction of Dimensionality* (Finansy i Statistika, Moscow, 1989), 607 pp.
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press, New York, 1972).
5. J. Tu and R. Gonsales, *Principles of Pattern Recognition* [Russian translation] (Mir, Moscow, 1978), 408 pp.
6. Ya.Z. Tsypkin, *Fundamentals of the Theory of Teachable Systems* (Nauka, Moscow, 1970), 252 pp.
7. K.T. Protasov, Atmos. Oceanic Opt. **7**, No. 6, 448–451 (1994).
8. K.T. Protasov, Izv. Vyssh. Uchebn. Zaved., Ser. Fizika **38**, No. 9, 59–64 (1995).
9. Yu.V. Gridnev and K.T. Protasov, Atmos. Oceanic Opt. **8**, No. 7, 574–578 (1995).
10. G. Han and S. Shapiro, *Statistical Models in Engineering* [Russian translation] (Mir, Moscow, 1969), 369 pp.