Greedy algorithms of feature selection for multiclass image classification

E F Goncharova[1] and A V Gaidel[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34, Samara, Russia, 443086
[2]Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, Molodogvardeyskaya str. 151, Samara, Russia, 443001

Abstract. To improve the performance of remote sensing images multiclass classification we propose two greedy algorithms of feature selection. The discriminant analysis criterion and regression coefficients are used as the measure of feature subset effectiveness in the first and second methods respectively. The main benefit of the built algorithms is that they estimate not the individual criterion for each feature, but the general effectiveness of the feature subset. As there is a big limitation on the number of real remote sensing images, available for the analysis, we apply the Markov random model to enlarge the image dataset. As the pattern for image modelling, a random image belonging to one of the 7 classes from the UC Merced Land-Use dataset has been used. Features have been extracted with help of MaZda software. As the result, the largest fraction of correctly classified images accounts for 95%. Dimension of the initial feature space consisting of 218 features has been reduced to 15 features, using the greedy strategy of removing a feature, based on the linear regression model.

CEUR-WS.org/Vol-2210/paper6.pdf

1. Introduction

Multiclass or multinomial classification is a significant and complicated step, which can be applied in solving various computer vision tasks. Large number of techniques has been developedto perform the task of multinomial image classification. Some of them apply neural networks, while the others tend to adapt the classical methods of machine learning to improve the quality of the classification results.

In this paper we present two greedy algorithms of feature selection to improve the performance of multiclass image classification. An image itself can be described by various numerical characteristics. For example, the MaZda software for texture analysis [1] estimates almost 300 histogram and texture features, moreover, it includes procedures for their reduction and classification. It should be noticed that not all the extracted features have similar influence on image distinguishing. Redundancy features can affect the performance of classification badly and require additional computational cost.

The feature selection methods have been widely developed in recent years. Some researchers propose feature selection methods based on clustering process. In paper [2] the algorithm is built in the following way: firstly, objects are clustered, than the features which provide the biggest distance between clusters' centroids areappended to the subset of the most informative features. In [3] authors present the novel approach of dimensionality reduction for hyperspectral image classification. To reduce the numbers of variables they use inter band block correlation coefficient technique and QR

decomposition. The support vector machines algorithm has beenapplied to fulfill the classification task. Classification accuracy for images from different databases is between 83 and 99%.

In this work we use MaZda software to extract more than 200 texture features per image. The most informative features are selected with the help of two greedy strategies, based on the discriminant analysis and linear regression model, respectively. The proposed algorithms enable us to select descriptors which have the strongest effect on multinomial image classification. As there is a huge limitation on the number of images, available for the analysis, we also consider the algorithm of image modelling based on the applying of Markov random fields [4].

The experiments are carried out on images belonging to 7 land use classes, 100 for each class, from the UC Merced Land-Use dataset, which provides aerial optical images. To measure the significance of feature subset we estimate the classification error, using *k*-nearest-neighbor scheme. To estimate the effectiveness of image synthesis, we compare the description of the generated and source image in the best feature subset, using the Euclidean distance between two feature vectors.

2. Feature extraction

An image is characterized by its intensity matrix $I^{(M\times N)}$, where $M \times N$ is an image size.

$$I(m,n) = \frac{R(m,n) + G(m,n) + B(m,n)}{3},\ m = \overline{1,M},\ n = \overline{1,N}, \quad (1)$$

R, G, B is an intensity of red, green, and blue component of the image resolution cell having coordinates (m,n) respectively. $I(m,n)$ ranges in value from 0 to $\mathbf{I}-1$, where $\mathbf{I}-$ is a maximum grey level.

To extract the features we compute numerical descriptors of an image, which, eventually, are going to be used to perform the feature selection procedure and further classification. The MaZda software is applied to form the set of features, describing input images [1].

The histogram is calculated via the intensity of each image pixel, calculated by (1), regardless to the spatial correlation among the pixels. The following descriptors are computed: mean intensity, variance, skewness, kurtosis, and percentiles.

The next type of features includes the textural characteristics, calculated with the gray-level spatial dependence matrix. It is built according to the following rule:

$$P_{d_1 d_2}(i,j) = \left\{(m,n) \in \{1,2,...,M\} \times \{1,2,...,N\} \mid I(m,n) = i, I(m+d_1, n+d_2) = j\right\}, i, j = \overline{0, L-1}.$$

Thus, the following features are calculated for five different distances in four directions: angular second moment, contrast, entropy, and correlation.

Features from the other group are calculated based on the autocorrelation function, which describes the dependency between image pixels. The calculated features, as well as the previous ones, are estimated for five different distances in four directions.

3. Methods of feature selection

3.1. Formulatin of feature selection task

The main idea of feature selection process is to improve the classification performance. Thus, let $\mathbf{\Omega}$ be a set of objects for recognition. The set $\mathbf{\Omega}$ is divided into L non-overlapping classes.

To fulfill the classification task we should create the mapping function $\tilde{\Phi}(x)$, which identifies the feature vector x, $x \in \Phi^K$ (K – number of features), with its class. $\tilde{\Phi}(x)$ should be as similar to the ideal mapping function $\Phi(x)$ as possible. $\Phi(x)$ is the ideal mapping function, which is aware of the information about the real object's class.Classification is considered an instance of supervised learning, that is why $\tilde{\Phi}(x)$ is created on the basis of a training set of data $\mathbf{U} \subseteq \mathbf{\Omega}$, containing object with the known class labels.

The aim of feature selection step is to extract the subset of the most informative features, which provides the least classification error.

To classify the feature vectors we apply*k*-nearest-neighbor scheme. According to this method, an object is classified by a majority vote of its neighbors. The classifier assigns the class of the x vector to the class of its *k*-nearest neighbors. The distance between two feature vectors is calculated as the Euclidean distance (2):

$$\rho(x,y)=\sqrt{\sum_{i=1}^{K}(x_i-y_i)^2},\ x\in \mathrm{R}^K ?y\in \mathrm{R}^K, \tag{2}$$

where K is a number of features.

The nearest neighbor error rate is assessed by the formula (3).

$$\varepsilon=\frac{\left|\left\{x\in\tilde{\mathbf{U}}\mid\Phi(x)\neq\tilde{\Phi}(x)\right\}\right|}{\left|\tilde{\mathbf{U}}\right|}, \tag{3}$$

We should notice that $\tilde{\mathbf{U}}$, which is a test set, should be independent of the training set, i.e. $\tilde{\mathbf{U}}\cap\mathbf{U}=\varnothing$. In order to avoid overfitting of classification model and get more accurate results, the leave-one-out cross-validation technique is applied.

Normalization of data is a crucial step of classification process. As different features can be measured in varied scales they affect the classification performance differently. To avoid this problem, all the features in dataset should be standardized. Therefore, the feature vectors get zero mean and unit variance. To achieve this goal we should estimate the expected value $\overline{x}(i)$ and variance $\sigma_{x(i)}$ for each feature.

$$\overline{x}(i)=\frac{1}{|\mathbf{\Omega}|}\sum_{x\in\mathbf{\Omega}}x(i),\ \overline{x}(i)\in\square\ ,$$

$$\sigma_{x(i)}=\frac{1}{|\mathbf{\Omega}|}\sum_{x\in\mathbf{\Omega}}\left(x(i)-\overline{x}(i)\right)^2,\ \sigma_{x(i)}\in\square\ .$$

Thus, each feature can be standardized by applying formula (4).

$$\forall x\in\mathbf{\Omega}\quad x(i)=\frac{x(i)-\overline{x}(i)}{\sqrt{\sigma_{x(i)}}},\ i=\overline{1,K}\ . \tag{4}$$

3.2. Greedy adding algorithm based on the discriminant analysis

When we have several classes, feature selection aims on choosing the features which provide the strongest class separability. In discriminant analysis theory the criterion of separability is evaluated using within-class, between-class, and mixture scatter matrices. Let x be a random vector, belonging to the feature space. Therefore, to measure the importance of the current feature space we should evaluate the degree of isolation of vectors, belonging to different classes [5].

The feature selection method based on the discriminant analysis criterion was proposed in paper [6]. There the separability of two classes was assessed with help of the discriminant criterion[7]. In this work we generalize that technique to the case of several classes.

A within-class scatter matrix (5) shows the scatter of points around their respective class mean vectors (6), and is calculated as follows:

$$\overline{x}_j=\frac{1}{\left|\mathbf{U}\cap\mathbf{\Omega}_j\right|}\sum_{x\in\mathbf{U}\cap\mathbf{\Omega}_j}x,\ \overline{x}_j\in\square^{|\mathbf{Q}|}. \tag{5}$$

$$R_j=\frac{1}{\left|\mathbf{U}\cap\mathbf{\Omega}_j\right|}\sum_{x\in\mathbf{U}\cap\mathbf{\Omega}_j}\left(x-\overline{x}_j\right)\left(x-\overline{x}_j\right)^T,\ R_j\in\square^{|\mathbf{Q}|\times|\mathbf{Q}|}. \tag{6}$$

Prior probability of class $\mathbf{\Omega}_j$ is expressed by $\mathrm{P}\left(\mathbf{\Omega}_j\right)=\dfrac{\left|\mathbf{U}\cap\mathbf{\Omega}_j\right|}{\left|\mathbf{U}\right|}$. The mixture scatter matrix is the covariance matrix of all samples among all the classes, it is defined by:

$$R_{mix}=\frac{1}{\left|\mathbf{U}\right|}\sum_{x\in\mathbf{U}}\left(x-\overline{x}_{mix}\right)\left(x-\overline{x}_{mix}\right)^T,\ R_{mix}\in\mathbb{R}^{|\mathbf{Q}|\times|\mathbf{Q}|},$$

where $\overline{x}_{mix}=\sum_{i=0}^{L-1}\overline{x}_i P\left(\mathbf{\Omega}_i\right),\ \overline{x}_{mix}\in\mathbb{R}^{|\mathbf{Q}|}$ is a mean vector of mixture distribution.

Thus, the discriminant criterion is formulated as

$$J\left(\mathbf{Q}\right)=\frac{\operatorname{tr}R}{\sum_{l=0}^{L-1}\mathrm{P}\left(\mathbf{\Omega}_l\right)\operatorname{tr}R_l}.$$

Criterion $J\left(\mathbf{Q}\right)$ tries to assess the influence of feature set $\mathbf{Q}$ on the within-class compactness and inter-class separability.

To select the most informative features we propose greedy adding strategy. On the first step of the algorithm current set of features is empty $\mathbf{Q}_{(0)}=\varnothing$. On the step i, we observe all the sets, formed as follows $\mathbf{Q}_{(i,j)}=\mathbf{Q}_{(i-1)}\cup\left\{j\right\}$, and calculate the criterion $J_{ij}=J\left(\mathbf{Q}\right)$. We choose the feature subset which provide the maximum value of criterion J_{ij}:

$$\mathbf{Q}_{(i)}=\mathbf{Q}_{(i-1)}\cup\left\{\underset{j\in[1;K]\cap\mathbf{Z}\setminus\mathbf{Q}_{(i-1)}}{\arg\max}\ J_{i,j}\right\}=\mathbf{Q}_{(i-1)}\cup\left\{\underset{j\in[1;K]\cap\mathbf{Z}\setminus\mathbf{Q}_{(i-1)}}{\arg\max}\ J\left(\mathbf{Q}_{(i-1)}\cup\left\{j\right\}\right)\right\}.$$

Then the above steps are repeated until we get the required number of features.

3.3. Greedy algorithm of feature removing based on the regression model

The second algorithm develops the method, examining in paper [6]. The regression analysis studies the relationship between the output (dependent) variable and one, or more, independent descriptors. For the binary classification the number of class can be considered as the dependent variable, which is influenced by feature vector. In the case of multinomial classification we cannot use the number of class as an output, thus we present the function $\Psi^l\left(x\right):\mathbf{\Xi}\to\left[0;1\right]\cap\mathbf{Z}$, which determines whether the feature belongs to the class l or not. The function is defined by:

$$\Psi^l\left(x\right)=\begin{cases}1, y\left(x\right)=l,\\ 0, y\left(x\right)\neq l.\end{cases}$$

Thereby, $\Psi^l\left(x\right)$ is a dependent variable, which is affected by the feature vector $x\in\mathbf{\Xi}\left(\mathbf{Q}\right)$. To assess the degree of feature vector influence we should build L linear regression equations:

$$\Psi^l=X\theta^l+\varepsilon^l,\ l=\overline{0,L-1},$$

where $\Psi^l=(\Psi_1^l\quad\Psi_2^l\quad\ldots\quad\Psi_n^l)^T$ – the output vector; X – "object-feature" matrix; $\theta^l=\left(\theta_0^l\quad\theta_1^l\quad\ldots\quad\theta_{|\mathbf{Q}|}^l\right)^T$ – regression weights; $\varepsilon^l=\left(\varepsilon_1^l\quad\varepsilon_2^l\quad\ldots\quad\varepsilon_n^l\right)^T$ – error vector.

The unknown parameters are estimated by applying the method of least squares:

$$\left(\Psi^l-X\theta^l\right)^T\left(\Psi^l-X\theta^l\right)\to\min_{\theta^l}.$$

Therefore, L vectors θ^l which characterize the coefficients in linear regression are found for each of L classes. Vector $\hat{\theta}=\left(\hat{\theta}_1\quad\hat{\theta}_2\quad\ldots\quad\hat{\theta}_{|\mathbf{Q}|}\right)^T$ is expressed by:

$$\hat{\theta}_i = \sum_{l=0}^{L-1} (\theta_i^l)^2,\ i = \overline{1, |\mathbf{Q}|}. \quad (7)$$

The measure of the influence of each feature is evaluated according to the vector $\hat{\theta}$ element. To select the most informative features we propose greedy removing strategy. The initial feature subset includes all the features $\mathbf{Q}_{(0)} = \mathbf{Q}$. Than we sequentially remove the worst feature from the current subset and rebuild the linear regression as follows: on the step j of the algorithm we create L linear regression models $\tilde{\Psi}^l_{(j)} = X^l_{(j)} \theta^l_{(j)}$, then vector $\hat{\theta}_{(j)}$ is calculated for the current feature subset $\mathbf{Q}_{(j)}$. The feature with the minimal value of $\hat{\theta}_{(j)}(k)$ is removed from the current subset:

$$\mathbf{Q}_{(j+1)} = \mathbf{Q}_{(j)} \setminus \left\{ \arg\min_{k \in [1;K] \cap \mathbf{Z} \cap \mathbf{Q}_{(j)}} \left| \hat{\theta}_{(j)}(k) \right| \right\}.$$

The steps of the algorithm are iterated until a required number of features is obtained.

4. Image modelling

4.1. Markov random fields

Imagemodellingisperformedwithhelpof Markovrandomfields. Let $F = \{F_i \mid i \in S\}$ be a multivariate random variable, which is defined on the discrete set of index $S = \{1, 2, ..., N\}$. F_i is a random variable that takes values $\{F_1 = f_1, F_2 = f_2, ..., F_n = f_n\}$. The probability of random variable F_i taking the value of f_i is denoted as $P(f_i)$. Thus, F is a random field.

Configuration $f = (f_1, f_2, ..., f_n)$ is a specific realization of random variable F. Let $\mathrm{N} = \{\mathrm{N}_i \mid \forall i \in S\}$ be a neighborhood system, where N_i – set of elements neigh bouring i. Thus, the nodes that influence the local characteristics of the i-node are included in its neighbourhood system.

Markov random fields satisfy the following formula (8):

$$\forall f \ P(F_i = f_i \mid F_j = f_j, i \neq j) = P(F_i = f_i \mid F_j = f_j, j \in \mathrm{N}_i). \quad (8)$$

Hence, Markov random fields imply conditional independence [8]. According to (8) F_i only depends on the nodes included in its neighbourhood N_i. Thus, if nodes in the neighbourhood are known than the values of F_j for $j \neq i$ and $j \notin \mathrm{N}_i$ do not affect F_i [9].

4.2. Image modelling

Suppose that the set of indices S defines the set of points on the 2D plane. The discrete image is a realization of 2D random variable F, defined in the points S. Following by the conditional independence of F, wecanassumethattheintensityvalueofeachimagepixel can be predicted on the basis of several nodes, included in its neighbourhood.

Thus, we can present the following strategy of image modelling using the Markov random fields. On the each step k of the algorithm the neighbourhood system N_i is created for i pixel of the image $G_{(k)}(i)$. Than this neighbourhood system is compared with the neighbourhood of the correspond pixel, belonging to the input image $G_{in}(i)$. $G_{in}(i)$ is a sample real image for synthesis.The pixel's values are set as follows:

$$G_{(k+1)}(i) = G_{in}\left(\arg\min_{i_* \in S} \left(\rho\left(\mathrm{N}_{i_*}(G_{in}), \mathrm{N}_i\left(G_{(k)}\right) \right) \right) \right).$$

The distance $\rho(x, y)$ is defined by formula (2). The initial image $G_{(0)}$ is approximated by the white noise.

In this paper we propose to use causal 5-neighbourhood system. This neigh bourhood pattern is shown in figure 1. The peculiarity of this type of neigh bourhood is that it contains only those nodes that precede the current output pixel. That means that $\mathrm{N}_i\left(G_{(k)}\right)$ includes already assigned pixels.

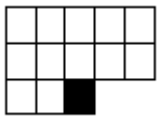

Figure 1. The instance of causal 5-neighbourhood system. The qurrently processing pixelis marked byblacksquare.

5. Experimental results

5.1. Experiments of feature selection

The experiments were carried out on the images from the UC Merced Land-Use dataset, which consists of the aerial optical images, belonging to different classes (agricultural field, forest, beach, etc.), 100 for each class. Each image measures 256×256 pixels. In this work we analyzed images, belonging to 7 classes (agricultural field, forest, buildings, beach, golf course, chaparral, and freeway). figure 2 illustrates sample images belonging to the mentioned above classes.

To get the correct classification results the Leave-one-out cross-validation technique was applied. The total number of features, extracted with the MaZda accounts for 218.

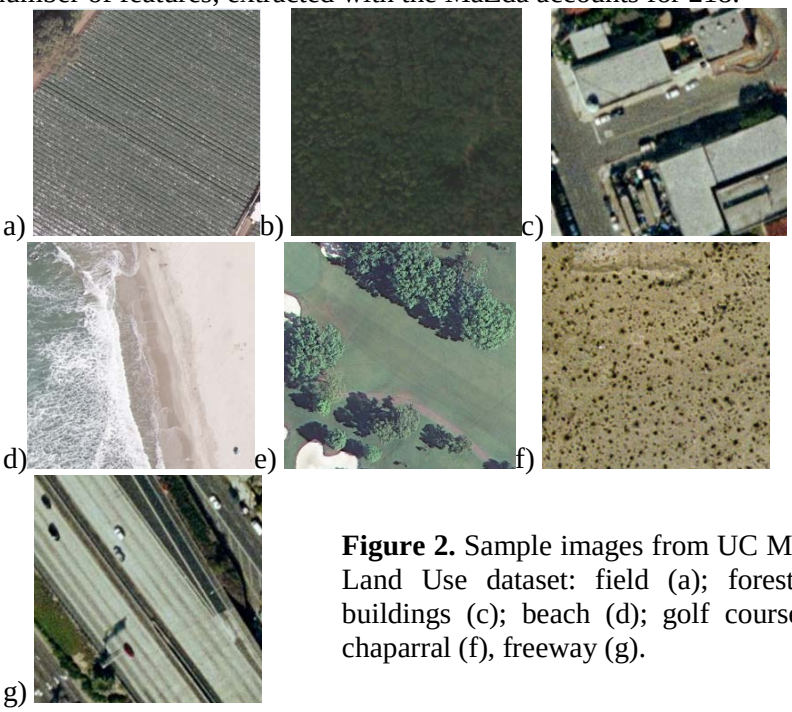

Figure 2. Sample images from UC Merced Land Use dataset: field (a); forest (b); buildings (c); beach (d); golf course (e), chaparral (f), freeway (g).

The results obtained with the discriminant and regression analysis methods are shown in table 1.

The most informative groups of features, selected with the two proposed strategies, along with the classification error (3), obtained on these groups, are presented in tables 2 and 3.

Having analyzed the results, we can conclude that the greedy removing algorithm, based on the linear regression model, performed best on this multinomial classification task. The lowest classification error rate of 0.05 was achieved in feature space, consisting of the 15 features from the 218 initial.

Table 1. The features selected with the greedy algorithms, based on discriminant and regression analysis respectively, in descending order of priority.

Discriminant analysis		Regression analysis	
Feature number	Feature name	Feature number	Feature name
37	S11SumVarnc	96	S202DifEntrp
30	S01DifEntrp	74	S02DifEntrp
24	S01InvDfMom	85	S22DifEntrp
40	S11DifVarnc	107	S30DifEntrp
…	…	…	…
79	S22InvDfMom	171	S44Entropy
34	S11SumOfSqs	215	S55Entropy
32	S11Contrast	217	S55DifEntrp

Table 2. Groups of the most informative features, selected with the discriminant analysis.

K	Features	ε
3	37, 30, 24	0.74
4	37, 30, 24, 40	0.47
5	37, 30, 24, 40, 38	0.63
6	37, 30, 24, 40, 38, 2	0.63
…	…	…
47	37, 30, 24, 40, 38, 2, 55, 13, 42, 209, 44, …, 91, 69, 80, 16, 118, 85	0.32
48	37, 30, 24, 40, 38, 2, 55, 13, 42, 209, 44, …, 91, 69, 80, 16, 118, 85, 5	0.32

Table 3. Groups of the most informative features, selected with the regression analysis.

K	Features	ε
3	96, 74, 85	0.16
4	96, 74, 85, 107	0.32
5	96, 74, 85, 107, 140	0.42
6	96, 74, 85, 107, 140, 63	0.16
…	…	…
15	96, 74, 85, 107, 140, 63, 151, 129, 173, 118, 110, 154, 66, 99, 143	0.05
16	96, 74, 85, 107, 140, 63, 151, 129, 173, 118, 110, 154, 66, 99, 143, 138	0.05

The best group includes various textural features, extracted for 4 dimensions: 2, 3, 4 and 5. The greedy adding algorithm maximized the discriminant analysis criterion provided worse results. The lowest classification error rate of 0.32 was achieved on the set, consisting of 47 features. We should notice that the fracture of the images that were classified correctly in the whole space of 218 features accounts for 63%. That means that both analyzed techniques have succeeded in dimension reduction and improving classification performance.

5.2. Experiments of image modelling

To carry out the experiments of image synthesis we have presented the initial images the UC Merced Land-Use dataset in the greyscale. The results of modelling are shown in figure 3.To check the quality of synthesized images we performed the comparison of the feature vectors for the input sample and the obtained image. The vectors in clude 15 best features, selected by the greedy removing algorithm. The measure of equality $\xi(x,y)$, $x \in \mathsf{R}^K$, $y \in \mathsf{R}^K$ isexpressedby

$$\xi(x,y) = 1 - \left(\frac{1}{K} \sum_{k=1}^{K} (x_k - y_k)^2 \right)^{\frac{1}{2}}.$$

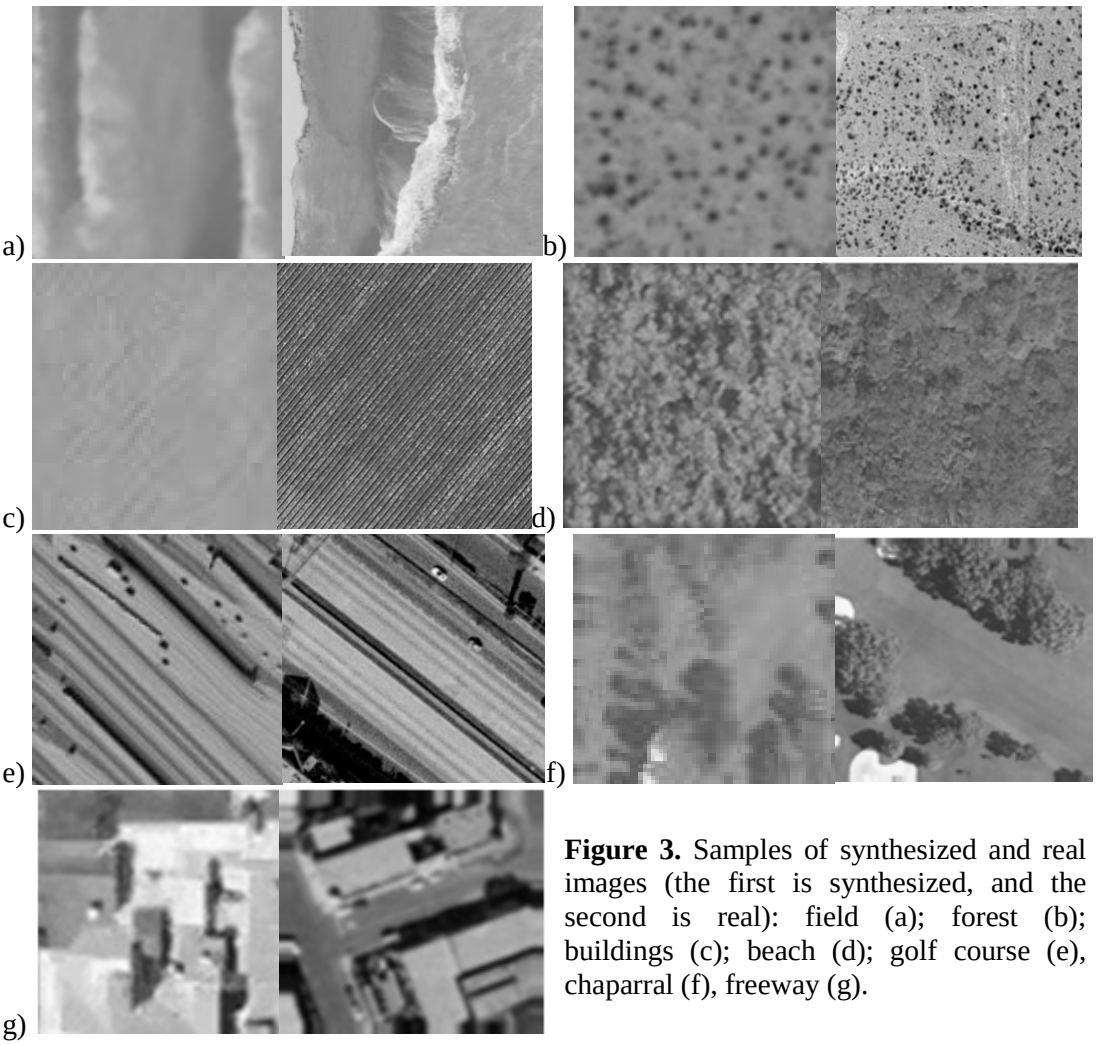

Figure 3. Samples of synthesized and real images (the first is synthesized, and the second is real): field (a); forest (b); buildings (c); beach (d); golf course (e), chaparral (f), freeway (g).

Table 4 presents the value of $\xi(x, y)$ for the images synthesized for 7 classes.

Table 4. Measure of equality for the synthesized images.

Class	$\xi(x, y)$
Beach	0.15
Chaparral	0.08
Field	0.14

Continuation of table 4

Forest	0.09
Freeway	0.09
Golf course	0.09
Buildings	0.13

The results shown in table 4 prove that the proposed method performs successfully for the images with small scale structure. For example, synthesized images belonging to the classes: chaparral and field, turned to be quite similar to the real images. However the quality of synthesized images containing large scale structure is lower. Its modelling demands large neighborhoods which leads to the increasing computational cost. To solve this problem, method, based on the multiresolution image

pyramids, is proposed in paper [4]. In that methodcomputation is saved because the large scale structures are presented more compactly by a few pixels in a certain lower resolution pyramid level.

6. Conclusion

Thus, for the task of the remote sensing images classification the subset of informative features was extracted. We proposed two greedy strategies for informative feature selection. The feature vector, selected with the greedy removing algorithm, based on building the regression model, produced the best classification performance (using the nearest-neighbor classification method) on the images from the UC Merced Land Use dataset. The minimal classification error rate made up 0.05. In comparison to that, the greedy adding algorithm maximized the discriminant analysis criterion provided worse results. The lowest classification error rate of 0.32 was achieved on the set, consisting of 47 features. We should notice that the fracture of the images that were classified correctly in the whole space of 218 features accounted for 63%.

Overall, applying the feature selection methods leads to improving the multinomial image classification performance and dimension reduction. Using only 15 (of 218 initial) descriptors allows to classify 95% of images correctly

To increase the number of images, available for analysis, we applied the algorithm of image modelling on the basis of Markov random fields. The experimental results showed that this technique can be applied for synthesis images with the low scale structure. To generate samples, containing large scale structure, the proposed algorithm should be adopted. One of the possible variants is to apply multiresolution image pyramids.

7. References

[1] Strzelecki M A, Szczypinski P, Materka A and Klepaczko A 2013 A software tool for automatic classification and segmentation of 2D/3D medical images *Nuclear Instruments and Methods in Physics Research* **702** 137-140

[2] Liu C, Wanga W, Zhao Q, Shen X and Konan M 2017 A new feature selection method based on a validity index of feature subset *Pattern Recognition. Leters.* **92** 1-8

[3] Reshma R, Sowmya V and Soman K P 2016 Dimensionality Reduction Using Band Selection Technique for Kernel Based Hyperspectral Image Classification *Proc. Computer Science* **93** 396-402

[4] Wei L Y and Levoy M 2000 Fast texture synthesis using tree-structured vector quantization *Proc. of the 27th annual conf. on Computer graphics and interactive techniques* 479-488

[5] Gaidel A V 2015 A method for adjusting directed texture features in biomedical image analysisproblems *Computer Optics* **39(2)** 287-293 DOI: 10.18287/0134-2452-2015-39-2-287-293

[6] Goncharova E and Gaidel A 2017 Feature Selection Methods for Remote Sensing Images Classification *3rd Int. conf. Information Technology and Nanotechnology* 535-540

[7] Kutikova V V and Gaidel A V 2015 Study of informative feature selection approaches for the texture image recognition problem using Laws' masks *Computer Optics* **39(5)** 744-750 DOI: 10.18287/0134-2452-2015-39-5-744-750

[8] Winkler G 2012 *Image Analysis, Random Fields and Dynamic Monte Carlo Methods* (Springer-Verlag) p 387

[9] Li S Z 2009 *Markov random field modeling in image analysis*(Springer-Verlag) p 356

Acknowledgments

This work was supported by the Federal Agency for Scientific Organizations under agreement No. 007-GZ/Ch3363/26.