A research of classification algorithm of spatial information on the basis of methods of persistent homology and random forest

S V Eremeev[1], K V Kuptsov[1] and Yu A Kovalev [1]

[1]Vladimir State University named after Alexander and Nikolay Stoletovs, Gorky street 87, Vladimir, Russia, 600000

Abstract. The classification problem of spatial data is one of the most difficult challenges in the field of the analysis and processing of spatial information. A new approach to a solution of the classification problem of spatial data is presented in article. The offered classification technology of objects will be based on algebraic topology, namely on methods of persistent homology. A barcode is a qualifier of a spatial object. It is determined by computation of topological features of a classified object. The distinctive feature of the offered algorithm is its invariance to affine and topological transformations. The research on results of classification algorithm operation on a set of spatial objects of different classes is carried out.

1. Introduction

Automatic digitization of maps is one of global problems in geographic information systems [1, 2, 3]. Questions of identification [4] and classification of cartographical information appear within this problem. The problem of classification of spatial data on object classes is one of the most difficult in the field of the analysis and processing of spatial information. Russian and world researchers try to solve this problem and propose a set of application-oriented solutions. Having studied scientific works on this subject it is possible to tell that they solve a problem of object classification with various degree of efficiency. There are different methods of classification of spatial objects.

The method intended for work with topographic maps of average scale is presented in [5].The main application is a classification of the area of objects under construction. The method is based on geometrical structures of data and spatial analytical methods. Advantage is improvement of quality of automation of cards with areas of objects under construction.

The problem of classification of spatial data is also relevant for control of information on exhaustion of reservoirs or, on the contrary, – about their degradation [6]. The technology is applied to spatial objects which have similar spectral features, but various form. The algorithm is realized for classification of reservoirs on Alaska and also is used in Bolivia for classification of pastures.

The analysis of the image is applied in [7] together with network methods of extraction of information within a problem of creation of digital tourist maps. The algorithm classifies spatial objects according to the developed rules of simplification and generalization of maps to emphasize reference points and to reduce a role of less significant objects. The technology is applied to creation of tourist maps of San Francisco.

Processing of satellite images or images with high resolution is made in [8] for classification of the objects which are contained in them. A classification is made for the main classes of objects which are presented on topographic maps of large scale.

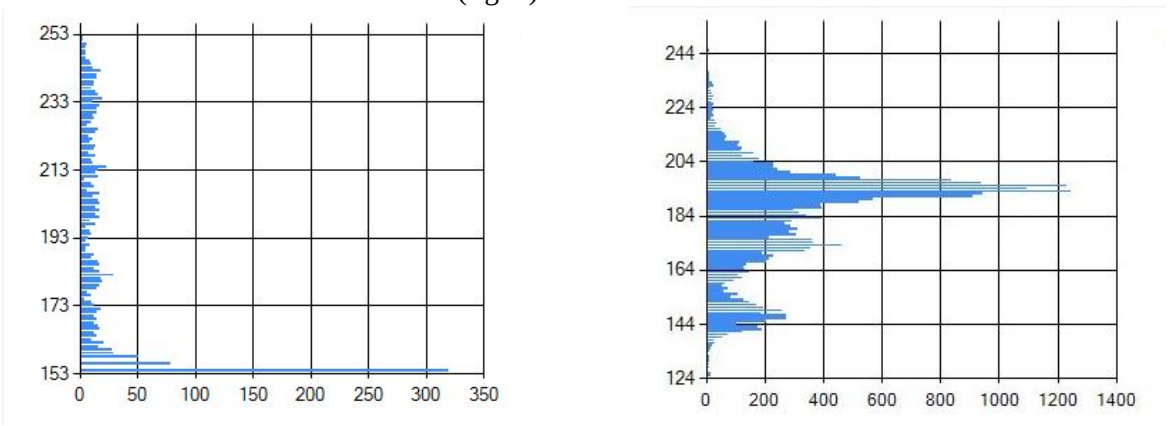
An approach using the example of digitizing distribution maps taken from plant-taxonomic atlases is described in [9]. In result, plant distributions over Europe and Asia have been digitized. The algorithm is a tool to capture data from maps based on obscure projections.

The purpose of work is creation of an algorithm for classification of cartographical information which will make high-quality object classification of various spatial classes and also is invariant [10] to affine transformations and changes of scale.

2. A classification algorithm of spatial data on the basis of methods of a persistent homology and random forest

The offered algorithm of classification of objects is based on algebraic topology, namely by methods of persistent homology. Application of topological characteristics and their analysis is new area of theoretical researches for tasks of the analysis and processing of spatial information. Information from aircraft is processed and analyzed. The allocated objects are distributed on spatial classes in accordance with the classification of spatial information.

A barcode is taken as the qualifier of a class of a spatial object. It is formed by calculation of topological features of the classified object. Set of values of color intensity of all object points is created. Sorting of this set of values according to increase is made. Search of vertices of some intensity is run step by step. It is noted in the list of vertices when finding such point. If this vertice appeared in the neighborhood of Moore of already noted point, then they are connected a line. The triangle is formed at emergence of three such vertices. The number of components (vertices, lines and triangles) at such approach can change on each step of an algorithm. Emergence of vertice adds a component. Emergence of the line connecting different components leads to disappearance of component (two components unite in one). Pass in reversed sequence (on decrease) is the following stage of an algorithm. At the same time the number of holes and their existence time is counted. The hole is formed at emergence of a triangle. The filtration list for holes turns out depending on emergence of new components, their association and other operations. Search of the maximum number of holes and lines is made by the following step. Barcode of the image of an object is calculated on the basis of these numbers (fig. 1).

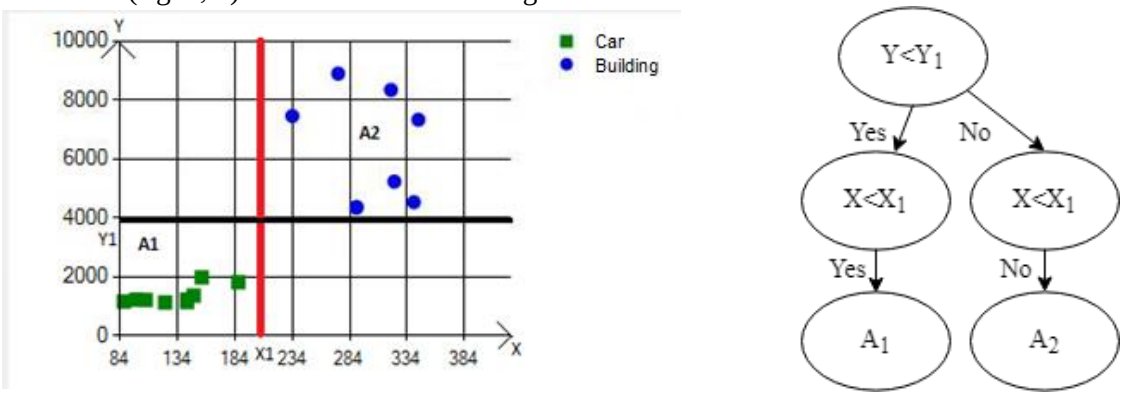


(a) (b)

Figure 1. Barcodes: (a) – the car and (b) – the P-shaped building. The quantity of holes and intensity of points on the color model RGB are displayed on axis X and on axis Y.

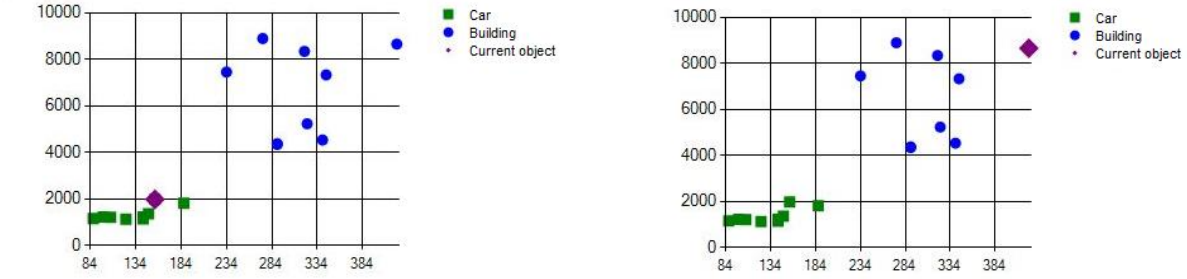
A belonging of object to a spatial class is defined by comparing of barcodes of two objects. Previously training on images of objects of different classes is made (fig. 2, a). Comparing represents check of inclusion of Bettie numbers (the maximum numbers of holes and lines of the image of an object) in the range which characterizes objects of a spatial class (fig.3). The algorithm is complemented with the random forest method [11-13] for optimization of work of an algorithm on

time. It allows to improve speed of an algorithm. The decision tree is formed on the basis of this distribution (fig. 2, b).It is result of work of algorithm.



(a) (b)

Figure 2. (a) Distribution of objects on classes (an algorithm training): A_1 –cars (if $Y < Y_1$and $X<X_1$); A_2 –buildings (if $Y \geq Y_1$and $X \geq X_1$); X – maximum quantity of components; Y – maximum quantity of holes. (b) Decision tree on the basis of this distribution.



(a) (b)

Figure 3. Distribution of objects on classes: (a) – current object is a car and (b) – current object is a building. The maximum quantity of edges and the maximum quantity of holes are displayed on axis X and on axis Y.

The random forest model is applied on the basis of values of Bettie numbers which are taken as features. Random forest is the algorithm consisting of a set of decisive trees. It has been offered by Leo Breiman and Adele Cutler [14]. The algorithm is implemented according to the following scheme.

1. Receiving subselection of the training model. Taking of selection of the training row happens at this stage. The tree is building on its basis.

Basic algorithms have to be unique. Each tree is formed on the training selection for this purpose. There is an element of randomness at the choice of splittings.

The more trees, the quality is better. But time of control and work of Random forest increase in proportion.

2. Viewing of set of random features. It is made for realization of splitting of each branch of a tree.

If the number of features for splitting increases, then time of creation of the forest increases, and trees become "more monotonous".

3. Choice of the best feature and of branch of a tree. Creation of trees is made until disappearance of elements from selection.

Creation of Random forest in direct ratio to selection depth. The selection is deeper, the algorithm is longer executed. The quality on training sharply increases at increase in selection depth. But it usually increases on control selection.

It is recommended to use the maximum depth of trees at realization of this algorithm. Change of the parameters connected with restriction of number of objects doesn't result in significant effect when using superficial trees. Each feature has different degree of importance therefore a part of features can not bear advantages.

3. A research of work of classification algorithm on real spatial objects

The research of the offered algorithm is carried out. Testing of spatial information on the basis of images from aircraft from some height is executed. Classes of spatial objects such as vehicles and buildings are considered. The detailing is made on each class. Vehicles such ascars (are presented on fig.4,a-f), light commercial vehicles and minibuses(fig.4,g-j), buses (fig.4,k-l) are considered. Rectangular, G-shaped, P-shaped, private types of buildings are presented (fig.5).The quantity of the training and test selections in the sum is equal 100 images of objects.

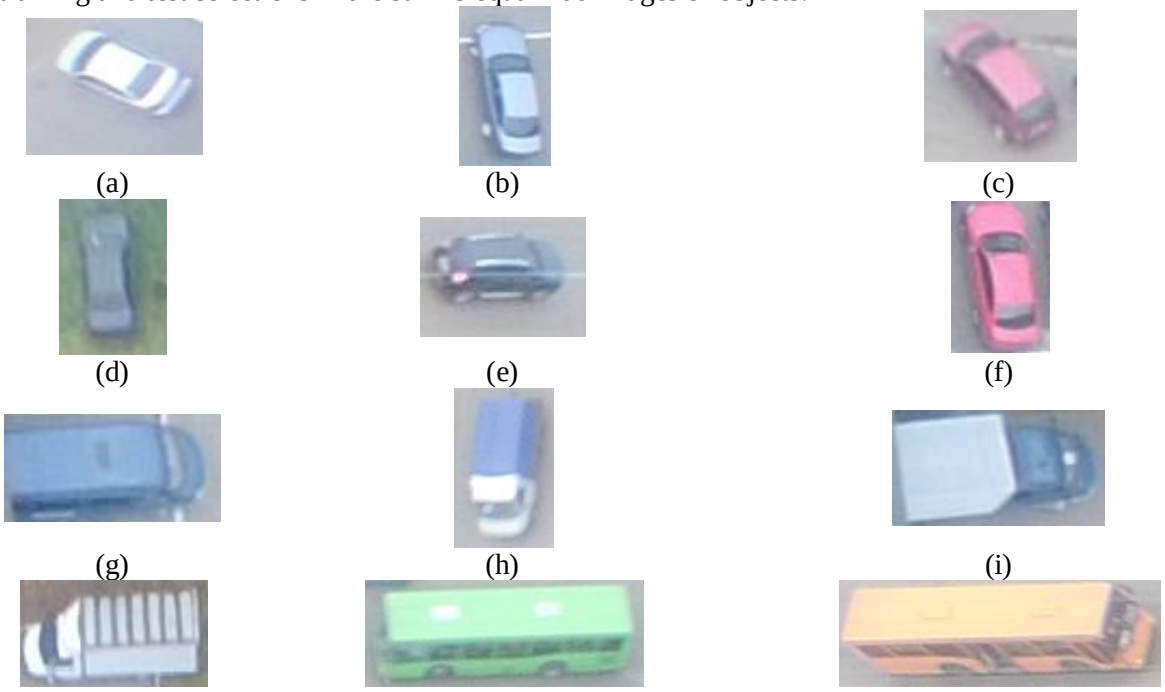


Figure 4. Spatial class of objects "Vehicles". (a-f) – cars; (g-j) – light commercial vehicles and minibuses; (k-l) – buses.

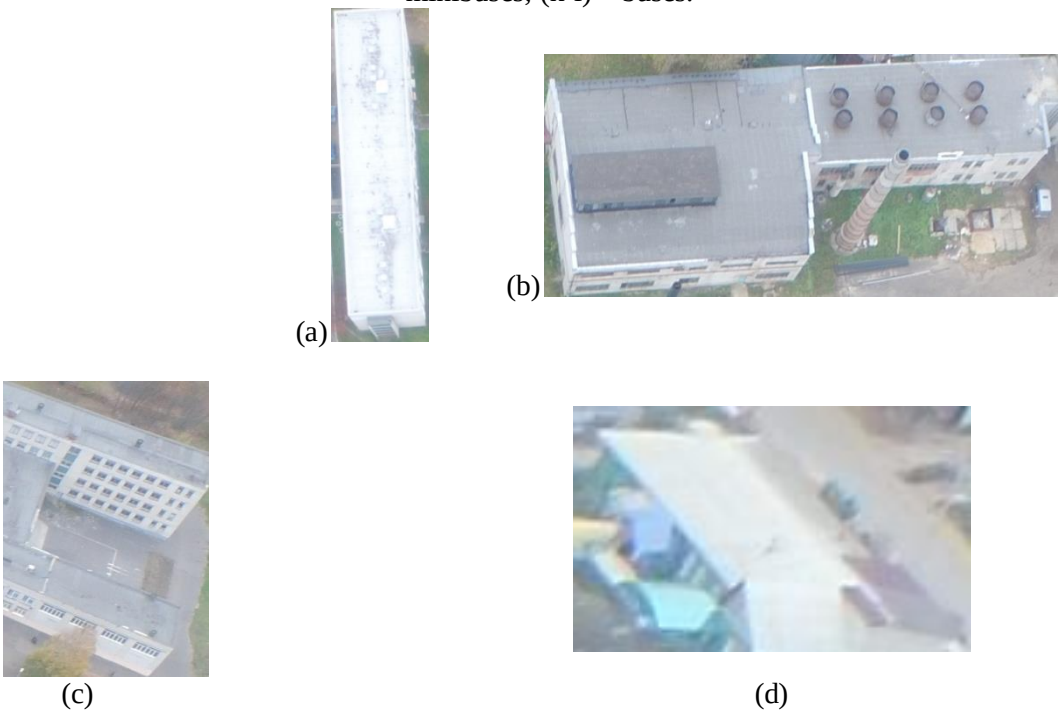


Figure 5. Spatial class of objects of "Building". (a) – rectangular; (b) – G-shaped; (c) – P-shaped; (d) – private houses.

Research results are presented in table 1. The algorithm showed the largest accuracy when determining light commercial vehicles and buses in category of vehicles and for rectangular, G-shaped and P-shaped buildings (100%).

Table 1. Classification of objects by spatial classes.

Object class	Classification accuracy, (%)
Vehicles	98.33
Cars	95.00
Light commercial vehicles and minibuses	100.00
Buses	100.00
Building	97.50
Rectangular;	100.00
G-shaped	100.00
P-shaped	100.00
Private	90.00

Algorithm is invariant to affine transformations. Therefore tests with the different angles of shooting are carried out for the offered types of an object (90˚,180˚,270˚, 360˚– fig.6). Research results of invariancy of an algorithm to affine transformations are presented in table 2.

(a) (b) (c) (d)

Figure 6. A spatial object of the class "Vehicle" in turn on (a) – 90˚; (b) – 180˚; (c) –270˚; (d) – 360˚.

Table 2. Classification of objects by spatial classes with turn of the image.

Object class	Classification accuracy for turn on 90˚, (%)	Classification accuracy for turn on 180˚, (%)	Classification accuracy for turn on 270˚, (%)	Classification accuracy for turn on 360˚, (%)
Vehicles	100.00	100.00	100.00	100.00
Cars	100.00	100.00	100.00	100.00
Light commercial vehicles and minibuses	100.00	100.00	100.00	100.00
Buses	100.00	100.00	100.00	100.00
Building	100.00	100.00	100.00	100.00
Rectangular;	100.00	100.00	100.00	100.00
G-shaped	100.00	100.00	100.00	100.00
P-shaped	100.00	100.00	100.00	100.00
Private	100.00	100.00	100.00	100.00

Also an algorithm is invariant to scale. Therefore tests in images of the same area of different scales are carried out (table 3). The algorithm hasn't made mistake for two subclasses of vehicles (from three) and for two subclasses of buildings (from four).

Table 3. Classification of objects by spatial classes with images of an object at different scales.

Object class	Classification accuracy, (%)
Vehicles	96.67
Cars	90.00
Light commercial vehicles and minibuses	100.00
Buses	100.00
Building	96.25
Rectangular;	100.00
G-shaped	95.00
P-shaped	100.00
Private	90.00

4. Conclusion

Existing approaches of classification of spatial data are considered in article. New approach of realization of an classification algorithm on the basis of topological characteristics of the image is offered. At the heart of an algorithm methods of a persistent homology and the Random Forest method are put. Results of researches have shown expediency of application of the developed classification algorithm of spatial information as it invariant to affine transformations and changes of scale.

A deformation (distortion) of source data is one of problems of map object classification. A stretching of images of spatial objects is one of types of deformation.The solution of this problem is important within the solution of a problem of classification of spatial information and automatic digitization of maps.Algorithm modernization in the sphere of processing of spatial data with deformations of various types is the following step in improvement of work of the offered algorithm.

5. References

[1] Andrianov D, Eremeev S and Kuptsov K 2017 Identification of spatial objects on digital maps *CEUR Workshop Proceedings* **1940** 1-7
[2] Shekhar S and Xiong H 2008 *Digitization of Maps Encyclopedia of GIS* (Springer, Boston, MA)
[3] Fursov V A, Goshin Ye V and Kotov A P 2016 The hybrid CPU/GPU implementation of the computational procedure for digital terrain models generation from satellite images *Computer Optics* **40(5)** 721-728 DOI: 10.18287/2412-6179-2016-40-5-721-728
[4] Vizilter Y V, Gorbatsevich V S, Vorotnikov A V and Kostromov N A 2017 Real-time face identification VIA CNN and boosted hashing forest *Computer Optics* **41(2)** 254-265
[5] Basaraner M and Selcuk M 2008 A structure recognition technique in contextual generalisation of buildings and built-up areas *Cartographic Journal* **45(4)** 274-285
[6] Frohn R C 2006 The use of landscape pattern metrics in remote sensing image classification *International Journal of Remote Sensing* **27(10)** 2025-2032
[7] Grabler F, Agrawala M, Sumner R W and Pauly M 2008 Automatic generation of tourist maps *ACM Transactions on Graphics* **27(3)** 100
[8] Guienko G and Doytsher Y 2003 Geographic information system data for supporting feature extraction from high-resolution aerial and satellite images *Journal of Surveying Engineering* **129(4)** 158-164
[9] Scholzel C A, Hense A, Hubl P, Kuhl N and Litt T 2002 Digitization and geo-referencing of botanical distribution maps *Journal of Biogeography* **29(7)** 851-856

[10] Fedotov N G, Syemov A A and Moiseev A V 2016 Analysis of conditions that influence the properties of the consructed 3D-image features *Computer Optics* **40(6)** 887-894 DOI: 10.18287/2412-6179-2016-40-6-887-894
[11] Abdulsalam H, Skillicorn D B and Martin P 2011 Classification Using Streaming Random Forests *IEEE Transactions on Knowledge and Data Engineering* **23** 22-36
[12] Biau G, Devroye L and Lugosi G 2008 Consistency of random forests and other averaging classifiers *Journal of Machine Learning Research* **9** 2015-2033
[13] Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag) p 746
[14] Breiman L 2001 Random Forests *Machine Learning* **45(1)** 5-32

Acknowledgment
The reported study was funded by RFBR and Vladimir region according to the research project № 17-47-330387. The reported study was funded by Vladimir region according to the research project № 326 of 29.09.2017.