Memoirs of the Scientific Sections of the Romanian Academy
Tome XXXIX, 2016

COMPUTER SCIENCE

BRAIN TUMOR SEGMENTATION BASED ON RANDOM FOREST

LÁSZLÓ LEFKOVITS[1], SZIDÓNIA LEFKOVITS[2] and MIRCEA-FLORIN VAIDA[3]

[1]*Department of Electrical Engineering, Faculty of Technical and Human Sciences, Sapientia University, Tg. Mureş, Romania*
[2]*Department of Informatics, Faculty of Science and Letters "Petru Maior" University, Tg. Mureş, Romania*
[3]*Department of Communications, Technical University of Cluj-Napoca, Romania*
Corresponding author: lefkolaci@ms.sapientia.ro

In this article we present a discriminative model for tumor detection from multimodal MR images. The main part of the model is built around the random forest (RF) classifier. We created an optimization algorithm able to select the important features for reducing the dimensionality of data. This method is also used to find out the training parameters used in the learning phase. The algorithm is based on random feature properties for evaluating the importance of the variable, the evolution of learning errors and the proximities between instances. The detection performances obtained have been compared with the most recent systems, offering similar results.

Keywords: brain tumor segmentation, random forest, feature selection, variable importance.

1. INTRODUCTION

A tumor is a mass of tissue formed by accumulation of malignant cells. The World Health Organization (WHO) defines four grades of tumor: I, II, III and IV. The higher the grade, the more malignant the tumor is. Grade I and II are the least malignant tumors, called low-grade (LG) tumors; grade III and IV are the most malignant tumors called high-grade (HG) tumors [7]. LG tumors are usually benign tumors, but they present the risk of growing into HG tumors. The main goal of the most efficient treatment of brain tumors is early discovery, identification and diagnosis.

In the following we will present the best-performing systems based on a discriminative model used in multimodal MR tumor segmentation. In this overview, we will analyze the number and type of features used and the classification algorithm applied, with the goal of comparing them with our model. The performances of the systems presented and of our own system are compared in the experimental results section.

D. Zikic [20] and his research team from Microsoft created a discriminative model that extracts the attributes from image intensities, as well as from a generative model. In his approach, 2,000 context-aware attributes are defined. As a classification ensemble, they use 40 decision trees, each having a depth of 20.

E. Geremia and the research group from INRIA Sophia-Antipolis, France [5] built a discriminative model that associates a vector of 412 features to each point. The classification algorithm is an ensemble of decision trees trained on a set of images containing 20 HG and 10 LG images. M. Goets [6] also created a discriminative model which does not rely on a generative probabilistic model based on *a priori* information from atlases. This model uses 208 attributes, 52 attributes for each of the 4 image types. The classifier is made up of an ensemble of Extra-Randomized Trees (ERT). S. Reza and K.M. Iftekharuddin [17] created a discriminative model which only processes planar images that are axial sections of 3D MRI, without using *a priori* information about the anatomical structure of the brain. The system works only with the intensity information of the pixels in multimodal images, extracting special attributes based on texton textures and fractal dimension. The classification algorithm is again the RF. The final decision is made by weighed voting. A remarkable performance is obtained from texture information.

2. DISCRIMINATIVE MODELS

Discriminative models create a decision function that describes the input vectors and assigns each vector to a class. The decision functions do not use *a priori* knowledge of the classification domain, instead, they try to build the necessary informational relation based on the training samples. Therefore, the models used in segmentation create the relational space based on the intensity information in labeled images. Segmentation quality depends on the quality of the images and of annotations; however, it is most dependent on the discriminative model created [1].

The general structure of such a model is given in Figure 1. In the following, we will describe the role of each part with regard to tumor segmentation.

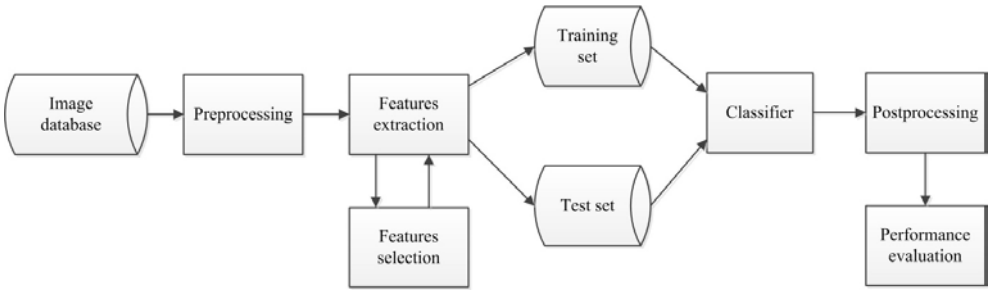


Fig. 1. Discriminative model.

2.1. DATABASE

The database is the first and most important step in the discriminative model, regarding the acquisition of image samples and the corresponding annotations.

The standard image database used for brain tumor segmentation is the BRATS 2013 [12] clinical image database, consisting of 51 high grade (HG) and 14 low grade (LG) cranial MRIs with multiform brain tumors. The images, acquired by specialists with 1.5T and 3T scanners, contain four types of modalities: T1, T1c (with contrast material Gadolinium), T2 and Flair. We have used BRATS annotation with four different tumor structures: the edema, the non-enhancing core, the enhancing core and the necrotic core (Fig. 2).

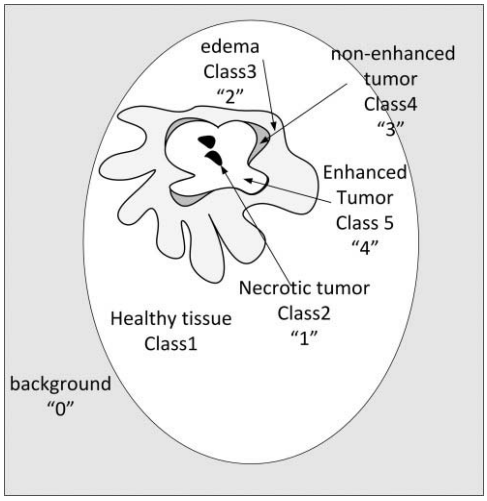


Fig. 2. Tissue annotation.

2.2. PREPROCESSING

Another important step is the necessity of image preprocessing. Preprocessing consists of noise filtering and standardization of luminosity and contrast; this means standardization of image pixel intensities. The preprocessing steps used for the images included in the training set have to be subsequently applied for every further application built on the model created.

MRI acquisition is associated with many artifacts. Some of them can be eliminated by the medical staff, by setting the acquisition parameters accordingly. The images acquired are sufficiently appropriate for human visual analysis, but the main problem is that automatic segmentation is significantly influenced by these artifacts.

In our work we have analyzed three important artifacts: inhomogeneity, noise and intensity nonstandardness.

In our previous work [10] we evaluated three inhomogeneity reduction methods. Each of these filters has its own advantages and disadvantages. The best-performing and most accepted algorithm is N4 filtering [19]. For inhomogeneity reduction in MR images, we have applied the N4 filter implemented in the ITK package [8].

The MRI image contains a significant amount of noise caused by human interaction with the equipment, equipment parameters and environmental changes. The denoising method improves image quality by reducing the noise component while preserving the quality of the image. We decided to use anisotropic diffusion filtering, implemented in the ITK package [8].

In MRI, a weighed image is acquired, in which tissues are distinguished based on their intensities. However, these intensity values do not have the same meaning in different images, even if the body and the protocol is the same. In segmentation and quantification, the absence of context meanings causes different problems. The task of MRI standardization is to unify the meanings of intensities. The aim is to transform the histogram in order to match it to a predetermined shape [14].

In preprocessing, we have to eliminate these three artifacts. The order of their application to a MRI was studied by D. Palumbo in [15]; he states that the adequate sequence for the best segmentation is the following: bias field correction, then noise filtering, and finally, intensity standardization.

2.3. FEATURE EXTRACTION

Image processing offers many procedures for the extraction of characteristics from images. In the field of tumor segmentation there are many studies trying to find certain characteristics with a high correlation to the appearance of the brain tumor in MRI. Despite these research efforts, no proper feature sets have been found yet. That is the reason for using a large feature set, with the features having little correlation to the goal of classification. At first, our approach defines a smaller feature set, but this is later enlarged for increasing classification performance. For each feature we defined many low-level characteristics [13] that describe the intensities in the neighborhood of the voxels studied. In our application we have used the following features:

– first order operators (mean, standard deviation, max, min, median, Sobel, gradient)

– higher order operators (Laplacian, Difference of Gaussians, entropy, curvatures, kurtosis, skewness)

– texture features (Gabor filter)

– spatial context features.

By extracting all these features for every voxel in all modalities, we transform the image segmentation task into a statistical pattern recognition problem. The segmentation process obtained with this statistical model also requires analysis on the importance of the variable. An appropriate selection of attributes has to be done according to the given objects.

2.4. CLASSIFIER

The classifier is the main part of a statistical pattern recognition system. In the field of data mining there are many well-known classification algorithms, such

as Naïve Bayes, C4.5 tree, k-NN, k-means, Neural Networks, SVM, AdaBoost, Random Forest (RF). The most important classifiers used in this field have been implemented in the WEKA Data Mining Toolkit [9]. Using this toolkit, we have compared several classifiers and have chosen to use RF for our application. The most important advantages of RF are:

– high accuracy
– easy handling of large databases
– estimating variable importance
– computing the proximities between instances
– generating the error as forest building progresses.

The RF classifier was introduced by L. Breiman [2]. This classifier builds a large collection of binary decision trees based on two random processes. First, the training set is randomly sampled with replacement for obtaining the bootstrap set. The second randomization is introduced in the building process of trees. In each node only a small part, the randomly chosen features, are used to search for the best split.

The training set and bootstrap set have the same size, N and, accordingly, the bootstrap set contains an instance of the training set that is different by approximately 2/3, while the rest is made up of repeated samples. Approximately 1/3 of the training samples are left out from the bootstrap set. These instances form the out-of-bag (OOB) set. Thus, every tree is grown on its own bootstrap set and tested on its OOB set. The overall OOB error is the average classification error on the OOB sets over all the trees in the forest. Breiman [2] shows that the upper bound for the generalization error is given by:

$$GE = \rho\left(\frac{1}{S^2} - 1\right) \tag{1}$$

ρ – mean value of correlation, S – strength of the ensemble. In order to reduce the error, the correlation should be decreased and the strengths increased. An interesting characteristic of RF is the general error (GE), which can be estimated *via* the OOB error.

In order to maintain the OOB error under a given value, we have to optimize the following three RF parameters: the number of trees K, the number of randomly selected features m and the number of nodes T in each tree [4].

For classification purposes, each tree produces a decision individually, after which the decisions of all trees of the ensemble are combined to generate a decision by [18]:

– vote of majority, where all trees are equal
– weighed sum of trees, weights depending on the individual training error
– Bayesian combination scheme, where weights are considered as approximations of prior probabilities of classifiers.

2.5. FEATURE SELECTION

In the field of image segmentation, discriminative classifiers are based on several local image features. Most authors create their model by using the features without any selection criteria, based on their intuition and/or previous experience. A more reliable model can be built by using a framework that selects the importance of variables from the point of view of classification.

In the construction of RF classifiers there are two possibilities to evaluate variable importance. The first is related to Gini importance, while the second is computed based on the variable permutation error. Because the two variable importance values depend on the forest structure, the values obtained are somewhat random. More confidence is given to the order of variable importance. This order is determined by the two measurements in a very consistent way (Fig. 3). By applying this conclusion, we can easily eliminate a part of the low-importance variables obtained by both measurements [3].

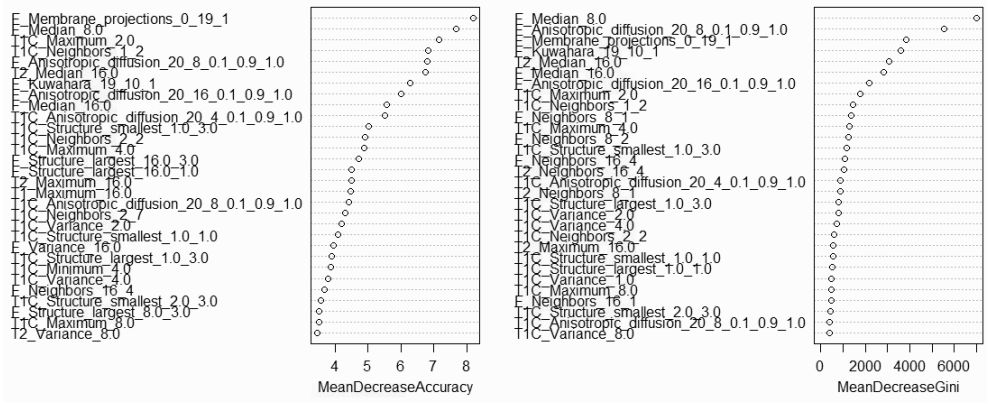


Fig. 3. Variable importance.

One main objective of variable selection is to find a small number of variables appropriate for a good prediction. For this task, we propose our feature selection algorithm, presented in detail in [11]. The main idea of the algorithm is to evaluate variable importance on a randomly chosen part of training set several times. It builds the cumulative order of variable importance and eliminates a significant portion of the most unimportant variables. The algorithm monitors the evolution of the OOB error and the selection of important variables stops when the OOB error reaches the given limit.

All these measurements and the RF classifier are in strong correlation with the given training set. The number of variables used in the end can be determined by evaluating classifier performances on a not previously used test database.

2.6. POST-PROCESSING

Given that the discriminative models use no context information or *a priori* information about the classified object, the results obtained have to be refined by post-processing. In this step, anatomic brain structure information from digital atlases can help to improve classification performances. This task cannot be accomplished without registering the atlas to the brain image analyzed. The digital atlas contains information about healthy brains. Registration in atlases is difficult because of the large variety of tumor appearance in real images.

In order to eliminate misclassified voxels, the following spatial relation between classes can be used: class2 $\subset$ class5 $\subset$ class4 $\subset$ class3 $\subset$ class1 $\subset$ class0. In this context, we can define morphological filters or apply the neighborhood condition provided by Markov Random Fields or Conditional Random Fields.

2.7. EVALUATION

The final step of segmentation is evaluation of the results obtained. Segmentation can be assessed by the following coefficients: Dice coefficient, Jaccard similarity, precision, sensitivity, specificity [16]. All these coefficients can be evaluated from the confusion matrix provided by the classifier. One of their drawbacks is the lack of spatial relation of misclassified voxels to their class. Evaluation can be done using the Hausdorff distance, which is the supremum of the minimum distance between real and segmented surfaces. A more realistic, but subjective evaluation can be done by visual inspection of results.

3. EXPERIMENTAL RESULTS

The distribution of the four analyzed tissues in one brain, from our experiments, is: healthy tissue: 1,500,000 voxels, edema tissue: about 100,000 voxels, tumor core: 50,000 voxels, non-enhancing core: 3,000 voxels and necrotic core: 500 voxels.

In the first step, we extracted 500 image features of each modality, a feature vector with 2,000 elements being thus obtained. A 3D brain image (240× 240×150) contains about 1,500,000 voxels and each voxel contains a vector of 2,000 features; with dimensions of 1,500,000×2,000, the space requirement for it will be 12 GB of memory. After balancing the database, the memory requirement for one image set will be of 2.4 GB. For analyzing all variations of brain tumor appearance, more cases should be considered. In the first step, we used 20 brain image sets for training, with memory requirements of 48 GB. This is too large to build the random forest classifier (on our hardware and software). Using the algorithm proposed in [11], we reduce the number of features to 100. On the first run, almost half of the noisy features were dropped, while, along the further 7 steps, the feature set was

reduced to 100, maintaining the OOB error under <10%. With this reduced feature set, the training database occupies 2.4 GB (for all voxels of 20 multimodal brain images). By randomly sampling 5:1 of this database, we obtain 480 MB for the final training set, which allows building and optimization of the RF classifier for segmentation tasks. The 2.4 GB training database is used in the final optimization phase, which consists of tuning the three important parameters of the RF classifier: K – number of trees, m – number of tries and T – number of nodes in each tree.

By accepting the 100 features for each voxel, a new brain will need about 600 MB of memory for the test set. In the test phase, each instance is sequentially classified by the RF classifier, the result being a 3D image with the voxels classified. In the post-processing and evaluation steps, only the classified image obtained was used.

Tables 1 to 3 list the performances obtained on the training set, whereas Tables 4 to 6 provide the performances for the previously unseen image set. Here, we used the following classes, which better present the clinical application task:

– whole tumor – contains all tumor structures

– tumor core – contains the entire tumor except the edema

– active tumor, which is the enhanced core.

The same results are presented graphically on a brain slice of the training set in Figure 4 and of the unseen set (Fig. 5). The black line is the contour of annotation. The light gray region is the detection of the edema and the white region is the result for the tumor core (containing the necrotic tissue – dark gray). The segmentation obtained shows performances comparable to the state-of-the-art systems (Table 7) [12].

We can assert that segmentation shows good results and detects well the edema and tumor zones. The first observation is that many errors occur at the delimitation surfaces between tissues. In order to avoid this shortcoming, we propose to use another segmentation method, that can delineate these tissues more accurately. Level set segmentation, which is highly sensitive to initialization, could be applied, the results obtained being surprisingly good. In a future work we will propose to apply this method by using the initial contour obtained by our statistical segmentation presented in this article. The second observation is that only two classes, the non-enhanced tumor and the necrotic tumor, are barely detected. To overcome this deficiency, we propose a hierarchical classification in which these classes are detected separately, after the classification of edema and core tumor voxels. This two-step detection is possible because non-enhanced and necrotic tumor tissues are effectively included in the edema and tumor region, respectively. The non-enhanced tissue is located between the edema, while the tumor and the necrotic tumor occur inside the tumor zone.

Table 1
Confusion matrix on training image

	Normal	Non-enh.	Edema	Necrotic	Enhanced
Normal	28,719	22	744	55	90
Non-enh.	19	1,313	2,117	73	1,003
Edema	1,104	254	22,014	191	503
Necrotic	390	63	1,116	651	553
Enhanced	263	545	4,334	252	2,752

Table 2
Detection of whole tumor on training image

	Normal	Whole tumor	Sensitivity	Precision	Dice
Normal	28,719	911	0.969	0.942	
Whole tumor	1,776	37,734	0.955	0.976	0.966

Table 3
Detection of tumor core edema on training image

	Normal	Edema	Whole tumor	Precision	Dice
Normal	28,719	744	167	0.942	
Edema	1,104	22,014	948	0.726	0.809
Tumor core	672	7,567	7,205	0.866	0.606

Table 4
Confusion matrix on a unseen test image

	Normal	Non-enh.	Edema	Necrotic	Enhanced
Normal	1,462,136	951	43,306	0	1,424
Non-enh.	0	4	199	0	1,049
Edema	8,920	208	131,674	0	7,342
Necrotic	108	78	1,860	0	1,116
Enhanced	545	262	13,303	0	7,140

Table 5
Detection of whole tumor on an unseen test image

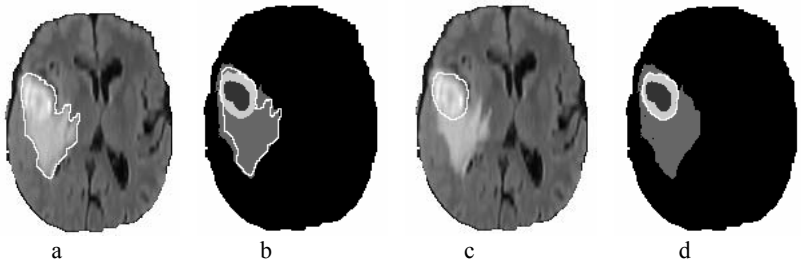
	Normal	Whole tumor	Sensitivity	Precision	Dice
Normal	1,462,136	45,681	0.970	0.993	
Whole tumor	9,573	164,235	0.945	0.782	0.856

Table 6
Detection of tumor core and edema on an unseen test image

	Normal	Edema	Whole tumor	Precision	Dice
Normal	1,462,136	43,306	2,375	0.993	
Edema	8,920	131,674	7,550	0.692	0.778
Whole tumor	653	15,362	9,649	0.493	0.427

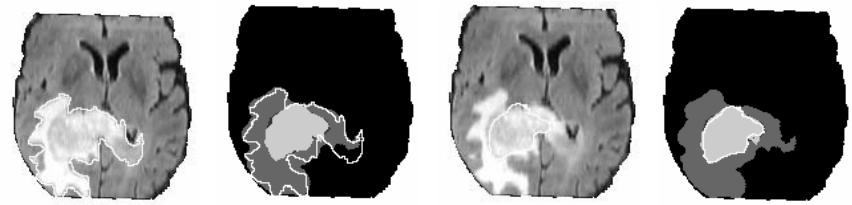
Table 7
Compared Dice indexes

	Our classifier	Brats2012 [12]	Brats 2013 [12]
Whole tumor HG	75–86	63–78	71–87
Core tumor HG	71–82	24–37	66–78



a b c d

Fig. 4. Segmentation results on training images.



a b c d

Fig. 5. Segmentation results on unseen test images.

4. CONCLUSIONS

The classification algorithm is the main part of the discriminative model. It has to be adapted to the purpose of classification and correlated with the available attribute in the training, as well as with the test phases.

The disadvantages of discriminative models are: the extensive training set, which has to cover, if possible, all different aspects of the detectable object; very time-consuming image annotation process for experts; uniform preprocessing of images; necessity of post-processing.

Each part of the model can influence the performances of final segmentation. The most critical parts are the preprocessing steps: inhomogeneity correction, intensity standardization and filtering. Another important part is the selection of adequate features defined for each voxel. Within this framework, we have proposed a feature selection algorithm that can evaluate the importance of new feature sets by comparing them with the existing ones. By this method, we can find the best feature set for the proposed task. In our opinion, the database used considerably limits segmentation performance. Furthermore, the system can be optimized with regard to processing time and efficient memory usage. All these ideas could constitute a meaningful foundation for future research.

Authors contributions: László Lefkovits (first author) – 60%; Szidónia Lefkovits (second author) – 25%; Mirces-Florin Vaida (third author) – 15%.

R E F E R E N C E S

1. BISHOP C.M. *Pattern Recognition and Machine Learning*, Springer, 2006.
2. BREIMAN L., *Random forests*, Machine learning, 2001, **45** (1), 5–32.
3. DÍAZ-URIARTE R., DE ANDRES S.A., *Gene selection and classification of microarray data using random forest.*, BMC bioinformatics, 2006, **7** (1), 1–13.
4. GENUER R., POGGI J. M. TULEAU-MALOT C. *Variable selection using random forests.* Pattern Recognition Letters, 2010, **31** (14), 2225–2236.
5. GEREMIA E., MENZE B.H., AYACHE N. *Spatial decision forests for glioma segmentation in multi-channel mr images*, MICCAI-BRATS Challenge on Multimodal Brain Tumor Segmentation, 2012.
6. GOETZ M., WEBER C., BLOECHER J., STIELTJES B., MEINZER H.-P., MAIER-HEIN M. *Extremely randomized trees based brain tumor segmentation*, MICCAI-BRATS Challenge on Multimodal Brain Tumor Segmentation, 2014.
7. http://www.who.int/en/ (Accessed March 2016).
8. http://www.itk.org/ (Accessed March 2016).
9. http://www.cs.waikato.ac.nz/ml/weka/ (Accessed March 2016).
10. LEFKOVITS L., LEFKOVITS SZ., VAIDA M., *An Atlas Based Performance Evaluation of Inhomogeneity Correcting Effects*, The 5th Int. Conf. on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics, 2015.
11. LEFKOVITS L., LEFKOVITS SZ., EMERICH S., VAIDA M., *Random Forest Feature Selection Approach for Image Segmentation*, under review Eusipco 2016.
12. MENZE B.H., JAKAB A., BAUER S., KALPATHY-CRAMER J., FARAHANI K., KIRBY J., *et al*. *The multimodal brain tumor image segmentation benchmark (BRATS).* Medical Imaging, IEEE Transactions on, 2015, **34** (10), 1993–2024.
13. NIXON M. S., AGUADO A. S., "Feature extraction & image processing for computer vision", Academic Press, 2012.
14. NYÚL L.G., UDUPA J.K., ZHANG X.. *New variants of a method of mri scale Standardization*. Medical Imaging, IEEE Transactions on, 2000, **19** (2), 143–150.
15. PALUMBO D., YEE B., O'DEA P., LEEDY S., VISWANATH S., MADABHUSHI A., *Interplay between Bias Field Correction, Intensity Standardization, and Noise Filtering for T2-weighted MRI*", 33rd Annual Int. Conf. of the IEEE EMBS, 2011.
16. POWERS D.M. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* Journal of Machine Learning Technologies, 2011, **2 (**1), 37–63.
17. REZA S., IFTEKHARUDDIN K. M. *Multi-class abnormal brain tissue segmentation using texture features*, MICCAI-BRATS Challenge on Multimodal Brain Tumor Segmentation, 2013.
18. SAZONAU V. *Implementation and evaluation of a random forest machine learning algorithm*, University of Manchester, 2012, 9.
19. TUSTISON N.J., AVANTS B.B., COOK P.A., ZHENG Y., EGAN A., YUSHKEVICH P.A., GEE J.C. *N4ITK: improved N3 bias correction*. Medical Imaging, IEEE Transactions, 2010, **29** (6), 1310–1320.
20. ZIKIC D., GLOCKER B., KONUKOGLU E., SHOTTON J., CRIMINISI A., YE D., *et al*. *Context-sensitive classification forests for segmentation of brain tumor tissues*, MICCAI-BRATS Challenge on Multimodal Brain Tumor Segmentation, 2012.

Received February 18, 2016