# MULTIPLE EXPERTS IMAGE SEGMENTATION FOR OBJECT DETECTION

*Bogdan-Cristian TALOI*[1]
*Alin-Gabriel DUMITRU*[2]
*Patricia-Steliana PENARIU*[3]
*Costin-Anton BOIANGIU*[4]

**Abstract**: *This paper explores ways of creating an image segmentation system based on voting. Using segmentation techniques ranging from Artificial Intelligence algorithms such as Mask-RCNN to classical approaches like Mean Shift Clustering and other custom-designed segmentation techniques, the purpose of this paper is to create a mixed algorithm for image segmentation and object detection. The proposed method accomplishes the task by generating good results that compare favorably to four other related stand-alone image segmentation methods.*

**Keywords:** *object detection, image segmentation, voting technology, Mask R-CNN, Mean Shift Clustering, K-Means Color Clustering, Edge-Based Region Growing Segmentation.*

## 1. Introduction

Recent history brought us essential advancements in the domain of image segmentation and object detection. These advancements are conducted by the favorable outcome of region proposal methods, as well as region-based convolutional neural networks (R-CNNs). Even though R-CNNs, introduced by R. Girshick in [1], were computationally expensive in their original implementation, the latest research [2][3] achieves a significant cost reduction due to sharing convolutions across proposals [4]. R-CNN method implies deep convolutional

[1]Engineer, Computer Science and Engineering Department, Faculty of Automatic Control and Computers, Politehnica University of Bucharest, Splaiul Independentei 313, Bucharest 060042, Romania, bogdan.taloi@stud.fils.upb.ro
[2]Engineer, Computer Science and Engineering Department, Faculty of Automatic Control and Computers, Politehnica University of Bucharest, Splaiul Independentei 313, Bucharest 060042, Romania, alin.dumitru@stud.fils.upb.ro
[3]PhD Student, Eng., Computer Science and Engineering Department, Faculty of Automatic Control and Computers, Politehnica University of Bucharest, Splaiul Independentei 313, Bucharest 060042, Romania, patricia.penariu@stud.acs.upb.ro, patriciapenariu@gmail.com
[4]Professor, PhD Eng., Computer Science and Engineering Department, Faculty of Automatic Control and Computers, Politehnica University of Bucharest, Splaiul Independentei 313, Bucharest 060042, Romania, costin.boiangiu@cs.pub.ro

neural networks (CNNs) to pair object proposals in order to obtain consistent object detection accuracy [2]. Compared to standard neural networks with similarly-sized layers, CNNs can be manipulated in training more quickly as they have fewer connections and parameters (the capacity of CNNs is managed by varying their depth and breadth) [5]. Despite these advantages, training is costly, and there is no positive impact on the speed of the object detection (R-CNN method performs a deep CNN forward pass for each object proposal, with no computing shared). Typically, the object detection task using this method runs in 47 s/image using a GPU [2].

A way to improve R-CNN's performance consists of using spatial pyramid pooling (SPP) networks [3]. SPP network, also called SPM or spatial pyramid matching [6][7] is an addition of Bag of Words model [8] and can reduce the training time of R-CNN method by 3 times and speed up test time by 10 to 100 times [2].

The input data of the Fast R-CNN network [2] is represented by an image and a group of object proposals. The first step is to create a convolutional feature map, based on the CNN image, considered as the entry date. The next step is identifying the region of proposals from the feature map, by Fast R-CNN, and transforms them into squares, which previously it reshaped them into a fixed size, using a Region of Interest (RoI) pooling layer, in order to be an integral part of the entire connected layer. The process continues by identifying the class of the proposed region accompanied by the bounding box offsets as can be observed in Figure 1. Even if Fast R-CNN brought many improvements, it still suffered from a couple of crucial drawbacks and limitations: the regions are still proposed by selective search, making up most of the algorithm's running time, and the process of learning was not end-to-end. The network was trainable after the selective search provided its output; however, the selective search algorithm could not be influenced to improve its predictions.

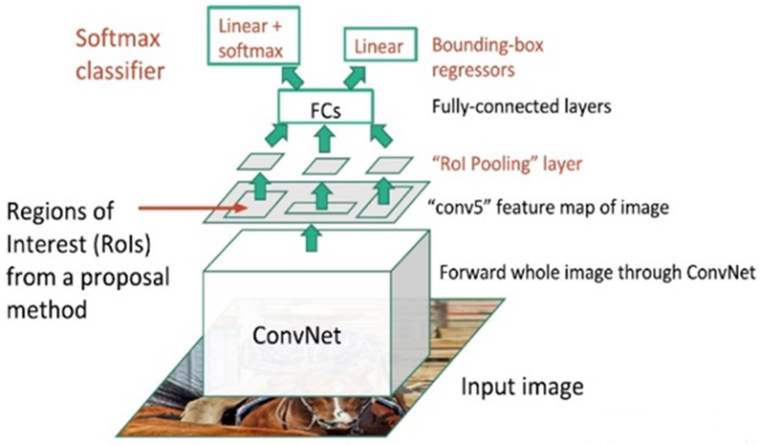


Fig.1. Fast R-CNN architecture; image taken from [2].

A more recent iteration of this approach, Faster R-CNN, achieves more accurate results than the Fast R-CNN method, S. Ren et al. introducing in [4] Region Proposal Network (RPN). This approach allows showing the convolutional features of the complete image among the detection network, the benefit consisting of the permissiveness for region proposals almost without costs. The ability to predict object bounds and unnecessary scores for each position is also a valuable asset to consider, RPN having the capacity to make proposals for quality regions in an end-to-end manner, which is an advantage in using Fast R-CNN for detection. The CNN layers provide a convolutional feature map on an image, furnished as the input data, in a similar way to Fast R-CNN. The innovative breakthrough is that the region proposals are predicted using a distinct network that has a classifier and a regressor. The schematic representation of the construction of the Faster R-CNN model is revealed in figure 2.

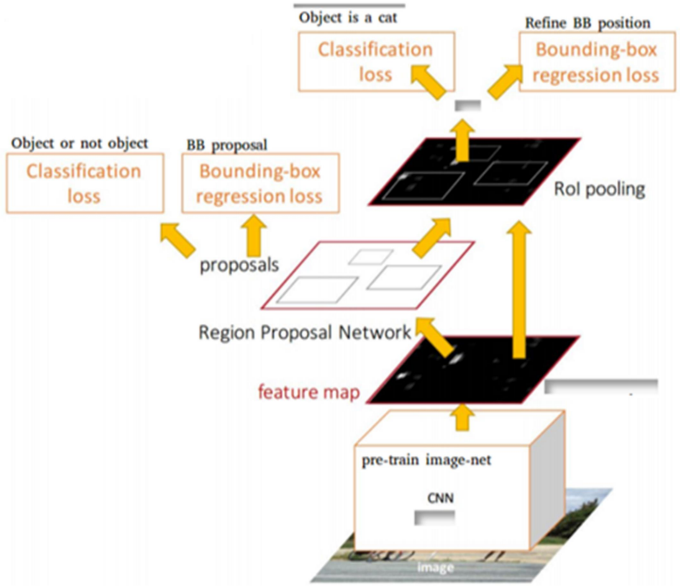


Fig.2. Faster R-CNN layout; image taken from [4].

This network works with the concept of anchors (shown in figure 3). An anchor is the central point of a sliding window. Anchor boxes are fixed-sized boundary boxes with different sizes and ratios that are placed throughout the image [4]. For every anchor, RPN anticipates the following: the probability that an anchor is an object versus the anchor is representing a background area and the bounding box regressor for adjusting the anchors to suit to the object. The classifier then determines the probability of a proposal being the target object. Regression generates the coordinates of the proposals. The remodeling of the envisaged region proposals is done using an RoI collection layer, useful in both predicting the offsets and in the proposed region's image classification.

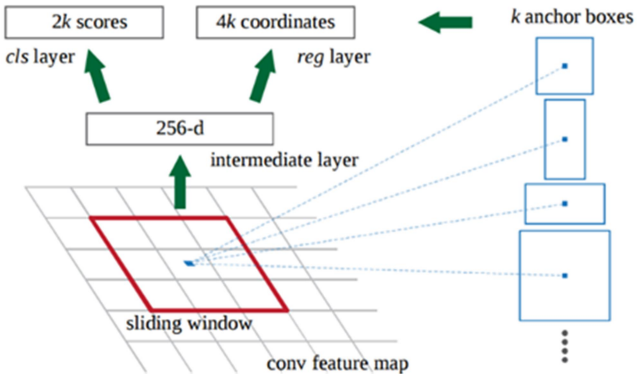


Fig. 3. Anchors representation in the Faster R-CNN model; image taken from [4].

The latest further iteration for object detection is Mask R-CNN [9], its architecture is illustrated in Fig.4. Mask R-CNN (Fig.4) is a framework made up of two steps. Firstly, the input image and the output proposals are processed (areas likely to include an object). Secondly, the proposals are classified, and the bounding boxes and masks are generated. Because Faster R-CNN works very well in tasks like image classification [5][10] and object detection [1][11][12], this method can be adapted to run pixel segmentation. Mask R-CNN accomplishes this by implementing a complementary addition (branch) to Faster R-CNN that outputs a binary mask that concludes if a given pixel represents a part of an object or not, the branch is represented by a fully CNN superimposed over a CNN based feature map.

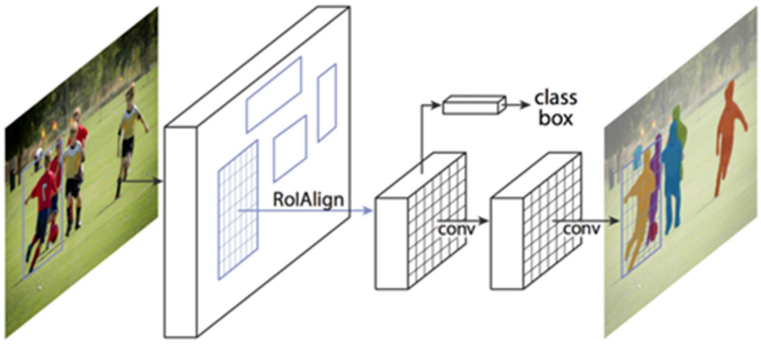


Fig. 4. Mask R-CNN architecture; image taken from [9].

### *A. Problem motivation*

The presented algorithms and, in general, most algorithms involved in segmentation and object detection, are self-standing algorithms, based on specific techniques or heuristics. There has been little research exploring the usage of multiple such algorithms in conjunction.

This paper aims to explore a voting-based system for image segmentation, to improve both segmentation results and object detection results by correlating the results of multiple neural networks, classical and custom algorithms. This concept is similar to the process of boosting in general Machine Learning (ML).

While image segmentation through voting has been approached before [13], algorithms based on neural networks have not been included in the plethora of approaches that were discussed. This study hypothesizes that using such ML models can further improve the results of voting-based image segmentation, while further contributing to the possibility of achieving object detection at the same time.

## 2. Proposed method

### *A. Selected segmentation algorithms*

For the proposed system, we handpicked several different methods for image segmentation, the first of them being Mask R-CNN [9], which extends Faster R-CNN [4] to pixel-level image segmentation. The main point in this network is to segment the classification and the pixel-level mask prediction tasks. There is a qualitative delimitation between the pixel-level segmentation and the bounding boxes, the first-mentioned presenting a more fine-grained alignment. The role of the R-CNN mask is to stimulate the RoI pooling layer (also called RoI align layer), having the effect of improving the region of interest and increasing the specificity of the mapping to the regions of the original image. For this voting system, there will make use of the head of Mask R-CNN that is responsible for the output of the masks in an image.

The second segmentation technique used in the system is Mean Shift Clustering [14]. Mean Shift is a hierarchical clustering algorithm that attempts to group data in an unsupervised manner. As opposed to K-Means, Mean Shift does not require the number of categories (clusters) to be known beforehand. It consists in several simple steps: the user defines a window (bandwidth of the kernel), and the algorithm places the window on a data point, it calculates the mean for all the points in the window, it moves the center of the window to the location of the mean, and then repeats the 2$^{nd}$ and the 3$^{rd}$ steps until convergence.

The third segmentation algorithm chosen for the proposed system is Color Clustering based on K-Means [15]. K-Means gathers among the most votes in terms of popularity and, at the same time, is the simplest from the class of unsupervised learning. For all the data points scattered in an n-dimensional space, K-Means groups the data points with some similarities in clusters. After randomly

initializing *k* cluster centroids, the algorithm performs two steps iteratively, in an Expectation-Maximization fashion. First, cluster assignment (each data point is assigned a cluster based on its distance from the cluster centroid), and secondly, it moves the centroids (the mean of all the points of a cluster is calculated and cluster centroid is relocated to the mean location). Based on the new centroid locations, the data points are reassigned to the clusters. After a certain number of iterations, if the centroids do not change positions any further, and also if the data points are not reassigned to the clusters, it is considered that the algorithm converged. In this case, the number of clusters specified at the beginning will correspond to the number of colors set to extract/cluster in the images. This color-based clustering approach creates *k* clusters, which are further split into connected components.

The fourth algorithm used here for segmentation is Edge-Based Region Growing Segmentation [16]. Among the best-known techniques used for image segmentation are edge detection and region growing. Edge-based approaches are proven to significantly reduce meaningless information while retaining the relevant information and properties coming from the structure of an image. Sometimes, region growing is often chosen on behalf of edge detection methods because it works better in cases that involve contrast issues and deals adequately with connectivity problems that edge detectors face. Recent studies provided that a combination of the two, region growing and edge detection, can lead to much better results. Some studies have highlighted, in the region growing approach, the importance of the similarity of the pixel or the neighborhood in making region joining decisions. As well as the fact that these decisions involve already extracted edges that complement each other. Also, in the case of two adjacent regions that could merge, the strength of the boundary is taken into account, as follows: a strong boundary induces the maintenance of the regions separately, while a weak boundary allows their combination.

### *B. Voting technique*

The voting technique chosen for this paper is similar to an existing method [13]. The proposed method uses a breadth-first search (BFS) to spread each segment of the image based on voting results. The voting phase consists of each algorithm deciding if the new pixel should or should not be part of the segment currently being expanded. This is done by considering whether the current pixel is in the same segment or different segments in the output of each algorithm.

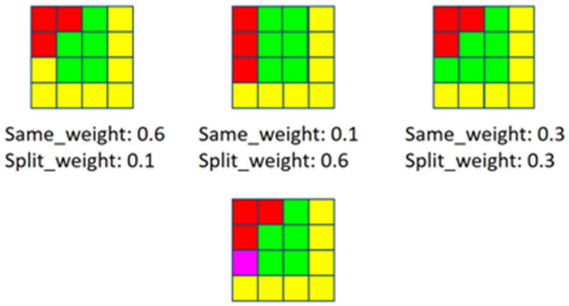


Fig.5.The voting algorithm (top: an over-segmenting algorithm, an under-segmenting algorithm, and a balanced algorithm; bottom: the voting result).

Usually, a weighted voting algorithm will assign a weight to each algorithm, which will decide how each algorithm influences the vote. The proposed method uses two sets of weights: "same" and "split". If one algorithm decides that the current pixel is in the same segment as the previous one, it contributes with its "same" weight. Otherwise, it contributes to the "split" weight (Fig.5). This modification was made because some algorithms are more credible when placing two pixels in the same segment, and other algorithms are better at splitting the pixels of different clusters. For example, if an algorithm that segments based on color clustering decides that two pixels are in different clusters, the voting should only be affected by a small margin as this algorithm is proven to create over-segmentation. Meanwhile, if this method places two pixels in the same segment, it is much more likely that these pixels belong together. Small clusters where the voting algorithm is unable to come to a decision are cleaned similarly to the outputs of the initial algorithms.

## 3. Results

The proposed method achieves successful object detection with excellent results by using different methods combined (Mask R-CNN, Mean Shift Clustering, Color Clustering based on K-Means and Edge-Based Region Growing Segmentation). The results of each algorithm, in parallel with the clean versions of the segmentation, are presented in figure 6 where on the first row there are found Mask R-CNN and Mean Shift, on the second row Edge-Based Region Growing and K-Means with $k$ = 3, followed by K-Means with $k$ = 4,5 on the third row. The output from each algorithm is cleaned, retaining only components having at least a minimum size. Some results of the proposed voting method and their corresponding original images are illustrated in figures 7 to 9. Therefore, the quality of the images is improved using the proposed method, compared to the images resulting from the application of the algorithms individually.

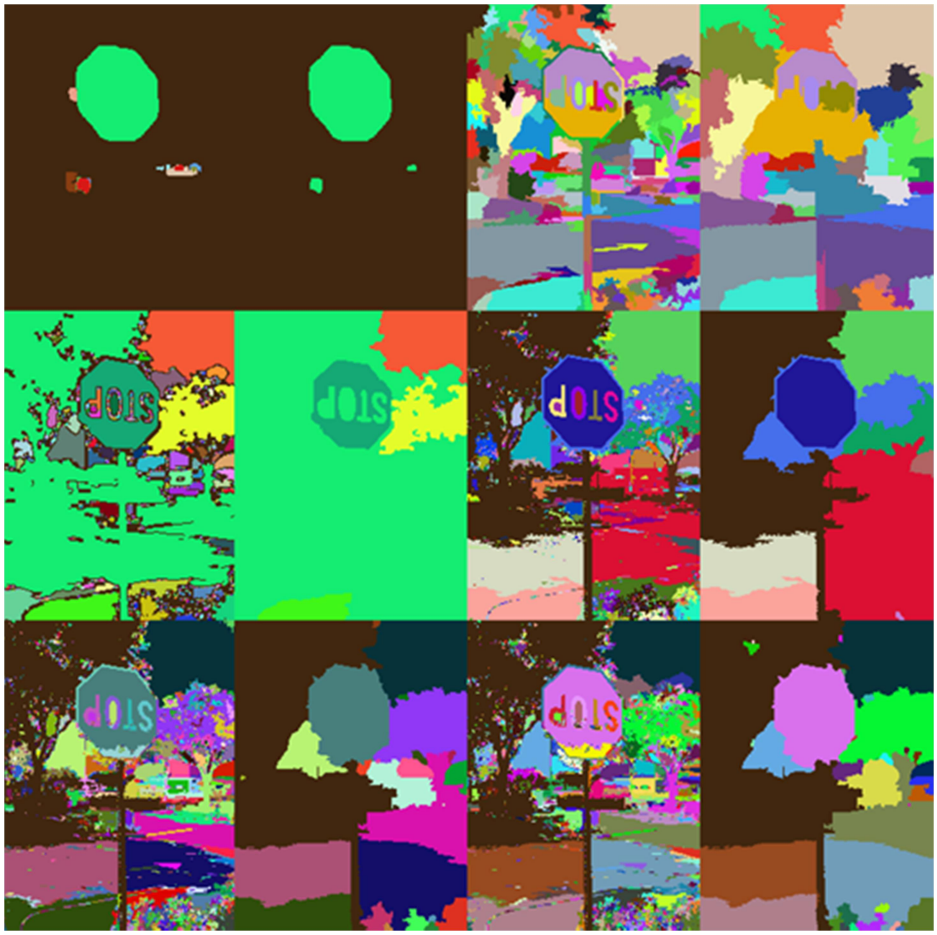


Fig. 6. Results of each algorithm, in parallel with the clean versions of the segmentation (Top: Mask R-CNN, Mean Shift; Middle: Edge-Based Region Growing, K-Means with $k = 3$; Bottom: K-Means with $k = 4, 5$).

Fig. 7. Bedroom results of the proposed method.

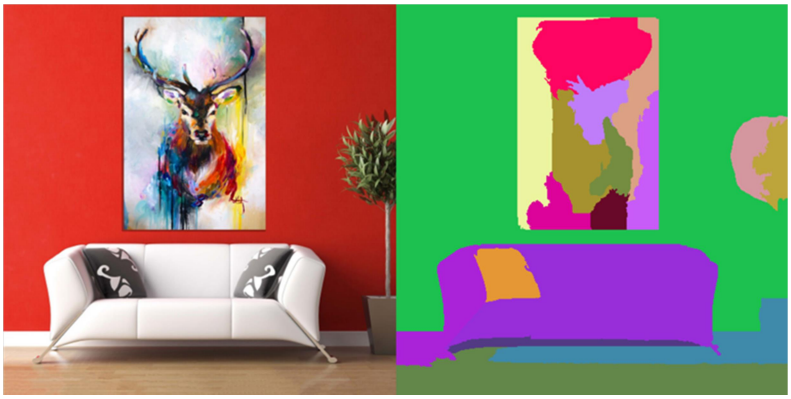

Fig. 8. Living room results of the proposed method.


Fig.9. STOP sign results of the proposed method.

## 4. Conclusion

While image segmentation has often been tackled through neural networks recently, classic methods are still relevant in some situations. If it is possible to combine the results of both these approaches successfully, there can be found ways for them to complement each other's deficiencies. This can be achieved through the process of voting. The proposed variant still supports improvements, the results obtained being encouraging compared to the other four methods studied in the current voting.

Weighted voting is already used in several scenarios. This paper proposed a variant of this method, based on a dual set of weights. The explanation for this is intuitive and can be further developed by optimizing these weights with a neural network or

another type of optimization algorithm. Further research into this subject may reward us with ways to segment images that are more accurate than the currently available algorithms.

A possible future development of the proposed technique will be the integration with other voting-based methods [17-19] in a system based on multiple experts' decision destined for automatic image document content conversion.

**Acknowledgement**

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0689/„Lib2Life–Revitalizarea bibliotecilor și a patrimoniului cultural prin tehnologii avansate"/"Revitalizing Libraries and Cultural Heritage through Advanced Technologies", within PNCDI III.

**References**

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation*, in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, pp. 580-587, DOI: 10.1109/CVPR.2014.81, June 2014.

[2] R. Girshick, *Fast R-CNN*, in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1440-1448, DOI: 10.1109/ICCV.2015.169, December 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, *Spatial pyramid pooling in deep convolutional networks for visual recognition*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 37, issue 9, pp. 1904-1916, DOI: 10.1109/TPAMI.2015.2389824, January 2015.

[4] S. Ren, K. He, R. Girshick, J. Sun, *Faster R-CNN: Towards real-time object detection with region proposal networks*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 39, issue 6, in Advances in neural information processing systems, pp. 91-99, DOI: 10.1109/TPAMI.2016.2577031, arXiv: 1506.01497, June 2015.

[5] A. Krizhevsky, I. Sutskeverand, G. Hinton, *ImageNet classification with deep convolutional neural networks*, in Advances in neural information processing systems, volume 25, issue 2, pp.1097-1105, DOI: 10.1145/3065386, January 2012.

[6] S. Lazebnik, C. Schmid, and J. Ponce, *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition

(CVPR'06), IEEE, volume 2, pp. 2169-2178, DOI: 10.1109/CVPR.2006.68, June 2006.

[7] K. Graumanand, T. Darrell, *The pyramid match kernel: Discriminative classification with sets of image features*, in Tenth IEEE International Conference on Computer Vision (ICCV'05), volume 1 (volume 2, pp. 1458-1465), Beijing, China, DOI:10.1109/ICCV.2005.239, October 2005.

[8] J. Sivicand, A. Zisserman, *Video Google: a text retrieval approach to object matching in videos*, in Proceedings Ninth IEEE International Conference on Computer Vision, volume 2, pp. 1470-1477, IEEE, Nice, France, DOI:10.1109/ICCV.2003.1238663, October 2003.

[9] K. He, G. Gkioxari, P. Dollár, R. Girshick, *Mask R-CNN*, in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988, Venice, Italy, DOI: 10.1109/ICCV.2017.322, October 2017.

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, *OverFeat: Integrated recognition, localization and detection using convolutional networks*, arXiv preprint arXiv: 1312.6229, 2013.

[11] W. Y. Zou, X. Wang, M. Sun, and Y. Lin, *Generic object detection with dense neural patterns and regionlets*, in arXiv preprint arXiv: 1404.4316, 2014.

[12] C. C. Nguyen, G. S. Tranetal., *Towards Real-Time Smile Detection based on Faster Region Convolutional Neural Network*, in 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR), pp. 1-6, IEEE, DOI:10.1109/MAPR.2018.8337524, April 2018.

[13] C. A. Boiangiu, R. Ioanitescu, *Voting-based image segmentation*, in The Proceedings of Journal of Information Systems & Operations Management, volume 7, issue 2, pp. 211-220, 2013.

[14] Y. Cheng, *Mean shift, mode seeking, and clustering*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 17, issue 8, pp. 790-799, DOI:10.1109/34.400568, August 1995.

[15] S. Lloyd, *Least squares quantization in PCM*, in IEEE Transactions on Information Theory, volume 28, issue 2, pp. 129-137, DOI:10.1109/TIT.1982.1056489, March 1982.

[16] N. Jamil, H. C. Soh, T. M. T. Sembok, Z. A. Bakar, *A modified edge-based region growing segmentation of geometric objects*, in International Visual Informatics Conference, IVIC 2011: Visual Informatics: Sustaining Research and Innovations, Lecture Notes in Computer Science, volume 7066, pp. 99-112, Springer, Berlin, Heidelberg, DOI:10.1007/978-3-642-25191-7_11, November 2011.

[17] Costin-Anton Boiangiu, Radu Ioanitescu, Razvan-Costin Dragomir, *Voting-Based OCR System*, in The Journal of Information Systems & Operations Management, volume 10, number 2, 2016, pp. 470-486.

[18] Costin-Anton Boiangiu, Mihai Simion, Vlad Lionte, Zaharescu Mihai, *Voting Based Image Binarization*, in The Journal of Information Systems & Operations Management, volume 8, number 2, pp. 343-351, 2014.

[19] Costin-Anton Boiangiu, Paul Boglis, Georgiana Simion, Radu Ioanitescu, *Voting-Based Layout Analysis*, in The Journal of Information Systems & Operations Management, volume 8, number 1, pp. 39-47, 2014.