IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Accurate Object Detection and Semantic Segmentation using Gaussian Mixture Model and CNN

Sakshi Jain[1], Satish Dehriya[2], Yogendra Kumar Jain[3]

Research Scholar, Computer Science & Engg, Samrat Ashok Technological Institute, Vidisha (M.P.), India[1]

Assist. Professor, Computer Science & Engg, Samrat Ashok Technological Institute, Vidisha (M.P.), India [2]

Head of the Department, Computer Science & Engg, Samrat Ashok Technological Institute, Vidisha (M.P.), India[3]

Abstract: Semantic segmentation and object detection are two most common tasks in the field of image processing, pattern recognition and classification. This paper presents a two stage procedure to perform these two tasks. The proposed work uses the Gaussian mixture model for image segmentation and identifies the segments by optimally searching the possible Gaussian distribution inside the image histogram. The optimal partition searching procedure uses the genetic algorithm. For the object detection, we apply the Convolution Neural Network (CNN) to extract the features of each segments and then apply them to pre-trained Support Vector Machine (SVM) to identify the object the segment belongs to. Finally the proposed system is developed using Matlab computational software and tested with different types of image datasets. The experimental results demonstrate encouraging performance of the proposed technique for both object detection and semantic segmentation tasks.

Keywords: Semantic segmentation, object detection, Gaussian mixture model, Genetic Algorithm, Support Vector Machine.

1. INTRODUCTION

Object detection and semantic segmentation are two basic requirements of visual perception. Object detection is frequently referred as searching of a rectangle box covering the object of interest while semantic segmentation for the most part intends to allocate a class name to every pixel from a predefined set. Despite the fact that emphatically related, these two tasks have commonly looked closely related but required generously different strategies. Layout based detection utilizing sliding window examining has long been the overwhelming methodology for object detection. In spite of the fact that great at finding the rough object positions, this methodology ordinarily neglects to precisely restrict the entire object through a tight jumping box. It has been found that the biggest problem of detection with these techniques is incorrect bouncing box positioning. This may emerge from the restricted representation capacity of template based classifiers for overlapping objects. This paper examines the issue of accomplishing automatic detection, recognition and segmentation of object classes in image, and presents a framework ought to automatically divide it into semantically significant segments each named with a specific item class. In this paper we combine several well tested techniques in different fields of image processing to develop the complete framework. For example, the Gaussian mixture model for image segmentation with Expectation Maximization (EM) is already tested with many segmentation problems, but for the present problem we modified it as the EM is replaced with the genetic algorithm which provides greater flexibility because the objective function can be modified as per the requirements and not specific to special conditions like in EM. Secondly, the object detection procedure uses the Convolution Neural Network (CNN) for feature extraction and Support Vector Machine for classification , however in our work the input image for the CNN in not of a fixed shape which reduces the detection efficiency instead the image is similar to object’s shape in the image.

2. LITERATURE REVIEW

Recently, a number of techniques for both the tasks have been presented. This section presents a brief overview of some of the related literatures. Dong has proposed a combined approach for simultaneous object detection and semantic segmentation in [5], they employ the hypotheses based approach for the segmentation component. Then, with a group of generated segment hypotheses, the segmentation problem is translated into selecting the best hypothesis. The algorithm performs sliding window scanning and hypothesis based semantic segmentation jointly on template based detection. The correct detection and segmentation must fulfill the both detection and segmentation predictions. This requirement is achieved by utilizing a consistency model and the algorithm is further modified by designing a context model to aggregate both local and global context information. Carreira et al. [2] concentrated on feature extraction, coding and pooling, considering them as most important process in many object recognition systems and explore novel pooling techniques that encode the second-order statistics of local

DOI 10.17148/IJARCCE.2015.41167

descriptors inside a region. This effect is achieved by introducing multiplicative second-order equivalents of average and max-pooling. The average and max-pooling together with appropriate non-linearities gives better performance on free-form region recognition, without need of any feature coding. The literature also presented that enriching local descriptors with provided image information further improves the performance while compared with the coding especially for the presented algorithm. Arbelaez et al. [4] addressed the problems of segmentation and recognition of objects in real world images they especially focused on humans and other animals like objects. Their proposed model is based on region-based scanning-windows object detectors that integrate efficiently top-down information from local and global appearance clues. Finally the detector generates class-specific scores for bottom-up regions, and then accumulates the scores of multiple overlapping candidates through pixel classification. Shotton has presented a new approach to learning for Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation in [6]. The presented discriminative model jointly model shape and texture by textons based features. The shared boosting is used to generate an efficient classifier for large number of classes. Furthermore these classifiers are incorporated in a conditional random field to achieve precise image segmentation.

3. GAUSSIAN MIXTURE MODEL (GMM) FOR IMAGE SEGMENTATION

The Gaussian Mixture Model (GMM) model assumed that there are a finite number of Gray-level probability density functions in the image, say k, and each pixel distribution can be modeled by one Gaussian function which represents the one object in the image.

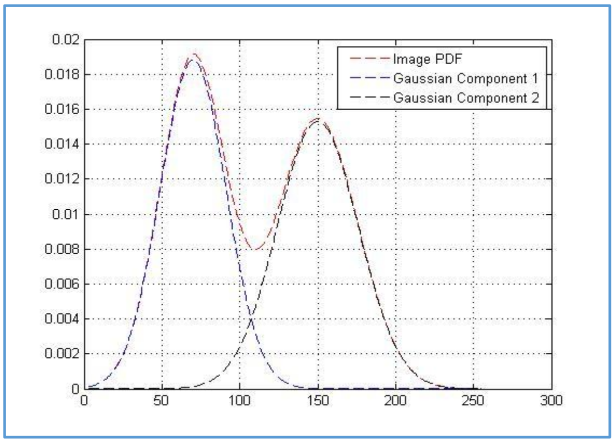

Figure 1: Gray-level probability density function approximated by two Gaussian Components.

With this assumption, the whole image can be modeled by a mixture of k component Gaussian distributions in some unknown proportions $\pi_i, i = 1,2,3,..,k$. The probability density function (PDF) of a data point x will be

$$f(x|\psi) = \sum_{i=1}^{k} \pi_i f(x, \mu_i, \sigma_i^2)\,, \dots\dots\dots\dots..\dots..(1)$$

where, $\pi_i$is the mixing weight such that $\sum_{i=1}^{k} \pi_i = 1$, and ψ is a vector containing parameters π_i, μ_i and $\sigma_i^2$for $i = 1,2,3,\dots,k$. Hence,

$$f(x, \mu_i, \sigma_i^2) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(X-\mu_i)^2}{2\sigma_i^2}}, \dots\dots\dots\dots\dots\dots\dots..(2)$$

Equation (2) describes the i^{th}component Gaussian distribution with mean $\mu_i$and varianceσ_i^2. The fit of a model to the data can be measured by the mean squared error between the PDF of original image and the sum of Gaussian distributions (equation (1)).

$$MSE = \frac{\left(PDF_{img} - f(x|\psi)\right)^{\frac{1}{2}}}{k}, \dots\dots\dots\dots\dots\dots\dots.(3)$$

From the equation (3) it is clear that the image can be portioned properly only if the value of MSE reached to approximately zero. Hence it is required to search the proper values for π_i, μ_i and σ_i^2.

In the above figure 1, the X axis represents the data point x and the Y axis represents

(a) $f(x|\psi)$ for the image PDF.

(b) $f(x, \mu_i, \sigma_i^2)$ for various Gaussian components.

4. GENETIC ALGORITHM

In the field of artificial intelligence, a genetic algorithm (GA) is a heuristic (also sometimes called a meta-heuristic routinely used to generate useful solutions to optimization and search problems) search algorithm that mimics the process of natural selection. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

4.1 Initialization

The initialization of the genetic algorithm is performed by arbitrary selecting (or generating) a number of (equal to the set value of initial population parameter) possible solutions for the given problem. Often, the initial population is generated randomly, allowing the entire range of possible solutions (the search space). However sometimes, the solutions may be "seeded" (or predicted) in areas where optimal solutions are likely to be found.

4.2 Selection

After each generation, some of the existing population is selected to create a new generation of solutions. The selection of the particular existing population for creation of new generation is done through a fitness-based process, where fitter solutions (as measured by a fitness function or objective function) are typically more likely to be selected.

4.3 Genetic operators

For each new solution to be produced, a pair of "parent" solutions (selected existing solutions) are pass through the crossover and mutation (while crossover operates on two or more parental chromosomes, mutation locally but randomly modifies a solution) operations to create the "child" solution (generated new solutions) using the above methods of crossover and mutation, a new solution is

DOI 10.17148/IJARCCE.2015.41167

created which typically shares many of the characteristics of its "parents".

4.4 Termination
This process from 4.1 to 4.3 is repeated until a termination condition has been reached. Common terminating conditions are:

- A solution is found that satisfies fitness criteria.
- Maximum number of generations reached.
- Maximum computation time reached
- The successive iterations no longer produce better results.

4.5 The pseudo-code for genetic algorithm

```
function GA( )
{
    Initialize population;
    Calculate fitness function;
    While(fitness value != termination criteria)
    {
        Selection;
        Crossover;
        Mutation;
        Calculate fitness function;
    }
}
```

5. CONVOLUTIONAL NEURAL NETWORKS

The problem of feature extraction for recognition in natural images is typically very difficult, because there are not enough training points in the space created by the input images in order to allow accurate estimation of class probabilities throughout the input space. Additionally, they are also sensitive to small variations. Convolutional networks (CN) incorporate constraints and achieve some degree of shift and deformation invariance using three ideas: local receptive fields, shared weights, and spatial subsampling. The use of shared weights also reduces the number of parameters in the system aiding generalization. Convolutional networks have been successfully applied to many characters and face recognition systems.

A typical convolutional network is shown in Figure 2. The network consists of a set of layers each of which contains one or more planes. Approximately centered and normalized images enter at the input layer. Each unit in a plane receives input from a small neighbourhood in the planes of the previous layer. The idea of connecting units to local receptive fields dates back to the 1960's with the perceptron and Hubeland Wiesel's discovery of locally sensitive orientation-selective neurons in the cat's visual system. The weights forming the receptive field for a plane are forced to be equal at all points in the plane. Each plane can be considered as a feature map which has a fixed feature detector that is convolved with a local window which is scanned over the planes in the previous layer. Multiple planes are usually used in each layer so that multiple features can be detected. These layers are called convolutional layers.

Once a feature has been detected, its exact location is less important. Hence, the convolutional layers are typically followed by another layer which does a local averaging and sub sampling operation (e.g., for a sub sampling factor of two:

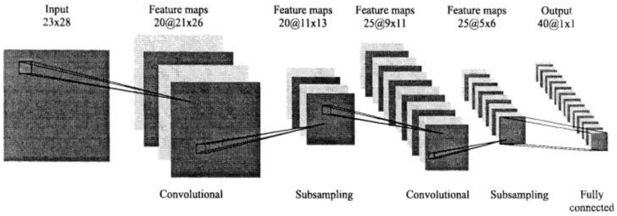

Figure 2: A typical convolutional neural network.

$$y_{ij} = \frac{x_{2i,2j} + x_{2i+1,2j} + x_{2i,2j+1} + x_{2i+1,2j+1}}{4}$$

where y_{ij} is the output of a subsampling plane at position and is the output of the same plane in the previous layer). The network is trained with the usual backpropagation gradient-descent procedure. A connection strategy can be used to reduce the number of weights in the network.

6. SVM CLASSIFICATION

In machine learning, Support Vector Machines (SVMs) are supervised learning models used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

Let $x_i \in R^m$ be a feature vector or a set of input variables and let $y_i \in \{+1, -1\}$ be a corresponding class label, where m is the dimension of the feature vector. In linearly separable cases a separating hyper-plane satisfies.

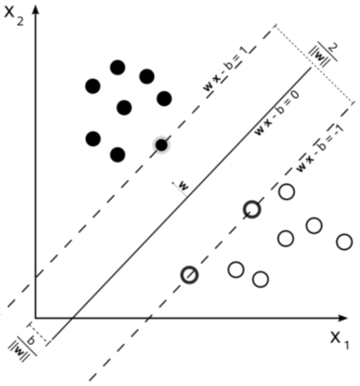

Figure 3: Maximum-margin hyper-plane and margins for an SVM trained with samples from two classes.

Samples on the margin are called the support vectors.

$$y_i(\langle w.x_i \rangle + b) \geq 1, i = 1 \dots n, \dots.(4)$$

Where the hyper-plane is denoted by a vector of weights w and a bias term b. The optimal separating hyper-plane, when classes have equal loss-functions, maximizes the margin between the hyper-plane and the closest samples of classes. The margin is given by

The optimal separating hyper-plane can now be solved by maximizing (4) subject to (5).

DOI 10.17148/IJARCCE.2015.41167

$$d(w,b) = \min_{x_i,y_i=1} \frac{|\langle w.x_i\rangle + b|}{\|w\|} + \min_{x_i,y_i=-1} \frac{\left|\langle w.x_j\rangle + b\right|}{\|w\|}$$
$$= \frac{2}{\|w\|} \ldots \ldots (5)$$

The SVM gives the following advantages over neural networks or other AI methods:

SVM training always finds a global minimum, and their simple geometric interpretation provides fertile ground for further investigation.

Most often Gaussian kernels are used, when the resulted SVM corresponds to an RBF network with Gaussian radial basis functions. As the SVM approach "automatically" solves the network complexity problem, the size of the hidden layer is obtained as the result of the QP procedure. Hidden neurons and support vectors correspond to each other, so the center problems of the RBF network is also solved, as the support vectors serve as the basis function centers.

Classical learning systems like neural networks suffer from their theoretical weakness, e.g. back-propagation usually converges only to locally optimal solutions. Here SVMs can provide a significant improvement.

The absence of local minima from the above algorithms marks a major departure from traditional systems such as neural networks.

SVMs have been developed in the reverse order to the development of neural networks (SVMs). SVMs evolved from the sound theory to the implementation and experiments, while the SVMs followed more heuristic path, from applications and extensive experimentation to the theory.

7. PROPOSED ALGORITHM

The proposed algorithm works in two parts. In first part the image segmentation is performed while in second part the object detection is performed.

Part 1: In image segmentation problem using Gaussian mixture model (GMM) we need to find the proper values of variables π_i, μ_i and $\sigma_i^2, i = 1,2,3, \ldots, k$. Hence the total number of variables in the problem is $\pi = \{\pi_1, \pi_2, \pi_3, \ldots \ldots \pi_k\}, \mu = \{\mu_1, \mu_2, \mu_3, \ldots \ldots \mu_k\}$, where k total number of segments.

Hence the vector of a gene is group of elements corresponding to $P = [\pi, \mu, \sigma^2]$. Therefore, the value of i^{th} gene at generation g can be represented as the vector

$$P_i^g = \{\pi_1^g, \pi_2^g, \ldots\ldots, \pi_k^g, \mu_1^g, \mu_2^g, \ldots\ldots, \mu_k^g, \sigma_1^{2^g}, \sigma_2^{2^g}, \ldots\ldots, \sigma_k^{2^g}\}$$

which sets the dimention of the vector equal to $3k$.

Now if the initial population be N then the complete set can be presented as:

$$P^g = \left\{p_1^g, p_2^g, p_3^g, \ldots\ldots p_N^g\right\}$$
$$p_i^g = \{\pi_1^g, \pi_2^g, \ldots\ldots, \pi_k^g, \mu_1^g, \mu_2^g, \ldots\ldots, \mu_k^g, \sigma_1^{2^g}, \sigma_2^{2^g}, \ldots\ldots, \sigma_k^{2^g}\}$$
$$\pi_{min} \le \pi_i^g \le \pi_{max}, \ \ for\ \forall i, g$$
$$\mu_{min} \le \mu_i^k \le \mu_{max}, \ \ for\ \forall i, g$$
$$\sigma_{min}^2 \le {\sigma^2}_i^k \le \sigma_{max}^2, \ \ for\ \forall i, g$$

Where $\pi_{min}, \mu_{min}, \sigma_{min}^2$ and $\pi_{max}, \mu_{max}, \sigma_{max}^2$ are the user defined lower and upper bounds for level of decompositions and threshold value.

Using the above mapping we can call the Genetic algorithm to find the best values for $\pi, \mu\ and\ \sigma^2$ as follows:

1. Initialize the max population, max iterations and fitness criteria.
2. Produce the initial population randomly within the defined constrains.
3. Extract the values of $\pi, \mu\ and\ \sigma^2$ for each Generation.
4. Calculate the fitness of each gene using MSE equation (3).
5. Produce the new generation by performing crossover and mutation operation.
6. Check for the stopping criteria. If satisfy then exit else repeat the procedure from Step 2.
7. On exit return the gene related to best solution.

Part 2: Once the segmentation is completed now we can proceed to object detection.

1. Feature Extraction: For the object we initially extract a feature vector from each segment region using CNN. Features are computed by forward propagating a mean-subtracted RGB image through five convolutional layers and two fully connected layers.
2. Once we get the feature, next the classification of the object can be performed using a pre-trained support vector machine (SVM).

8. SIMULATION RESULTS

The proposed work is implemented using the Matlab environment and PASCAL VOC 2010 image dataset is used for the extensive testing of the algorithm. Finally the result of the simulation is presented in tabular form for comparison.

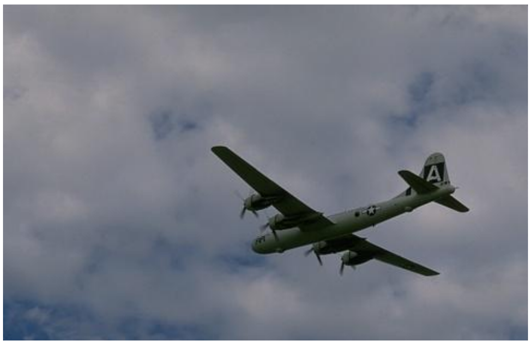

Figure 4: Image showing the plane as object.

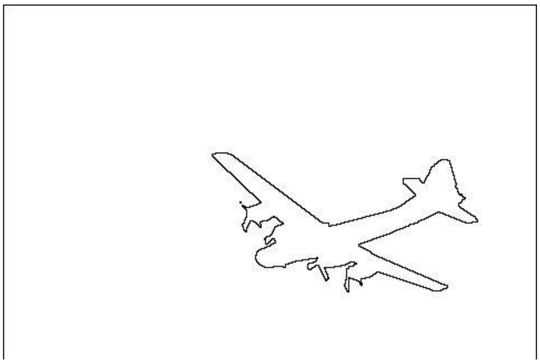

Figure 5: Detection of plane as object by the proposed algorithm.

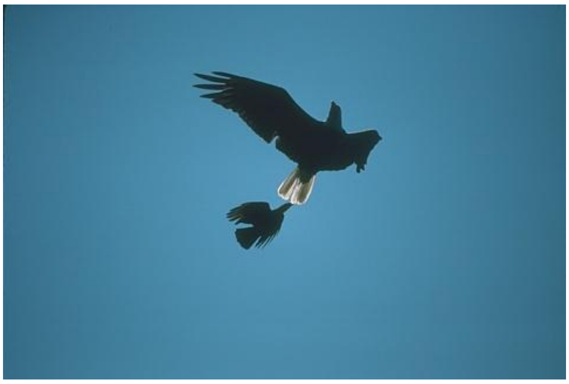

Figure 6: Image showing the bird as object.

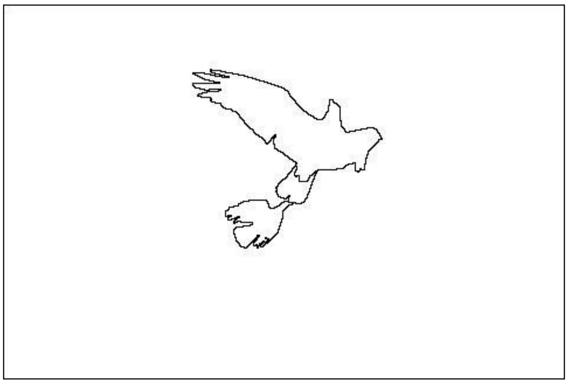

Figure 7: Detection of bird as object by the proposed algorithm

VOC 2010	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47	44.8	35.1
Region lets	65	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54	45.9	39.7
Seg DPM	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32	30.5	56.4	57.2	65.9	27	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	71.8	65.8	53	36.8	35.9	59.7	60	69.9	27.9	50.6	41.4	70	62	69	58.1	29.5	59.4	39.3	61.2	52.4	53.7
Proposed	72.3	65.1	55.6	38.2	34.9	61.8	59.4	71.1	30.5	51.5	44.7	68.2	60.1	73.4	57.9	32.3	65.8	45.2	66.1	51.7	58.3

Table 1: Results Comparison for PASCAL VOC 2010 dataset

9. CONCLUSION

This paper presents the Gaussian mixture model based image segmentation using genetic algorithm for optimal partitioning with convolution neural network as feature extractor which at last is applied to support vector machine for object classification. The results show that the presented algorithm gives a 10% relative improvement over existing algorithm. This achievement is possible because of proper segmentation done by the Gaussian mixture model consideration with optimal parameters searching by genetic algorithm and efficient feature extraction by convolution neural network which helps in proper identification of object through SVM. In future, we can plan to integrate explicit context information, to improve the accuracy of the classification.

REFERENCES

[1] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur and Roberto Cipolla, “Segmentation and Recognition Using Structure from Motion Point Clouds”,In: 10th European Conference on Computer Vision, Springer Berlin Heidelberg, vol. 5302, pp. 44-57, 2008.

[2] Joao Carreira, Rui Caseiro, Jorge Batista and Cristian Sminchisescu, “Semantic segmentation with second-order pooling”, In: Proceedings of the 12th European conference on Computer Vision, Springer-Verlag Berlin, Heidelberg , Vol. Part VII, pp. 430-443, 2012.

[3] R. Arandjelovi´c and A. Zisserman, “Smooth object retrieval uses a bag of boundaries”, In: Computer Vision (ICCV), IEEE International Conference, pp. 375-382, 2011.

[4] Pablo Arbelaez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev and Jitendra Malik, “Semantic segmentation using regions and parts”, Computer Vision and Pattern Recognition (CVPR),IEEE Conference, pp. 3378-3385, 2012.

[5] Jian Dong, Qiang Chen, Shuicheng Yan and Alan Yuille, “Towards Unified Object Detection and Semantic Segmentation”, In: 13th European Conference on Computer Vision, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, Springer International Publishing, vol. 8693, pp. 299-314, 2014.

[6] Jamie Shotton, John Winn, Carsten Rother and Antonio Criminisi, “TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation”, In: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part I, Springer Berlin Heidelberg, vol. 3951, pp. 1-15, 2006.

[7] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun and Alan Yuille “The Role of Context for Object Detection and Semantic Segmentation in the Wild”, In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference , pp. 891-898, 2014.

[8] Joao Carreira, Fuxin Li and Cristian Sminchisescu, “Object Recognition by Sequential Figure-Ground Ranking”, International Journal of Computer Vision, Springer US, Vol. 98, pp 243-262, 2012.

[9] T. Brox, L. Bourdev, S. Maji, and J. Malik, “Object segmentation by alignment of poselet activations to image contours”, In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference, pp. 2225-2232, 2011

[10] L. Bourdev and J. Malik, “Poselets: Body part detectors trained using 3d human pose annotations”, In: Computer Vision, 2009 IEEE 12th International Conference, pp. 1365-1372, 2009.

[11] P. Arbelaez, M. Maire, C. Fowlkes and J. Malik, “Contour detection and hierarchical image segmentation”, In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol. 33, pp. 898-916, 2011.

[12] E. Borenstein and S. Ullman, “Combined top-down/bottom-up segmentation”, In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, pp. 2109-2125, 2008.

DOI 10.17148/IJARCCE.2015.41167