ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 3, No.15 , May 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Indoor Scene Recognition Using Local Semantic Concepts

Elham Seifossadat[1], Niloofar Gheissari[2] and Ali Fanian[3]

[1] Electrical and Computer Department,Isfahan University of Technology
Isfahan, Iran
e.seifossadat@ec.iut.ac.ir

[2] Electrical and Computer Department,Isfahan University of Technology
Isfahan, Iran
n.gheissari@cc.iut.ac.ir

[3] Electrical and Computer Department,Isfahan University of Technology
Isfahan, Iran
a.fanian@cc.iut.ac.ir

Abstract

This paper introduces a system for recognizing indoor scene images. The system aims at recognizing the environment illustrated in an image and assigning an appropriate semantic label to it. Developing such systems is one of the most important issues in the field of machine vision and robotics. They are extensively used in object recognition, image and video semantic recognition, motion detection, positioning, and robot direction. The overall algorithm involves three major steps. The first is to extract local information and spatial relationships in the images, modeling the environment based on those pieces of information, and using this model for the semantic sectioning of the image and labeling local areas using graph-based energy minimization algorithm. Results show that this system can perform as well as other object-based and 3D-image features-based environment recognition systems.

Keywords: Computer Vision, Scene Recognition, Semantic Segmentation, Local Semantic Concepts.

1. Introduction

Scene is a perspective of the real world, which is comprised of multiple objects and levels which are meaningfully positioned alongside one another. Scene could be categorized into indoor and outdoor scenes. Indoor scene includes covered spaces, such as kitchen, hospital, while outdoor scenes include open spaces, such as streets, beaches, and mountains. The proposed system aims at recognizing the scene illustrated in images and assigning appropriate semantic labels to them[1]. Developing such systems is one of the most important issues in machine vision and robotics. They are extensively used in object recognition, image and video semantic recognition, motion detection, positioning, and robot direction.

Multiple methods have been proposed in the field of scene recognition. Bosch [2], considers representing scene image information and describing it, and learning scene model based on those representations as the most important steps in scene recognition algorithms. According to this, scene recognition methods are divided into two categories:

- Modeling based on low-level features
- Semantic modeling

Low-level features-based modeling methods use low-level features, such as color, texture, edge, and at times, entropy, and pixel or shape illumination in inferring higher-level information. These methods assume that the type of scene could be recognized directly based on low-level features.

Other modeling methods based one low-level features include the ones introduced in [3] and [4]. They exploit the prominent direction of the edge and joints in images to recognize the scene. In [5], color histograms was used. Then, these histogram classes were used for classifying indoor and outdoor scenes. The methods introduced in [6] and [7] used spectral analysis for recognizing outdoor scenes. In [8], illumination of the image was used in recognizing the environment.

The main issue in aforementioned methods is that they cannot be applied to unseen images. Due to inter-class diversity and similarity between various scene

classes, using low-level features of the image cannot produce a comprehensive model for complex scenes. The other problem with such methods is that they do not exploit important semantic information of the image which can help in scene recognition. Although they have a high performance in recognizing outdoor images and landscapes, they have a low performance in recognizing indoor scene images.

In order to overcome the problems of scene modeling based on low-level features, modeling methods based on semantic information were introduced. In semantic modeling, semantic concepts, such as sky, lawn, water, and etc. are used alongside low-level features in scene recognition. These methods are classified into the following:

- Semantic objects
- Local semantic concept
- Semantic properties

The methods, which are based on semantic objects, use the objects in the scene to describe them. One of these methods is introduced in [9], which recognizes indoor scene images. It combines global and local features for detecting the objects in the scene. In [10] and [11], common objects in images were used for representing central semantic for recognizing the indoor scene images. In this method, visual features of different areas in the image and spatial features are extracted using a 3D sensor for identifying objects and associating objects to scenes. In [12], [13], [14], [15], and [16], depth features of images are extracted using RGB-D sensors for recognizing objects and, subsequently, the scene. In [17], shape properties, such as the reflection, and illumination caused by the existing objects in the image are used in recognizing objects and scenes. Although using objects for recognizing indoor scenes seems to be efficient, in complex and disordered scenes, it is impossible to properly identify objects, even using state-of-the-art and highly successful detection methods.

As an alternative, methods have been proposed to model the scene based on local semantic concepts. In these methods, central features are extracted using local descriptors around specific points in the image. Based on those features, the meaning of the scene is represented. An example of these methods could be found in [18], where indoor and outdoor scenes were recognized based on global geometric correspondence estimation. In this method, the image is incrementally divided into sub-sections. Then, for each section, local feature histogram is extracted. This structure is called "spatial pyramid." Using such structures does not constantly give correct results since coded spatial information is very limited in this method. This limitation stems from the fact that it is assumed that major section related to the images with similar scenes should happen in cells that are similar to the pyramid. In [19], alongside feature histograms, directions in 3D space were also exploited. In [20], for adding spatial data of major points of the image, their relative position was used, rather than spatial pyramid. As with spatial pyramid method, this method, which was used for recognizing indoor scenes, have some limitation. In [21], for recognizing outdoor scenes, other than using the relative positions of major points, their co-occurrence frequency was also used.

In contrast to previous methods, the methods which are based on semantic features use the visual features of images, rather than object information. In fact, they use naturalness, openness, expansion, ruggedness, and roughness features, rather than man-madness, the existence of horizontal line, perspective in man-made scenes, deviation from horizon in natural scene images, and fractal complexity, respectively, which are shared by images from similar classes. Each feature constructs one dimension of the environment and all features introduce the dominant spatial structure of the scene. Scene type is determined based on the membership in each of these features [22]. While this method performs well in recognizing outdoor images, it performs poorly in recognizing indoor scenes. It is because most indoor scenes contain complex structures which are not identifiable by such features.

Since semantic modeling exploits the highest amount of information in recognition, this paper used it for solving scene recognition problem. Since indoor scenes are mostly complex and lack an overall structure, using object or semantic features is not useful. Therefore, the aim of this paper was to propose a method for recognizing indoor scenes based on local semantic concepts, which can overcome indoor scene problems and are efficient.

2. Proposed method

Semantic labeling of various sections of the image was used in the proposed method. Using semantic segmentation [23], each section of the image is labeled by the name of an appropriate object. The overall stages involved in the proposed method are described later.

2.1 Making training images

First, training images are divided into smaller sections, called turbo pixels using hyper segmentation algorithms. These turbo pixels are moderately homogeneous sections of an image, each of which correspond to the semantic objects of the image [24]. Then, each section of the image is labeled appropriately. Each class of object labels are recognized by a unique color. Figure 1. An example of hyper segmented image along with its corresponding labeled image.

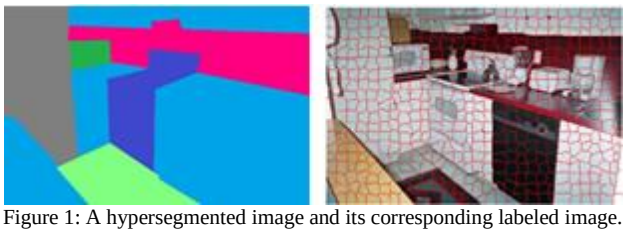

Figure 1: A hypersegmented image and its corresponding labeled image.

2.2 Image feature extraction

In the next stage, image features are extracted using histogram of oriented gradients (HoG), scale-invariant feature transform (STIF), and local binary patterns (LBP) descriptors. SIFT descriptors are one of the most powerful tools in extracting key points of an image. These key points are mostly located at the corners, troughs, and T-shaped joints of the image. The features extracted using this descriptor are tolerant of changes in image scale, skewness, and rotation [25]. HoG descriptor is one of the descriptors of machine vision which describes the edge features of images by calculating the number of gradient direction occurrences in various sections [26]. LBP algorithm is also used for extracting image texture features. The most important features of this algorithm is its being tolerant to changes in scale, direction, and illumination [27]. Figure 2 shows the extracted features of the images.

2.3 Forming a library of image features

The feature vectors determined using each descriptor is separately classified using K-means algorithm. The center of each cluster is named "visual word." The determined visual word for each descriptor is stored in separate libraries. Based on the number of their occurrences in each image, a histogram is formed for the image using which the image is represented. Hence, there will be three distinct libraries with Vvisual words. By using them for each picture, three different representations of $V = [v_1, v_2, \dots, v_T]$ are formed based on the used descriptors. The most important advantage of creating distinct histograms for each descriptor is the ability to learn distinctively and storing all obtained data from images.

2.4 Semantic segmentation

Assume that each image is segmented into Tsections and that there are L classes of different labels for each image. Semantic segmentation of each image determines the labeling vector $L = (l_1, l_2, \dots, l_T)^T$ which assigns a single label $l_i \in \{1,2, \dots, L\}$ to each section i. Having obtained feature vector $V = [v_1, v_2, \dots, v_T]$, labeling probability L for each section is estimated as below:

$$P(L|V) = \frac{P(V|L)P(L)}{P(V)} \quad (1)$$

For estimating labeling probability L, maximum aposterior probability (MAP) is used as follows:

$$\arg max_L \, P(L|V) = \arg max_L \, P(V|L)P(L) \quad (2)$$

Probability $P(V|L)$ could be expressed as follows using chain rule:

$$P(v_1|l_1)P(v_2|v_1, l_1, l_2) \dots P(v_T|v_1, v_2, \dots, v_{T-1}, L)$$

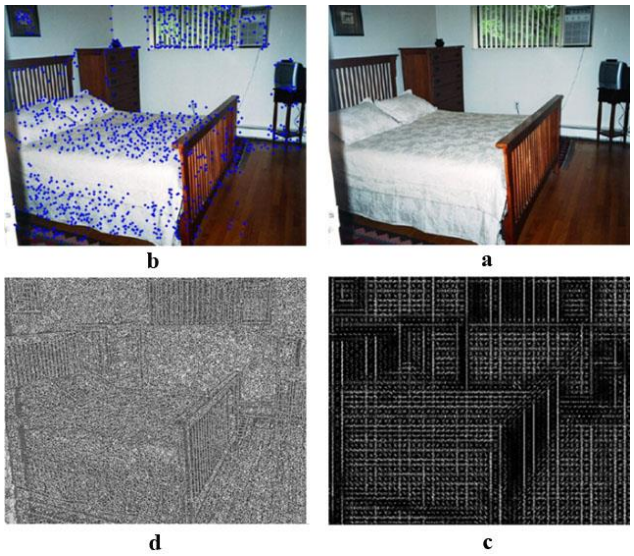

Figure 2: Data extracted from images a) the main image, b) key points extracted by SIFT descriptor which are marked with blue points, c) edge direction extracted using HoG descriptor, d) texture information extracted using LBP descriptor

Therefore, in order to estimate probability$P(V|L)$, the probability of adjacent areas should be also considered. Therefore, probability $P(V|L)$ is calculated as follows:

$$P(V|L) \cong \prod_{i=1}^{T} \prod_{j=1}^{8} P(v_i|B_{i,j}, l_i) \quad (3)$$

where $B_{i,j}$ is a sub-set of visual words that emerged in the neighborhood in the direction $j$with a depth of 2 and there are i. Neighborhood with depth 2 of section i includes direct neighbors and the neighbors of direct neighbors. Two sections are neighbors if they have at least one pixel in common in 8 neighborhood of each other. Therefore, in training, for each class, the co-occurrence probability label of different visual words in 8 directions of neighborhood are calculated. Therefore, conditional probability $P(v_i|B_{i,j}, l_i)$ is calculated as follows:

$$P(v_i|B_{i,j}, l_i) = \left(\frac{1}{|B_{i,j}|+1}\left(P(v_i|l_i) + \sum_{k \in B_{i,j}} P(v_k|l_i)\right)\right) \quad (4)$$

Probability value $P(v_i|l_i)$ is calculated based on the occurrence of visual word i with label l_i in visual words library. Note that the probability $P(V|L)$ for each visual word using each descriptor is calculated separately.

For calculating the co-occurrence probability of visual words, co-occurrence matrix is used. To do this, 8 co-occurrence matrixes C with dimensions of $|V| \times |V| \times |L|$ are used, where V is the total number of visual words in the library and L is the number of labels. *This matrix is formed as follows:*

1. Repeating steps 2 and 3 for training images and their neighborhood with specific label.
2. For section i, all k of the neighborhoods in the direction j at depth 2 are found. Visual words v_i and v_k for these sections are extracted using libraries.
3. One is added to the value of element $C_j(v_i, v_k, l_i)$ for all k which have $l_i = l_k$.
4. Matrix C_j should be normalized after formation so much so that each row of the matrix represents probability $P(v_k|v_i, l_i, l_i = l_k)$

As mentioned earlier, the set of visual words V contains the visual words related to SIFT, HoG, and LBP descriptors. Therefore, probability $P(v_i|B_i, l_i)$for each set of visual words is estimated separately.

$$P(v_{Si}|B_i, l_i) = \prod_{r=1}^{8} P(v_{Si}|B_{ri}, l_i) \quad (5)$$

$$P(v_{Hi}|B_i, l_i) = \prod_{r=1}^{8} P(v_{Hi}|B_{ri}, l_i) \quad (6)$$

$$P(v_{Li}|B_i, l_i) = \prod_{r=1}^{8} P(v_{Li}|B_{ri}, l_i) \quad (7)$$

where v_{Si}, v_{Hi}, and v_{Li} are visual words corresponding to descriptors HoG, SIFT, and LBP. In order to increase the generalization of scene recognition system, these three probabilities are considered to be independent of one another.

Assume that the sections related to l in training images contain visual words v_S, v_H, and v_L, and the sections related to label l' contain visual words $v_S{}'$, $v_H{}'$, $v_L{}'$, and $P(v_L{}'|B, l) \ll P(v_L|B, l)$. Therefore, if a section in the testing image is related to section l containing visual words v_S, v_H, and $v_L{}'$, and the probability of labeling of this section depends on the co-occurrence of three probabilities $P(v_L{}'|B, l)$, $P(v_H|B, l)$, and $P(v_S|B, l)$:

$$P(V|l) = P(v_L{}'|B, l) \times P(v_H|B, l) \times P(v_S|B, l) \ll P(v_l|B, l) \times P(v_H|B, l) \times P(v_S|B, l) \quad (8)$$

Equation (8) shows that if there is at least one difference between the data obtained from training images and the data obtained from testing images related to a similar section, the probability of labeling this section with the same label decreases. Therefore, the results of scene labeling will have errors.

In order to solve this problem, the probability of labeling a section in terms of the weighted total of obtained probabilities from different descriptors is calculated. Weights were calculated using K-fold cross validation.

$$P(V|B, l) = \frac{1}{3} \times \frac{1}{8}(1.5\, P(v_H|B, l) + 1.2\, P(v_S|B, l) + 0.3\, P(v_l|B, l)) \quad (9)$$

Using Equation (9), the effect of the existence of difference between visual words obtained from testing and training images corresponding to a similar section decreases in labeling that section. Therefore, generalization of training images to cover testing images increases.

2.5 Final labeling

In final labeling, energy minimization is carried out using second-order Markov random field (MRF). In order to do this, following equation is used:

$$arg\ min_L\left(\sum_{i=1}^{T} E_{app} + \sum_{|i,k| \in \varepsilon} E_{smooth}\right) \quad (10)$$

Where $E_{app} = -\log P_S(v_i|B_{i,j}, l_i)$ and $P_S(v_i|B_{i,j}, l_i)$ are the sum of probability$P(v_i|B_{i,j}, l_i)$, which is obtained from three descriptors.

E_{Smooth} $P(L)$ is also calculated using a posterior probability. Probability P(L) or smoothness equation is approximately obtained from the relationship between both neighboring sections:

$$P(L) \approx exp\left(\sum_{(i,j) \in \varepsilon} g(i,j)\right) \quad (11)$$

Function $g(i,j)$ is calculated as follows:

$$g(i,j) = \begin{cases} 1 - e, & if\ l_i = l_j \\ \delta + e, & otherwise \end{cases} \quad (12)$$

where $e = \exp(-\|d_i - d_j\|^2/2\partial^2)$. In these relationships, d_i and d_j are feature vectors of the texture of both neighboring sections of i and j, respectively. σ

equals 1.0 and ε is the set of all pairs of neighboring sections. This smoothness equation is a combination of Potts model for punishing neighboring section pairs with different labels with factor δ and the similarity of neighboring sections. Using this equation, on one hand a similar label is maintained for neighboring sections, and on the other hand, similar labels which have different texture are punished. Therefore, E_{Smooth} was roughly taken to be equal to $g(i,j)$. In order to calculate the minimum energy, variables δ and λ were taken to be 8.0 and 2.0.

3. Experimental results

In order to test the proposed method, the following dataset was used.

d_1:The first set contains 800 colored images which are divided into 8 scene classes, such as the internal part of kitchens, bedrooms, toilets, halls, meeting rooms, bookstores, PC rooms, and stores. Each set of scenes in this set contains 100 images with the minimum resolution of 500*300.

d_2:The second set contains 4485 images which are classified into 10 outdoors scenes, such as, beaches, forests, highways, urban buildings, mountains, open spaces, streets, skyscrapers, suburban buildings, and factories. It also contains 5 sets of indoor scenes, such as toilets, kitchens, bedrooms, offices, and stores. The images in each scene have a resolution of at least 300*250 with each class containing 210 to 410 images [18].

d_3:This dataset contains 2688 colored images with a resolution of 250*250. It is divided into 3 classes of natural outdoor scenes, such as beaches, forests, and mountains, 4 classes of urban scenes, such as highways, urban buildings, streets, and skyscrapers, and another class named open spaces. The last class, in terms of naturalness, is divided into urban and natural scenes, such as farms, villages, extensive landscapes, and aerial images [6]. Each section contains 260 to 410 images.

d_4: NYUD2 dataset which contains 1449 RGB-D images of 27 classes of indoor scenes [16].

The proposed method was applied on these three sets. Figure 3 shows the labels corresponding to each scene class from the image set d_1 and Figure 4 shows an example of training and testing images for each class of this image set. Figure 5 shows finalized labeled images in this set using the proposed algorithm.

The results of indoors scene recognition in image set d_1 are shown in Table 1.

Table 1: Results of scene recognition in 8 indoor scene methods

Store	Hall	Office	Kitchen	Bookstore	Bedroom	Toilet	Meeting room
100%	100 %	100 %	95 %	100 %	98 %	95 %	100 %

As shown in Table 2, the proposed method is more accurate, compared to other methods. As this table shows, the proposed method has an appropriate accuracy rate, compared to other methods.

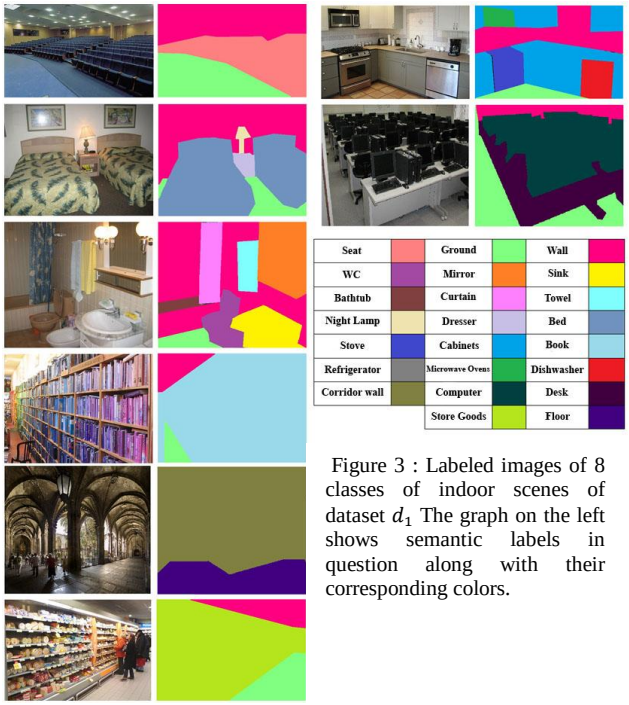

Figure 3 : Labeled images of 8 classes of indoor scenes of dataset d_1 The graph on the left shows semantic labels in question along with their corresponding colors.

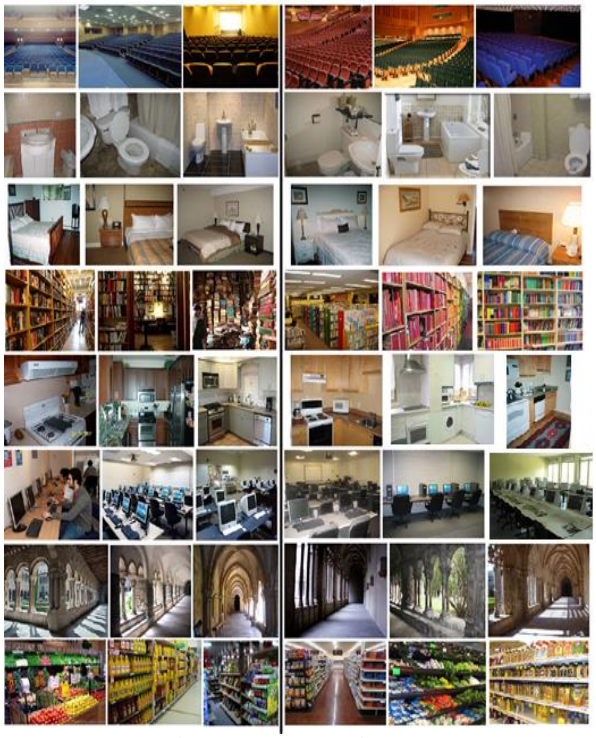

Figure 4: Examples of training images (left) and testing images (right) from dataset d_1.

Table 2: Comparing the results of the proposed method on three image sets.

	First set	Second set	Third set	Fourth set
Proposed method	**98.5%**	97%	88%	**39.5%**
Terrabla [18]	-	82.6%	73%	-
Bosch et al. [28]	-	86.65%	-	-
Elfiky [29]	-	97.4%	-	-
Lazebnik et al. [18]	-	-	64.6%	-
Ulusoy [20]	-	-	82.8%	-
Gupta [12]	-	-	-	39.5%
Gupta [16]	-	-	-	39.9%

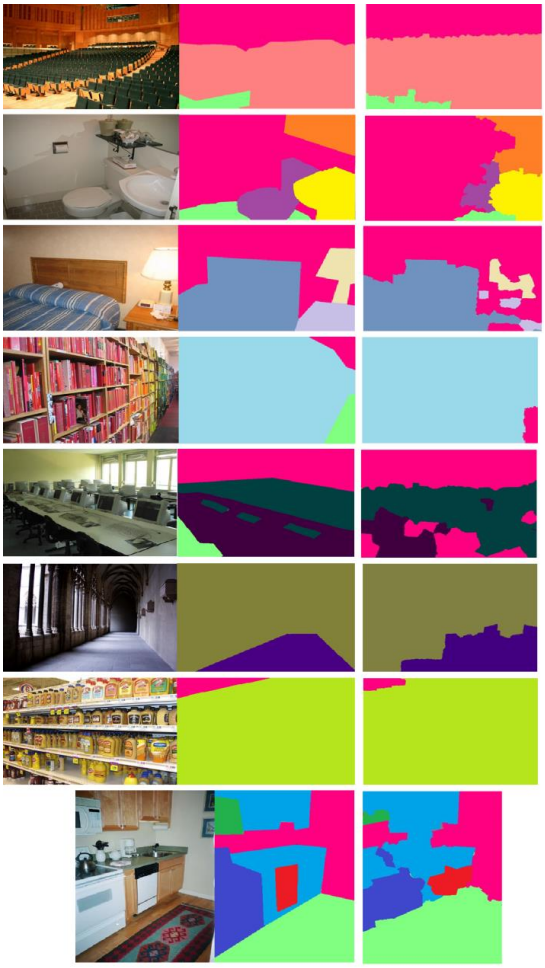

Figure 5: Segmentation results of the testing images for each class of indoor scene from image set d_1. The first column shows testing images, the second one is the output images of desired labeling for testing images, and the third one is the labeled output testing image.

References

[1] M. Szummer, R.W. Picard, "Indoor–outdoor image classification", IEEE International Workshop Content Basee Access Image Video Databases, ICCV '98, 1998.

[2] A. Bosch, X. Mu noz, and R. Marti," A review: Which is the best way to organize/classify images by content?" Image and Vision Computing (25), 778–791, 2007.

[3] A. Guerin-Dugue, A. Oliva, "Classification of scene photographs from local orientation features", Pattern Recognition (21), 1135–1140, 2000.

[4] S. Ramalingam, J.K. Pillai, A. Jain, Y. Taguchi, "Manhattan Junction Catalogue for Spatial Reasoning of Indoor Scenes", CVPR, IEEE, 3065-3072, 2013.

[5] I. Ulrich, I. Nourbakhsh, "Appearance-based place recognition for topological localization", IEEE International Conference on Robotic and Automation, 2000.

[6] A.Oliva , A.Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope", International Journal Computer Vision(42), 145–175, 2001.

[7] Z. Niu, G. Hua, X. Gao, Q. Tian, "Context aware topic model for scene recognition" , Conference Computer Vision and Pattern Recognition (CVPR), IEEE , 2743-2750, 2012.

[8] S. Achar, S.G. Narasimhan, "Multi Focus Structured Light for Recovering Scene Shape and Global Illumination", ECCV (1), 205-219, 2014.

[9] A. Quattoni, A.Torralba, "Recognizing indoor scenes", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[10] P. Espinace, T. Kollar, A. Soto, N. Roy, "Indoor scene recognition through object detection", Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2010.

[11] F. J. Kämäräinen, J.K. Buch, A. Krüger, "Indoor objects and outdoor urban scenes recognition by 3d visual primitives", *Computer Vision–ACCV Workshop on Big Data in 3D Computer Vision*, Springer, 2014.

[12] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation", ECCV, 2014.

[13] S. Gupta, P. Arbelaez, J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images", Computer Vision and Pattern Recognition (CVPR), IEEE, 564-571, 2013.

[14] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, "Indoor segmentation and support inference from RGBD images Computer Vision" ,ECCV, 746-760,2012.

[15] S. Wan, C. Hu, J. K. Aggarwal, "Indoor Scene Recognition from RGB-D Images by Learning Scene Bases", ICPR, 3416-3421, 2014.

[16] S. Gupta, P. Arbeláez, R.B. Girshick, J. Malik," Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation" , International Journal of Computer Vision 112(2), 133-149,2015.

[17] J.T. Barron,J. Malik, "Intrinsic Scene Properties from a Single RGB-D Image", CVPR,17-24, 2012.

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

[19] L. Xie, J. Wang, B. Guo, B. Zhang, Q. Tian, " Orientation Pyramid Matching for Recognizing Indoor Scenes" ,CVPR, 3734-3741,2014.

[20] F. Cakir, U. Güdükbay, O. Ulusoy, "Nearest-Neighbor based Metric Functions for indoor scene recognition", CVIU 115(11), 1483-1492, 2011.

[21] C. Galleguillos, A. Rabinovich, S. Belongie, "Object categorization using co-occurrence, location and appearance", CVPR 2008.

[22] A. Torralba, A. Oliva, "Semantic organization of scenes using discriminant structural templates", International Conference on Computer Vision, Korfu, Greece, 1253–1258, 1999.

[23] B. Micusik, J. Kosecka , "Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry" , IEEE Workshop on Video-Oriented Object and Event Classification (VOEC), held jointly with International Conf. on Computer Vision (ICCV), Japan, 2009.

[24] Y. Boykov, O. Veksler, R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2(11), 1222-1239, 2001.

[25] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features", International Conference on Computer Vision, 1999.

[26] N. Dalal, NB. Triggs, "Histograms of Oriented Gradients for Human Detection", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[27] T. Ojala, M. Pietikäinen, D. Harwood, "Comparative Study of Texture Measures with Classification Based on Feature Distributions", Pattern Recognition (29), 51-59, 1996.

[28] A. Bosch, A. Zisserman, X. Mun Oz, "Scene classification via plsa" , European Conference on Computer Vision, vol. 4, Graz, Austria, 517–530, 2006.

[29] N.M. Elfiky, et al., "Compact and adaptive spatial pyramids for scene recognition", Image and Vision Computing (30), 492-500, 2012.