A Contextual Image Descriptor for Scene Classification

Kelly Assis de Souza Gazolli
Universidade Federal do Espírito Santo
Vitória, ES, Brazil
Email: kasouza@ifes.edu.br

Evandro Ottoni Teatini Salles
Universidade Federal do Espírito Santo
Vitória, ES, Brazil
Email: evandro@ele.ufes.br

***Abstract*—Image representation is an important issue for many visual computing tasks such as image classification, understanding, feature extraction and retrieval. This paper presents a new image descriptor for scenes classification, called CMCT (Contextual Mean Census Transform), that combines distribution of local structures with contextual information. Experimental results on commonly used datasets demonstrate that the proposed method could achieve competitive performance against previous methods.**

***Keywords*-visual descriptor; scene classification; contextual information; census transform;**

I. Introduction

Classifying scenes involves associating a label with an image based on its content. This task is useful service for image retrieval systems based on content, creating photo albums automatically and to help the construction of interfaces which allows computer use by visually impaired people. In many situations, scene classification helps intelligent agent acquire knowledge of the environment. That is essential to an appropriate agent interaction with the environment. However, the scene classification task is quite challenging because there are a great number of possible classes. Also, some scenes are ambiguous and indeed, even human beings can be unsure about these classifications. In addition, the variation in illumination and scale could be daunting.

This paper proposes a visual descriptor that captures structural properties, by modeling distribution of the image local structures, and adds contextual information, by modeling the distribution of structures formed by neighbor local structures. The inclusion of contextual information can help differentiate patches in a scene that are equal but have very different neighbors.

A. Related Work

An important tasks in scene recognition is the image representation. A common approach is to employ global features, which is inspired by literature on human perception. These methods consider the image as a whole and use low-level features, such as color [1] [2], edge response [2], texture [1], to represent the characteristics of the scene. A global feature called ”gist” [3] employs a visual attention model to combine global color, intensity and orientation features to represent the scene what could be sufficient for separating scenes with significant differences in the global properties. Oliva and Torralba [4][5] proposed a formal approach to build the ”gist” of the scene and provided a statistical summary of the spatial layout properties (naturalness, openness, expansion, depth, roughness, complexity, ruggedness, symmetry) of the scene. However, if scenes with similar global characteristics (e.g. bedroom vs. sitting room) are to be differentiated, then global features may not be discriminative enough [6].

Another way to represent an image is employing local features [7] [8] [9] [10]. In this approach, the image is divided into parts or regions on which individual features are computed. The collection of these local descriptors shape the final representation. One advantage in this approach is that a change of an image part does not affect all the representation components, what could be useful when parts of a scene are lost by, for example, occlusion. In this sense, the Scalar Invariant Feature Transform (SIFT) [11] [12] became a very popular local descriptor. This method transforms an image into a large collection of local features vectors, each of which is invariant to translation, scaling and rotation and partially invariant to illumination changes.

A common problem in scene classification is that the collection of local features usually has a large size. Therefore, to reduce the final size of the image representation, generally, the bag-of-word approach is used. This method models an image by the co-occurrences of a number of visual components or topics. Many variants of this model has been proposed. Lazebnik et al. [13] proposed a spatial pyramid, a technique which works by partitioning the image into increasingly fine sub-regions. Quelhas et al. [14] showed that a textlike bag-of-visterms (histogram of quantized local visual features) is suitable for scene classification and used probabilistic latent semantic analysis (pLSA) to find intermediate topics. Ergul and Arica [15] proposed a pyramid of Latent Topics. Despite good results, this approach has some disadvantages. The codebook should be large enough so that each image could be properly represented by the histogram [16], thus, the codebook size depends on the dataset. Furthermore, the codebook-building process is often computationally intensive, which limits efficiency of its application [16].

Recently, Wu and Rehg [17] proposed CENTRIST (Cen-

sus Transform Histogram), a holistic representation that captures structural properties, rough geometry and generability by modeling distribution of local structures. CENTRIST is easy to implement, has nearly no parameter to tune, and is invariant to illumination. However, is not invariant to rotation. Also recently, Qin and Yung [6] proposed a method based on contextual visual words, in which the contextual information from neighbor region and the regions from coarser scales are included. The results reported in [6] show that the use of such extra information improves the final values.

B. Technique Overview

For recognizing a scene, firstly, local structures are extracted using Modified Census Transform (MCT). Then, the distribution of the local structures is modeled. A new image is generated in which the pixel value at a given position is the MCT value of the pixel at the correspondent position in the original image. Then, the local structures are extracted from this new image and their distribution is modeled. The distributions obtained from the two images are concatenated to form the new features. The whole process is schematized in Figure 1.

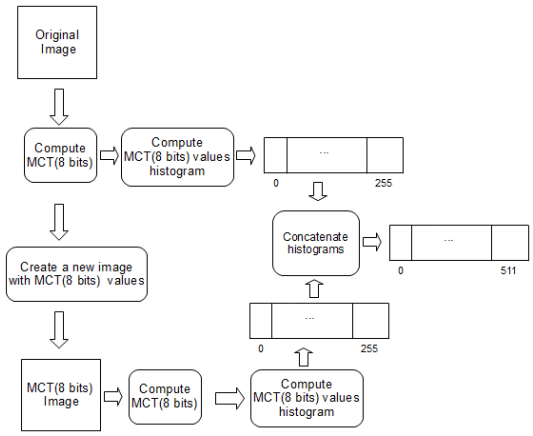

Figure 1. Contextual descriptor extraction process.

The rest of this article is structured as follows: in section II a description of Census Transform and Modified Census Transform are presented, Section III describes the proposed method, CMCT, Section IV presents the experimental results and finally in Section V the conclusions are provided.

II. Technical Background

A. Census Transform Histogram

Census Transform is a nonparametric local transform originally designed for establishing correspondence between local patches [18]. Census Transform, $\mathcal{C}(x)$, compares the intensity value of a pixel with its neighborhood pixels. If the center pixel is bigger than or equal to one of its neighbors, a bit 1 is set in the corresponding location. Otherwise, a bit 0 is set, as follow

$$\mathcal{C}(x) = \bigotimes_{y \in \mathcal{N}(x)} \zeta(I(x), I(y)), \;\; \zeta(m,n) = \begin{cases} 1, & m \geq n \\ 0, & otherwise \end{cases} \tag{1}$$

where $I(x)$ is the gray value of the central pixel at x position, $I(y)$ is the gray value of the pixel at y position, $\bigotimes$ denotes the concatenation operation and $\mathcal{N}(x)$ defines a local spatial neighborhood of the pixel at x so that $x \notin \mathcal{N}$. If a 3 x 3 neighborhood size is utilized, the 8 bits generated from intensity comparisons can be put together in any order and converted to a decimal number in [0, 255], namely CT. A histogram of CT values for an image, namely CENTRIST [17], can be used as visual descriptor.

B. Modified Census Transform Histogram

Census Transform has been shown to be invariant to camera bias, gamma correction and illumination and is useful in order to extract edges in images [19]. However, in some cases, Census Transform could not capture the edge information [20]. To overcoming this shortcoming Bernard and Andreas [20] proposed the Modified Census Transform (MCT), $\Gamma(x)$, which is computed in the following manner. A 3 x 3 window of pixels is considered and the mean $\overline{I}(x)$ of the pixels is computed. Every pixel in the 3 x 3 window is then compared with $\overline{I}(x)$. If the pixel is bigger than or equal to $\overline{I}(x)$, a bit 1 is set in the corresponding location. Otherwise, a bit 0 is set, as follows

$$\Gamma(x) = \bigotimes_{y \in \mathcal{N}'(x)} \zeta(I(y), \overline{I}(x)), \;\; \zeta(m,n) = \begin{cases} 1, & m \geq n \\ 0, & otherwise \end{cases} \tag{2}$$

where $\overline{I}(x)$ is the mean of the intensity values in the 3 x 3 window of pixels centered at x, $I(y)$ is the gray value of the pixel at y position and $\mathcal{N}'(x)$ is a local spatial neighborhood of the pixel at x so that $\mathcal{N}'(x) = \mathcal{N}(x) \cup x$. In the Modified Census Transform technique, 9 bits are generated and converted to a decimal number in [0, 511], namely, here, MCT. Thus, a 512 bins histogram of MCT values can be used as visual descriptor.

In this work, we adopt Modified Census Transform. However, although comparing pixels in a 3 x 3 window with $\overline{I}(x)$ of the pixels, $\overline{I}(x)$ is compared only with the center pixel neighbors. That is, we do not compare the mean with the center pixel. Thus, we generate 8 bits, instead of 9, which are converted to a decimal number in [0, 255], i.e., we generate a smaller descriptor. Moreover, we adopt $\zeta(m,n) = 1$, if $m > n$. In order to differentiate Modified Census Transform with 9 bits of Modified Census Transform with 8 bits and with $\zeta(m,n) = 1$, if $m > n$, we refer to this last as MCT(8 bits). Finally, A histogram of MCT(8 bits) for an image is used as a visual descriptor.

In Figure 2 one can see that the original Census Transform does not capture the local image structure correctly in some cases while the MCT(8 bits) can do it. Note that, although the sample patches have different structures, they generate the same CT value, while the MCT(8 bits) generates different values.

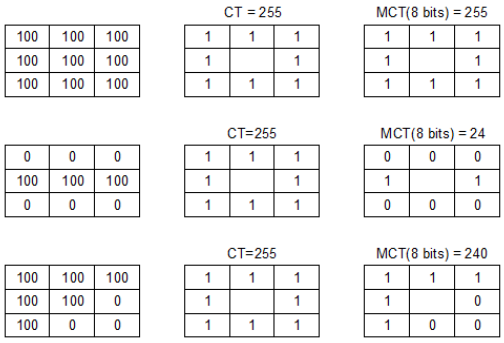

Figure 2. Examples where CT does not capture different structures while MCT(8 bits) does.

III. Contextual Mean Census Transform Histogram

The Contextual Mean Census Transform (CMCT) integrates contextual information with local structures information for differentiating patches that have similar structures, but have significant difference in their neighborhood. For accomplishing this task, this approach takes into consideration information of neighborhood windows in the MCT(8 bits) computation, by creating a new local structure from the local structure of the patch and from the local structures of its neighboring patches. We believe these additional information can improve the image representation. We call these informations coming from outside windows by *context*.

The steps for the Contextual Mean Census Transform generation are as follow. First, MCT(8 bits) is computed for all pixels. Then, a histogram of MCT(8 bits) is obtained. A new image is created in which the original image pixels are replaced by the correspondent MCT(8 bits) values as shown in Figure 3.

MCT(8 bits) is computed on the new image pixels and a new histogram is generated. Then, MCT(8 bits) histogram for the original image and the MCT(8 bits) histogram for the new image are concatenated, generating a new descriptor.

The MCT(8 bits) maps a 3 x 3 image patch to one of 256 possible values. The MCT(8 bits) value acts as an index to different local structures. When pixel value is replaced by the MCT(8 bits) value, as illustrated in Figure 4, the information type changes from intensity to a local structure index. In this way, when the Modified Census Transform operation is applied on the new image, it performs comparisons between local structures, obtaining a different type of information: the

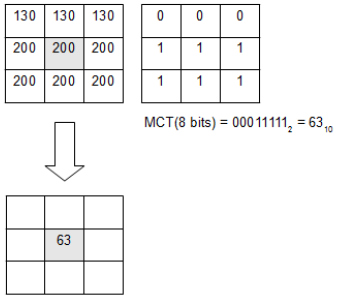

Figure 3. Pixel value is replaced by the MCT(8 bits) for obtaining contextual information.

relationship between local structures. So, when the MCT(8 bits) is computed for a patch in the new image, the resultant value refers to a local structure formed from the local structures of the original image.

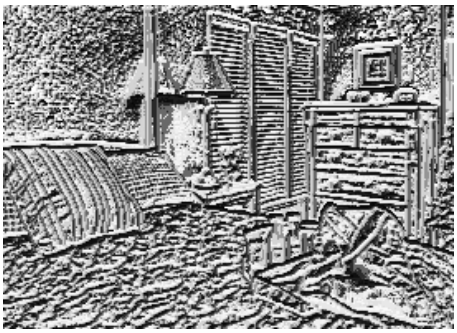

Figure 4. An example of an image with the gray values replaced by MCT(8 bits). This image is taken from 15-category dataset.

IV. Experiments

In this section, we investigate the effectiveness of our representation and compare it with existing works.

A. Datasets and Setup

Our descriptor has been tested on three public data sets, which including indoor scenes and outdoor scenes. These datasets are described bellow:

- 8-category scenes provided by Oliva and Torralba [4]. This dataset contains 2,688 color images, divide into 8 categories; with the number of images in each category ranging from 260 to 410. The 8 category are: coast (360 images), forest (328 images), mountain (274 images), opencountry (410 images), highway (260 images), insidecity (308 images), tallbuilding (356 images) and street (292 images). The size of each image is 256 x 256.
- 15-category dataset [13], which is an extension of dataset described above by adding 7 new scene categories: bedroom (216 images), kitchen (210 images),

livingroom (289 images), office(215 images), suburb (241 images), industrial (311 images) and store (315 images). This dataset contains 4,486 gray-values images in total. The image size is approximately 300 x 250. Figure 5 depicts the samples from this dataset.

- 67-class indoor scene [21]. This dataset contains 15,620 images. The scenes varies from corridor to bakery. This dataset poses a challenging classification problem [21].

In the experiment, each category in a data set is split randomly into a training set and a test set. The random splitting is repeated 5 times, and the average accuracy is reported, as adopted by [17]. We employed linear kernel SVM (Support Vector Machine), a pattern classifier introduced by [22], to make scene classification. For 8-category scenes and 15-category scenes, an amount of 100 images in each category are used for training and the remaining images constitute the testing set. Following [21], in 67-class indoor dataset, we use 80 images in each category for training and 20 images for testing. All color images were converted to gray scale. The log frequency weighting was applied in the histogram values. The log frequency weighting is a technique used in Information Retrieval [23] whereas relevance does not increase proportionally with term frequency. The log frequency weight of term t in a document d ($W_{t,d}$) is

$$W_{t,d} = \begin{cases} 1 + log(tf_{t,d}), & tf_{t,d} > 0, \\ 0, & otherwise \end{cases} \quad (3)$$

where $tf_{t,d}$ is number of occurrences of term t in a document d.

V. Results and Discussion

The first dataset we consider is the 15-class scene dataset. We achieve 76.87 $\pm$ 0.58% accuracy in this dataset. Figure 6 presents the confusion matrix from one run on 15-class scene dataset. We observe that the highest recognition rate is 95% for Suburb class. The biggest confusion happens between bedroom and livingroom, which have similar elements. We human may confuse them due to the small inter-class variation.

Table I compares the classification performance of the proposed method on 15-category dataset with existing results in literature and with MCT(8 bits). Since the goal is to compare the efficiency of the descriptors, only the results of experiments in which the images were not partitioned into increasingly fine subregions, i.e., with levels number equal to zero, were considered. In [13], two kinds of features were used in the experiments: weak features, points whose gradient magnitude in a given direction exceeds a minimum threshold, and strong features, SIFT features computed from 16 x 16 image patches. Strong features showed to have a higher accuracy than weak features. CMCT, in its turn, outperforms the weak features and the strong features (SIFT 200 clusters centers and SIFT 400 cluster centers). CMCT also outperforms CENTRIST, since the CMCT provides more information about the local image structures than CENTRIST. In [24], a method for scene categorization by integrating region contextual information into the bag-of-words approach is used. This method also is overcomed by CMCT. By comparing MCT(8 bits) and CMCT, one can see that the addition of contextual information improves performance, since using only MCT(8 bits) we achieve 73.71 $\pm$ 0.30%.

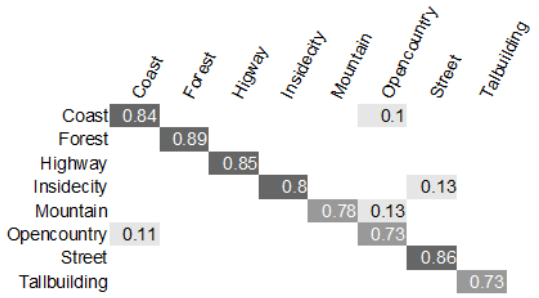

Figure 7. Confusion matrix from one run for 8-class scene recognition experiment.

Table I
Comparison classification results for 15-categories with existing works

Method and Feature Type	Accuracy(%)
CENTRIST[17] - CENTRIST not using PCA	73.29 $\pm$ 0.96
SPM [13] - 16 channel weak features	66.80 $\pm$ 0.6
SPM [13] - SIFT 200 cluster centers (0 levels)	72.20 $\pm$ 0.6
SPM [13] - SIFT 400 cluster centers (0 levels)	74.80 $\pm$ 0.3
Spatial Envelope [4] - gist	73.28 $\pm$ 0.67
RCVW [24] - Region Contextual Visual Words	74.5
MCT(8 bits) - Modified CT Histogram	73.71 $\pm$ 0.30
Ours - CMCT	**76.87 $\pm$ 0.58**

In the 8-category scene class we achieve 79.91 $\pm$ 0.99% accuracy. Figure 7 presents the confusion matrix from one run on 8-category dataset. We observe that the highest recognition rate is 89% for Forets class. Confusion happens between insidecity and street, opencountry and mountain and opencountry and coast, wich have similar spatial layout.

Table II shows experimental results for 8-category dataset. Using gist descriptors the recognition accuracy is 82.60 $\pm$ 0.86%, which is greater than the results achieved by the CMCT. However, on the 15-category dataset which adds several indoor categories, the accuracy using gist dropped to 73.28 $\pm$ 0.67%, which is lower than CMCT accuracy. As in 15-category dataset, CMCT outperforms MCT(8 bits) and CENTRIST.

In the 67-class indoor scene, the experiments performed by [21] with gist achieved about 21% average recognition accuracy. When it is used local and global information to represent the scenes, the accuracy was improved to 25%. By

Figure 5. Three images from each 15 scene categories. The categories are: coast, forest, opencountry, mountain, insidecity, tallbuilding, highway, bedroom, street, kitchen, livingroom, office, store, suburb and industrial (from top to bottom and left to right).

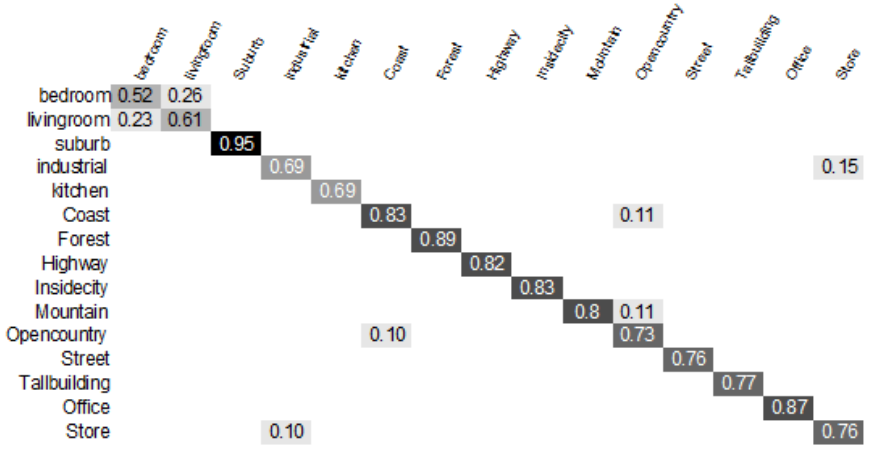

Figure 6. Confusion matrix from one run for 15-class scene recognition experiment.

Table II
EXPERIMENTAL RESULTS FOR 8 SCENE CATEGORIES DATASET

Feature type	Accuracy (%)
gist	**82.60 ± 0.86**
CENTRIST (0 levels)	76.49 ± 0.84
MCT(8 bits)	77.07 ± 0.68
CMCT	**79.91 ± 0.99**

using CMCT we achieve 25.82 ± 0.72%. The experiments performed using CENTRIST with 0 levels achieved 22.46 ± 0.84%. As one can see, in this challenging dataset, CMCT reaches better results than the presented experiments.

VI. CONCLUSION

In this paper we proposed a visual descriptor, the Contextual Mean Census Transform, that captures structural properties by modeling distribution of local structures and combines it with contextual information, obtained from the distribution of local structures formed from local structures in the original image, to perform scene recognition task. For capturing structural properties and contextual information CMCT uses a modification of CENTRIST, since the original method does not capture the local image structure correctly in some cases. Comparing the results of MCT(8 bits) and CMCT, one can see that the introduction of contextual information improves the image representation. Furthermore, our method preserves the advantages of CENTRIST (easy to implement, almost no parameter to tune, low illumination dependence) as showed in the Tables I and II. As CENTRIST, our method is not invariant to rotation. In our future research, we intend to use some form of associating spatial layout information, as subregions of different resolution levels, and include another type of information to improve the classification performance.

ACKNOWLEDGEMENTS

Kelly Gazolli gratefully acknowledge the support from IFES - Instituto Federal do Espírito Santo.

REFERENCES

[1] E. Chang, K. G. K. Goh, G. Sychay, and G. W. G. Wu, "Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines," pp. 26–38, 2003.

[2] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang, "Content-based hierarchical classification of vacation images," *Proceedings IEEE International Conference on Multimedia Computing and Systems*, vol. 1, no. 2, pp. 518–523, 1999.

[3] C. Siagian and L. Itti, "Gist: A mobile robotics application of context-based vision in outdoor environment," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05 Workshops*, vol. 3, pp. 1063–1069, 2005.

[4] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[5] A. Oliva, "Gist of the scene," *Nature*, vol. 17, pp. 251–257, 2005.

[6] J. Qin, "Scene categorization via contextual visual words," *Pattern Recognition*, vol. 43, no. 5, pp. 1874–1888, 2010.

[7] Y. K. Y. Ke and R. Sukthankar, "Pca-sift: a more distinctive representation for local image descriptors," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2004 CVPR 2004*, vol. 2, no. 3, pp. 506–513, 2004.

[8] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan, "Categorizing nine visual classes using local appearance descriptors," *ICPR Workshop on Learning for Adaptable Visual Systems*, vol. 17, p. 21, 2004.

[9] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *Tenth IEEE International Conference on Computer Vision ICCV05 Volume 1*, vol. 2, pp. 1458–1465, 2005.

[10] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05*, vol. 2, pp. 524–531, 2005.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, no. 8, pp. 1150–1157 vol.2, 1999.

[12] ——, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Volume 2 CVPR06*, vol. 2, pp. 2169–2178, 2006.

[14] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575–1589, 2007.

[15] E. Ergul and N. Arica, "Scene classification using spatial pyramid of latent topics," *2010 20th International Conference on Pattern Recognition*, pp. 3603–3606, 2010.

[16] S. K. Wei Liu and M. Gabbouj, "Robust scene classification by gist with angular radial partitioning," *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, pp. 2–4, 2012.

[17] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2010.

[18] J. Zabih, Ramin Wood, "Non-parametric local transforms for computing visual correspondence," *Science*, vol. 801, pp. 151–158, 1994.

[19] S. Bhavani, A. Thawani, V. Sridhar, and K. R. Ramakrishnan, "Illumination invariant face recognition for frontal faces using modified census transform," *TENCON 2007 2007 IEEE Region 10 Conference*, vol. 2, pp. 1–4, 2007.

[20] B. Froba and A. Ernst, "Face detection with the modified census transform," *Gesture*, pp. 91–96, 2004.

[21] A. Quattoni and A. Torralba, "Recognizing indoor scenes," *Proceedings IEEE CS Conf. Computer Vision and Pattern Recognition*, pp. 413–420, 2009.

[22] V. Vapnik, "The support vector method of function estimation," *Nonlinear Modeling advanced blackbox techniques Suykens JAK Vandewalle J Eds*, pp. 55–85, 1998.

[23] G. Salton and M. J. McGill, *Introducrion to modern information retrieval*. New York: McGraw-Hill, 1983.

[24] S. Liu, D. Xu, and S. Feng, "Region contextual visual words for scene categorization," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11 591–11 597, 2011.