IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. VI (Sep. - Oct. 2016), PP 31-36
www.iosrjournals.org

Image Multi-Classification using PHOW Features

Shereen A. Hussein[1], HowidaYoussryAbd El Naby[2], Aliaa A. A. Youssif[3]
[1]*Computer Science Department, Helwan University, Cairo, Egypt*
[2]*Computer Science Department, Misr University for Science & Technology, Cairo, Egypt*
[3]*Computer Science Department, Helwan University, Cairo, Egypt*

Abstract: *Automatic labeling and classification of a vast number of images is a huge challenge, so machines are used as a part of image classification and annotation is turned into a prerequisite to adapt to the high improvement of advanced digital image innovations consistently. Scale Invariant Feature Transform (SIFT) is an image descriptor for image-based matching and recognition; this descriptor is used for computer vision purposes like point-matching between different views and object recognition in the same view. SIFT features are regarded as an efficient way for image classification due to its usefulness demonstration under real-world conditions. Also, representing these features in bag-of-words (BOW) model and spatial pyramid model adds the ability to distinguish spatial distribution to the former.*

Keywords: *SIFT Features, Bag of words, Spatial Pyramid Description, Classification, Multi-SVM*

I. Introduction

The main characteristic of data classification is dealing with plenty of class labels and a small number of samples. This aspect makes feature extraction and selection a vital strategy to guarantee reliable and meaningful results for data classification alongside different advantages such fewer data storage and computation cost [1-4]. For any object to be recognized, some point features must be extracted then described in a proper model. For features extraction, SIFT features are the best descriptors [5] because of its robustness in translation, rotation and scaling transformations and illumination varieties. For image description, a spatial pyramid of visual words model is used to overcome the problem of dismissal of spatial information of local descriptors in the basic bag-of-words model [6]. For image labeling prediction, multi-class classification methods [7-9] is used which depend on the combination of several two-class (binary) classifiers each with its feature variables. One-versus-all (OVA) binary classifiers are the most common strategy developed to separate one class from different classes. The rest of the paper is organized as follows. Section 2 gives a short review of automatic image classification. Section 3 presents the multi-class SVM classification using SIFT-PHOW features; Section 4 shows the experimental results and analysis of the classification on Corel5k dataset. Section 5 contains the conclusions and discussions. Section 6 represents future work.

II. Related Work

Automatic image annotation is regarded as a kind of multi-class image classification. Image classification needs various types of features to describe the image contents. Such classification techniques based on low features (colors, textures, and boundaries) have been studied for years in the area of image retrieval. Sometimes, global features such as color and texture cannot successfully recognize objects at the same kind [4]. These works usually perform supervised learning these image features, but image classes are semantically different. Also, objects with the same semantic may have different colors such as cars. Because of the low accuracy of image object recognition based on global features, researchers have changed the focus to the local image features [10-18]. Although there are three kinds of local features based on points, boundaries and regions, most researchers today focus on those based on points. The extraction of local features based on points is partitioned into two stages: 1) keypoint detection and 2) generation of feature descriptor. In later studies of descriptors, David Lowe in 1999 [19] proposed the big scale invariant feature transformation description. SIFT is proved to be the best through literature. Choi, M. J. et al. in 2010 [20] introduced a technique for creating fuzzy multimedia ontologies automatically using SIFT feature extraction and BOW for feature quantization. Determining the number of visual words to quantize image feature vectors into during codebook improvement is a controversial subject. Tsai in 2012 [21] clarified that although most implementations of the BOW modeling depend on 1000 visual-words, the number of visual words is reliant on the dataset. Bosch et al. in 2007 [22] used 1500 as the number of visual words developed from SIFT features vectors of sample images in all experimentation including BOW. The utilization of these methodologies uncovered the classification process to constrained distinctiveness because of a small number of visual words in the codebook, and high processing overhead when a codebook with an excess of visual words is used [23]. Consequently, research into the determination of the number of visual words required during BOW modeling will provide a method for dispensing with some unnecessary calculation overhead. Another issue about BOW modeling of images is the loss of classification accuracy because of the negligence for the spatial area of the visual words during the modeling process [24]. Verbeek et al. in 2007 [25] used Random Field theory to provide spatial data alongside the BOW Models for Probabilistic Latent Semantic Analysis (PLSA) classification of image regions. David M. Blei in 2003 [26] proposed Latent Dirichlet Allocation (LDA) for image classification. Zhang et al. in 2010 [27] used at the searching step of a retrieval system the Geometry-preserving visual phrase that encodes more spatial data into the BOW models, so local spatial interactions between the visual words can be represented. Lazebnik et al. in 2006 [28] proposed the Spatial Pyramid in which histograms are computed for multi-level image regions, and after that connected to form a single spatial histogram. Bosch et al. in 2007 [29] extended the idea of a spatial pyramid to the development of an image signature known as Pyramid Histogram of Oriented Gradient (PHOG).

III. Proposed System (Phow-Msvm)

In the algorithm, the keypoints and descriptors of all training images are extracted. Then these descriptors are clustered into N centroids. For executing this procedure, the K-means clustering algorithm is used. This paper focuses on the independent descriptors extraction and can be used as Bag-of-Words (BOW) in the image. The multiclass SVMs are trained by the pyramid histogram of words (PHOW). For an image to be queried, descriptors are extracted. The dictionary formed from PHOW is used as the basis to map each descriptor to its equivalent visual word. This result is passed to SVM to classify and annotate the image. The proposed framework is shown in Fig.1. There are certain modules involved in the algorithm which are:
1- Computation of SIFT features.
2- Compute histograms based on bag-of-words.
3- Compute PHOW Descriptors.
4- Classification using SVM

1. The PHOW Descriptors

Image description by Pyramid of Histograms of Visual Words (PHOW) method is an extension to the bag-of-words (BOW) model in which the extracted SIFT image features treated as words. It considers the local information feature of the image. The method is implemented as in Fig. 2. PHOW [28, 29] overcomes the drawback of BOW of unavailability of spatial image features information by dividing the image into fine sub-regions (pyramids) and concatenating the histogram of each of these regions to the histogram of the original image with a suitable weight.

When color images are processed, they are converted from RGB space to HSV color space with the SIFT feature extracted from each channel. For grayscale images, only the intensity is used. Hence, the resulting SIFT feature dimension is 128*3 for color images and 128 for grayscale images. Once the SIFT features are obtained, “bag of words” model is used to quantize them into visual words by k-means clustering. Thus the image is represented by a histogram of visual word occurrences.

1.1 SIFT Features Extraction

There are numerous features for the object, but important ones are extracted to provide an object feature description. This description can be used in locating the object in an image containing many other objects. So many methods used in feature extraction but SIFT is the most well-known one in computer vision. As SIFT ability to detect the salient keypoints and extract the discriminative descriptions of the appearance. SIFT keypoints are not influenced by a considerable lot of the complexities experienced in different techniques such as translation, rotation, scaling and also the noise effects. As an object could be recognized among other objects in a large image, the same object of multiple images could be recognized.

In the SIFT algorithm [4, 13, 17], the key points are detected via a DOG (Difference of Gaussian) pyramid created using a Gaussian filtered copy of the image. Each of the detected key points is then described by a 128-dimensioned histogram of the gradients and orientations of pixels within a box centered on the key point. Despite the fact that SIFT stays one of the best descriptors as far as exactness, the 128-measurements of the descriptor vector makes its feature extraction process relatively computationally costly [30, 31]. As known that the features dimensionality reduction proportionally reduces the required computations for classification which improves the speed of the process but it reduces the accuracy.

1.2 Image Description

There is a need for image description after the feature extraction and before the classification. Thus, the performance of image classification and annotation is dependent on the reliability of the image feature representation (image mathematical model) [28]. Using a normalized histogram or a vector for sometimes quantized features occurs on an image representation are the most common approaches. The most popular of these methods is the Bag-of-Visual words (BOW) image model.

1.3 Bag-of-Words

The Bag-of-word model (BOW) model is a popular image representation for classification purposes, which uses a visual-words histogram for an efficient image representation [6, 17, 18]. An essential part of BOW image representation is the visual codebook. It is the process that uses k-means clustering with a Euclidean distance metric for the vectors quantization. These vectors represent the features of an image into visual words. The computational requirement of this stage is very high and therefore regarded as the most costly part of the BOW modeling process, and the computation time reduction trails often lead to noisy visual-words [21, 23]. For our purposes, we choose a value of N (number of words) = 1500. This parameter provides our model with a balance between underfitting and overfitting.

2. Multi-Class SVM

Support Vector Machines (SVMs) are one of the most well-known supervised classification methods because of its high classification performance with less sensitivity of the dimensionality curse of large scale data in numerous applications. It uses set of labeled training data where are several features and one class label to generate input-output mapping function. SVM objective is a model development based on the training data which predicts the test data label given only the features [7-9]. SVMs were initially developed for two-class classification by determining the isolating hyperplane with maximum distance nearest to training set points which accomplish most extreme separation. As a part of machine learning, SVMs learn from the notable cases represented as data points for data classification.

SVMs extended from two-class to multi-class problems. A multiclass SVM classifier can be obtained by training several classifiers and combining their results. Two of the most used ways for multi-class SVM development [32,33] are OVO "one-versus-one" or OVA "one-versus-all". They decompose the multiclass problem and construct the binary classifiers for the combination. These binary classifiers in combination are built based on the way OVO-SVM or OVA-SVM. OVA-SVM in classification problems constructs M binary classifiers and each binary classifier classifies one class (positive) versus all other categories (negative). The most common combination strategy for OVA-SVM is Winner-Takes-All (WTA) [34] which assigns a sample to the class with the maximum decision function among all the M binary classifiers. In this paper, once the pyramid histogram of words features for all training images are obtained, they are given into OVA-SVM with WTA combination strategy as it achieves comparable performance with faster speed than other strategies.

IV. Experimental Results & Discussion

The used dataset with specified category names and also shows the experimental results of the proposed approach (PHOW-MSVM) accompanied with discussion of the classification accuracy using hardware configuration as follows:-

Processor (CPU):- Intel Core i7, ***Operating System: -*** Microsoft Windows 7, ***System Type: -*** 64-bit operating system, ***Memory:*** - 8GB RAM, ***Storage: -*** 500 GB internal hard disk.

1. Dataset

The Corel dataset is the most widely used benchmark for image classification and retrieval. As this study focuses on classification than retrieval, a limited number of categories is used. Corel5k dataset is used with 20 visual concepts/categories and a total of 5000 images (50 concepts). Under each concept category, 100 images are present. The concepts are very different such as African, Beach, Bus, Car, Dinosaur, Dog, Elephant, Fashion, Flower, Food, Historical, Horse, Lazard, Mountain, Sunset, Antique, Battleship, Skiing, Waterfalls, and Dessert. Each category in the dataset is divided into 75 training and 25 testing subsets and given to multi-SVM after computing PHOW image descriptors..

2. Precision & Recall

As the Corel5K dataset has 50 visual concepts, just 20 are used. For measuring the classification accuracy for each concept as in Fig.3, precision and recall [35-36] are used. The equations are in (1), (2) and (3) and the true positives, false positives, true negatives and false negatives are as follows:

True positives (tp):- the number of images correctly labeled as belonging to this class.
False positives (fp):- the number of images incorrectly labeled as belonging to this class.
True negatives (tn):- the number of images correctly not labeled as belonging to this class.
False negatives (fn):- the number of images incorrectly not labeled as belonging to this class

$$\text{Recall} = \frac{tp}{tp+fn} \qquad (1)$$

$$\text{Precision} = \frac{tp}{tp+fp} \qquad (2)$$

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn} \qquad (3)$$

3. Discussion

For the previous considered 20 category classification of the corel5K dataset, BOW model achieved 64.6% accuracy. Spatial Pyramid proposed which divide the image into some cells and then concatenate the histogram for each one to form one histogram. It solves the problem missing the spatial image information in BOW model and improves the classification accuracy. Confusion matrix of two dimensions (actual and predicted labels) with an identical set of values is allowing the visualization of the algorithm performance. While applying PHOW on gray-level images achieved accuracy 83%, the proposed PHOW-MSVM on color images with 128*3 resulting dimensions achieved accuracy 88.8% as shown in Fig.4. Since these promising experimental results, PHOW-MSVM highly recommended in the classification especially for semantic purposes with considering high hardware configuration.

V.Figures And Tables

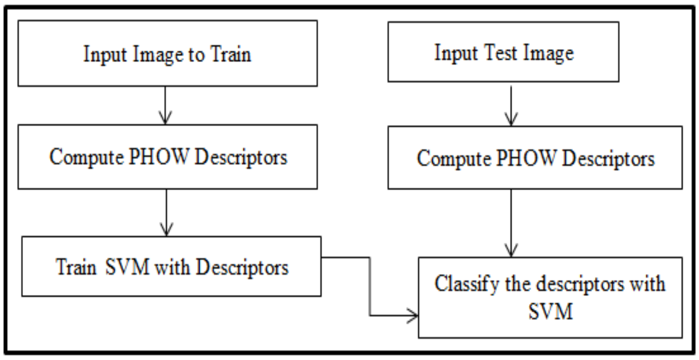


Figure1. Proposed PHOW-MSVM algorithm

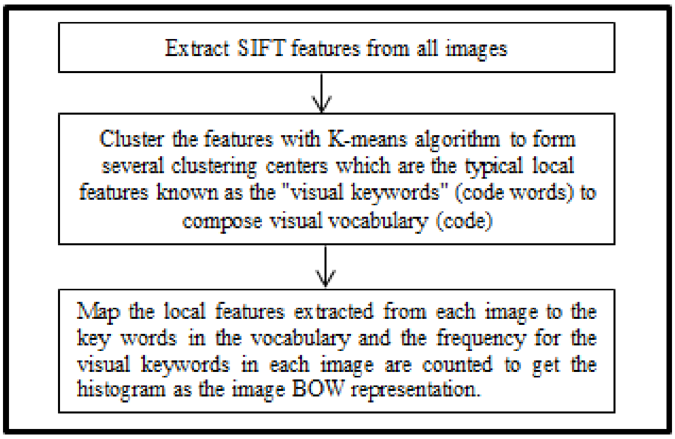
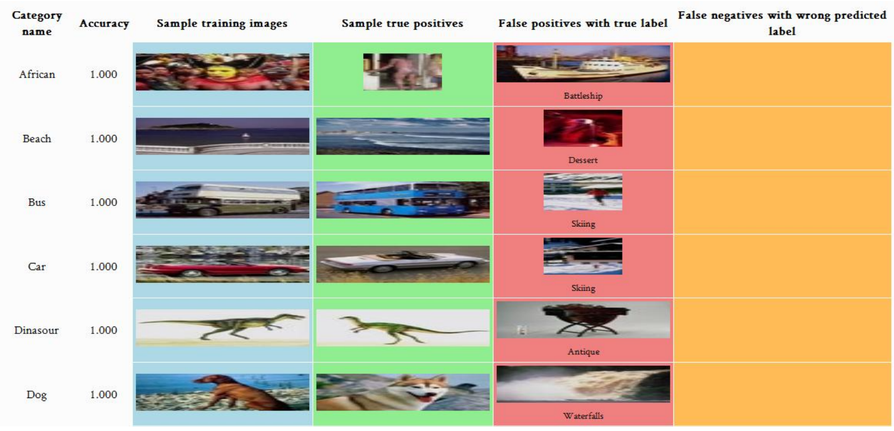
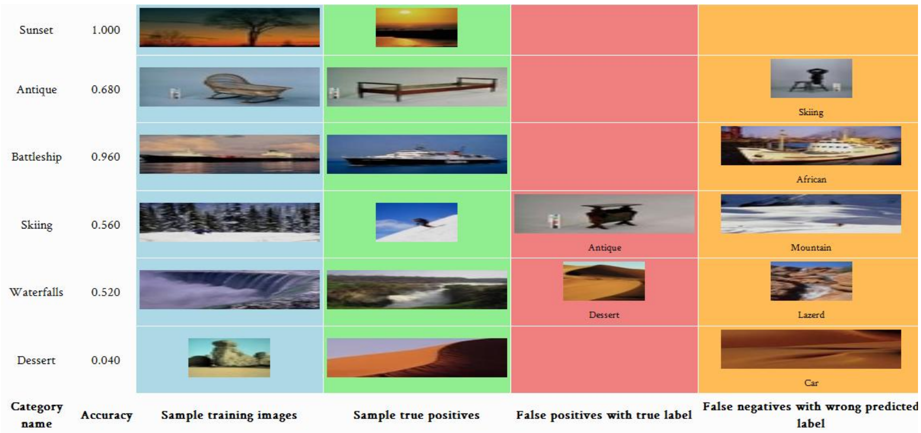

Extract SIFT features from all images
Cluster the features with K-means algorithm to form several clustering centers which are the typical local features known as the "visual keywords" (code words) to compose visual vocabulary (code)
Map the local features extracted from each image to the key words in the vocabulary and the frequency for the visual keywords in each image are counted to get the histogram as the image BOW representation.

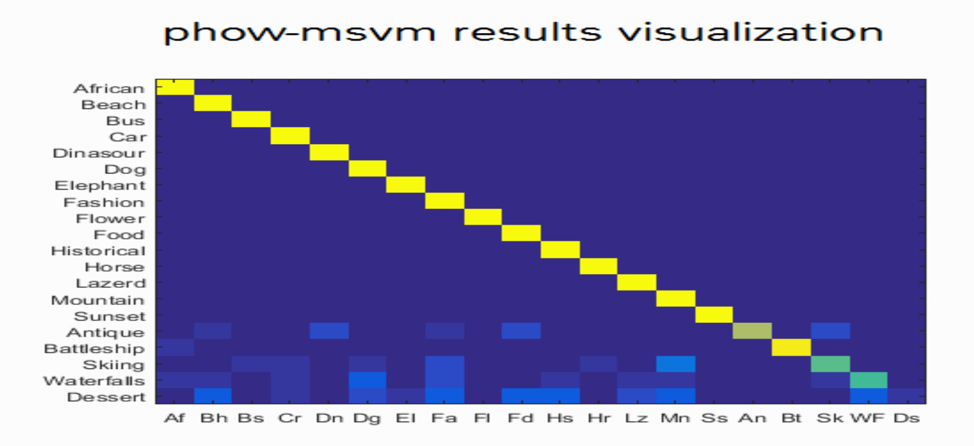
Figure2. PHOW Descriptor Steps



Category name	Accuracy	Sample training images	Sample true positives	False positives with true label	False negatives with wrong predicted label
African	1.000			Battleship	
Beach	1.000			Dessert	
Bus	1.000			Skiing	
Car	1.000			Skiing	
Dinasour	1.000			Antique	
Dog	1.000			Waterfalls	



Figue3. Classification Accuracy for categories



List of shortened category names:- Af-->African, Bh-->Beach, Bs-->Bus, Cr--> Car, Dn-->Dinasour, Dg-->Dog, El-->Elephant,Fa-->Fashion, Fl-->Flower, Fd-->Food, Hs-->Historical, Hr-->Horse, Lz-->Lazerd, Mn-->Mountain, Ss-->Sunset, An-->Antique, Bt-->Battleship, Sk-->Skiing, WF-->Waterfalls, Ds-->Dessert

Figure4. Accuracy (mean of diagonal confusion matrix) is 0.888

VI. Conclusion

This paper demonstrated that the PHOW-MSVM method could operate on a large-scale image dataset with an efficient image labeling inference due to its ability to combine the images visual and textual information. The automatic image classification using OVA-MSVM on the PHOW model of extracted SIFT features of images set accomplishes the best performance. The main topic of image annotation and classification in recent studies is BOW image model for feature representation and how to enhance its discriminative power by such techniques as image segmentation, vector quantization, and visual vocabulary construction. From comparisons of related work, the most used method for keypoint detection is Difference of Gaussian (DOG) and then each key point is represented by the SIFT feature. The extracted feature vectors quantized using k-means clustering algorithm with 1500 visual words which depend on the dataset used. Finally, image representation by mapping the feature vectors to visual words with spatial pyramid information.

VII. Future Work

According to the comparative results, there are some future research directions. First, trying to use other region-based features rather than point-based SIFT feature descriptor for vector quantization. Second, assigning some rules of how to decide the number of visual words based on datasets used. Thirdly, improve the learning models performance over different kinds of datasets, such as different dataset sizes and different image contents (a single object per image and multiple objects per image).

References

[1]. Olson, David L., Delen, Dursun, Advanced Data Mining Techniques, XII, Springer Publications, 2008.

[2]. Han, Jiawei, MichelineKamber, and Jian Pei. Data mining: concepts and techniques. Elsevier, 2011.

[3]. Li Xirong, SnoekCees, Worring Marcel, “Learning Social Tag Relevance By neighbor voting”, IEEE TRANSCATION MM, 11(7):1310-1322, 2009.

[4]. Toews Matthew; Wells, William M. "SIFT-Rank: Ordinal Descriptors for Invariant Feature Correspondence", IEEE International Conference on Computer Vision and Pattern Recognition. pp. 172–177, 2009

[5]. Lowe, David G. "Distinctive image features from scale-invariant key-points," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[6]. Faheema, A.G.; Rakshit, Subrata "Feature Selection using Bag-Of-Visual-Words Representation," in Advance Computing Conference (IACC), IEEE 2nd International, Patiala, 2010.

[7]. Qi Xiaojun and Han Yutao., "Incorporating multiple SVMs for automatic image annotation", Pattern Recognition.40, 2, Pages.728-741, 2007.

[8]. GohKheng-Swee.; Chang Edward.; Li Bing; , "Using one-class and two-class SVMs for multiclass image annotation," Knowledge and Data Engineering, IEEE Transactions on , vol.17, no.10, pp. 1333- 1346, 2005.

[9]. Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[10]. Shi Rui., FengHuamin, Chua Tat-Seng., Lee Chin-Hui, "An adaptive image content representation and segmentation approach to automatic image annotation", Proceedings of the International Conference on Image and Video Retrieval, pp. 545–554, 2004.

[11]. Li Xirong, Chen Le, Zhang, Lei, Lin Fuzong, and Ma Wei-Ying, "Image annotation by large scale content based image retrieval", 2006.

[12]. Wang Xin-Jing, Zhang Lie, Jing Feng, and Ma Wei-Ying., "Annosearch: Image auto annotation by search In IEEE Conference on Computer Vision and Pattern Recognition", New York, USA, 2006.

[13]. Kalyani, B.Bobade; Shital, V.Jagtap, "Automatic Image Annotation by classification using SIFT features", International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 0882, Volume 3, Issue 3, 2014.

[14]. Christian, Hetschel; Sebastian, Stober; Andreas, Nurnberger and Marcin, Detyniecki, "Automatic Image Annotation using a visual dictionary based on reliable Image Segmentation", LNCS 4918,pp. 45-56, 2008.

[15]. Jeon, Jiwoon, Victor Lavrenko, and RaghavanManmatha, "Automatic image annotation and retrieval using cross-media relevance models", Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003.

[16]. Olaode, Abass; Naghdy, Golshah; Todd, Catherine, "Unsupervised Classification of Images: A Review", International Journal of Image Processing (IJIP), 2014.

[17]. Mansourian, Leil; Abdullah, Muhamad Taufik; Abdullah, Lilli Nurliyana; and Azman, Azreen; "Evaluating classification strategies in Bag of SIFT Feature method for Animal Recognition", Research Journal of Applied Sciences, Engineering and Technology 10(11):1266-1272, 2015.

[18]. Liu, Jialu,; "Image Retrieval based on Bag-of-Words model",. arXiv preprint arXiv:1304.5168, 2013.

[19]. Dengsheng Zhang, Md. Monirul Islam, GuojunLu,"A review on automatic image annotation techniques", Pattern Recognition, Volume 45, Issue 1, Pages 346-362, 2012.

[20]. Choi, Myung Jin; Lim, Joseph J.; Torralba, Antonio; and Willsky, Alan S.; "Exploiting hierarchical context on a large database of object categories", IEEE Computer vision and pattern recognition (CVPR) conference, pp. 129-136, 2010.

[21]. Tsai Chih-Fong, "Bag-Of-Words Representation in Image Annotation: A Review," International Scholarly Research Network, vol. 2012, pp. 1-19, 2012.

[22]. Bosch Anna, Zisserman Andrew and Munoz Xavier, "Scene Classification via PLSA," Computer Visionand Robotics Group, University of Girona, Girona, 2006.

[23]. Yang, Jun; Jiang, Yu-Gang; Hauptmann, Alexander G.; Ngo, Chong-Wah;" Evaluating Bag-of-Visual-Words Representations in Scene Classification", MIR 07 Proceedingd of the international workshof on Workshop on multimedia information retrieval pages 197-206, 2007.

[24]. Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Fretias, David M. Blei,and Michael I. Jordan, "Matching words and pictures". Journal of Machine Learning Research, 3:1107–1135, 2003

[25]. Verbeek, Jakob; and Triggs, Bill; "Region Classification with Markov Field Aspect Models," in Computer Vision and Pattern Recognition, 2007. CVPR '07.IEEE Conference, Minneapolis, MN, 2007.

[26]. Blei, David; Ng, Andrew; and Jordan Michael, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[27]. Zhang Yimeng, JiaZhaoyin and T. Chen Tsuhan, "Image Retrieval with Geometry-Preserving Visual Phrases," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference, Providence, RI, 2010.

[28]. Lazebnik Svetlana, SchmidCordelia and Ponce Jean, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference, Illinois, 2006.

[29]. Bosch Anna, Zisserman Andrew and Munoz Xavier, "Representing shape with a spatial pyramid kernel," in CIVR, Amsterdam, 2007.

[30]. Tuhin, Shukla; Nishchol, Mishra and Sanjeev, Sharma, "Automatic Image Annotation using Surf Features", International Journal of Computer Applications, vol. 68, no. 4, 2013.

[31]. Bay, Herbert, TinneTuytelaars, and Luc Van Gool. "Surf: Speeded up robust features", Computer Vision–ECCV 2006, Pages.404-417, 2006.

[32]. Chapelle Olivier; Haffner Patrick; Vapnik, Vladimir, "Support vector machines for histogram-based image classification", Neural Networks, IEEE Transactions on , vol.10, no.5, pp.1055-1064, 1999.

[33]. Cusano Claudio, CioccaGianluigi, SchettiniRaimondo, "Image annotation using SVM", Proceedings of the Internet Imaging IV, vol. 5304, SPIE, 2004.

[34]. Gonzalez R., Woods R, Eddins S, Digital Image Processing Using MATLAB, 2nd ed., Pearson Education.

[35]. Lei Yinjie, Wong Wilson, Liu Wei, and Bennamoun Mohammed. "An HMM-SVM-based automatic image annotation approach", Computer Vision–ACCV 2010, Pages.115-126, 2011.

[36]. Powers, David M W , "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness& Correlation", Journal of Machine Learning Technologies,vol.2, no.1, 2011.