A Performance Evaluation of Content-Based Image Retrieval for Scene Categorization

Ahsiah Ismail, Mohd Yamani Idna Idris, Mohamad Nizam Ayub, Lip Yee Por

***Abstract*—**Content-based image retrieval remains a critical problem in computer vision**.** In this paper, we study the performance of various content-based image retrieval technique for recognizing the object and scene. We conduct the comparative survey to compare the state of the art bag of words (BOW) framework with other method to help the researcher to understand more and enable researcher selecting the most suitable technique. We carried out the experiment and tested with 3 publicly dataset that are Caltech 101, Caltech 256 and 15-Scene Category dataset with BOW method. We also compare and evaluate the effect on the number of cluster toward the computational time and the accuracy. In addition, the significant of different feature extraction method applied in BOW performance is analyzed. In conclusion, we discuss on several key potential research topics towards the content-based image retrieval.

Keywords **—** Bag of words, Content-based image retrieval, Image classification, Object recognition, Scene categorization

I. INTRODUCTION

Content-Based Image Retrieval (CBIR) is an active and important research area in computer vision [1, 2]. CBIR is the application of computer vision technique to the image retrieval problem. CBIR refers to image retrieval according to its content from a collection by similarity [1, 2]. The images mainly can be described based on their numerical information basis, which can be obtained by object recognition techniques [3]. There are many CBIR system has been develop, however, only a small number of research on retrieval based on object recognition [4]. The prior research on retrieval based on the object recognition only limited to classify a single class objects such as horse or people [4]. In this paper, we study on the crowding, occlusions, and cluttered scene images. The recognizing task becomes even more difficult when it is involved crowding, occlusions, cluttered in background environment, noise, poor quality and deformable object and also with the present of many objects in the same scene [5]. We evaluate on the performance evolution of content-based Image retrieval for scene categorization using the object recognition technique for scene categorization. The object recognition technique is required to determine the common object that usually exists in the same scene. Thus, this can be used to categorize the scene accordingly. Object recognition in an image is the fundamental challenge [6] in computer vision. In this work, we interested in recognizing scene categories using the image taken by normal rectilinear camera lens. In order to successfully categorize the scene, features of the images should be well extracted and the images descriptors should be presented well to describe the object image.

There are various content-based image retrieval technique have been exploit. Despite all of the efforts, the current CBIR technique still have limitation which is the object image must be the same viewpoint as the image use for the training[4]. The emergence of a large number of feature analysis techniques and machine learning classifiers reported a year has increase the content-based image retrieval research. The comparative study on existing technique is necessary to help and guide the researchers in comparing or selecting the most suitable content-based image retrieval technique. In this paper, we thoroughly review and discuss the existing technique for content-based image retrieval, which certainly vital for further progress in image retrieval area.

II. BAG OF WORD TECHNIQUES

There are many content-based image retrieval technique has been proposed, however the problem of retrieving images remain largely unsolved. The technique in content-based image retrieval as shown in Fig. 1.Generally, recent content-based image retrieval methods mostly rely on bag of word model [7].

The Bag of word framework is among the popular and well-known feature representation in information retrieval [8]. This method had been applied by J. Sivic and A. Zisserman in image and video retrieval field [9]. This method not only has shown promising result for object categorization[10-13] but also for image annotation and retrieval tasks [11, 14-16]. The latest research find that the Bag of word is the most popular image classification [17]. Generally this method shown promising result and successful in classifying images for object categorization [10, 12, 18]. Besides its efficiency in recognizing the object, this method is fast and easier to implement. It can be improve so that it can be robust to the occlusion object, clutter, non-rigid deformation and viewpoint change [19, 20].

Despite all the advantages that had been discussed, nevertheless the bag of word method disregard all the spatial

Ahsiah Ismail, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.
ahsiahismail15@siswa.um.edu.my
Mohd Yamani Idna Idris, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.
yamani@um.edu.my
Mohamad Nizam Ayub, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.
nizam_ayub@um.edu.my
Lip Yee Por, Faculty of Computer Science and Information Technology, University of Malaya , Malaysia
porlip@um.edu.my

layout information[21]. Disregard completely the spatial layout between visual words leads to the missing information of the image composition and the features spatial arrangement, which is useful as a powerful cues for the scene classification task[17, 19].

A part from that, in bag of word method, detecting and selecting the keypoints from images is one of the process involve in recognizing the object. The BOW feature usually required the large number of keypoints [22-24]. Due to the large number of detected keypoints in bag of word, this lead to high computational cost in the vector quantization stage. The most crucial part in the bag of word framework is the high computational cost in vector quantization stage [23, 24] [22].

This study, review some of the latest improvement towards BOW method. The techniques are spatial pyramid matching, sparse coding and Power Normalization.

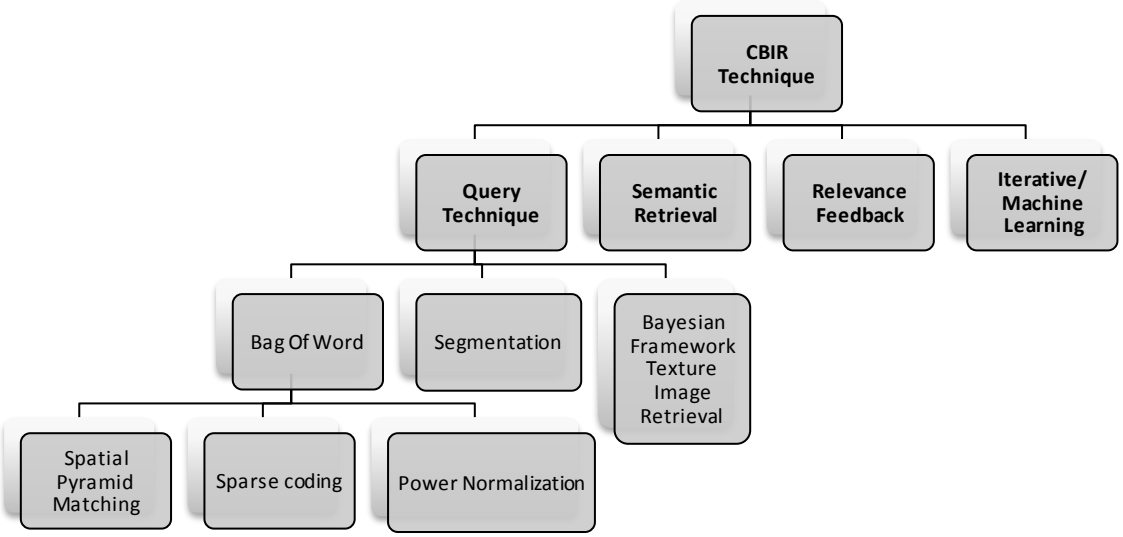

Fig. 1 Content-based image retrieval technique

III. SPATIAL PYRAMID MATCHING

In bag of word framework, the visual words is generate without taking into account its spatial location information. However, the spatial information is really important in object recognition especially in scene categorization. This problem had been address by Lazebnik et. all by introducing the spatial pyramid matching [21] . These methods add the spatial information in the unstructured bag of word model.

IV. DATASET

In previous section, we have discussed numerous content-based image retrieval techniques. The existence of a great number of competitive techniques causes the natural problem in selecting and choosing the best technique. Generally, to address this problem, the empirical comparison is required. Thus, in order to examine and compare the performance for each of the technique, the dataset and experimental setting standardization is extremely crucial.

There are some public dataset available on the Internet. Different types of dataset normally serve different task in the computer vision field such as object detection, object categorization, object segmentation and others. In this section, we present three diverse standard datasets that are the most widely used, which are Caltech 101, Caltech 256, and 15-Scene categories datasets. The Caltech 101 and Caltech 256 datasets used for object categorization while 15-scene category dataset are used for scene categorization. Each dataset properties are shown in TABLE 1.

TABLE 1: COMPARISON DATASET PROPERTIES

Dataset	Categories	Image per category		Total	Resolution (Pixel)
		Min	Max		
Caltech 101	102	31	800	9248	300*x*200
Caltech 256	257	80	827	30868	450 *x* 600
15- Scene	15	211	411	4551	300*x*250

In Caltech 101 dataset, the images are lack of clutter and clean images with most of object are centered and occupy the images. While in Caltech 256, the dataset are large and more cluttered images compared to Caltech 101 and intended to address limitation in Caltech 101. The object images not left right aligned. In contrast to both Caltech dataset, the images in dataset in 15-scene categories consist scene images including wide range both indoors and outdoor environment.

V. COMPARISON OF OBJECT RECOGNITION TECHNIQUES

In this section, we evaluate and present in detail the bag of word framework performance and some of the latest improvement. We also thoroughly evaluate on the impact number of cluster toward the computational time and the accuracy in bag of word approaches.

A. Experimental Setup

Based on the Bag of word framework specified before, we carried out the experiment to evaluate the accuracy and computational time. We tested with the most widely used standard dataset Caltech 101, Caltech 256 and 15-Scene Category. The objective of this work is to classify the object and scene images according to their category classed.

In this paper, we follow the experiment setup from the previous studies [21, 24, 25] by randomly split the images in every category into two sets, training and testing sets. The average accuracy value for each run was calculated and recorded. The experiments were carried out ten times using the same setup for each category in this datasets. In each experiment, the training set consists of ten images randomly selected from each category and the remaining --images are the testing set.

B. Performance of Bag of Word

Given an input of object and scene images, we categorize the object by extracting the keypoints from the images. We used the bag of word model with surf as a feature extraction since SURF is the fast feature extraction method and has good performance [26].

This approach followed by creation of the visual codebook using the K-Means algorithm. Then, we generate the codebook from the images patch sampled in the training set. Based on the result obtain in [21], 200 visual words will perform the best. In this work, we generate the codebook with three different visual vocabulary sizes, which are 100, 200 and 400 clusters to compare the effect on the precision and time performance. We used the K-means algorithm to cluster the SURF vector into code words. K-means algorithm is the vector quantization method to cluster the N descriptor into k cluster. Each descriptor is belongs to nearest mean cluster, serving as a prototype of the cluster. This led to the division of data space into Voronoi cells.

To predict the classification score for each category of images respectively, the Support Vector Machine (SVM) is used in this work. We choose this technique as the classifier since it is one of the most popular classifier for image classification and widely used as classifier in Bag of Word framework [27]. The multi class classification used in this work is one-versus-all rule. The classifier learned to classify the tested images between each class and assign to the highest respond label of the classifier.

Our first experiment is tested on Caltech 101 dataset and the second dataset tested is Caltech 256. The third dataset involve in this experiment are 15-scene category dataset. The examples of images including the category name are shown in Fig. 2. The average classification accuracy and computational time for the experiment tested on each dataset are summarizing in Table 2, Table 3 and Table 4.

TABLE 2: VISUAL WORDS VOCABULARY SIZE AGAINST CLASSIFICATION ACCURACY AND COMPUTATIONAL TIME OVER CALTECH 101

Method	Feature Type	Visual Words	Average Accuracy Rates (%)	Training Time (s)	Testing Time (s)
BOW	Surf	100	26.2	3686	4992
BOW	Surf	200	27.2	3728	5102
BOW	Surf	400	29.2	4181	5058

TABLE 3: VISUAL WORDS VOCABULARY SIZE AGAINST CLASSIFICATION ACCURACY AND COMPUTATIONAL TIME OVER CALTECH 256.

Method	Feature Type	Visual Words	Average Accuracy Rates (%)	Training Time (s)	Testing Time (s)
BOW	Surf	100	8	13721	65391
BOW	Surf	200	10	16246	107178
BOW	Surf	400	10	30256	96531

TABLE 4: VISUAL WORDS VOCABULARY SIZE AGAINST CLASSIFICATION ACCURACY AND COMPUTATIONAL TIME OVER 15 SCENE CATEGORY.

Method	Feature Type	Visual Words	Average Accuracy Rates (%)	Training Time (s)	Testing Time (s)
BOW	Surf	100	57	1484	744
BOW	Surf	200	60	1603	821
BOW	Surf	400	63	1216	759

Fig. 2 Example one of the images from each category in 15-Scene category dataset

The confusion matrix of the experiment tested on 15-scene category dataset using 400 visual words is shown in Fig. 3.

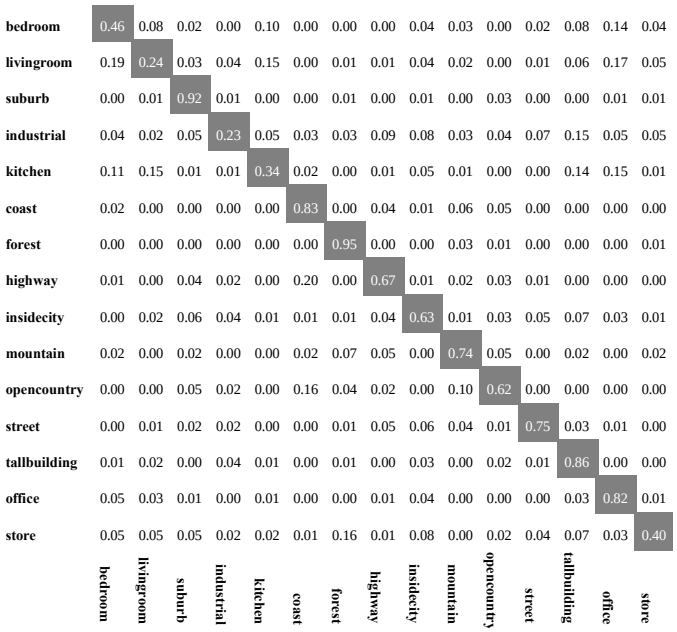

Fig. 3 The confusion matrix for 15 scene Categories dataset

As shown in TABLE 2, the accuracy gradually improves when increase the number of visual words. However the computational time of training increase when increase the number of visual words. In this table, we can see that when we change the parameter of visual words to 400, the training time is high compared to 200 visual words but the testing time is less. This is because it is consume a lot of time during the training to form the visual words based on the large number of cluster. However it is reduce the time in the classification process. Almost similar result obtains in Caltech 256 dataset with our first experiment using Caltech 101. The accuracy improves as increase the number of visual words. However it also increase the computational time. In TABLE 3,we can see that less accuracy obtained in Caltech 256. This is due to the very large number of category involved in this dataset, which are 257 categories. This may greatly effect the misclassification during the classification process. Similarly, the result obtained for 15-scene categories in

TABLE 4 also shows when we increase the number of visual words, the accuracy increases and also the computational time become high. From Fig. 3 we can see that, the accuracy for the outdoor environment category is high compared to indoor environment. This is due to the clutter and occlusion and diverse viewpoint object found in the indoor scene images.

To further evaluate the effect on the visual vocabulary size towards the accuracy and time performance, we present the evaluation score on the accuracy and time while increasing the number of visual words in Fig. 4 and Fig. 5.

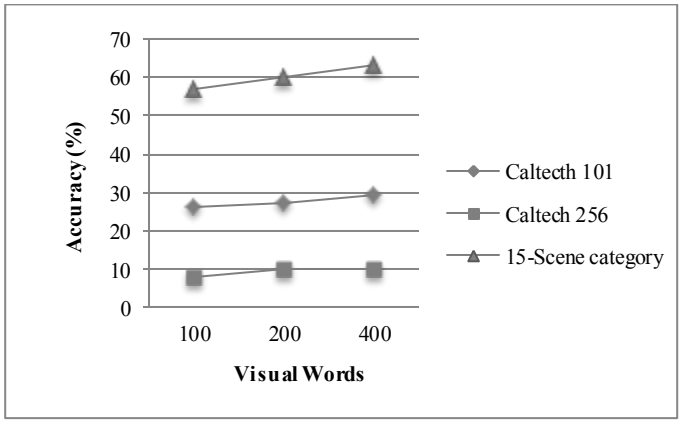

Fig. 4. Evaluation of the visual words vocabulary size against accuracy over Caltech 101, Caltech 256 and 15 Scene Category dataset

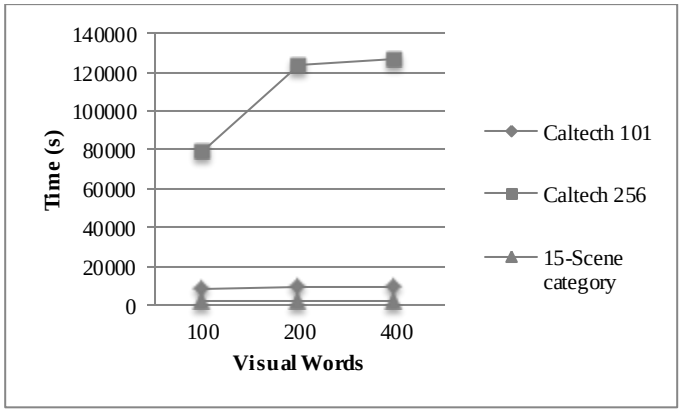

Fig. 5. Evaluation of the visual words vocabulary size against time over Caltech 101, Caltech 256 and 15 Scene Category dataset

C. *Performance of Spatial Pyramid Matching*

After evaluating the state of the baseline bag of word, we evaluate on the importance of the spatial information since the Bag of Word framework discard the spatial information. The Spatial Pyramid Matching (SPM) framework with SIFT feature extraction is tested with Caltech 101 and 15-Scene Categories dataset with two different parameter number of visual words. We present the accuracy result in TABLE 5.

TABLE 5: ACCURACY PERFORMANCE OF SPATIAL PYRAMID MATCHING

Method	Dataset	Feature Type	100 Visual Words (%)	200 Visual Words (%)	400 Visual Words (%)
SPM [21]	Caltech 101	SIFT	Not-Reported	41.2	Not-Reported
SPM[21]	15-scene	SIFT	Not-Reported	72.2	74.8
SPM [28]	15-scene	SIFT	Not-Reported	Not-Reported	83.3

From the result obtained in TABLE 5, it shows that by include the spatial information in the bag of word can greatly increase the accuracy performance.

D. *Comparison performance with other method*

In our work, we have shown that the accuracy performance

and computational time of Bag of Word framework using SURF as feature extraction varies depending on the number of visual words. In order to demonstrate that our proposed method are competitive against other method, we present inTABLE 6, TABLE 7 and TABLE 8 the summary of the result obtained from previous study using the same dataset that we had tested in our work to access their performance.

TABLE 6: COMPARISON ACCURACY PERFORMANCE OF CALTECH 101 DATASET USING OTHER METHOD

Method	Feature Type	100 Visual Words (%)	200 Visual Words (%)	400 Visual Words (%)
BOW	Surf	26.2	27.2	29.2
BOW IKS1[24]	SIFT	20.77	23	Not-Reported
BOW IKS2[24]	SIFT	24.99	25.92	Not-Reported
SPM IKS1[24]	SIFT	38.98	41.91	Not-Reported
SPM IKS2[24]	SIFT	34.61	33.71	Not-Reported
SPM [21]	SIFT	Not-Reported	41.2	Not-Reported

TABLE 7: COMPARISON ACCURACY PERFORMANCE OF CALTECH 256 DATASET USING OTHER METHOD

Method	Feature Type	100 Visual Words (%)	200 Visual Words (%)	400 Visual Words (%)
BOW	Surf	8	10	10
BOW IKS1[24]	SIFT	5.9	6.7	Not-Reported
BOW IKS2[24]	SIFT	8.21	8.84	Not-Reported
SPM IKS1[24]	SIFT	9.22	11.23	Not-Reported
SPM IKS2[24]	SIFT	12.23	11.53	Not-Reported

TABLE 8: COMPARISON ACCURACY PERFORMANCE OF 15-SCENE CATEGORY DATASET USING OTHER METHOD.

Method	Feature Type	100 Visual Words (%)	200 Visual Words (%)	400 Visual Words (%)
BOW	Surf	57	60	63
SPM[21]	SIFT	Not-Reported	72.2	74.8
CENTRIST[25]	Without PCA	Not-Reported	73.29	Not-Reported

As can be seen from both TABLE 6 and TABLE 7, the Bag of Word with SURF descriptor obtained high accuracy results compared to SIFT descriptor. However, the SPM method obtained better accuracy result compared to Bag of Word model as shown in TABLE 5, TABLE 6, TABLE 7 and TABLE 8.

VI. DISCUSSION

Overall, as we can see that the accuracy for all of the dataset tested including scene categorization in both place indoor and outdoor increases as the large number of visual words is used. However, the computational time required increase when we increased the number of visual words. This is due to the large visual words size that sheer amount of time for the process of clustering in the vector quantization step. Based on some comparison from the previous work towards Bag of Word model, we believe that this model still have some limitation in terms of accuracy and time performance. Hence, this can lead to significant degradation performance in a more challenging task. Further, the vital research towards this model is required.

VII. CONCLUSION

In this paper, we presented a comprehensive survey emphasizing the recent achievement in bag of word in details as well as the evolution on latest improvement towards bag of word framework. The efficient features extraction from content images incorporates with the optimum classification method are required for the image retrieval process. The creation of generic category object in the large number of category dataset is important as the classification accuracy rates degrade when tested to the large number of category in the dataset. The CBIR technology appears to be interesting research area in developing commercial CBIR application for efficient an accurate image retrieval for management of image collections and drawing archives, electronic publishing and multimedia content creation. From the comprehensive survey that had been discussed in this paper, it shows that the BOW method still have some limitation. The continue and vibrant research is required in order to gradually approaching it.

ACKNOWLEDGMENT

Special thanks to Postgraduate Research Grant (PPP) – Research (PG040-2015B), University Malaya Research Grant (RP036(A,B,C)-15AET) and University of Malaya for providing support and material related to educational research.

REFERENCES

[1] F. Zhang, Y. Song, W. Cai, A. G. Hauptmann, S. Liu, S. Pujol, *et al.*, "Dictionary pruning with visual word significance for medical image retrieval," *Neurocomputing*, vol. 177, pp. 75-88, Feb 2016.

[2] M. Alkhawlani, M. Elmogy, and H. Elbakry, "Content-Based Image Retrieval using Local Features Descriptors and Bag-of-Visual Words," *International Journal of Advanced Computer Science and Applications*, vol. 6, pp. 212-219, Sep 2015.

[3] H. Chougrad, H. Zouaki, and O. Alheyane, "Bag of Features Model Using the New Approaches: A Comprehensive Study," *International Journal of Advanced Computer Science and Applications*, vol. 7, pp. 226-234, Jan 2016.

[4] Y. Li, "Object and concept recognition for content-based image retrieval," Citeseer, 2005.

[5] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, pp. 712-722, 2010.

[6] X. Zhang, Y.-H. Yang, Z. Han, H. Wang, and C. Gao, "Object class detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 46, p. 10, 2013.
[7] Y. Ren, "A comparative study of irregular pyramid matching in bag-of-bags of words model for image retrieval," *Signal Image and Video Processing*, vol. 10, pp. 471-478, Mar 2016.
[8] D. Aldavert, M. Rusinol, R. Toledo, and J. Llados, "A study of Bag-of-Visual-Words representations for handwritten keyword spotting," *International Journal on Document Analysis and Recognition*, vol. 18, pp. 223-234, Sep 2015.
[9] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 1470-1477.
[10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 2005, pp. 1816-1823.
[11] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, pp. 316-336, 2010.
[12] H.-L. Luo, H. Wei, and L. L. Lai, "Creating efficient visual codebook ensembles for object categorization," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 41, pp. 238-253, 2011.
[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," 2005.
[14] J. Fan, Y. Gao, and H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic concepts," in *Proceedings of the 12th annual ACM international conference on Multimedia*, 2004, pp. 540-547.
[15] E. Hörster and R. Lienhart, "Fusing local image descriptors for large-scale image retrieval," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1-8.
[16] G. Wang, Y. Zhang, and L. Fei-Fei, "Using dependent regions for object categorization in a generative framework," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 1597-1604.
[17] W. C. Lin, C. F. Tsai, Z. Y. Chen, and S. W. Ke, "Keypoint selection for efficient bag-of-words feature generation and effective image classification," *Information Sciences*, vol. 329, pp. 33-51, Feb 2016.
[18] C. Schmid, "Bag-of-features for category classification," *ENS/INRIA Visual Recognition and Machine Learning Summer School Lecture 25-29 July*, 2011.
[19] S. Lazebnik, C. Schmid, and J. Ponce, "Spatial pyramid matching," *Object Categorization: Computer and Human Vision Perspectives*, vol. 3, 2009.
[20] N. Kejriwal, S. Kumar, and T. Shibata, "High performance loop closure detection using bag of word pairs," *Robotics and Autonomous Systems*, vol. 77, pp. 55-65, Mar 2016.
[21] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 2169-2178.
[22] K. E. Van de Sande, T. Gevers, and C. G. Snoek, "Empowering visual categorization with the GPU," *Multimedia, IEEE Transactions on*, vol. 13, pp. 60-70, 2011.
[23] Z. W. Lu, L. W. Wang, and J. R. Wen, "Image classification by visual bag-of-words refinement and reduction," *Neurocomputing*, vol. 173, pp. 373-384, Jan 2016.
[24] W.-C. Lin, C.-F. Tsai, Z.-Y. Chen, and S.-W. Ke, "Keypoint selection for efficient bag-of-words feature generation and effective image classification," *Information Sciences*, vol. 329, pp. 33-51, 2016.
[25] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, pp. 1489-1501, 2011.
[26] P. Panchal, S. Panchal, and S. Shah, "A comparison of SIFT and SURF," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 1, pp. 323-327, 2013.
[27] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," 2009.
[28] J. Liu and M. Shah, "Scene modeling using co-clustering," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1-7.

About Author (s):

Ahsiah Ismail is a Ph.D student at Faculty of Computer Science and Information Technology, University of Malaya. She received her Master Degrees in Computer Science from University of Malaya. Her research interests are in the area of Image Processing.

Mohd Yamani Idna Idris is a senior lecturer at Department of Computer System and Information Technology, University of Malaya. He received a Ph.D from University of Malaya. His research interests including the Sensor Network, Computer Vision and Embedded System, Digital Signal Processing, and Image Processing.

Mohamad Nizam Ayub currently is a senior lecturer at Department of Computer System and Information Technology, University of Malaya. He's a Ph.D holder from Paisley University. His research interest includes Edutainment and Interactive Multimedia.

Lip Yee Por received his Ph.D. from University of Malaya, Malaysia. Currently, he is a Senior Lecturer at Department of System and Computer Technology, Faculty of Computer Science and Information, University of Malaya. In general, his research interests are Bio-Informatics, Computer Security, Neural Network, Grid Computing, and e-Learning Framework.