

Research Article

Evaluating Classification Strategies in Bag of SIFT Feature Method for Animal Recognition

Leila Mansourian, Muhamad Taufik Abdullah, Lilli Nurliyana Abdullah and Azreen Azman

Department of Multimedia, Faculty of Computer Science and Information Technology,

Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Abstract: These days automatic image annotation is an important topic and several efforts are made to solve the semantic gap problem which is still an open issue. Also, Content Based Image Retrieval (CBIR) cannot solve this problem. One of the efficient and effective models for solving the semantic gap and visual recognition and retrieval is Bag of Feature (BoF) model which can quantize local visual features like SIFT perfectly. In this study our aim is to investigate the potential usage of Bag of SIFT Feature in animal recognition. Also, we specified which classification method is better for animal pictures.

Keywords: Bag of feature, Content Based Image Retrieval (CBIR), feature quantization, image annotation, SIFT feature, Support Vector Machines (SVM)

INTRODUCTION

In Content Based Image Retrieval (CBIR) (Qi and Snyder, 1999) proposed in the early 1990s, images are automatically indexed by extracting their different low level features such as texture, color and shape. Semantic gap is a well-known problem among Content Based Image Retrieval (CBIR) systems. This is caused by humans tendency to use concepts, such as keywords and text definitions, to understand images and measure their resemblance. Although low-level features (texture, color, spatial relationship, shape, etc.) are extracted automatically by computer vision techniques, CBIR often fails to describe the high-level semantic concepts in user's mind (Zhou and Huang, 2000). These systems cannot effectively model image semantics and have many restrictions when dealing with wide ranging content image databases (Liu *et al.*, 2007).

Another problem caused by using low level features like texture, color and shape is that they need image digestion. But Scale-Invariant Feature Transform (SIFT) (Lowe, 1999) is a robust feature in scaling, rotation, translation, illumination and partially invariant to affine distortion. Also, there is no need to digest images. The only thing we need to do is to quantize SIFT features by well-known Bag of Feature (BoF) technique.

Furthermore, in most of the previous works we observed that there isn't any appropriate investigation on animal annotation and animal picture recognitions because they have the same environments which caused low accuracy. For this reason, our objective in this study, is to investigate the potential usage of bag of

SIFT feature in animal recognition. And find out which kind of classification is more suitable to our animal recognition system.

LITERATURE REVIEW

At the starting point of BoF methodology we must identify local interest regions or points. Then we can extract features from these points, both of which described in the following section.

Interest point detection: There are several distinguished methods which are listed below (Mikolajczyk *et al.*, 2005).

Harris-Laplace regions: In this method corners are detected by using Laplacian-of-Gaussian operator in scale-space.

Hessian-Laplace regions: Are localized in space at the local maxima of the Hessian determinant and in scale at the local maxima of the Laplacian-of-Gaussian.

Maximally Stable External Regions (MSERs): Are components of connected pixels in a threshold image. A water-shed-like segmentation algorithm is applied to image intensities and segment boundaries which are stable over a wide range of thresholds that define the region.

DoG regions: This detector is appropriate for searching blob-like structures with local scale-space maxima of the difference-of-Gaussian. Also it is faster and more

Corresponding Author: Muhamad Taufik Abdullah, Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

This work is licensed under a Creative Commons Attribution 4.0 International License (URL: <http://creativecommons.org/licenses/by/4.0/>).

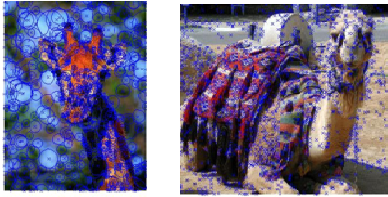


Fig. 1: Detected SIFT features of Harris-Laplace key points as circles

compact (less feature points per image) than other detectors.

Salient regions: In circular regions of various sizes, entropy of pixel intensity histograms is measured at each image position.

In our study we used Harris-Laplace for finding key points.

SIFT feature descriptors: After interest Points are detected we can describe them by their features like SIFT. SIFT is an algorithm published by Lowe (1999) for detecting and describing local features in images. Each SIFT key point is a circular image region with an orientation. It is described by four parameters: key point center (x and y coordinates), its scale (the radius of the region) and its orientation (an angle expressed in radians). SIFT detector is invariant and robust to translation, rotations, scaling and partially invariant to affine distortion and illumination changes.

Four steps involved in SIFT algorithm.

Scale-space extrema detection: Which identify those locations and scales that are identifiable from different views (Gaussian blurring and sigma) of the same object.

Keypoint localization: Eliminate more points from the list of keypoints by finding those that have low contrast or are poorly localized on an edge.

Orientation assignment: Assign a consistent orientation to the keypoints based on local image properties.

Keypoint descriptor: Keypoint descriptors typically uses a set of 16 histograms, aligned in a 4×4 grid, each with 8 orientation bins, one for each of the main compass directions and one for each of the mid-points of these directions. This result come up in a feature vector containing 128 elements.

In other words, each pixel in an image is compared with its 8 neighbors as well as 9 pixels in next scale and 9 pixels in previous scales. If that pixel is a local extrema, it means that the keypoint is best represented in that scale.

Figure 1 shows 2 examples of SIFT features of Harris-Laplace key points which are generated by our experiment.

Visual word quantization: After extracting features, images can be represented by sets of keypoint descriptors. But they are not meaningful. For fixing this problem Vector Quantization techniques (VQ) are presented to cluster the keypoint descriptors into a large number of clusters by using the K means clustering algorithm and then convert each keypoint by the index of the cluster to which it belongs. By using Bag of Feature (BoF) method we can cluster similar features to visual words and represent each picture by counting each visual word. This representation is similar to the bag-of-words document representation in terms of semantics. There is a complete definition of BoW in the next part.

Bag of Words (BoW) model: Bag of Words (BoW) model is a popular technique for document classification. In this method a document is represented as the bag of its words and features are extracted from frequency of occurrence of each word. Recently, the Bag of Words model has also been used for computer vision (Perona, 2005). Therefore instead of document version name (BoW) Bag of Feature (BoF) will be used which is described below.

Bag of Feature (BoF) model: These days, Bag of Feature (BoF) model is widely used for image classification and object recognition because of its excellent performances.

Steps of BoF method are listed as follows:

- Extract Blobs and features (e.g., SIFT) on training and test Blobs of images
- Build visual vocabulary using a classification method (e.g., K-mean) and descriptor quantization
- Represent images with BoF histograms
- Image classification (e.g., SVM)

The related works in this area by Choi *et al.* (2010) presented a method for creating fuzzy multimedia ontologies automatically. They used SIFT feature extraction for their feature extraction and BoF for their feature quantization. Zhang *et al.* (2012) analyzed key aspects of the various Automatic Image Annotation (AIA) method, including both feature extraction and semantic learning methods. Also major methods are discussed and illustrated in details. Tousch *et al.* (2012) re-viewed structures in the field of demonstration and analyzed how the structure is used. They first demonstrated works without structured vocabulary and then showed how structured vocabulary started with introducing links between categories or between features. Then reviewed works which used structured vocabularies as an input and analyzed how the structure is exploited. Jiang *et al.* (2012) proposed Semantic Diffusion (SD) approach which enhanced the previous annotations (may be done manually or with machine

learning techniques) by using a graph diffusion formulation to improve the stability of concept annotation. Hong *et al.* (2014) proposed Multiple-Instance Learning (MIL) method by performing feature mapping MIL to change it to a single-instance learning problem for solving the problem of MIL method. This method is able to explore both the positive and negative concept correlations. It can also select the effective features from a large and diverse set of low-level features for each concept under MIL settings. Liu *et al.* (2014), presented a Multi-view Hessian Discriminative Sparse Coding (MHDSC) model which mixed Hessian

regularization and discriminative sparse coding to solve the problem of multi-view difficulties. Chiang (2013) offered a semi-automatic tool, called IGAnn (interactive Image Annotation), that assists users in annotating textual labels with images. By collecting related and unrelated images of iterations, a hierarchical classifier related to the specified label is built by using proposed semi-supervised approach. Dimitrovski *et al.* (2011) presented a Hierarchical Multi-label Classification (HMC) system for medical image annotation, where each case can be in multiple classes and these classes/labels are organized in a hierarchy.

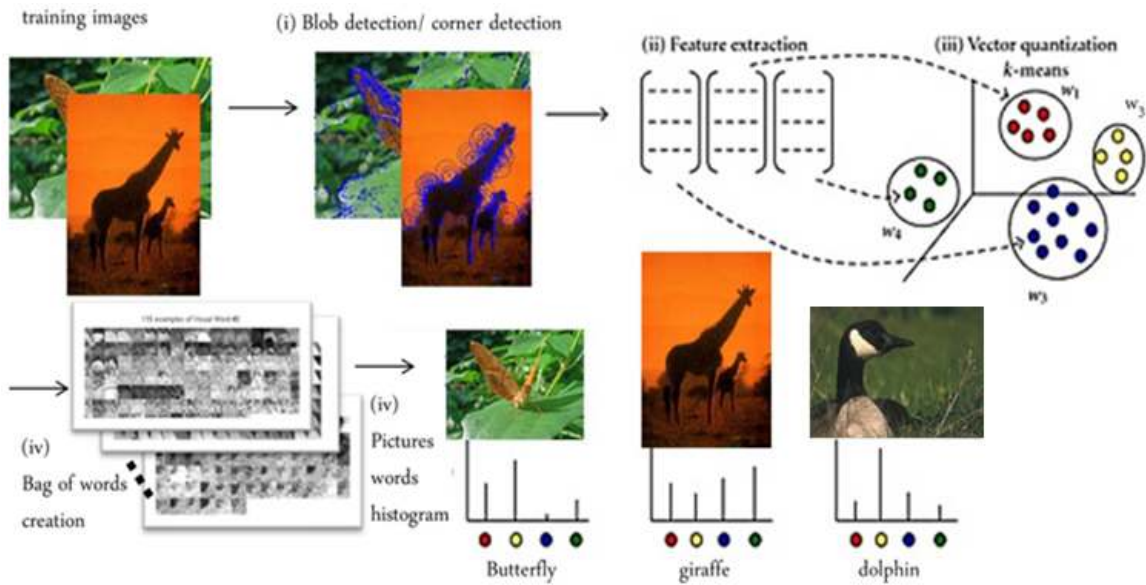


Fig. 2: Animal recognition using BoF model training stages

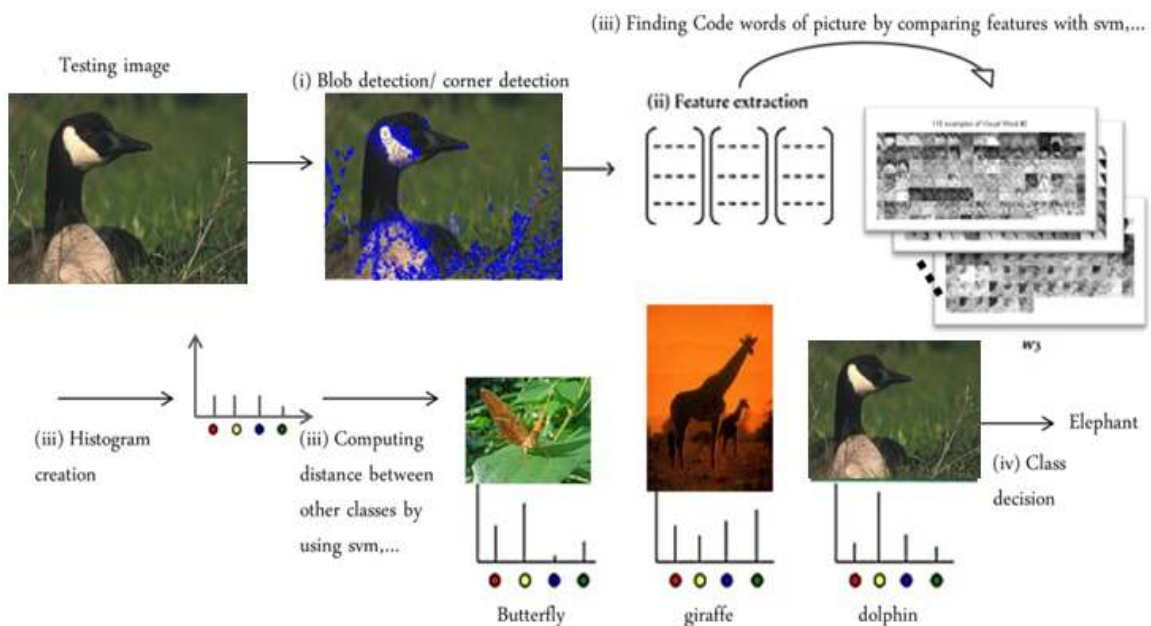


Fig. 3: Animal recognition using BoF model testing stages

In most of the reviewed literature, BoF with SIFT feature has the key role in feature extraction and quantization and shows better results in comparison with using other low level feature like color or texture alone (Tsai, 2012).

Figure 2 and 3 depict the stages of Animal recognition using BoF model for training and testing, respectively.

METHODOLOGY

In this study we will investigate the potential and accuracy of BoF model with SIFT feature, K-mean method for clustering and quantization of words and 6 different kinds of classification (NN L2, NN Chi², SVM linear, SVM LLC, SVM IK and SVM chi²) in a special domain (animal) to find which one is more effective.

Because of the variety of animal pictures and natural environment, our dataset is Caltech 256 (Griffin *et al.*, 2007). We investigate 20 different animals or 20 concepts from different kinds of animals (bear, butterfly, camel, dog, house fly, frog, giraffe, goose, gorilla, horse, humming bird, ibis, iguana, octopus, ostrich, owl, penguin, starfish, swan and zebra) in different environments (lake, desert, sea, sand, jungle, bushy, etc.). For each animal, 40 images are randomly selected for training and 10 images are randomly selected for testing. The total number of images is 800 for training and 200 for testing, The number of extracted code words is 1500 and for evaluating the accuracy of each concept we used a well-known formulas Precision, Recall and Accuracy (Tousch *et al.*, 2012; Chiang, 2013; Fakhari and Moghadam, 2013; Lee *et al.*, 2011).

Although, we have just focused on 20 different animals, this method can be used for other animals or

other categories rather than animals. All we need to do is to separate the folder of new concept and change its name. Then all the stages can be automatically done by our algorithm.

Essential equipments for running our program are MATLAB 2013a/ 2014a, Microsoft Windows SDK 7.1, C++ Compiler for running MATLAB and C++ files inside MATLAB.

Stages for recognizing animals by BoF are:

- Extract Blobs with Harris-Laplace key points
- Extract SIFT features on training and test Blobs of images
- Build visual vocabulary using k-means
Figure 4 shows a visual word example in Animal BOF dictionary out of 1500 visual words
- K-means descriptor quantization (quantize SIFT descriptors in all training and test images using the visual dictionary)
 - Compute Euclidean distances between visual dictionary and all descriptors
 - Compute visual word ID for each feature by minimizing the distance between feature SIFT descriptors and visual dictionary
- Visualize visual words (i.e., clusters)
To visually verify feature quantization computed above (or showing image patches corresponding to the same visual word), represent images with BOF histograms of visual word labels of its features. Compute word histogram over the whole image, normalize histograms
- Six kinds of image classification
- **Nearest Neighbor classification (NN L2) (Deza and Deza, 2009):** Nearest Neighbor classification (1-NN) using L2 distance



Fig. 4: Visual word example in animal BOF

- **Nearest Neighbor image classification using chi² distance (NN chi²) (Fakhari and Moghadam, 2013):** Nearest Neighbor classification with Chi² distance Compute and compare overall and per-class classification accuracies to the L2 classification above
Pre-computed linear kernels by SVM classification (using LIBSVM (Chang and Lin, 2013))
- Linear SVM (Corinna Cortes, 1995)
- LLC Linear SVM
Pre-computed non-linear kernel/intersection kernel
- SVM Intersection Kernel (IK) try a non-linear SVM with the histogram intersection kernel
- **SVM chi² pre-compute kernel:** Experiment with Chi² non-linear kernel. Chi² pre-computed in the past section b
- Compute classification accuracy based on precision, recall and accuracy

Figure 2 Illustrates training model of Bag of SIFT Feature in animal pictures which was implemented by MATLAB 2014. Then we tested Bag of SIFT Feature with test model which is shown in Fig. 3. All the pictures for both models are generated by our experiment.

Accuracy: For measuring the accuracy we used 2 famous methods: Precision, Recall and accuracy which are used in Tusch *et al.* (2012), Chiang (2013), Fakhari and Moghadam (2013) and Lee *et al.* (2011). Their formulas are in (1), (2) and (3) and also the definition of tp, tn, fp and fn are as follows.

True positives (tp): The number of items correctly labeled as belonging to this class.

False positives (fp): Items incorrectly labeled as belonging to this class.

False negatives (fn): Items which were not labeled as belonging to this class but should have been.

True negative (tn): The number of items correctly not labeled as belonging to this class:

$$Recall = \frac{t_p}{t_p + f_n} \quad (1)$$

$$Precision = \frac{t_p}{t_p + f_p} \quad (2)$$

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (3)$$

DISCUSSION

Normalized confusion matrix is a n×n matrix for showing how many test images are correctly classified

and how many are misclassified in other classes. Which means it can find in each concept how many of them are classified by the others. Therefore by using this matrix we can analyze and find the reason for the misclassification of some pictures and find a good solution for it. Figure 5 shows our final experimental results for 20 concepts (bear, butterfly, camel, dog, house fly, frog, giraffe, goose, gorilla, horse, hummingbird, ibis, iguana, octopus, ostrich, owl, penguin, starfish, swan and zebra), 40 images are randomly selected for training and 10 images are randomly selected for testing. It means the total number of images is 800 for training and 200 for testing. The number of extracted code words is 1500 and for computing the accuracy of each concept, we used well-known formulas Precision, Recall and Accuracy in six kinds of image classification methods (NN L2, NN Chi², SVM linear, SVM LLC, SVM ik, SVM chi²). All of them are respectively depicted in Fig. 6 to 8. Although we have just focused on 20 different animals, this method can be scalable to other concept. And all we need is to separate the folder of new concept and change its name to that new one. Then all the stages can be automatically done by our experiment.

Clearly, the results of SVM Chi-square are better than other ones which are shown in Fig. 5. Therefore

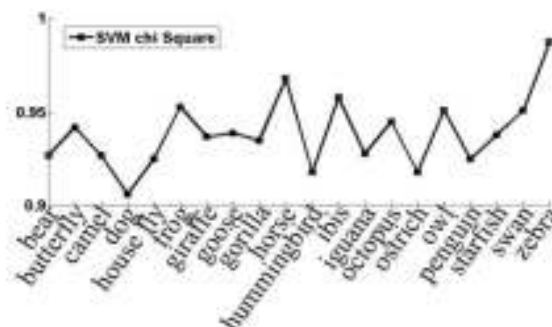


Fig. 5: Accuracy of SVM chi-square

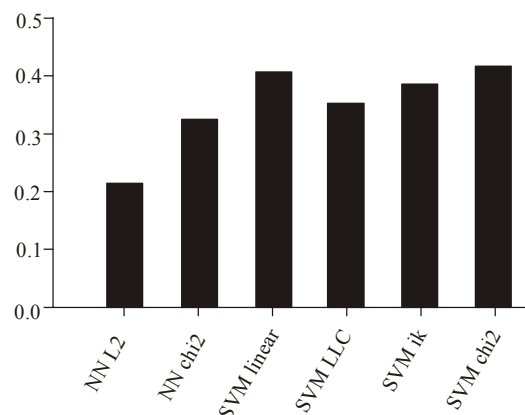


Fig. 6: Mean precision for each kind of classification

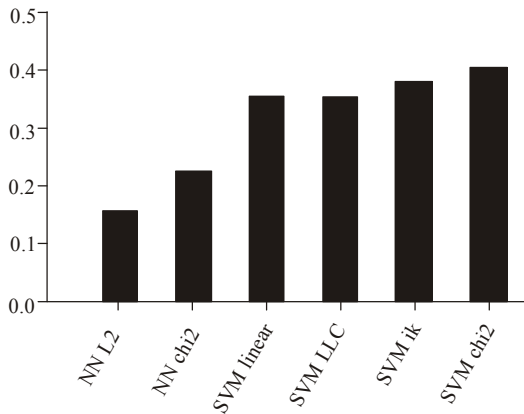


Fig. 7: Mean recall for each kind of classification

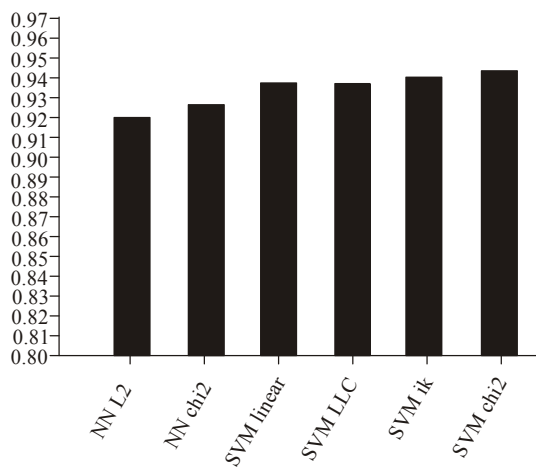


Fig. 8: Mean accuracy for each kind of classification

SVM Chi-square is a better classifier. The running of our code provides better results for three specific animals: zebra, horse and starfish. This is probably the result of a better distinguishing pattern in these animals. So if we can omit the unimportant parts of our dataset pictures, we will get more accurate results.

CONCLUSION

Our objective in this research was to find the potential usage of bag of feature in animal recognition and other concepts within recognition category. After implementation of our experiment we got reasonable results which show BoF is a good selection for finding animals in nature. Also, SVM Chi-square has a better accuracy in comparison with NN L2, NN Chi², SVM linear, SVM LLC, SVM IK. But most of the animals are the same as their environment because nature wants to protect them against enemies. In future if we omit the background parts we can definitely get better result. Therefore in future we want to extract regions for addressing the location of objects and extract other

features as well (Color, Texture, Shape and Spatial location etc.) to get better results.

REFERENCES

- Chang, C. and C. Lin, 2013. LIBSVM: A library for support vector machines. *ACM T. Intell. Syst. Technol.*, 2(3): 1-39.
- Chiang, C.C., 2013. Interactive tool for image annotation using a semi-supervised and hierarchical approach. *Comp. Stand. Inter.*, 35(1): 50-58.
- Choi, M.J., J.J. Lim, A. Torralba and A.S. Willsky, 2010. Exploiting hierarchical context on a large database of object categories. *Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp: 129-136.
- Corinna Cortes, V.V., 1995. Support-vector networks. *Mach. Learn.*, 20(3): 273-297.
- Deza, M.M. and E. Deza, 2009. *Encyclopedia of Distances*. Springer-Verlag, Berlin, Heidelberg, pp: 1-583.
- Dimitrovski, I., D. Kocev, S. Loskovska and S. Džeroski, 2011. Hierarchical annotation of medical images. *Pattern Recogn.*, 44(10-11): 2436-2449.
- Fakhari, A. and A.M.E. Moghadam, 2013. Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. *Appl. Soft Comput.*, 13(2): 1292-1302.
- Griffin, G., A. Holub and P. Perona, 2007. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology.
- Hong, R., M. Wang, Y. Gao, D. Tao, X. Li and X. Wu, 2014. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *IEEE T. Cybern.*, 44(5): 669-680.
- Jiang, Y.G., Q. Dai, J. Wang, C.W. Ngo, X.Y. Xue and S.F. Chang, 2012. Fast semantic diffusion for large-scale context-based image and video annotation. *IEEE T. Image Process.*, 21(6): 3080-3091.
- Lee, C.H., H.C. Yang and S.H. Wang, 2011. An image annotation approach using location references to enhance geographic knowledge discovery. *Expert Syst. Appl.*, 38(11): 13792-13802.
- Liu, Y., D. Zhang, G. Lu and W.Y. Ma, 2007. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40(1): 262-282.
- Liu, W., D. Tao, J. Cheng and Y. Tang, 2014. Multiview Hessian discriminative sparse coding for image annotation. *Comput. Vis. Image Und.*, 118: 50-60.
- Lowe, D.G., 1999. Object recognition from local scale-invariant features. *Proceeding of the 7th IEEE International Conference on Computer Vision*, 2: 1150-1157.

- Mikolajczyk, K., B. Leibe and B. Schiele, 2005. Local features for object class recognition. Proceeding of the 10th IEEE International Conference on Computer Vision. Washington, DC, USA, pp: 1792-1799.
- Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories. Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2: 524-531.
- Qi, H. and W.E. Snyder, 1999. Content-based image retrieval in picture archiving and communications systems. *J. Digit. Imaging*, 12(2): 81-83.
- Tousch, A.M., S. Herbin and J.Y. Audibert, 2012. Semantic hierarchies for image annotation: A survey. *Pattern Recogn.*, 45(1): 333-345.
- Tsai, C.F., 2012. Bag-of-words representation in image annotation: A review. *ISRN Artif. Intell.*, 2012: 1-19.
- Zhang, D., M.M. Islam and G. Lu, 2012. A review on automatic image annotation techniques. *Pattern Recogn.*, 45(1): 346-362.
- Zhou, X.S. and T.S. Huang, 2000. CBIR: From low-level features to high-level semantics. Proceeding of the SPIE, Image and Video Communication and Processing. San Jose, CA, 3974: 426-431.