CRIMSON PUBLISHERS
Wings to the Research

Open Access
Biostatistics & Bioinformatics

ISSN 2578-0247

Mini Review

A New Strategy of the Bag-of-Word Method with a Multi-scale Representation for Evaluating Color Pairwise Descriptors

Fatma Abdedayem[1], Ahmed Kaffel[2]* and Ahmed Ben Hmida[1]

[1]*Department of Technologies for Medecine and Signals, National School of Engineers of Sfax, Tunisia*

[2]*Department of Mathematics, Statistics & Computer Science, Marquette University, USA*

***Corresponding author:** Ahmed Kaffel, Mathematics, Statistics & Computer Science Department, Marquette University, USA.

Submission: March 05, 2018; **Published:** April 25, 2018

Abstract

In the field of image categorization, the Bag-Of-Word has proved to be successful. It treats local image features as visual words. After collecting all local features, each image is represented by a histogram of occurrences of visual words. In this work, we propose an extension to the Bag-Of-Words (BOW) by integrating the spatial relationships information between local features. In a first step, we extract local features by using both multi-scale representation and color descriptors based on HSV-SIFT, opponent-SIFT, RGB-SIFT, rg-SIFT and transformed-color-SIFT. In a second step, and in order to represent the relationships between local features, we form pairwise color descriptors by joining pairs of spatially neighbor SIFT color features. In a third step, we encode the histograms which involve the occurrence of pairwise color descriptors by applying the BOW strategy and the Spatial Pyramid Representation (SPR). Finally, image classification is carried out by using Support Vector Machine (SVM) on the generated histograms. Our proposed method is tested and validated using the standard image datasets "Pascal Voc 2007".

Keywords: Bag-of-Word (BOW); Color local SIFT descriptor SIFT descriptor; Support Vector Machine; Spatial Pyramid Representation; Pairwise descriptors

Background

The era of information age has led to an exponential growth of the available image databases. Consequently, a huge number of image processing methods have been proposed. One of the most important issues related to image processing is automatic indexation which mainly needs the extraction of low-level features from the image. These features are obtained by performing some low-level processing on the pixels to generate values that are more indicative to the image appearance and structure (shape, texture, color [1], etc...).

Few years ago, Bag of word (BOW) method [2-8] was proposed by Sivic. It was considered as a simple and an efficient method of image categorization. The BOW approach represents an image with a histogram of visual features generally quantized with the K-Means algorithm [9,10]. This produces a quantified feature vector with a reduced size in comparison with the concatenation of all extracted features which accelerate the image classification. Many methods have used only the classical SIFT [11,12] to represent an image. The major disadvantage of the Bag of words (BOW) approach is ignoring the spatial relationship between local features [13,14] in images. However, it has been argued before by several researchers those different parts of an object do not exist independently from each other. Geometric relationships of these parts are important information to be included in image representation.

In order to overcome the problem due to the absence of spatial relationships between features caused by the quantization of features with the BOW approach, a Spatial Pyramid Representation (SPR) [15,16] has been proposed as a solution. This method consists in extracting the local features [17] and partitioning the image into different regions at different spatial levels. Despite that the SPR produces an accurate image representation by providing the spatial information obtained by concatenating a series of spatial regions in different levels, it provides only the information about the global layout which is not enough to get accurate image description.

One of the successful proposed methods to take into account of the local spatial relationships between feature vectors was Quadratic Pairwise Codebook (QPC) [18]. It operates by joining each pair of visual words. The inconvenience of this approach is the exponential increase in the number of pairs involved. Later, Morioka & Satoh [5] have proposed a new approach based on the Local Pairwise Codebook (LPC) [6] which uses a compact codebook

consisting of cluster centers of spatially close SIFT descriptors pairs [7,9]. LPC approach has given better performance in comparison with both the classical BOW [2] and QPC approaches. However, in one hand this presented method doesn't take into account of the color feature information [5,19]. On the other hand, it doesn't ensure the image scale invariance. The goal of this study is using a new approach based on the following:

A. The color feature description is based on different color descriptors like HSV-SIFT, opponent-SIFT, RGB-SIFT, rg-SIFT and transformed-color-SIFT.

B. The spatial relations between local features are taken into account by joining local pairwise and using Spatial Pyramid Representation (SPR) to capture successively the local spatial information and the global spatial information.

C. The feature extraction is based on a multi-scale representation in order to consider the scale invariance.

Bag-of-Word

Among recently developed image classification methods, Bag-of-visual-Words (BOW) [20,21] methods have gained a significant interest. The BOW approach, called also Bag-of-Features (BoF), is a strategy inspired from the text retrieval community where each document is represented with its word frequency. It has been applied to a variety of applications such as web applications, smart phone applications, surveillance, robotics and others different multimedia applications. Bag-Of-Words (BOW) [3] was considered as a successful way in image clustering and retrieval settings. In comparison with method based on local features matching, it is considered as a simple method since it guaranties fast run time and less storage requirements. In contrast, there are some parameters that affect the performance of the model such as dictionary and histograms generation methods, dictionary size, normalization, distance functions, visual words generation methods and clustering methods.

Typically, BOW needs to apply three main steps: image representation, visual words dictionary generation and construction of the histogram of visual words. The first step consists of characterizing the image by a set of local features [22]. In the second step, in order to create a dictionary of visual words, currently, most visual codebooks in state-of-the-art are built using some forms of K-Means [10,22] clustering. In fact, K-Means clustering takes a large number of local features in n-dimensional space and divides them into a smaller number of clusters minimizing the distance between the local feature and the assigned cluster center. The center of each cluster will be taken as a visual word to create next the codebook. Finally, the third step in this method consists on computing histogram of visual words for each image.

Approach

In this section we present the basic strategy of our proposed method as described in the Figure 1. First, we extract multi-scale color descriptors [1,5,8] with different patch size. Second, we joint each pair of color SIFT descriptors. Third, we form a codebook of visual words by applying K-Means clustering on these pairs. Then, by using Spatial Pyramid Representation (SPR) each image is represented by a concatenation of histograms of visual words over three various pyramid levels. Finally, we apply SVM [4] learning algorithm to generate the model in the learning phase and to predict the image category in the test phase.

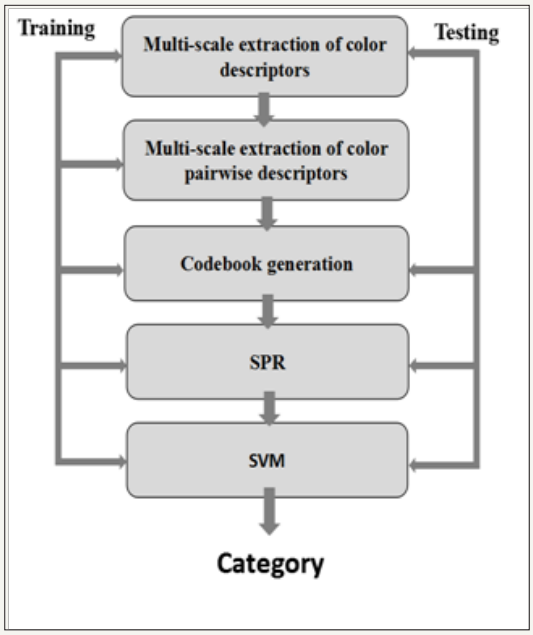

Figure 1: Implementation strategy of our proposed approach.

Multi-scale extraction of color descriptors

The idea of this step is to first detect the dense key points in the image in order to identify their descriptions using particular color SIFT descriptors [8]. For that, the image is sampled uniformly by using different scale values (4*4, 8*8, 16*16) at fixed locations forming a grid of rectangular windows. At this stage, we can identify the size and the step between the grid patches for each image. The selected values of the step are 2,4 and 8 and represent the double of the patch size. As a result, each image is divided in patches represented by a centralized point called key point. In the implementation, each feature point fi in the image is encoded as (xi,yi,di) where xi and yi denote the feature location and di is the feature color descriptor vector. Some popular color descriptors are chosen from other previous work results. The employed color descriptors are: RGB-SIFT [23], HSV-SIFT [24], Opponent-SIFT [25,26], rg-SIFT[27] and transformed-Color-SIFT [28].

Multi-scale extraction of color pairwise descriptors

After densely sampling the image features and describing them by color SIFT descriptors, we represent for each scale (4*4, 8*8, 16*16) each pair of spatially close descriptors di and dj as a joint descriptor [di,dj] by simple vector concatenation. Note that the

concatenation is achieved over the horizontal, vertical, and diagonal directions [6,7]. We extract at this stage color pairwise descriptors. Then, all possible color local pairwise descriptors over the scales are collected in a set of pairwise descriptors like presented in the following Figure 2.

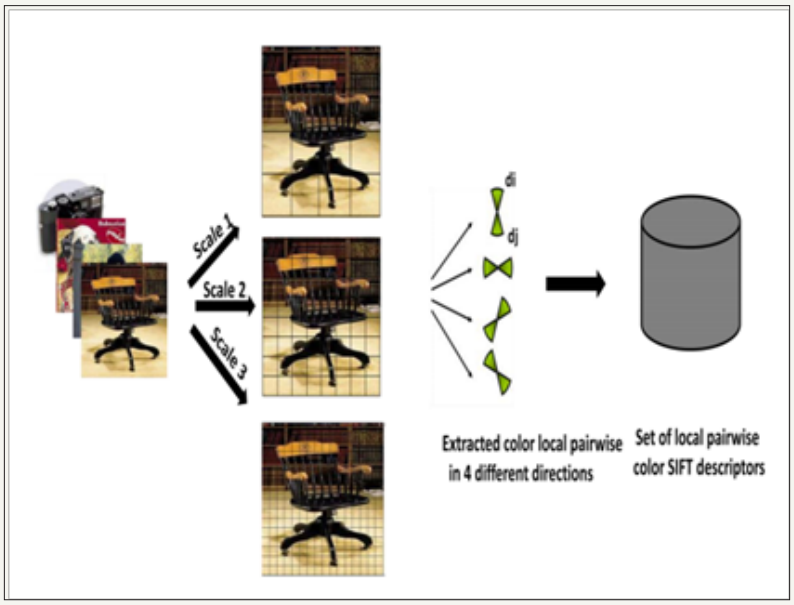

Figure 2: Collection of all possible pairwise color descriptors over 3 scales.

Code book generation

In our approach, we extract pairwise color SIFT descriptors in multi-scale (4*4, 8*8, 16*16). Then we collect them and we construct the dictionary codebook of visual words. In this step, we can generate k visual words that represent all collected color pairwise SIFTS descriptors. Then we apply the unsupervised K-Means algorithm on the set of local pairwise color SIFT descriptors (Figure 3).

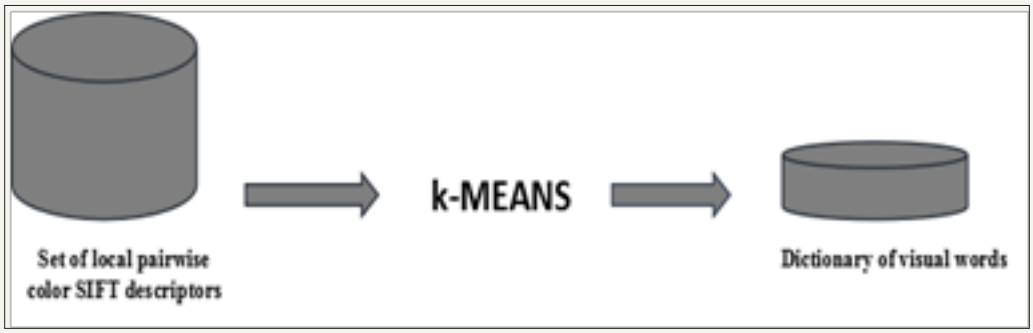

Figure 3: Dictionary generation from multi-scale.

Below we present the steps of the K-MEANS algorithm applied on the pairwise color descriptors:

Parameters: The number k of clusters.

Inputs: A set of M pairwise color descriptors $x_1,...x_M$ selected from the training images

A. Select k initial centers $c_1,...c_k$

B. For each pairwise descriptors, assign it to the cluster i corresponding to the nearest center

C. If all descriptors remain in the same cluster then exit.

D. Compute the new Centers: for each i, the center c_i is the mean of the descriptors belonging to the cluster i.

E. Go to (2).

Outputs: The visual word (cluster) assigned to each pairwise color descriptor, the center of each cluster.

Each obtained cluster corresponds to a visual word Vi. These visual words are collected to form the codebook that will be used to index the image. For a given image, the generated system extracts its pairwise color descriptors and assigns each descriptor to nearest K-Means cluster regarding the Euclidian distance. Thus,

How to cite this article: Fatma A, Ahmed B H, Ahmed K. A New Strategy of the Bag-of-Word Method with a Multi-scale Representation for Evaluating Color Pairwise Descriptors. Open Acc Biostat Bioinform. 1(4). OABB.000516. 2018. DOI: 10.31031/OABB.2018.01.000516

each local pairwise descriptor is associated to the visual word that corresponds to the assigned K-means cluster. Finally, we describe the image by a histogram of visual words where each bin counts the occurrences of each visual word (generated k-mean cluster) in the image.

Application of the Spatial Pyramid Representation (SPR) [29]

The Bag-Of-Words (BOW) model does not take into account the spatial information in the image. One of the most popular methods proposed to add global spatial constraints in the image description is Spatial Pyramid Representation (SPR). SPR starts by representing the image into three level of resolutions (0,1 and 2). For each level, the image is divided into a number of cells. Next, the histograms of visual words are calculated for each cell by computing the frequency of each visual word present in the cell. Afterwards, at each level, the cell histograms are concatenated horizontally to form the histogram related to this level. Then each histogram is Li normalized by dividing by the sum of the histogram components. This is achieved in order to ensure the scale invariance. After forming three different normalized histograms that correspond to all the three levels previously defined, we concatenate them horizontally in order to form the final image histogram.

Image classification

Generally, the goal of supervised learning is to predict the value of an outcome measure based on a set of input measure. Support Vector Machine (SVM) is one of the efficient image classification algorithm. The two principal steps in image classification [16,30] are the training and the testing phases.

The training phase (classification phase)

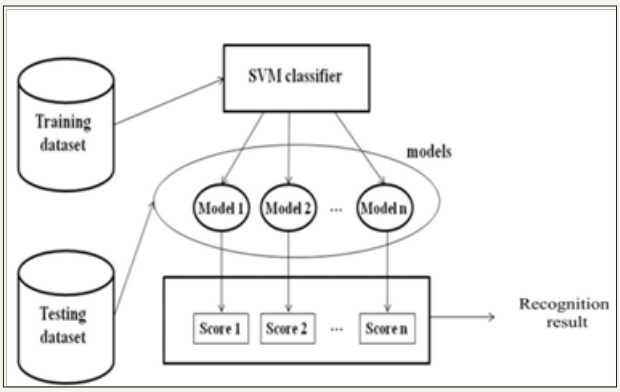

Figure 4: Image classification/image recognition.

In the training phase a supervised learning algorithm is applied on the computed pairwise histograms are presented to the multi-class SVM [4] in order to learn a model for each category (see figure 4). In our approach, the commonly used SVM algorithm with an RBF-$\chi 2$ kernel is used since it has given promising results in the image categorization systems. Using $\chi 2$ the SVM algorithm, we compute the $\chi 2$ distance for each pair of training color pairwise descriptors and obtain the kernel matrix. We normalize this kernel matrix using A which is the mean pairwise distance between the training samples. We then map this kernel matrix using exponential function exp^{-x}:

$$K(h1,h2) = \exp\left(-\frac{1}{A}\sum_{i=1}^{N}\frac{(h1(i)-h2(i))^2}{h1(i)+h2(i)}\right)$$

Where h1 and h2 are a pair of training descriptors (the frequency histograms of word occurrences) of N-dimensions. We next feed this transformed kernel matrix to an SVM with a Radial Basis Kernel.

The testing phase (recognition phase)

In the testing phase, the models built in the training phase are used to recognize the images from the tested set. Each generated model estimates the posterior class probability (score), then the test image is assigned to the category (class) associated to the high posterior probability value. In this section, we present some experimental results on widely used dataset in image research named Pascal VOC 2007 [31]. We present the impact of different color descriptors applied in our work and we show how the use of multi-scale in feature extraction can improve the final result of classifier. For the datasets, SIFT descriptors are densely sampled at every 2,4 and 8 pixel step between patches. The values taken for the size of the patches are successively 4,8 and 16. The number of vocabulary of all generated dictionaries used in this experiment was for every color descriptor 3200 visual words. In addition, we used 5011 training images and 4952 test images for the PASCAL VOC 2007.

In this section, we present some details of the experimental setup used to evaluate the impact of the used color descriptors. We tested our approach for different color spaces. In order to extract local features, we have applied classic SIFT [32] and SIFT based on color descriptors which are HSV-SIFT, Opponent-SIFT, RGB-SIFT, rg-SIFT and transformed-color-SIFT. All results are represented in Figure 5 for PASCAL VOC 2007.

After using the proposed color extensions, we can show the following results:

How to cite this article: Fatma A, Ahmed B H, Ahmed K. A New Strategy of the Bag-of-Word Method with a Multi-scale Representation for Evaluating Color Pairwise Descriptors. Open Acc Biostat Bioinform. 1(4). OABB.000516. 2018. DOI: 10.31031/OABB.2018.01.000516

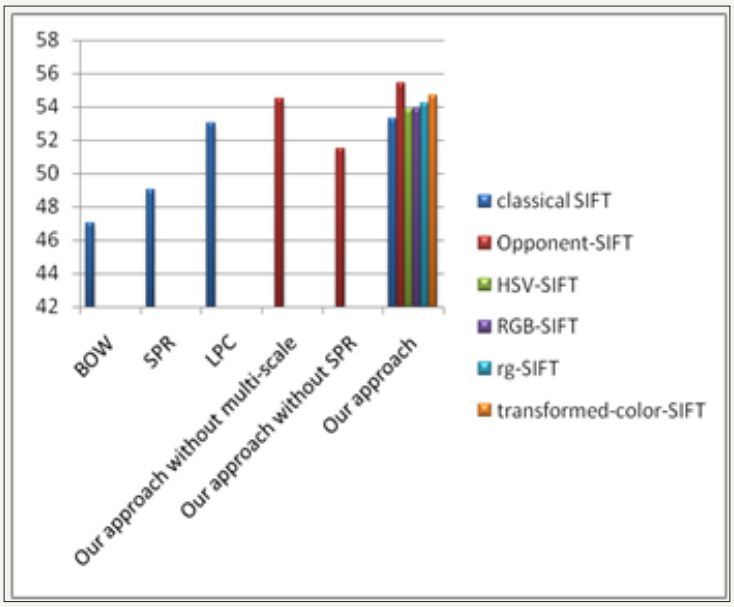

Figure 5: Comparison results of Mean Average Precision.

a. Color descriptors (HSV-SIFT, RGB-SIFT, rg-SIFT, transformed-color SIFT, Opponent-SIFT) get the (Mean Average Precision values) or MAP values of 53% 53,8%, 54,7%, 53,9% 54,2 % and 55,4% respectively on PASCAL VOC 2007. These color descriptors outperform (that applied in classic BOW method). This indicates that they truly benefit from the additional color information.

b. Compared to the final results of these color descriptors, opponent-SIFT is slightly better than the others. It outperforms SIFT descriptor in PASCAL VOC 2007 datasets.

c. By taking opponent-SIFT as an example of color descriptors, we make comparison (without SPR/with SPR) between the two results. We note an improvement is guaranteed in the final result of the image categorization which is about 4%. This demonstrates that combining local spatial information (obtained by the color pairwise descriptor) with global information (obtained by SPR) can considerably improve the classification performance.

d. We have compared opponent-SIFT without multi-scale representation and opponent-SIFT with multi-scale representation for PASCAL VOC 2007 dataset. We showed the effect of using multi-scales with different size patches (4*4, 8*8, 16*16). It's clear that there is an improvement of result values of mean average precision (MAP) for PASCAL VOC 2007.

e. We found that image representation based on Bag-Of-Word (BOW) [33,34] can be improved by adding spatial pyramid representation (SPR) [24] to classical BOW method. The results also showed an improvement of MAP value by applying LPC [6] method. Indeed, N. Morioka has demonstrated that using local pairwise of features with spatial pyramid representation captures more local and global information about relationship between features and improves the final result of image categorization. Then, our extensions which add both color SIFT descriptors and multi-scale representation to local pairwise codebook in step of feature extraction guarantees also more improvement to the image categorization process.

Conclusion

Image understanding methods allow photos and videos to be found and used by a user-oriented and semantically meaningful way based on their visual content. The Bag of Words (BOW) model was employed, but needs improvements. Recent studies suggested incorporating the local spatial constraints in the BOW model to improve the final classification results, which is one of the objectives in our proposed approach.

We first integrated color information in the SIFT descriptors by using different color space to get better performance. We also followed a multi-scale representation in order to tackle the problem of scale invariance. Then, we extended the BOW approach by building pairwise multi-scale descriptors and joining spatially nearby features. In order to improve the final result, we also created an image by using the Spatial Pyramid Representation (SPR) approach, which, in the recent years, has been proved important and useful in the field of image recognition. All Our experimental results show the benefits of our extensions on Pascal Voc 2007.

References

1. Sande KEA, Gevers T, Snoek CGM (2008) Evaluation of color descriptors for object and scene recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 32(9): 1582-1596.

2. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of key points. In workshop on Statistical Learning in Computer Vision, ECCV, pp. 1-22.

3. Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-

features image classification. In Proceedings of European Conference on Computer Vision, pp. 490-503.

4. Morioka N, Satoh S (2010) Building compact local pairwise codebook with joint feature space clustering. ECCV Vol. 6311.
5. Morioka N, Satoh S (2010) Learning directional local pairwise bases with sparse coding. BMVC.
6. Curka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. Workshop on Statistical learning in compute Vis ECCV 1: 1-22.
7. Tsai CF (2012) Bag-of-words representation in image annotation: A review. ISRN Artificial Intelligence.
8. Singh D, Roy D, Mohan CK (2017) DiP-SVM: Distribution preserving kernel support vector machine for big data. In IEEE Transactions on Big Data 3(1): 79-90.
9. Chang RI, Lin SY, Ho JM, Fann CW, Wang, YC (2012) A novel content based image retrieval system using k-means/knn with feature extraction. Computer Science and Information Systems 9(4): 1645-1661.
10. Hartigan JA, Wong MA (1979) Algorithm AS 136: A k-means clustering algorithm. Applied Statistics, pp. 100-108.
11. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int Journal of Compute Vis 60: 91-110.
12. Zhu C, Bichot C, Chen L (2011) Visual object recognition using daisy descriptor. IEEE International Conference on Multimedia and Expo.
13. Li J, Allinson NM (2008) A comprehensive review of current local features for computer vision. Neurocomputing 71(10-12): 1771-1787.
14. Passalis N, Tefas A (2016) Entropy optimized feature-based bag-of-words representation for information retrieval. IEEE Transactions on Knowledge & Data Engineering 28(7): 1664-1677.
15. Huang X, Xu Y, Yang L (2016) Compact spatial pyramid representation: The viewpoint of codeword. Proceedings of IEEE 2016 International Conference on Optoelectronics and Image Processing (ICOIP), pp. 40-43.
16. Krishnan P, Dutta K, Jawahar CV (2016) Deep feature embedding for accurate recognition and retrieval of handwritten text. IEEE 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 289-294.
17. Mikolajczyk K, Leibe B, Schiele B (2005) Local features for the object class recognition. In Computer Vision ICCV 2005. Tenth IEEE International Conference 2: 1792-1799.
18. Herve N, Boujemaa N (2009) Visual word pairs for automatic image annotation. In Proceedings of the 2009 IEEE international conference on Multimedia and Expo ICME 09.
19. Sande KEA, Gevers T, Snoek CGM (2008) A comparison of color features for visual concept classification. CIVR.
20. Sande KEA, Gevers T, Snoek CGM (2008) Evaluation of colors descriptors for object and scene recognition. IEEE conf computer vision and pattern recognition CVPR.
21. Wang JZ, Li J, Wiederhold G (2000) Simplicity: Semantics sensitive integrated matching for picture libraries. In VISUAL '00: Proceedings of the 4th International Conference on Advances in Visual Information Systems, London, UK, pp. 360-371.
22. Everingham M, VanGool L, Williams CKI, Winn J, Zisserman A (2007) The PASCAL visual object classes challenge 2007 results.
23. Sergyán S (2007) Color content-based image classification. 5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics, pp. 25-26.
24. Geusebroek JM, Boomgaard R, Smeulders AWM, Geerts H (2001) Color invariance. IEEE Trans Pattern Analysis and Machine Intelligence 23(12): 1338-1350.
25. Tuytelaars T, Mikolajczyk K (2008) Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3): 177-280.
26. Bosch A, Zisserman A, Muñoz X (2008) Scene classification using a hybrid generative/discriminative approach. IEEE Trans Pattern Analysis and Machine Intelligence 30(4): 712-727.
27. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Proceedings of the IEEE Computer society Conference on Computer Vision and Pattern Recognition, pp. 2169-2178.
28. Firuzi K, Vakilian M, Darabad VP, Phung BT, Blackburn TR (2017) A novel method for differentiating and clustering multiple partial discharge sources using S transform and bag of words feature. IEEE Transactions on Dielectrics and Electrical Insulation 24(6): 3694-3702.
29. Liu ZJ (2015) Image classification method based on visual saliency and bag of words model. IEEE 8th International Conference on Intelligent Computation Technology and Automation (ICICTA), pp. 466-469.
30. Xiaojing W, Yuanbo L (2018) Research on improved K-Means algorithm based on hadoop. 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha, Hunan, China, pp. 593-598.
31. Liu L, Shen Hengel A (2017) Cross-convolutional-layer pooling for image recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(11): 2305-2313.
32. Zhuang Y, Hanqi W, Jun X, Fei W, Yang Y, et al. (2017) Bag-of-discriminative-words (BoDW) representation via topic modelling. IEEE Transactions on Knowledge & Data Engineering 29(5): 977-990.
33. Feng C, Wang X (2016) Image retrieval system based on bag of view words model. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, Japan, pp. 1-4.
34. Gauen K, Dailey R, Laiman J, Zi Y, Asokan N, et al. (2017) Comparison of visual datasets for machine learning. IEEE International Conference on Information Reuse and Integration (IRI), USA, pp. 346-355.