IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

A Review on Image Classification Using Bag of Features Approach

Santosh Kumar Panda[1*], Chandra Sekhar Panda[2]

[1,2] Department of Computer Science, Sambalpur University, Jyotivihar, Burla, Odisha, India

Corresponding Author: sonupanda1995@gmail.com, Tel.: +918144101447

DOI: https://doi.org/10.26438/ijcse/v7i6.538542 | Available online at: www.ijcseonline.org

Accepted: 09/Jun/2019, Published: 30/Jun/2019

***Abstract*—**Bag of Features or BoF approach has been used in many computer vision tasks, including image classification, video search, robot localization, and texture recognition. It is so widely popular because of its simplicity. These methods are based on unordered collections of image descriptors which are then quantized and are discarded spatial information, therefore conceptually and computationally simpler than many alternative methods, because of this; BoF based systems have set new performance standards on popular image classification benchmarks and have achieved scalability breakthroughs in image retrieval. This paper reviews related works based on the issues of improving and/or applying BoF. Emphasis is placed on recent techniques that mitigate quantization errors, improve feature detection, and speed up image retrieval. Meanwhile, unresolved issues and fundamental challenges are also raised. Among those issues the best techniques for sampling images, describing local image features, and evaluating system performance. Among those the fundamental challenges are how the BoF methods can contribute in localizing the objects in more complex images, or associating high-level semantics with natural images. Moreover, many recent works are compared in terms of the methodology of BoF feature generation and experimental design. Different Classification Models are also discussed.

***Keywords*—** Bag Of Features, Feature Extraction, Quantization, Clustering, Image Representation, Image Classification, Support Vector Machine

I. INTRODUCTION

Large collections of images are becoming available to the public, from photo collections to Web pages or even video databases. To index or retrieve them is a challenge which is the focus of many research projects (like IBM's QBIC) [1]. Classification refers to the process of classifying images into several categories, based on their similarities. Earlier systems were developed to look out databases for images on the basis of color, texture and some other information. Classification system consists of database that contains predefined patterns and features that compares with the detected object to classify it into accurate category. Image classification is an important and challenging duty in a variety of application domains, including biomedical imaging, biometry, video surveillance, vehicle navigation, industrial visual inspection, robot navigation, and remote sensing.

Bag of Features Approach:-

The bag-of-Features (BoF) methodology was first proposed in the text retrieval domain problem for text document analysis, and in texture recognition and then it was further adapted for computer vision applications. It is also called as Bag of Words (BOW). BoF approaches are characterized by the use of an order less collection of image features.

For image analysis, a visual analogue of a word is used in the BoF model, which is based on the vector quantization process by clustering low-level visual features of local regions or points, such as color, texture, and so forth. To extract the BoF feature from images involves the following steps: (i) automatically detect regions/points of interest, (ii) compute local descriptors over those regions/points, (iii) quantize the descriptors into words to form the visual vocabulary, and (iv) find the occurrences in the image of each specific word in the vocabulary for constructing the BoF feature (or a histogram of word frequencies)[3]

The BoF model can be defined as follows. Given a training dataset D containing n images represented by D = d1, d2, ..., and dn, where d is the extracted visual features, a specific unsupervised learning algorithm, such as k-means, is used to group D based on a fixed number of visual words W (or categories) represented by W = w1,w2, ..., and wv, where V is the cluster number. Then, we can summarize the data in a V ×N co-occurrence table of counts Nij = n(wi, dj), where n(wi, dj) denotes how often the word wi occurred in an image di.[4]

Steps of BOF Approach:-

1) Feature Extraction
   - Interest Point Detection
   - Local Descriptors
2) Visual Word Generation/ Quantization
3) Clustering
4) Classification

Feature Extraction:

Feature extraction is most important step in the procedure of the Classification. Features are classified into three types that is low, middle and high level. Low level features are color, texture and Middle level feature is shape and High level feature is semantic gap of objects [2]. Color is by far the most common visual feature used, primarily because of the simplicity of extracting color information from images. Texture and shape are also key component of human visual perception. The main features are described in detail below:

- **Interest Point Detection**

The first step of the BoF methodology is to detect local interest regions or points. For feature extraction of interest points (or key points), they are computed at predefined locations and scales. Several well-known region detectors that have been described in the literature are discussed below [5,6]

(i) Harris-Laplace regions are detected by the scale-adapted Harris function and selected in scale-space by the Laplacian-of-Gaussian operator. Harris-Laplace detects corner-like structures.

(ii) DoG regions are localized at local scale-space maxima of the difference-of-Gaussian. This detector is suitable for finding blob-like structures. In addition, the DoG point detector has previously been shown to perform well, and it is also faster and more compact (less feature points per image) than other detectors.

(iii) 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.

(iv) Salient regions are detected in scale-space at local maxima of the entropy. The entropy of pixel intensity histograms is measured for circular regions of various sizes at each image position.

(v) Maximally stable external regions (MSERs) are components of connected pixels in a threshold image.

- **Local Descriptors**

In most studies, some single local descriptors are extracted, in which the Scale Invariant Feature Transform (SIFT) descriptor is the most widely extracted [17]. It combines a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a 3D histogram of gradient locations and orientations. The dimensionality of the SIFT descriptor is 128.

Quantization

When the key points are detected and their features are extracted, suchas with the SIFT descriptor, the final step of extracting the BoF feature from images is based on vector quantization. In general, the k-means clustering algorithm is used for this task, and the number of visual words generated is based on the number of clusters (i.e., k).

Clustering

Clustering is a common method for learning a visual vocabulary or codebook – Unsupervised learning process – Each cluster center produced by k-means becomes a codevector – Codebook can be learned on separate training set – Provided the training set is sufficiently representative, the codebook will be "universal"

The codebook is used for quantizing features – A vector quantize takes a feature vector and maps it to the index of the nearest code vector in a codebook – Codebook = visual vocabulary – Code vector = visual word.

K-Means Clustering

It is used for classification of objects based on certain features into k number of classes or we can say it can be used for segmenting the image according to their features. One important advantage of K-means clustering is that , with larger data set it gives better efficiency as it minimizes the sum of squared Euclidean distance between points and their nearest cluster centers .

Algorithm:-

- Randomly initialize K cluster centers.
- Iterate until convergence:
  - Assign each data point to the nearest center
  - Recomputed each cluster center as the mean of all points assigned to it.

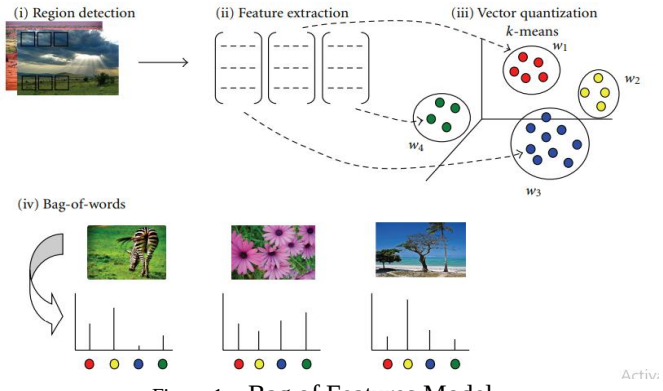

Figure 1. Bag of Features Model

II. RELATED WORK

In Mikolajczyk et al. [7], they compare six types of well-known detectors, which are detectors based on affine normalization around Harris and Hessian points, MSER, an edge-based region detector, a detector based on intensity extreme, and a detector of salient regions. They conclude that the Hessian-Affine detector performs best.

On the other hand, according to Horster and Lienhart ¨ [8], interest points can be detected by the sparse or dense approach. For sparse features, interest points are detected at local extremes in the difference of a Gaussian pyramid [9]. A position and scale are automatically assigned to each point and thus the extracted regions are invariant to these properties. For dense features, on the other hand, interest points are defined at evenly sampled grid points. Feature vectors are then computed based on three different neighborhood sizes, that is, at different scales, around each interest point.

Some authors believe that a very precise segmentation of an image is not required for the scene classification problem [10], and some studies have shown that coarse segmentation is very suitable for scene recognition. In particular, Bosch et al. [11] compare four dense descriptors with the widely used sparse descriptor (i.e., the Harris detector) [12, 13] and show that the best results are obtained with the dense descriptors. This is because there is more information on scene images, and intuitively a dense image description is necessary to capture uniform regions such as sky, calm water, or road surface in many natural scenes.

Similarly, Jurie and Triggs [15] show that the sampling of many patches on a regular dense grid (or a fixed number of patches) outperforms the use of interest points. In addition, Fei-Fei and Perona [16],and Bosch et al. [14] show that dense descriptors outperform the sparse ones.

In order to reduce the dimensionality of the SIFT descriptor, which is usually 128 dimensions per keypoint, principal component analysis (PCA) can be used for increasing image retrieval accuracy and faster matching [18]. Specifically, Uijlings et al. [19] show that retrieval performance can be increased by using PCA for the removal of redundancy in the dimensions.

SIFT was found to work best [20, 21]. Specifically, Mikolajczyk and Schmid [20] compared 10 different descriptors extracted by the Harris-Affine detector, which are SIFT, gradient location and orientation histograms (GLOH) (i.e., an extension of SIFT), shape context, PCASIFT, spin images, steerable filters, differential invariants, complex filters, moment invariants, and cross-correlation of sampled pixel values. They show that the SIFT-based descriptors perform best.

In addition, Quelhas et al. confirms in practice that DoG + SIFT constitutes a reasonable choice. Very few consider the extraction of different descriptors. For example, Li et al. [22] combine or fuse the SIFT descriptor and the concatenation of block and blob based HSV histogram and local binary patterns to generate the BoF.

Jiang et al. [23] conducted a comprehensive study on the representation choices of BoF, including vocabulary size, weighting scheme, such as binary, term frequency (TF) and term frequency-inverse document frequency (TF-IDF), stop word removal, feature selection, and so forth for video and image annotation.

T. de Campos, G. Csurka, and F. Perronnin in their paper "Images as sets of locally weighted features" used SHIFT as local descriptors and Logistic Regression as the classification model.

B. Fernando, E. Fromont, D. Muselet, and M. Sebban, in their paper "Supervised learning of Gaussian mixture models for visual vocabulary generation" used PCASIFT/SIFT/SURF as local descriptors and k-means as clustering algorithm and svm as classification model.

N. M. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzalez, in their paper "Discriminative compact pyramids for object and scene recognition" used SIFT/HSV color + SIFT as local descriptors and k-means as clustering algorithm and svm as classification model.

H. L. Luo, H. Wei, and L. L. Lai, in their paper "Creating efficient visual codebook ensembles for object categorization," used SIFT as local descriptors and k-means as clustering algorithm and svm as classification model.

K. T. Chen, K. H. Lin, Y. H. Kuo, Y. L. Wu, and W. H. Hsu, in their paper "Boosting image object retrieval and indexing by automaticallydiscoveredpseudo-objects," used SIFT as local descriptors and GMM-BIC as clustering algorithm.

J. S. Hare, S. Samangooei, and P. H. Lewis, in their paper " Efficient clustering and quantisation of SIFT features: exploiting characteristics of the SIFT descriptor and interest region detectors under image inversion," in Proceedings of the 1st ACM International Conference on Multimedia Retrieval , used SIFT as local descriptors and CPM and Adaptive Refinement as clustering algorithm and SVM as classification model.

J. Stottinger, A. Hanbury, N. Sebe, and T. Gevers, in their paper "Sparse color interest points for image retrieval and object categorization," on IEEE Transactions on Image Processing, used RGB Harris with Laplacian scale selection as local descriptors and k-means as clustering algorithm and SVM as classification model.

III. METHODOLOGY

Appropriate classification method will be used on the data. Some of the methods discussed are as follows:-

1) Artificial Neural Network(ANN):-

ANN is a type of artificial intelligence that imitates some functions of the person mind. ANN has a normal tendency for storing experiential knowledge. An ANN consists of a sequence of layers, each layer consists of a set of neurons. All neurons of every layer are linked by weighted connections to all neurons on the preceding and succeeding layers.

It uses Nonparametric approach. Performance and accuracy depends upon the network structure and number of inputs

2) Decision Tree:-

DT calculates class membership by repeatedly partitioning a dataset into uniform subsets Hierarchical classifier permits the acceptations and rejection of class labels at each intermediary stage. This method consists of 3 parts: Partitioning the nodes, find the terminal nodes and allocation of class label to terminal nodes

DT is based on hierarchical rule based method and use nonparametric approach.

3) Support Vector Machine:-

A support vector machine builds a hyper plane or set of hyper planes in a high- or infinite dimensional space, used for classification. Good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (functional margin), generally larger the margin lower the generalization error of the classifier.

SVM uses Nonparametric with binary classifier approach and can handle more input data very efficiently. Performance and accuracy depends upon the hyperplane selection and kernel parameter.

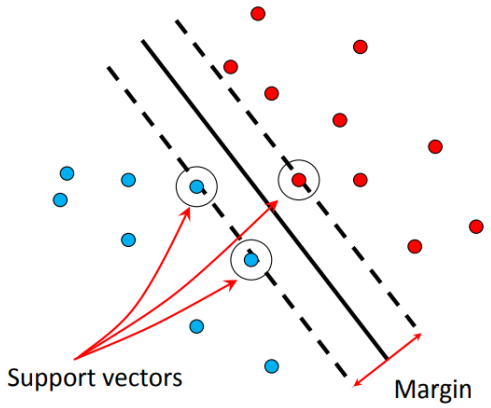

Figure 2. Support vector machine model

4) Fuzzy Measure

In Fuzzy classification, various stochastic associations are determined to describe characteristics of an image. The various types of stochastic are combined (set of properties) in which the members of this set of properties are fuzzy in nature. It provides the opportunity to describe different categories of stochastic characteristics in the similar form.

It uses stochastic approach. Performance and accuracy depends upon the threshold selection and fuzzy integral.

Comparison between different classification techniques:-

A brief comparison between all the classification models is done along with its some advantages and disadvantages in the following table:

Table 1. Comparison of different Classification methods

Classification Methods	Advantages	Disadvantages
Artificial Neural network	• It is a non-parametric classifier • It is a universal functional approximate with arbitrary accuracy. • capable to present functions such as OR, AND, NOT • It is a data driven self-adaptive technique • Efficiently handles noisy inputs.	• It is semantically poor. •The training of ANN is time taking. • Problem of over fitting. •Difficult in choosing the type network architecture.
Decision tree	• Can handle nonparametric training data • Does not require an extensive design and training. • Provides hierarchical associations between input variables to forecast class membership and provides a set of rules n are easy to interpret. • Simple and computational efficiency is good.	•The usage of hyper plane decision boundaries parallel to the feature axes may restrict their use in which classes are clearly distinguishable. •Becomes complex calculation when various values are undecided and/or when various outcomes are correlated.
Support Vector Machine	It gains flexibility in the choice of the form of the threshold. • Contains a nonlinear transformation. • It provides a good generalization capability. • The problem of over fitting is eliminated.	•Result transparency is low. •Training is time consuming. • Structure of algorithm is difficult to understand •Determination of optimal parameters is not easy when there

	• Reduction in computational complexity. • Simple to manage decision rule complexity and Error frequency.	is nonlinearly separable training data.
Fuzzy Measure	• Efficiently handles uncertainty. • properties are describe by identifying various stochastic relationships.	Without prior knowledge output is not good • precise solutions depends upon direction of decision.

IV. CONCLUSION

The Bag of Features representation is notable because of its relative simplicity and strong performance in a number of vision tasks. This Paper describes the detail steps included in Bag of Features approach for Image Classification also the different classification methods being used along with its advantage and disadvantages. Also we concluded that the mostly used local descriptors is SIFT algorithm. The mostly used clustering algorithm is K-means clustering algorithm and the mostly used classification method is SVM due to its simplicity and less time consumption than the neural network methods described above.

V. FUTURE WORK

According to the comparative results, there are some future research directions. First, the local feature descriptor for vector quantization usually by point-based SIFT feature can be compared with other descriptors, such as a region based feature or a combination of diff erent features. Second, a guideline for determining the number of visual words over what kind of datasets should be provided. The third issue is to assess the performance of generative and discriminative learning models over diff erent kinds of datasets, such as diff erent dataset sizes and diff erent image contents, for example, a single object per image and multiple objects per image. Finally, it is worth examining the scalability of BoF feature representation for large scale image annotation.

REFERENCES

[1] Olivier Chapelle, Patrick Haffner, and Vladimir N. Vapnik,"Support Vector Machines for Histogram-Based Image Classification",IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 5, SEPTEMBER 1999.

[2] T. Dharani, I. Laurence Aroquiaraj IEEE Member "Content Based Image Retrieval System using Feature Classification with Modified KNN Algorithm", arXiv preprint:130.4717,2013

[3] A. Bosch, X. Munoz, and R. Mart , "Which is the best way to organize/classify images by content?" Image and Vision Computing, vol. 25, no. 6, pp. 778–791, 2007.

[4] Chih-Fong Tsai, "Bag-of-Words Representation in Image Annotation: A Review", International Scholarly Research Network, Volume 2012, Article ID 376804, 19 pages doi:10.5402/2012/376804.

[5]K. Mikolajczyk, B. Leibe, and B. Schiele, "Local features for object class recognition," in Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05), pp. 1792–1799, October 2005.

[6] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," Foundations and Trends in Computer Graphics and Vision, vol. 3, no. 3, pp. 177–280, 2007.

[7] K. Mikolajczyk, T. Tuytelaars, C. Schmid et al., "A comparison of affine region detectors," International Journal of Computer Vision, vol. 65, no. 1-2, pp. 43–72, 2005.

[8] E. Horster and R. Lienhart, "Fusing local image descriptors ¨ for large-scale image retrieval," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07), pp. 1–8, June 2007.

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

[10] D. Gokalp and S. Aksoy, "Scene classification using bag-of- ¨ regions representations," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07), pp. 1–8, June 2007.

[11] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in European Conference on Computer Vision, pp. 517–530, 2006.

[12] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03), pp. 1470–1477, October 2003.

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05), pp. 370–377, October 2005.

[14] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in European Conference on Computer Vision, pp. 517–530, 2006.

[15] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05), pp. 604–610, October 2005.

[16] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in Proceedings of the 6th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05), pp. 524–531, June 2005.

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

[18] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04), pp. 506–513, July 2004.

[19] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "Real-time visual concept classification," IEEE Transactions on Multimedia, vol. 12, no. 7, pp. 665–681, 2010.

[20] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp. 1615–1630, 2005.

[21] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," International Journal of Computer Vision, vol. 73, no. 2, pp. 213–238, 2007.

[22] Z. Li, Z. Shi, X. Liu, Z. Li, and Z. Shi, "Fusing semantic aspects for image annotation and retrieval," Journal of Visual Communication and Image Representation, vol. 21, no. 8, pp. 798–805, 2010.

