# An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges

Wisam A. Qader
*Computer Engineering Dep.*
*Tishk International University*
Erbil, Iraq
wisam.softeng@gmail.com

Musa M.Ameen
*Computer Engineering Dep.*
*Tishk International University*
Erbil,Iraq
musa.m.ameen@gmail.com

Bilal I. Ahmed
*Information Technology Dep.*
*Tishk International University* Erbil, Iraq
bilalismael91@gmail.com

***Abstract*—In the past fifteen years, the grow of using Bag of Words (BoW) method in the field of computer vision is visibly observed. In addition, for the text classification and texture recognition, it can also be used in classification of images, videos, robot localization, etc. It is one of the most common methods for the categorization of text and objects. In text classification, the BoW method records the number of occurrences of each bag that is created for each instance type or word disregarding the order of the words or the grammar. And in visual scene classification it is based on clusters of local descriptors which are taken from the images disregarding the order of the clusters. The key idea is to generate a histogram for the words in the documents or the features in the images to represent the specified document or image. The BoW method is computationally and even conceptually is simpler than many other classification methods. For that reason, BoW based systems could record new and higher performance scores on common used benchmarks of text and image classification algorithms. This paper presents an overview of BoW, importance of BoW, how does it work, applications and challenges of using BoW. This study is useful in terms of introducing the BoW method to the new researchers and providing a good background with associated related works to the researchers that are working on the model.**

***Keywords—Bag of Words, Image Classification, Text Classification, Visual Scene Classification***

## I. INTRODUCTION

In image processing and computer vision, the Bag of Words (BoW) or Bag of Features (BoF), is also known as variable subset selection, attribute selection, or simply variable selection [1]. It is a multi-purpose model that can be used as feature selection algorithm, and classification of documents and images. In classification of images it can be processed by creating bags for each image features, and in classification of documents, a BoW is a vector of number of word occurrences, which is also called histogram of that document.

In BoW technique, the images is treated same as documents, and the features in the images are the same as the words in the documents. There are three main steps to achieve the histogram of the features in the images, which are: "feature detection", "feature description", and "codebook generation" [2].

In other terms, the BoW method can be defined as the "histogram representation based on independent features". After the process of feature detection, each image is isolated into a number of patches. And in feature representation process, the effort is for representing the patches as "numerical vectors", and those numerical vectors are called "feature descriptors" [3]. A great feature descriptor should have potential to treat processing features with different scales, rotations, illuminations, and deformations [4]. One of the common and widely used feature descriptors is "Scale-Invariant Feature Transform" (SIFT), it transforms each patch into 128 dimensional vectors. So, each image in SIFT can be divided into a collection or group of 128 dimensional vectors [5][6].

The last step of the BoW method for image classification is to produce "Code Books" (similar to a word dictionary in document classification) from transforming the vector represented patches into "Code Words" (similar to the words in text documents). In which the representation of a number of similar patches is called a Code Word. The size of the Code Book which is formed from a set of Code Words is equal to the number of the centers (similar to the size of the word dictionary). So, in the process of clustering, each of the patches in an image is designed to a certain Code Word and the histogram of the Code Words is used to represent the images in the process of visual scene classification [7].

This paper is organized as follows: section II mentions the works done previously on the BoW. Section III describes the importance of using BoW. Section IV explains how does BoW work as a classifier for text, images and videos. Also the applications, and challenges are presented in this section. Finally, some points are concluded in section V.

978-1-7281-4377-4/19/$31.00 ©2019IEEE

## II. Related Works

BoW is one of the methods that is used for feature selection and classification. This method is hot and has a great capability for selecting and classifying the features by creating bags for each instance type [1].

The researchers in [3] mentions that BoW model is a simplified representation used in Natural Language Processing (NLP) and Information Retrieval (IR). In this model, texts such as sentences, or documents are represented as the bag of its words, it is only considering the word duplicates, and ignoring the grammar and order of the words. The BoW model is mainly used in document classification methods, in which the features to train a classifier are produced from the occurrence or frequency of each word [8].

Classification of images and recognition of objects are part of the major interest areas for researchers ranging from classification of the image features to recognition of objects in the images including moving models [9]. One of the areas that is related to the BoW is local patch appearances, because of great functionalities such as robustness, simplicity, and good practical performances [10]. One of the pioneers in BoW is the authors in [3], which are applied BoW in analyzing documents and images, each document or image is treated as an unstructured set (“bag”) of those words or patches [4][2].

Texture recognition is done by local histograms of Texton codes. Retrieving content from images can be performed by using Textons [11], e.g. the researchers in [10] use a sparser bag of features model and SIFT descriptors over Harris-affine keypoint and avoiding global quantization by comparing the histograms.

## III. Importance of Bag of Words

It can be used for many purposes, because of its capability in different fields. Some of those purposes are:

A. *Text classification:* Classifying or categorizing texts and calculating weights for each word, which are the number of occurrences for each of the words.

B. *Image Recognition and Classification:* Classifying visual scenes (images and videos) by designing techniques that are capable to:

- Classify visual scenes: To detect a specified image, or to check whether an object on an image exists or not.
- Detect and localize objects: to find out which object or objects located on the image and where they are located on the image.
- Estimate semantic and geometrical attributes: To find out the orientation of the object (e.g.: To which direction does the man goes?) and to estimate the distance of a specified object to another.
- Classify human activities and events: To find out the movement of the objects (e.g.: What are the people in the image doing?).

## IV. Implementation of BoW Classification

In the BoW model, the images are represented as order-less groups of local features as the same as the words in the documents. The BoW name comes from the representation of bags of words in the retrieval of textual information, or bags of features from the visual scenes [12]. BoW model is getting more popular because of its high performance and simplicity. Because of the great success of BoW model in classification, it has been used in image processing field, for classifying images and video purposes [5].

This section explains the steps of implementation of text and image classifications and showing one example for each of them. Here, it has to be mentioned that videos are a sequence of frames, so classifying a video has the same process as classifying a set of images. Firstly, representation of images in the BoW model should be understood, and for explaining the document and image representation in the BoW model, there are two main perspectives, which are:

*A. Future Representation:* Represents a document as a vector of word occurrences for the specified text, or a vector of feature occurrences for an image, it is also called histogram of that text or image. Some of the non-informative words, such as; a, an, the, and, .etc will be removed from the dictionary after counting all the words from the dictionary which appear on the document, and the value of each element in the vector is the number of times a term or a word occurs in that document divided by the total number of the elements in the dictionary of that document. The same concept is available in representation of images in BoW model, a visual vocabulary is derived to show the dictionary by collecting the extracted features from a set of training images. Features representing the local areas of the images are the same as document words [13].

*B. Codebook Generation:* In this perspective, features will be extracted from the training images, and vectors will be quantized to develop a codebook. The nearest code in the codebook will be assigned for the features of an image. The set of codes will be used to reduce the features of an image and represented as a histogram. The normalized histogram of visual words is the same as the normalized histogram of codes [13].

Producing codewords (words in text documents) from vector represented patches is a very important step in BoW, and in turn the codebook (word dictionary in text documents) from the codewords. One of the simple and easy techniques in BoW is applying K-means clustering on all the vectors. There is a center for each cluster, and the codewords will be defined as those centers, and the size of the codebook will be equal to the number of the clusters. Finally, each image is formed from several patches, and by using the process of clustering each of the patches are mapped to a specific codeword, and the histogram of the codewords can be used to represent the images [7].

In other words, each image is treated as a word in the BoW model. This can be achieved in three basic steps which are: feature detection, feature description, and codebook generation [2]. After feature detection completed, each image or visual scenes are isolated by local patches for representing the features. Representing the patches as vectors of numbers is the key idea of the method of feature representation. The numerical vectors are called descriptors of the features. A descriptor with good performance should guarantee to treat texts and images with different scales, intensities, illuminations, and rotations.

The representation of term vectors of images in BoW method is very tight and compact process, because it discards the relative locations, spatial information, feature orientations, and feature scales. Briefly, the process of image representation in BoW is summarized as follows and shown in Fig. 1.

*A. Generating Vocabulary:* is the process of feature extraction from an image. Quantizing or clustering vectors of the features into a "visual vocabulary", each of the clusters represent a "visual word" or "term". In some researches, "visual codebook" is also defined as the vocabulary. Terms in the vocabulary are the codes in the "codebook".

*B. Defining Terms:* is the process of assigning the features to the closest terms in the vocabulary by using the Nearest Neighbors technique or a strategy that is related to it.

*C. Generating Term Vector:* is the process of creating a histogram which is representing a "term vector" by recording the count or numbers of each term that appears in the image. That "term vector" is the representation of BoW of an image.

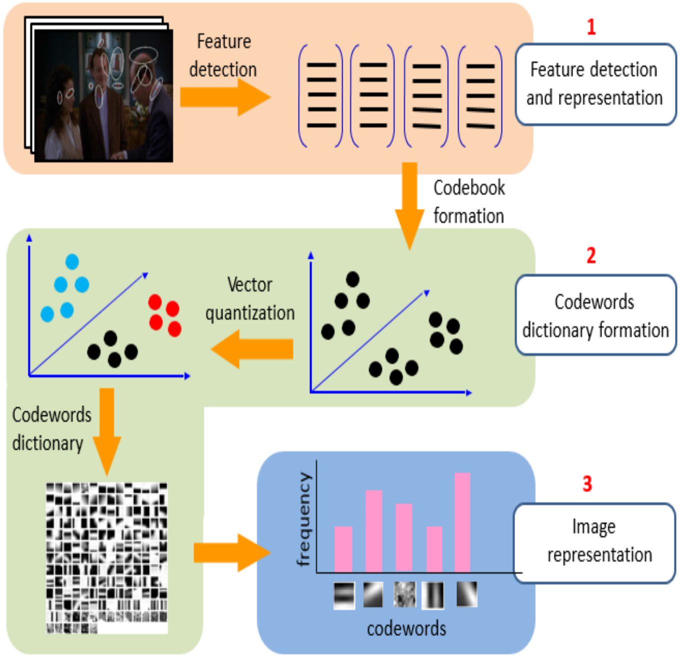


Fig. 1. The process of visual scene classification in BoW model

Visual scenes can be classified by applying the abovementioned three steps, the final step of the classification is the output from the histogram which is an image that is classified by the BoW.

Selecting the feature detection and feature representation methods is one of the key decisions in using BoW technique. For that purpose, one of the detectors that is using commonly and has good performance is Harris-Affine detector, and interest point operators as the detectors can also be used in this process [14]. Usually to describe the local image patches, a high-dimensional vector for the features is used. One of the popular and commonly used methods is the 128-dimensional SIFT descriptor.

The process of image or scene classification has the same concept as the document classification with some changes in the practical process only. For example, a filter is used to break the images into feature patches instead of the tokenizer to break the documents or sentences into words, the features are used instead of the words, that is why Bag of Words is also called Bag of Features, and the dictionary contains the features instead of the words [15].

In this part of the Implementation section, an example of document classification using BoW is explained:

*First Document:* Dara likes to go to cinema. Going to cinema is one of the hobbies of Azad.
*Second Document:* Dara also wants to go to play football and to go to swim

Based on the abovementioned text documents or sentences, a group of words is produced for each sentence by tokenizing the sentences to produce a dictionary of the words, as follows:

First Document: "Dara", "likes", "to", "go", "to", "cinema", "Going", "to", "cinema", "is", "one", "of", "hobbies", "of", "Azad"

First Document: "Dara", "also", "wants", "to", "go", "to", "play", "football", "and", "to", "go", "to" "swim"

Each bag of the words is represented as following:

BoW 1 = {"Dara":1, "likes":1, "to":3, "go":1, "cinema":2, "Going":1, "is":1, "one":1, "of":2, "the":1, "hobbies":1, "Azad:1"};

BoW 2 = {"Dara":1, "also":1, "wants":1, "to":4, "go"2, "play":1, "football":1, "and":1, "and":1, "swim:1"};

In BoW model, the order of the elements is free, and the number of duplications of each word is the value of that word in the document. And the "Union" of two text documents is the combination of the words of the both text documents as shown in " (1)".

$$\text{BoW } 3 = \text{BoW } 1 \text{ U BoW } 2 \quad (1)$$

The result of the "union" of the two documents will be:

*Third Document:* Dara likes to go to cinema. Going to cinema is one of the hobbies of Azad. Dara also wants to go to play football and to go to swim.

Which can be represented in the format of BoW as the following:

BoW3 = {“Dara”:2, “likes”:1, “to”:7, “go”:3, “cinema”:2, “going”:1, “is”:1, “one”:1, “of”:2, “the”:1, “hobbies”:1, “Azad”:1, “also”:1, “wants”:1, “play”:1, “football”:1, “and”:1, “swim”:1}

In practical experiments, the BoW model is basically used as a tool to generate and classify features. Various measures of text categorization can be generated by transforming text using BoW. The number of occurrences of a word or a term in a text document is known as term frequency, which will be used to specify the category of the text or simply classifying the text. As an example, for the above mentioned two text documents, the term frequencies of the duplicated words can be used to create histograms for each document, in which the BoW 3 is used to specify the term frequencies for the BoW 1 and BoW 2.

1) [1, 1, 3, 1, 2, 1, 1, 1, 2, 1, 1, 1, 0, 0, 0, 0, 0, 0]
2) [1, 0, 4, 2, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1]

Each of the elements in the sequence represents the number of the occurrences of the specified entries in the sequence, which is also called “Histogram Representation” of the document. For example, in the first sequence which defines the first document, the first two elements are “1, 1” which can be explained as follows:

- The first word in the document or the sentence corresponds to “Dara” which is the first element. The value of “Dara” is “1”, since in the first document it occurs only once.
- The second word in the document corresponds to “likes” which is the second element. The value of “likes” is “1”, since in the first document it occurs only once.

The representation of the sequences or vectors does not consider the sequence of the words or orders in the documents, which is one of the major properties of the BoW model. The Term frequency of documents are not the best representation, for that reason common terms like “the”, “a”, “to” has usually the maximum frequency terms in the documents. So, the terms with the highest frequencies does not mean that those words are the most important terms in the document. For that reason, the technique of “term frequency - inverse document frequency” is used to solve that kind of problem, which is one of the most common ways to “normalize” the term frequencies by weighting a term by the inverse of the frequency of that document. So, after discarding the common terms in the previously example, the most duplicated terms are “go” and “cinema”. It can be said that “go to cinema” is the main subject in the documents.

## A. *Applications*

Bag of Words can be used in different fields and in too much applications, some of the applications that are used in different fields are mentioned in this section, which are:

1) *Text classification:*
- Put a weight for each word (as shown in Fig. 2).
- Determine document’s topic.

2) *Image classification:*
- Is the image a city, a river, or a forest?
- Is any car existing in the image or not?

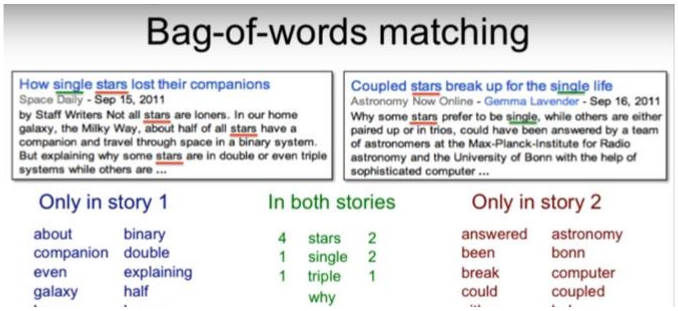
3) *Object detection:*
- Where is the car in the image?
- Which objects does this image contain?
- Detecting and reading the car plates (as shown in Fig. 3).
- Detecting specified objects (as shown in Fig. 4).

4) *Detecting orientations:*
- Is the car entering the park or exiting?
- Where the man looks for?

5) *Distances determination:*
- How much meters the car is far from the sea?
- What is the length of this building?

6) *Instance recognition:*
- Where is he waffle cake located in the market?
- Capturing car plates by radar on street.

7) *Event recognition:*
- Is the man walking or running?
- Is the man sitting or sleeping?



Bag-of-words matching

How single stars lost their companions
Space Daily - Sep 15, 2011
by Staff Writers Not all stars are loners. In our home galaxy, the Milky Way, about half of all stars have a companion and travel through space in a binary system. But explaining why some stars are in double or even triple systems while others are ...

Coupled stars break up for the single life
Astronomy Now Online - Gemma Lavender - Sep 16, 2011
Why some stars prefer to be single, while others are either paired up or in trios, could have been answered by a team of astronomers at the Max-Planck-Institute for Radio astronomy and the University of Bonn with the help of sophisticated computer ...

| Only in story 1 | | In both stories | | | Only in story 2 | |
|---|---|---|---|---|---|---|
| about | binary | 4 | stars | 2 | answered | astronomy |
| companion | double | 1 | single | 2 | been | bonn |
| even | explaining | 1 | triple | 1 | break | computer |
| galaxy | half | | why | | could | coupled |

Fig. 2. Determining weight for specified words.

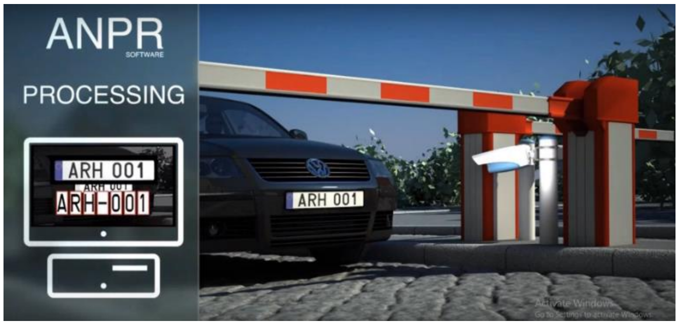
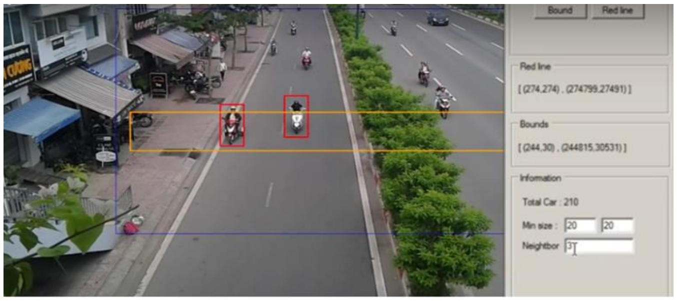


Fig. 3. Capturing car plate and translate to digital text.



Red line
[(274,274) . (274799,27491)]

Bounds
[(244,30) . (244815,30531)]

Information
Total Car : 210
Min size : 20 20
Neighbor 3

Fig. 4. Counting the motorbikes and cars differently that are in the range between the two yellow lines.

### B. Challenges

In this section the main challenges of BoW will be highlighted, in which the image will be difficult to be detected or fully unrecognized in the following cases:

- Viewpoint variation: the same image that is captured from different positions.
- Illumination: the same image with lighting from different locations.
- Deformation: the same object in which its shape deformed.
- Occlusion: two objects that are put behind each other and the result will be a new combined image.
- Background clutter: the object with a color that is the same or very near to the background color.
- Intra- class variation: an object with different kinds of the same type (e.g.: a chair with different kinds).

## V. CONCLUSION

The world goes through automatic and technological era, with lesser amount of people but faster and better work, and if we want a reader device to compute correctly and in a right manner, so they should be programed as well as they could work with one time train and work harder and faster. For that reason, BoW model have recently become more popular for content based text and visual scene classifications, because of its relative simplicity and good performance in a number of vision tasks. They evolved from Texton methods in texture analysis to image and even videos through frame classifications and representations. Algorithms with good performances should guarantee for the detection, determination, classification and recognition of objects. Finally, according to the good performance and capability of the BoW, it can be concluded that BoW can be used as classification algorithm for document and visual scenes, because it provides successful and validated outputs.

## REFERENCES

[1] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," Knowledge and Information Systems, vol. 53, pp. 551-577, 2017.

[2] M. Gabryel and R. Damaševičius, "The image classification with different types of image features," International Conference on Artificial Intelligence and Soft Computing, pp. 497-506, Springer, June 2017.

[3] A.A.A. Karim and R.A. Sameer, "Image Classification Using Bag of Visual Words (BoVW)," Journal of Al-Nahrain University-Science, vol. 21, pp. 76-82, 2018.

[4] N.M. Ali, S.W. Jun, M.S. Karis, M.M. Ghazaly, and M.S.M. Aras, "Object classification and recognition using Bag-of-Words (BoW) model," IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA), pp. 216-220, March 2016.

[5] W. Li, P. Dong, B. Xiao and L. Zhou, "Object recognition based on the region of interest and optimal bag of words model," Neurocomputing, vol. 172, pp. 271-280, 2016.

[6] M.M. Ameen, B. Ahmed, M. Anwar, and P.M. Hussein, "Wavelet Transform based Score Fusion for Face Recognition using SIFT Descriptors," Eurasian Journal of Science and Engineering, vol. 2, pp. 48-55, 2017.

[7] J. Cao, T. Chen, and J. Fan, "Landmark recognition with compact BoW histogram and ensemble ELM," Multimedia Tools and Applications, vol. 75, pp. 2839-2857, 2016.

[8] H.K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," Neurocomputing, vol. 266, pp. 336-352, 2017.

[9] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Local binary features for texture classification: Taxonomy and experimental study," Pattern Recognition, vol. 62, pp. 135-160, 2017.

[10] N. Passalis and A. Tefas, "Learning bag-of-features pooling for deep convolutional neural networks," Proceedings of the IEEE International Conference on Computer Vision, pp. 5755-5763, 2017.

[11] H. Mittal, and M. Saraswat, "Classification of histopathological images through bag-of-visual-words and gravitational search algorithm," Soft Computing for Problem Solving, pp. 231-241, Springer 2019.

[12] K. Li, F. Wang, and L. Zhang, "A new algorithm for image recognition and classification based on improved Bag of Features algorithm," Optik, vol. 127, pp. 4736-4740, 2016.

[13] N. Martinel, G.L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 567-576, March 2018.

[14] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," ISPRS journal of photogrammetry and remote sensing, vol. 138, pp. 281-294, 2018.

[15] S. Santurkar, D. Budden, and N. Shavit, "Generative compression," 2018 Picture Coding Symposium (PCS), pp. 258-262, June 2018.