Image Classification Using Bag of Visual Words (BoVW)

Abdul Amir Abdullah Karim[1] and Rafal Ali Sameer [2]
[1]Department of Computers, University of Technology, Baghdad-Iraq.
[2]Department of Computers, Collage of Science, University of Baghdad, Baghdad-Iraq.
Corresponding author: Ameer_aldelphi@yahoo.com
Corresponding author: Rafalali@scbaghdad.edu.iq

Abstract

In this paper two main stages for image classification has been presented. Training stage consists of collecting images of interest, and apply BOVW on these images (features extraction and description using SIFT, and vocabulary generation), while testing stage classifies a new unlabeled image using nearest neighbor classification method for features descriptor. Supervised bag of visual words gives good result that are present clearly in the experimental part where unlabeled images are classified although small number of images are used in the training process.
[DOI: 10.22401/ANJS.21.4.11]

Keywords: SIFT, Euclidean distance, classification, k-nearest neighbor, Bag of Visual Words.

1. Introduction

Recognition is the main problem of learning visual categories and classify new instances to those categories. Vision task almost relies on the capability to identify objects, scenes, and categories. Visual recognition has different applications that contact with many areas of artificial intelligence and information retrieval e.g. content based image, data mining, or object identification for mobile robots [1].

Content based image retrieval (CBIR) make it possible to search for and classify images. Images can be analyzed based on their features (such as color, textures, shape or edge). Keypoints are salient image patches that contain rich local information of an image, and they can be automatically detected using various detectors [2].

Local features have been widely used, the most well-known local features detection and description approaches are Speed Up Robust Feature (SURF) and Scale Invariant Feature Transform (SIFT). To find images similar to a query image, all images feature descriptors must be compared using some distance measures. Bag of Words (BoW) method has gained popularity. In BoW method, get clustered vectors of image features and create histograms (number of features occurrence) based on features descriptor. All the obtained descriptors in the histogram must be compared [3].

Bag of Visual Words (BoVW) model in computer vision represents image as visual words. The concept of BoVW is extract from the idea of Bag of Words (BoW) in the text document, therefore this techniques for text classification easily viable to the problem of image classification [4].

The remaining of this paper includes the following: section 2 presents a review about the existing works on image classification and Bag of Word. Section 3 presents the concept of Scale Invariant Feature Transform method. Section 4 presents a general concept of clustering and k-means clustering algorithm. Section 5 presents Euclidean distance metric for corresponding features comparison process. Section 6 presents K-Nearest Neighbor classification algorithm. Section 7 presents Bag of Visual Words approach. Section 8 presents image classification based on bag of visual words algorithm. Section 9 presents images of interest and the experimental results when running the algorithm on unlabeled images. Section 10 presents conclusion of this work.

2. Related Work

Various surveys for image classification using BoVW can be found, in literatures below some of those which are most related to this work:

1. In 2007, Jun Yang, Yu-Gang Jiang, Alexander Hauptmann, and Chong-Wah Ngo used text categorization steps to create different representations of visual word and studied their impact to classification performance on the TRECVID and

PASCAL collections. The Empirical study gives basis for representing visual word that is likely to create high classification performance [2].

2. In 2012, Mingyuan Jiu, Christian Wolf, Christophe Garcia, and Atilla Baskurt offers a novel method for learning supervised codebook and optimization bag of words approach. The proposed approach allows to evolve or keep the distinctive power of an unsupervised learning codebook while reducing of the learned codebook size. The codebook learning and recognition process are integrated to update the cluster centers through the back propagated errors: one is based on classical error backpropagation. The drawback of a gradient descent algorithm applied to a nonlinear system is difficult to learn a set of optimal parameters, the algorithms mostly converge to local minima and sometimes even diverge. The other is based on cluster reassignments algorithm which adjusts the cluster centers indirectly by rearranging the cluster labels for all the feature vectors. It needs more iterations to converge to a better solution [5].
3. In 2015, Marcin Korytkowski, Rafał Scherer, Paweł Staszewski, and Piotr Woldan presents method to classify and retrieve visual words using a novel relational database architecture. This work created a special database indexing algorithm, which will significantly speed up answering to visual query-by-example SQL queries in relational databases. The proposed method tested on three classes of visual objects and divided them into learning and testing examples. The testing set consists of 15% images from the whole dataset. Local keypoint generated before the learning procedure for all images using the SIFT algorithm. All simulations were performed on a hyper virtual machine [3].

3. Scale Invariant Feature Transform (SIFT)

SIFT is a local features detection and description algorithm, it is able to provide steady point for matching image. SIFT is popular algorithm for detecting important points which are invariant to image translation, image rotation, image scaling, and image lightening variation. SIFT is patent algorithm and take dense processing cost that make it too slow [6].

SIFT composed of four main stages: (a) detect scale space, (b) localize keypoints, (c) assign orientation, and (d) describe keypoints. The first step is define a location and a scales of the interest points using the extrema of scale space in the DoG (Difference of Gaussian) functions with various values of σ. Different scale of images created by using different value of σ in Gaussian function (σ in every scale separated by k that is constant value), then Subtract consecutive images to create DoG pyramid. DOG was used instead of Gaussian to increase the processing speed. After that the Gaussian image down sampled by 2 and create DoG to down sampled image. Gaussian function shown in equation (1) and DoG shown in equation(2) [7][8].

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left[- \frac{x^2+y^2}{2\sigma^2}\right] \quad \text{.....(1)}$$

Where
G (x,y,σ) represents a changing scale Gaussian,
σ represents the scale variable of the consecutive scale space,
x represents horizontal coordinates in Gaussian window,
y represents vertical coordinates in Gaussian window,
π = 3.14

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \quad \text{.....(2)}$$

Where
* represents the convolution operation,
k represents scaling factor,
G(x,y,σ) represents a changing scale Gaussian function,
I(x, y) represents an input image,
D(x,y,σ) represents Difference of Gaussians have k times scale,
x represents horizontal coordinate in image (I(x,y)) with corresponding horizontal coordinate in Gaussian window (G(x,y,σ)),
y represents vertical coordinate in image (I(x,y)) with corresponding vertical coordinate in Gaussian window (G(x,y,σ)).

Local extrema obtained by comparing every pixel after DoG with 26 other pixels (eight neighbor pixels at the current pixel's level and nine pixels in the upper level and nine pixels in the lower level. When the compared pixel is extrema (minimum than all 26 pixels or maximum than all 26 pixels), pixel position and scale are saved. In the keypoint localization step, low contrast points and points at edge are eliminated. Intervention point is also eliminated by using (2×2) Hessian matrix [7][9].

The descriptors build by calculating the gradient strength and orientation strength for each neighbor of a keypoint. The neighborhood of every keypoint are characterized by creating 8 bins gradient and orientation histogram for 16×16 region of neighbors around keypoint. The region is split up into 4×4 sub regions and each sub region have 8 directions this will produce 4×4×8= 128 dimensional vector to give description for every keypoint [9][10].

The existence of large number of features will produce irrelevant or redundant features that increasing the processing time and can also affect the accuracy. The aim of feature selection is to reduce feature space dimensionality and to keep the distinctive features [11].

4. K-means clustering Algorithm

Clustering is unsupervised iterative method that classify group of points into clusters based on similarity. The similarity measure frequently depending on distance methods e.g. Euclidean distance to classify points in groups (or clusters) [12].

K-means clustering algorithm is an unsupervised classification procedure which classifies or groups the objects automatically into K number of group where each group contain points that have minimum distance between them. K-means is also called C-means or ISODATA clustering method. K-means algorithm initialize clusters center (or centroids) by selecting samples at random from training vectors. K-means is repeated method which used to collect data into groups and these groups change every iteration [13].
K-means algorithm described as follows [12]:

1) Select integer number k at random that represent the number of centroid.
2) Compute the average distance between every data point and centroids using Euclidean distance equation (4).
3) The data point assigned to the cluster when distance between data point and cluster center is minimum than the distances with other centroids.
4) Repeat calculation of the new centroid using:

$$V_i = \left(\frac{1}{ci}\right) \sum_{j=1}^{ci} Xi \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

5) Repeat calculation of the distance between every data point and the new obtained centroids.
6) If the data point was not reassigned then break, otherwise repeat from step 3.

Note that:
K represents positive integer number,
x represents a group of data points $\{x_1, x_2, x_3, \ldots, x_m\}$,
c_i represents number of data points in *i-th* cluster,
V represents set of centers $\{v_1, v_2, \ldots, v_c\}$.

5. Euclidean distance

Euclidean distance is considered as the standard metric for geometrical problems. It is simply the ordinary distance between two points. Euclidean distance is extensively used in clustering and classification problems. It is the default distance measure used with the K-means and k-nearest neighbor algorithms. The Euclidean distance determines the root of square differences between the coordinates of a pair of objects as shown in equation (4) [12]:

$$\text{Dist}_{(P1,\,P2)} = \sqrt{\sum((x2 - x1)^2 + (y2 - y1)^2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Where
$P_1\,(x_1, y_1)$ is the First point,
$P_2\,(x_2, y_2)$ is the second point.

6. K-Nearest Neighbor

The K Nearest Neighbor classifier (or called instance based classifier) is a traditional nonparametric classification algorithm that gives good performance for best value of k. The KNN classifier performs classification of

unlabeled image by relating the unlabeled image's features to the labeled features depending on distance function (equation (4)) or similarity measure. In the k nearest neighbor a test sample allocates to the class that frequently describe among the k nearest training samples. If two or more such classes exist, then the test sample is assigned the class with minimum average distance to it [14].

7. Bag of Visual words (BoVW)

The image has keypoints or local features identified as prominent image regions that have rich local information (such as color or texture) and these features can be detected using different detection and description method. Detected features are then split to a number of clusters using the K-means clustering algorithm where each cluster will have features with similar descriptors and encodes each keypoint by the index of the cluster to which it belongs this is called *vector quantization (VQ)* technique [2].

The VQ can be considered as a generalization of scalar quantization to the quantization of a vector. The VQ encoder encodes a given set of k-dimensional data vectors with a much smaller subset. The subset C is called a codebook and its elements C_i are called codewords, codevectors, reproducing vectors, prototypes or design samples. The commonly used vector quantizers are based on nearest neighbor called Voronoi or nearest neighbor vector quantizer [13].

Each cluster represented by a visual word that represents the specific local pattern participated by the keypoints in that cluster, so a visual word vocabulary identifies all types of local patterns of image. The image can be identifies as a bag of visual words, or in other words, as a visual-word vector containing the number (weight) of each visual word in image (i.e., the number of keypoints in the corresponding cluster), which in classification task can be used as a feature vector [2].

BoVW approach in general creates supervised classifiers depend on visual words taken from labeled images for label prediction of a new image. Therefore the clustering method creates a visual words vocabulary to describe different local patterns in images. The clusters number defines the size of the vocabulary that can vary from hundreds to more than tens of thousands. By mapping the keypoints to visual words each image can be represented as a "bag of visual words" [2].

BoVW for new unlabeled images calculated in a similar way: local features extraction from image and features description, projection of these descriptors on the dictionary calculated previously by the training set, and histogram calculation of each visual word appearance of the dictionary [5].

8. Proposed Algorithm

The proposed Algorithm of image classification using bag of visual words can be described by two main algorithms training algorithm and testing algorithm as follows:

Training Algorithm
Input (collection of image**)**
Output (k - clusters, k - visual word**)**

Step 1: Collect set of images for each class of interest (in this paper the experimental class of interest are Car, Motor, and Ship).

Step 2: Apply BoVW on collected images. BoVW consists of three main steps:

1. Extract keypoints from images using SIFT feature detection and description algorithm.
2. Create descriptor for each extracted keypoints.
3. Clustering features using k-means clustering algorithm (Create visual vocabulary using vector quantization of descriptor space) and save the resulting "visual words".

Testing Algorithm
Input (k - visual word**)**
Output (labeled image**)**

Step 1: Open unlabeled new image.

Step 2: Extract and describe features of unlabeled image using SIFT.

Step 3: Extract visual word (centroid) for testing image.

Step 4: Calculate the nearest neighbor using Euclidean distance between visual word of tested image and visual words of training images.

Step 5: Take the decision: compare extracted features of unlabeled image with visual words extracted in training stage.

The new image classified by finding the smallest distance between the new image's feature vector and the feature vectors of clusters obtained in the learning phase (unlabeled image belong to the cluster that have smallest distance with it based on their feature vectors or the differencing between histograms of unlabeled image and training cluster are the smallest histogram differencing which indicate that the two clusters are Converged).

9. Experimental Results

Quantitative performance of algorithms is reported in terms of sensitivity, specificity. Sensitivity is the fraction of positive class sample correctly classified (the ability of the classifier to find all the positive samples). Specificity is the fraction of negative class sample correctly identified. Accuracy is the proportion of true results, either true positive or true negative, in a population. It measures the degree of veracity of a diagnostic test on a condition [15].

$$\text{Sensitivity} = \frac{\text{Tp}}{\text{Tp} + \text{FN}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Tp+TN+FP+FN}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \ldots (7)$$

Note that,

TP (true positive): represents the number of images correctly labeled with corresponding class by the algorithm, FP (false positive): represents the number of images not exist in training set but labeled as one of the clusters (unexpected result), FN (false negative): represents the number of Missing images, TN (true negative): represents the number of images not exist in training set and not labeled correctly [15].

The program written using VisualBasic.net programming language, data set for three different classes was used for training process (car, ship, and motor) every data set have 16 images.

First data set (car class):

Second data set (ship class):

Third data set (Motorbike class):

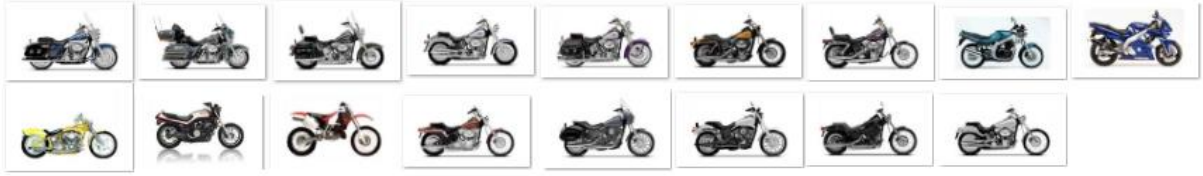

Table (1)
Classification performance.

Class	Sensitivity	Specificity	Accuracy
Car	0.875	0.875	0.9
Ship	0.83	0.7	0.8
Motorbike	0.7	0.83	0.8

The experimental results of (16) unlabeled images (tested images) present in Table (1) shows the performance of nearest neighbor classification method based on bag of visual words.

Nearest neighbor based on Euclidean distance for (16) tested image will present in Table (2), where the distance calculated using x,y position of tested image centroid and x,y position of training images centroid in each cluster. At the right side of Table (2) the original image type while at the left side the

distance with each cluster the minimum distance describe the class of tested image.

Table (2)
K-Nearest Neighbor.

Original image	Car class	Ship class	Motorbike class
Car	14	154	154
Car	30	154	17
Car	18	30	154
Car	10	16	19
Car	154	574	518
Ship	21	15	24
Ship	24	11	51
Ship	17	9	18
Ship	18	30	37
Ship	15	18	23
Motorbike	8	44	7
Motorbike	23	154	17
Motorbike	14	16	6
Motorbike	9	16	9
Motorbike	16	21	17
Motorbike	14	17	11

The ratio of sensitivity and specificity is limited between 0 and 1; that is the ratio of true classification, high value of sensitivity and specificity give impression of good method performance.

10. Conclusion

Bag of visual word (BoVW) technique is an efficient image representation in the classification task. In this paper there are two main stage, the first is the training stage and the second is the testing stage, every stage have number of steps. In general the first stage create visual vocabulary from training images. The information that extracted in the first stage used to classify new unlabeled image based on bag of features created using supervised BoVW approach on set of training images. This approach gives very good results although small number of images using in training process.

References

[1] Kristen G., "Visual Object Recognition", thesis, 2010.

[2] Jun Y., Yu-Gang J., Alexander H., Chong-Wah N., "Evaluating Bag-of-Visual-Words Representations in Scene Classification", Proceedings of the international Workshop on Workshop on Multimedia information Retrieval, vol. 2, pp. 197-206, 2007.

[3] Marcin K., Rafał S., Paweł S., Piotr W., "Bag-of-Features Image Indexing and Classification in Microsoft SQL Server Relational Database", IEEE, 46, 746-751, 2015.

[4] Pornntiwa P., Emmanuel O., Olarik S., Lambert S., Marco W., "Comparing Local Descriptors and Bags of Visual Words to Deep Convolutional Neural Networks for Plant Recognition", 6th International Conference on Pattern Recognition Applications and Methods, 1, 886-893, 2017.

[5] Mingyuan J., Christian W., Christophe G., Atilla B., "Supervised Learning and Codebook Optimization for Bag-of-Words Models", Springer Science Business Media, 4, 409-419, 2012.

[6] Yi H., Guohua D., Yuanyuan W., Ling W., Jinsheng Y., Xiqi L., Yudong Z., "Optimization of SIFT algorithm for fast-image feature extraction in line-scanning ophthalmoscope", Optik journal, 152, 21-28, 2017.

[7] El-gayar M., Soliman H., Meky N., "A comparative study of image low level feature extraction algorithms", Egyptian Informatics Journal, 14, 175-181, 2013.

[8] Panchal P., Panchal S., Shah S., "A Comparison of SIFT and SURF", International Journal of Innovative Research in Computer and Communication Engineering, 1, 323-327, 2013.

[9] Jian W., Zhiming C., Victor S., Pengpeng Z., Dongliang S., Shengrong G., "A Comparative Study of SIFT and its Variants", Measurement Science Review, 13, 122-131, 2013.

[10] Pedro J., "Contribution to the completeness and complementarity of Local Image Features", thesis, 2013.

[11] Soumyadeep G., Tejas I., Rohit K., Richa S., Mayank V., "Feature and Keypoint Selection for Visible to Near-infrared Face Matching", International Conference on

Biometrics: Theory, Applications & Systems, 978, 1109-1119, 2015.

[12] Jasmine I., Nitin P., Madhura P., "Clustering Techniques and the Similarity Measures used in Clustering: A Survey", International Journal of Computer Applications, 134, 93-103, 2016.

[13] Balwant A., Doye D., "Speech Recognition Using Vector Quantization through Modified K-means LBG Algorithm", Computer Engineering and Intelligent Systems, 3, 137-144, 2012.

[14] Aman K., Singh M., "A Review of Data Classification Using K-Nearest Neighbour Algorithm", International Journal of Emerging Technology and Advanced Engineering, 3, 354-360, 2013.

[15] Chamaa C. , Mukhopadhyaya S., Biswasa P., Dharaa A., Madaiahb M., Khandelwalb N., "Automated Lung Field Segmentation in CT images using Mean Shift Clustering and Geometrical Features", Medical Imaging 2013: Computer-Aided Diagnosis, 8670, 867032-867042, 2013.