ICIC Express Letters
Volume **11**, Number **10**, October **2017**
 ISSN 1881-803X
pp. **1479–1488**

CONTENT BASED IMAGE RETRIEVAL USING BAG OF VISUAL WORDS AND MULTICLASS SUPPORT VECTOR MACHINE

Suharjito, Andy and Diaz Djaja Santika

Master in Computer Science, Binus Graduate Program
Bina Nusantara University
Jl. Kebon Jeruk Raya No. 27, Jakarta, Indonesia
suharjito@binus.edu; andy.ang1991@gmail.com; ddsantika@binus.ac.id

Received April 2017; accepted June 2017

Abstract. *Content Based Image Retrieval (CBIR) is an image retrieval method that is based on the meta content of an image which is proven to be more efficient than image retrieval method that is based on text or key words. Bag of Visual Words (BoVW) is a popular approach and the most widely used in CBIR problem. In this research, a CBIR method using BoVW and multiclass SVM classifier will be proposed. In BoVW, Scale Invariant Feature Transform (SIFT) will be used as the local features descriptor. This research will use Gaussian Mixture Model (GMM) as the method to do the visual vocabulary generation and Fisher Vector (FV) to create the encoder. The multiclass SVM classifier will use linear kernel, Hellinger's kernel, and chi-square kernel to do the classification. After the classification, the proposed CBIR will use the classification model to classify query image. After the query image class is known, then the color histogram features will be extracted from query image and dataset which only consists of image in the same class as query image. Datasets used in the research are Corel and Guang-Hai Liu (GHIM-10K). The experimental results show that BoVW with GMM and FV produces higher accuracy than other encoding methods in BoVW which produce a good retrieval result in the CBIR.*
Keywords: Content Based Image Retrieval, Bag of Visual Words, Image classification

1. **Introduction.** The rapid development of digital technology has caused the amount of image data to increase at a rapid pace almost equal to the number of text data. However, searching data in image data is not as easy as searching data in text data. Searching in image (image retrieval) becomes one of the most commonly researched topics because of the need of an effective method to do searching in image data [1]. In general image retrieval is divided into 2 methods, which are text based and content based. Recent research shows that text-based method is not effective to represent an image into a small set of keywords which then shifts recent research topics to content-based method [2].

The idea of Content-Based Image Retrieval (CBIR) is searching an image using Meta data from the image. CBIR uses low level features in the searching method such as color, texture, or shape [3]. In general, the features which CBIR uses can be categorized into 2 key features: global feature descriptors and local feature descriptors. Global feature descriptors describe an image as a whole, while local feature descriptors describe an image by image patch (a small pixel). Global feature descriptors have an advantage in terms of its computational speed but the drawback is the low accuracy. Global feature descriptors tend to fail in identifying important visual characteristics from an image. On the other hand, local feature descriptors produce better accuracy than the global feature descriptors mainly because the image is represented by features calculated from the patch in the image. The drawback of the local features is that the size of feature space is large and will become very huge for large image database [4].

Recent research in CBIR shows that the approach using Bag of Visual Words (BoVW) produces better results than the other approaches. BoVW is a popular method for solving CBIR problems. BoVW imitates the method of text retrieval which is called bag of words. Research on BoVW shows that BoVW performs well not only in CBIR but also in object recognition, image classification, and image annotation [2]. BoVW is a method that treats an image as a distribution of local feature descriptors where each descriptor was given a label further called as visual words. A set of visual words is called visual vocabulary or codebook. An image is represented by a set of visual words. The proccess of representing an image into a set of visual words is usually called encoding image [3]. One of local feature decriptors that is popular and shows an excellent result in BoVW is Scale Invariant Feature Transform (SIFT). SIFT shows good performance in the change of rotation and scale. It is capable of capturing the edges of an image. The other advantage of SIFT is its excellent performance in images that have a simple background; SIFT represents them without noise [5].

In general, the approach to visual vocabulary generation process consists of hard assignment and soft assignment. Most visual vocabulary uses hard assignment to generate the visual vocabulary, but soft assignment method shows better results in [6]. Soft assignment method uses probability distribution from the features of space to determine the probability value of visual words in feature spaces. Soft assignment does not only convey the mean value of visual word like hard assignment, but also conveys the shape of the distribution of the visual words. K-mean clustering algorithm is one of the most used methods in hard assignment while GMM is one of the commonly used methods in soft assignment [6].

The encoding image step in BoVW uses an encoder to encode the image. The approach used in the step to create an encoder depends on the method used in visual vocabulary generation. Vector Quantization (VQ) is a widely used method in feature encoding which uses k-means algorithm. Research on feature encoding in BoVW indicates that Fisher Vector (FV) encoding method performs better than VQ [8]. FV uses GMM in the visual vocabulary generation and GMM model in the encoding method. The result of the research indicating that FV performs better than VQ is in line with GMM method, which is better than k-means in visual vocabulary generation. In FV, each feature descriptor is not assigned to only one quantization cell (visual word) but to several quantization cells by using the probability value for each quantization cell [7].

The focus of this research is to increase the retrieval accuracy in CBIR. The approach used to increase the retrieval accuracy is using BoVW [2] and SVM [8] as classifier model. SIFT features will be used as the local features descriptor [5] and Fisher Vector (FV) encoding as the encoding method. By using Fisher Vector, GMM will be used as a method to do visual vocabulary generation. The feature space then will be fed to SVM classifier to train the classifier. The trained SVM classifier model will be used to classifiy the query image class. After classification, color histogram which is the most widely used color features to represent an image [9] will be used to calculate the degree of similarity between query image and dataset images. The dataset used in the similiarity measurement only consists of the same image class as the query image.

The paper is organized as follows. Section 2 describes the datasets used in this work, proposed set of features and its feature extraction method. Section 3 outlines the metodology used in this work. Section 4 presents the result of the experiments and comparison with other methods. At last, the conclusions of the research are presented in Section 5.

2. **Literature Review.**

In this section, the theoretical background of the methods used in the experiment is described. Scale Invariant Feature Transform (SIFT) is used as local features descriptor in this work due to the SIFT advantages and because SIFT is one of the popular features used for BoVW [10]. SIFT shows excellent results for BoVW because

SIFT has an advantage in the change of rotation, and the scale is capable of capturing the edges of an image. Other advantages of SIFT are its excel in images that have a simple background and representing them without noise. The disavdvatage of SIFT is in complex background and illumination in image [5]. In general SIFT consists of 4 steps [4].

- Extrema Detection: In extrema detection, the image will be checked using different octave and scale to isolate points from the image that are different from its surrounding. Those points are called extrema, potential candidates to become image features.
- Keypoints Localization: In this step, some of the extrema will be chosen to become keypoints. The keypoints will further be refined by rejecting extrema that are caused by edges and low contrast point.
- Orientation Assingment: In this step, magnitude and direction are used to represent every key point and its neighborhoods as a set of vectors.
- Keypoint Descriptor: In this step, descriptors for the image are created. Descriptor is a combination with a set of eight vectors from a collection of vectors in the neighborhood of every keypoint.

Gaussian Mixture Model (GMM) is a parametric probability density function that represents the total weight of density for the Gaussian component. In general, the GMM equation is [11]:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g\left(x|\mu_i, \Sigma_i\right) \tag{1}$$

where x is a D-dimensional of continue vector, w_i, $i = 1, \ldots, M$ is the mixture weight, $(x|\mu_i, \Sigma_i)$, $i = 1, \ldots, M$ is the component of the Gaussian densities, and M is the number of Gaussian components. Every densities component is Gaussian function that follows the following equation:

$$g\left(x|\mu_i, \Sigma_i\right) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'{\Sigma_i}^{-1}(x-\mu_i)\right\} \tag{2}$$

where μ_i is the mean vector and Σ_i is the covariance matrix which must fulfill

$$\sum_{i=1}^{M} w_i = 1 \tag{3}$$

The mean vector, covariance matrix, and mixture weight are calculated in a density component which uses the following equation:

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M \tag{4}$$

In GMM, the parameter λ is usually learned by using Maximum Likelihood (ML). The purpose of ML is to determine a model of parameter to maximize the GMM. For a T sequence training of vector $= \{X_i, \ldots, X_T\}$, the GMM ML uses the following equation:

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_T|\lambda) \tag{5}$$

The most used approach in ML is Expectation-Maximization (EM) algorithm. The idea of EM is to estimate the value of λ from the initial value of λ so that it achieves condition $p\left(X|\overline{\lambda}\right) \geq p(X|\lambda)$ [11].

Fisher Vector (FV) is a framework deriving from fisher kernel which combines the advantages of generative and discriminative approaches [12]. The construction of FV will start by learning the GMM model. FV will capture the average of first and second order

differences between the image descriptor and centers of GMM. The FV method uses the following equation [8]:

$$u_k = \frac{1}{1\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \Sigma_k^{-\frac{1}{2}} (x_t - \mu_k) \tag{6}$$

$$v_k = \frac{1}{1\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \left[(x_t - \mu_k) \Sigma_k^{-1} (x_t - \mu_k) - 1 \right] \tag{7}$$

x is a set of descriptors $x_1, \ldots, x_N$ from an image, q_{ik} which $k = 1, \ldots, K$ and $i = 1, \ldots, N$ are the soft assignment of the N descriptor to the K Gaussian components, and π_k is the probability value.

Support Vector Machine (SVM) is a classifier that is widely used to classify an image and shows an excellent result in image classification. SVM has high effectiveness in high dimensional data [1]. The SVM classifier model in this work will use linear kernel and non-linear kernel approaches. Non-linear kernel tends to produce better accuracy than the linear kernel but it has low efficiency. There are non-linear kernels that not only perform better than linear kernels but also have high efficiency, which are called additive homogeneous kernels. Hellinger's kernel and chi-square kernel are examples of additive homogeneous kernels which are widely used [8]. In this work, linear kernel, Hellinger's kernel, and chi-square kernel will be used as the kernels in SVM classifier. The linear kernel performs no transformation to the input vector; meanwhile the other 2 kernels will perform transformation according to the following equation:

$$\text{Hellinger's Kernel: } \sum_{i=1}^{d} 2\sqrt{x_i y_i} \tag{8}$$

$$\text{Chi-square Kernel : } \sum_{i=1}^{d} 2\frac{x_i y_i}{x_i + y_i} \tag{9}$$

Color features are very stable and robust compared to the other features such as texture and shape. Color features are not sensitive to rotation, translation, and scale changes. Color histogram is one of the color features that can describe the global color distribution of an image; it is a method mostly used in color features [9]. Color histogram has high effectiveness, simplicity and low storage requirement which makes color histogram perform better than other color features [13]. The color space of the image will be converted to HSV color space from RGB color space because it does not meet the visual requirement in the image retrieval. The image will be quantified according to the following equation:

$$\begin{array}{rll} & 0 & h \in [316, 360] \\ & 1 & h \in [1, 25] \\ & 2 & h \in [26, 40] \\ H = & 3 & h \in [41, 120] \\ & 4 & h \in [121, 190] \\ & 5 & h \in [191, 270] \\ & 6 & h \in [271, 295] \\ & 7 & h \in [295, 315] \\ & 0 & s \in [0, 0.2] \\ S = & 1 & s \in [0.2, 0.7] \\ & 2 & s \in [0.7, 1] \\ & 0 & v \in [0, 0.2] \\ v = & 1 & v \in [0.2, 0.7] \\ & 2 & v \in [0.7, 1] \end{array} \tag{10}$$

After the color histogram is extracted, the similarity of the query image and the images in the dataset is calculated using Euclidean distance [14], which uses the following equation [9]:

$$D = \sum_{i=1}^{n} (A_i - B_i)^2 \tag{11}$$

where A and B are the 2 features vectors and n is the dimensional size of the features vector. The images retrieved then will be ranked according to the similarity to the query image.

3. **Research Methods.** In this research, the methods used in the experiment consist of feature extraction using SIFT, visual vocabulary generation using GMM, the encoder generation using fisher vector, training the SVM classifier, and then retrieved images by its similarity using trained classifier to classify query image by using color histogram as the features and Euclidean distance as the similarity measurement. The experiment process can be seen in Figure 1. The dataset used in this work can be downloaded from http://www.ci.gxnu.edu.cn/cbir/Dataset.aspx. The database used consists of Corel

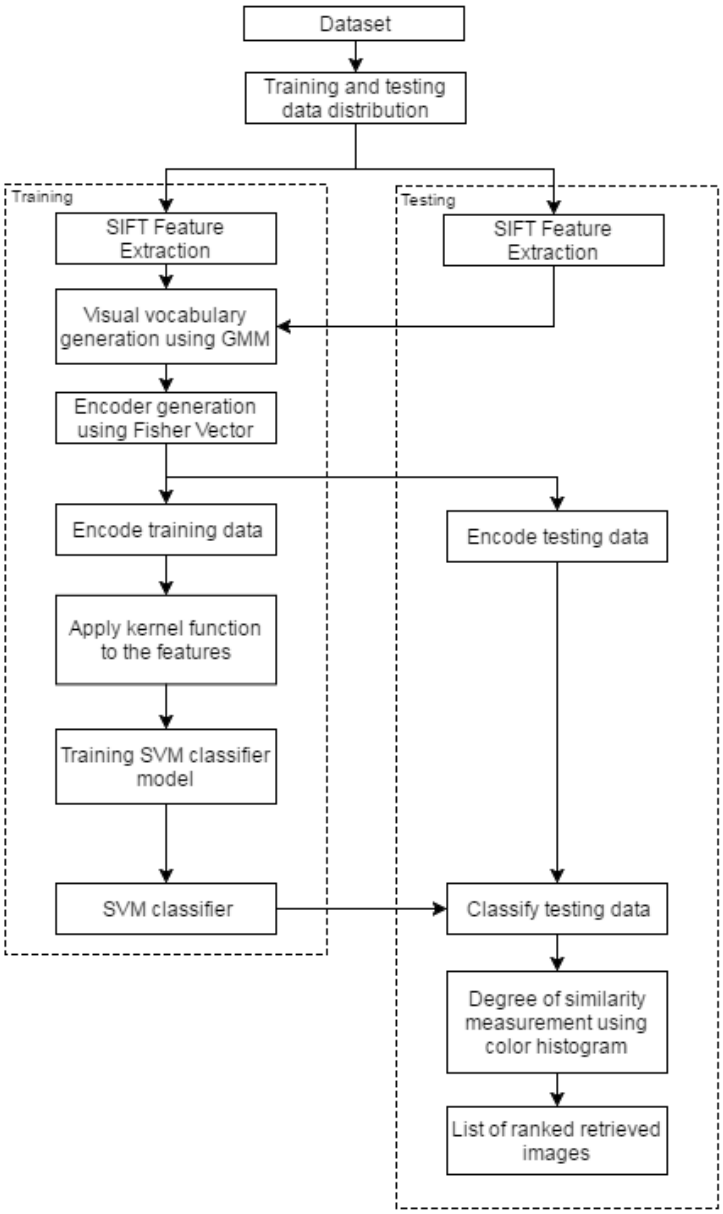

FIGURE 1. The research methods

FIGURE 2. Samples of Corel and GHIM-10K dataset

image and Guang-Hai Liu (GHIM-10) captured by the author. The Corel dataset has 10000 images and 100 classes. GHIM-10K dataset has the same 10000 images as Corel but fewer classes, which are 20 classes. The dataset will be further organized into folders according to their class. The Corel image resolution consists of 126×187 for portrait orientation and 187×126 for landscape orientation. Meanwhile, GHIM-10K dataset has 300×400 for portrait orientation and 400×300 for landscape orientation.

The datasets are divided into 2, which are training dataset and testing dataset. Training dataset consists of 2 images of each class in the dataset. In total, the training dataset has 240 images and the testing dataset has 19760 images. The experiment will also use other testing dataset that consists of 2 images of each class from the training dataset. We name this dataset as the non-exclusive dataset because the dataset consists of the same population of the training dataset while the former dataset is exclusive testing dataset.

The main idea in visual vocabulary generation is to assign features space into a collection of visual words by quantizing into some clusters. The center of each cluster is then called visual word. Meanwhile, the purpose of encoder in BoVW is to map local features from an image to the corresponding visual word from the created visual vocabulary. The process is often called encode image. Encoder in BoVW depends on the method which the visual vocabulary creates. In this work, the module in VLFeat library will be used to do the experiment.

3.1. **Training phase.** For training phase, the first step is conducting SIFT features extraction from the images. The features extracted are clustered using GMM method to create visual words. The collection of visual words is called visual vocabulary; hence, this step is also called visual vocabulary generation. The next step is to create the encoder based on the clustering method used in the visual vocabulary generation step. The created encoder then is used to encode the images to make a representation of images into a collection of visual words.

Before the image feature spaces are used to train the SVM classifier, the feature space is transformed using the concerned kernel equation. In this work, linear, Hellinger's, and chi-square kernels are used to identify the effectiveness of the kernel usage in SVM. After the kernel function is applied to the feature spaces, then SVM classifier will be trained using the training dataset which will produce an SVM classifier to use later in testing phase.

3.2. **Testing phase.** In testing phase, the first step is similar to the training phase, which is SIFT feature extraction. Because the visual vocabulary is already known in the training phase, in the testing phase, the visual vocabulary is not generated again. The same goes for the encoder generation. The testing phase will use the encoder created from the testing phase to encode the testing data. The feature spaces from the encoding are going to be used to classify the images class using the trained SVM classifier. After an image class is known, the next step is to do retrieval images. The retrieval will calculate the degree of similarity between the query image and the dataset using color histogram.

The measurement uses Euclidean distance. The dataset used in the measurement consists of only the images that have the same class as the query image. After the measurement, the retrieved images will be shown in a ranked list based on the similarity with the query images.

4. **Results and Discussion.** In this work, confusion matrix and mean average precision (mAP) [15] are used to evaluate the method used in experiment. The Vector of Locally Aggregated Descriptors (VLAD) encoding and Vector Quantization (VQ) are used to show the result comparison to the Fisher Vector encoding method. VLAD and VQ use k-means algorithm for the visual vocabulary generation. Each encoding method will be tested on 3 linear kernels for SVM classifier. The number of visual words for VQ is set to 1024. VLAD and FV are set to 64. The SIFT is computed at five scales with a factor $2–\sqrt{2}$ between successive scales, bins 8 pixel wide, and computed with a step of 3 pixels. All experiment will use SIFT as the local features descriptor. After the classification experiment, CBIR will use 2 images as the query to show the retreived images by using the classification result before doing retrieval.

Table 1 and Table 2 respectively show the accuracy of confusion matrix for exclusive test dataset and non-exclusive test dataset. As shown by the accuracy of the confusion matrix result, Fisher Vector (FV) encoding method that uses GMM to create visual vocabulary performs better than the other 2 encoding methods that only use k-means to create the visual vocabulary. The results also show that chi2 kernel produces better result when paired with fisher vector. Meanwhile, linear kernel performs better when paired with the 2 other encoding methods as linear kernel has the worst performance in FV. The small differences between exclusive and non-exclusive test dataset show that the performance of SVM classifier is only slightly reduced when classifying new images.

Table 1. Confusion matrix with exclusive test dataset

Confusion Matrix Accuracy – 120 class Exclusive			
Kernel	**Encoder Method**		
	Fisher Vector	**VQ**	**VLAD**
Linear	71.67%	62.50%	67.08%
Hellinger's	73.30%	58.75%	64.58%
Chi2	73.30%	62.08%	65.00%

Table 2. Confusion matrix with non-exclusive test dataset

Confusion Matrix Accuracy – 120 class Non-Exclusive			
Kernel	**Encoder Method**		
	Fisher Vector	**VQ**	**VLAD**
Linear	74.58%	65.00%	69.17%
Hellinger's	74.58%	61.25%	66.67%
Chi2	75.83%	65.00%	69.17%

Table 3 and Table 4 respectively show the value of mean Average Precision (mAP) for exclusive test dataset and non-exclusive test dataset. As shown in the results, Fisher Vector (FV) encoding method still outperforms the other 2 encoding methods. The difference between FV and 2 other encoding methods is about 10% which is larger than the difference between FV and other encoding methods in the confusion matrix, which is only about 5%. The results also show that chi2 kernel still produces better result when paired with fisher vector. Meanwhile, the linear kernel's performance still shows a better result when paired with the other 2 encoding methods.

TABLE 3. mAP results with exclusive test dataset

mAP – 120 class Exclusive			
Kernel	**Encoder Method**		
	Fisher Vector	**VQ**	**VLAD**
Linear	79.65%	69.28%	69.17%
Hellinger's	79.56%	68.24%	66.67%
Chi2	80.17%	70.15%	69.17%

TABLE 4. mAP results with non-exclusive test dataset

mAP – 120 class Non-Exclusive			
Kernel	**Encoder Method**		
	Fisher Vector	**VQ**	**VLAD**
Linear	81.74%	71.78%	75.94%
Hellinger's	82.83%	72.24%	74.55%
Chi2	83.59%	73.87%	75.04%

TABLE 5. Result comparison with previous research

Method	**Dataset**	**mAP**
[8]	PASCAL VOC 2007	61.69%
	Clactehc-101	77.78%
[3]	Wang	63.39%
[5]	COREL-1000	70.58%
	COREL-1500	68.05%
	COREL-2000	58.31%
	Olivia and Torralba	69.75%
	Groundtruth	83.53%
Proposed Method	**COREL & Guang-Hai Liu**	**80.17%**

For comparing with previous research in BoVW, this research uses [3], [8], and [5] as shown in Table 5. The result shows that the proposed method performs better than the previous research except for dataset Groundtruth in [5] which yields 83.53% as the result. This is because the Groundtruth dataset only has 228 images and 5 classes, far different from the dataset used in this research, which has 20,000 images and 120 classes.

From the classification result, CBIR then will classify the query image. After the class of query image is determined, the color histogram of the query image and the images in dataset which have the same class as the query image will be extracted. The 2 features vector of query image and images dataset will be calculated using Euclidean distance. The query images used are the images of lion and bear.

Figure 3 and Figure 4 show that only the same class images as the query image are retrieved. This is because of the correct classification of the query image. If the classification of the query image classifies the wrong class, the entire result of the retrieval will show images that have a different class but have a similarity in terms of color histogram features space.

5. **Conclusions.** This research proposes a new approach for encoding method in CBIR using BoVW with SIFT features and Fisher Vector (FV). The FV encoding uses GMM in the process of visual vocabulary generation. SVM classifier is then used to classify the image first before the retrieval process is executed. The results show that FV outperforms the other encoding methods for BoVW and that chi2 kernel is a good kernel to use in SVM when using FV encoding. The accuracy of the SVM classifier is 73.3% which is good

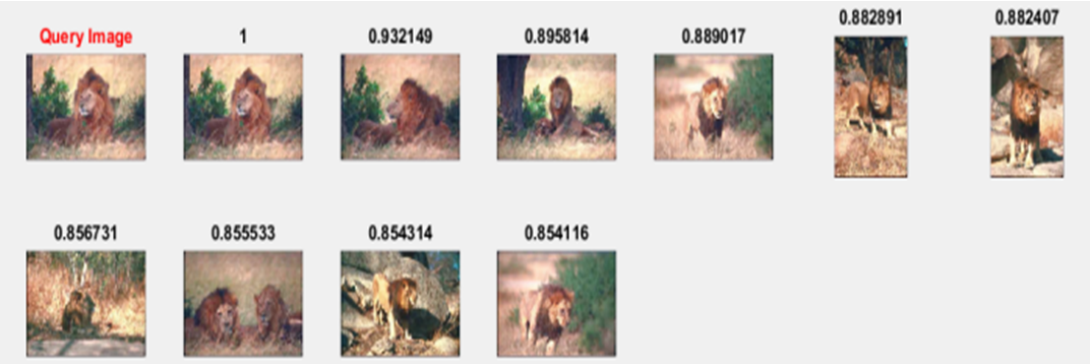

FIGURE 3. CBIR result using lion class image

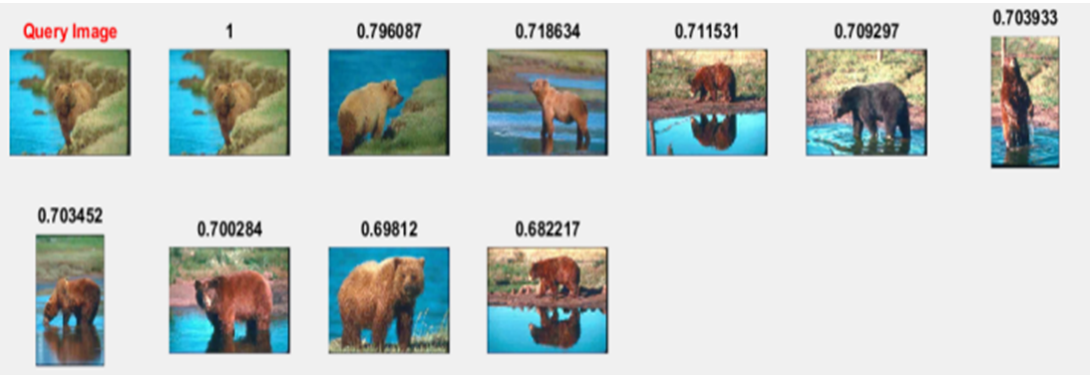

FIGURE 4. CBIR result using bear class image

accuracy to classify the query image in CBIR. The future research will be better if we can apply some other classifiers like ANN or ANFIS during training to see other classifier ability to extract data with GMM.

Acknowledgment. This work is partially supported by Bina Nusantara University. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] D. Zhang, M. M. Islam and G. Lu, A review on automatic image annotation techniques, *Pattern Recognition*, vol.45, no.1, pp.346-362, 2012.
[2] J. Liu, *Information Retrieval*, https://arxiv.org/abs/1304.5168.
[3] N. Mansoori, M. Nejati, P. Razzaghi and S. Samavi, Image retrieval by bag of visual words and color information, *The 21st Iranian Conference on Electrical Engineering (ICEE)*, Mashhad, Iran, 2013.
[4] M. Alkhawlani, M. Elmogy and H. Elbakry, Content-based image retrieval using local features descriptors and bag-of-visual words, *International Journal of Advanced Computer Science and Applications*, vol.9, no.6, pp.212-219, 2015.
[5] A. Nouman, K. B. Bajwa, R. Sablatnig, S. A. Chatzichristofis, Z. Iqbal, M. Rashid and H. A. Habib, A novel image retrieval based on visual words integration of SIFT and SURF, *PLOS ONE*, vol.11, no.6, 2016.
[6] X. Wang, L. M. Wang and Y. Qiao, A comparative study of encoding, pooling and normalization methods for action recognition, *Computer Vision – ACCV 2012*, Daejeon, Korea, 2012.
[7] D. Oneate, J. Verbeek and C. Schmid, Action and event recognition with fisher vectors on a compact feature set, *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.
[8] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman, The devil is in the details: An evaluation of recent feature encoding methods, *The 22nd British Machine Vision Conference*, Dundee, UK, 2011.
[9] J. Yue, Z. Li, L. Liu and Z. Fu, Content-based image retrieval using color and texture fused features, *Mathematical and Computer Modelling*, vol.54, pp.1121-1127, 2011.

[10] W. Yang, Z. Lu, M. Yu, M. Huang, Q. Feng and W. Chen, Content-based retrieval of focal liver lesions using bagof-visual-words representations of single- and multiphase contrast-enhanced CT images, *Journal of Digital Imaging*, vol.25, no.6, pp.708-719, 2012.
[11] D. Reynolds, Gaussian mixture models, in *Encyclopedia of Biometrics*, Springer, New York, 2009.
[12] X. Peng, L. Wang, X. Wang and Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, *Computer Vision and Image Understanding*, pp.109-125, 2016.
[13] K. Iqbal, M. O. Odetayo and A. James, Content-based image retrieval approach for biometric security using colour, texture and shape features controlled by fuzzy heuristics, *Journal of Computer and System Sciences*, vol.78, pp.1258-1277, 2012.
[14] Y.-S. Lin, J.-Y. Jiang and S.-J. Lee, A similarity measure for text classification and clustering, *IEEE Trans. Knowledge and Data Engineering*, vol.26, no.7, pp.1575-1590, 2014.
[15] D. Hoeim, Y. Chodpathumwan and Q. Dai, Diagnosing error in object detectors, *European Conference on Computer Vision*, Florence, Italy, 2012.