

# Images Similarity based on Bags of SIFT Descriptor and K-Means Clustering

**Banu Wirawan Yohanes**

Department of Computer Engineering,  
Faculty of Electronic and Computer Engineering,  
Universitas Kristen Satya Wacana, Salatiga  
banu.yohanes@uksw.edu

## **Abstract**

The content based image retrieval is developed and receives many attention from computer vision, supported by the ubiquity of Internet and digital devices. Bag-of-words method from text-based image retrieval trains images local features to build visual vocabulary. These visual words are used to represent local features, then quantized before clustering into number of bags. Here, the scale invariant feature transform descriptor is used as local features of images that will be compared each other to find their similarity. It is robust to clutter and partial visibility compared to global feature. The main goal of this research is to build visual vocabulary dictionary and use it to measure image similarity across two small image datasets. K-means clustering algorithm is used to find the centroid of each cluster at different  $k$  values. From experiment results, the bag-of-keypoints method has potential to be implemented in the content based information retrieval.

**Keywords:** scale invariant feature transform, k-means clustering, image similarity

## **1. Introduction**

In recent years, there have been ubiquity of Internet and digital devices. One of the popular media at web, instant messaging, and social network is an image. Thus emerging a new challenge on how to retrieve image precisely and quickly. Started with indexing meta data information, such as image titles and annotation tags. However, textual information may not describe image content accurately. Then, content based image retrieval (CBIR) is developed and receives many attention from computer vision communities.

CBIR retrieves images based on a query image or low-level image feature descriptors. The main concern of CBIR is how to measure images similarity [1]. At first, CBIR analyzes images color, shape, and texture features. Images visual features can be classified into two main classes, which are global features and local features. For instance, image color histogram, shape feature, and texture distribution are global features that represent a whole image and affected by several image transformations, such as translation and scaling. On the other hand, local features represent orientation, position, or scale of points or image regions.

Recently, several researches borrow Bag-of-Words (BoW) method from text-based image retrieval to adapt for scalability issues [2]. Similar to term frequency-inverse document (Tf-Idf) method, BoW trains images local features to build visual vocabulary.

These visual words are used to represent local features, then quantized before clustering into number of bags.

Zhu et al. [3] argued that small bounding boxes within image were quantized to produce more semantics results compared to color and texture based approaches. However, their approach did not consider any invariance properties at all. Later, the image patches might be clustered based on their invariant descriptors for solving texture classification problems [4].

Here, the scale invariant feature transform (SIFT) descriptor is used as local features of images that will be compared each other to find their similarity. The reasons are local feature is more robust to clutter and partial visibility compared to global feature. Specifically, SIFT is not only scale and affine invariant, but also resistant to geometric and illumination variation. The descriptors effectively characterized both objects within images and entire images. Other alternatives would be Gradient Location and Orientation Histogram (GLOH) and Speed Up Robust Feature (SURF). SIFT and SURF integration in CBIR was proposed to combine advantages of both descriptors, with scale and rotation, and illumination invariance respectively [5]. Tian et al. [6] proposed Edge Oriented Difference Histogram (EODH) as the rotation and scale invariant descriptor. The vector sum and steerable filter were applied to obtain each pixel's orientation. A weighted word distribution was obtained by applying of color SIFT and EODH integration.

The main contribution of this research is to analyze that image retrieval performance is affected by the number of image features cluster. Then finding such fixed number of clusters was not a trivial task. In order to achieve the good trade-off between images discrimination and number of clusters, a simple clustering algorithm named k-means is preferred. K-means is easy to use in spite of large computational cost for finding the nearest neighbor repeatedly.

A heuristic may be applied for solving k-means, and here Lloyd's algorithm [7] is preferred. Given any set of  $k$  center  $Z$ , for each center  $z \in Z$ , let  $V(z)$  define its neighborhood, that is a set of data in which  $z$  is the nearest neighbor. In geometric terminology,  $V(z)$  is located in the Voronoi cell of  $z$ . Each iteration of Lloyd's algorithm moves center point  $z$  to the centroid of  $V(z)$ , then updates  $V(z)$  by re-calculate distance from each data point to its nearest center. These steps are repeated, until some convergence condition is met.

## 2. Methodology

There are three basic questions in CBIR. First, how to describe image content using a set visual features? Second, how to transform variety length feature sets into fixed-length vector? And third, how to compute similarity between a pair of fixed-length vectors efficiently? This research emphasizes on a viable solution for the third question, that is to calculate image similarity in an efficient manner without losing much accuracy.

The quantized feature vectors extracted from images is called keypoints [2], as it came from keywords in text categorization. A common CBIR system consists of several phases, that are to detect and describe image patches in training dataset, to build vocabulary which is a set of cluster center of keypoints, to extract bag of keypoints for the visual vocabulary, to train a classifier using bag of keypoints, and to report image similarity measurement for the dataset.

The main goal of this research is to build visual vocabulary dictionary and use it to measure image similarity across two small image datasets. The research methodology used are as simple as well to cope with the purpose to introduce that a method in text information processing may be applied for a computer vision scheme. They are SIFT descriptors extraction as the image feature vector, build visual vocabulary as visual codewords generated from image feature vector, and image similarity measurement.

### 2.1. SIFT Descriptors Extraction

Harris affine detector [7] is preferred to provide stability of keypoint's location, inspite of two observations, that are real world objects have diverse three-dimensional (3D) structure and increasing feature invariance may reduce discriminative information.

Harris affine points are detected iteratively. The interest points location and scale are described both as local maxima of an adapted Harris function scale, and as local extrema in Laplacian operator scale. Then, affine neighborhood is determined by the eigenvalues of image's second moment matrix. Those selections are iterated until some converge points achieved within limited number of iteration. Karakasis et al. [8] proposed to use affine moment invariants as descriptors for CBIR. In contrast, the affine moment invariants were extracted with the help of the SURF detector.

SIFT descriptors are extracted from a circular region, as the mapping result from Harris affine regions. Those descriptors are Gaussian derivatives computed at 8 orientation planes over  $4 \times 4$  spatial grid locations, hence resulting 128-dimension vector. For instance, a SIFT descriptor composition within an image area is shown in Fig. 1. In other words, SIFT descriptors are multi-image representations of an image neighborhood. SIFT is preferred because of its simplicity and more discriminative representations over either steered Gaussian derivatives or differential invariants of a local jet [9]. Recently, a simpler version of SIFT named dense-SIFT (D-SIFT) was introduced. D-SIFT excluded keypoint detection and extracts features only for some fixed intervals and space [10]. However, this research conducted on small number of images only to show the potential of an automated categorization system.

### 2.2. Build Dictionary of Visual Vocabulary

Rather than comparing each query descriptor to all training descriptors extracted in previous phase, it is more efficient to identify a small number of vast clusters. Those clusters are able to differentiate given classes, with each cluster represent a visual word. Then, a histogram of visual words occurrence is constructed. Giveki et al. argued that SIFT descriptors are not able to describe scenes and landscape images well [11]. This research is also trying to support this statement.

K-means clustering algorithm is utilized to identify the center set of SIFT descriptors or called bag-of-keypoints within training image dataset. K-means is chosen since it is a trivia square-error partitioning method, instead of its locality result. In consequence, this research is to find a relation between number of representative vectors,  $k$ , with the accuracy of measuring images similarity across the dataset. In this phase, k-means is run several times with various number of  $k$ . To improve image retrieval speed, specifically in clustering phase, Kieu et al. utilized a hierarchical tree for k-means [12]. Liu et al. [13] organized local features into many groups by applying k-means. A compact descriptor was selected to describe each group's visual information. In consequence, it reduced complexity for a large-scale image search.

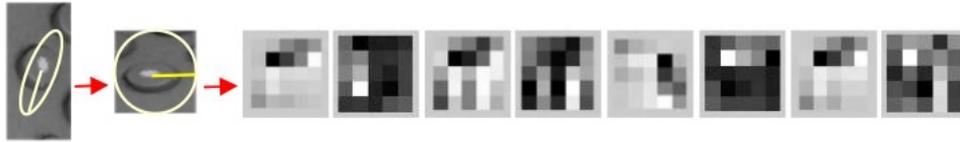


Figure 1. (from left to right) An example of Harris affine region, normalized region, and 8 maps of gradient magnitude describing SIFT descriptor [2]

### 2.3. Image Similarity Measurement

Image similarity measurement is the key operation in CBIR. With regards to vast image database and fast query response, it is necessary to make visual features as compact as possible while maintain enough discriminative information. For each pair of image within dataset, calculate the Chi-square distance between their histogram to represent image similarity. Two exactly same images will result 0, on the other hand different images will result greater than 0. Intuitively, more larger than 0 result means that the two images are far more different each other. Euclidean distance was used as similarity metric in similar scenario [14].

## 3. Experiment Result and Analysis

The methods explained in previous section were implemented using OpenCV C++ code which might take either one single image or multiple images altogether. Whenever a single image was given, the program would convert it into grey-scale image and rescale it into VGA size, 480 rows x 600 columns, to reduce computation time. Note that the image ratio is kept originally. Inter area interpolation was used to resampling images using pixel area relation. Then SIFT descriptors were extracted from Y image component. For each detected keypoint, there would be marked '+' at the location and drawn a circle around it whose radius is proportional to the keypoint scale. At last, both input image and highlighted image were displayed side-by-side. The algorithm flowcharts, for both the main program and the build or load dictionary procedure, are displayed in Fig. 2 and Fig. 3 respectively.

Two parameters set for SIFT descriptor are `contrastThreshold = 0.1` and `edgeThreshold = 1.0`. The `contrastThreshold` was used to filter out weak features in low contrast regions. The larger the `contrastThreshold`, the less features are produced by the detector. While, the `edgeThreshold` was used to filter out edge-like features, the larger the `edgeThreshold`, the less features are filtered out, or in other words more features are retained. Other parameters such as number of features to retain, number of layers in each octave, and sigma of Gaussian were set default.

Besides the program was utilized `BOWKMeansTrainer` to train the dataset. The `BOWKMeansTrainer` class took dictionary size,  $K$ , as the number of cluster. Termination criteria were set for the algorithm, that are the maximum number of iterations = 100, the desired accuracy = 0.001, retries or number of attempt = 3, and a flag noted `kmeans++` center initialization. Then, it clustered the feature vectors into each dictionary. Finally, the program would create 3 dictionaries in YAML format, one for each value of  $K = 5\%$ ,  $10\%$ , and  $20\%$  of total keypoints extracted. The dictionaries are stored locally to be reused in the future.

If the input of program is a single image, then to display SIFT keypoints extracted using `drawKeypoints` with a proportional radius circle surrounding each keypoint. The

parameters used are set. Then, both the original input image and the processed image with SIFT keypoints embedded are displayed together in a window by merging two matrices into one big size matrix sufficient to contain both images and the frame as shown in Fig. 4.

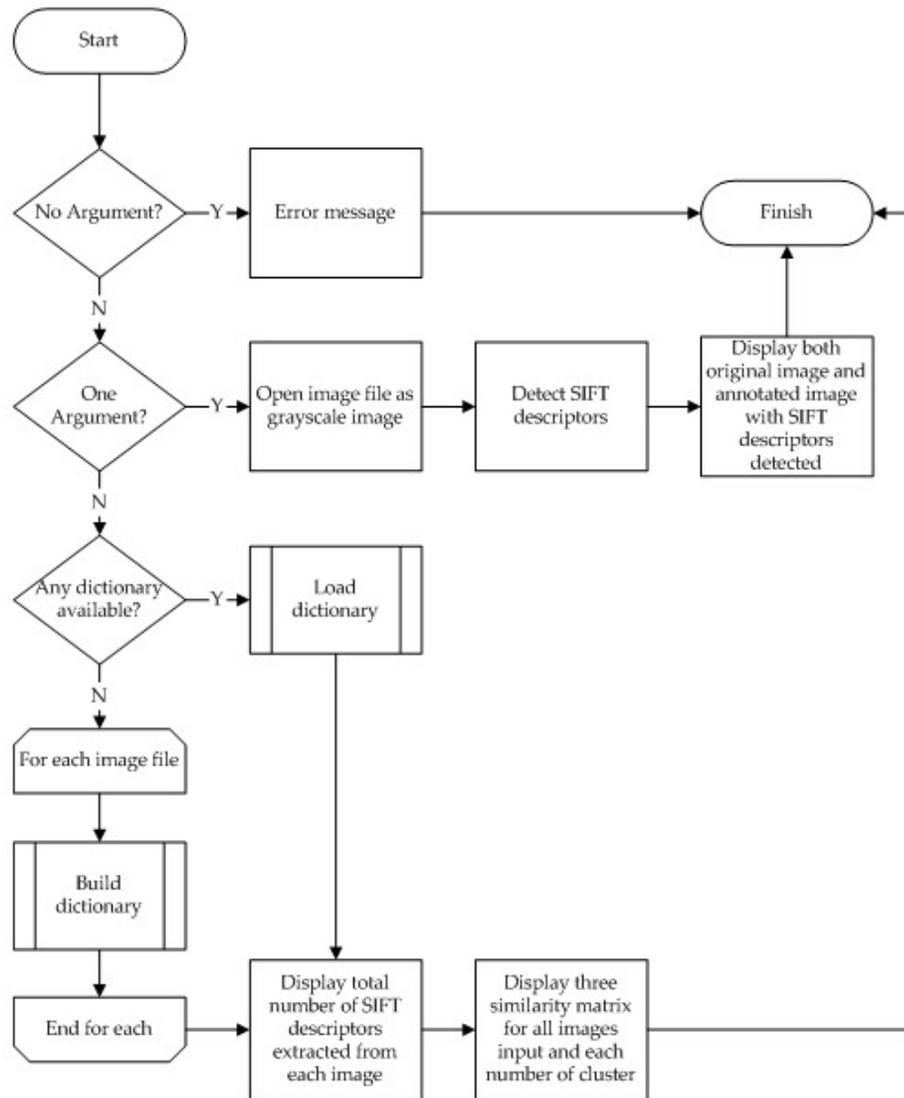


Figure 2. Main program flowchart to receive either a single image file or several image files concurrently

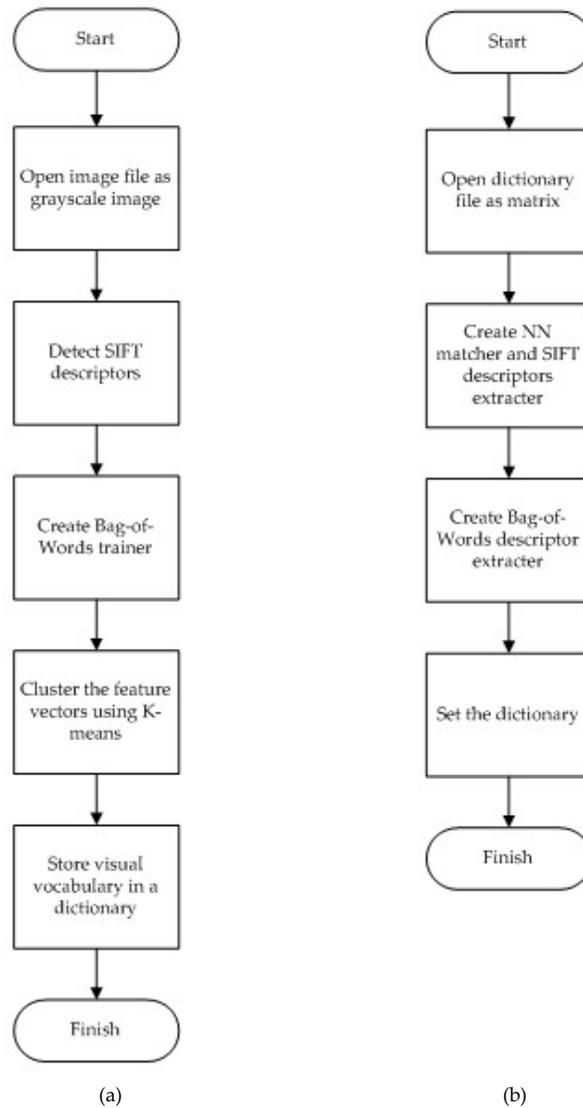


Figure 3. (a) Build dictionary procedure, and (b) load dictionary procedure flowcharts



Figure 4. Keypoints from a cruiser ship image (top), and keypoints from a lighthouse image (bottom)

Whenever multiple images are fed into the program, it will compare each pair of images using bag-of-keypoints model constructed from SIFT descriptors. Similar like one single image processing, each image is converted into grey-scale image and rescaled into VGA size, 480 rows x 600 columns. Inter area interpolation is used to resampling images using pixel area relation. SIFT descriptors are well extracted from Y image component. Then cluster the keypoints from all images using k-means algorithm whose  $K$  is determined to be 5%, 10%, and 20% of the total number of keypoints extracted from all images within dataset. For each image, construct a histogram of its visual words by counting number of visual words occurrence within it. At last, the program will calculate images similarity measurement for each pair of images within dataset using Chi-square distance.

The program loads a nearest neighbor matcher using `FlannBasedMatcher` interface from Fast Approximate Nearest Neighbor Search Library (FLANN) to perform a quick and an efficient image matching. It also utilizes SIFT feature point extractor and SIFT descriptor extractor with same parameters value as the program module for one single image. Then, it computes image descriptor using bag-of-keypoints implemented in `BOWImgDescriptorExtractor` class. First, it computes keypoints sets of given images, then find the nearest visual words from the vocabulary which loaded from a visual vocabulary, YAML files within current directory, before composing bag-of-keypoints in a normalized histogram. At last, the program uses chi-square distance to measure dissimilarity between any pair of images. Whenever the value of Chi-square distance nearer to zero, it means that a pair of images is similar each other.

Two small image datasets were used in experiments, scene and landscape pictures, and vocational school student face images while they were on the job training in the Faculty of Electronic and Computer Engineering, UKSW. Sample images from both datasets are displayed in Fig. 5. There are 10 classes in the scene and landscape datasets, such as cruiser ship, lighthouse, kids playground, and campus library. Each class consists of two or more images from different scenes and point of views. Total number of detected keypoints is 6843. There are 9 classes in the student facial datasets, with all girls wearing hijab and either little bit looked to the left, right, or down for approximately at 5-10 degrees. Each class consists of three images with different pose. Total number of detected keypoints is 1079. The experiment results are presented in Table 1 and Table 2, in the scene and landscape datasets and student facial datasets, respectively.

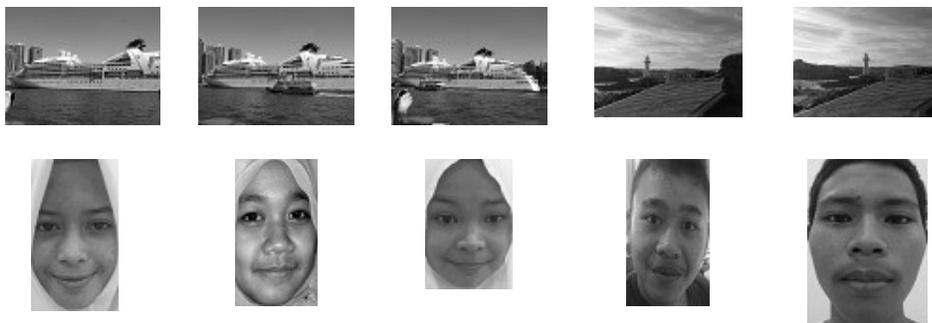


Figure 5. Sample images from scene and landscape (first row), and student facial (second row) datasets

Table 1. Scene and landscape dataset similarity results

Total number of detected keypoints: 6843									
K = 5% * 6843 = 342									
Dissimilarity matrix:									
	img01.jpg	img02.jpg	img03.jpg	img04.jpg	img05.jpg	img06.jpg	img07.jpg	img08.jpg	img09.jpg
img01.jpg	0	0.59	1.2	1.1	1.2	1.2	0.95	1.1	1.5
img02.jpg	0.59	0	1.1	1.1	1.5	1.7	1.1	1.1	2.8
img03.jpg	1.2	1.1	0	0.66	0.58	0.85	0.78	0.69	2.7
img04.jpg	1.1	1.1	0.66	0	0.77	1	0.87	0.85	2.7
img05.jpg	1.2	1.5	0.58	0.77	0	0.66	0.59	0.56	1.5
img06.jpg	1.2	1.7	0.85	1	0.66	0	0.51	0.56	1.7
img07.jpg	0.95	1.1	0.78	0.87	0.59	0.51	0	0.48	2.7
img08.jpg	1.1	1.1	0.69	0.85	0.56	0.56	0.48	0	2.2
img09.jpg	1.5	2.8	2.7	2.7	1.5	1.7	2.7	2.2	0

K = 10% * 6843 = 684									
Dissimilarity matrix:									
	img01.jpg	img02.jpg	img03.jpg	img04.jpg	img05.jpg	img06.jpg	img07.jpg	img08.jpg	img09.jpg
img01.jpg	0	0.49	0.92	0.8	1.1	0.92	0.9	0.89	1.9
img02.jpg	0.49	0	0.81	0.77	1.1	0.98	1	0.96	2.1
img03.jpg	0.92	0.81	0	0.54	0.74	0.74	0.71	0.81	2.3
img04.jpg	0.8	0.77	0.54	0	0.62	0.91	0.95	1.2	2.2
img05.jpg	1.1	1.1	0.74	0.62	0	0.68	0.76	0.82	1.7
img06.jpg	0.92	0.98	0.74	0.91	0.68	0	0.54	0.66	1.2
img07.jpg	0.9	1	0.71	0.95	0.76	0.54	0	0.56	1.9
img08.jpg	0.89	0.96	0.81	1.2	0.82	0.66	0.56	0	1.9
img09.jpg	1.9	2.1	2.3	2.2	1.7	1.2	1.9	1.9	0

K = 20% * 6843 = 1368									
Dissimilarity matrix:									
	img01.jpg	img02.jpg	img03.jpg	img04.jpg	img05.jpg	img06.jpg	img07.jpg	img08.jpg	img09.jpg
img01.jpg	0	0.52	0.89	0.82	0.9	0.92	0.82	0.86	1.6
img02.jpg	0.52	0	0.97	0.81	0.96	0.91	0.84	0.96	1.1
img03.jpg	0.89	0.97	0	0.67	0.91	0.84	0.87	0.91	1.5
img04.jpg	0.82	0.81	0.67	0	0.81	0.81	0.76	0.85	1.6
img05.jpg	0.9	0.96	0.91	0.81	0	0.57	0.69	0.81	1.1
img06.jpg	0.92	0.91	0.84	0.81	0.57	0	0.55	0.66	1
img07.jpg	0.82	0.84	0.87	0.76	0.69	0.55	0	0.65	1.6
img08.jpg	0.86	0.96	0.91	0.85	0.81	0.66	0.65	0	1.2
img09.jpg	1.6	1.1	1.5	1.6	1.1	1	1.6	1.2	0

It can be seen from Table 1 and Table 2 with various  $K$  value, that for each  $K$  value there were different image similarity results. Using  $K = 10\%$  would produce the best Chi-square distances for each pair of similar images, compared to both  $K = 5\%$  and  $K = 20\%$ . It was consistent with the hypothesis for k-means algorithm, where  $k$  value does not guarantee best results for distinct experiments. However, there are a rule of thumb that define that for increasing  $K$  values, the error rate will be decreased significantly until some fixed number of  $K$ , then the error rate will be steady.

Another finding in these experiments were SIFT descriptors do not take into account for the location of objects within images. It can be seen that, different scenes and slightly facial variations were undetected by the descriptors. In contrast, SIFT is only extract keypoints from interest region defined by a region finder, such as Harris affine detector.

Table 2. Student facial image dataset similarity results

K = 5% * 1079 = 53						
Dissimilarity matrix:						
img01.JPG	img01.JPG	img02.JPG	img03.JPG	img10.JPG	img11.JPG	img12.JPG
	0	0.6078	0.6666	0.5843	0.6982	0.5278
img02.JPG	0.6078	0	0.9393	0.606	0.5792	0.4587
img03.JPG	0.6666	0.9393	0	0.6905	0.6911	0.5718
img10.JPG	0.5843	0.606	0.6905	0	0.7768	0.7156
img11.JPG	0.6982	0.5792	0.6911	0.7768	0	0.7678
img12.JPG	0.5278	0.4587	0.5718	0.7156	0.7678	0
K = 10% * 1079 = 107						
Dissimilarity matrix:						
img01.JPG	img01.JPG	img02.JPG	img03.JPG	img10.JPG	img11.JPG	img12.JPG
	0	0.4777	0.6155	0.5948	0.6905	0.5479
img02.JPG	0.4777	0	0.9188	0.569	0.5806	0.6125
img03.JPG	0.6155	0.9188	0	0.5629	0.5734	0.65
img10.JPG	0.5948	0.569	0.5629	0	0.7684	0.8549
img11.JPG	0.6905	0.5806	0.5734	0.7684	0	0.7509
img12.JPG	0.5479	0.6125	0.65	0.8549	0.7509	0
K = 20% * 1079 = 215						
Dissimilarity matrix:						
img01.JPG	img01.JPG	img02.JPG	img03.JPG	img10.JPG	img11.JPG	img12.JPG
	0	0.5438	0.7354	0.7012	0.6381	0.6115
img02.JPG	0.5438	0	0.8922	0.7054	0.6894	0.6176
img03.JPG	0.7354	0.8922	0	0.6375	0.687	0.6431
img10.JPG	0.7012	0.7054	0.6375	0	0.6233	0.6273
img11.JPG	0.6381	0.6894	0.687	0.6233	0	0.8047
img12.JPG	0.6115	0.6176	0.6431	0.6273	0.8047	0

## 4. Conclusion

From experiment results, it can be concluded that bag-of-keypoints method borrowed from text information retrieval has potential to be implemented in CBIR. Particularly, when combined with a clustering algorithm to provide image similarity measurement tool. Moreover, this research provide benchmarking to guide for k-means implementation with regard to define  $K$  value in other datasets.

## Acknowledgement

This research was conducted as an extension from an UKSW internal research grant in 2017. Thank you to the Vice Rector V office for supporting this research.

## References

- [1] W. Zhou, H. Li, and Q. Tian, Recent Advance in Content-based Image Retrieval: A Literature Survey, arXiv preprint arXiv:1706.06064, 2017. [Online], <https://arxiv.org/pdf/1706.06064>, diakses tanggal 19 Juli 2018.
- [2] G. Csurka , C. R. Dance , L. Fan , J. Willamowski, and C. Bray, Visual categorization with bags of keypoints, In Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, pp. 1-22, 2004.
- [3] L. Zhu, A. Rao and A. Zhang, Theory of Keyblock-based image retrieval, ACM Transactions on Information Systems, vol. 20, no. 2, pp. 224-257, 2002.
- [4] M. Varma and A. Zisserman, Classifying materials from images: to cluster or not to cluster?, European Conference on Computer Vision, 2002.
- [5] N. Ali, et al., A Novel Image Retrieval Based on Visual Words Integration of SIFT and SURF, PLoS ONE, Vol. 11, No. 6, 2016.

- [6] X. Tian, L. Jiao, X. Liu, X. Zhang, Feature integration of EODH and Color-SIFT: Application to image retrieval based on codebook, *Signal Processing: Image Communication*, Vol. 29, No. 4, pp. 530–545, 2014.
- [7] K. Mikolajczyk and C. Schmid, An affine invariant interest point detector, *European Conference on Computer Vision*, 2002.
- [8] E. Karakasis, A. Amanatiadis, A. Gasteratos, S. Chatzichristofis, Image moment invariants as local features for content based image retrieval using the Bag-of-Visual-Words model, *Pattern Recognition Letters*, Vol. 55, pp. 22-27, 2015.
- [9] K. Mikolajczyk and C. Schmid, A performance evaluation of local descriptors, *Conference on Computer Vision and Pattern Recognition*, 2003.
- [10] X. Lou, D. Huang, L. Fan, and A. Xu, An Image Classification Algorithm Based on Bag of Visual Words and Multi-kernel Learning, *Journal of Multimedia*, Vol. 9, No. 2, 2014.
- [11] D. Giveki, M. A. Soltanshahi, F. Shiri, H. Tarrah, A New SIFT-Based Image Descriptor Applicable for Content Based Image Retrieval, *Journal of Computer and Communications*, Vol. 3, pp. 66-73, 2015.
- [12] M. Kieu, K. D. Lai, T. D. Tran, and T. H. Le, A Fusion of Bag of Word Model and Hierarchical K-Means++ in Image Retrieval, *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, LNCS, Vol. 9978, 2016.
- [13] Z. Liu, H. Li, W. Zhou, R. Hong, Q. Tian, Uniting Keypoints: Local Visual Information Fusion for Large-Scale Image Search, *IEEE Transactions on Multimedia*, Vol. 17, No. 4, pp. 538–548, 2015.
- [14] P. Górecki, K. Sopyła, and P. Drozda, Ranking by K-Means Voting Algorithm for Similar Image Retrieval, *International Conference on Artificial Intelligence and Soft Computing*, LNCS, Vol. 7267, pp. 509-517, 2012.