Performance Enhancement of Image Clustering Using Singular Value Decomposition in Color Histogram Content-Based Image Retrieval

Catur Supriyanto, Guruh Fajar Shidik, Ricardus Anggi Pramunendar, and Pulung Nurtantio Andono

***Abstract*—This paper presents an enhancement of the performance of image clustering. K-Means has been chosen as our clustering technique. We applied Hue Saturation Value (HSV) color histogram as features to retrieve image information. Singular Value Decomposition (SVD) technique is employed to enhance the performance of image clustering by reducing features that not useful to be proceed in clustering process. We evaluated the image clustering using Recall, Precision, F-Measure and computational time. The obtained result indicates that SVD in HSV color histogram Content-Based Image Retrieval (CBIR) is promising.**

***Index Terms*—Content based image retrieval, HSV color histogram, k-means algorithm, singular value decomposition**

I. Introduction

Currently the increasing of image data on internet is giving opportunity for researcher to working in searching and classifying the content of image. The example of image management tool that capable to search and classifying image is Google image as image searching application usually used for image mining and image search in web context. The importance of searching image information in web context given many research communities has produced many method algorithms with tools for image retrieval and clustering, where this technique also include in technique Content-Based Image Retrieval (CBIR).

There are many applications of CBIR such as biomedicine, military, commerce, education, web image classification and searching [1]. CBIR is a technique to searching by analyzes the actual content (feature) of image. In CBIR there are two techniques that should strong to achieve accurate results; there are a technique to retrieve information and a technique to cluster for classifying image.

Image retrieval is a process to retrieve image features that include in each image. In [2] has stated two types of images features there are low level feature and high level feature, where high level feature is difficult to extract like emotion or other human behavior activities. There are several image features at low level feature usually used to retrieve image information such as feature color, shape and texture. In this study we only used color feature to retrieve the information of image based on Hue Saturation Value (HSV) space format.

Manuscript received July 24, 2012; revised September 2, 2012.

The authors are with the Dept. of Computer Science Dian Nuswantoro University, Semarang, Indonesia (e-mail: caturs@research.dinus.ac.id, guruh.fajar@research.dinus.ac.id, ricardus.anggi@research.dinus.ac.id, pulung@research.dinus.ac.id).

This research applied K-Means as clustering technique, where this method commonly used for partitioning [3], [4]. Each cluster in K-Means method is represented by its centroids or the mean value of all data in the cluster.

We applied color histogram technique that usually used to retrieve information image. The weakness applicable of this features extraction for information retrieval has been state by [5] where the implementation of color histogram does not give relevant image as seen by an algorithm with human visual. The aim of this paper is to adapt Singular Value Decomposition (SVD) as feature transformation in CBIR. SVD is a Latent Semantic Analysis (LSA) approach. The reason of using SVD is provide useful information of the color.

CBIR system can be improved by using image clustering [6]. The problem of image clustering is high number of image in clustering leads to more computational time [7]. To overcome this problem, SVD also can be proposed to reduce the computational time of image clustering.

The outline of this paper is as follows: section 2 describes the related work. Section 3 describes the methodology of research. Section 4 describes the dataset and shows the performance analysis of proposed approach. Section 5 presents the conclusion and future work.

II. Related Work

There are several research has been conduct to improve the performance of CBIR in image retrieval clustering area. The previous work in [8]-[10] have been analyzed the performance of clustering algorithm for image retrieval. Kucuktunc and Zamalieva [11] proposed fuzzy color histogram for CBIR. Mamdani fuzzy inference system was used as fuzzy technique to link L*a*b* to fuzzy color space. Their work show that fuzzy color histogram performed better than conventional methods.

Other study in [12] was compare the using of Conventional Color Histogram (CCH), Invariant Color Histogram (ICH) and Fuzzy Color Histogram (FCH) of an images in CBIR system. ICH and FCH has been used to address the problem of rotation, translation and spatial relationship of ICH.

Tonge [13] proposed a K-means clustering algorithm to grouping the collection of images. Image is clustered based on the query image. Color was used to be features in their CBIR system.

Sakthivel *et al*. [14] proposed modified K-Means clustering to group similar pixel in CBIR. Their purpose is to improve retrieval performance by capturing the regions and

also to provide a better similarity distance computation.

III. Methodology

Fig.1 depicts the general overview of our proposed image clustering. First is the collection of images. Then, extract the feature by using HSV color histogram. Next, reduce the dimensionality of feature by using Singular Value Decomposition. Finally, cluster the image collection using K-Means algorithm. Detailed description of each phase is explained in the next subsection.

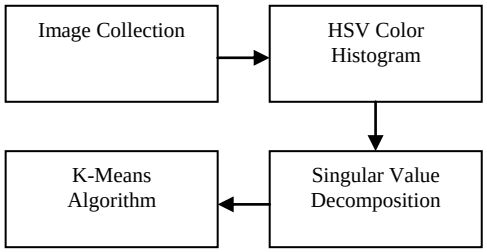

Fig. 1. Proposed image clustering system

A. HSV Color Histogram

Hue Saturation Value (HSV) has been chosen as color histogram feature, since HSV color space gives better result for CBIR [2]. HSV is often used because of its accordance with human visual feature [15]. The conversion of Red Green Blue (RGB) space into HSV space can be seen in formulae (1), (2), and (3).

$$H = \cos^{-1}\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right\} \quad (1)$$

$$S = 1-\frac{3}{R+G+B}[\min(R,G,B)] \quad (2)$$

$$V = \frac{1}{3}(R+G+B) \quad (3)$$

B. Singular Value Decomposition

Singular Value Decomposition (SVD) is a feature transformation technique which reduces the high dimensional matrix into small dimensional matrix. Let A is the features-images matrix of size $m \times n$ where m is the number of features and n is the number of images. The singular value decomposition of features-images matrix A can be defined as (4).

$$A_{m\times n} = U_{m\times k}\,\Sigma_{k\times k}\,V^T{}_{k\times n} \quad (4)$$

where U is features vector, Σ is the diagonal matrix of singular value and V^T is images vector. Next, matrix V^T is used to cluster the images collection. The value of rank k is $k \le \min(m,n)$. Small k of SVD was enough to generate high F-Measure value [16].

C. K-Means Algorithm

The clustering technique in CBIR is required to classify number of images into several group based on its cluster that have similarity. Clustering algorithm is classified into five types there are partitional, hierarchical, density-based, grid-based, and model-based clustering. The most popular clustering technique that widely used is partition and hierarchical clustering [17]. This research applied K-Means as clustering technique that used for classifying, where this technique quite popular for this purpose. K-Means is unsupervised learning which means there is no data training in the process of clustering. K-Means clustering is carried out in four steps [18]:

1) Choose objects to be k initial seeds (basically is random)
2) Calculate the distance of each seed to the each object using distance or similarity metric; assign each object to the cluster with the nearest seed point.
3) Compute the new seed point.
4) Return to the step 2 if the current seeds are different to the previous seeds.

For distance metric, we used city block metric is defined as (5). Since, city block metric gave the best precision for content-based image retrieval [19].

$$d = \sum_{i=0}^{n} |\, x_i - y_i \,| \quad (5)$$

where d is the distance value, x_i and y_i are vector of image x and image y, respectively.

IV. Experiment

A. Dataset

This paper used 150 images from corel-princeton dataset which divided into five classes: column, flower, horse, model and sea. The dimension of images is 128×85 pixels. The sample image of each class can be seen in Fig 2.

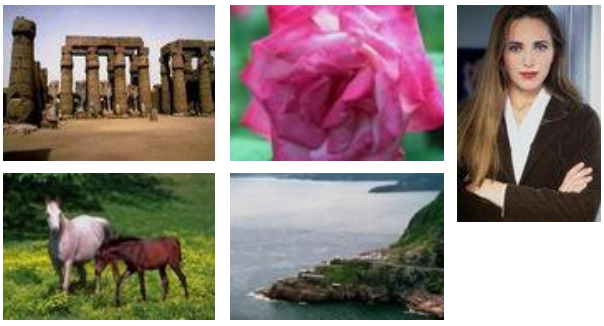

Fig. 2. Sample images of each class

B. Evaluation Measure

In order to evaluate the quality of images clustering, this paper employed F-measure as the standard evaluation measurement widely used in clustering. F-measure is the combination between recall and precision. The Precision, Recall, and F-Measure are defined as (6), (7) and (8) respectively.

$$\text{Precision} = \frac{Total \text{ number of retrieved relevant images}}{\text{Total number of retrieved images}} \quad (6)$$

$$\text{Recall}=\frac{\text{Total number of retrieved relevant images}}{\text{Total number of relevant images}} \quad (7)$$

$$\text{F-Measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision + Recall}} \quad (8)$$

C. Experiment Result

This paper used K-Means algorithm as the clustering algorithm. We randomly selected 5 initial seed of each class and averaged 30 times as the final performance of images clustering. MATLAB was used to perform the experiment. The experiment was executed on Intel Core i7-2600 CPU 3.4 GHz with 16 GB of memory. The performance was measured by comparing Recall, Precision, F-Measure, and computational time of proposed method to the K-Means image clustering.

In our experiment, we clustered the images on the rank of sizes of 2 to 100. Fig. 3 shows the Recall, Precision and F-Measure of our proposed image clustering system.

From Fig. 3, we can see that the using of SVD can improve the accuracy of image clustering. Respectively, for Recall, Precision, and F-Measure, the best performance of 0.5869, -0.1275, and 0.2094 are obtained when the number of rank *k* SVD is 5. The F-Measure can improve to be 25.94% compare to the F-Measure of K-Means 0.1663. Detailed improvement of Recall, Precision and F-Measure can be seen in the Table I. Respectively, the Recall, Precision, and F-Measure decrease to 0.4278, 0.0962, and 0.1570 for rank SVD *k*=12.

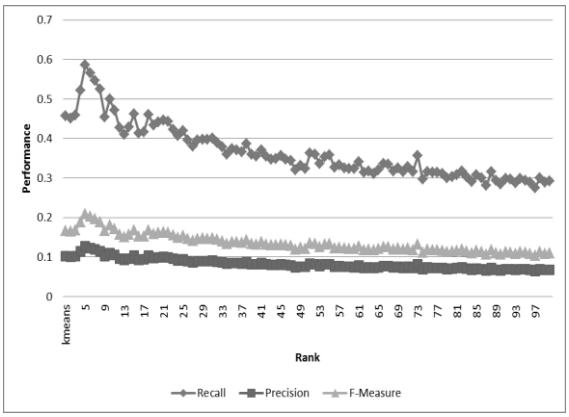

Fig. 3. Accuracy of proposed image clustering system

TABLE I: THE IMPROVEMENT OF USING SVD

	Recall	Precision	F-Measure
K-Means	0.4567	0.1016	0.1663
K-Means in rank 5 SVD	0.5869	0.1275	0.2094
Improvement	28.51%	25.47%	25.94%

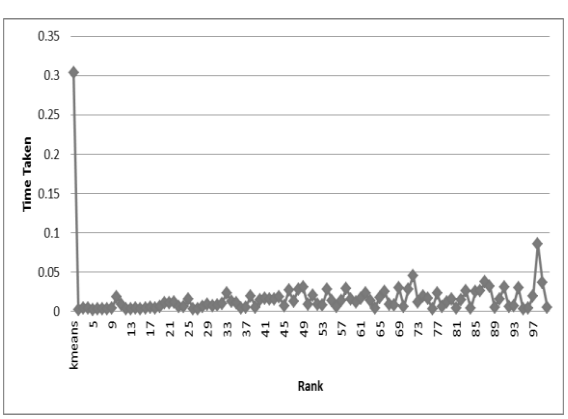

Fig. 4. Time taken of proposed image clustering system

In order to the time computational, SVD can minimize the time taken of image clustering. From the Fig. 4, SVD gave the best computational time with the rank of SVD k=4 which the computational time can decrease into 0.002 second. This achievement was compared to the computational time of K-Means (0.3031 second).

V. CONCLUSION

The objective of this paper is improves the performance of color histogram feature in image clustering system. In this work, Singular Value Decomposition (SVD) has been proposed for color histogram image clustering. The computation of image clustering is faster by performing SVD. SVD is not only reduces the computational of image clustering, but also SVD improve the accuracy of image clustering in the small rank of SVD. High number of rank SVD can decrease the accuracy of image clustering.

The content based image retrieval in this paper use only HSV color feature. Future work will be applied on another feature such as texture and shape feature. The expectation is improve the accuracy of image clustering.

REFERENCES

[1] M. Braveen and P. Dhavachelvan, "Evaluation of content based image retrieval based on color feature," *International Journal of Recent Trend in Engineering*, vol. 1, no. 2, May 2009.

[2] G. Deshpande and M. Borse, "Image retrieval with the use of different color spaces and texture feature," in *Proc. International Conference on Software and Computer Application*, 2012, pp. 273-277.

[3] E. Forgy, "Cluster analysis of multivariate data: Efficiency vs interpretability of classification," in *Biometrics*, 1965, pp. 768-780.

[4] J. MacQueen, "Some method for classification and analysis of multivariate observation," in *barkeley symposium*, 1967, pp. 281-297..

[5] R. Chakravarti and X. Meng, "A study of color histogram based image retrieval," in *Proc. 6th International Conference on Information Technology: New Generation*, 2009, pp. 1323-1328.

[6] D. Kinoshenko, V. Mashtalir, and E. Yegorova, "Clustering method for fast content-based image retrieval," *Computer Vision and Graphics*, 32, Mar. 2006, pp. 946-952.

[7] H. Liu and X. Yu "Application research of k-means clustering algorithm in image retrieval system," in *Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)*, 2009.

[8] G. Liu and B. Lee, "A color-based clustering approach for web image search results," in *Proc. 2009 International Conference on Hybrid information Technology (ICHIT '09)*, Daejeon, Korea, 2009, vol. 321, pp. 481-484.

[9] K. Jarrah, S. Krishnan, and L. Guan, "Automatic content-based image retrieval using hierarchical clustering algorithms," in *Proc. International Joint Conference on Neural Ne tworks (IJCNN '06)*, Oct. 2006, Vancouver, BC pp. 3532 - 3537.

[10] P. J. Dutta, D. K. Bhattacharyya, J. K. Kalita, and M. Dutta, "Clustering approach to content based image retrieval," in *Proc. Conference on Geometric Modeling and Imaging: New Trends (GMAI)*, 2006, pp. 183-188.

[11] O. Kucuktunc and D. Zamalieva, "Fuzzy color histogram-based CBIR system," In *Proc. 1st Int'l Fuzzy Systems Symp. (FUZZYSS'09)*, Ankara, Turkey, 2009.

[12] P. S. Suhasini, K. S. R. Krishna, and I. V. M. Krishna, "CBIR using color histogram processing," *Journal of Theoretical and Applied Information Technology*, vol. 6, no. 1, pp. 116-122.

[13] V. G. Tonge, "Content Based Image Retrieval by K-Means Clustering Algorithm," *International Journal of Engineering Science and Technology (IJEST)*, pp. 46-49, Feb 2011.

[14] K. Sakthivel, T. Ravichandran, and C. Kavitha, "Performance enhancement in image retrieval using modified k-means clustering algorithm," *Journal of Mathematics and Technology*, pp. 78-85, Feb 2010.

[15] Z. Zhang, W. Li, and B. Li, "An improving technique of color histogram in segmentation-based image retrieval," in *Proc. Fifth International Conference on Information Assurance and Security*, 2009.
[16] H. Kim and H. Park, "Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations using a priori knowledge of gene relationships," *ACM First International Workshop on Text Mining in Bioinformatics (TMBio)*, 2006.
[17] R. Xu and D. Wunsch, "Survey of clustering algorithm," *IEEE Transaction on Neural Network*, vol. 16, pp. 645-678, May 2005.
[18] M. H. Dunham, *Data mining introductory and advanced concepts*. Pearson Education, 2006.
[19] A. Vadivel, A. K. Majumdar, and S. Sural, "Performance comparison of distance metrics in content-based Image retrieval applications," in *Proc. of Int'l. Conf. on Information Technology*, Bhubaneswar, India, 2003, pp. 159–164.

Catur Supriyanto has received Bachelor of Information Technology from Dian Nuswantoro University, Semarang, Indonesia in 2009, M.CS (Master of Computer Science) from Universiti Teknikal Malaysia Melaka (UTeM) in 2011. Currently, he is a lecturer in Dian Nuswantoro University. He has guided several Master of Computer students and he has published 2 international papers. His area of interest is information retrieval, machine learning and content-based image retrieval.

Guruh Fajar Shidik has received Bachelor of Information Technology from Dian Nuswantoro University, Semarang, Indonesia in 2009, M.CS (Master of Computer Science) from Universiti Teknikal Malaysia Melaka (UTeM) in 2011. Currently, he is a lecturer in Dian Nuswantoro University. He has guided several Master of Computer students. His area of interest is computer vision, machine learning and content-based image retrieval.

Ricardus Anggi Pramunendar has received Bachelor of Information Technology from Dian Nuswantoro University, Semarang, Indonesia in 2009, M.CS (Master of Computer Science) from Universiti Teknikal Malaysia Melaka (UTeM) in 2011. Currently, he is a lecturer in Dian Nuswantoro University. He has guided several Master of Computer students and he has published 1 international paper. His research interests have included the image processing, machine learning and computer vision.

Pulung Nurtantio Andono has received Bachelor of Engineering from Trisakti University, Jakarta, Indonesia in 2006 and Master of Computer from Dian Nuswantoro University in 2009. Currently, he is a lecturer in Dian Nuswantoro University and he is now a Ph.D. student of Tenth of November Institute of Technology, Surabaya, Indonesia. He has published 1 international paper. His area of interest is 3D image reconstruction and computer vision.