Journal of Theoretical and Applied Information Technology
20th May 2013. Vol. 51 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN IMPROVED TECHNIQUE OF COLOR HISTOGRAM IN IMAGE CLUSTERING USING IMAGE MATTING

[1]CATUR SUPRIYANTO, [2]RURI SUKO BASUKI, [3]GURUH FAJAR SHIDIK, [4]RICARDUS ANGGI PRAMUNENDAR, [5]ABDUL SYUKUR

[1, 2,3,4,5] Department of Computer Science, Postgraduate Program, Dian Nuswantoro University, Semarang, Indonesia

E-mail: {[1]caturs, [2]rurisb, [3]guruh.fajar, [4]ricardus.anggi, [5]abdul.syukur}@research.dinus.ac.id

ABSTRACT

Image clustering is one of the applications of content based image retrieval. The purpose of image clustering is to group the object inside the images into different cluster. Using all color of images lead to reduce the accuracy of image clustering. Therefore, this research focuses on the color of the foreground in an image. Image matting has been proposed to subtract between foreground and background of an image. Then, the color histogram of foreground object was used as the feature in image clustering. By applying image matting, the accuracy of image clustering can be improved.

Keywords: *Content Based Image Retrieval, Image Clustering, Image Matting, Singular Value Decomposition*

1. 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.

This research applied K-Means as clustering algorithm, 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.

This research is based on Supriyanto *et al* [5] that proposed Singular Value Decomposition as feature reduction in color feature image clustering. The using of all color in an image may lead to reduce the accuracy of image clustering. Therefore, this research focuses on the color of the foreground inside the images. Grouping object or foreground inside the image is the purpose of image clustering.

This research proposes image matting to subtract between foreground and background. Further, the color of foreground is used as the feature of image clustering. Our image matting is based on Basuki *et al* [6] which proposed Fuzzy C-Means algorithm to produce adaptive threshold on alpha matting.

2. RELATED WORK

There are several research has been conduct to improve the performance of CBIR in image clustering area. The previous work in [7],[8],[9] have been analyzed the performance of clustering

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

algorithm for image retrieval. Kucuktunc and Zamalieva [10] 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 [11] compared the using of Conventional Color Histogram (CCH), Invariant Color Histogram (ICH) and Fuzzy Color Histogram (FCH) of an image in CBIR system. ICH and FCH has been used to address the problem of rotation, translation and spatial relationship of ICH.

Tonge [12] proposed k-means algorithm to cluster the collection of images. These images are grouped based on the query image. Color was used to be features in their CBIR system.

Sakthivel *et al*. [13] 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.

3. IMAGE CLUSTERING PROPOSED METHOD

The proposed image clustering comprises several steps as illustrated in Figure 1. First, image matting is applied to obtain foreground object. Then, perform HSV color histogram as the feature of CBIR. Next, Singular Value Decomposition is performed to overcome the sparse matrix that produced from color feature extraction. Finally, k-means algorithm groups the image collection into different cluster. Detailed explanation of each step is described in the next subsection.

3.1 Image Matting

The content of image contains of foreground, background, boundary area and noisy [6]. First of all, manual scribble needs to be performed to dataset. Scribble color (white for foreground and black for background) was used to calculate the alpha value of the closed-form method. Alpha value is generated by adaptive threshold using Fuzzy C-Means (FCM) clustering.

FCM is performed to determine the threshold between foreground and background. FCM is popular fuzzy clustering algorithm. FCM produces a membership matrix, which contains the degree of membership of a pixel to all the clusters. Therefore, FCM is soft clustering algorithm. FCM attempts to minimize the sum of square error (SSE).

$$SSE = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , 1 \le m \le \infty \quad (1)$$

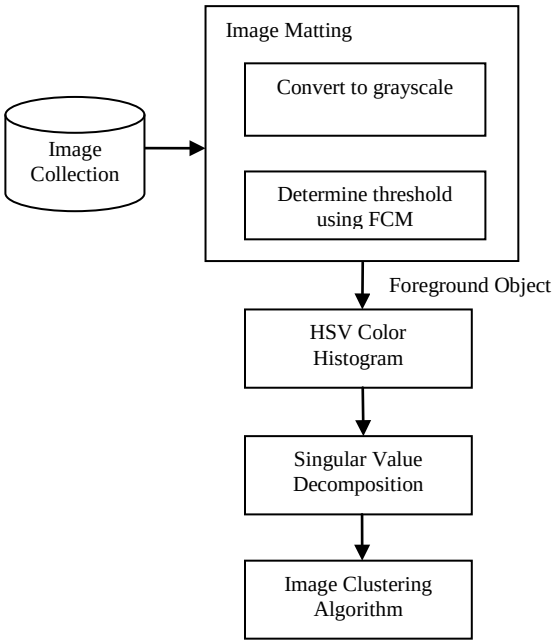

Figure 1. Proposed image clustering method

Where u_{ij} represents the membership of pixel x_i in the j^{th} cluster, c_j is the j^{th} cluster center.

$$\sum_{i=1}^{c} u_{ij} = 1, 1 \le j \le n \quad (1a)$$

$$u_{ij} \ge 0, 1 \le i \le c, 1 \le j \le n \quad (1b)$$

$$\sum_{i=1}^{n} u_{ij} = 1, 1 \le i \le c \quad (1c)$$

The FCM algorithm is composed of the following steps.

1. Get the input data from an image.
2. Choose the number of cluster and the value of ε ($\varepsilon > 0$).
3. Compute partition matrix using (2).

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}} \quad (2)$$

4. Update the cluster center using (3).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$c_j = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} \quad (3)$$

5. Repeat step 3 to 4 where $\left\| c^k - c^{k+1} \right\| < \varepsilon$.

3.2 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 [14]. The conversion of Red Green Blue (RGB) space into HSV space can be seen in formulae (4), (5), and (6).

$$H = \cos^{-1}\left\{ \frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}} \right\} \quad (4)$$

$$S = 1 - \frac{3}{R+G+B}[\min(R,G,B)] \quad (5)$$

$$V = \frac{1}{3}(R+G+B) \quad (6)$$

3.3 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 (7).

$$A_{m\times n} = U_{m\times k} \sum_{k\times k} V^T{}_{k\times n} \quad (7)$$

where U is features vector, $\sum$ 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 [15].

3.4 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 [16]. 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 [17]:

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 (8). Since, city block metric gave the best precision for content-based image retrieval [18].

$$d = \sum_{i=0}^{n} | x_i - y_i | \quad (8)$$

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

4. EXPERIMENTAL RESULT

This research used 120 images from corel-princeton dataset which divided into four classes: column, flower, horse, and model. The dimension of images is 128×85 pixels. The sample image of each class can be seen in Figure 2.

In order to evaluate the quality of images clustering, this research employed Recall (R), Precision (P) and F-Measure (F) 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 (9), (10) and (11) respectively.

$$\text{Precision} = \frac{Total \text{ number of retrieved relevant images}}{\text{Total number of retrieved images}} \quad (9)$$

$$\text{Recall} = \frac{\text{Total number of retrieved relevant images}}{\text{Total number of relevant images}} \quad (10)$$

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \quad (11)$$

Journal of Theoretical and Applied Information Technology
20th May 2013. Vol. 51 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 2. Sample images of each class

In our experiment, Figure 3 shows the manual scribble and separated foreground image from the background. Further, this foreground image was used to group the images.

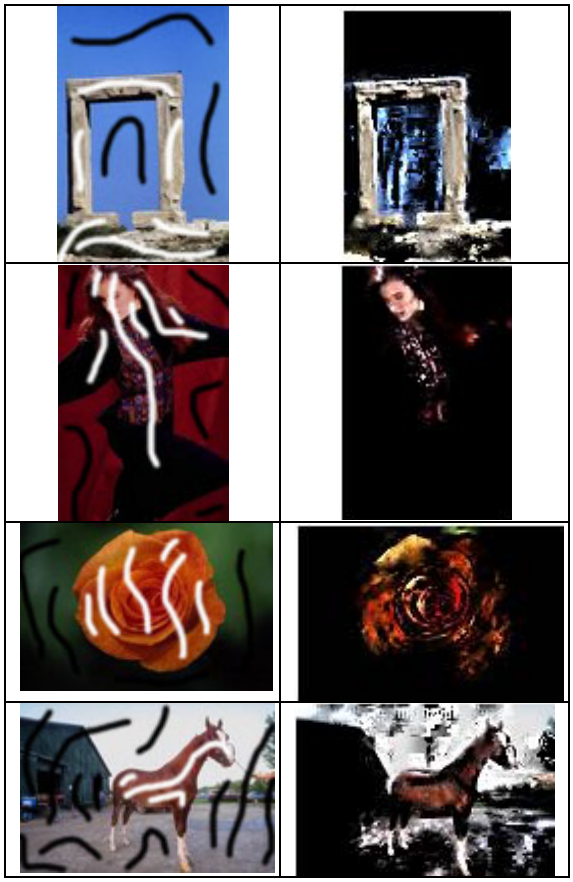

Figure 3. **Left Side.** Manual scribble. **Right Side.** Collected foreground from image matting.

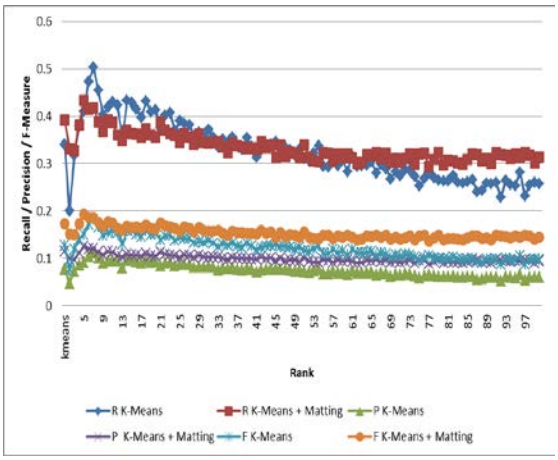

Figure 4. A Comparison Performance of the using image matting in image clustering.

Figure 4 shows the performance of using image matting in image clustering. Based on the result in Figure 4, the better accuracy of image clustering without image matting were obtained when the rank of SVD between $k = 4$ and $k = 7$. The proposed image matting in image clustering improved the Precision and F-Measure in all rank of SVD. Meanwhile, the Recall was improved when the rank of SVD $k = 50$ or higher. Overall, the accuracy of image clustering can be improved by the using of image matting.

5. CONCLUSSION AND FUTURE WORK

Image matting was used to subtract the foreground and background of an image. The using of color in foreground image evidently improves the performance of image clustering. SVD was used to reduce the large number of zeros in feature-image matrix.

It still human intervention in image matting. Therefore, in future work, automatic scribble can be used to overcome the shortcoming of our model.

REFRENCES:

[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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[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] C. Supriyanto, G. F. Shidik, R. A. Pramunendar, and P. N. Andono, "Performance enhancement of image clustering using singular value decomposition in color histogram content based image retrieval," *International Journal of Computer and Communication Engineering*, Vol. 1, No. 4, 2012, pp. 317-320.

[6] R. S. Basuki, M. Hariadi, and R. A. Pramunendar, "Fuzzy c-means algorithm for adaptive Threshold on alpha matting," in *Proc. of International Conference on Information Technology and Electrical Engineering*, Yogyakarta, Indonesia, 2012.

[7] 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, vol. 321, 2009, pp. 481-484.

[8] 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.

[9] 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.

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

[11] 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, 2009, pp. 116-122.

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

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

[14] 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.

[15] 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.

[16] R. Xu and D. Wunsch, "Survey of clustering algorithm," *IEEE Transaction on Neural Network*, vol. 16, 2005, pp. 645-678.

[17] M. H. Dunham, *Data mining introductory and advanced concepts*. Pearson Education, 2006.

[18] 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.