Received: 3 March 2016 | Revised: 19 February 2017 | Accepted: 24 March 2017

DOI: 10.1111/exsy.12220

WILEY Expert Systems

ARTICLE

Content-based image retrieval based on binary signatures cluster graph

Thanh The Van[1,2] | Thanh Manh Le[1]

[1] Faculty of Information Technology, Hue University of Sciences, Hue, Vietnam

[2] Center for Information Technology, HCMC University of Food Industry, Ho Chi Minh City, Vietnam

Correspondence
Thanh The Van, Faculty of Information Technology, Hue University of Sciences, Hue, Vietnam.
Email: vanthethanh@gmail.com

Abstract

In this paper, we approach a method of clustering binary signature of image in order to create a clustering graph structure for building the content-based image retrieval. First, the paper presents the segmentation method based on low-level visual features including colour and texture of image. On the basis of segmented image, the paper creates binary signature to describe location, colour, and shape of interest objects. In order to match similar images, the paper presents a similarity measure between the images based on binary signature. From that, the paper proposes the method of clustering binary signature to quickly query similar images. At the same time, the graph data structure is built using the partition cluster technique and the rules of binary signatures' distribution of images. On the basis of data structure, we propose a graph creation algorithm, a cluster splitting/merging algorithm, and a similarity image retrieval algorithm. To illustrate the proposed theory, we build an image retrieval application and assess the experimental results on the image datasets including COREL (1,000 images), CBIR images (1,344 images), WANG (10,800 images), MSRDI (15,720 images), and ImageCLEF (20,000 images).

KEYWORDS

binary signature, cluster graph, image mining, image retrieval, similarity measure

1 | INTRODUCTION

Nowadays, image data are applied widely in many fields such as digital library, geographic information system, satellite observation system, criminal investigation system, astronomical researches, bio-medical application, and so forth (Marques & Furht, 2002; Muneesawang, Zhang, & Guan, 2014; Wang, 2001). On the other hand, image is a non-structured data because their content has intuitive nature (Acharya & Ray, 2005). So there are many challenges in looking for utility information from large image dataset in the problem of image mining. Moreover, image retrieval is an important problem in the field of visual computer and image processing (Acharya & Ray, 2005).

On the other hand, image data have become familiar with daily life and are used in many different devices. The digitization of image data created colossal dataset, which leads to the problem of searching object becomes more complicated and has more challenges such as applying automatic classification, querying by the content of object, creating index, and query quickly relevant objects, reducing storage in the process of query, and so forth.

According to IDC report, in 2015, people shared over 1.6 trillion images all over the world, 70% of which is generated from mobile devices (IDC, 2016). The digitized multimedia data have created many huge datasets leading to the complexion of searching problems, so there are many challenges, such as classifying automatically according the objects' content, creating index to retrieve quickly relevant objects, reducing the searching space, and so forth. The similar images retrieval from large image datasets is an important problem in computer vision (Acharya & Ray, 2005; Deligiannidis & Arabnia, 2015). According to the survey results of the recent works, the finding problem of relevant images is consistent with the needs of modern society (ACI, 2015). The problem is that we should build an effective retrieval method, mean to search quickly similarity images from the set of large images with high accuracy.

The image retrieval problem is divided into two main classifications (Acharya & Ray, 2005; Marques & Furht, 2002; Muneesawang et al., 2014; Wang, 2001) including (a) image retrieval based on key word TBIR (text-based image retrieval), that is, image's index on the base of description in form of text is defined by user; therefore, it is time-consuming to describe image's content, and there are particular limitations because of

subjectivity of people; (b) image retrieval based on content-based image retrieval (CBIR) is presented in 1980; the CBIR finds a set of similar images with query image on the basis of automatic extraction about content of image (Wang, 2001). So the CBIR overcomes the restriction of TBIR. However, it has many difficult matters such as extracting feature, creating multi-dimension indexes, and giving the similar image retrieval method.

CBIR architectural includes two parts: (a) extracting feature to create index for the image; (b) implementing the similar image retrieval based on index. The result of image retrieval returns the most similar images in the given measure (Acharya & Ray, 2005; Muneesawang et al., 2014). However, if we search similar images that rely on direct matching by content images, it takes query storage and query time. So we need to describe the content image in form index; then, we query similar images via this index.

The first step of the problem of CBIR performs image preprocessing such as colour analysis, segmentation, filter, de-noise, and so forth. With a processed image, we extract features for creating image signature. The signature is a form of index that describes features of image (Acharya & Ray, 2005; Le & Van, 2013; Van & Le, 2014a,b,c). It is used for automatic classification and semantic classification according to image's content. With each signature of query image, we find a set of relevant signatures (or non-relevant) in image dataset. On the basis of the set of this signature, we retrieve information to find out a set of image similar to query image. The final step of the problem brings an ordered set of images according to user requirement.

We need to describe the content image in form of metadata. After that, we query similar images via this metadata. The paper approaches the CBIR and uses binary signature to create index for image object. The binary signature of image has length n as a vector in space $\sum^n$ (with as the number of dimensions and $\sum = \{0,\ 1\}$ as a set of basis symbols) to describe visual interest of image (Manolopoulos, Nanopoulos, & Tousidou, 2003).

When using binary signature, we will reduce the storage space and simplify data with complicated calculation models. From there, it reduces significantly the number of comparisons when performing the problem of similar images retrieval. Furthermore, binary signature applies easily logical operations (AND, OR, NOT) for implementing inference on the basis of logical rules. In this case, the problem of image retrieval becomes the image mining on a set of binary signatures.

The content of the paper presents the method of clustering binary signature to create a structure of cluster graph, from that we can build an image retrieval system effectively. Therefore, the paper approaches the method of image segmentation for creating binary signature based on *CIE* $L^*a^*b^*$ colour space and Wavelet transform in order to extract colour and texture of image. This binary signature describes location, colour, and shape of interest object. From there, the paper proposes the similarity measure and performs clustering binary signature to quickly query similar images. This paper is inspired from our earlier work in 2014 and 2016 (Van & Le, 2014a; Van & Le, 2016).

The main contributions of the paper include as follows: (a) create the binary signature and similarity measure based on interest objects and colours of images; (b) approach the clustering method for binary signature; (c) create the cluster graph to store binary signatures of images. After that, we propose the creating cluster graph algorithm and improve this algorithm, which relies on feature vector of cluster, the splitting/merging algorithm; (d) give the image retrieval algorithm and build the CBIR. From that, we assess experimental results and compare with other methods.

The rest of the paper is organized as follows: Section 2 mentions related works to prove the feasibility and improvement of the proposed method; Section 3 performs the image segmentation to extract the interest object in order to create binary signature, concurrently describe the similar measure between two binary signatures for evaluating the similarity between two images; Section 4 presents the method of building the cluster of the binary signature as well as to describe the relationship between the images. Then, we build cluster graph to image retrieval system; Section 5 presents the algorithm of image retrieval and describes the empirical experiment of image retrieval on cluster graph; conclusions and discussions of future works are given in Section 6.

2 | RELATED WORK

Many applications related to image retrieval are developed and applied in many different fields such as applying in digital library including CIRES, C-BIRD, Photo-File, iMATCH, and so forth (Muneesawang et al., 2014); Image retrieval application IRMA in medicine on the base of support vector machine (Huang, Zhang, Zhao, & Ma, 2010), medical image retrieval CBMIR (content-based medical image retrieval) on computed tomography image (Jin, Hong, & Lianzhi, 2009), medical image retrieval system on wavelet transform (Rajakumar & Muttan, 2010), image retrieval application on Geographic Information System (Shea & Cao, 2012), and so forth.

Many works related to CBIR are published such as extracting object on image based on histogram value (Wang, Wu, & Yang, 2010), similar image retrieval based on matching interest regions (Bartolini, Ciaccia, & Patella, 2010; Wang et al., 2010), colour image retrieval based on bit plane and colour space L*a*b* (Wang, Yang, Li, & Yang, 2013), converting colour space and building hashing in order to retrieve content of colour images (Tang, Zhang, Dai, Yang, & Wu, 2013), and so forth.

There are many feature detection methods introduced (Wang et al., 2013), including angle and edge detector method introduced in 1998 by Harris & M. Stephens, scale-invariant feature transform (SIFT) detector method introduced in 2003 by D. Lowe based on the filter of convolution mask between image and difference of Gaussians to approximate Laplacian operator of Gaussian function, speeded up robust features introduced in 2006 by Bay et al., the point detector method based on Laplacian operator of Gaussian function in 2001 by Mikolajczyk & C. Schmid, and so forth.

In 1973, Haralick et al. introduced co-occurrence matrix to describe feature texture (Muneesawang et al., 2014). In this method, the co-occurrence matrix is built on the base of direction and distance among pixels. The texture is extracted from co-occurrence matrix via frequency of grey level.

Tamura proposed the method of approximate texture on the basis of human visual system. Wavelet transform is applied in analysing texture and classifying images on the basis of decomposition of multi-resolution of images (Acharya & Ray, 2005).

Kumar, Raja, Venugopal, and Patnaik 2009 proposed automatic segmentation method on the basis of wavelet transform in order to create segmentation quickly and easily. The paper shows that this method segments effectively on large images and implements more easily than other methods.

Chitkara, Nascimento, and Mastaller (2000) presented the technical report about CBIR on the basis of binary signature at Alberta University, Canada. The content of the report proposed the method of creating binary signature of colour image and gave out the similarity measure among binary signatures applying in image retrieval problem. That document evaluated the accuracy of experiment from large image dataset to demonstrate method's feasibility.

Manolopoulos et al. (2003) described the binary signature to query similar images based on S-Tree. Besides, Nascimento and Timothy Chappell approached the similar image retrieval method based on binary signature. The experiment shows the effectiveness when querying on large image datasets. (Chappell & Geva, 2013; Nascimento & Chitkara, 2002; Nascimento, Tousidou, Chitkara, & Manolopoulos, 2002).

El-Kwae and Kabuka (2000) proposed a method of image retrieval based on binary signature and a multi-level file structure. In order to evidence method's effectiveness, El-Kwae and Kabuka (2000) analysed the theory and build an experiment application about image retrieval, which has a large image dataset.

Snášel (2000) applied fuzzy signature and S-Tree for similarity image retrieval problem. The experiment in this paper is compared with other methods to show the effectiveness of proposed method. El-Kwae (2000) used binary signature and hierarchical index file to increase the efficiency for image retrieval problem.

Ahmad and Grosky (2003) used binary signature as an index for image and applied in image retrieval problem.

Nascimento and Chitkara (2002) approached image retrieval technical based on binary signature. The experimental results show that the efficiency for image retrieval problem has a large image dataset.

Prasad, Biswas, and Gupta (2004) announced the work about image retrieval using binary index to describe the low-level feature. They experimented on the accuracy of the image retrieval method.

Landre and Truchetet (2007) presented the image retrieval method on the basis of binary signature and Hamming similarity measure. The experimental results of the paper show the effect about retrieval speed and efficiency.

Abdesselam, Wang, and Kulathuramaiyer (2010) built a CBIR system based on binary bit-string. The paper proposed a similarity measure for bit-string. It assessed the effectiveness of retrieval performance and time.

Chappell and Geva (2013) approached the searching method based on binary signature. Their paper showed the effectiveness and increased speed of image retrieval in the Hamming measure application and then assessed the similarity measure among binary signatures.

Ren, Cai, Li, Yu, and Tian (2014) proposed the method of searching similarity images on the basis of bit-string to describe the SIFT feature of image. The experiment of the paper evidenced the effectiveness of given method on different image datasets. Cai, Liu, Chen, Joshi, and Tian (2014) used bit-string to describe the visual feature of image. The method decreased the query space and increased the retrieval time. Özkan, Esen, and Akar (2014) introduced the cluster method for video data using binary signature on the basis of the visual feature of image. The experiment of the paper shows the effectiveness on reducing query space and increasing query speed.

Liu, Lu, and Suen (2015) used image signature on the basis of bit-string and applied earth mover distance (EMD) to match images. The paper proved the efficiency for variant signature in the experiment.

Zhou, Li, Wang, Lu, and Tian (2012), Zhou, Lu, Li, and Tian (2012), Zhou, Li, Lu, and Tian (2013), Zhou, Li, and Tian (2014) announced many works about image retrieval on the basis of binary signature to describe the SIFT feature of image. The paper's experiment proved the effectiveness on large image datasets (Zhou, Li, Lu, & Tian, 2011; Zhou, Li, & Lu, 2015).

Recently, many works of similar image retrieval based on binary signature have been published such as image retrieval based on index and S-Tree (Nascimento et al., 2002), image retrieval based on EMD measure and S-Tree (Le & Van, 2013; Van & Le, 2014), image retrieval based on binary signature (Nascimento & Chitkara, 2002), image retrieval based on signature graph (Van & Le, 2014; Van & Le, 2014), and so forth.

CBIR based on the graph data structure has been applied widely, such as building the graph data structure based on the low-level feature of image to search neighbour pixels and apply in image retrieval problem (Xu, Bu, Wang, & He, 2015), making image retrieval using relevance feedback based on graph data structure to sort the images in accordance with user's feedback (Kundu, Chowdhury, & Bulò, 2015), clustering the image set based on the low-level feature and build the cluster graph to apply in the CBIR (Yan, Liu, Wang, Zhang, & Zheng, 2014), searching similar images by the graph structure based on semantic measure between images (Zhao et al., 2013), creating the index graph structure for image retrieval problem (Demirci, 2012), creating the cluster graph structure based on interest region of image and applying in content-based medical image retrieval (Li, Zhang, Pan, Han, & Feng, 2012), creating CBIR based on the graph structure using grey level of image (Li & Lu, 2011), create the cluster graph structure based on K-mean algorithm (Hlaoui & Wang, 2003), and so forth.

According to the survey of related works, the method of CBIR based on binary signature is very effective to develop the tool of similarity image retrieval. So the paper approaches the method that creates binary signature based on interest object of segmented image. The binary signature

describes the shape, colour, and location of interest objects. From there, we cluster binary signatures to query quickly for similar images. Moreover, the graph structure is also applied in many different image retrieval problems. Then, the paper creates the cluster graph to store binary signature, from that we build the CBIR system.

3 | CREATING BINARY SIGNATURE OF IMAGE

3.1 | Image segmentation

In order to increase accuracy and query speed in image retrieval problem, we create binary signature to describe colour and shape of interest object of image. In this case, the binary signature includes the following: (a) a binary signature describes shape of interest objects; (b) a binary signature describes colours of image.

Each image is extracted from the colour palette such as MPEG7 (Wang et al., 2010), the colour palette from image collections using K-mean algorithm, and CIE-$L^*a^*b^*$ colour space (Wengert, Douze, & Jégou, 2011) including 16 colours, 32 colours, 64 colours, 128 colours, and 256 colours. Each image is quantized as n fixed colours $c_1, c_2, ..., c_n$; each colour c_j is described by a bit-string $b_1^j b_2^j ... b_t^j$. On the basis of a document of reference (Nascimento et al., 2002), the binary signature describes colours of image as a bit-string as follows:

$$S = b_1^1 b_2^1 ... b_t^1 ... b_1^2 b_2^2 ... b_t^2 ... b_1^n b_2^n ... b_t^n \tag{1}$$

To recognize the texture vector of neighbouring pixels, we use the discrete wavelet frames transform (Unser, 1995) to convert the intensity into sub-samples. The DWT is executed through the low-pass filter $H(z)$ to decompose the intensity. The high-pass filter $G(z) = zH(-z^{-1})$ is defined relying on $H(z)$. The bank filter $H_V(z)$, $G_i(z)$, $i = 1, ..., V$ is created by $H(z)$, $G(z)$, with $H_{k+1}(z) = H\left(z^{2^k}\right)H_k(z)$, $G_{k+1}(z) = G\left(z^{2^k}\right)H_k(z)$, with $k = 0, ..., V-1$, $H_0(z) = 1$.

The standard deviation value reflects entropy around the expectation value, and this entropy describes the texture of discrete signals. So the standard deviation of all detailed components on discrete wavelet frames transform is used as the feature texture. For this reason, the texture vector corresponding to pixel p is $T(p) = [\sigma_1(p), \sigma_2(p), ..., \sigma_{9\times V}(p)]$, which is calculated on neighbouring square.

The colour vector $I(p) = (I_L(p), I_a(p), I_b(p))$ is created using the colour space *CIE* $L^*a^*b^*$, which is endorsed as international standard in 1970 and uniform perception of human. The Euclidean distance between two points on this colour space corresponds to the perception distance between two colours on the human visual system (Acharya & Ray, 2005).

After extracting the texture and colour of image, we implement the process of cluster of all pixels on the image by K-means method. The first step is choosing the centre cluster relied on the contrast C of the image. To quickly execute, the image I is divided into non-overlap blocks called *supper pixels*. Figure 1 describes that an image is divided into 7x11 blocks. Therefore, the texture vector $T^b(b_l)$ and colour vector $I^b(b_l)$ of the block b_l are average values of texture vectors and colour vectors of all pixels on the block.

Definition 1. *Given the two arbitrary blocks b_l, b_n, the contrast of image is*

$$C = \max\{d = \alpha||I^b(b_l) - I^b(b_n)|| + \beta||T^b(b_l) - T^b(b_n)||\} \tag{2}$$

(In the experiment, we use $\alpha = \beta = 0.5$) The background and the foreground of image correspond to the blocks that have low energy and high energy, respectively.

In the next step, we find the set of complement centre O (i.e., we search the nearest blocks with foreground relying on the measure d). In the experiment, we find the centres that have $d > \mu C$ (with $\mu = 0.4$).

FIGURE 1 C blocks of image

Algorithm 1. *Segmentation*

Input: The color image I

Output: The mask M

begin

Step 1: Extract texture vector $T(p)$ and intensity vector $I(p)$ for each pixel on image I.

Step 2: Compute the center of blocks by calculating the average of texture vectors and color vectors of all pixels in each block.

Step 3: Calculate the contrast C to form background and foreground.

Step 4: Find the set of complement center O based on the contrast C.

Step 5: Cluster all the pixels on the image I rely on the set of center O.

Step 6: Create the mask M with clustered pixels.

Step 7: Remove the regions have small area rely on the mask M.

Step 8: Return the mask M.

End

The result of process of segmentation is a mask (i.e., grey image) to describe interest objects of image. On the basis of the mask, we compute the connected regions and remove the regions that have small area (the experiment removes the regions that have area less than 5% of image). In Figure 2, we describe the masks and segmented images.

In the last step, we remove the connected regions which have area less than threshold θ. The computing of area of regions is done by 4-neighbouring algorithm as follows:

Algorithm 2. *Compute the area of region*

Input: The mask M and the location (r, c)

Output: The area value S

Begin

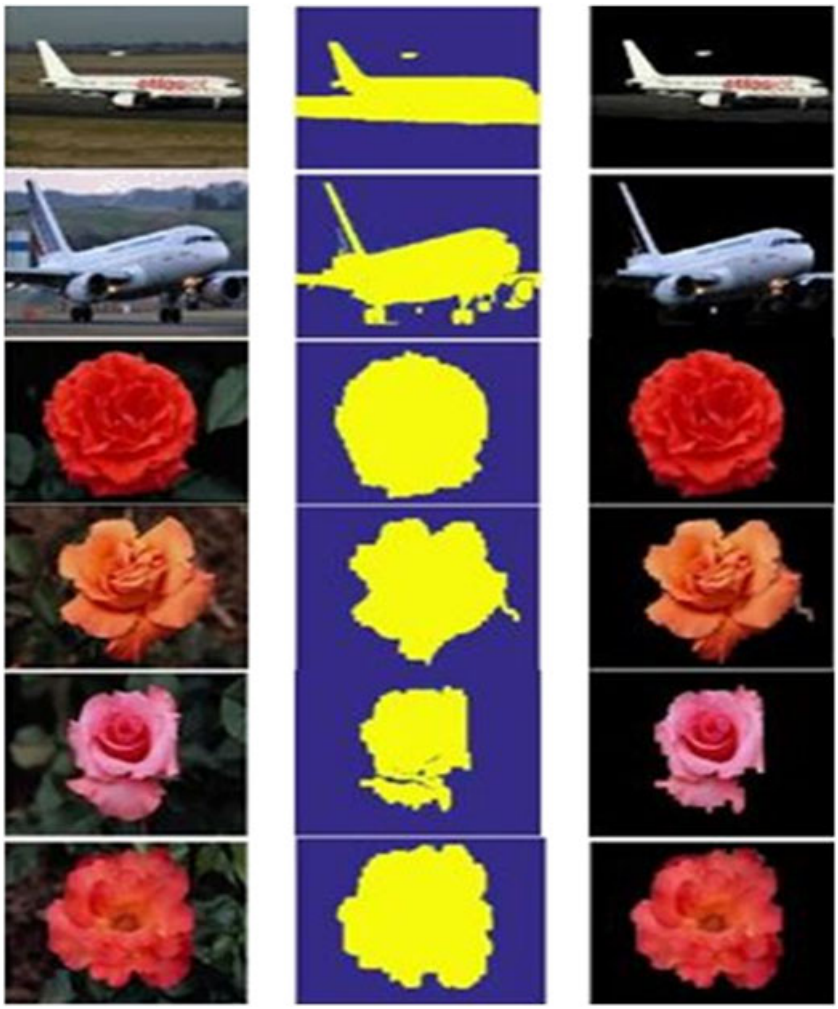

FIGURE 2 Some results of segmented images

Step 1: Initialize $S = 0$; Stack =∅;
Step 2: Push(Stack, r, c);
Step 3: **while** Stack ≠∅ **do**
(r, c)= Pop(Stack);$S = S + 1$;
If$(r>1)$**and**$(M(r, c) == M(r-1, c)$ **then** $Push(Stack, r-1, c)$;
End If
If$(r<rows)$**and**$(M(r, c) == M(r + 1, c)$**then** $Push(Stack, r + 1, c)$;
End If
If$(c>1)$**and**$M(r, c) == M(r, c-1)$**then** $Push(Stack, r, c-1)$;
End If
If$(c<columns)$**and**$M(r, c) == M(r, c + 1)$**then** $Push(Stack, r, c + 1)$;
End If
End While
Return*S*;
End

3.2 | Creating binary signature of image

After extracting the interest regions of image, we create binary signature to describe content of image. Each image is divided into *n*×*n* equal cells. Let *O* be an interest region of image, the binary signature of the interest region is defined as follows:

Definition 2. The binary signature of the interest object *O on image I is a bit-string* $Sig_I = b_1b_2...b_N$, *where* $b_i = 1$ *if the i*th *cell of image I has overlap with the interest region O; otherwise* $b_i = 0$, *with* $N = n{\times}n$ *be the number of cells of image.*

Each image is quantized in the form of *M* colours. The binary signature describing colours of image can be defined as follows.

Definition 3. Let $c_1, c_2, ..., c_m$ *be indexes of dominance colours of interest region* $O_1, O_2, ..., O_m$ *with* $c_i{\in}\{1, 2, ..., M\}$. *Then, the binary signature describing colours of image I based on interest regions* $O_1, O_2, ..., O_m$ *is a bit-string* $Sig^I_C = b_1b_2...b_M$, *where* $b_i = 1$ *if* $i{\in}\{c_1, c_2, ..., c_m\}$; *otherwise* $b_i = 0$.

On the basis of the binary signature of interest regions and the ones of dominance colours, the binary signature of image is defined as follows:

Definition 4. Giving image *I*, *let* $Sig^I_O = b^O_1...b^O_N$ *and* $Sig^I_C = b^C_1...b^C_M$ *be the binary signature of interest regions and the ones of dominance colours of image* $I^b(b_I)$, *respectively. Then, the binary signature of image* b_I *is defined as follows:*

$$Sig = Sig^I_O \oplus Sig^I_C = b^O_1...b^O_N b^C_1...b^C_M \quad (3)$$

The formula 3 describes image's binary signature including two parts: the first part is a string with *N*-bit to describe binary signature of object; the second part is a string with *M*-bit to describe binary signature of colours of image. An example about binary signature of image is show in Figure 3.

3.3 | The similarity measure

The similarity measure between two images is a necessary component when we perform the process of similar image retrieval. However, each image is described in the form of binary signature, so the problem needs to build the similarity measure between binary signatures from that to assess the similarity of images. Then, the paper defines the similarity measure as follows:

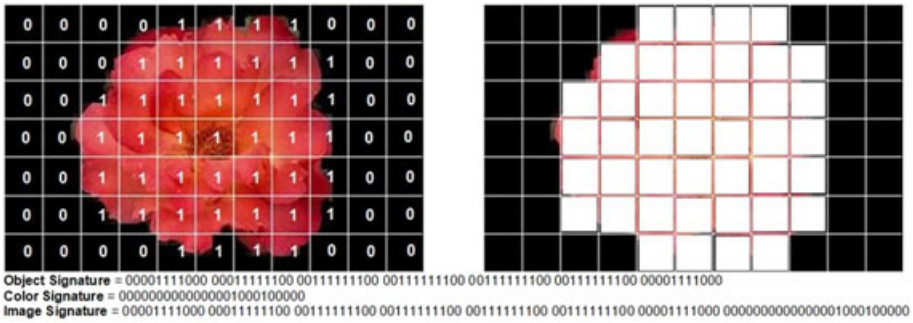

FIGURE 3 An example about binary signature of image

Definition 5. Let $Sig(I) = Sig_O^I \oplus Sig_C^I$ and $Sig(J) = Sig_O^J \oplus Sig_C^J$ be binary signatures of images I and J. Then, the similarity measure between two images I and J is $\delta(I,J) = \alpha \times d(sig_O^I, sig_O^J) + \beta \times d(sig_C^I, sig_C^J)$, with $\alpha, \beta \in (0,\ 1)$ as adjustment coefficients, where $\alpha + \beta = 1$; N is the number of cells of image; and M is the number of colour to quantize image.

$$d(sig_O^I, sig_O^J) = \frac{|\,NOT\left(sig_O^I XORsig_O^J\right)|}{N} = \frac{\sum_{i=1}^{N} NOT\left(sig_O^I[i] XORsig_O^J[i]\right)}{N} \in [0,\ 1] \tag{4}$$

$$d(sig_C^I, sig_C^J) = \frac{|\,NOT\left(sig_C^I XORsig_C^J\right)|}{M} = \frac{\sum_{i=1}^{N} NOT\left(sig_C^I[i] XORsig_C^J[i]\right)}{M} \in [0,\ 1] \tag{5}$$

The function of distance *O* matches on each image's cell on the basis of two criteria, which are colour and interest region's structure. On the basis of operator with each bit of binary signatures of image including XOR and NOT operator, the function *M* returns the number of same components between two binary signatures. Therefore, the function *M* matches the same components based on colours, location and shape of interest regions of image. In order to calculate quickly, we apply the Hamming measure (Alzu'bi, Amira, & Ramzan, 2015; Zhou, Li, et al., 2012; Zhou, Lu, et al., 2012; Zhuang, Zhang, & Li, 2013) to create object measure *M* and colour measure θ as follows:

Definition 6. Let $Sig(I) = Sig_O^I \oplus Sig_C^I$ and $Sig(J) = Sig_O^J \oplus Sig_C^J$ be binary signatures of image I and J. The dissimilarity measures based on interest regions and colour in turn $\mu_O(sig_O^I, sig_O^J)$ and $\mu_C(sig_C^I, sig_C^J)$ are defined as follows:

$$\mu_O(sig_O^I, sig_O^J) = \frac{|\left(sig_O^I XORsig_O^J\right)|}{N} = \frac{\sum_{i=1}^{N} \left(sig_O^I[i] XORsig_O^J[i]\right)}{N} \in [0,\ 1] \tag{6}$$

$$\mu_C(sig_C^I, sig_C^J) = \frac{|\left(sig_C^I XORsig_C^J\right)|}{M} = \frac{\sum_{i=1}^{N} \left(sig_C^I[i] XORsig_C^J[i]\right)}{M} \in [0,\ 1] \tag{7}$$

N is the number of image cells and M is the number of colours to quantize the image.

On the basis of the combination of similarity measure between interest object and colours, the dissimilarity measure ϕ between two images is defined as follows:

Definition 7. Let $Sig(I) = Sig_O^I \oplus Sig_C^I$ and $Sig(J) = Sig_O^J \oplus Sig_C^J$ be binary signatures of image I and J. Then, the dissimilarity measure between two images I and J is defined as follows:

$$\phi(I,J) = \alpha \times (\mu_O(sig_O^I, sig_O^J)) + \beta \times (\mu_C(sig_C^I, sig_C^J)) \tag{8}$$

with $\alpha, \beta \in (0,\ 1)$ as adjustment coefficients, $\alpha + \beta = 1$.

Theorem 1. Let $Sig(I) = Sig_O^I \oplus Sig_C^I$ and $Sig(J) = Sig_O^J \oplus Sig_C^J$ be binary signatures of image I and J. The function of similarity relies on interest regions and colours in turn $d(sig_O^I, sig_O^J)$ and $d(sig_C^I, sig_C^J)$. Then, the dissimilarity measure based on interest regions and colours in turn $\mu_O(sig_O^I, sig_O^J) = 1 - d(sig_O^I, sig_O^J)$ and $\mu_C(sig_C^I, sig_C^J) = 1 - d(sig_C^I, sig_C^J)$.

***Proof*:**

We have

$$\mu_O(sig_O^I, sig_O^J) + d(sig_O^I, sig_O^J) = \frac{\sum_{i=1}^{N} \left(sig_O^I[i] XORsig_O^J[i]\right)}{N} + \frac{\sum_{i=1}^{N} NOT\left(sig_O^I[i] XORsig_O^J[i]\right)}{N}$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left[\left(sig_O^I[i] XORsig_O^J[i]\right) + NOT\left(sig_O^I[i] XORsig_O^J[i]\right)\right] = 1$$

Infer $\mu_O(sig_O^I, sig_O^J) = 1 - d(sig_O^I, sig_O^J)$.

With similar method, we have $\mu_C(sig_C^I, sig_C^J) = 1 - d(sig_C^I, sig_C^J)$. ■

Theorem 2. *The function of dissimilarity measure* μ_α *is a metric because of properties as follows:*

1. *Non-negative:* $\mu_\alpha(sig_\alpha^I, sig_\alpha^J) \geq 0$ *and* $\mu_\alpha = 0 \Leftrightarrow sig_\alpha^I = sig_\alpha^J$
2. *Symmetry:* $\mu_\alpha(sig_\alpha^I, sig_\alpha^J) = \mu_\alpha(sig_\alpha^J, sig_\alpha^I)$
3. *Triangle inequality:* $\mu_\alpha(sig_\alpha^I, sig_\alpha^J) + \mu_\alpha\left(sig_\alpha^J, sig_\alpha^K\right) \geq \mu_\alpha\left(sig_\alpha^I, sig_\alpha^K\right)$

***Proof*:**

1. ***Non-negative***

Let $sig_\alpha^I, sig_\alpha^J$ be two binary signatures of images *I* and *J*, with$sig_\alpha^I[i], sig_\alpha^J[i] \in \{0, 1\}$, $i = 1, ..., N$.

Then, $sig_\alpha^I[i] XOR sig_\alpha^J[i] \geq 0$. Infer $\mu(sig_\alpha^I, sig_\alpha^J) \geq 0$.

Hence, the function of dissimilarity measure $\mu_\alpha(sig_\alpha^I, sig_\alpha^J)$ is non-negative. Assume that

$$\mu_\alpha(sig_\alpha^I, sig_\alpha^J) = 0 \Leftrightarrow \frac{\sum_{i=1}^{N}(sig_\alpha^I[i] \text{XOR} sig_\alpha^J[i])}{N} = 0 \Leftrightarrow sig_\alpha^I[i] = sig_\alpha^J[i], i = 1, ..., N$$

$$\Leftrightarrow \frac{\sum_{i=1}^{N}(sig_\alpha^I[i] \text{XOR} sig_\alpha^J[i])}{N} = 0.$$

So the function of dissimilarity measure $\mu_\alpha(sig_\alpha^I, sig_\alpha^J)$ is unique.

2. ***Symmetry***

The XOR operator is commutative, then

$$\mu_\alpha(sig_\alpha^I, sig_\alpha^J) = \frac{\sum_{i=1}^{N}\left(sig_\alpha^I[i] XOR sig_\alpha^J[i]\right)}{N} = \frac{\sum_{i=1}^{N}\left(sig_\alpha^J[i] XOR sig_\alpha^I[i]\right)}{N} = \mu_\alpha(sig_\alpha^J, sig_\alpha^I)$$

So $\mu(sig_\alpha^I, sig_\alpha^J) = \mu(sig_\alpha^J, sig_\alpha^I)$. Hence, the function of dissimilarity measure $\mu_\alpha(sig_\alpha^I, sig_\alpha^J)$ is symmetric.

3. ***Triangle inequality***

Let $sig_\alpha^I, sig_\alpha^J, sig_\alpha^K$ be three binary signatures of images I, J, and K. So $\mu_\alpha(sig_\alpha^I, sig_\alpha^J) + \mu_\alpha\left(sig_\alpha^J, sig_\alpha^K\right) = \frac{1}{N}\sum_{i=1}^{N}\left(sig_\alpha^I[i] XOR sig_\alpha^J[i]\right) + \left(sig_\alpha^J[i] XOR sig_\alpha^K[i]\right)$

We can use a truth table, infer $\left(sig^I[i] XOR sig^J[i]\right) + \left(sig^J[i] XOR sig^K[i]\right) \geq \left(sig^I[i] XOR sig^K[i]\right)$. Then, $\mu_\alpha(sig_\alpha^I, sig_\alpha^J) + \mu_\alpha\left(sig_\alpha^J, sig_\alpha^K\right) \geq \mu_\alpha\left(sig_\alpha^I, sig_\alpha^K\right)$.

Therefore, the function of dissimilarity measure $\mu_\alpha(sig_\alpha^I, sig_\alpha^J)$ satisfies the condition of the triangle inequality. ■

The paper uses the dissimilarity measure ϕ to assess the similar level between two images. So the dissimilarity measure ϕ is used as the similarity measure.

4 | IMAGE RETRIEVAL USING CLUSTER GRAPH

4.1 | Binary signature clustering

After creating binary signature and similarity measure between the images, querying quickly similar images needs to be solved. The target of clustering aims to reduce searching space and increase the query speed of similarity image.

Fuzzy c-means is a clustering method that allows items to belong to clusters based on minimized objective function. In this method, the value of membership function with a power parameter of each item has to be defined (Alnihoud, 2012; Bhanu & Dong, 2002; Shambharkar & Tirpude, 2011; Wang et al. 2013). Each item has a degree of membership corresponding to each current cluster, so we determine the items in clusters and define relevant clusters corresponding to each item. However, the fuzzy clustering method has disadvantages as follows:

1. If the number of items of a cluster has a change, the centre cluster will be updated. This leads to the membership function of all items, which has to update on all clusters. Therefore, this method is costly to update membership function of items of cluster;
2. If the value of membership of items has a change, the number of each cluster will change because the items can move to another cluster. So this method can be costly to restructure all clusters. Furthermore, if the number of cluster grows, we have to update again all the membership function for each item in all clusters. Therefore, the result of fuzzy clustering method can change according to the number of item leading to the result of searching a set of similar items on clustering structure, which can be different by the time.
3. So the membership function belonging to power parameter can lead to many errors when clustering item based on objective function.

On the basis of the above analysis, we do not perform the fuzzy clustering method of binary signature in image retrieval problem. Instead, the paper will improve the clustering method on the base of K-mean clustering algorithm.

According to Jain's (2010) cluster method, this paper describes the cluster method, which groups the objects on a given similarity measure. This method is divided into two forms: hierarchical cluster and partition cluster; The K-mean algorithm is the most popular partition cluster, announced in 1955. The K-mean algorithm has to determine three input parameters including the number of cluster K, the number of initial centre, and a similarity measure. Moreover, if we add new item into cluster, we redetermine the new centre of cluster.

The cluster methods have been applied in many image retrieval problems such as apply K-mean algorithm and Euclidean distance to cluster and apply in image retrieval (Lin, Chen, Lee, & Liao, 2014), build CBIR using K-mean algorithm and Mahalanobis distance between colour feature vectors of images (Banerjee, Bandyopadhyay, & Pal, 2013), use the K-mean algorithm and MPEG7 palette to cluster and retrieve similarity images (Saboorian, Jamzad, & Rabiee, 2010), combine the scale feature of colours, texture, shape to cluster, and query similar images (Zakariya, Ali, & Ahmad, 2010), cluster images based on colours and K-mean algorithm (An, Baek, Shin, Chang, & Park, 2008), query image by content based on clusters of images and the unsupervised learning (Chen, Wang, & Krovetz, 2005), and so forth.

The image data are grown up by the time, so the predetermination of the number of centre clusters does not accommodate because the distance between items is long. So if we cluster to apply in image retrieval problem based on binary signature, we have to improve as follows:

4. Determine the number of centre clusters, which can grow to ensure the similarity images in a cluster.
5. Determine the similarity measure between items to assess the degree of similarity between two images.
6. Select the centre item as a basis for performing a partition of new items into the clusters based on a given measure.

The paper proposes the clustering binary signature method on the basis of the similarity measure ϕ, in such a way that each item in cluster as a set of $I = sig, oid$ including a binary signature sig and unique indentify oid of a corresponding image. Let V_m be a cluster of items I, where V_m has a centre as I_m and a radius as $k_m\theta$. Each item I is assigned to the cluster V_m if the condition $\phi(I, I_m) \le k_m\theta$ is satisfied. Otherwise, we consider the rules of distribution as shown in **Definition** 9 to classify the image I into appropriate cluster.

In the above proposed clustering method, it is not necessary to predetermine the number of centres of clusters. Moreover, the number of cluster can grow in the increase of the number of image. So our paper builds the cluster of the binary signature as well as describes the relationship between the images. This cluster has an item as a centre. Then, each cluster including similar images is defined as follows:

Definition 8. *A cluster V has a centre I with $k\theta$ as a radius, which is defined as* $V = \{J | \phi(I, J) \le k\theta, J \in \Im, k \in N^*\}$.

With each image, we need to classify in clusters. So we need to have the rule of distribution in clusters. This rule is developed from our own paper (Van & Le, 2014) and defined as follows:

Definition 9. *(Van & Le, 2014) Giving set* $\Omega = \{V_i | i = 1, ..., n\}$ *is a set of clusters, with* $V_i \cap V_j = \varnothing, i \ne j$, let I_0 *be an image needing to distribute in a set of clusters Ω, let* I_m *be a centre of cluster* V_m, *so that* $(\phi(I_0, I_m) - k_m\theta) = \min\{(\phi(I_0, I_i) - k_i\theta), i = 1, ..., n\}$, *with* I_i *is a centre of cluster* I_i. *There are three cases as follows:*

1. *If* $\phi(I_0, I_m) \le k_m\theta$, then the image I_0 is distributed in cluster V_m.
2. *If* $\phi(I_0, I_m) > k_m\theta$, then setting $k_0 = [(\phi(I_0, I_m) - k_m\theta)/\theta]$, at that time:

2.1. *If* $k_0 > 0$, then creating cluster V_0 with centre $\mu_C(sig_C^I, sig_C^J) = 1 - d(sig_C^I, sig_C^J)$ and radius $k_0\theta$, at that time $\Omega = \Omega \cup \{V_0\}$.

2.2. *Otherwise (i.e.,* $k_0 \le 0$), the image I_0 is distributed in cluster V_m and $\phi(I_0, I_m) = k_m\theta$.

On the basis of the similarity measure ϕ and the rules of distribution of image are shown in ***Definition*** 9. The paper proposes the method to create clusters of binary signatures. With the input image dataset $\Im$ and the threshold $k\theta$, the algorithm returns the set of clusters *CLUSTER*. First, we initialize the set of cluster $CLUSTER = \varnothing$, then create the first cluster. With each image I, we evaluate the distance $\mu_\alpha(sig_\alpha^I, sig_\alpha^J)$ with the centre of cluster and find out the nearest cluster according to $(\phi(I, I_0^m) - k_m\theta) = \min\{(\phi(I, I_0^i) - k_i\theta), i = 1, ..., n\}$. If the condition $\phi(I, I_0^m) \le k_m\theta$ is satisfied, the image I is distributed in cluster V_m in Figure 4, we show an illustration of the distribution rules of images.

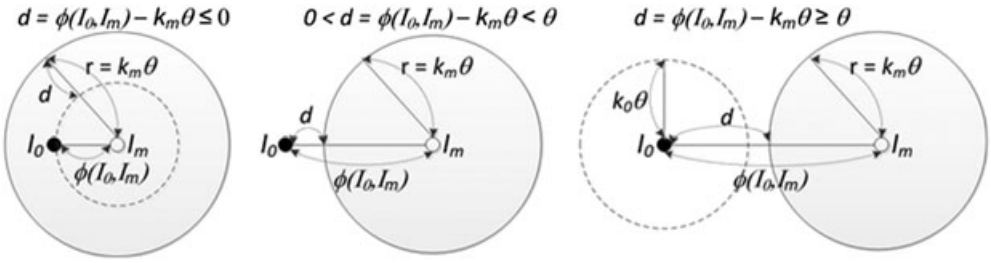

FIGURE 4 An illustration of the distribution rules of images

4.2 | Creating clustering graph

On the basis of the binary signature clustering algorithm (Van & Le, 2016), the cluster graph algorithm is proposed as follows:

Algorithm 3. *Creating cluster graph.*

Input: Set of image data $\Im$ and threshold value $k\theta$

Output: Cluster Graph (V, E)

Begin

$V = \varnothing$; $E = \varnothing$; $k_I = 1$; $n = 1$;

For $(\forall I \in \Im)$ **do**

Begin

If $(V = \varnothing)$ **then**

$I_0^n = I; r = k_I\theta$;

Creating cluster $C_n = I_0^n, r, \phi = 0$;

$V = V \cup C_n$;

Else

$(\phi(I, I_0^m) - k_m\theta) = \min\{(\phi(I, I_0^i) - k_i\theta), i = 1, ..., n\}$;

If $(\phi(I, I_0^m) \le k_m\theta)$ **then**

$V_m = V_m \cup I, k_m\theta, \phi(I, I_0^m)$;

Else

$k_I = [(\phi(I, I_0^m) - k_m\theta)/\theta]$;

If $(k_I > 0)$ **then**

$I_0^{n+1} = I; r = k_I\theta$;

Creating cluster $C_{n+1} = I_0^{n+1}, r, \phi = 0$;

$V = V \cup C_{n+1}$;

$E = E \cup \{C_{n+1}, C_i | \phi(I_0^{n+1}, I_0^i) \le k\theta, i = 1, ..., n\}$;

$n = n + 1$;

Else

$\phi(I, I_0^m) = k_m\theta$;

$C_m = C_m \cup I, k_m\theta, \phi(I, I_0^m)$;

End If

End If

End If

End For

End

The cost for the algorithm of creating clustering graph is $O(m \times n)$, with n as a number of binary signatures and m as a number of clusters. According to the experiment, if we create the clustering graph on COREL dataset, then $n = 1,000$ and $m = 12$; if we experiment upon CBIR images, then $n = 1,344$ and $m = 14$; if we experiment upon WANG image dataset, then $n = 10,800$ and $m = 42$; if we experiment upon MSRDI image dataset, then $n = 15,270$ and $m = 24$; if we experiment upon ImageCLEF dataset, then $n = 20,000$ and $m = 8$. According to the experiment, the average value of m is smaller than the value Ω. This shows that the clustering process reduces considerably the search space, which help quick query the similarity images.

The process of clustering can lead to unequal distribution about the number of images in each cluster. So we need to split the cluster that has a large radius. Let $\zeta = \{\zeta_1, \zeta_2, ..., \zeta_m | \zeta_i \in [0, 1]\}$ be a scale sequence that is ordered descending. The split algorithm of a cluster $V = \{I | \phi(I_0, J_i) \le k\theta, J_i \in \Im, i = 1, ..., n, k \in N^*\}$ is done as follows:

Algorithm 4. *Splitting cluster*

Input:$\zeta = \{\zeta_1, ..., \zeta_m | \zeta_i \in [0, 1]\}$, $V = \{I | \phi(I_0, J_i) \le k\theta, J_i \in \Im, k \in N^*\}$

Output: A set of cluster C

Begin

Step 1: Initializing $C = \varnothing$;

Step 2: Setting $\zeta_{\max} = \max\{\zeta_i \in \zeta\}$;

Step 3: Choosing $I_M \in V$ C, so that $d = k\theta - \phi(I_0, I_M) \ge k\zeta_{\max}\theta$ and $\phi(I_M, I_m) > k\zeta_{\max}\theta$ where I_m is a center of cluster of $C(m = 1, ..., M-1)$;

If exist I_M **then**

$V_M = \{I | \phi(I_M, I) \le k\zeta_{\max}\theta\}$; $C = C \cup V_M$; go to *Step 3*;

Else $\zeta = \zeta \backslash \zeta_{\max}$; go to *Step 2*;

End If

Step 4: Each $J \in V$is not distributed yet, we cluster it into C based on the rules of distribution by ***Definition 9***;

End

Theorem 2. *Let* $C = \{V_1, V_2, ..., V_M\}$ be a set of clusters that is split from $V = \{I | \phi(I_0, J_i) \leq k\theta, J_i \in \Im, i = 1, ..., n\}$. We have the results as follows:

1. The set of clusters of C includes the items of V, and this set is non-overlapping, that is, $V_i \cap V_j = \varnothing$, with $i \neq j$ and $i, j \in \{1, 2, ..., M\}$.
2. Each item of V belongs to a unique cluster of C.
3. All of the items in V are distributed into C.

In Figure 5, we describe an illustration about the process of splitting cluster.

Proof:

1. Let V_i, V_j be two arbitrary clusters in C and they have two centres I_i, I_j, respectively. Because the condition of the centres must satisfy $d = k\theta - \phi(I_0, I_i) \geq k\zeta_{\max}\theta$, so both of I_i and I_j belong to V. . On the other hand, radius of cluster is $k\zeta_{\max}\theta$, then the clusters V_i, V_j must belong to V. Moreover, the distance of two centres I_i and I_j is $\phi(I_i, I_j) > k\zeta_{\max}\theta$, so the clusters V_i, V_j are non-overlapping, that is,$V_i \cap V_j = \varnothing$.
2. Because $C = \{V_1, V_2, ..., V_M\}$ is non-overlapping, so each item of V belongs to a unique cluster of C.
3. Assume that $\exists I \in V$, which is not distributed yet, then the ***Algorithm*** 4 will cluster based on ***Definition*** 9. Therefore, all of the items in V are distributed into C.■

Giving the set of cluster $C = \{V_1, V_2, ..., V_M\}$, in which V_i has a centre I_i. The group algorithm is done as follows:

Algorithm 5. *Group a set of clusters*

Input: $C = \{V_1, V_2, ..., V_M\}$, threshold δ and a scale ζ

Output: A set of centers of clusters V and the group of clusters G

Begin

Step 1: Initializing $G = \varnothing$; $V = \varnothing$;

Step 2: Choosing $V_\alpha, V_\beta \in C$: $\phi(I_\alpha, I_\beta) = \max\{\phi(I_i, I_j) | i \neq j \wedge i, j \in \{1, ..., M\}\}$;

Step 3:

If $\phi(I_\alpha, I_\beta) < \delta$ **then** $G = C$; **return**;

Else $V = \{V_\alpha, V_\beta\}$; $C = C\ V$;

End If

Step 4:

For $V_i \in C$**do**

If $d = \min\{\phi(I_i, J_j)\} \geq \zeta\phi(I_\alpha, I_\beta)$ **then**

$V = V \cup V_i$; $C = C\ V_i$;

End If

End For

$G = \{G_i = V_i | V_i \in V, i = 1, ..., |V|\}$;

Step 5:

For $V_j \in C$ **do**

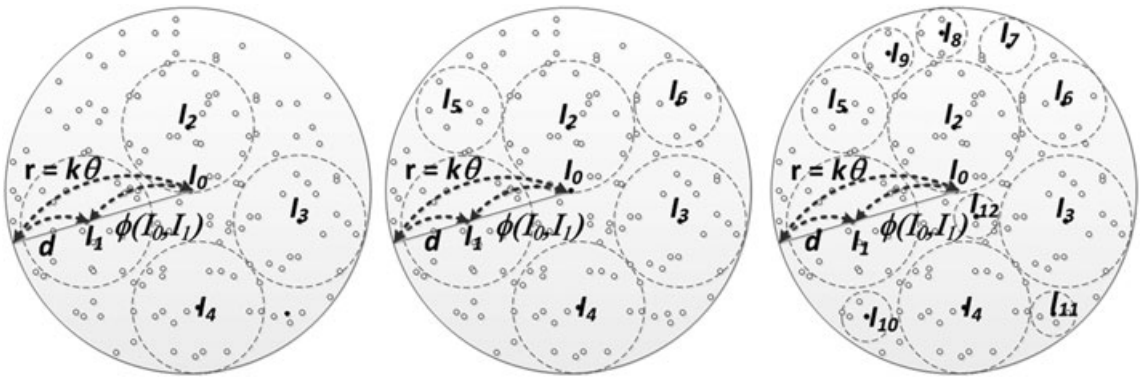

FIGURE 5 An illustration about the process of splitting cluster

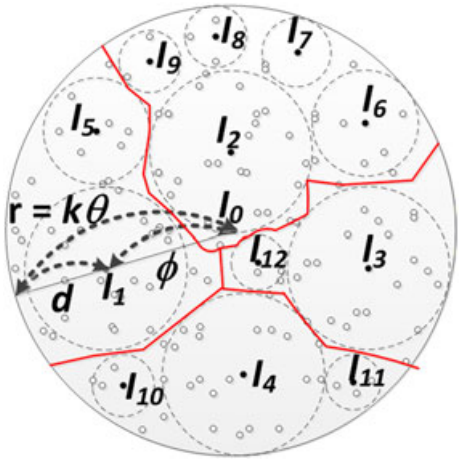

FIGURE 6 A sample about group of clusters

$\phi(l_j, l_m) = \min\{\phi(l_j, l_k)\}$, with $V_k \in V$; $G_m = G_m \cup V_j$;

End For

End

In Figure 6, we give a sample about group of clusters.

5 | IMAGE RETRIEVAL

5.1 | Image retrieval algorithm

After creating cluster graph, the paper performs similarity image retrieval on the basis of this graph. With each query image I_Q, implementation is done to query set of similarity images *IMG*. The query process needs to find out the nearest cluster in graph, which means $\phi_{\min} = \phi(I_Q, I_0^m) = \min\{\phi(I_Q, I_0^i), i = 1, ..., n\}$. Otherwise, we need to do similarity image retrieval at vertexes near to graph that has a distance smaller than given threshold $k\theta$.

Algorithm 6. *Image Retrieval based on Cluster Graph*

Input: Query Image I_Q, cluster graph, threshold value $k\theta$.

Output: A set of similar images *IMG*

Begin

$IMG = \varnothing$; $V = \varnothing$;

Step 1: Searching the nearest cluster

$\phi_{\min} = \phi(I_Q, I_0^m) = \min\{\phi(I_Q, I_0^i), i = 1, ..., n\}$;

$V = V \cup V_m$;

Step 2: Searching neighbor cluster

For ($V_i \in V_{SG}$) **do**

If ($\phi(I_0^m, I_0^i) \le k\theta$) **then**

$V = V \cup V_i$;

End If

End For

Step 3: Searching similarity image set

For ($V_j \in V$) **do**

$IMG = IMG \cup \left\{ I_k^i, I_k^i \in V_j, k = 1, ..., |V_j| \right\}$;

End For

Return *IMG*;

End

5.2 | Experiment model

The model of applying binary signature and clustering in CBIR system is illustrated as in Figure 7. The process of experiment is implemented including two phases. The first phase is preprocessing to create the input data. The second phase executes the process of image retrieval with the query image.

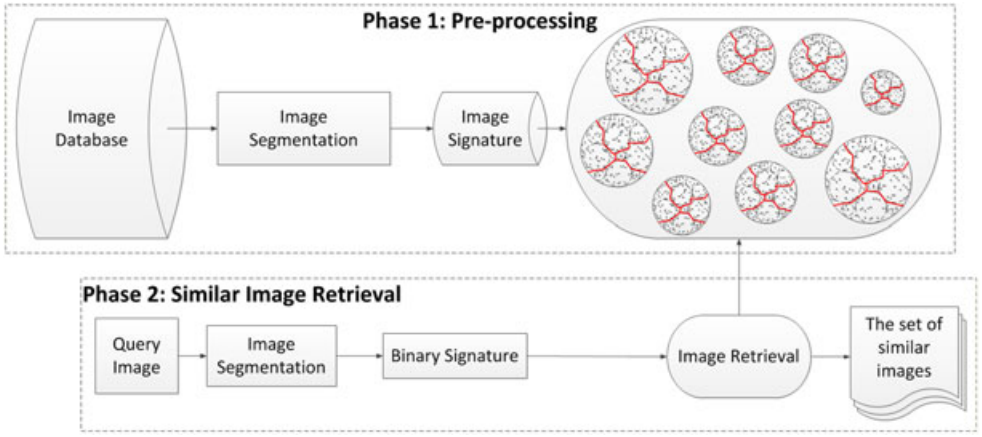

FIGURE 7 The model of image retrieval

The steps of preprocessing include (a) segment images on image dataset; (b) create the binary signature of each segmented image; (c) build the clusters including the binary signatures of similar images.

The steps of image retrieval includes (a) segment the query image; (b) create the binary signature of query image; (c) find the nearest cluster based on group of clusters; (d) sort the similar image's signatures by similarity measure; (e) retrieve the similar images.

5.3 | Experiment results

The application is build on dotNET Framework 3.5, C# programming language. The preprocessing stage is experimented upon the computer with Intel(R) X3440 @ 2.53 GHz × 2, Windows Server 2008 R2 Enterprise 64-bit, RAM 8.00GB. The image retrieval stage is implemented on the computer with Intel(R) CoreTM i7-2620 M, CPU 2.70 GHz, RAM 4 GB, and Windows 7 Professional operating system.

In order to evaluate the performance of the proposed CBIR system, we calculate the values including precision, recall, F-measure, and the true positive rate of receiver operating characteristic (ROC) curve. According to Alzu'bi et al. (2015), precision is the ratio of the number of relevantimages within the first *k*-results to the number of total retrieved images. Recall is the ratio of the number of relevant images within the first *k*-results to the number of total relevant images. F-measure is the harmonic mean of precision and recall. The formulas of these values are defined as follows:

$$precision = \frac{(relevant\ images \cap retrieved\ images)}{retrieved\ images} \tag{9}$$

$$recall = \frac{(relevant\ images \cap retrieved\ images)}{relevant\ images} \tag{10}$$

$$F\text{-}measure = 2\times\frac{(precision \times recall)}{(precision + recall)} \tag{11}$$

The experimental processing is done on image datasets including COREL, CBIR images, Wang, MSRDI, and ImageCLEF in Table 1, we describe these image datasets. For any query image, we retrieve the most similar images on dataset. Then, we compare the list of subjects of images to evaluate the accurate method. Figure 8 shows a result of the proposed retrieval method.

The COREL dataset has 1,000 images and is divided into 10 subjects, each subject has 100 images. Figure 9 describes the precision-recall curve and ROC curve.

TABLE 1 Image datasets

Image dataset	No. images	No. subjects	Size
COREL	1,000	10	30.3 MB
CBIR images	1,344	22	225 MB
WANG	10,800	80	69.2 MB
MSRDI	16,710	31	269 MB
ImageCLEF	20,000	39	1.64 GB

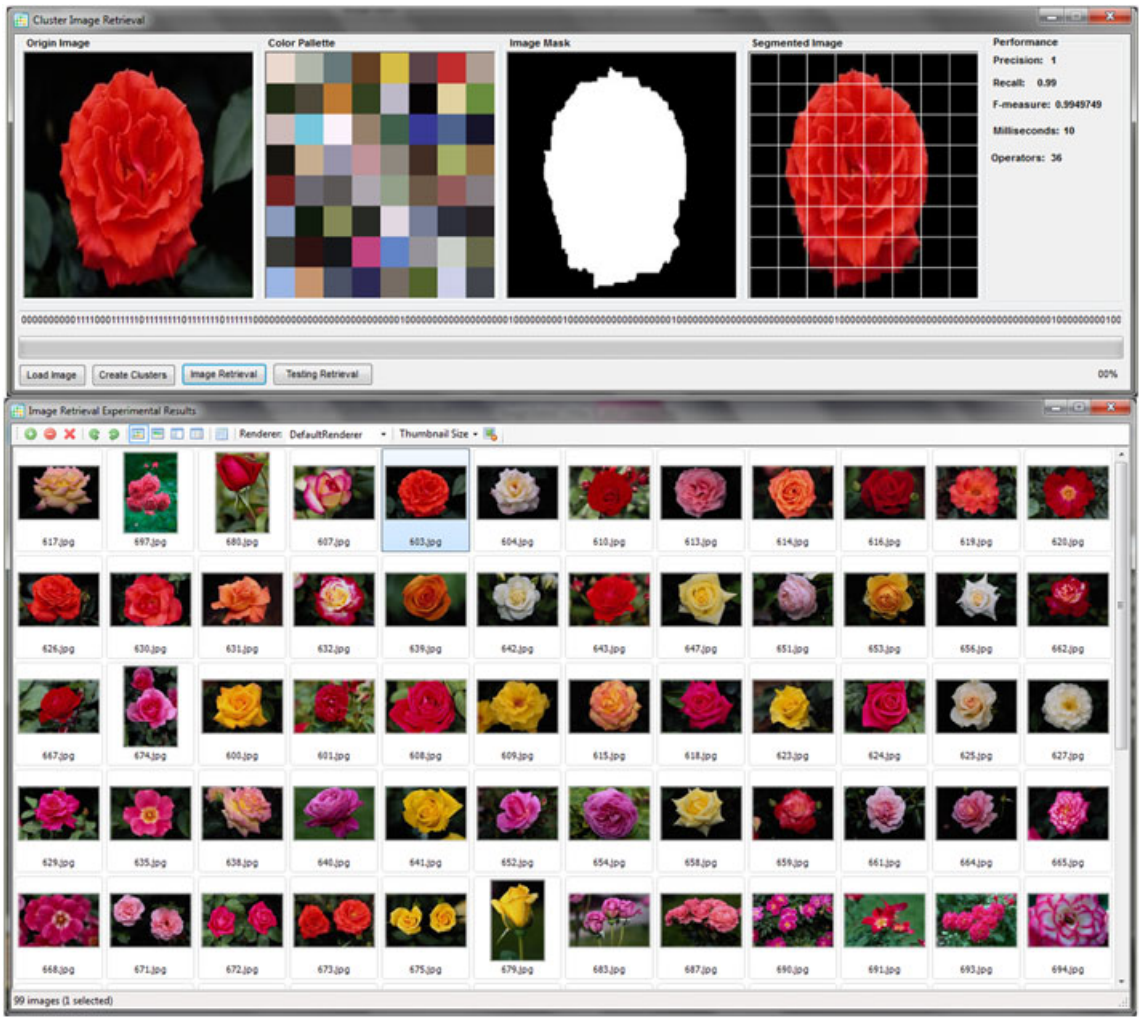

FIGURE 8 A result of image retrieval based on binary signature cluster graph

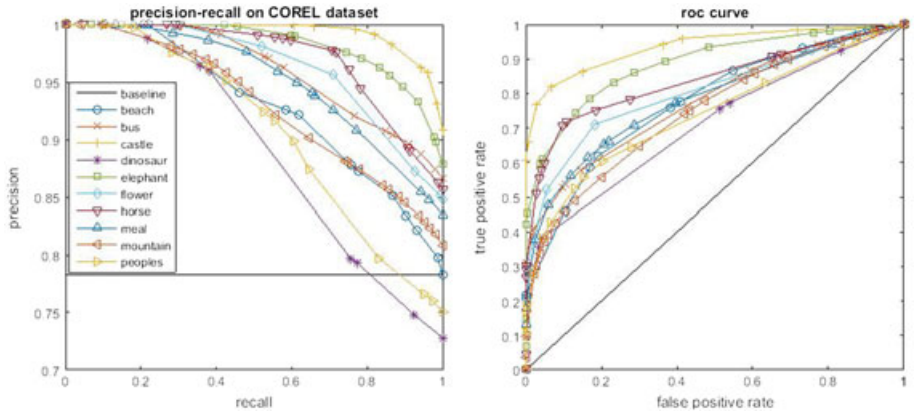

FIGURE 9 Accuracy of system on COREL dataset

The CBIR images dataset (1,344 images) is divided into 22 subjects. The precision-recall curve and ROC curve of query process are described in Figure 10.

The MSRDI dataset (16,710 images) is divided into 31 subjects. The precision-recall curve and ROC curve of query process are described in Figure 11.

The ImageCLEF dataset (20,000 images) is divided into 40 subjects; each subject has from 200 images to 900 images. We use the proposed method to classify automatically into 115 clusters. The precision-recall curve and ROC curve of query process are described in Figure 12. These precision-recall curves show that the proposed method is very effective.

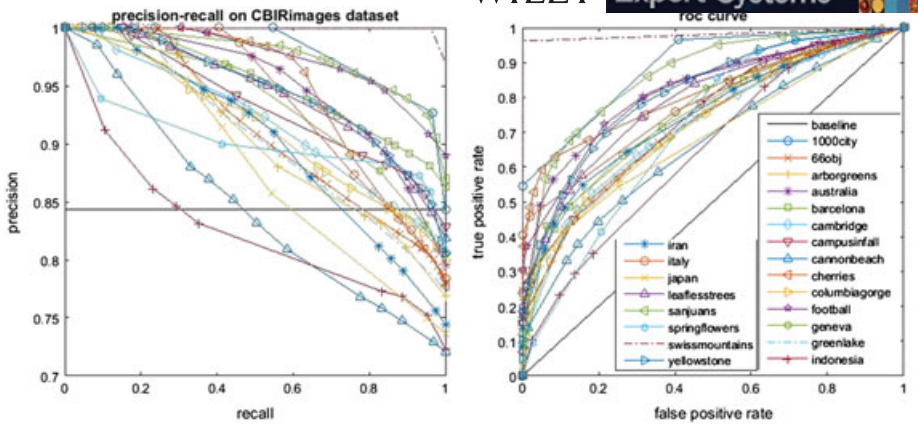

FIGURE 10 Accuracy of system on CBIR images

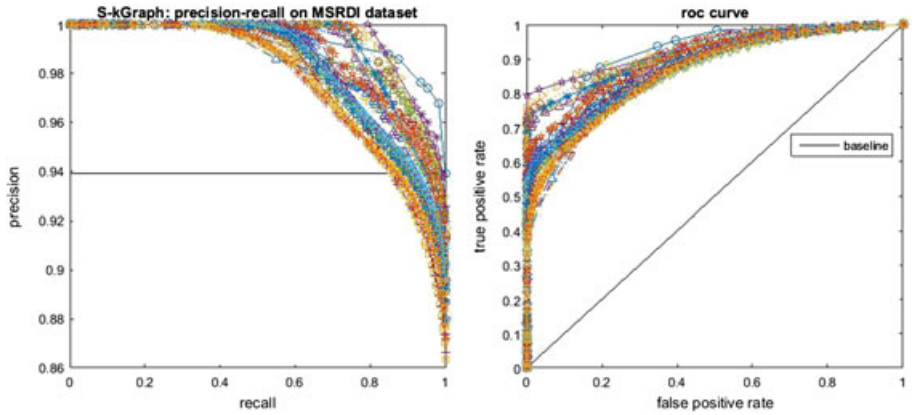

FIGURE 11 Accuracy of system on MSRDI dataset

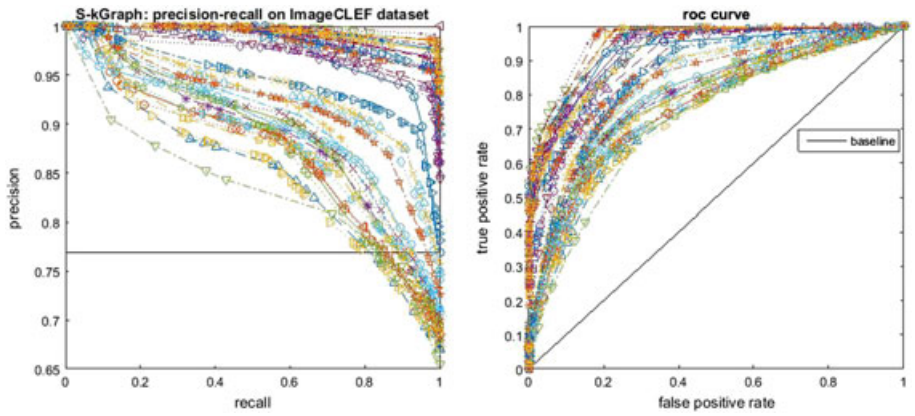

FIGURE 12 Accuracy of system on ImageCLEF dataset

The Wang dataset (10,800 images) is divided into 80 subjects; each subject has from 100 images to 400 images. The precision-recall curve and ROC curve of query process are described in Figure 13.

The figure describing precision-recall corresponds with each image dataset described in Figures 9, 10, 11, 12, and 13. Each figure describes the precision and the recall for subjects in every image dataset. So each subject is described as a curve to assess the precision in image retrieval.

The average retrieval time is described in Figures 14, 15, 16, 17, and 18, where retrieval time of each image subject is measured and its average value is calculated. The values of performance, retrieval time of each subject, and comparative assessment are also presented from Tables 2 to 4.

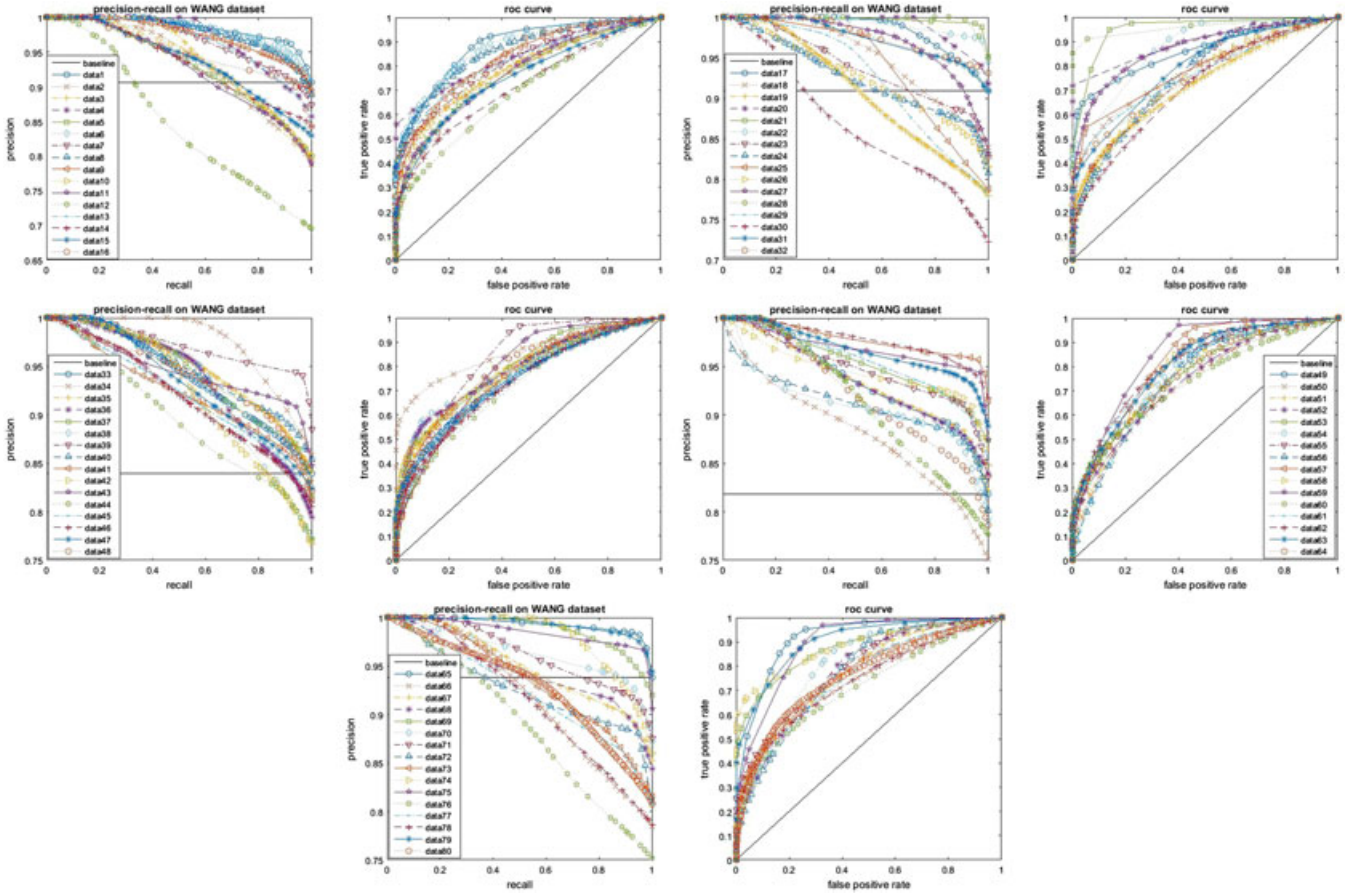

FIGURE 13 Accuracy of system on Wang dataset

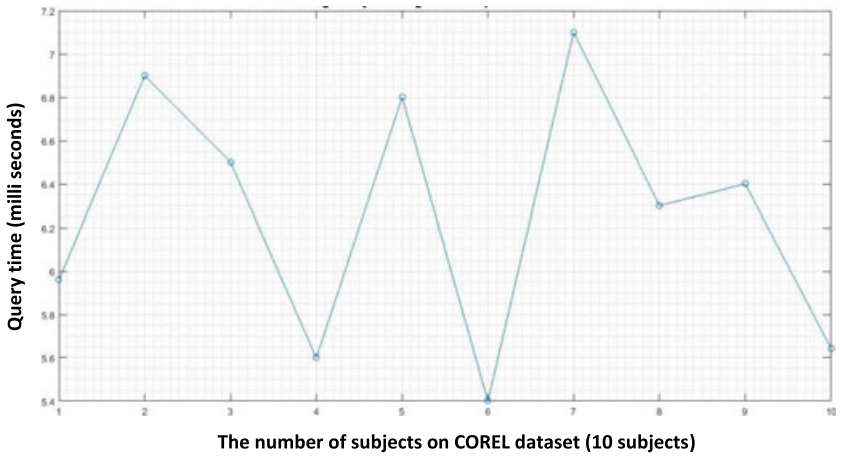

FIGURE 14 Query time on COREL dataset

Experimental application of image retrieval based on cluster graph is described in Figure 8 including two stages as Figure 7. In the first stage, the application performs the process of creating cluster graph from image dataset. In the second stage, it performs image retrieval and prints an output of a set of image similar to query image on the basis of matching binary signature in accordance with similarity measure ϕ on cluster graph. Figure 8 is a result of image retrieval on COREL image dataset, including a set of similar images, which is queried.

Experiment carries out on image datasets including COREL, CBIR images, WANG, MSRDI, and ImageCLEF. With each queried image, the application performs to assess the productivity and the query time. The average query time is described as graph at Figures 14, 15, 16, 17, and 18. The form of the figures shows the variation about query time. The reason is the number of signatures is distributed in clusters unequally, and the number of neighbouring clusters corresponding to each cluster is different. That leads to the retrieval speed corresponding to each image is different.

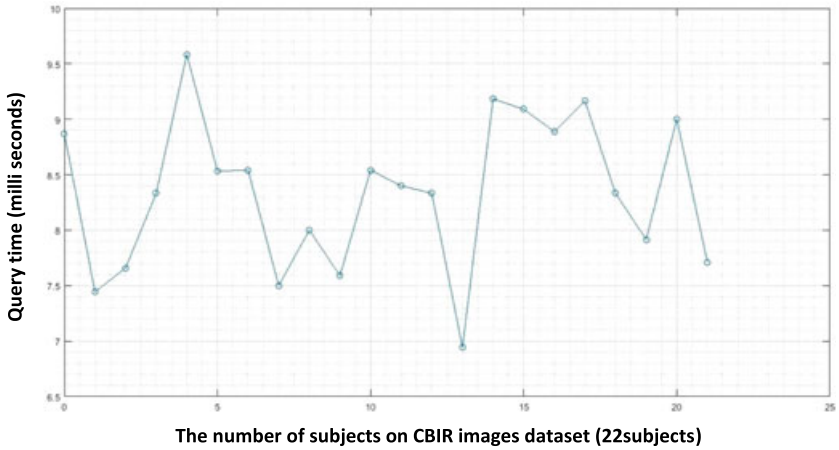

FIGURE 15 Query time on CBIR images dataset

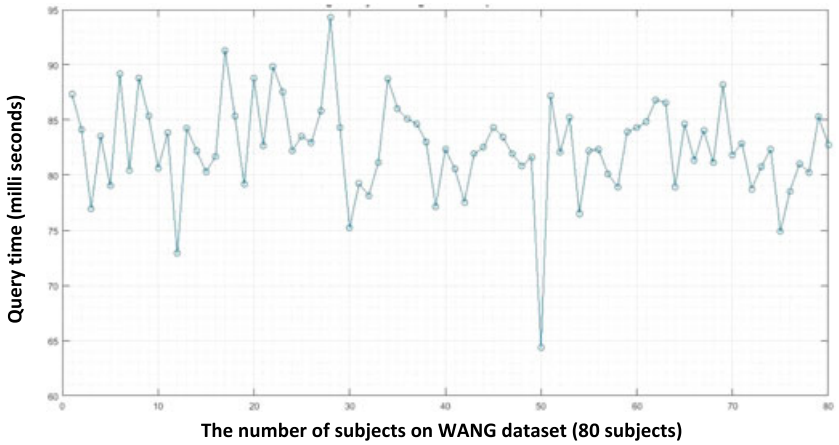

FIGURE 16 Query time on WANG dataset

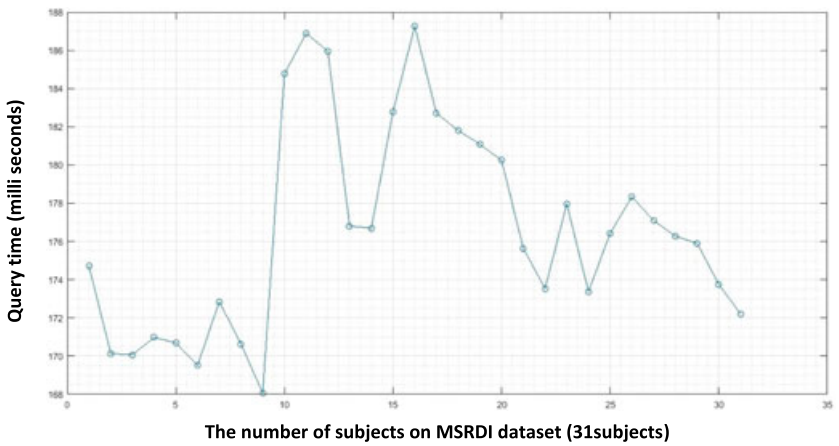

FIGURE 17 Query time on MSRDI dataset

The retrieval productivity is described in Figures 9, 10, 11, 12, and 13; every figure describes the precision of the query in the subject of the image. The values of the precision of each subject are described as a curve on a figure. The shape of curves that shows the image retrieval on the cluster graph has a relative high and effective precision. This demonstrated the correctness of proposed image retrieval method.

On the basis of the values of productivity, Table 2 shows a synthetic of the values of average query productivity for each image dataset. The figures of this table show the proposed retrieval method that have picked up speed and ensured the precision of retrieval process. Tables 3 and 4 include figures of assessment and comparison among methods in order to demonstrate the effectiveness of proposed method.

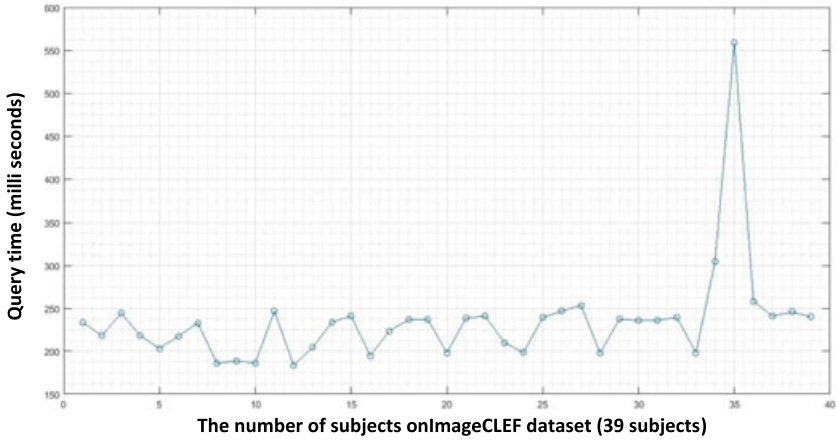

FIGURE 18 Query time on ImageCLEF dataset

TABLE 2 Average accuracy image datasets rely on cluster graph

Dataset	No. images	The time to create cluster (ms)	No. clusters	Average precision	Average recall	Average F-measure	Average query time (ms)
COREL	1,000	3,884.4075	12	0.826185	0.790653	0.808028	6.260010
CBIR images	1,344	3,088.8046	14	0.802329	0.759366	0.780256	8.199417
WANG	10,800	71,884.9236	42	0.832874	0.794125	0.813038	82.637803
MSRDI	15,270	18,4579.5275	24	0.896505	0.86401	0.879957803	176.5775989
ImageCLEF	20,000	698,850.0235	8	0.794522	0.75786	0.775758287	226.641098

TABLE 3 Comparison of average accuracy on COREL dataset

Methods	Beach	Bus	Castle	Dinosaur	Elephant	Flower	Horse	Meal	Mountain	Peoples	Average
M. K. Kundu without RF (Kundu et al., 2015)	0.6	0.69375	0.6125	0.99375	0.8125	0.8625	0.9	0.75625	0.59375	0.7375	0.75625
M. K. Kundu with RF (Kundu et al., 2015)	0.9625	0.9625	0.9375	0.96875	0.98125	0.98125	0.99375	0.996875	0.996875	0.8125	0.959375
Y. Yang et al. (Yang et al., 2012)	0.8375	0.875	0.85	0.8875	0.89375	0.9	0.925	0.9375	0.9375	0.7875	0.883125
Chu-Hui Lee (Lee & Lin, 2010)	0.8	0.825	0.81875	0.8375	0.85	0.875	0.8875	0.9	0.9	0.7625	0.845625
N. Shrivastava (Shrivastava & Tyagi, 2014)	0.582	0.802	0.621	1.000	0.751	0.923	0.896	0.803	0.561	0.748	0.769
ElAlami (ElAlami, 2011)	0.561	0.876	0.571	0.987	0.675	0.914	0.834	0.741	0.536	0.703	0.739
Wang et al. (Wang, Yua, & Yanga, 2011)	0.400	0.500	0.600	0.950	0.600	0.800	0.630	0.400	0.300	0.720	0.590
Muhammad Imran (Imran, Hashim, & Khalid, 2014)	0.62	0.45	0.37	1.00	0.56	0.74	0.69	0.32	0.37	0.77	0.589
Chuen et al. (Lin, Chen, & Chan, 2009)	0.540	0.888	0.562	0.992	0.658	0.891	0.803	0.733	0.522	0.683	0.727
A. Huneiti (Huneiti & Daoud, 2015)	0.5420	0.5260	0.3440	0.5260	0.5560	0.8280	0.7480	0.3040	0.5000	0.2780	0.5588
S. Thirunavukkarasu (Thirunavukkarasu, Ahila Priyadharshini, Arivazhagan, & Mahalakshmi, 2013)	0.3340	0.3220	0.3500	0.3220	0.3840	0.2960	0.3460	0.4000	0.3800	0.2700	0.3386
H. K. Bhuravarjula (Bhuravarjula & Kumar, 2012)	0.2615	0.1725	0.1105	0.1725	0.3490	0.4950	0.2080	0.1560	0.2590	0.1330	0.3109
V. S. Thakare (Thakare & Patil, 2014)	0.9625	0.8750	0.9437	0.9937	0.7563	0.9437	0.8937	0.7625	0.8125	0.9500	0.8894
P. Shrinivasacharya (Shrinivasacharya & Sudhamani, 2013)	0.5550	0.5500	0.4925	0.9330	0.6370	0.7560	0.8710	0.6200	0.3800	0.7433	0.6537
R. Mostafa (Mostafa & Mohsen, 2011)	0.9000	0.9000	0.8500	0.8000	0.6250	0.8125	0.8000	0.6250	0.5750	0.8937	0.7781

(*Continues*)

TABLE 3 (Continued)

Methods	Beach	Bus	Castle	Dinosaur	Elephant	Flower	Horse	Meal	Mountain	Peoples	Average
T. W. S. Chow (Chow, Rahman, & Wu, 2006)	0.3970	0.8173	0.4466	0.9978	0.5185	0.6510	0.9045	0.6696	0.7171	0.8018	0.6696
C.-H. Lin (Lin et al., 2009)	0.5400	0.8880	0.5615	0.9925	0.6580	0.8910	0.8025	0.7325	0.5215	0.6830	0.7270
Jehad Alnihoud (Alnihoud, 2012)	0.68	0.84	0.70	1.00	0.50	0.97	0.85	0.63	0.32	0.87	0.742
Our method	0.7831	0.8662	0.908	0.728	0.878	0.8481	0.8565	0.835	0.808	0.750	0.826

TABLE 4 Comparison of query time on COREL dataset

Methods	Query time
N. Shrivastava (Shrivastava & Tyagi, 2014)	1.2 s
ElAlami (ElAlami, 2011)	0.6 s
Wang et al. (Wang et al., 2011)	3.5 s
M. K. Kundu (Kundu et al., 2015)	0.434375 s
Y. Yang et al. (Yang et al., 2012)	0.58875 s
Chu-Hui Lee (Lee & Lin, 2010)	0.3284375 s
Our method	6.260010 ms

6 | CONCLUSION AND FUTURE WORK

The paper presents the method of creating binary signature based on colour and shape of interest item on image. On the basis of this binary signature, the paper proposes the method of clustering binary signature in order to build the CBIR. To improve clustering process, the paper designs the cluster graph to improve the speed of clustering as well as to build the CBIR system. On the basis of proposed theory, the paper builds image retrieval systems based on binary signature. According to the experimental results, the method of clustering binary signature creates initial cluster at great cost. Moreover, the figures in comparison with the other methods show the proposed method of building image retrieval systems effectively, that is, query time is quick and precision is high. This proved correctness in accordance with experiment of proposed methods. From that, the method of image retrieval according to content based on binary signature is an effective solution to build an image retrieval system that meets the requirements of user. In the next development, the paper will build a network structure self-organizing map in order to design structure of cluster graph, which describes relationship among images at the same time performs image retrieval based on interest vectors of winner cluster on network structure self-organizing map.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous reviewers for their helpful comments and valuable suggestions. We would also like to thank the Faculty of Information Technology, University of Sciences - Hue University, Vietnam, and the Center for Information Technology, HCMC University of Food Industry, Vietnam.

REFERENCES

Abdesselam, A., Wang, H. H., & Kulathuramaiyer, N. (2010). Spiral bit-string representation of color for image retrieval. *The International Arab Journal of Information Technology*, *7*(3), 223–230.

Acharya, T., & Ray, A. K. (2005). *Image processing: Principles and applications*. Hoboken, New Jersey: John Wiley & Sons Inc. Publishers.

ACI (2015). "http://www.aci.aero/."

Ahmad, I., & Grosky, W. I. (2003). Indexing and retrieval of images by spatial constraints. *J. Vis. Commun. Image R.*, *14*, 291–320.

Alnihoud, J. (2012). Content-based image retrieval system based on self organizing map, fuzzy color histogram and subtractive fuzzy clustering. *The International Arab Journal of Information Technology*, *9*(5), 452–458.

Alzu'bi, A., Amira, A., & Ramzan, N. (2015). Semantic content-based image retrieval: A comprehensive study. *Journal of Visual Communication and Image Representation*, *32*, 20–54.

An, Y., Baek, J., Shin, S., Chang, M., & Park, J. (2008). Classification of feature set using K-means clustering from histogram refinement method. Fourth International Networked Computing and Advanced Information Management (NCM '08). Gyeongju, IEEE.

Banerjee, M., Bandyopadhyay, S., & Pal, S. K. (2013). A clustering approach to image retrieval using range based query and Mahalanobis distance. In A. Skowron, & Z. Suraj (Eds.), *Rough sets and intelligent systems* (Vol. 2) (pp. 79–91). New York Dordrecht London: Springer Berlin Heidelberg.

Bartolini, I., Ciaccia, P., & Patella, M. (2010). Query processing issues in region-based image databases. *Springer-Verlag, Knowl Inf Syst*, *25*, 389–420.

Bhanu, B., & Dong, A. (2002). Concepts learning with fuzzy clustering and relevance feedback. *Engineering Applications of Artificial Intelligence*, *15*(2), 123–138.

Bhuravarjula, H. K., & Kumar, V. N. S. V. (2012). A novel content based image retrieval using variance color moment. *International Journal of Computational Engineering Research*, *1*, 93–99.

Cai, J., Liu, Q., Chen, F., Joshi, D., & Tian, Q. (2014). Scalable image search with multiple index tables. Proceedings of International Conference on Multimedia Retrieval, ACM.

Chappell, T., & Geva, S. (2013). Efficient top-K retrieval with signatures. In *ADCS '13* (pp. 10–17). Brisbane, QLD: Australia, ACM.

Chen, Y., Wang, J. Z., & Krovetz, R. (2005). CLUE: Cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing*, *14*(8), 1187–1201.

Chitkara, V., Nascimento, M. A., & Mastaller, C. (2000). *Content-based image retrieval using binary signatures*. Edmonton, Alberta, Canada: Department Of Computing Science, University of Alberta.

Chow, T. W. S., Rahman, M. K. M., & Wu, S. (2006). Content-based image retrieval by using tree-structured features and multi-layer self-organizing map. *Pattern Analysis and Applications*, *9*(1), 1–20.

Deligiannidis, L., & Arabnia, H. R. (2015). *Emerging trends in image processing, computer vision, and pattern recognition*. Waltham, MA 02451, USA: Elsevier Morgan Kaufmann.

Demirci, M. F. (2012). Graph-based shape indexing. *Machine Vision and Applications*, *23*(3), 541–555.

ElAlami, M. E. (2011). A novel image retrieval model based on the most relevant features. *Knowledge-Based Systems*, *24*(1), 23–32.

El-Kwae, E. A. (2000). Signature-based indexing for retrieval by spatial content in large 2D-string image databases. 12th International Symposium, Charlotte, NC, USA, Springer Berlin Heidelberg.

El-Kwae, E. A., & Kabuka, M. R. (2000). Efficient content-based indexing of large image databases. *ACM Transactions on Information Systems*, *18*(2), 171–210.

Hlaoui, A., & Wang, S.-R. (2003). A graph clustering algorithm with applications to content-based image retrieval. International Conference on Machine Learning and Cybernetics, IEEE.

Huang, Y., Zhang, J., Zhao, X., & Ma, D. (2010). Medical image retrieval with query-dependent feature fusion based on one-class SVM. In *Computational science and engineering (CSE)* (pp. 176–183). Hong Kong: IEEE Xplore.

Huneiti, A., & Daoud, M. (2015). Content-based image retrieval using SOM and DWT. *Journal of Software Engineering and Applications*, *8*(2), 51–61.

IDC (2016). "https://www.idc.com."

Imran, M., Hashim, R., & Khalid, N. E. A. (2014). Content based image retrieval using MPEG-7 and histogram. The First International Conference on Soft Computing and Data Mining (SCDM), Universiti Tun Hussein Onn Malaysia, Johor, Malaysia, Springer International Publishing.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Jin, L., Hong, L., & Lianzhi, T. (2009). A mapping modelling of visual feature and knowledge representation approach for medical image retrieval. International Conference on Mechatronics and Automation (ICMA 2009) Changchun, IEEE: 1778–1783.

Kumar, H. C. S., Raja, K. B., Venugopal, K. R., & Patnaik, L. M. (2009). Automatic image segmentation using wavelets. *International Journal of Computer Science and Network Security*, *9*(2), 305–313.

Kundu, M. K., Chowdhury, M., & Bulò, S. R. (2015). A graph-based relevance feedback mechanism in content-based image retrieval. *Knowledge-Based Systems*, *73*, 254–264.

Landre, J., & Truchetet, F. (2007). Fast image retrieval using hierarchical binary signatures. 9th International Symposium on Signal Processing and Its Applications, Sharjah, IEEE.

Le, T. M., & Van, T. T. (2013). Image retrieval system based on EMD similarity measure and S-Tree. In J. Juang, & Y.-C. Huang (Eds.), *Intelligent technologies and engineering systems* (Vol. 234) (pp. 139–146). Taiwan: Springer Verlag, Lecture Notes in Electrical Engineering.

Lee, C.-H., & Lin, M.-F. (2010). Ego-similarity measurement for relevance feedback. *Expert Systems with Applications*, *37*(1), 871–877.

Li, C.-H., & Lu, Z.-M. (2011). Graph-based features for image retrieval. Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), Dalian, IEEE.

Li, J., Zhang, M., Pan, P., Han, Q., & Feng, X. (2012). Graph-based medical image clustering. International Conference on Computing and Networking Technology (ICCNT), Gueongju, IEEE.

Lin, C.-H., Chen, R.-T., & Chan, Y-K. (2009). A smart content-based image retrieval system based on color and texture feature. *Image and Vision Computing*, *27*(6), 658–665.

Lin, C.-H., Chen, C-C, Lee, H.-L, & Liao, J.-R. (2014). Fast K-means algorithm based on a level histogram for image retrieval. *Expert Systems with Applications*, *41*(7), 3276–3283.

Liu, L., Lu, Y., & Suen, C. Y. (2015). Variable-length signature for near-duplicate image matching. *IEEE Transactions on Image Processing*, *24*(4), 1282–1296.

Manolopoulos, Y., Nanopoulos, A., & Tousidou, E. (2003). *Advanced signature indexing for multimedia and web applications*. Springer Science+Business Media New York: Kluwer Academic Publishers.

Marques, O., & Furht, B. (2002). *Content-based image and video retrieval*. Springer Science+Business Media New York: Kluwer Academic Publishers.

Mostafa, R., & Mohsen, E. M. (2011). A texture based image retrieval approach using self-organizing map pre-classification IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, IEEE.

Muneesawang, P., Zhang, N., & Guan, L. (2014). *Multimedia database retrieval: Technology and applications*. Springer Cham Heidelberg New York Dordrecht London: Springer International Publishing Switzerland.

Nascimento, M. A., & Chitkara, V. (2002). Color-based image retrieval using binary signatures. In *SAC 2002* (pp. 687–692). Madrid, Spain: ACM.

Nascimento, M. A., Tousidou, E., Chitkara, V., & Manolopoulos, Y. (2002). Image indexing and retrieval using signature trees. *Data & Knowledge Engineering*, *43*, 57–77.

Özkan, S., Esen, E., & Akar, G. B. (2014). Visual group binary signature for video copy detection. International Conference on Pattern Recognition, Stockholm.

Prasad, B. G., Biswas, K. K., & Gupta, S. K. (2004). Region-based image retrieval using integrated color, shape, and location index. *Computer Vision and Image Understanding*, *94*, 193–233.

Rajakumar, K., & Muttan, S. (2010). Medical image retrieval using energy efficient wavelet transform. International Conference on Computing Communication and Networking Technologies (ICCCNT). Karur, IEEE: 1–5.

Ren, G., Cai, J., Li, S., Yu, N., & Tian, Q. (2014). Scalable image search with reliable binary code. Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, ACM.

Saboorian, M. M., Jamzad, M, & Rabiee, H. R. (2010). User adaptive clustering for large image databases. International Conference on Pattern Recognition (ICPR), Istanbul, IEEE.

Shambharkar, S., & Tirpude, S. (2011). Fuzzy C-means clustering for content based image retrieval system International Conference on Advancements in Information Technology, Singapore, IACSIT Press.

Shea, G. Y. K., & Cao, J. (2012). Geo-Planar Indexing (GPI) - An efficient indexing scheme for fast retrieval of raster-based geospatial data in mobile GIS applications. In *International Congress on Image and Signal Processing (CISP)* (pp. 1047–1052). Chongqing: IEEE.

Shrinivasacharya, P., & Sudhamani, M. V. (2013). Content based image retrieval using self organizing map. the Fourth International Conference on Signal and Image Processing (ICSIP), Coimbatore, India, Springer India.

Shrivastava, N., & Tyagi, V. (2014). An efficient technique for retrieval of color images in large databases. *Computers & Electrical Engineering*, *46*, 314–327.

Snášel, V. (2000). Fuzzy Signatures for Multimedia Databases. In *Advances in Information Systems, ADVIS 2000*. Izmir, Turkey: Springer Berlin Heidelberg.

Tang, Z., Zhang, X., Dai, X., Yang, J., & Wu, T. (2013). Robust image hash function using local color features. *International Journal of Electronics and Communications*, *67*, 717–722.

Thakare, V. S., & Patil, N. N. (2014). Image texture classification and retrieval using self-organizing map. International Conference on Information Systems and Computer Networks (ISCON), Mathura, IEEE.

Thirunavukkarasu, S., Ahila Priyadharshini, R., Arivazhagan, S., & Mahalakshmi, C. (2013). Content based image retrieval based on dual tree discrete wavelet transform. *International Journal of Research in Computer and Communication Technology*, *1*, 473–477.

Unser, M. (1995). Texture classification and segmentation using wavelet frames. *IEEE Transactions on Image Processing*, *4*(11), 1549–1560.

Van, T. T., & Le, T. M. (2014a). Image retrieval based on binary signature and S-kGraph. *Jour. of Annales Univ. Sci. Budapest., Sect. Comp., 43*, 105–122.

Van, T. T., & Le, T. M. (2014b). RBIR based on signature graph. International Conference on Computer Communication and Informatics. Coimbatore, India, IEEE Xplore.

Van, T. T., & Le, T. M. (2014c). RBIR using interest regions and binary signatures. *Journal of Annales Univ. Sci. Budapest., Sect. Comp., 43*, 105–122.

Van, T. T., & Le, T. M. (2016). Clustering binary signature applied in content-based image retrieval. World Conference on Information Systems and Technologies (WorldCist'16). Recife, PE, Brazil, Springer.

Wang, F., Lu, Y, Zhang, F., & Sun, S. (2013). *A new method based on fuzzy C-means algorithm for search results clustering ISCTCS*. Beijing, China: Springer-Verlag Berlin Heidelberg.

Wang, J. Z. (2001). *Integrated region-based image retrieval*. Springer Science Business Media New York: Kluwer Academic Publishers.

Wang, X.-Y., Wu, J.-F., & Yang, H.-Y. (2010). Robust image retrieval based on color histogram of local feature regions. *Springer Science, Multimed Tools Appl*, *49*, 323–345.

Wang, X.-Y., Yua, Y-J., & Yanga, H.-Y. (2011). An effective image retrieval scheme using color, texture and shape features. *Computer Standards & Interfaces*, *33*(1), 59–68.

Wang X.-Y., Yang H.-Y., Li Y.-W., & Yang F. (2013). Robust color image retrieval using visual interest point feature of significant bit-planes. *Digital Signal Processing*, *23*(4), 1136–1153.

Wengert, C., Douze, M., & Jégou, H. (2011). Bag-of-colors for Improved image search. Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, Arizona, USA, ACM.

Xu, B., Bu, J., Wang, C., & He, X. (2015). EMR: A scalable graph-based ranking model for content-based image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *27*(1), 102–114.

Yan, Y., Liu, G., Wang, S., Zhang, J., & Zheng, K. (2014). Graph-based clustering and ranking for diversified image search. *Multimedia Systems*, (Special Issue Paper), 1–12.

Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., & Pan, Y. (2012). A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 723–742.

Zakariya, S. M., Ali, R., & Ahmad, N. (2010). Combining visual features of an image at different precision value of unsupervised content based image retrieval. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, IEEE.

Zhao, N., Dong, Y., Bai, H., Wang, L., Huang, C., Cen, S., & Zhao, J. (2013). A semantic graph-based algorithm for image search reranking. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, IEEE.

Zhou, W., Li, H., Lu, Y., & Tian, Q. (2011). Large scale image search with geometric coding. Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, Arizona, USA, ACM.

Zhou, W., Li, H., Wang, M., Lu, Y., & Tian, Q. (2012). Binary SIFT: Towards efficient feature matching verification for image search Proceedings of the 4th International Conference on Internet Multimedia Computing and Service, Wuhan, Hubei, China, ACM.

Zhou, W., Lu, Y., Li, H., & Tian, Q. (2012). Scalar quantization for large scale image search. Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, ACM.

Zhou, W., Li, H., Lu, Y., & Tian, Q. (2013). SIFT match verification by geometric coding for large-scale partial-duplicate web image search. *ACM Transactions on Multimedia Computing, Communications and Applications*, *9*(1), 1–18.

Zhou, W., Li, H., & Tian, Q. (2014). Chapter 12–Multimedia content-based visual retrieval. In *Academic Press Library in signal Processing Image and Video Compression and Multimedia* (Vol. 5) (pp. 383–416). Waltham, MA, USA: Elsevier.

Zhou, W., Li, H., & Lu, Y. (2015). Visual word expansion and BSIFT verification for large-scale image search. *Multimedia Systems*, *21*(3), 245–254.

Zhuang, D., Zhang, D., & Li, J. (2013). Improved binary feature matching through fusion of hamming distance and fragile bit weight. Proceedings of the 3rd ACM international workshop on Interactive multimedia on mobile & portable devices, Barcelona, Spain, ACM.

Thanh The Van was born in 1979. He received the BSc degree in mathematics and computer science from the University of Science/HCMC National University, Vietnam, in 2001. In 2008, he obtained an MSc degree in computer science from Vietnam National University. Since 2012, he has been a PhD candidate at Hue University, Vietnam. His research interests include image processing and image retrieval.

Thanh Manh Le was born in 1953. He received a PhD degree in computer science from Budapest University (ELTE), Hungary, in 1994. He became an associate professor at Hue University, Vietnam, in 2004. His research interests include databases, knowledge bases and logic programming.

How to cite this article: Van TT, Le TM. Content-based image retrieval based on binary signatures cluster graph. *Expert Systems*. 2017; e12220. https://doi.org/10.1111/exsy.12220