Concentric Circle-Based Image Signature for Near-Duplicate Detection in Large Databases

Ayoung Cho, Won-Keun Yang, Weon-Geun Oh, and Dong-Seok Jeong

Many applications dealing with image management need a technique for removing duplicate images or for grouping related (near-duplicate) images in a database. This paper proposes a concentric circle-based image signature which makes it possible to detect near-duplicates rapidly and accurately. An image is partitioned by radius and angle levels from the center of the image. Feature values are calculated using the average or variation between the partitioned sub-regions. The feature values distributed in sequence are formed into an image signature by hash generation. The hashing facilitates storage space reduction and fast matching. The performance was evaluated through discriminability and robustness tests. Using these tests, the particularity among the different images and the invariability among the modified images are verified, respectively. In addition, we also measured the discriminability and robustness by the distribution analysis of the hashed bits. The proposed method is robust to various modifications, as shown by its average detection rate of 98.99%. The experimental results showed that the proposed method is suitable for near-duplicate detection in large databases.

Keywords: Content-based image signature, duplicate detection, image partitioning, hash generation.

Manuscript Oct. 22, 2009; revised June 8, 2010, accepted June 24, 2010.

This work was supported by the IT R&D program of MKE/MSCT/IITA, Rep. of Korea [2010-S-024-01, Development of the Rich UCC Technology].

Ayoung Cho (phone: +82 32 860 7415, email: ayoung@inhaian.net), Won-Keun Yang (email: aida@inhaian.net), and Dong-Seok Jeong (email: dsjeong@inha.ac.kr) are with the Department of Electronic Engineering, Inha University, Incheon, Rep. of Korea.

Weon-Geun Oh (email: owg@etri.re.kr) is with the Contents Research Laboratory, ETRI, Daejeon, Rep. of Korea.

doi:10.4218/etrij.10.0109.0623

I. Introduction

Multimedia contents are being actively exchanged due to the growth of the internet. The widespread use of such content is also affected by the supply of portable devices, such as cameras, media players, and flash memory. Web users readily share various digital contents on the internet. They sometimes download and keep such content individually or upload their own content onto the web. Content is frequently exchanged between online and offline environments, and is increasing exponentially. In addition, multimedia content can be modified to adapt to digital device settings or to the particular objective of the user. The content can be rotated, resized, or changed to allow for enhancement by an editing program, and then this modified content is released again. The modified content imported into a database becomes the near-duplicate of the existing content in the database. Consequently, there exist a number of near-duplicates of the same content in the database.

This new database paradigm demands a simple and fast method of handling a large number of contents. This method should include the detection of near-duplicates for the efficiency and accuracy of content handling. There is a particularly strong need for near-duplicate detection in certain applications, such as digital content management and copyright protection. Digital content management is composed of several procedures: support of the content organization, search of archived content, removal of repeated data, and other processes. That is, the primary task of content management is to cluster related versions or eliminate identical content in a database. Copyright protection is conducted by near-duplicate detection. A pirate intentionally transforms copyrighted content to avoid exposure to a digital rights management (DRM) system. Although this content is modified by transformation, the DRM

system should track down the pirate who illegally uses the copyrighted content.

Content-based methods are proposed as solutions to image search and illegal content detection. Content-based methods extract a signature from the content itself without any additional information. In content-based methods of searching for digital content, there are two distinct applications; namely, content-based retrieval and content-based copy detection. Some papers have dealt with these two content-based methods without differentiation. Other papers have considered that content-based copy detection is a subset of content-based retrieval. These different treatments bring about confusion. Content-based retrieval and content-based copy detection are not exactly identical. Content-based retrieval targets similar images, while content-based copy detection targets near-duplicates differentiated by modifications. The most representative algorithms for content-based image retrieval are MPEG-7 visual descriptors, which utilize such features as color, texture, and shape. The MPEG-7 descriptors were used for copy detection in [1]. The edge histogram descriptor [2] shows comparatively good performance, but the MPEG-7 descriptors are not sufficient to detect near-duplicates. We need a distinct content-based copy detection method for near-duplicate detection. The final feature extracted by a content-based copy detection method is usually referred to as a signature instead of a descriptor.

Content-based copy detection methods should satisfy the requirements of efficiency, accuracy, and scalability. These abilities mean fast execution time, secure detection of modified contents, and similar performance regardless of the size of the database. To evaluate these requirements, many researchers used recall-precision [3]-[8] or the average normalized modified retrieval rank (ANMRR) of MPEG-7 [6], [9]. Other papers evaluated their methods in terms of their robustness to various modifications and discriminability between different images [4], [6], [8].

In this paper, we target near-duplicate image detection. The purpose of the proposed algorithm is to detect various modified images in a large database. We reform the signature extraction process on the basis of a concentric circle-based signature [10]. We evaluate the performance in terms of robustness, discriminability and the ANMRR values. The performance of this algorithm is also investigated by analyzing the signature components in large image sets. The remainder of this paper is organized as follows. Section II presents the existing copy detection algorithms. Section III discusses previous work and the details of the proposed algorithm. Section IV provides experimental results for the proposed algorithm and other algorithms with the test conditions. Analysis and discussions follow in section V. Finally, our conclusions are drawn in section VI.

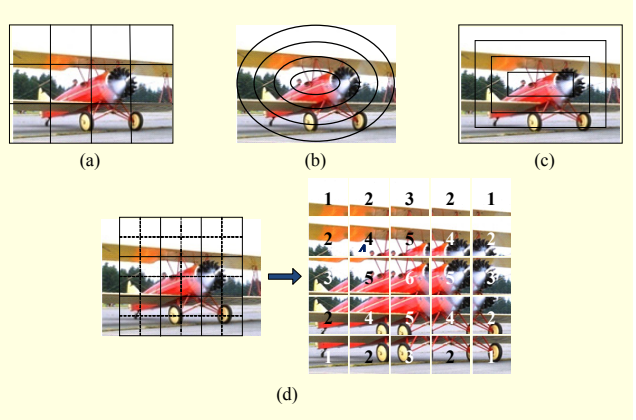

Fig. 1. Image partitioning methods: (a) non-overlapping blocks, (b) elliptical tracks, (c) rectangle rings, and (d) half-overlapping blocks and clustering.

II. Related Work

Content-based copy detection systems search for duplicates or near-duplicates of the query image. The first paper on content-based copy detection, the replicated image detector (RIME), was published in 1998 [11]. RIME was suggested as an alternative approach to watermarking for copyright protection. RIME extracts two types of feature vectors using wavelet coefficients in three color channels. The first one is made of high-frequency coefficients and serves as a shape filter, and the second one is made of low-frequency coefficients and acts as a color filter. Replicated image detection in RIME performs a coarse-to-fine matching procedure with the feature vectors. However, this method is only robust to slight modifications.

Most algorithms apply an image partitioning method in order to use the position information in feature extraction. Kim [12] proposed a DCT-based ordinal measure for content-based copy detection. The input gray image is divided into 8×8 non-overlapping blocks (Fig. 1(a)) and each block value takes an average intensity. After applying the 8×8 two-dimensional DCT, a rank matrix is generated in descending order of the AC magnitudes as the image signature. As the DCT-based ordinal measure is simple and robust, several papers have referred to it for the development of new techniques [4], [6], [7], [13]. The DCT-based ordinal measure was also utilized as a comparative method in other papers [14]-[16]. However, the DCT-based ordinal measure is weak when it comes to detecting near-duplicates rotated by an arbitrary angle.

Wu and others [16] proposed a method which uses elliptical track division (Fig. 1(b)) to overcome this weakness concerning rotation. However, this method cannot handle rotations of 22.5° and 45°. Lin and others [15] proposed an

edge-based image signature with image division of a rectangular ring shape (Fig. 1(c)). The edge information used in the image signature is damaged by modifications, such as blurring and the addition of noise. It is difficult to distinguish the images clearly, even though they are different.

Another trend is the use of hash generation in signature extraction [3], [8], [17]. It is beneficial to apply hashing to a large database since a hashed signature has a small signature size and uses a simple similarity measure. Lu and others [8] proposed a mesh-based hash algorithm for error-resilient and fast matching. The extraction method consists of Harris detection, mesh generation, and hash extraction. It is difficult to guarantee the repeatability from various modified images in the preceding two steps. That is, the accuracy of the mesh-based hash algorithm is reduced despite the advantages of hash generation. Wnukowicz and others [17] proposed a trajectory signature which is applied to hash generation. The input image is normalized and divided into half-overlapping blocks. These blocks are grouped by their distance from the center of the image (Fig. 1(d)). The trajectory signature method extracts the group features, and then the feature values are converted into hashed bits by hash generation. The trajectory signature is robust to various modifications, including image compression, noise addition, and basic rotations of 90°, 180°, and 270°.

III. Concentric Circle-Based Signature

Figure 2 shows the extraction process of the concentric circle-based signature. The resized gray image is divided by radius and angle levels. The feature values are calculated in the partitioned regions, and then the image signature is made by hash generation. The proposed algorithm alters two schemes in the existing concentric circle-based signature method. The first is the formula used for angle partitioning, and the second is the types of extracted features. The details of the previous work and the proposed method are given below.

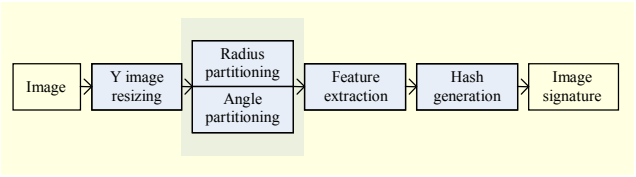

Fig. 2. Extraction process of concentric circle-based signature.

1. Previous Work on Concentric Circle-Based Signature

Signature extraction only uses the Y component in an input image converted to the YUV color model. This Y component image is resized to 256×N or N×256 while maintaining the aspect ratio of the image (256≤N). Prior to image feature extraction, the preliminary work required is to split the image by radius partitioning and angle partitioning. In radius partitioning, the origin is the center of the resized image and the largest circle is established as the one with the maximum radius of 128. The inner region of the largest circle is divided into R(=128/k) concentric circles at intervals of k pixels. Angle partitioning divides these concentric circles by the angle level. It produces sub-regions of the concentric circles at regular intervals. To simplify the image partitioning procedure, we use the polar coordinates (r, θ) instead of the Cartesian coordinates (x, y). As shown in Fig. 3, the inner part of a circle region is converted into the polar coordinate plane. Then, the polar coordinate image map is divided into non-overlapping blocks. We use an angle level of 36 and multi-radius levels of 32 and 16 in this paper. The image features are extracted from the 36×32 grids and 36×16 grids.

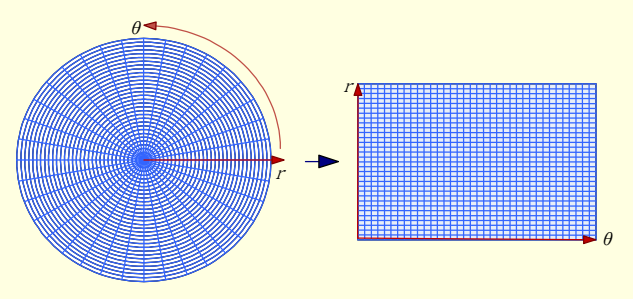

Fig. 3. Polar coordinate conversion.

The previous concentric circle-based signature uses four types of feature distributions; the average intensity distribution $f1$, difference distribution of the average intensity $f2$, symmetrical difference distribution $f3$, and circular difference distribution $f4$. In every concentric circle region, these four types of feature values are calculated by

$$f1(r) = \frac{1}{A}\sum_{a=1}^{A} x_a, \quad (1 \le r \le R), \tag{1}$$

$$f2(r) = \left| f1(r+1) - f1(r) \right|, \quad (1 \le r \le R-1), \tag{2}$$

$$f3(r) = \frac{1}{A/2}\sum_{a=1}^{A/2} \left| x_a - x_{a+A/2} \right|, \quad (1 \le r \le R), \tag{3}$$

$$f4(r) = \frac{1}{A}\sum_{a=1}^{A} \left| x_{(a+1) \bmod A} - x_a \right|, \quad (1 \le r \le R), \tag{4}$$

where A and R are the angle level and radius level, respectively, and x_a is the average intensity of the a-th sub-region in the r-th concentric circle.

The values of each feature are distributed in concentric circle order. $f1(r)$ represents the average intensity computed in the r-th concentric circle, and $f2(r)$ is obtained by calculating the

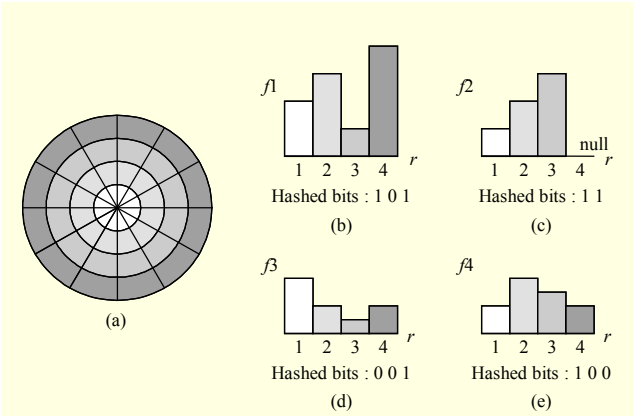

Fig. 4. Example of signature extraction: (a) radius and angle partitioning of image, and (b)-(e) four types of feature distributions and hashed bits.

absolute difference between the neighboring feature values from $f1$. The others, namely $f3$ and $f4$, use the difference between the sub-regions partitioned by the angle level in the concentric circle. As shown in (3) and (4), the average absolute difference is calculated using the opposite sub-regions in $f3$ and the neighboring sub-regions in $f4$.

The distribution of each feature is altered to generate the distribution of the hashed bits. The value of hashed bit B_r is decided by the transition from the current feature value M_r to the next one M_{r+1}, upward or downward. The hash function can be expressed as

$$B_r = \begin{cases} 1, & M_{r+1} > M_r, \\ 0, & M_{r+1} \le M_r. \end{cases} \quad (5)$$

Figure 4 illustrates a simple example of signature extraction in the previous work. Supposing the four features are distributed as shown in Fig. 4, the image signature becomes a bit-stream arranged in the form 10111001100.

The signature size is determined by the radius level R through the following procedure.

$$\begin{aligned} \text{Data size} &= L(f1) + L(f2) + L(f3) + L(f4) \\ &= (R-1) + (R-2) + (R-1) + (R-1) \\ &= 4R - 5 \text{ (bits)}. \end{aligned}$$

Here, $L(*)$ is the length of the hashed bits used for the feature distribution *. In the example of Fig. 4, the signature size is 11 bits when R is 4.

The hashed image signature allows for fast matching by using the Hamming distance measure. The Hamming distance between the two signatures is computed using a simple XOR bit operation. That is, the dissimilarity D is denoted by the average number of different bits between the signatures as shown by

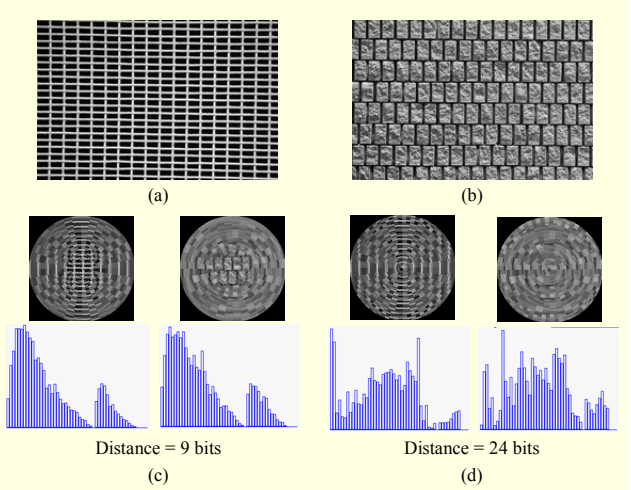

Fig. 5. Comparison of angle partitioning methods: (a) and (b) test images, (c) test results of previous work, and (d) test results of proposed method.

$$D = \frac{1}{N}\sum_{i=1}^{N} T_i \oplus Q_i, \quad (6)$$

where T_i and Q_i are the i-th bits of the target and query signatures, and N is the number of hashed bits.

2. Proposed Image Signature Algorithm

To achieve higher performance in terms of both robustness and discriminability, the new signature uses adaptive angle levels and more delicate features. The rest of the signature extraction process is the same as the existing method.

In previous work, the angle partitioning procedure with regular intervals affects the feature values of the sub-regions in each concentric circle. The closer the concentric circle is to the origin, the smaller the area is; whereas, the farther one has the larger area. This interferes with the extraction of a unique image feature value. This effect is salient in the case of $f3$ and $f4$ extracted using the relation between the sub-regions. Figure 5(c) shows the concentric circle partitioned maps and $f4$ feature distributions in the previous work. It gives similar feature distributions for different images, and the distance between signatures is rather small.

To solve this problem, the proposed algorithm uses an adaptive angle level for the area of the concentric circle. The area of the concentric circle linearly increases in proportion to its radius. For concise computation, we used six kinds of angle levels. The following angle-partitioning pseudocode shows how to select the angle level according to the area of the concentric circle.

```
Angle_level[] = {4, 8, 12, 24, 36, 72}
for i = 1 to Radius_level
```

```
AR[i] = Area(C(i))/Area(C(1))
for j=0 to 5
  if |AR[i]–Angle_level[j]| < Min_dist
    Min_index = j
AL[i] = Angle_level[Min_index]
```

Figure 5(d) shows the result of the proposed angle partitioning method. Feature distributions are different, and the distance between signatures is larger than in the previous one.

Therefore, the proposed method has better discriminative power than the angle partitioning of previous work.

In this proposed algorithm, we extract three types of features from each concentric circle, and two additional features are generated from each feature. The features are the mean intensity, $MI(r)$, the circular difference, $CD(r)$ and the circular variance, $CV(r)$, where r is the index of the concentric circle. The mean intensity $MI(r)$ is the average gray-value of the r-th concentric circle. $CD(r)$ and $CV(r)$ use the relation between the sub-regions by angle partitioning within the r-th concentric circle. $CD(r)$ is the average absolute difference between the neighboring sub-regions in a concentric circle, and $CV(r)$ is the variance of the sub-regions as follows:

$$CD(r) = \frac{1}{A_r}\sum_{a=1}^{A_r}\left|s(r,a) - s(r,\ a-1)\right|, \quad (1 \le r \le R), \tag{7}$$

$$CV(r) = \frac{1}{A_r}\sum_{a=1}^{A_r}\left\{s(r,\ a) - MI(r)\right\}^2, \quad (1 \le r \le R), \tag{8}$$

where A_r is the angle level when in the r-th concentric circle, and $s(r, a)$ is the average intensity of the a-th sub-region in the r-th concentric circle.

The additional features, the first variation and the second variation for each feature, are generated by

$$V^1_{fv}(r) = fv(r+1) - fv(r), \quad (1 \le r \le R-1), \tag{9}$$

$$V^2_{fv}(r) = V^1_{fv}(r+1) - V^1_{fv}(r), \quad (1 \le r \le R-2), \tag{10}$$

where $fv(r)$ is the feature value of the r-th concentric circle. The second variation is only generated for $MI(r)$ and $CD(r)$. The circular variance $CV(r)$ has variation information on itself. As the second variation of the circular variance is highly sensitive to modification, it can have a negative effect on the near-duplicate detection.

In the proposed method, the method of hash generation used differs from (5) used in previous work. An attempt is made to determine the hashed value impartially, since the concentric circles are partitioned by non-uniform angle levels. The hashed value is decided by comparing features extracted using the same angle level. For example, M_r is calculated by angle level 4 and M_{r+1} is calculated by angle level 8. In this case, another value M'_{r+1} is calculated using the same angle level 4, and then M'_{r+1} and M_r are compared. This method reduces that difference between the feature values when they are increased or decreased by an abrupt change of angle level.

The following is the total length of the image signature extracted by the proposed method. The signature size is determined by the radius level R, as in the previous work.

$$\begin{aligned}\text{Data size} &= L(MI) + L(V^1_{MI}) + L(V^2_{MI}) \\ &\quad + L(CD) + L(V^1_{CD}) + L(V^2_{CD}) + L(CV) + L(V^1_{CV}) \\ &= (R-1) + (R-2) + (R-3) \\ &\quad + (R-1) + (R-2) + (R-3) + (R-1) + (R-2) \\ &= 8R - 15 \text{ (bits)}.\end{aligned}$$

IV. Experimental Results

The performance evaluation of the near-duplicate detection algorithm should reflect its suitability in practical applications. We used the image database and modification tool of MPEG-7 VCE-6 for the performance evaluation [18].

Our near-duplicate detection method is compared with two algorithms, the DCT-based ordinal measure [12] and trajectory signature method [17]. The DCT-based ordinal measure (OM) is the most popular and representative algorithm for content-based copy detection. The algorithm in [12] showed the highest discriminability when the ordinal measure used 35 low-frequency magnitudes from 63 AC coefficients. The trajectory signature (TS) method resembles the proposed algorithm in terms of the feature extraction and hash generation. The trajectory signature method extracts group features after three types of block features have been computed. These block features are the mean level Y, mean energy E, and singular energy S. The group features are the mean value $M(.)$ and standard deviation value $D(.)$ for the three types of features. The six types of features use only 119 bits per feature. We also compare the proposed method with the previous concentric circle-based signature method (PreCC). These four algorithms are evaluated in terms of their discriminability, robustness, the ANMRR and complexity, as follows.

1. Discriminability Test

In the ideal near-duplicate detection system, the signature distance between near-duplicate images should be small, while that between different images should be as large as possible. The image signatures were extracted from 135,609 different images of various sizes. The discriminability test uses the normalized distances of (0, 1) for all possible pairs. Figure 6

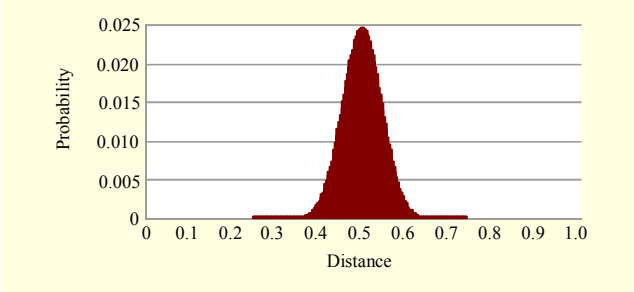

Fig. 6. Normalized distance histogram for all image pairs in proposed method.

Table 1. Normalized distance thresholds for false alarm rate of 0.05 ppm.

Algorithm	Maximum distance	Distance in 0.05 ppm	Normalized distance threshold
OM	612	123	0.20098
TS	714	128	0.17927
PreCC	182	21	0.11538
Proposed	354	83	0.23446

represents the normalized distance histogram of the proposed method. In Fig. 6, the horizontal axis is the normalized distance, and the vertical axis is the occurrence probability. The normalized distance histogram between the different images has a Gaussian distribution around a mean of 0.5.

Table 1 presents the normalized distance thresholds for various algorithms. The distance measures used are the L1 norm in the OM and the Hamming distance in the other algorithms. The distance of the OM algorithm is the maximum when the permutations of the rank matrices are the reverse of each other. The maximum distance of the other algorithms is the number of signature bits. We select a false alarm rate of 0.05 ppm (parts per million) as the distance threshold. This is an error tolerance. This threshold value is applied in the robustness test.

2. Robustness Test

In this experiment, we use 10,000 original images (750×600 pixels or 600×750 pixels) and 28 types of modifications per image. The modification types are noise addition, blurring, JPEG compression, scaling, rotation, and slight skew. The entire set of types of modifications is listed in Table 2.

The robustness is evaluated for every type of modification. The original image is used as the query, and modified images are used as targets in the matching process. A query is compared with the targets, that is, the modified versions of all of the original images. The matching process for the query

Table 2. Robustness test results for 28 modifications.

Modifications \ Algorithms	OM	TS	PreCC	Proposed
Gaussian noise (σ=4)	9997	9977	9974	9995
Gaussian noise (σ=8)	9994	9962	9939	9992
Gaussian noise (σ=12)	9992	9931	9890	9989
Blurring (3×3)	9996	9998	9998	10000
Blurring (5×5)	9998	9991	9990	10000
Blurring (7×7)	9998	9985	9984	9999
Brightness (+10%)	9980	9996	9972	9998
Brightness (+20%)	9885	9985	9809	9986
Brightness (+25%)	9739	9973	9602	9981
16-bit color (bmp)	9992	9978	9966	9996
8-bit color (gif)	9991	9893	9850	9988
JPEG (QF=80)	9999	10000	9994	9999
JPEG (QF=60)	9994	9996	9978	9999
JPEG (QF=30)	9988	9981	9939	9994
Monochrome	9989	9996	9988	9998
Auto-levels	9987	9995	9988	9998
Histogram equalization	6210	7186	5709	9614
Scaling (90%)	9995	10000	10000	10000
Scaling (70%)	9996	9998	9991	10000
Scaling (50%)	9992	9994	9982	9997
Flip (left-right)	9999	9990	9484	9990
Rotation (10°)	30	0	9987	9997
Rotation (25°)	0	0	6385	9351
Rotation (45°)	0	0	4848	9797
Rotation (90°)	1	9992	9796	9998
Rotation (180°)	9997	9979	8920	9994
Rotation (270°)	1	9989	9850	9994
Skew (4°)	8460	1460	1043	8531
Average	8007	8508	9102	9899

should satisfy two conditions simultaneously to determine 'success' in near-duplicate detection. First, the query and its modified version should have the shortest distance. Second, the shortest distance should be less than the distance threshold taken from the discriminability test. This operation is repeated for every query, and the robustness is expressed as the number of successes among the 10,000 queries. Although the robustness test uses a smaller number of images than the discriminability test, the application of the distance threshold takes effect at nearly the equivalent evaluation as for a large database.

Table 3. ANMRR test results.

Algorithm	ANMRR
OM	0.01450824
TS	0.00568134
PreCC	0.00052462
Proposed	0.00046866

Table 4. Signature size and matching time.

Algorithm	Signature size (bit)	Matching time (ms)
OM	280 (35 bytes)	50.31
TS	714	74.06
PreCC	182	20.16
proposed	354	39.37

Table 2 shows the experimental results of the robustness test. As can be seen in Table 2, the proposed algorithm shows the best result with an overall accuracy of 98.99% in the robustness test. The robustness of PreCC is lower than that of the proposed algorithm. The distance threshold of PreCC is the cause of failure in the second condition. The test result also shows that the OM and TS algorithms are very weak in terms of rotation modifications.

3. ANMRR Test

The ANMRR defined in the MPEG-7 visual group measures the recall and precision information and the rank of the retrieval image [19]. This test uses the same test database as used in the robustness test, that is, 10,000 original images and 28 types of modifications per image. In Table 3, the proposed method shows the best performance with the smallest ANMRR value.

4. Signature Size and Matching Complexity

In terms of the complexity, the near-duplicate detection system requires a fast matching method, whereas the requirement for fast extraction of the signature is less stringent. Therefore, we only estimate the matching speed.

Table 4 shows the variation of the matching time with the signature size for each algorithm. OM generates a [1×35] rank matrix for an image. The OM signature size is 35 bytes since each element of the matrix is allocated 1 byte. TS has a signature consisting of 714 bits since 119 bits per feature are extracted. The concentric circle-based signature uses two radius levels of $R1$=32 and $R2$=16. Therefore, PreCC extracts 182 bits per signature, and the proposed method has 354 bits per signature.

The matching speed is related to both the distance measurement and signature size. The matching of OM uses the L1 norm distance, and the other algorithms use the Hamming distance. In our experiment, the Hamming distance is calculated by means of an 8-bit distance look-up table. The matching time is measured using 100 queries and 10,000 targets. According to the results, the Hamming distance measure is beneficial in terms of achieving fast matching for a given signature size.

V. Analysis and Discussion

The proposed concentric circle-based signature uses image partitioning and hash generation for robust and stable near-duplicate detection. This method improves on the existing algorithm through the alteration of the angle partitioning and feature extraction procedures. Angle partitioning is achieved using a non-uniform angle level. We extract the distributions for three types of features and their variations. Then, these distributions are converted into a hashed image signature. This process is repeated twice using radius levels of 32 and 16 as the multiresolution signatures.

As described above, the algorithms were evaluated by discriminability and robustness tests. These two tests evaluated their ability to discriminate different images and detect near-duplicates. A larger signature size seems to contribute to the robustness and discriminability in the comparison between the results obtained for PreCC and the proposed algorithm. However, this finding is in disagreement with the results obtained for TS. To clarify the cause of the performance enhancement, we analyzed the hashed bits of the image signature through a detailed examination. The signature type of the OM method is not a hashed bit, but a scalar value, so we excepted OM from this signature analysis.

If a bit has almost the same value in every image, it has no discriminatory power. Thus, a good signature is one in which each bit value is generated nearby into half 0 and half 1 in probability of occurrence. Figure 7 shows the distributions of the hashed bits for the various algorithms. The length of the horizontal axis of the graph is arranged according to the number of hashed bits per feature. The vertical axis expresses the occurrence probability of bit value “1” in the 135,609 different images used in the discriminability test. Both PreCC and the proposed method have multiresolution signatures using radius levels of 32 and 16. In Figs. 7(a) and 7(b), the horizontal axis is placed into two radius levels. The discrepancy in these graphs is due to the null value(s) generated by feature extraction and hash generation. In Fig. 7(c), the length of the horizontal axis is the number of hashed bits per feature, that is, 119 bits.

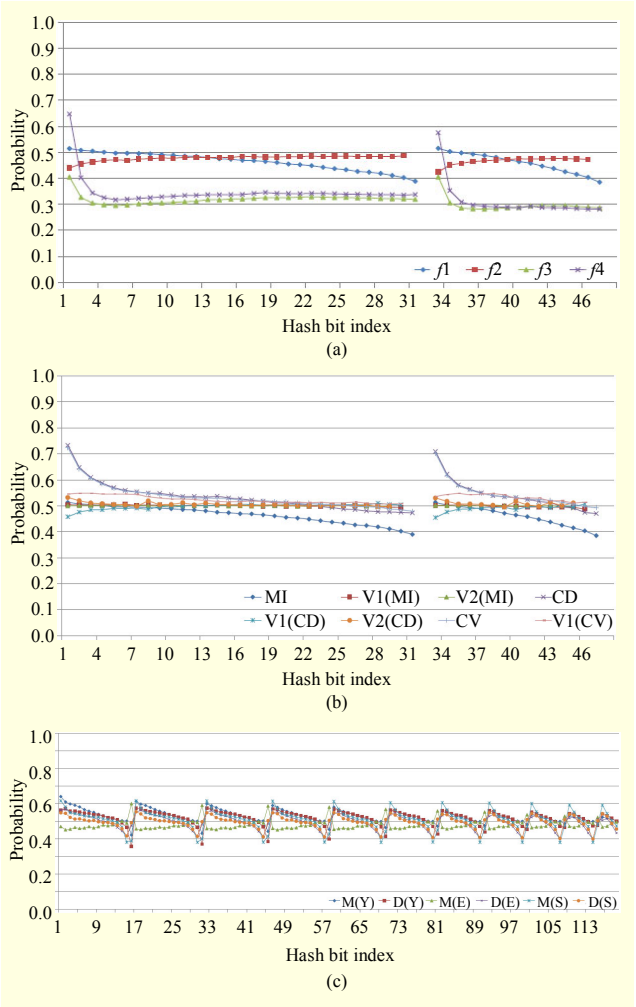

Fig. 7. Analysis results for discriminability of hashed bits: (a) PreCC, (b) proposed method, and (c) TS.

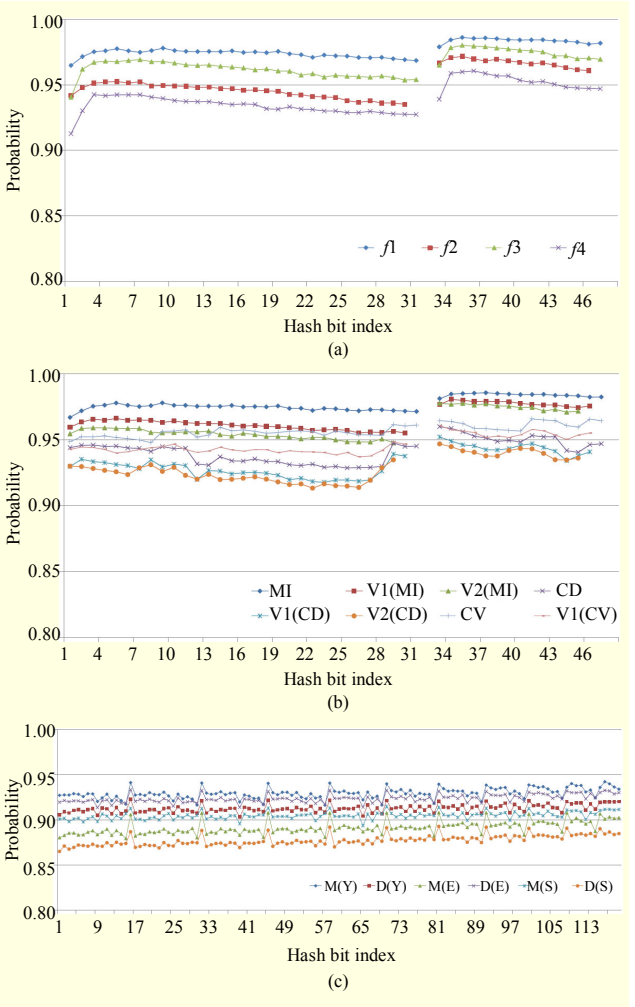

Fig. 8. Analysis results for robustness of hashed bits: (a) PreCC, (b) proposed method, and (c) TS.

The analysis results given in Fig. 7 is connected to the distance thresholds in Table 1. PreCC shows a biased distribution in Fig. 7(a) and the lowest distance threshold of 0.11538 in the discriminability test. In the proposed method, even though the probabilities of some bits deviate from 50% in several features, most bits are close to half. The distance threshold of 0.23446 is also the best in the discriminability test. The hashed bits distribution of TS repeatedly deviates from the middle. This has a bad effect on the discriminability, so being around 0.17927, the distance threshold of TS is comparatively low.

The next analysis represents the robustness of each hashed bit using 10,000 original images and 28 types of modifications.

Each bit of the image signature extracted from the original image is compared with the same bit extracted from a modified version. If the bit value from every modified version is equal to the one from the original image, this hashed bit plays the role of the image signature. Figure 8 presents the analysis results of the hashed bits in terms of the robustness. In the results obtained for all of the algorithms, it is found that the mean and variance of the intensity are comparatively robust features. The hashed bits of the concentric circle-based signature are more stable and robust to image modifications.

The hashed bits of some features exhibit good discriminability, while some exhibit good robustness, and others exhibit both or neither. As can be seen from the analysis results, the proposed method has a better image signature than the others in terms of the discriminability and robustness.

VI. Conclusion

We proposed a concentric circle-based image signature for near-duplicate detection. The proposed algorithm uses concentric circle partitioning and extracts an image signature using the relation of the partitioned local region. The experimental results show that our signature is more robust to various modifications. This robustness is particularly

prominent for rotation modifications in comparison to the other competing algorithms. In the analysis results for the feature components, the features of our signature show better performance. In addition, the proposed signature supports faster matching with an affordable signature size.

In the analysis of the various algorithms, it was confirmed that the performance is related to two important points in the signature extraction process. The first is how to partition an image, and the second is how to process the selected features. The image partitioning method affects the robustness against specific modifications, such as geometrical transformations. The image features are retouched by the ranking of the feature values or hash generation. This process contributes to the size reduction and stability of the image signature.

We used various modifications in the robustness test. The center of the modified image is not changed. The proposed method encounters difficulty for geometrical modifications which change the center. Generally, feature point-based methods [9], [20]-[23] and region-based methods [5] are robust to geometrical modifications. Such image signature detection methods commonly employ a complicated process and need a large storage capacity. Also, the feature point-based method is not robust to modifications, such as noise addition and blurring. Naturally, a single algorithm which can deal with all types of modifications would the best, but it is difficult to satisfy both robustness and complexity. Therefore, we will consider a synthetic method that could be incorporated into the proposed algorithm and other algorithms robust to geometric modification.

References

[1] J.G. Choi et al., "Further Feasibility Test Results of MPEG-7 Visual Descriptors as a Visual Identifier Descriptor," MPEG Doc. No. M12683, Nice, Oct., 2005, pp. 1-6.

[2] C.S. Won, D.K. Park, and S.J. Park, "Efficient Use of MPEG-7 Edge Histogram Descriptor," *ETRI J.*, vol. 24, no.1, Feb. 2002, pp. 23-30.

[3] L. Wu et al., "Query Oriented Subspace Shifting for Near-Duplicate Image Detection," *IEEE Int. Conf. Multimedia Expo.*, 2008, pp. 661-664.

[4] C. Kim and B. Vasudev, "Spatiotemporal Sequence Matching for Efficient Video Copy Detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, Jan. 2005, pp. 127-132.

[5] P. Ghosh et al., "Duplicate Image Detection in Large Scale Databases," *Indian Statistical Institute Platinum Jubilee Volume*, Kolkata, Oct. 2007, pp. 1-17.

[6] A. Chalechale, A. Mertins, and G. Naghdy, "Edge Image Description Using Angular Radial Partitioning," *IEE Proc. Vision, Image Signal Process.*, vol. 151, no. 2, Apr. 2004, pp. 93-101.

[7] J.H. Hsiao et al., "A New Approach to Image Copy Detection Based on Extended Feature Sets," *IEEE Trans. Image Process.*, vol. 16, no. 8, Aug. 2007, pp. 2069-2079.

[8] C.S. Lu and C.Y. Hsu, "Geometric Distortion-Resilient Image Hashing Scheme and Its Applications on Copy Detection and Authentication," *Multimedia Syst.*, vol. 11, no. 2, Dec. 2005, pp. 159-173.

[9] G. Roth and W. Scott, "Efficient Indexing for Strongly Similar Subimage Retrieval," *4th Canadian Conf. Computer Robot Vision,* May 2007, pp. 440-447.

[10] I.H. Cho et al., "Very Fast Concentric Circle Partition-Based Replica Detection Method," *Adv. Image Video Technol.*, *LNCS*, vol. 4872, 2007, pp. 905-918.

[11] E.Y. Chang et al., "RIME: A Replicated Image Detector for the World-Wide Web," *Proc. SPIE Multimedia Storage Archiving Syst.*, vol. 3527, Nov. 1998, pp. 58-67.

[12] C. Kim, "Content-based Image Copy Detection," *Signal Process. Image Commun.*, vol. 18, no. 3, Mar. 2003, pp. 169-184.

[13] Li Chen and F.W.M. Stentiford, "Video Sequence Matching Based on Temporal Ordinal Measurement," *Patt. Recog. Lett.*, vol. 29, no. 13, Oct. 2008, pp. 1824-1831.

[14] M.N. Wu, C.C. Lin, and C.C. Chang, "A Robust Content-Based Copy Detection Scheme," *Fundamenta Informaticae*, vol. 71, 2006, pp. 351-366.

[15] C.C. Lin and S.S. Wang, "An Edge-Based Copy Detection Scheme," *Fundamenta Informaticae*, vol. 83, 2008, pp. 299-318.

[16] M.N. Wu, C.C. Lin, and C.C. Chang, "Novel Image Copy Detection with Rotating Tolerance,"*J. Syst. Software*, vol. 80, no. 7, July 2007, pp. 1057-1069.

[17] K. Wnukowicz, G. Galinski, and R. Tous, "Still Image Copy Detection Algorithm Robust to Basic Image Modifications," *Int. Symp. ELMAR*, Sept. 2008, pp. 455-458.

[18] M. Bober, K. Iwamoto, and P. Brasnett, "Description of MPEG-7 Visual Core Experiments," MPEG Doc. No. N9582, Antalya, January 2008, pp. 1-15.

[19] S.J. Park, D.K. Park, and C.S. Won, "Core Experiments on MPEG-7 Histogram Descriptors," MPEG Doc. No. M5984, Geneva, May, 2000, pp. 1-13.

[20] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, Nov. 2004, pp. 91-110.

[21] H. Bay et al., "Speeded-Up Robust Features (SURF)," *Comput. Vision Image Understanding*, vol. 110, no. 3, June 2008, pp. 346-359.

[22] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors," *Int. J. Comput. Vision*, vol. 60, no.1, Oct. 2004, pp. 63-86.

[23] L. Hyston, Y. Ke, and R. Sukthankar, "Efficient Near-Duplicate Detection and Sub-image Retrieval," *Proc. ACM Multimedia Conf.*, Aug. 2004, pp. 869-876.

Ayoung Cho received her BS in electronic engineering and her MS in information engineering from Inha University, Korea, in 2003 and 2005, respectively. She is currently working toward a PhD in information engineering at the same university. Her research interests include machine vision, watermarking, and image and video signature.

Won-Keun Yang received his BS in electronic engineering and his MS in information engineering from Inha University, Korea, in 2004 and 2006, respectively. He is currently working toward a PhD in information engineering at the same university. His research interests include image processing, visual search, and image and video signature.

Weon-Geun Oh received his BS from Chungbuk National University, Korea, in 1979 and his MS from Yeungnam University, Korea, in 1981. He received his PhD from Osaka University, Japan, in 1988. He now works at ETRI as a senior research engineer. His research interests include computer vision, pattern recognition, and digital rights management.

Dong-Seok Jeong became a member of the IEEE in 1983 and a senior member in 2000. He got his BSEE from Seoul National University in 1977 and his MSEE and PhD from Virginia Polytechnic Institute and State University, USA, in 1985 and 1988, respectively. He is also the member of SPIE and HKN. From 1977 to 1982, he was a researcher at the Agency for Defense Development of Korea. Since 1988, he has been a professor at Inha University, Korea. He served as the president for the Institute of Information and Electronics Research from 2000 to 2004. His research interests include image and video processing, image and video signature, and forensic watermarking.