MPEG-7 Homogeneous Texture Descriptor

Yong Man Ro, Munchurl Kim, Ho Kyung Kang, B.S. Manjunath, and Jinwoong Kim

MPEG-7 standardization work has started with the aims of providing fundamental tools for describing multimedia contents. MPEG-7 defines the syntax and semantics of descriptors and description schemes so that they may be used as fundamental tools for multimedia content description. In this paper, we introduce a texture based image description and retrieval method, which is adopted as the homogeneous texture descriptor in the visual part of the MPEG-7 final committee draft. The current MPEG-7 homogeneous texture descriptor consists of the mean, the standard deviation value of an image, energy, and energy deviation values of Fourier transform of the image. These are extracted from partitioned frequency channels based on the human visual system (HVS). For reliable extraction of the texture descriptor, Radon transformation is employed. This is suitable for HVS behavior. We also introduce various matching methods; for example, intensity-invariant, rotation-invariant and/or scale-invariant matching. This technique retrieves relevant texture images when the user gives a querying texture image. In order to show the promising performance of the texture descriptor, we take the experimental results with the MPEG-7 test sets. Experimental results show that the MPEG-7 texture descriptor gives an efficient and effective retrieval rate. Furthermore, it gives fast feature extraction time for constructing the texture descriptor.

Manuscript received October 7, 2000; revised May 10, 2001.

Yong Man Ro and Ho Kyung Kang are with the Information and Communications University, Taejon, Korea.
(phone: +82 42 866 6129, e-mail: yro@icu.ac.kr)
Munchurl Kim and Jinwoong Kim are with the Broadcasting Media Technology Department, ETRI, Taejon, Korea.
(phone: +82 42 866 5847, e-mail: mckim@etri.re.kr)
B.S. Manjunath is with the University of California, Santa Barbara, USA. (phone: +1 1 805 7112, e-mail: manj@ece.ucsb.edu)

I. INTRODUCTION

Recently, there has been an overwhelming increase in the amount of digital multimedia information going over the Internet and broadcasting systems. And users need new method to organize, manipulate and transmit the data they want.

The current technologies for representing multimedia in the forms of texts have many limitations. They cannot efficiently represent and retrieve various types of multimedia contents. Also, international standards such as JPEG, MPEG-1, MPEG-2 and MPEG-4 have been developed only for compression of data. These standards are created for efficient storage and transmission, not for the representation of the contents.

At the 36th MPEG Chicago meeting in September 1996, MPEG members first discussed an “Audiovisual Content Description Interface” for efficient representation of multimedia information. They wanted to make an international standard that became MPEG-7. Its official name is “Multimedia Content Description Interface” of ISO/IEC JTC1 SC29/WG11 and work began at the 37th MPEG Maceio meeting in November 1996. MPEG-7 standardization reached the final committee draft (FCD) level at the MPEG Singapore meeting in March 2001. This is a technically stable stage.

MPEG-7 standard specification defines the syntax and semantics of describing the multimedia contents and consists of 7 parts: Systems, Description Definition Language (DDL), Visual descriptor, Audio descriptor, Multimedia Description schemes (MDS), Reference software, and Conformance testing.

In the visual part of the MPEG-7 standard, visual descriptors are specified as normative descriptors, basic descriptors, and descriptors for localization. Normative descriptor describes the color, shape, texture and motion features of visual data [1], [2].

In this paper, we introduce a texture-based image description and retrieval method which we proposed and adopted as the Homogeneous texture descriptor in the Visual part of the MPEG-7 FCD. Our proposal was adopted.

The texture information of an image is a fundamental visual feature, which has been studied during the last decade to analyze images in the areas of medical imaging and satellite imaging, etc. [3]-[5]. This contains structureness, regularity, directionality and roughness of images, which are important properties of the content-based indexing of the image [6].

Previous works such as probability distribution of pixels [3], directional filtering [3] and Markov random field have been studied. More recently, spatial Gabor filters and wavelet transformation have been studied to extract texture information. In [5], Gabor, Pyramid structured Wavelet Transform (PWT), Tree structured Wavelet Transform (TWT), and Multiresolution Simultaneous Autoregressive Model (MRSAR) methods have been compared. In that paper, Gabor and MRSAR methods show good performance of relevant texture image retrieval. However, the methods require high computational complexity to extract the texture information. The MPEG-7 homogeneous texture descriptor we invented is efficient not only for computing texture features but also in representing texture information. Through the core experiments in the MPEG-7 Visual group, the MPEG-7 homogeneous texture descriptor that we describe in this paper had been severely tested and compared with other proposed texture descriptors in terms of computation complexity and retrieval accuracy. It outperformed the other by showing fast feature extraction and compact representation of texture information. It provided higher retrieval accuracy for the testing data sets. Therefore, the homogeneous texture descriptor described in this paper was selected as the normative MPEG-7 homogeneous texture descriptor in the Visual part of the MPEG-7 final committee draft [11]-[14].

In this paper, we present technical details of the MPEG-7 homogeneous texture descriptor and its feature extraction method. In addition to the feature extraction method, similarity measuring criteria are presented for rotation-, scale- and intensity-invariant matchings [15]-[18].

The homogeneous texture descriptor in this paper consists of the mean and standard deviation values of an image. It also includes the energy and energy deviation values of the Fourier transform of the image. In order to explain the texture representation based on energy and energy deviation features, the texture feature extraction is explained in Section II . Texture indexing and retrieval algorithms are then presented in Section III. Section III also addresses various image matching criteria for intensity-, rotation- and/or scale-invariant matching in retrieval. In Section IV, the experimental results are provided for the MPEG-7 data set of the texture experiment.

II. TEXTURE DESCRIPTOR EXTRACTION ALGORITHM FOR THE MPEG-7 HOMOGENEOUS TEXTURE DESCRIPTOR

1. Human Visual System for the Texture Descriptor

Recently, texture-featuring and description techniques based on the HVS have been proposed [4]. Texture featuring based on the HVS corresponds well to some results from psychophysical experiments. In these experiments, the response of the visual cortex is turned to a band-limited portion of the frequency domain. The human brain decomposes the spectra into perceptual channels that are bands in spatial frequency [5], [7]. For texture featuring, the best sub-band representation of HVS is a division of the spatial frequency domain in octave-bands (4~5 divisions) along the radial direction and in equal-width angles along the angular direction. These sub-bands are symmetrical with respect to the origin of the Polar coordinate. In this section, a frequency layout is designed. The frequency layout allows extracted texture information to be matched with human perception system. The frequency layout consists of sub-bands. In these bands, the texture descriptor components such as energy and energy deviation are extracted.

According to the HVS properties mentioned above, the sub-bands are designed by dividing the frequency domain to compute texture feature values. The frequency space from which the texture descriptor in the image is extracted is partitioned in equal angles of 30 degrees along the angular direction and in octave division along the radial direction. The sub-bands in the frequency domain are called feature channels indicated as C_i in Fig. 1. The frequency space is therefore partitioned into 30 feature channels as shown in Fig. 1. In the normalized frequency space $(0 \le \omega \le 1)$, the normalized frequency ω is given by $\omega = \Omega / \Omega_{\max}$. $\Omega_{\max}$ is the maximum frequency value of the image. The center frequencies of the feature channels are spaced equally in 30 degrees along the angular direction such as $\theta_r = 30° \times r$. Here r is an angular index with $r \in \{0, 1, 2, 3, 4, 5\}$. The angular width of all feature channels is 30 degree. In the radial direction, the center frequencies of the feature channels are spaced with octave scale such as $\omega_s = \omega_0 \cdot 2^{-s}, s \in \{0, 1, 2, 3, 4\}$ where s is a radial index and ω_0 is the highest center frequency specified by 3/4. The octave bandwidth of the feature channels in the radial direction is written as $B_s = B_0 \cdot 2^{-s}, s \in \{0, 1, 2, 3, 4\}$ where B_0 is the largest bandwidth specified by 1/2.

Figure 1 shows the two-dimensional (2D) frequency layout configured by the above division scheme. As shown in Fig. 1, each partitioned region corresponds to a band-limited portion of the frequency domain that is the response of the visual cortex in the HVS. Therefore, the region can be denoted as a channel to transfer the response of the visual cortex. The channels located in the low frequency areas are of smaller sizes while those of the high frequency areas are of larger sizes. This corresponds to the human vision that is more sensitive to the change of low frequency area. Also, note that half of the entire frequency space is used because images are assumed to be of real value.

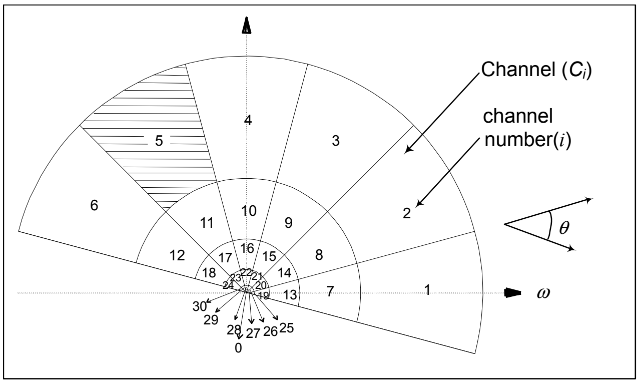


Fig. 1. Frequency region division with HVS filter.

2. Data Sampling in Feature Channel

As shown in Fig. 1, the channel layout in the spatial frequency domain is center-symmetrical. Since the partitioned frequency regions are relatively small compared with those in the high frequency regions in the Cartesian coordinate system, the frequency samples are sparse in the low frequency regions where the texture information is insufficient. In order to avoid this, we employ Radon transform for images, which allows Fourier transform of image in Cartesian to be represented in the Polar coordinate system. Using the Radon transformation, 2D image can be transformed to one-dimensional (1D) projection data, *i.e.*, Cartesian space *(x, y)* will be mapped to Radon space *(R, θ)* as shown in Fig. 2.

The line integral along the line L(R, θ) at angle θ in counterclockwise direction from y-axis and at a distance R from the origin can be written as

$$p_\theta(R) = \int_{L(R,\theta)} f(x,y)dl$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - R)dxdy\,, \qquad (1)$$

where $f(x,y)$ is an image function, R is projection axis and $\delta(\cdot)$ is delta function. The function $p_\theta(R)$ is a projection, since it collapses a 2D image to a 1D projection for each angle. The complete collection of line integrals is called the Radon transform of $f(x,y)$ and also called the *Sinogram*. The frequency properties in Radon transformation can be explained by "*central slice theorem*" in which the 1D Fourier transform of a projection of image at angle θ equals the slice at angle θ through the 2D Fourier transform of that image (see Fig. 3).

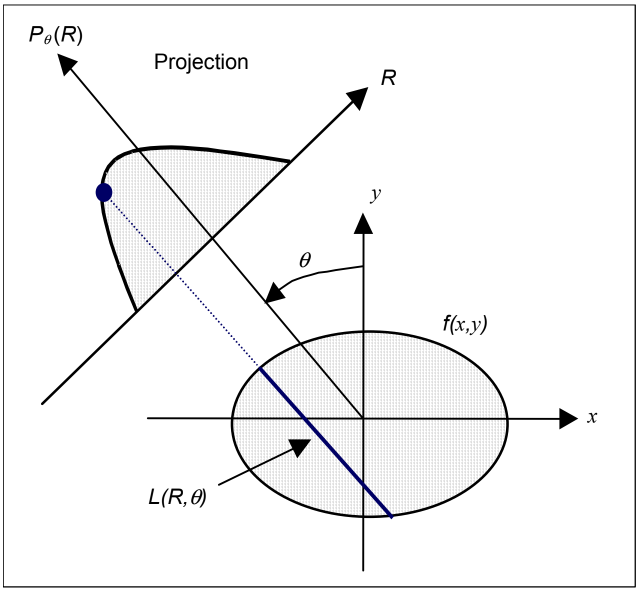


Fig. 2. Radon transform scheme. Image *f(x,y)* is transformed to $p_\theta(R)$ in Radon space *(R, θ)*.

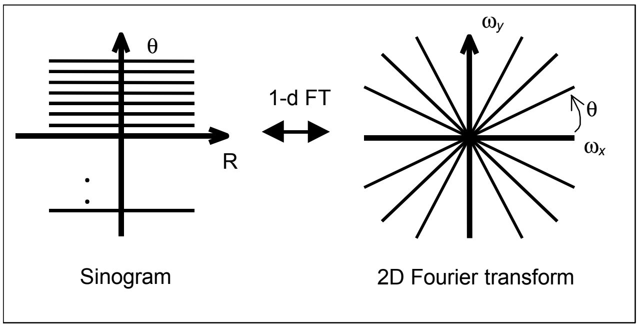


Fig. 3. Relationship between sinogram and 2-dimensional fourier domain.

One-dimensional Fourier transform of a projection can be written as

$$\int p_\theta(R)\exp(-j2\pi R\omega)dR$$
$$= \iint f(x,y)\exp[-j2\pi\omega(x\cos\theta + y\sin\theta)dxdy\,, \qquad (2)$$

where $\omega = \sqrt{\omega_x^2 + \omega_y^2}$ and $\theta = \tan^{-1}(\omega_y/\omega_x)$.

The Radon transform is suitable for the HVS since each central slice in Fourier domain is fit to the data representation in the HVS frequency layout mentioned previously. Data acquisition in the HVS-based frequency layout is done with polar-oriented sampling scheme.

Figure 4 shows a sampling grid structure of a Polar frequency domain after the Radon transform followed by its Fourier transform. As shown in the figure, sampling density is dense in the low and middle frequency areas while sparse in the high frequency area. This property corresponds to the human visual properties such that the human vision is more sensitive in the low frequency area than in the high frequency area. This property supports that the Radon transformation of image is suitable to the HVS [15].

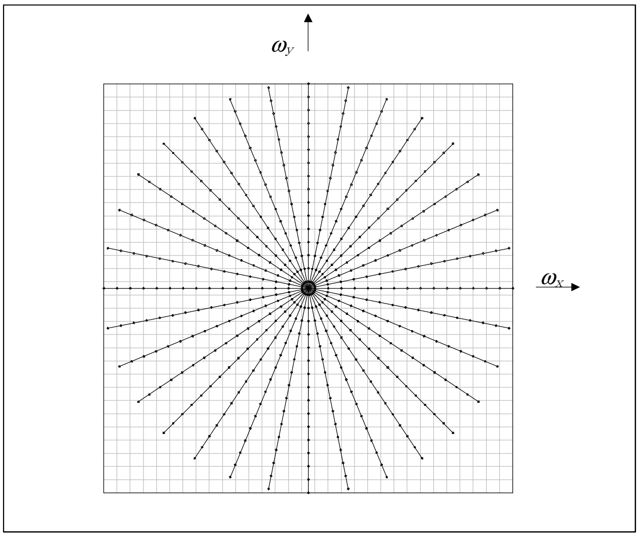


Fig. 4. Sampling grid of a Polar frequency domain after the Radon transform followed by Fourier transform.

In signal processing perspective, the frequency layout in Fig. 1 is actually realized with a set of ideal filter banks that have abrupt channel boundaries. In order to relax the sharpness of the pass band edges of the ideal filters between channels, Gabor filter banks are instead used to construct the frequency layout. By applying the Gabor filter banks, the channels are overlapped so that the channels can affect neighbor channels each other (each other is redundant) at the boundary areas. The Gabor function defined for Gabor filter banks is written as

$$G_{P_{s,r}}(\omega,\theta) = \exp\left[\frac{-(\omega-\omega_s)^2}{2\sigma_{\omega_s}^2}\right]\cdot\exp\left[\frac{-(\theta-\theta_r)^2}{2\sigma_{\theta_r}^2}\right] \qquad (3)$$

where $G_{P_{s,r}}(\omega,\theta)$ is Gabor function at s-th radial index and r-th angular index. σ_{ω_s} and σ_{θ_r} are the standard deviations of the Gabor function in the radial direction and the angular direction, respectively. The standard deviations of the Gabor function are determined by touching the Gabor function with its neighbor functions at half the maximum (1/2) in both radial and angular directions. Figure 5 shows the Gabor filters on top of the frequency layout, which are 6 partitions in the angular direction and 5 partitions in the radial direction.

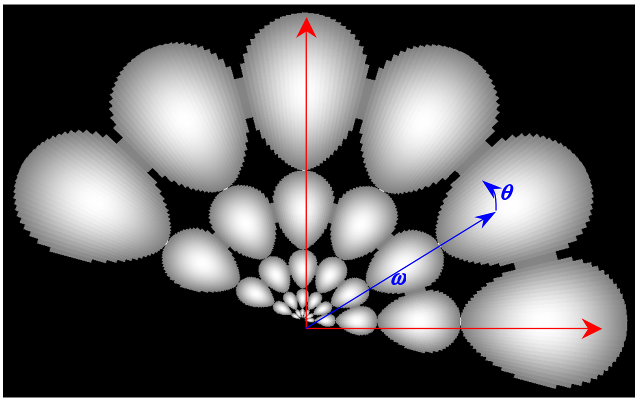


Fig. 5. 5×6 Gabor filters in polar coordinate system.

For the frequency layout shown in Fig. 1, σ_{θ_r} is a constant value of $15^\circ/\sqrt{2\ln 2}$ in the angular direction. In the radial direction, σ_{ω_s} is dependent on the octave bandwidth and is written as

$$\sigma_{\omega_s} = \frac{B_s}{2\sqrt{2\ln 2}}. \qquad (4)$$

Tables 1 and 2 show parameters in the feature channels and the Gabor functions.

Table 1. Parameters of octave band in the radial direction.

Radial index (s)	0	1	2	3	4
Centre frequency (ω_s)	$\frac{3}{4}$	$\frac{3}{8}$	$\frac{3}{16}$	$\frac{3}{32}$	$\frac{3}{64}$
Octave bandwidth (B_s)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
σ_{ω_s}	$\frac{1}{4\sqrt{2\ln 2}}$	$\frac{1}{8\sqrt{2\ln 2}}$	$\frac{1}{16\sqrt{2\ln 2}}$	$\frac{1}{32\sqrt{2\ln 2}}$	$\frac{1}{64\sqrt{2\ln 2}}$

Table 2. Parameters of angular band in the angular direction.

Angular Index (r)	0	1	2	3	4	5
Centre frequency (θ_r)	0°	30°	60°	90°	120°	150°
Angular bandwidth	30°	30°	30°	30°	30°	30°
σ_{θ_r}	$\frac{30^\circ}{2\sqrt{2\ln 2}}$	$\frac{30^\circ}{2\sqrt{2\ln 2}}$	$\frac{30^\circ}{2\sqrt{2\ln 2}}$	$\frac{30^\circ}{2\sqrt{2\ln 2}}$	$\frac{30^\circ}{2\sqrt{2\ln 2}}$	$\frac{30^\circ}{2\sqrt{2\ln 2}}$

3. MPEG-7 Homogeneous Texture Descriptor

To extract the texture feature values, we take the Radon transform on an image and subsequent 1D Fourier transform on the data. Then, we can obtain a central slice of $F(\omega,\theta)$ in 2D frequency domain. The texture descriptor consists of feature values extracted from each channel shown in Fig. 1. In this paper, the texture descriptor components are the first and second moments of energy in channels, *i.e.*, energies and energy deviations. The energies and energy deviations that constitute the texture descriptor are written as $[e_1,e_2,\ldots,e_{30}]$ and $[d_1,d_2,\ldots,d_{30}]$, respectively. Here the indexes from 1 to 30 indicate the feature channel numbers.

Based on the frequency layout (partitioned frequency domain) and the Gabor functions, the energy e_i of the *i*-th feature channel is defined as the log-scaled sum of the squares of Gabor-filtered Fourier transform coefficients of an image:

$$e_i = \log[1+p_i], \tag{5}$$

where

$$p_i = \sum_{\omega=0+}^{1} \sum_{\theta=0^\circ+}^{360^\circ} [G_{P_{s,r}}(\omega,\theta)\cdot|\omega|\cdot F(\omega,\theta)]^2 , \tag{6}$$

where $|\omega|$ is Jacobian term between Cartesian and Polar frequency coordinates and can be written as $|\omega| = \left|\sqrt{\omega_x^2+\omega_y^2}\right|$. $F(\omega,\theta)$ is Fourier transform of the image $f(x,y)$. The summation is taken over the entire frequency domain except the DC component. Note that $i=6\times s+r+1$ where s is the radial index and *r* is the radial index. The energy deviation d_i of the *i*-th feature channel is defined as the log-scaled standard deviation of the squares of Gabor-filtered Fourier transform coefficients of the image:

$$d_i = \log[1+q_i] \tag{7}$$

where

$$q_i = \sqrt{\sum_{\omega=0+}^{1} \sum_{\theta=0^\circ+}^{360^\circ} \left\{ \left[G_{P_{s,r}}(\omega,\theta)\cdot|\omega|\cdot F(\omega,\theta)\right]^2 - p_i \right\}^2} \tag{8}$$

Further, brightness information of texture (mean denoted by f_{DC}) and standard deviation (f_{SD}) of the entire image pixels are added as the texture feature values in the texture descriptor. Finally, the image intensity average f_{DC}, standard deviation f_{SD}, energies e_i, and energy deviations d_i of the channels constitute the homogeneous texture descriptor (TD) in the order as follows:

$$TD = [f_{DC}, f_{SD}, e_1, e_2, \ldots, e_{30}] \tag{9}$$

at the base layer, and

$$TD = [f_{DC}, f_{SD}, e_1, e_2, \ldots, e_{30}, d_1, d_2, \ldots, d_{30}] \tag{10}$$

at the enhancement layer.

The texture descriptor can be represented at two different layers: base layer and enhancement layer. The texture descriptor only consists of f_{DC}, f_{SD}, and 30 energy values (e_i) of the Fourier transform of the image. In the enhancement layer, the texture descriptor (additionally or) adds 30 (or additional) energy deviation values of the Fourier transform of the image in the texture descriptor vector. The layering scheme of the texture descriptor provides scalability of representing image texture depending upon applications. For the delivery of limited bandwidths, only texture descriptor components at the base layer may be transmitted. Also, fast matching can be performed at the base layer by satisfying retrieval accuracy.

4. Quantization of Texture Descriptor

In this paper, the quantization levels of the texture descriptor values set to 256. Eight bits are assigned and used for linear quantization of each descriptor value. The linear quantization used in the paper is written as

$$D_{quant} = \left[\frac{D_{nonquant}-\beta_{\min}}{\beta_{\max}-\beta_{\min}} \times q\_level\right] \times \frac{\beta_{\max}-\beta_{\min}}{q\_level} + \beta_{\min} \tag{11}$$

where D_{quant} is a quantized descriptor value, $D_{nonquant}$ is the value of the feature, and $\beta_{\max}$ and $\beta_{\min}$ are the maximum and minimum values of features in the database. Note that *q_level* is set to 255.

After the quantization, each feature has 1 byte in size. With many MPEG-7 core experiments, 1 byte was good enough not to lose the texture information. Further, an entropy coding is possible to reduce the bits more. But, this can be considered for coding efficiency in further work. Total texture descriptor length is therefore reduced to 32 bytes for base layer and 62 bytes for the enhancement layer.

III. RETRIEVAL ALGORITHM FOR THE TEXTURE DESCRIPTOR

1. Similarity Measurement

To retrieve similar texture images for a query, a matching procedure should be performed. The matching procedure is as follows. First, Radon transform of querying image is performed so that 1D-projection signals are obtained. Using "*central slice theorem*", frequency data in the polar space are obtained. For the texture descriptor, energy and deviation mentioned in the previous section are calculated. Then, the similarity between a querying image and images in the database is measured. The feature of a querying image *i* is denoted by TD_i while the feature of an image *j* in the database by TD_j. The similarity measured by calculating the distance between the two feature vectors is as follows:

$$d(i,j) = \text{distance}\,(TD_i, TD_j) = \sum_k \left| \frac{w(k)[TD_i(k) - TD_j(k)]}{\alpha(k)} \right|, \qquad (12)$$

where $w(k)$ is the weighting factor of *k*-th descriptor value. The normalization values $\alpha(k)$ are standard deviations of texture descriptor values for a reference database. (During MPEG-7 core experiments, T1 dataset was used as reference database). The weighting parameter $w(k)$ and the normalization value $\alpha(k)$ are calculated in advance so that they are independent on the database. These values could be obtained *a priori* at the beginning of establishing the database.

2. Intensity-Invariant Matching

For the intensity invariance that is usually required for most applications, f_{DC} is eliminated from the feature vector when the similarity measurement is performed.

3. Scale-Invariant Matching

For a given querying image, querying image is zoomed in and out with *N* different zooming factors. The distance $d(i,j)$ between the querying image *i* and the image *j* indexed in database is obtained by

$$d(i,j,n) = \text{distance}\,(TD_i(k)\,|_n, TD_j(k)) \qquad (13)$$

$$d(i,j) = \text{minimum of } \{d(i,j,n)\,|\, n = 1 \text{ to N}\} \qquad (14)$$

where *N* is the number of scaled (zoom-in and zoom-out) versions of the querying feature. *N* is usually 3 so for example, the original and two scaled versions of the querying image are 30% zoom-in and 30% zoom-out. One can use different zoom-in and zoom-out.

4. Rotation-Invariant Matching

Since the frequency space division for the texture descriptor is made in the polar domain as shown in Fig. 2, the texture descriptor of a rotated image is an angular-shifted version of the original image. By using the rotational property, we propose rotation invariant similarity matching method. We first measure the distance between texture descriptor vectors in the database and a querying texture descriptor vector by shifting the querying texture vector in the angular direction such as

$$d(i,j,m\phi) = \text{distance}\,(TD_i(k)\,|_{m\phi}, TD_j(k)) \qquad (15)$$

where $\phi = 30$ degrees. Then, for rotation invariant descriptor, distance is calculated as

$$d(i,j) = \text{minimum of } \{d(i,j,m\phi)\,|\, m = 1 \text{ to } 6\}\ . \qquad (16)$$

5. Layered Texture Descriptor

For an efficient storage or transmission, 62 features can be assigned with priority. Namely, with limited storage or bandwidth, the texture descriptor can be reduced without degrading retrieval accuracy significantly. Especially, in wireless Internet which has poor network environment, only a part of the texture descriptor components could be transmitted to the MPEG-7 database. In this case, the entire texture descriptor components are not used for the content-based indexing. To meet the above requirements, the layered configuration of the texture descriptor is helpful for the better retrieval performance. The texture descriptor is layered as follows:

$$TD \ = TD^{base-layer} + TD^{enhancement-layer} \qquad (17)$$

where $TD^{base-layer}$ is the texture descriptor at the base layer , which is represented with the first and second moments of the image pixels and channel energy (e_i) as

$$TD^{base-layer} = [f_{DC}, f_{SD}, e_1, e_2, \ldots, e_{30}]. \qquad (18)$$

$TD^{enhancement-layer}$ is an extended texture descriptor at the base layer to enhance the retrieval efficiency, that is, it uses full feature values in the descriptor. It can be written as

$$TD^{enhancement-layer} = [f_{DC}, f_{SD}, e_1, e_2, \ldots, e_{30}, d_1, d_2, \ldots, d_{30}]. \qquad (19)$$

IV. MEASUREMENT OF RETRIEVAL PERFORMANCE

To verify the performance of the texture descriptor mentioned above, experiments have been performed with the test data sets for the homogeneous texture descriptor in order to measure feature extraction time and retrieval accuracy. Retrieval performance of the texture descriptor is measured by retrieval rate (*RR*) which is a ratio between the number of relevant images and the number of ground truth image for a given querying image. Similar images, which are of the same number of the ground truth, are selected by measuring distance from the querying image. The relevant images are those belonging to the ground-truth images among the similar images. The *RR* can be written as

$$RR = \frac{\#\ of\ \ relevant\ retrieved\ images}{\#\ \ of\ \ ground\ \ truth} \tag{20}$$

The average retrieval rate for a data set (*AVRR*) is, therefore, denoted by

$$AVRR = {\left[\sum_{i=1}^{number\,of\,query} RR_i\right]}\Big/{number\ of\ query} \tag{21}$$

The MPEG-7 test data sets for the texture descriptor have 7 different kinds of test data sets, which are T1, T2, T3, T4, T5, T6, and T7 data sets. The following subsections explain MPEG-7 test data sets used in the core experiments of the homogeneous texture descriptor in detail.

1. T1 Data Set

T1 data set contains texture pattern images which have been used popularly as a test image set for the texture experiments in many literatures. It consists of 1856 images with matrix size of 128×128. 1856 images are made from 116 Brodatz images with matrix size of 512×512. Each Brodatz image with matrix size of 512×512 is divided into 16 non-overlapped partitions, *i.e.,* 16 images with matrix size of 128×128. In the T1 data set, one image has 15 ground truths since 16 images are generated from one Brodatz pattern. Therefore, the relevant images in (20) belong to the ground truth images as well as the first 15 retrieved images having minimum distance. The querying images for T2 data set are the original patterns. So the number of query in (20) is 116. Figure 6 shows an example of the retrieval by a querying image with the T1 data set. As shown in the figure, an image at top-left is the query and remaining 15 images are retrieved ones. The right side of the querying image has the smallest distance value.

2. T2 Data Set

T2 data set consists of real patterns taken from outdoor and indoor scenes. It consists of 832 images with size of 128×128. Like the T1 data set, 832 images in T2 data set are made from 52 images with matrix size of 512×512 such that an image of matrix size of 512×512 is divided into 16 non-overlapped partitions. In the T2 data set, one image has 15 ground truths and the number of querying images is 52 images. The relevant images are ground truth images which belong to the first 15 images with minimum distance. Figure 7 shows one example of the retrieval by a querying image with the T2 data set.

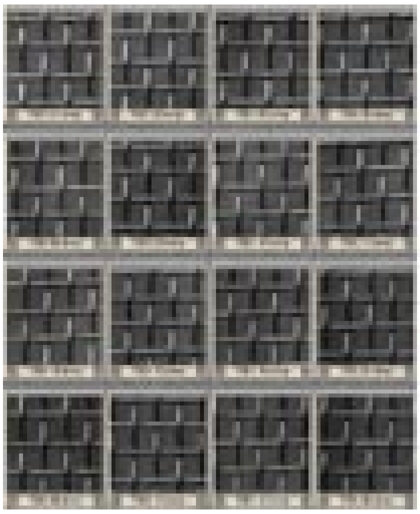


Fig. 6. An example of retrieved images in T1 data set. The upper-left image is a query image. The other 15 images are retrieved images for the query.

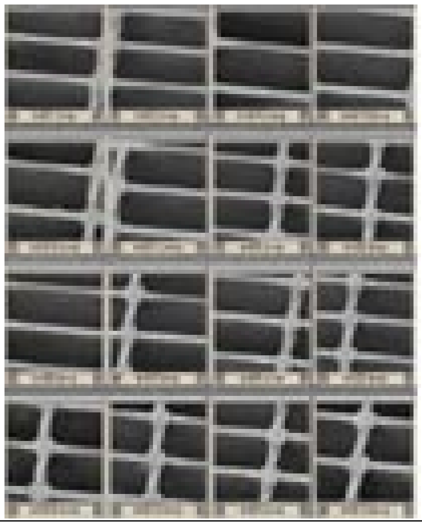


Fig. 7. An example of retrieved images in T2 data set. The upper-left image is a query image. The other 15 images are retrieved images for the query.

3. T3 Data Set

T3 data set is the rotated version of the T1 and T2 data sets. Fifty five original images with matrix size of 512×512 are taken from the T1 and T2 data sets such that 30 patterns are from the T1 data set and 25 patterns from the T2 data set. Then, the 55 images are rotated by 10, 15, 20, 30, 40, 50, 70, 75, 80, 100, 110, 130, 135 140, 160, and 170 degrees. Finally, the T3 data set is constructed by taking 128×128 size image at arbitrary position from the rotated images. So the total number of image in the T3 data set is 880. To evaluate the performance with the T3 data set, *RR* and *AVRR* are measured. The number of ground truth is 16 and the querying images are 55 images which are rotated by 30 degrees. Figure 8 shows one example of the retrieval by a querying image with the T3 data set.

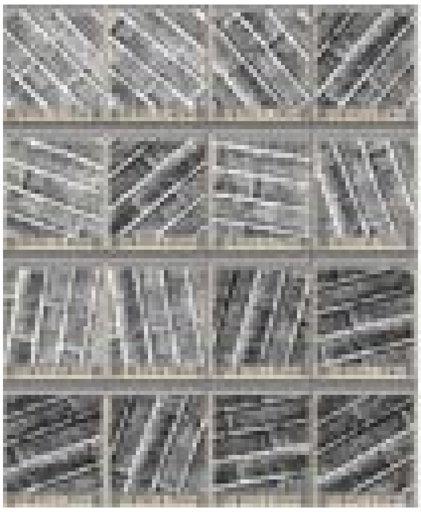


Fig. 8. An example of retrieved images in T3 data set. The upper-left image is a query image. The other 15 images are retrieved images for the query.

4. T4 Data Set

T4 data set is the scaled version of the T1 data set. One hundred and sixteen original images with matrix size of 512×512 are taken from the T1 data set. Then, the 116 images are scaled up and down by 5 %, *i.e.*, 95%, 100% and 105% scaled images are obtained. Then, 128×128 size of images are taken at arbitrary positions from those scaled images and composed of a data set which is called as T4a data set. Next images with scaling up and down with 10% are added onto the T4a data set and 90%, 95%, 100%, 105% and 110% scaled images are composed of T4b data set. Above procedure are repeated until the scales reach 50% and 150%. Then, 10 data sets are generated from T4a to T4j. The T4j data set is composed of 50% to 150% scaled images with increment of 5% scaling factor. So the total number of images for the T4j data set is 2436.

The *RR* and *AVRR* are measured in T4a to T4j data sets, respectively. 116 images with 100% scaling factor are querying images. Figure 9 shows an example of the retrieval by a querying image with the T4j data set.

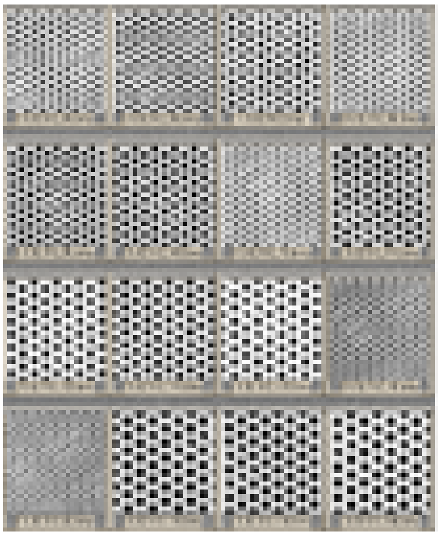


Fig. 9. An example of retrieved images in T4 data set. The upper-left image is a query image. The other 15 images are retrieved images for the query.

5. T5 Data Set

T5 data set includes images from Corel® album. The ground truth is selected by taking similar images with the querying image. The querying images are chosen so that they have relatively large texture patterns among the data sets. The number of chosen queries is 16. Total 2400 images constitute the data set. Fig. 10 shows an example of the retrieval by a querying image with the T5 data set. The query in the figure has 4 ground truths.

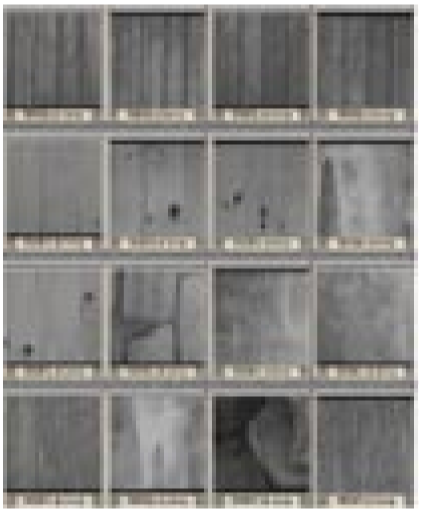


Fig. 10. An example of retrieved images in T5 data set. The upper-left image is a query image. The other 15 images are retrieved images for the query.

6. T6 Data Set

T6 data set consists of aerial images with 34,000 images of 128×128. For a query, ground truth is determined by taking similar images. Figure 11 shows an example of the retrieval by a querying image with the T6 data set.

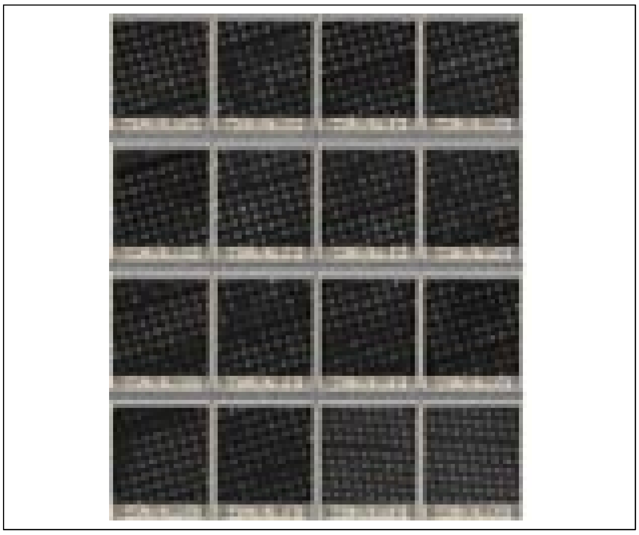


Fig. 11. An example of retrieved images in T6 data set. Upper-left image is a query image. The other 15 images are retrieved images for the query.

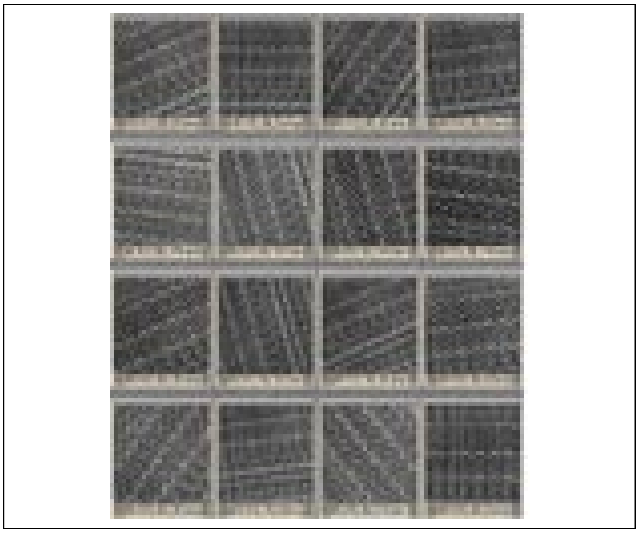


Fig. 12. An example of retrieved images in T7 data est. Upper-left image is a query image. The other 15 images are retrieved images for the query.

7. T7 Data Set

T7 data set is both scaled and rotated version derived from T1, T2, and T5 data sets. Seventy original images with matrix size of 512×512 are taken from the T1, T2 and T5 data sets such as 30 patterns from the T1 data set, 25 patterns from the T2 data set, and 15 patterns from T5 data set. The 70 images were rotated with 0, 8, 25, 55, 107, 131, and 174 degrees. Then, the rotated images are scaled with 90%, 80%, 70%, 60% and 50%. Finally, the T7 data set is constructed by taking 128×128 size image at arbitrary position from both rotated and scaled images. Total number of images in the T7 data set is 2400. The *RR* and *AVRR* are measured with 34 ground truths. The queries are 70 images with no rotation and 70% scaling. Figure 12 shows an example of the retrieval by a querying image with the T7 data set.

V. EXPERIMENTAL RESULTS

To verify the performance of the MPEG-7 texture descriptor, experiments were performed with test data sets of the homogeneous texture descriptor. These are T1, T2, T3, T4, T5, T6, and T7 data sets. The constitution of databases and performance test procedure are mentioned in the previous sections. Table 3 shows the average retrieval rates for the texture descriptor over T1, T2, T3, T4, and T7 data sets. As we can see, more than 75 % of AVRR have been achieved for T1 data set. The performance of the proposed methods can be compared to the reported results in the literatures. This is because the T1 data set is widely used for texture description-experiments. Our results were the best among participants of the MPEG-7 texture-core experiments. Furthermore, as shown in T3, T4 and T7 data sets, our proposed method shows good results for rotated and/or scaled images.

In Table 4, experimental results are shown to verify the effectiveness of the scalability of the layered feature descriptor for T1 data set. Scalable representation of feature description (the meaning of allow for is not fit here) provides the flexibility for transmission bandwidth and database storage. As shown in Table 4, 76.39 % of AVRR was obtained with the base layer in the T1 data set. Only 32 components of the descriptor were used. 77.32 % of AVRR was the result for the enhancement layer. 62 components of the descriptor were used. For the T1 data set, half of the description size could be saved with only about 1% loss of AVRR.

Furthermore, the texture descriptor is easy to compute because it is directly extracted in the frequency domain. We measured feature extraction time in a PC (Pentium II system with a 400 MHz CPU and an NT operating system). It takes around 0.14 seconds per one image query (128x128 image size).

Table 5 shows a comparison of the average retrieval rates with those of other texture descriptor extraction methods which are available in literatures. The average retrieval rates for other methods are referred to in [5]. There the same experiments

Table 3. AVRR on the unified texture descriptor.

Data set		*AVRR* (%)
T1		77.32
T2		90.67
T3		92.00
T4	T4a	86.21
	T4b	87.50
	T4c	88.94
	T4d	87.07
	T4e	84.91
	T4f	84.55
	T4g	83.87
	T4h	82.60
	T4i	79.60
	T4j	76.72
T5		60.46
T6		75.18
T7		78.66

Table 4. AVRR at the first layer and the second layer.

Layer	Features used	*AVRR*(%)
Base	$f_{DC}, f_{SD}, e(0), \ldots, e(30)$	76.39
Enhancement	$f_{DC}, f_{SD}, e(0), \ldots, e(30), ed(1), \ldots, ed(30)$	77.32

were performed with the same database of the T1 data set. The *AVRR* was reported as 74.37% in the T1 data set using the Gabor spatial filtering method. And it was reported as less than 70% in the T1 using wavelet related methods. These are the pyramid wavelet transform method (PWT) and the tree wavelet transform method (TWT) [5], [20]. With this T1 data set, the proposed algorithm of the texture descriptor extraction gives 77.32% of *AVRR*.

Table 5. AVRR on Brodaz album.

Texture descriptors	MPEG-7 texture descriptor	Gabor	PWT	TWT	MRSAR
AVRR (%)	77.32	74.37	68.70	69.41	73.18

VI. CONCLUSIONS

Texture is one of the salient features representing image contents. In this paper, we present a texture description method for images. These feature vectors are made up of an image intensity mean, a standard deviation 30 energy values and 30 energy deviations. The Polar frequency domain is partitioned based on the human visual system. From the feature channels within this domain, we can extract the energy values and energy deviations. We have shown this to be a very effective texture description. For fast and reliable feature extraction, Radon transform is used to obtain Fourier transform of the image in the Polar domain. Radon transform provides dense sampling in low frequency regions and sparser sampling away from the origin of the Polar frequency domain. This is well suited to the Human visual system. The Human visual system is more sensitive to signal variation in low frequencies and less sensitive in higher frequencies.

Our texture descriptor is compact in representation regardless of image size and is shown to be effective in relevant image retrieval. Furthermore, the intensity-, scale-, and rotation-invariant matching methods provide effective retrieval metrics for various applications.

Our proposed texture description method can be utilized to index and retrieve image and video. Some examples of applications are fast video searching and video parsing. Another example is contents-based image retrieval of aerial photos, fabric images, and electronic photo albums. The texture descriptor is a very effective way to describe object segmentation and image and video contents.

REFERENCES

[1] ISO/IEC JTC1 SC29 WG11 (MPEG), *MPEG-7 Visual part of Experimentation Model Version 4.0*, m3068, Maui, December 1999.

[2] ISO/IEC JTC1 SC29 WG11 (MPEG), *MPEG-7 Visual part of XM and WD*, N3335, Noordwijkerhout, May 2000.

[3] O. D. Faugeras and W. K. Partt, "Decorrelation methods of texture feature extraction," *IEEE Trans. Pattern Anal. and Machine Intell.*, Vol.2, July 1980, pp. 323-332.

[4] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological Cybernetics*, Vol.61, 1989, pp.103-113.

[5] B. S. Manjunath and W. Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 8, August 1996.

[6] Y. S. Kim, Y. S. Kim, W. Y. Kim and M. J. Kim, "Development of Content-Based Trademark Retrieval System on the World Wide Web", *ETRI Journal*, Vol. 21, No. 1, March 1999, pp. 40-54.

[7] R. Chellappa, "Two-dimensional discrete Gaussian Markov random field models for image processing," *Pattern Recognition*, Vol. 2, 1985, pp.79-112.

[8] P. Wu, W. Y. Ma, B. S. Manjunath, H. D. Shin and Y. L. Choi, "A texture descriptor for MPEG-7," *ISO/IEC JTC1 SC29 WG11*

(MPEG), P77, Lancaster, 1999.

[9] Y. M. Ro, “Matching Pursuit : Contents featuring for Image Indexing,” *Proceedings of SPIE*, Vol. 3527, 1998, pp. 89-100.

[10] J. R. Ohm and F. Bunjamin, “Descriptor for texture in wavelet domain,” *ISO/IEC JTC1 SC29 WG11 (MPEG)*, P566, Lancaster, 1999.

[11] A. Saadane, H. Senane and D. Barba, “On the Design of Psychovisual Quantizers for a Visual Subband Image Coding,” *SPIE*, Vol. 2308, 1994, pp. 1446.

[12] A. Saadane, H. Senane and D. Barba, “An Entirely Psychovisual based Subband Image Coding Scheme,” *SPIE*, Vol. 2501, 1995, pp.1702.

[13] J. G. Daugman, “High Confidence Visual Recognition of Persons by a Test of Statistical Independence,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.15, No.11, November, 1993, pp. 1148-1161.

[14] C. J. Lambrecht, “A Working Spatio - Temporal Model of Human Vision System for Image Restoration and Quality Assessment Applications,” *IEEE International Conference on ASSP*, New York, NY, USA, Vol. 4, 1996, pp. 2291-4.

[15] Y. M. Ro, S.Y. Kim, K.W. Yoo, M. Kim and J. Kim, “Texture descriptor using atoms for matching pursuit,” *ISO/IEC JTC1 SC29 WG11 (MPEG)*, P612, Lancaster 1999.

[16] Y. M. Ro, K.W. Yoo, M. Kim and J. Kim, “Texture Description using Radon transform,” *ISO/IEC JTC1 SC29 WG11 (MPEG)*, m4703, Vancouver, 1999.

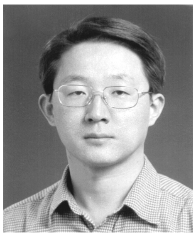
[17] Y. M. Ro, K.W. Yoo, M. Kim and J. Kim, “Texture description using Radon transform and experimental results on CT-5 core experiment using atoms for matching pursuit,” *ISO/IEC JTC1 SC29 WG11 (MPEG)*, m5152, Melbourne, 1999.

[18] Y. M. Ro, K.W. Yoo, M. Kim, J. Kim, B. S. Manjunath, D. G. Sim, H. K. Kim and J. R. Ohm, “An unified texture descriptor,” *ISO/IEC JTC1 SC29 WG11 (MPEG)*, m5490, Maui, 1999.

[19] J. G. Daugman, “Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression,” *IEEE Trans. ASSP*, Vol.36, July 1988, pp. 1160-1179.

[20] T. Chang and C. J. Kuo, “Texture analysis and classification with Tree-Structured Wavelet Transforms,” *IEEE Trans. Image Processing*, Vol.2, No.4, Oct. 1996, pp. 429-441.

Yong Man Ro received the B.S. from Yonsei University, Seoul, Korea in 1981 and the M.S. and Ph.D. degrees from the Korea Advanced Institute in Science and Technology (KAIST), in 1987 and 1992, respectively. In 1987, he was a staff associate at Columbia University, and from 1992 to 1995, he was a visiting researcher in University of California at Irvine and KAIST. In 1996, he was a research fellow at department of electrical engineering and computer sciences in University of California at Berkeley. In 1997, he joined Information and Communication University, Korea where he is currently associate professor and director of Image Video System Lab. His research interests include image/video processing, MPEG-7, feature recognition, image/video indexing, and spectral analysis of image signal. He received the Young Investigator Finalist Award in ISMRM in 1992. He is a senior member of IEEE and member of SPIE and ISMRM.



Munchurl Kim has received the B.E. degree in electronics from Kyungpook National University, Korea in 1989, and M.E. and Ph.D. degrees in electrical and computer engineering from University of Florida, Gainesville, USA, in 1992 and 1996, respectively. After his graduation, he joined Electronics and Telecommunications Research Institute (ETRI) where he had worked in the MPEG-4 standardization related research areas. Since 1998, he has been involved in MPEG-7 standardization works. In the course of MPEG standardization, he has been contributing more than 30 proposals in the areas of automatic/semi-automatic segmentation of moving objects, MPEG-7 visual descriptors and Multimedia Description Schemes, and served as the team leader on evaluation of Video Description Scheme proposals in MPEG-7 in Lancaster U.K., 1999. In 2001, he joined, as assistant professor in school of engineering, the Information and Communications University (ICU) in Taejon, Korea. His research areas of interest include multimedia computing, communications and broadcasting, and multimedia interactive services.

Ho Kyung Kang received the B.S. degree in electronic engineering from Korea University, Korea, in 1998 and M.S. degree in image processing from Information and Communication University, Korea, in 2000. Since 2000, He has been a Ph.D. candidate in the same university. His research interests include contents-based multimedia information retrieval, watermarking and image/ video processing.

Jinwoong Kim received the B.S. and the M.S. degrees from Seoul National University, Seoul, Korea, in 1981 and 1983, respectively, and the Ph.D. degree in the department of electrical engineering from Texas A&M University, United States in 1993. Since 1983, he has been a research staff in Electronics and Telecommunications Research Institute(ETRI), Korea. He is currently a director in the broadcast media technology department. He has been engaged in the development of TDX digital switching system, MPEG-2 video encoder, HDTV encoder system, and MPEG-7 technology. His research interests include digital signal processing in the field of video communications, multimedia systems, and interactive broadcast systems.