Computing and Informatics, Vol. 27, 2008, 757–767

SVM BASED INDOOR/MIXED/OUTDOOR CLASSIFICATION FOR DIGITAL PHOTO ANNOTATION IN A UBIQUITOUS COMPUTING ENVIRONMENT

Chull Hwan Song, Seong Joon Yoo

School of Computer Engineering, Sejong University, 98, Gunja, Gwangjin
Seoul, Korea
e-mail: sjyoo@sejong.ac.kr

Chee Sun Won

Department of Electronic Engineering
Dong Guk University, Seoul, Korea

Hyoung Gon Kim

Imaging Media Research Center, KIST, Seoul, Korea

Revised manuscript received 10 January 2008

Abstract. This paper extends our previous framework for digital photo annotation by adding noble approach of indoor/mixed/outdoor image classification. We propose the best feature vectors for a support vector machine based indoor/mixed/outdoor image classification. While previous research classifies photographs into indoor and outdoor, this study extends into three types, including indoor, mixed, and outdoor classes. This three-class method improves the performance of outdoor classification. This classification scheme showed 5–10 % higher performance than previous research. This method is one of the components for digital image annotation. A digital camera or an annotation server connected to a ubiquitous computing network can automatically annotate captured photos using the proposed method.

Keywords: Image classification, support vector machine, low-level feature extraction

1 INTRODUCTION

The term "ubiquitous computing environment" describes the fact that the user can reach the network through a computer freely at any time and in any place. Todays digital devices commonly include cameras and internet connection functions. People want to take digital photographs frequently, which produces numerous photographs through a lifetime. It would be quite tedious to annotate the images one by one. The present study describes an automatic image annotation system and explains its structure.

The application of image classification technologies is recommended for the annotation of digital photos in a simple and rapid manner. In our previous study [1], we used support vector machine (SVM) [2] to identify methods to classify seven nature objects. More to the point, the previous study utilized 9 different feature vectors to compare image classification performances. A comparison of the findings of that particular study with those of previous research revealed that the 576 bin joint histogram fared best in terms of performance. The current study revolves around the introduction of methods to heighten digital photo annotation performance created through the addition of indoor/mixed/outdoor classification functions to the image classification technology suggested in [1]. The performance of the indoor/mixed/outdoor classification technology was benchmark-tested using the data set employed in the previous study [3]. Overall performance was found to have improved by some 5–10 %. A digital photo annotation system with a structure created through the combination of the seven nature scene object classification methods enumerated in the previous study with the indoor/mixed/outdoor classification method introduced in the current study may result in the emergence of Figure 1.

2 RELATED WORK

Most of researches on image annotation have been limited to specific objects. A representative case is a natural scene annotation. Among researches on natural scenes, Fredembach et al. [3] formed eigen regions using principal component analysis (PCA) for the image classification. Specifically, 3 types of objects were defined – skin, vegetable and blue sky. The number of types classified in the research is smaller than our research, which identifies 7 types of objects. Researches by Vogel et al. [4] and Ren et al. [5] annotated natural scenes using semantic model step algorithm and SVM. However, like [3], these researches are applicable only to natural scene images, unlike our proposed method. Research by Cusano et al. [6], which is also for only nature scenes, annotates nature objects into 7 types and, for the classification, forms HSV

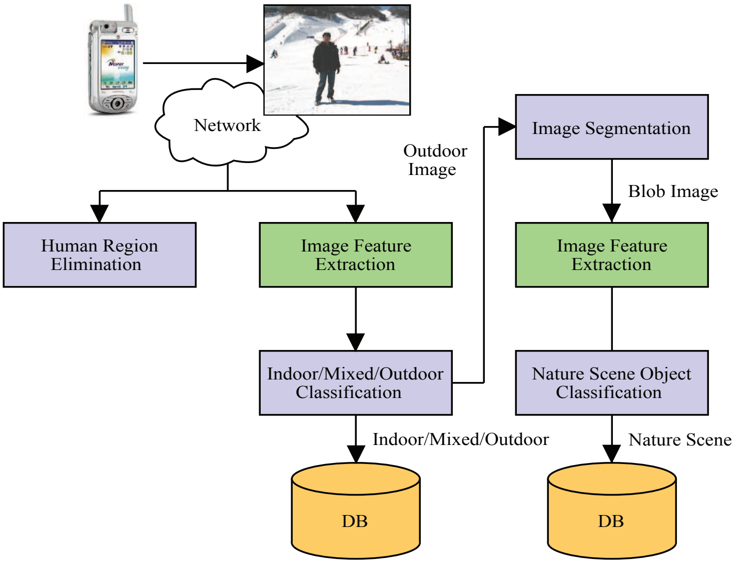


Fig. 1. A digital photo annotation with a new indoor/mixed/outdoor image classification method

color histograms and gradient histograms and applies them to SVM. [6] is similar to our research in terms of natural scene annotation, but is different in the application of image segmentation and the composition of input vector. Experimental results demonstrate that our proposed method shows better performance over the one in [6].

For indoor/mixed/outdoor classification, Szummer et al. [7] inferred high-level scenes from low-level image features. Fitzpatrick [8] extracted image features such as color, texture, and DCT coefficients and applied them to K-nearest neighbor algorithm. Payne et al. [9] classified indoor/mixed/outdoor scene images through a 2-pass classification system using contour tracking and straightness. Traherne et al. [10] extracted the line strength of image edges using an edge detection algorithm and, like [8], applied the strength to the K-nearest neighbor algorithm. Previous researches studied binary classification, which distinguishes between indoor and outdoor scenes. However, in this paper, we make a ternary classification of indoor/mixed/outdoor scenes. We also demonstrate the superiority of our method through an experiment that compares the proposed method with [10].

3 FEATURE VECTOR MODEL

We used 9 feature vectors in our previous study [1] for image classification. This chapter summarizes these feature vectors again.

3.1 Color and Edge Feature Vector Model

To express colors, the hue-saturation-value (HSV) color space is used. HSV color space is less sensitive to illumination than RGB and can separate HS (color element) easily. In [11], an 11-bin histogram was composed using HSV for the color feature. We expand it to 36 bins. The bins are composed of a 36-bin HSV color features such that each of 12 bin colors – 'black', 'grey', 'white', 'red', 'orange', 'yellow', 'green', 'cyan', 'blue', 'purple', 'magenta' and 'brown' – is divided into three bins. The 36-bin histogram can obtain various color ranges easily and is normalized into values between 0 and 1. In order to extract an edge-based histogram, we apply the Canny edge detection algorithm [12]. The Canny edge detection algorithm removes the noise of the original input image through Gaussian smoothing filtering. That is, the gray level at a pixel (m,n), $f(m,n)$ is first smoothed by the following Gaussian filter $G_\sigma(m,n)$

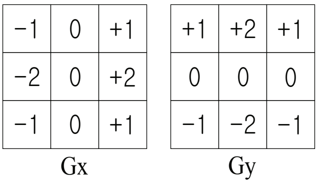
$$g(m,n) = G_\sigma(m,n) * f(m,n) \tag{1}$$

where

$$G_\sigma(m,n) = \frac{1}{2\pi\sigma^2}\exp\Big[-\frac{m^2+n^2}{2\sigma^2}\Big]. \tag{2}$$

After Gaussian smoothing filtering, we apply the Canny edge detection algorithm with masks such as Sobel, Priwitt and Roberts. Here, we use the Sobel mask (see Figure 2) that can detect edges relatively more efficiently. Specifically, for the horizontal and vertical elements, we have the edge magnitude on the pixel as follows:

$$M(m,n) = \sqrt{g_h^2(m,n) + g_v^2(m,n)}. \tag{3}$$



−1	0	+1
−2	0	+2
−1	0	+1

Gx

+1	+2	+1
0	0	0
−1	−2	−1

Gy

Fig. 2. Sobel mask

Also, we can obtain the direction of an edge as follows:

$$\theta(m,n) = tan^{-1}\Big[-\frac{g_v(m+n)}{g_h(m,n)}\Big]. \tag{4}$$

In Equation (4), we can obtain 4 edge directions based on 0° (horizontal direction), 45° (positive diagonal), 90° (vertical direction) and 135° (negative diagonal).

So far, we obtain magnitude and edge direction elements from the Canny edge detection algorithm using the *Sobel* mask. Now, we can form a histogram for the edge features. Specifically, 4 bins are composed for each of g_h and g_v in Equation (4), and 4 bins are composed using the 4 types of directions in θ value in Equation (4). Thus, an edge histogram with a total of 12 bins is formulated. Figure 3 shows the composition of the edge histogram.

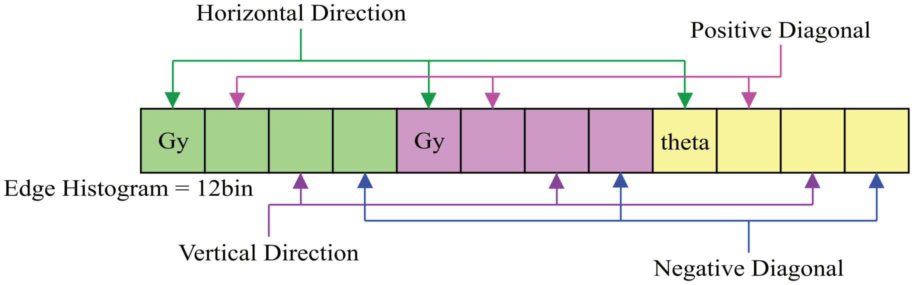


Fig. 3. Composition of edge histogram

The preprocessing to be applied to image classification is the same in two-stage classification models. This process is used ultimately to be applied to SVM. In the classification of indoor/mixed/outdoor (1st stage classification), low-level image features are extracted from the entire image and are used in classification without applying the image segmentation algorithm [13]. On the contrary, in natural scene classification (2nd stage classification), low-level image features extracted from segmented images are used in the classification.

3.2 MPEG-7 Visual Descriptor

MPEG-7 [15] features various types of visual descriptors such as color, texture, and shape. To this end, this study will use the color layout, edge histogram, color structure, homogeneous texture, region based shape, and scalable color descriptor. The color layout descriptor (CLD) can, when using a Y, Cb or Cr color space, be obtained by applying a DCT transform to the 2-D array of a locally re-expressed color. Meanwhile, the edge histogram descriptor (EHD) re-expresses local edge distribution. What's more, it consists of five types of histograms, namely vertical, horizontal, 45 diagonal, 135 diagonal, and non-directional edges. For its part, the color structure descriptor (CSD) re-expresses images by making use of the color distribution and local spatial structure of the relevant colors. The homogenous texture descriptor (HTD) characterizes the local texture by using the mean energy and energy deviation taken from the frequency channels. Here, the 2-D frequency channels should be regarded as consisting of 30 channels, with the mean energy and energy deviation calculated for each. The region based shape descriptor (SD) analyzes the shape created by the boundary and interior's pixels. This descriptor is in turn defined by the angular radial transform (ART). The scalable color descriptor (SCD) can be interpreted with an encoding scheme (Harr transform coefficient encoding) developed based on the Harr transform, where the value of the color histogram is applied to the HSV color space. Normalized values of 0 and 1 should be applied to each of these features.

3.3 Composite Feature Vector Model

In this section, we explain how the color and the edge histograms are combined to form a composite feature vector. There are two methods for the vector formation, namely combined input vector and joint input vector.

3.3.1 Combined Input Vector

Let us denote the normalized color histogram with 48 bins as $C = \{c_1, c_2, \ldots, c_{48}\}$ and the edge histogram with 12 bins as $E = \{e_1, e_2, \ldots, e_{12}\}$. Then, a simple composite histogram is the union of the two histograms such that the composite histogram has the form of

$$x = \{x_1, x_2, \ldots, x_{60}\} = \{c_1, \ldots, c_{48}, e_1, \ldots, e_{12}\}. \tag{5}$$

Note that the elements of equation (5) are normalized into values between 0–1.

3.3.2 Joint Input Vector

Another method to express the two histograms is to jointly consider the two elements as a pair of the histogram bins, (c_i, e_j) , at the same time. Then, we have the joint histogram as

$$x = \{f(c_1, e_1), f(c_1, e_2), \ldots, f(c_{48}, e_{12})\} \tag{6}$$

where $f(c_i, e_j)$ is the frequency that the two attributes (c_i, e_j) appear at the same time. Thus, the joint histogram is composed of $36 \times 12 = 432$ bins, which requires around 9 times larger input data and much longer input vector formation time than the combined histogram in equation (5).

4 INDOOR/OUTDOOR CLASSIFICATION USING SUPPORT VECTOR MACHINE

The feature vectors explained in Section 3 are used as input vector for SVM to classify the images. Specifically, the color and the edge histograms are formed from the unsegmented indoor/mixed/outdoor image and are applied to SVM. On the contrary, for natural scene images, features are extracted from blob images. Then, the same process as that for indoor/mixed/outdoor images is applied to SVM for the classification with various classes. Basically, SVM is a supervised learning and binary classification (±1) algorithm with the following equation

$$f(x) = sign(w^T x + b) \tag{7}$$

where w and b are the parameters of hyper-plane, and $f(x)$ has a value of ±1 for an input vector x. Since our input vectors are not linearly separable, SVM uses a mapping function (Φ) in order to expand the low-level input vector space to a high-level one and then classifies by finding the largest margin using the support vector. Here, various types of kernel functions $\Phi()$ such as linear and radial basis function(RBF) are applied.

$$K(u, v) = \Phi(u) \cdot \Phi(v) \tag{8}$$

Then, a decision can be made by the function defined as follows

$$f(x) = \text{sign}\left(\sum_i y_i \alpha_i K(x, x_i) + b\right) \tag{9}$$

where y_i are the binary classes and α_i is the weight parameter. Based on this binary classification, the process is expanded to multi-classification. For this, we use LIBSVM library [14]. The weak point of SVM is that its classification time is somewhat longer than other types of machine learning but is advantageous in that its classification performance is higher than other classifiers. Using this, individual digital images are expanded to semantic words of specific class. In application for classification, '*2-stage classification model*' is adopted. That is, for the whole image, three types of classes (indoor/mixed-outdoor/nature-outdoor-scene) are classified (Stage 1), and for images classified as '*outdoor*', blob images to which the image segmentation algorithm has been applied for the classification with 7 classes (Stage 2), namely '*Sea*', '*Vegetable*', '*Ground*', '*Sky*', '*Maple*', '*Glow*' and '*Snow*'. Based on the classes, digital photographs are annotated.

5 PERFORMANCE EVALUATION

In our experiments, we formed and applied three types of input vectors. The first is an input vector that has separate color and texture attributes, the second is a combined input vector composed through Equation (5), and the third is a joint input vector that expresses two low-level image features at the same time as in Equation (6). These vectors were used as input vectors of SVM. Our experiments in Section 3 show that the combined input vector, which is simple in structure and has two attributes, is superior in performance.

The indoor/mixed/outdoor experiment was conducted in two ways. First, indoor and outdoor scenes went through binary classification. Then, the results of ternary classification for indoor, mixed outdoor, and natural outdoor scenes are presented. The 11-bin HSV histogram [11] explained in Section 3 is expanded to 36-bin HSV histogram and this is combined with 12-bin edge histogram and combined 48-bin histogram are formed. In addition, 6 MPEG-7 visual descriptors [15] are composed and applied to SVM. Traherne et al. [10] experimented with 2-class (Type A) composed of indoor images and outdoor images without man-made objects, and another

2-class (Type B) composed of indoor images and outdoor images containing man-made objects and nature objects. We performed ternary classification including these experiments. The results showed the concept of 3-class (Type C) composed of indoor images, outdoor images containing man-made objects and nature objects, and nature outdoor images without man-made objects. Here we used a total of 150 datasets. Of them, 75 were training datasets, and 75 were test datasets.

First, in Type A, when the color structure of MPEG-7 visual descriptors was used, high classification performance was observed for both indoor and outdoor test datasets (100 %). Second, in Type B, when the color structure of each MPEG-7 visual descriptor and scalable color descriptors were applied, the classification rate was 100 % for indoor test datasets and 98 % for outdoor test datasets. Third, Type C was multi-class (3-class) classification, which is not in the research of Traherne et al. [10]. The result of this case showed highest performance when the color descriptor of each MPEG-7 visual descriptor and 48-bin histogram(combined HSV+Canny edge histogram) were used. Figure 4 shows the result of our experiment using these datasets.

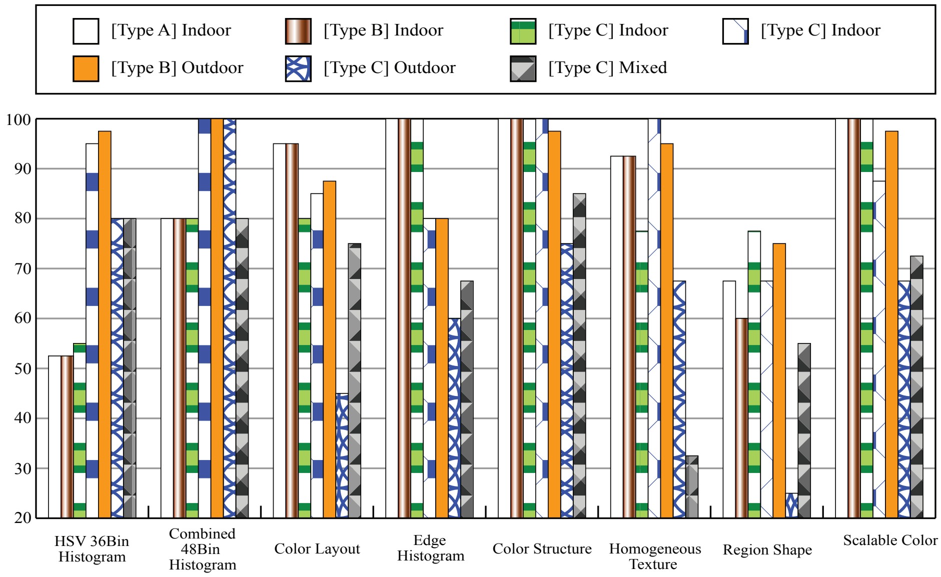


Fig. 4. Results of experiments with indoor/mixed/outdoor scenes using various image features

Figure 5 compares our results with the result of [10] which used the same datasets as ours. In the figure, [10] classified Type A around 95 % but in our research the classification rate was 100 % when the color structure descriptor of MPEG-7 visual descriptor was used. In addition, in case of Type B, [10] showed performance of 88 % but in our method the classification rate was 98.7 % when color structure and scalable descriptors were used and 93.3 % when combined 48-bin histogram was used, showing higher performance than [10].

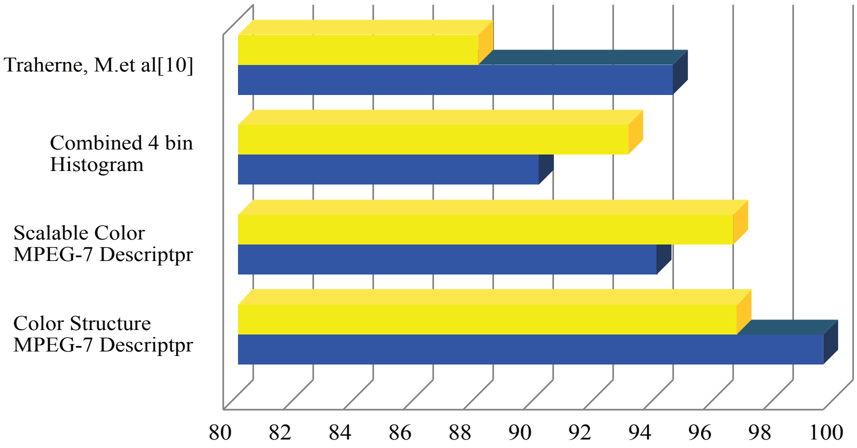


Fig. 5. Comparative classification accuracy for indoor/mixed/outdoor scenes using benchmark datasets

6 DISCUSSION AND CONCLUSION

In this study, a new approach to indoor/mixed/outdoor image classification was suggested for application using digital photo annotation under a ubiquitous computing environment. SVM was utilized in order to derive a more exact means of classification. The selection of feature vectors represents an important element in terms of heightening the precision of classification when using SVM. This study conducted a comparative analysis of the performances on the indoor/mixed/outdoor image classification using eight MPEG-7 descriptors as feature vectors. To this end, the edge histogram, scalable color histogram, color structure, and combined 48 bin histogram were found to exhibit relatively good performances. As far as indoor/outdoor scene classification was concerned, this study proved that the performance was more outstanding when using the Scalable Color MPEG-7 Descriptor and Color Structure MPEG-7 Descriptor as feature vectors than the performance using the method developed by Trahene et al. [10]. Data identical to that used in the method developed by Trahene et al. [10] was employed in this experiment.

Acknowledgements

We wish to thank Prof. Yong-Guk Kim and Mr. Young-Hyun Kim for their guidance, encouragement and support throughout this research.

This work was supported by the IT R & D program of MIC/IITA. [2006-S-032-01, Development of an Intelligent Service technology based on the Personal Life Log]

REFERENCES

[1] Song, C. H.—Yoo, S. J.—Won, C. S.: SVM Based Classification of Seven Nature Objects for Anytime, Anywhere Digital Photo Annotation. International Conference on Multimedia and Ubiquitous Engineering (MUE '07), Seoul, 2007, pp. 1249–1254.

[2] Vanpik, V.: Statistical Learning Theory. Willy, New York, 1998.

[3] Fredembach, C.—Schroder, M.—Susstrunk, S.: Eigenregions for Image Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, 2004, pp. 1645–1649.

[4] Vogel, J.—Schiele, B.: Natural Scene Retrieval Based on a Semantic Modeling Step. CIVR 2004, LNCS 3115, 2004, pp. 207–215.

[5] Ren, J.—Shen, Y.—Ma, S.—Guo, L.: Applying Multi-class SVMs into Scene Image Classification. IEA/AIE 2004, LNAI 3029, 2004, pp. 924–934.

[6] Cusano, C.—Ciocca, G.—Schettini, R.: Image Annotation Using SVM. Proceedings of the SPIE, Vol. 5304, 2003, pp. 330–338.

[7] Szummer, M.—Picard, R. W.: Indoor-Outdoor Image Classification. IEEE International Workshop on Content-based Access of Image and Video Databases (CAIVD 98), Bombay, India, 1998, pp. 42–51.

[8] Fitzpatrick, P.: Indoor/outdoor scene classification project. Pattern Recognition and Analysis. Available on: `http://people.csail.mit.edu/paulfitz/pub/indoor-outdoor.pdf`.

[9] Payne, A.—Singh, S.: A Benchmark for Indoor/Outdoor Scene Classification. ICAPR 2005, LNCS 3687, 2005, pp. 711–718.

[10] Traherne, M.—Singh, S.: An Integrated Approach to Automatic Indoor Outdoor Scene Classification in Digital Images. IDEAL 2004, LNCS 3177, 2004, pp. 511–516.

[11] Cox, I. J.—Miller, M. L.—Omohundro, S. M.—Yianilos, P. N.: Target Testing and the PicHunter Bayesian Multimedia Retrieval System. In the Proceedings of the 3rd Forum on Research and Technology Advances in Digital Libraries, DL '96, 1996, pp. 66–75.

[12] Canny, J.: A Computational Approach to Edge Detection. IEEE Transaction Pattern Analysis Machine Intelligence 86, 1986, pp. 679–698.

[13] Felzenszwalb, P. F.—Huttenlocher, D. P.: Efficient Graph-Based Image Segmentation. Intl. Journal of Computer Vision, Vol. 59, 2004, No. 2, pp. 167–181.

[14] Chang, C. C.—Lin, C. J.: LIBSVM: A Library for Support Vector Machines. Available on: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[15] ISO/IEC JTC1/SC29/WG11/N4579: DTR 15938-8 Extraction and Use of MPEG-7 Descriptors – Part 3 Visual. Jeju, March 2002.

Chull Hwan Song received the B. Sc., M. Sc. degrees from the School of Computer Engineering, Sejong University, Seoul, Korea, in 2002, 2004. He is currently working toward the Ph. D. degree in computer engineering at same university. His research interests are in machine learning, image processing, information retrieval and image classification/recognition.

Seong Joon Yoo received the Ph. D. degree in computer and information science from Syracuse University, Syracuse, New York, in 1996. Since 2002, he has been with the School of Computer Engineering, Sejong University, where he is currently a Professor. Before joining the university, he was a research member of senior technical staff at the ETRI, Korea. His research interests include learning systems, pattern recognition, data mining, and image processing.

Chee Sun Won received the B. Sc. degree in electronic engineering from Korea University, Seoul, in 1982 and M. Sc. and Ph. D. degrees in electrical and computer engineering from the University of Massachusetts at Amherst in 1986 and 1990, respectively. From 1989 to 1992 he was a senior engineer with GoldStar Co., Ltd. (LG electronics), in Seoul. In 1992, he joined Dongguk University, where he is currently a Professor in the Electronic Engineering Department. He was a Visiting Associate Professor at Stanford University from July 2001 to August 2002. His research interests include MRF image modeling, image segmentation, content-based image retrieval, and image watermarking.

Hyoung Gon Kim received the B. Sc. in electronic engineering from National Korea Civil Aviation College, Seoul, in 1974, and M. Sc. and Ph. D. in Electronics from University of Kent at Canterbury, England in 1982 and 1985, respectively. His current research interest includes multi-modal sensory information system, tele-immersive augmented reality systems, real-time visual tracking and tangible space applications. Currently he is working with Imaging Media Research Center at KIST.