Visual Modeling and Content-based Processing for Video Data Storage and Delivery

Jae-Jeong Hwang and Sang-Gyu Cho, *Member, KIMICS*

Abstract—In this paper, we present a video rate control scheme for storage and delivery in which the time-varying viewing interests are controlled by human gaze. To track the gaze, the pupil's movement is detected using the three-step process : detecting face region, eye region, and pupil point. To control bit rates, the quantization parameter (QP) is changed by considering the static parameters, the video object priority derived from the pupil tracking, the target PSNR, and the weighted distortion value of the coder. As results, we achieved human interfaced visual model and corresponding region-of-interest rate control system.

Index Terms—HCI, face detection, pupil detection, gaze tracking, rate control, content-based coding, object-based coding

I. INTRODUCTION

Most of useful information might be obtained through the eyes. Gazing attitude reflects human intentions and desires that can be achieved by HCI (Human Computer Interaction), detecting and tracking the gaze direction. In order to detect gaze direction in the image comprising a complex background and a human face, the first stage is to detect the face and the second is to detect the eye region in the face image. This method is more effective than that acquiring eye region directly[1].

Recently, the face detection techniques have been developed since Chellappa's research[2]. Some useful results have been published in literature[3-10]. Some of the techniques were tested on the big database, e.g., example-based learning[2], neural network[7] and feature-based[4,5,13].

The detection of the eye region has been studied using the geometrical connection[8], PCA (Principal Component Analysis)[9] and template matching[10,11]. Although there are many techniques to detect the eye gaze direction that have been suggested, none of them shows the perfect capability satisfied in all the situations.

Rate control techniques have been intensively studied for various standards and applications, such as H.261, H.263[14], MPEG-1, MPEG-2[15] and MPEG-4[16]. For different coding schemes, different coding parameters may be employed. In most of standards, the most influential parameter with regard to picture quality is the quantization parameter used for texture coding. It is selected based on a measure of the buffer fullness so that the target bit rate can be achieved. The rate control can not resort to changing the temporal coding parameter for buffer control. In contrast to this, H.263 scheme does allow enhanced techniques such as variable frame skip to avoid overflow or underflow.

MPEG-4 rate control considers spatial and temporal coding parameters. However, since MPEG-4 allows the coding of arbitrarily shaped objects, the encoder must consider the significant amount of bits which are used to code the shape information[17,18]. Therefore rate control is an important issue in most of object-oriented schemes.

In MPEG-4, objects are coded separately, called video object plane (VOP) made up of multiple video objects (MVO). They can be natural or synthetic video. Each VO is individually coded and corresponds to an elementary bit-stream that can be individually accessed and manipulated, while composition information is sent in a separate stream. An instance of a VO at a given time frame is called a VOP.

However image properties often vary from time to time are not homogeneous and fail to get correct boundary of object. Size or moving speed of an object can be changed from frame to frame, resulting in failed model adaptation. Therefore we design a human eye interfaced bit rate controller which will be used for segmenting the VOs, importance of a certain VO, and determining coding bit rates.

In subsequent sections, we analyze the gaze tracking algorithm and the human interfaced video rate control scheme, i.e., content-based video data processing controlled by human interests. The three-step gaze tracking system is proposed in Section II. Resulting human interest region is considered as weighting factor for video rate control which is one of the most important factor in a video coder in Section III. Some results are simulated in Section IV and concluded in Section V.

II. GAZE TRACKING SCHEME

Detection of pupils in the moving pictures can make problems of spending much time on a real-time system and falling in an unstable state. Those are solved by the hierarchical approach from detecting the face region to the pupil and gaze point as shown in Fig. 1. After getting the center of the pupil, the gaze points are tracked on the screen.

Manuscript received February 10, 2005.

Jae-Jeong Hwang and Sang-Gyu Cho are with the School of Electronic and Information Engineering, Kunsan National University, 573-701, Republic of Korea (Tel : +82-63-469-4855, Fax : +82-63-469-4699, Email : hwang@kunsan.ac.kr, sgcho@kunsan.ac.kr)

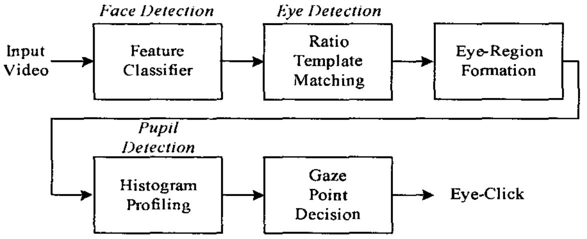


Fig. 1 Block diagram of the gaze tracking scheme.

A. Face, eye, and pupil detection

There are many kinds of faces of size and position in an image. Detecting the faces uses the sub-sampling method with the (24×24) sub-image. The pyramid method[7] decreases the original image size to get a smaller image with a certain ratio (1.1, 1.2, etc.) until the specific image size is achieved. The sub-image and the features prototypes[4,5] are passed to AdaBoost[13] technique to judge the face existence in an image. Then, the sub-image is judged to be a face region if the output, which is resulted from the AdaBoost, is over the particular threshold. The black part of the prototypes is the negative weight and the white part is the positive weight.

Eye detection process is followed. The detection of specific regions in the face is generally possible by using the intensity information of the gray scale images. For example, eyebrows and pupils in a face are darker than the other regions. These characteristics of the intensity can be used to classify each face region. Sinha[11,12] proposed the face ratio-template which is statistically analyzed and modeled.

A facial image is divided into 16 regions and set the ratio of the regions (eyes, nose, mouth, etc). The eye's ratio is 5:3, but if this ratio is adopted to detect the pupil, errors are experienced because the rectangle of eye region may include the eyebrows. To avoid this mismatching problem, the eye-region ratio is changed to 4:3 that do not include eyebrows' region. The proposed template detects only the eye regions discarding eyebrows.

The eye image obtained by the proposed ratio-template is colored image, which takes the information of RGB. Since the process is in gray scale image, it is easier and faster than the color image. The RGB image is converted to the gray scale image. Because the gray-scaled eye image includes noise of eye-brush and Purkinje image etc., significant errors may happen to the system. Threshold (T), which is dissimilar in all eye images, is taken between two peak histogram values as depicted in Fig. 2.

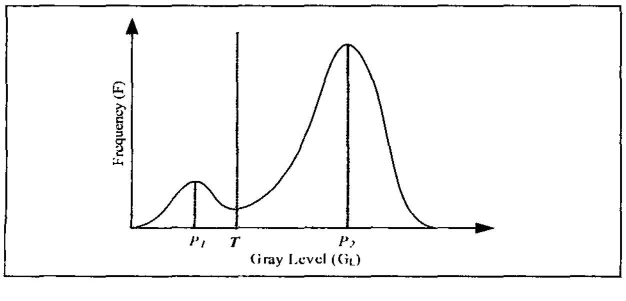


Fig. 2 Example of a histogram in the eye region.

If each pixel in the gray-scaled image is greater than the particular threshold, the pixel value is replaced to 255, otherwise 0, as defined by:

$$B(x,y)=\begin{cases}0, & G(x,y)\geq T\\ 255, & G(x,y)<T\end{cases} \tag{1}$$

The dilation processing is useful to expand the outside pixels of the object. The size of object is expanded, while the size of background is contracted, so that the empty spaces are filled and the disconnected spaces are connected in the pupil. The erosion processing gets rid of too small objects and increases the background in the image.

One of approaching methods to detect object's position in an image is to decide the image projection profiles. The profile in the image can be obtained by the sum of ON pixels taken across each axis. The ON pixels are composed of the objects.

Frequently, the method of using the image profiles is the vertical and horizontal profile. This profile is the sum of ON pixels on each row or column in the image. Then, if $I(x,y)$ is the input image, $H_v(i)$ is the sum of ON pixels of the i-th column while $H_h(i)$ is the sum of ON pixels of the j-th row.

$$\begin{aligned}H_v(i)&=\sum_{k=1}^{V} I(x,k)\\ H_h(i)&=\sum_{k=1}^{H} I(k,y)\end{aligned} \tag{2}$$

An image can be scanned to get profiles in both directions: vertical and horizontal. The profile of the vertical and the horizontal direction takes the half circle shape in an eye image as shown in Fig. 3.

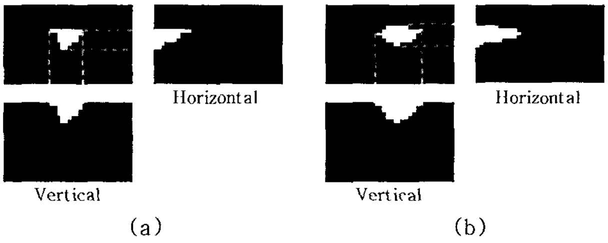


Fig. 3 Example of histogram profiling for (a) left eye and (b) right eye.

B. Gaze direction tracking

Getting the center of the pupil could be the preprocessing to get the gaze direction. So it is adapted to the change of gaze on the screen using the change of the motion vector of the pupil's center.

Even though the vector of the center of the pupil can be tracked using the PC camera, there are many problems on how far between the screen and eyes, how to compensate the head motion and so on. First of all, two points are arbitrary selected by mouse clicks for the initialization of relative displacement on the screen to be decided by the pupil movement as in Fig. 4.

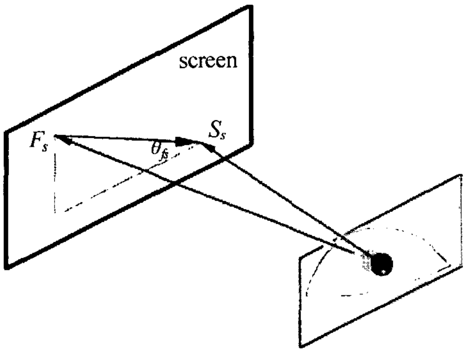


Fig. 4 Initialization for the gaze tracking process.

At first, get the first temporary point, $F_s(x_1, y_1)$, on the screen as doing the mouse click and synchronously take the first center of the pupil, $F_e(x_3, y_3)$, in one of eye regions. Next, with the same measurement, get the second point, $S_s(x_2, y_2)$, to the temporary diagonal direction and center of the pupil changed, $S_e(x_4, y_4)$. Each scale is computed by Eq. 3 and 4. Each direction (θ) is computed by Eq. 5.

$$S_{vs} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3}$$

$$E_{vs} = \sqrt{(x_4 - x_3)^2 + (y_4 - y_3)^2} \tag{4}$$

$$\theta_{fs} = \tan^{-1}\frac{x_2 - x_1}{y_2 - y_1} = \tan^{-1}\frac{x_4 - x_3}{y_4 - y_3} \tag{5}$$

In getting the information, the vectors of the center changed and the points on the screen can be obtained. If the screen would be flated and the center of the pupil would be moved to a straight line, these vector's attributes are of same direction and different scale so the coefficient which makes each other vector scale will be obtained the same. This coefficient means the distance and the angle between the screen and eyes. The coefficient(α_{se}) is computed by Eq. 6.

$$\alpha_{se} = \frac{S_{vs}}{E_{vs}} \tag{6}$$

As the proposed measurement, if the coefficient is adapted to the vector of the pupil, it is possible to track the gaze on the screen.

III. MVO RATE CONTROL

The MVO coding is performed for each VO instead of coding a sequence frame by frame, and the coding is made adaptive on an object by object basis that can be considered as a separate coding process. It has to meet overall constant bit rate and every coding process has to be related to each other.

The distribution of the target bits for each VO is done according to certain human interests or interactivities. Generally human eye tends to follow an important object instantaneously. We describe this in terms of semantic importance as : foreground objects are normally moving while background is mainly still and foreground objects receive most attention from the observer while background attract less attention.

Through the whole image sequence, significant objects appear and disappear from frame to frame. The systems discussed in the previous section can not adaptively respond to the time varying frame properties. Parameters are predetermined passively by the user and the available bits are allocated constantly. Generally target bits are calculated and the quantization parameter is varied in the coder.

In Fig. 5, we propose a human eye interfaced video coding scheme. Eye movement is detected by a camera installed in front of an observer. The observing point can be different from person to person and from time to time. Visual importance for a VO represented as a coordinate in a frame is obtained in the process. The importance can be converted to a priority of a VO in the coder and input to the block of quantization and encoding process. Another input can be statistics based target bits calculated using the statistical parameters, size, motion and distortion. Therefore the scheme can be considered as a combination of priority-based and statistical-based schemes. The main objective is to acquire human visual importance or region of interest (ROI) of each VO for each frame and to apply it to the coder.

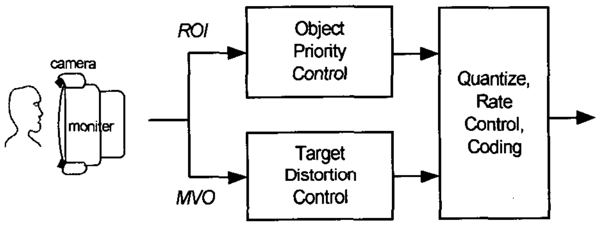


Fig. 5 Human eye interfaced encoding process. ROI and MVO are input to the encoder.

A. Visual importance measurement

The proposed gaze detection system detects gaze position in steps. First, gaze direction is initialized at the center of a screen, setting the motion vector to zero. Next, the pupil area is detected from the captured image. The image is converted to a binary image to detect the pupil area. Then, the gaze position is computed from the changed angle from the previous position. Binarization is used to make easy for the eye detection. Position is represented by pixel coordination and the motion vector for the pupil is obtained using the coordinate. We define the area pointed by human eye as an ROI in the frame if the gaze position is inside of a significant object, that is, boundary of image frame or outside may not be important.

In general, the number of MVO in an image is not large, e.g., 3-4 object areas are enough to detect those significant objects. It is difficult to detect if the size of object is too small, since the detection system is not as accurate and viewing angle is too small. The smaller the number of image pixels (low resolution), the more difficult to detect eye movement.

B. Calculation of target distortion

We define distortion ratios P_i among VOs in terms of the PSNR, to regard that the target distortion. The priority can be varied by the object property in time as

$$P_i = w_s SIZE_i + w_m MOT_i + w_v VAR_i \qquad (7)$$

where all parameters are defined as in Section 2. The value of P_i is the relative value to compute the PSNR of a VO. Sum of weight values should be unity for normalization. The target PSNR for the i-th VO is derived using the distortion ratio as

$$PSNR_i = P_i \sum_{i=0}^{n} PSNR_i \qquad (8)$$

To prevent overflowing bit rate usage for a particular object, maximum target PSNR is limited by

$$\begin{aligned} &\text{if } (PSNR_i > \max PSNR) \\ &\quad PSNR_i = \max PSNR \end{aligned} \qquad (9)$$

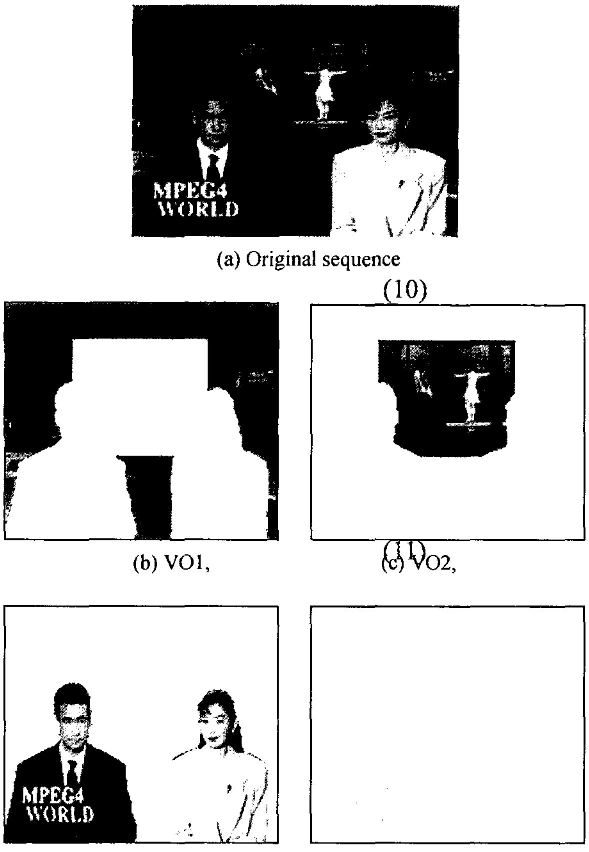
Thus, a VO or ROI with the highest importance or priority in a VOP is obtained by human eye interface. The obtained priority is applied to code the jointed video objects.

IV. SIMULATION AND RESULTS

Test sequences are chosen from the MPEG-4 test sequences which have a number of video objects, including Akiyo, News, and Coastguard, with the spatial resolution of QCIF(176 x 144), the frame rate 30 fps, and the length of frames 300. Akiyo is composed of 2 VOs of the anchor woman and the background, News (Fig. 6) has 4 VOs of background, inserted video with dancers, two anchors, and text, i.e., mixed all different properties. Coastguard shows nearly static sea water background, boat coming from left side, small boat moving left side from center of screen, and upper coast background.

Test sequences are viewed by a number of observers and their interests in each object are analyzed. The interests are measured by the number of changed view points and the importance of each video object is a normalized value by the total number of changes. As shown in Fig. 7, the monitor scene positioned behind the anchor is shown bear more visual importance given by the viewers than the foreground anchormen who were considered as the most important object in general.

Independent rate control using size, motion, and distortion results are shown in Fig. 8. Importance of a VO is defined by the three parameters independently. More bits are allocated to VO2 which has high motion. VO1, the background, shows higher quality in terms of the PSNR, though smaller number of bits is allocated, while the VO2 does not show the highest quality with the largest number of bits allocated.



(a) Original sequence

(10)

(b) VO1,

(c) VO2,

(11)

(d) VO3,

(e) VO4

Fig. 6 Simulation sequence "News" and four VOs.

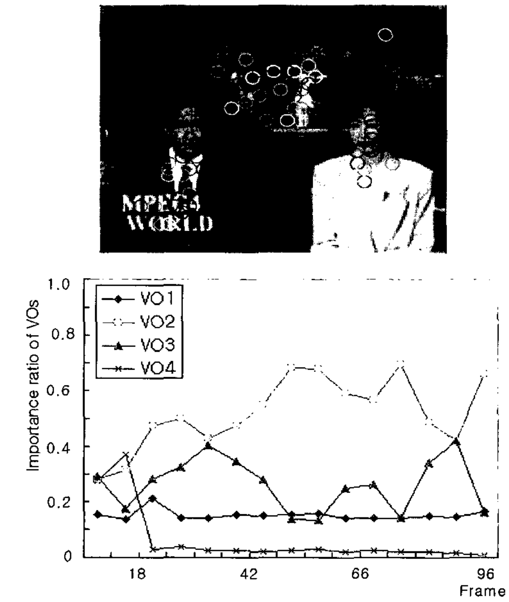


Fig. 7 ROI results, a) gaze position in a sequence and b) importance ratio of VOs.

Fig. 9 shows weighted distortion results. Weighting values are constant for VOs, for example, values of 0.5, 0.8, 1.0, 0.3 are given to the four objects in this example. Deviation of PSNRs decreases, indicating more constant quality in overall sense. However, it has a problem of constant weighting values regardless of varying importances of VOs.

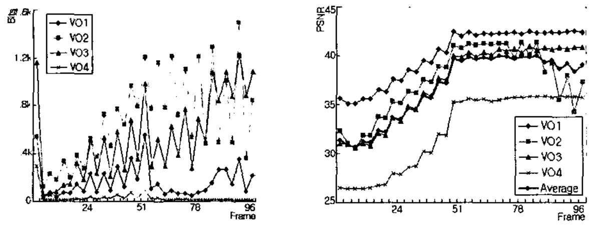


Fig. 8 Independent rate control results, a) bitrate and b) PSNR.

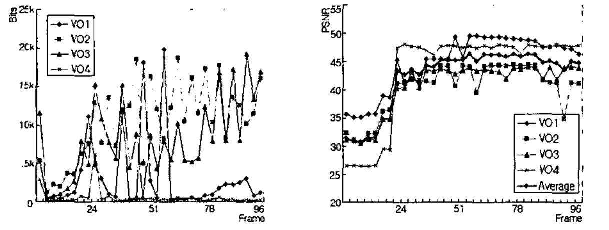


Fig. 9 Weighted distortion based control results, a) bitrate and b) PSNR.

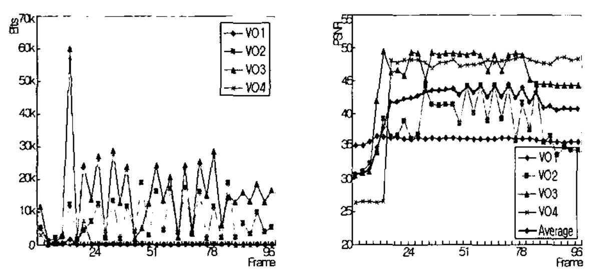


Fig. 10 Constant priority based control results, a) bitrate and b) PSNR.

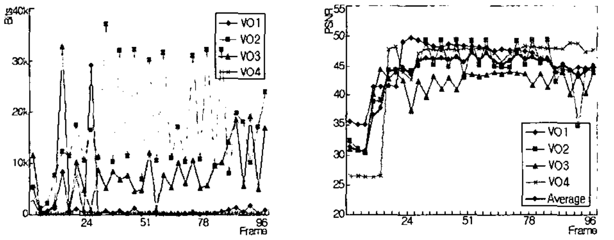


Fig. 11 Proposed control results with variable priority at each frame, a) bitrate and b) PSNR.

Fig. 10 shows the priority based control results. Priority and target distortion are given in a video sequence constantly, e.g., VO3 has the highest priority and target distortion of 35dB and followed by VO2, VO1, VO4, respectively. Once the system fulfilled these constraints, next object can be coded using bitrates remained in other VOs and adjusting to target quality. As results, VO4 shows again the highest quality with the smallest number of bits. The scheme is useful for emphasizing particular object quality but hard to get optimized quality for the whole sequence.

Fig. 11 shows the proposed variable priority control results. In view of human visual response, an object moving moderately has higher sensitivity. This means the VO2 (the dancing scene in the monitor) might be more important than the VO3(two speakers) and more bits are allocated, showing better results. Similar or less number of bits is allocated to VO1 and VO4 that are observed with less interest by the normal human eye interfacing. This is an advantage of this scheme that cannot be obtained by other existing techniques.

Another point for discussion is the deviation of quality among VOs. Regardless the number of bits allocated to each VO, the picture quality has to be nearly constant or with only a little deviation. The proposed scheme shows the smallest deviation of quality. Stability of the system is also important. The scheme reaches to the stable state in 15 frames and shows stable quality later on. Although the PSNR is not the only way of quality assessment, the system shows higher quality of more than 2dB than other bitrate control techniques.

IV. CONCLUSIONS

In this paper, we proposed an eye gaze tracking system which matched on the screen based on a real-time pupil detection using PC camera.

To detect the pupils, the vertical and horizontal profile and Gaussian filter were used. The pupil detection performance presented the pupil's detection-ratio of about 91.0% to detect the pupils in the face image accurately. Although every face detection system is much sensitive to illumination change, displaced head and distance between observer and camera, the pupil detection system shows sufficient performance not only in the stable situation, but also in the abnormal environment.

In initializing with two points clicked on the screen, the distance and angle between eyes and the screen are not necessary to get the interaction of computing of gazing on the screen. The eye gaze tracking result shows approximately a good performance, but as more interest regions were divided, the accurateness of the gaze tracking is worst.

We proposed a content-based bit-rate controller interfaced with human eye. By controlling bits allocated to each VO in a MVO encoder, the most relevant quality can be interactively obtained in this system. In this paper, we analyze the independent rate control algorithm and global algorithm where the QP value is controlled by the static parameters, the object importance or priority, the target PSNR, and the weighted distortion. The priority among static parameters is analyzed and adjusted into dynamic parameters according to the visual interests or importance obtained by a camera interface. The target PSNR and the weighted distortion are proportionally derived by using magnitude, motion, and distortion. We apply these parameters to the weighted distortion control and the priority-based control resulting in an efficient bit-rate distribution. As a result, we achieved an effective rate control that fewer bits are allocated for video objects which have less importance and more bits for those which have higher visual importance. The period to reach stability in the visual quality is reduced to less than 15 frames of the coded sequence. With respect to the PSNR, the proposed scheme shows higher quality of over 2dB than that of the conventional schemes. Thus the coding scheme interfaced to human-eye proves to be an efficient video coding technique for dealing with the multiple number of video objects.

REFERENCES

[1] S. Chien and I. Choi, "Face and Facial Landmarks Location Based on Log-Polar Mapping", *First IEEE International Workshop*, BMCV 2000, 2000.

[2] R. Chellappa, Wilson, C. L. Sirohey, S. Sirohey, "Human and machine recognition of faces: A Survey", *Proc. of IEEE*, vol. 83, 1995, pp. 705-740.

[3] K. Sung and T. Poggio, "Example-based learning for view-based human face detection", *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 20, 1998, pp. 39-51.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *IEEE CVPR*, 2001.

[5] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection", *IEEE ICIP* 2002, vol. 1, 2002, pp. 900-903.

[6] E. Hjelmas and B. K. Low, "Face detection: A survey", *Computer Vision and Image Understanding*, vol. 83, 2001, pp. 236-274.

[7] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection", *IEEE Trans PAMI*, vol. 20, 1998, pp. 23-38.

[8] R. Rrunelli and T. Poggio, "Face recognition: features versus templates", *IEEE Trans. PAMI*, vol. 20, 1993, pp. 1042-1052.

[9] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method", *IEEE. Trans. on Neural Network*, vol. 10, 1999, pp. 439-443.

[10] P. Sinha, "Object recognition via image invariant : A case study", *Investigative Oph-thalmology and Visual Science*, vol. 35, 1994, pp. 1735-1740.

[11] M. H. Yang, D. Kriegman and N. Ahuja, "Detecting faces in images: A survey", *IEEE Trans. PAMI*, vol. 24, 2002, pp. 34-58.

[12] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *Computational Learning Theory: Eurocolt'95*, 1995, pp. 23-37.

[13] R. Lienhart, A. Kuranov and V. Ppsarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection", *DAGM'03, 25th Pattern Recognition Symposium*, 2003.

[14] T. Weigand, et al., "Rate-distortion optimized mode selection for very low bit-rate video coding and emerging H.263 standard," *IEEE Trans. on CSVT*, vol. 6, Apr. 1996, pp. 182-190.

[15] J. Lee and B. W. Dickenson, "Rate-distortion optimized frame type selection for MPEG encoding," *IEEE Trans. on CSVT*, vol. 7, June 1997, pp. 501-510.

[16] T. Chiang and Y. Q. Zhang, "A new rate control scheme using quadratic rate-distortion modeling," *IEEE Trans. on CSVT*, vol. 7, Feb. 1997, pp. 246-250.

[17] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 rate control for multiple video objects," *IEEE Trans. on CSVT*, vol. 9, Feb. 1999, pp. 186-199.

[18] J. I. Ronda, M. Eckert, F. Jaureguizar, and N. Garcia, "Rate Control and Bit Allocation for MPEG-4", IEEE Trans. Circuits Syst. Video Technol., vol. 9, Dec. 1999, pp. 1243-1258.

Jae-Jeong Hwang
He received his B.S., M.S., and Ph.D. degrees from the Chonbuk National University, Korea, in 1983, 1986, and 1992, respectively. He is a professor in the School of Electronic & Information Engineering at the Kunsan National University, Korea. He spent the academic year 1990 as a visiting professor at the Hannover and Wuppertal University in Germany, the year 1993 at the University of Texas at Arlington and the year 2003 at Monash University in Australia. He is the coauthor of Techniques and Standards for Image, Video and Audio Coding (Prentice Hall, Inc., 1996) and Digital Video Engineering (Ajin Pub., 1999). His current research interests lie in low bit rate image and video coding, broadcasting of digital data, and interactive multimedia processing.

Sang-Gyu Cho
He received his B.S. and M.S. degrees from the Kunsan National University, Korea, in 2002 and 2004, respecttively. He is currently working for Ph.D. degree at the School of Electronic & Information Engineering, Kunsan National University, Korea. His current research interests lie in software implementtation of multimedia data compression, human-computer interaction for visual communication, and DSP realization of video codec.