Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 10(3): 300-306, Sep. 2012 **Regular Paper**

Visible Distortion Predictors Based on Visual Attention in Color Images

Sang-Gyu Cho[1], Jae-Jeong Hwang[1*], and Nae-Joung Kwak[2], *Member, KIICE*

[1]Department of Radio Communication Engineering, Kunsan National University, Kunsan 573-701, Korea
[2]Department of Information and Communication Engineering, Chungbuk National University, Cheongju 361-763, Korea

Abstract

An image attention model and its application to image quality assessment are discussed in this paper. The attention model is based on rarity quantification, which is related to self-information to attract the attention in an image. It is relatively simpler than the others but results in taking more consideration of global contrasts between a pixel and the whole image. The visual attention model is used to develop a local distortion predictor, named color visual differences predictor (CVDP), in color images in order to effectively detect luminance and color distortions.

Index Terms: Visual attention, CVDP, Visible distortion, Human visual system, Cortex transform, Inverse contrast

I. INTRODUCTION

An image is composed of a number of regions that contain background, foreground, textures, and meaningful objects. In other words, an image is a set of attention objects (AO) that capture the user's attention and interests. For example, human-related objects, such as a human face, a flower, a house, or a text sentence, usually attract high attention. Humans view an image and adapt in a short time interval to a handful of attention objects. Thus, image adaptation is treated as a way of manipulating AOs to provide as much information as possible in the image.

The attention model is generally classified into two categories: top-down approaches and bottom-up approaches. The first is task-driven, where prior knowledge of the target is known before the analysis or detection process. This is based on the cognition of the human brain [1]. For example, a bird is flying in the sky. The observer is already aware of the bird′s next action and anticipates the continuous flap of its wings. The second form of attention is the bottom-up approach, which is usually referred to as the stimuli-driven technique. This is based on human sensitivity to image features, such as the bright color, distinctive shape, or the orientation of objects.

Some investigation of computational attention methodologies has been performed. Itti and Koch [2] have worked on computational models of visual attention and presented a bottom-up, saliency- or image-based visual attention system. By combining multiple image features into a single topographical saliency map, the locations that draw attention are detected in the order of decreasing saliency by a dynamical neural network [3]. The main parameters that are extracted from the attention system are low-level features such as color, intensity, and orientation. Other possible parameters are stereo disparity in stereo images and shape information. Each feature is computed by a set of linear center-surround operations using visual receptive fields, since visual neurons are typically most

Received 29 June 2012, **Revised** 25 July 2012, **Accepted** 01 August 2012
***Corresponding Author** E-mail: hwang@kunsan.ac.kr

Open Access **http://dx.doi.org/10.6109/jicce.2012.10.3.300** print ISSN: 2234-8255 online ISSN:2234-8883

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ⓒ The Korea Institute of Information and Communication Engineering

sensitive in a small (center) region. Center-surround is modeled as the difference between a center fine scale and a surrounding coarser scale, yielding feature maps. The first set of feature maps is concerned with intensity contrast. The second set of maps is similarly constructed for color channels, which are represented by using the color double-opponent system. Neurons are excited by one color (e.g., red) and inhibited by another (e.g., green), while the converse is true in the surrounding area. The third set of maps is concerned with local orientation obtained by oriented Gabor pyramids [4]. A Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave, approximating the human receptive field sensitivity of orientation-selective neurons in the primary visual cortex. Li et al. [5] developed upon Itti's work by using a hierarchical architecture for saliency search on the basis of multi-scale saliency maps.

Human visual system (HVS)-based quality assessment is implemented in two modeling approaches: single-channel models and multi-channel models. Single channel models regard the human visual system as a single spatial filter, whose characteristics are defined by the contrast sensitivity function (CSF). The input is filtered by the CSF model. A single-channel model is unable to cope with more complex properties in the HVS, such as adaptation and inter-channel masking. These can be explained quite successfully by a multi-channel model, which assumes a whole set of different channels instead of just one. Daly [6] proposed the visual differences predictor (VDP), a rather well-known image distortion metric. The underlying vision model includes amplitude nonlinearity to account for the adaptation of the visual system to different light levels, an orientation-dependent two-dimensional CSF, and a hierarchy of detection mechanisms. These mechanisms involve a decomposition process similar to the above-mentioned cortex transform and a simple intra-channel masking function. The responses in the different channels are converted to detection probabilities by means of a psychometric function and finally combined according to the rules of probability summation. The resulting output of the VDP is a visibility map indicating the areas where two images differ in a perceptual sense.

The remainder of the paper is organized as follows. The luminance VDP proposed by Daly [6] is extended to include color components and to detect color distortions in Section II. The concept of attention modeling based on entropy and inverse contrast is developed and proposed in Section III. In Section IV, simulation results for the proposed visual attention and color VDP (CVDP) based image assessment model are presented. Finally, Section V draws major concluding remarks.

II. VISIBLE DISTORTION PREDICTOR

A. Luminance VDP

The VDP is a relative metric since it does not describe an absolute metric of image quality, but instead addresses the problem of describing the visibility of differences between two images. The VDP can be seen to consist of components for calibration of the input images, an HVS model, and a method for displaying the HVS predictions of the detectable differences. The input to the algorithm includes two images and parameters for viewing conditions and calibration, while the output is a third image describing the visible differences between them. Typically, one of the input images is a reference image, representing the image quality goal, while the other is a distorted image, representing the system's actual quality. The block components outside of the VDP generically describe the simulation of the distortion under study. The VDP is used to assess the image fidelity of the distorted image as compared to the reference. Its output image is a map of the probability of detecting the differences between the two images as a function of their location in the images. This metric, probability of detection, provides an accurate description of the threshold behavior of vision, but does not discriminate between different suprathreshold visual errors. The VDP can therefore be summarized as a threshold model for suprathreshold imagery, capable of quantifying the important interactions between threshold differences and suprathreshold image content and structure.

The HVS model addresses three main sensitivity variations. These are the variations as a function of light level, spatial frequency, and signal content. Sensitivity can be thought of as a gain, though the various nonlinearities of the visual system require caution in this analogy. The variations in sensitivity as a function of light level are primarily due to the light adaptive properties of the retina, and we shall refer to this overall effect as the amplitude nonlinearity of the HVS. The variations as a function of spatial frequency are due to the optics of the eye combined with the neural circuitry, and these combined effects are referred to as the CSF. Finally, the variations in sensitivity as a function of signal content are due to the post-receptoral neural circuitry, and these effects are referred to as masking. The HVS model consists of three main components that essentially model each of these sensitivity variations. In the current state of development of the VDP, these three components are sequentially cascaded in the straightforward manner shown in Fig. 1. The first component is the amplitude nonlinearity, implemented as a point process, while the second is the CSF, implemented as a filtering process. The final component in the cascade is the detection process, which models the masking effects. It is implemented as a combination of filters and nonlinearities.

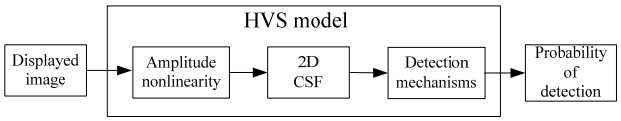


Fig. 1. The luminance visual differences predictor components. HVS: human visual system, 2D CSF: 2 dimensional contrast sensitivity function.

B. Color VDP

Color components are not as sensitive as luminance components. However, the image quality assessment has to describe the chromatic distortion that cannot be done by a luminance-based system. Thus, we designed an assessment system extended from Daly's work, as shown in Fig. 2.

For the spatial filtering of opponent channels, white-black (W-Bk), red-green (R-G), and blue-yellow (B-Ye), it is performed using a simple multiplication on a series of convolution. In order to filter the image, the opponent channels must be first transformed into their respective frequency representation using a discrete Fourier transform (DFT). Specifications of the contrast sensitivity functions are used for the general shape of the luminance and chrominance models, as defined by many researchers [6-8]. We use the spatial filters of the S-CIELAB model [9] that are designed to approximate the human contrast sensitivity function for the luminance component as given by:

$$CSF_{luma}(f) = a \cdot f^{c} e^{-b \cdot f} \tag{1}$$

where the parameters a, b, and c are 75, 0.2, and 0.8, respectively.

For the chrominance channels, R-G and B-Ye, we use the chrominance CSF filters in Eq. (2), and they are also depicted in Fig. 3. The six parameters are empirically defined.

$$CSF_{chroma}(f) = a_1 \cdot e^{-b_1 \cdot f^{c_1}} + a_2 \cdot e^{-b_2 \cdot f^{c_2}} \tag{2}$$

In digital imaging applications, spatial frequency in cycles per degree of the visual angle is a function of both addressability and viewing distance. For example, if a computer monitor is capable of displaying 72 pixels per inch (ppi) and is viewed at 18 inches, then there are roughly 23 digital samples per degree of visual angle. The calculation is shown in Eq. (3) [10]. However, Johnson's equation should be read in pels/deg instead of cycles/deg.

$$f_s = \frac{ppi}{2 \cdot \frac{180}{\pi} \cdot \tan^{-1}\left(\frac{1\ inch}{viewing\ distance}\right)}\ [\text{cyc/deg}] \tag{3}$$

Since the Poison opponent color space was designed for pattern-color separability, it is useful for implementation of separate contrast sensitivity for each color channel. When applying the CSF to the opponent transformed data, spatial frequencies are calculated by Eq. (3). The luminance data and two color difference data are processed in parallel in the domain of the cortex transform with a reasonable parameter setting.

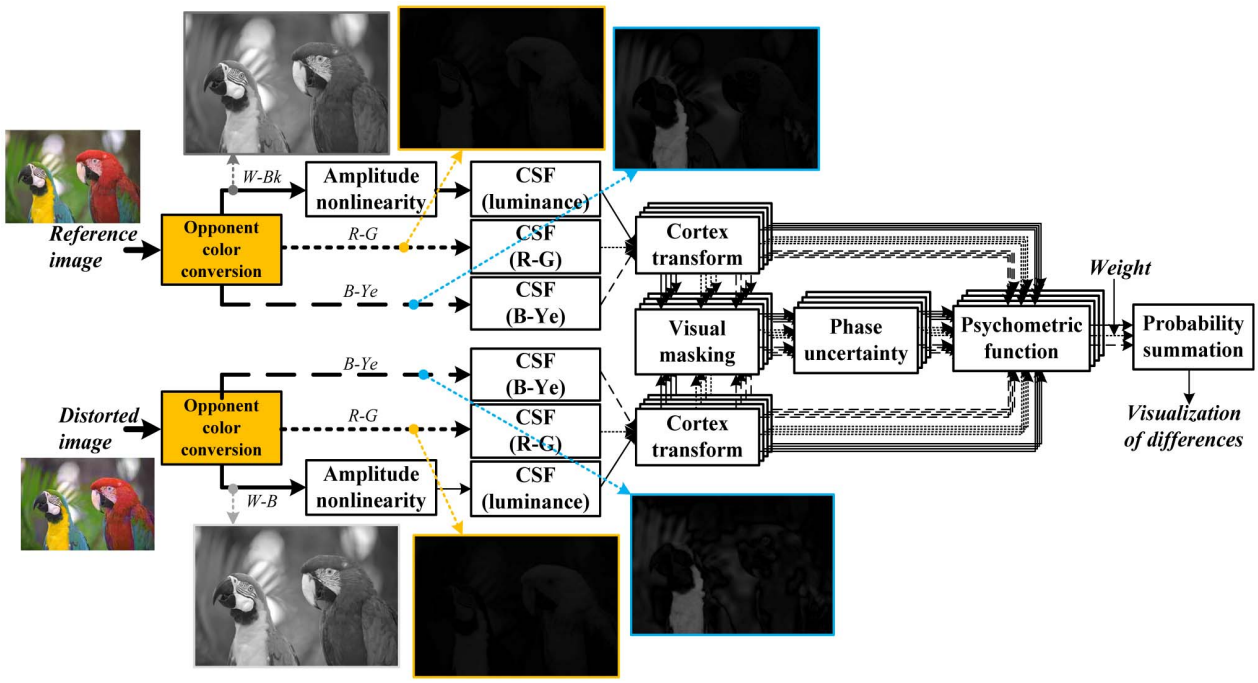


Fig. 2. Proposed color visual differences predictor block diagram. CSF: contrast sensitivity function, W: white, Bk: black, B: blue, R: red, G: green, Ye: yellow.

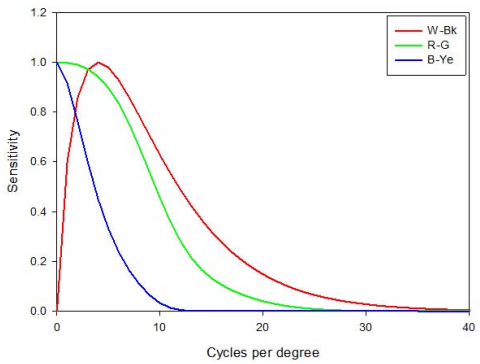


Fig. 3. Contrast sensitivity function for the opponent channels. W: white, Bk: black, B: blue, R: red, G: green, Ye: yellow.

III. VISUAL ATTENTION MODELING

A. Attention Modeling Based on Entropy and Inverse Contrast

It is believed that visual attention is not driven by a specific feature that has dedicated low-level properties that can be treated by the HVS-based assessment. Visual attraction can be induced by either heterogeneous or homogeneous, dark or bright, symmetric or asymmetric objects [11], which is determined by higher level processing than the existing HVS-based system. It can be assumed that image information that is rare in the image, such as high frequency areas or higher contrast areas, will draw more attention. Thus, visual attention may be ascertained by modeling and quantifying the rarity of image information, called entropy [12].

It is well-known that self-information is a function of probability of a symbol. Larger probability gives lower self-information by taking a logarithm as

$$I(m_i) = -\log(p(m_i)) \tag{4}$$

where $p(m_i)$ denotes the probability of a message, $m_i, 0 \le i \le G$. In image processing, the probability density function can be estimated by the histogram that shows the distribution of probabilities of all image levels.

A pixel is conspicuous if its gray level is significantly different from the neighboring pixel value. The larger the difference between the two levels, the higher the saliency that it represents. Thus, the saliency value of a level intensity I_k, $0 \le k \le G$ can be calculated from a contrast map that is constructed prior to the saliency map computation [1]. The maximum intensity G is chosen as 255 in this work. For an image of size $N \times M$, the global contrast value of I_k is defined as

$$C(I_k) = \frac{1}{N \times M} \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{|I_k - I(m,n)|}{I(m,n)} \tag{5}$$

where $I(m,n)$ denotes the intensity value at pixel location (m, n) in an image in the range [0, G]. While the global contrast dominates over the whole image, the inverse contrast [13] is defined as the reciprocal of $C(I_k)$,

$$C_I(I_k) = 1 / C(I_k). \tag{6}$$

The histogram of a digital image with $G+1$ total possible intensity levels is defined as the discrete function

$$H(I_k) = n_k \tag{7}$$

where n_k is the number of pixels in the image whose intensity level is I_k. The histogram affecting the attention probability is multiplied by the inverse contrast, resulting in the combined probability of the message as shown by

$$p(I(m,n)) = H(I(m,n) \times C_I(I(m,n)) \tag{8}$$

where $H(m,n)$ and $C_I(I(m,n))$ denote the histogram and the inverse contrast of the intensity level at the pixel location (m, n), respectively. Then, the visual attention is obtained by logarithmic operation as

$$A(m,n) = -\log(p(I(m,n)). \tag{9}$$

If a message is very different from all the others, $C_I(I(m,n))$ will be low so that the occurrence $p(I(m,n))$ will be lower and the message attention will be higher. Thus, instead of computing the saliency values of all the image pixels, only the saliency values of the intensity levels are necessary for the generation of the final saliency map. One example of the pixel-level spatial saliency computation is shown in Fig. 4.

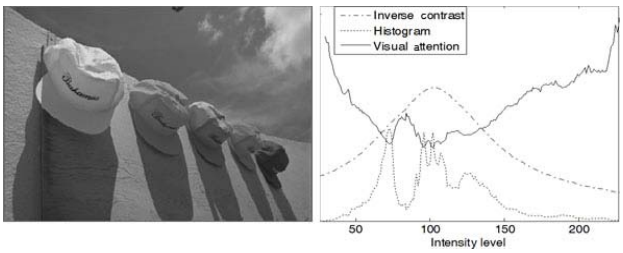


Fig. 4. Relative value of visual attention (right) for Y component of 'caps' image (left).

The left and right portions of Fig. 4 show, respectively, the luminance component of the input image and the resulting spatial saliency values, compared with image

histogram and global inverse contrast. Note that the scales for three plots are adjusted to represent them on a graph. The lowest saliency is found in the range of frequent occurrences and high global inverse contrast. The saliency values are close to what a human would expect because higher occurrence indicates redundant information in the image, which is, therefore, relatively unattractive (unattended).

The output saliency map shows some important objects obtained by a relatively simple algorithm. However, there may be too many salient objects in the complex images since the map is based on the histogram method. It does not distinguish the semantic meaning of the pixels, size or shape of objects, or texture information. In spite of these limitations, it is still useful for detecting the most salient pixels, which correspond to the pixels of visual attention.

B. Attention-based CVDP Modeling

The CVDP utilizes luminance and color responses in HVS and spatial frequency channel sensitivities. However, attention is focused on certain areas of an image that can be modeled by information theory rather than amplitude or frequency responses. Both attention and CVDP systems can be merged in serial or parallel connection. Fig. 5 shows a parallel combination in which visual attention results are weighted to a CVDP map to derive a final visible distortion map.

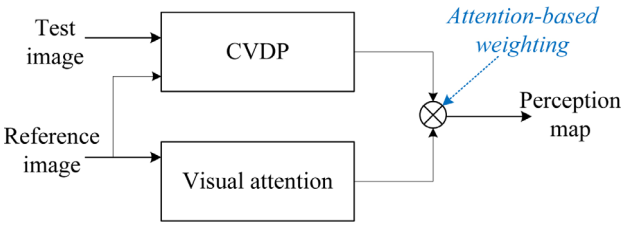


Fig. 5. Proposed attention-based color visual differences predictor (CVDP) quality metric overview.

IV. SIMULATION RESULTS

The purpose of visual attention is to determine which objects are the most interesting to the human eye at the moment. However, it is not always object-based. Some regions are most attractive depending on image properties. Itti's model works well for simple images like Fig. 6, which includes only two aircraft objects. Conspicuity maps for intensity, color, and orientation are separately drawn by considering relevant features. The final saliency map is obtained by combining the three conspicuity maps. However the test image contains only two significant objects, that is, it is so simple. If the complexity of an image increases, Itti's model does not provide meaningful results as shown in Fig. 7, which contains five cap objects. Although the human eye would be expected to latch on to one or more of the caps that are most attractive, the cap objects are hardly discernible in the final saliency map.

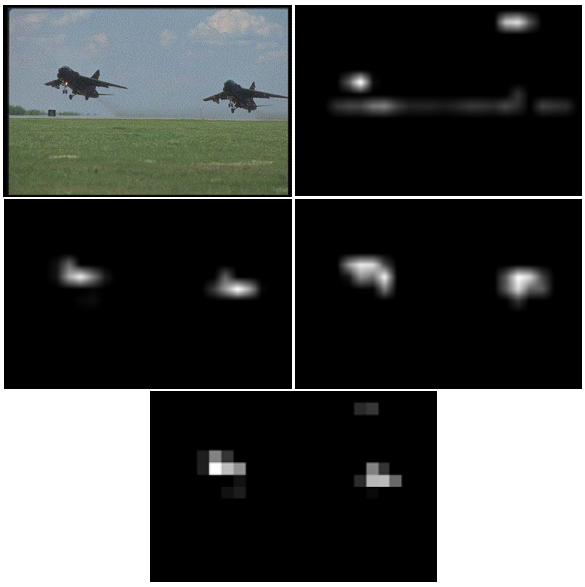


Fig. 6. Saliency map of Itti model: aircrafts image (top left), color conspicuity map (CM, top center), intensity CM (top right), orientation CM (bottom left), and saliency map (bottom right).

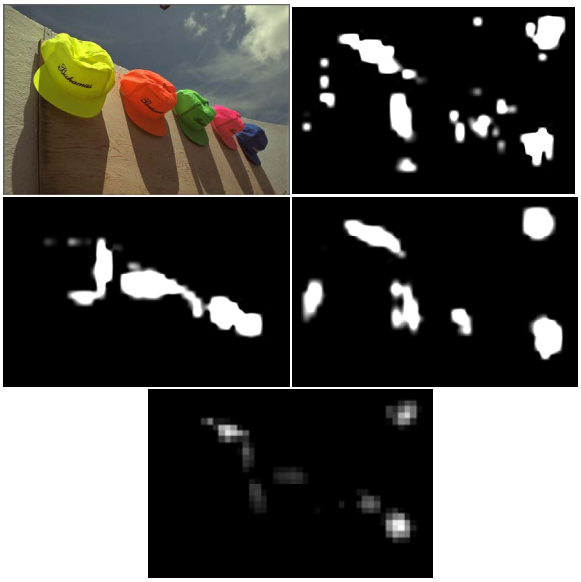


Fig. 7. Saliency map of Itti model: caps image (top left), color conspicuity map (CM, top center), intensity CM (top right), orientation CM (bottom left), and saliency map (bottom right).

Performance of the proposed attention-based CVDP (ACVDP) scheme is shown in Fig. 8, presenting an original image, CVDP perception map, and attention map obtained by rarity quantified modeling, and the final ACVDP result. It is worth noting that there are some clouds in the top-right part of the distorted "caps" image that draw quite high attention, but distortions existing inside are not detectable in terms of human sensitivity. On the other hand, the shadow areas below the caps do not draw much attention, but distortions are quite detectable. This means the two different approaches should be closely related to obtain optimum performance.

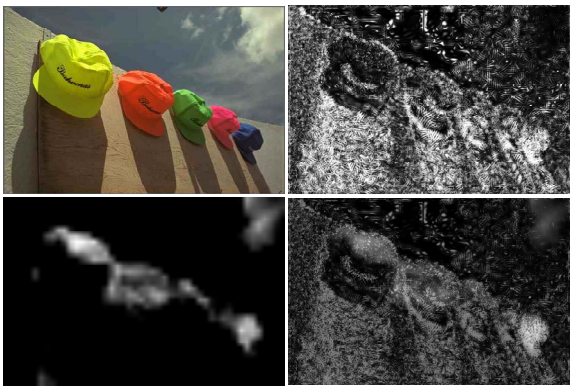


Fig. 8. Attention-based color visual differences predictor (CVDP) results: distorted image of caps (top left), CVDP perception map (top right), proposed attention map (bottom left), and perception map by combining the attention and CVDP maps (bottom right).

V. CONCLUSIONS

Image attention models and their application to image quality assessment were discussed. First, the luminance VDP developed by Daly was extended to cover color responses. Color distortions are detected by CVDP. Next, we reviewed the well-organized Itti's model for image attention modeling. Three parameters, intensity, color, and orientation, are used to generate a saliency map in a hierarchical structure and center-surround technique. Our attention model is based on rarity quantification, which is related to self-information or entropy in information theory. It is based on the fact that more important things are rare. The algorithm is relatively simpler than Itti's model but results in more consideration of global contrasts from a pixel to a whole image.

The two human eye-related schemes are combined to obtain a more efficient image quality assessment algorithm. The results show that objects that are important in view of one scheme are not really important from the perspective of another scheme. Therefore, the two schemes should be used to compensate for each other.

REFERENCES

[1] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," *Proceedings of the 14th Annual ACM International Conference on Multimedia*, Santa Barbara, CA, pp. 815-824, 2006.

[2] L. Itti and C. Koch, "Computational modelling of visual attention," *Natural Reviews Neuroscience*, vol. 2, no. 3, pp. 194-203, 2001.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[4] C. C. Taylor, Z. Pizlo, J. P. Allebach, and C. A. Bouman, "Image quality assessment with a Gabor pyramid model of the human visual system," *Proceedings of IS&T/SPIE International Symposium on Electronic Imaging Science and Technology (vol. 3016)*, San Jose, CA, pp. 58-69, 1997.

[5] Q. Li, S. Wang, and X. Zhang, "Hierarchical identification of visually salient image regions," *Proceedings of International Conference on Audio, Language and Imaging Processing*, Shanghai, China, pp. 1708-1712, 2008.

[6] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, Cambridge, MA: MIT Press, pp. 179-206, 1993.

[7] P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*, Bellingham, WA: SPIE Optical Engineering Press, 1999.

[8] J. A. Movshon and L. Kiorpes, "Analysis of the development of spatial contrast sensitivity in monkey and human infants," *Journal of the Optical Society of America A. Optics and Image Science*, vol. 5, no, 12, pp. 2166-2172, 1988.

[9] X. Zhang and B. A. Wandell, "A spatial extension of CIELAB for digital color-image reproduction," *Journal of the Society for Information Display*, vol. 5, no. 1, pp. 61-63, 1997.

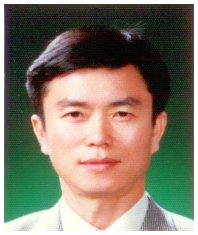
[10] G. M. Johnson and M. D. Fairchild, "A top down description of S-CIELAB and CIEDE2000," *Color Research and Application*, vol. 28, no. 6, pp. 425-435, 2003.

[11] J. W. Crabtree, P. D. Spear, M. A. McCall, K. R. Jones, and S. E. Kornguth, "Contributions of Y- and W-cell pathways to response properties of cat superior colliculus neurons: comparison of antibody- and deprivation-induced alterations," *Journal of Neurophysiology*, vol. 56, no. 4, pp. 1157-1173, 1986.

[12] A. K. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice Hall, 1989.

[13] J. J. Hwang and H. R. Wu, "Stereo image quality assessment using visual attention and distortion predictors," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 9, pp. 1613-1631, 2011.

Sang-Gyu Cho
received the B.S., M.S., and Ph.D. degreesfrom the Department of Electronic and Information Engineering of Kunsan National University in 2002, 2004, and 2010, respectively. He worked as a research engineer at DICS Vision from Sept. 2010 to March 2012. He currently teaches at Kunsan National University in Korea. His research interests are digital image/video coding and processing, computer vision, object segmentation and tracking, and 2D/3D visual quality assessment and evaluation.



Jae-Jeong Hwang
received the B.S., M.S., and Ph.D. degrees in electronic engineering from Chonbuk National University in 1983, 1986, and 1992, respectively. He is currently a full professor at Kunsan National University, Korea, and an adjunct professor at RMIT University, Australia. His research interests are digital image/video coding and processing, information theory, object segmentation and tracking, and 2D/3D visual quality assessment and evaluation. He is a coauthor of *Techniques and standards for image, video and audio coding* (Prentice Hall, 1996) and *Fast Fourier transform – Algorithms and applications*.

Nae-Joung Kwak
received the B.S. in February1993, M.S. in February 1995, and Ph.D in February 2005 from the Department of Computer and Communication Engineering, Chungbuk National University. She was a contract professor at Mokwon University from March 2005 to February 2009. She currently teaches at Hanbat University and Chun gbuk National University in Korea. Her research interests include multimedia communication, multimedia signal processing, video surveillance systems, and MPEG.