Stereo Image Quality Assessment Using Visual Attention and Distortion Predictors

Jae Jeong Hwang[1] and Hong Ren Wu[2]
[1] Dept. of Radiocommunication Eng., Kunsan National University
Kunsan, 573-701, Korea
[e-mail: hwang@kunsan.ac.kr]
[2] School of Electrical and Computer Eng., RMIT University
Melbourne, Victoria 3000, Australia
[e-mail: henry.wu@rmit.edu.au]
*Corresponding author: Jae Jeong Hwang

Received May 30, 2011; revised July 25, 2011; accepted August 18, 2011; published September 29, 2011

Abstract

Several metrics have been reported in the literature to assess stereo image quality, mostly based on visual attention or human visual sensitivity based distortion prediction with the help of disparity information, which do not consider the combined aspects of human visual processing. In this paper, visual attention and depth assisted stereo image quality assessment model (VAD-SIQAM) is devised that consists of three main components, i.e., stereo attention predictor (SAP), depth variation (DV), and stereo distortion predictor (SDP). Visual attention is modeled based on entropy and inverse contrast to detect regions or objects of interest/attention. Depth variation is fused into the attention probability to account for the amount of changed depth in distorted stereo images. Finally, the stereo distortion predictor is designed by integrating distortion probability, which is based on low-level human visual system (HVS), responses into actual attention probabilities. The results show that regions of attention are detected among the visually significant distortions in the stereo image pair. Drawbacks of human visual sensitivity based picture quality metrics are alleviated by integrating visual attention and depth information. We also show that positive correlation with ground-truth attention and depth maps are increased by up to 0.949 and 0.936 in terms of the Pearson and the Spearman correlation coefficients, respectively.

Keywords: Image quality assessment, stereo image processing, visual attention, 3D depth, distortion predictor

This research was supported by the Korea Research Foundation Grant (KRF-2009-013-D00084).

DOI: 10.3837/tiis.2011.09.007

1. Introduction

With the advanced development of data compression, visualization and display technologies, and the availability of ever increasing transmission channel bandwidth, multiview or 3-D (three-dimensional) imaging has been deployed to enhance the viewing experience or sense of realism that is comparable to the natural scene. This trend towards immersive media is going to have a wide range of applications such as 3-D TV, 3-D cinemas, 3-D gaming, virtual reality, etc. [1][2]. Though many efforts have been made to develop 2-D objective image quality measurements and metrics, relatively few reported in the literature have concentrated on 3-D objective quality assessment [3]. Picture quality assessment of coded video sequences can, currently, only be performed reliably using expensive and inconvenient subjective tests [4][5], notwithstanding a number of objective video quality metrics reported most recently [6][7][8][9]. Furthermore, the analysis of the obtained results is not straightforward. To enable 3-D imaging systems to provide a realistic 3-D information in a timely fashion, it is essential that reliable objective measures are found.

Up to now, only a few objective assessment methods for stereo images have been reported, which use depth map to assess stereo perception. There are two issues associated with depth map computation. First, computing depth maps for images is a highly computationally intensive and time consuming process. Second, it is hard to determine what degree of depth is good or bad. The sense of stereo vision is obtained from the difference between the view points of the two eyes. The vector between two corresponding points in the left and the right images of a stereopair is called disparity [10]. Finding the best matching points between two images is known as the correspondence problem. Depth value is the distance between a scene point and the camera baseline. It is inversely proportional to disparity value.

A viewer's 3-D perception is different for individual stereoscopic displays and human viewers can cope with a large variation of the perceived depth range. Thus, the depth mapping needs to be carried out dynamically to avoid excessive perceived depth [11]. Cepstrum filtering can be used in finding disparity in the frequency or the spatial domain [12]. Visual perception is the result of the integration of not only binocular disparity, but also motion parallax and image-realism cues [13]. Display duration is also a factor affecting visual perception [14].

Quality assessment metrics used for 2-D images can be applied to 3-D images with careful consideration and depth information [15]. Benoit et al. proposed a stereo quality metric by measuring, first, the difference between original images and the corresponding distorted version [16]. Corresponding left and right images are assessed by a 2-D metric such as the SSIM (Structural SIMilarity index) [17] or the C4 [18]. The two measures per pair are averaged in order to get the global 2-D image distortion measure. The second step is to assess the 3-D picture quality by measuring the difference between the disparity map of the original images and that of the distorted images. It is plausible to argue that perceptual-based distortion metrics cannot be applied here, since disparity maps are not natural images. Hewage introduced a "color plus depth map" based stereoscopic video assessment in order to design a scalable video codec that subjective and objective evaluations are optimally correlated [19]. Shen also proposed an enhanced SSIM metric by considering physiological and psychological factors of visual information [20].

Yang et al [21] presented a simple method that does not use depth map. Two types of objective assessment are taken into account, i.e., the image quality assessment (IQA) and the

stereo sense assessment (SSA). The former is performed by the arithmetic mean of the left and the right images assessed in term of the PSNR (peak signal to noise ratio), while the latter deals with stereo distortions. Disparity information is still measured by the absolute difference image between left and right image pairs. Disparity with low magnitude is removed, since it does not affect the outcome of quality assessment. Finally, the SSA is performed by only considering meaningful points in term of the PSNR. In [22], quality scores on both left and right images are evaluated by means of conventional metrics. Then, in order to obtain a single measurement for the quality assessment of the stereo image, the two scores are combined taking account of average, main eye and visual acuity.

Depth information can be used to calculate sharpness of edge distortion, since the distorted edge results in blurred artifact when viewing 3-D synthesis image. In [23], CSED (color and sharpness of edge distortion) metric is implemented. Color distortion measures the luminance loss of the test image pair and sharpness of edge distortion calculates the proportion of the remaining edge to the original edge along the boundary of the whole artifact region in the 3-D image rendering.

With respect to the block-based image coding, the stereo quality can be predicted to enhance the codec performance [24]. Blocking artifacts are measured, first, based on segmentation of the left and the right images separately into either edge, flat, and texture blocks. Second, zero crossing rate within each (8x8) pixel block is measured and a differencing zero crossing rate is determined between the same corresponding block of the left and the right images. The blockiness and the zero crossing disparity measurements are combined for final prediction of the mean opinion score (MOS) in a nonlinear equation. Stereo image quality prediction is useful to reconstruct the right view, if it is absent, since the left view shares a lot on common or similar information with its right view [25] in the depth image based rendering (DIBR) [26]. One view is encoded in high quality as the key view and the non-key views are encoded by inter-view prediction, which is called asymmetric view coding [27]. When viewing distorted stereo images, HVS may choose one of two views as a dominant view. Experiments show that images with high spatial frequencies are more dominant than the images with very low spatial frequencies [28], caused by the HVS spatial frequency sensitivity. That is, binocular perception of a stereo image pair is dominated by the high quality component. The level of dominance of the stereo image pair can be measured to design an objective stereo quality metric and data compression codec [29].

HVS sensitivity is also useful to assess stereo images. Gorley [30] developed an HVS-based metric that uses Peli's Local Band-Limited Contrast (BLC) algorithm. Local band-limited contrast of images is defined as a contrast value that is assigned to every pixel in the image as a function of the spatial frequency band. For each frequency band, the contrast is defined as the ratio of the bandpass-filtered image to the lowpass-filtered image to an octave below the same frequency [31]. Improved version for stereo images is called SBLC (Stereo Band-Limited Contrast), which is calculated from the mean of the ratio of stereo matched regions in both the left and the right views to the mean luminance of the whole image for every matched point. Goley's work is based on matching the regions of high spatial frequency between the left and the right views of the stereo pair and accounting for HVS sensitivity to contrast and luminance changes in regions of high spatial frequency. Matching algorithm uses SIFT (scale invariant features) [32] to extract local features (e.g., edges, corners) and RANSAC (RANdom SAmple Consensus) [33] algorithm to match the regions. An extensive survey of modeling the stereoscopic HVS is reported in [34].

In [35], a stereoscopic visual attention model is devised by integrating depth information with other low-level features, including motion, intensity, color and orientation contrast. Itti's

bottom-up attention model [36] is used to implement the spatial attention model. Local spatial discontinuities are detected by seven multi-scale low-level feature maps using simulated center-surround neurons. The seven neuronal features are sensitive to opponent color contrast, intensity contrast and four orientations (0^o, 45^o, 90^o, and 135^o). Finally, depth map, static saliency and motion saliency are integrated into a unique saliency map based on Treisman's feature integration theory [37].

A number of research investigations have been conducted in this arena to detect the most essential parts in images which are original or distorted in 2-D or 3-D spaces. The HVS is considered as the final decision making factor. However, most of the reported approaches utilize one or two aspects of three possible techniques (the HVS-based IQA (image quality assessment), attention modeling, and disparity in 3-D images), in spite of the fact that perceived stereo picture quality is most dependent up on their combined responses. In this paper, we consider different roles of the image assessment factors, as compared in **Table 1** in terms of the HVS-based IQA and attention-based modeling. Hence, we focus our attention on devising an integrated model for stereo color image quality assessment.

Table 1. Comparison of HVS-based IQA with attention-based modeling.

	HVS-based IQA	Attention Modeling
Processing images	Reference image & Distorted image	Reference image
Detection target	Unwanted components	Interested components
Processing direction	Bottom-up	Bottom-up & Top-down
Assessment level & tool	Low level contrast & masking	Low to high level contrast & masking

The remainder of the paper is organized as follows. Stereo attention predictor (SAP) predicts more visually attended regions based on rarity information described in Section 2. The concept of depth variation (DV) due to distortions in stereo images is also introduced, since it has more significance than depth itself. Stereo distortion predictor (SDP) is designed in Section 3 by integrating all three parameters. The resultant visibility map for stereo images is presented in Section 4 using test images. Section 5 draws major concluding remarks.

2. Stereo attention predictor

The human visual system refers to biological vision system that consists of several pathways from low level image input through the eye to high level information analysis and understanding in the brain. Visual attention is considered as one the most important tasks of the HVS, which is to extract interesting features from the surrounding images. The features are driven either by specific objects or by regional textures by means of parallel or serial processing. A model is required to approximate and to mimic some of these HVS processes.

According to studies on afferent superior colliculus (SC) pathways in the brain, the direct path from the retina to V1 cortex is responsible for spatial and temporal processing and the indirect pathway via the retina – SC – V1 cell is mainly responsible for spatial and motion direction (orientation) and color processing [38]. Since the spatial processing mainly deals with luminance and frequency contrasts, it is possible to model them separately from orientation and color processing.

2.1 Attention modeling based on entropy and inverse contrast

It is believed that visual attention is not driven by a specific feature which 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 [39], that is determined by higher level processing than the existing HVS-based system. Assume that image information which is rare in the image will be more attractive. Thus, visual attention may be ascertained by modeling and quantifying the rarity of image information, called entropy [40].

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

$$I(m_i) = -\log(p(m_i)) \tag{1}$$

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 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 two levels, the higher the saliency which 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 [41]. 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

$$C(I_k) = \frac{1}{N \times M} \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{\left| I_k - I(m,n) \right|}{I(m,n)} \tag{2}$$

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 is defined as the reciprocal of $C(I_k)$,

$$C_I(I_k) = 1 / C(I_k). \tag{3}$$

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

$$H(I_k) = n_k \tag{4}$$

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 by

$$p(I(m,n)) = H(I(m,n)) \times C_I(I(m,n)) \tag{5}$$

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

$$A(m,n) = -\log\big(p(I(m,n))\big). \tag{6}$$

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 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. 1**.

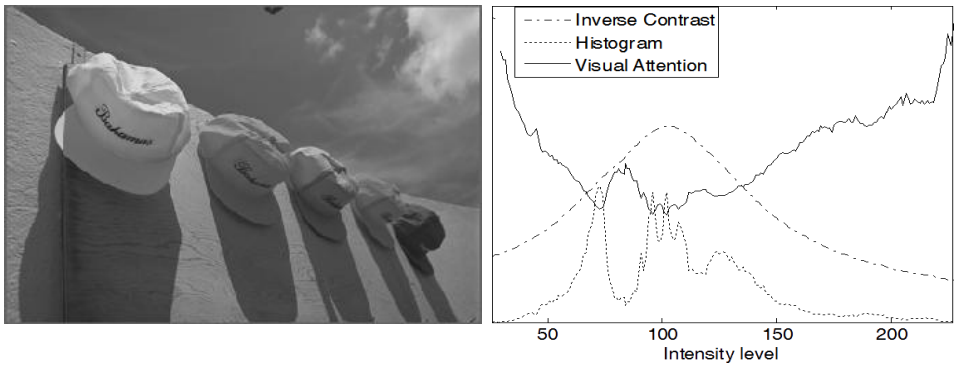


Fig. 1. Relative value of visual attention for Y component of 'caps.bmp' image (a), compared with the image histogram and the global inverse contrast.

Fig. 1 (a) and **(b)** 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 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 human expects, since higher occurrence indicates redundant information in the image, and therefore, relatively unattractive (unattended).

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

3. Stereo distortion predictor

A depth assisted attention model is proposed in this section for quality assessment of stereoscopic images. **Fig. 2** presents the architecture of the proposed model including a depth assisted attention model which will be used as a weighting factor for the quality assessment tool to derive the final SDP (Stereo Distortion Predictor) output in the form of a map or a

single number. More detailed operations and notations in the visual attention and depth assisted stereo image quality assessment model (VAD-SIQAM) are depicted in **Figs. 3** and **4**.

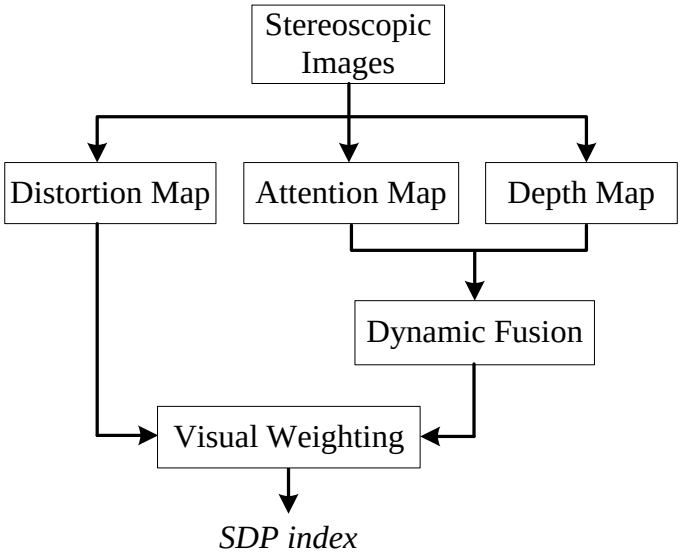


Fig. 2. Flow diagram of the proposed visual attention and depth assisted stereo image quality assessment model (VAD-SIQAM).

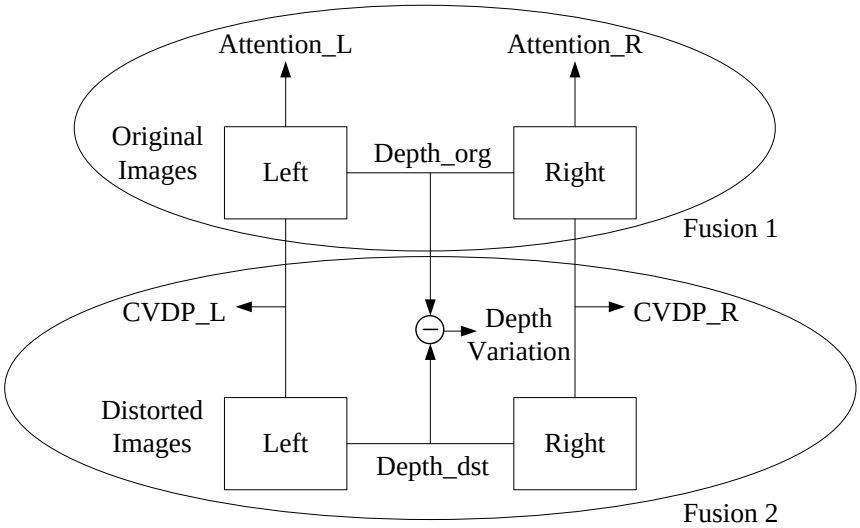


Fig. 3. Detailed operations and notations in the visual attention and depth assisted stereo image assessment model.

For *distortion map*, the color VDP (CVDP) metric is used which extends Daly's model [42] for the stereoscopic inputs. Daly's system consists of four main components including luminance nonlinearity, CSF processing, directional processing and masking by cortex filtering to deal with luminance component only. All algorithms are organized to match human sensitivity to detect distinguished distortions in an image. However, it only deals with luminance image. Thus, an extended system to color image assessment is proposed in this

work. First, RGB components are converted to the opponent color space [43]. Each component is processed by relevant contrast functions for luminance, color, and orientation. Orientational masking is applied in the omnidirectional cortex domain. The system yields detection capability to chromatic distortion, resulting in the visible distortion probability, $p_{vdp}(m,n)$, at each pixel location. This is corresponding to low-level processing in terms of distortion sensitivity in the visual system which has multi-level hierarchy to derive a final decision, while the attention modeling is with regard to low-level processing in terms of interest sensitivity.

For *attention map*, an attention model is motivated by Itti model [36][44]. The purpose of visual attention is to obtain the most interesting objects of the moment as viewed by human eye. It is not always object-based as it has been commonly done in the IQA applications. Some regions are most attractive depending on image properties. Itti's model is designed to detect most salient spot or object. Conspicuity maps for intensity, color, and orientation are separately generated by considering entropy and inverse contrast features. Final saliency map is obtained by combining the three conspicuity maps. However, as the complexity of an image increases, it becomes less meaningful, since the saliency map is used to highlight mainly some hot spots which stand out attracting visual attention.

It is necessary to equalize the attention values for the purpose of indicating the most attentive region throughout the image as the maximum level 1 and the least attentive region as the minimum level 0, meaning the probability of attention at the pixel location, defined by

$$p_a(m,n) = \frac{A(m,n) - \min(A)}{\max(A) - \min(A)} \tag{7}$$

where A denotes the set of attention in an image.

For *depth map*, it can be generated from the disparities between corresponding points in two images. The Zitnick-Kanade algorithm [45] is used and its results are converted to probability, $p_d(m,n)$, in the same vein as Eq. (7). Depth maps are derived from the original data set and the distorted images. However, the ground-truth depth map is used for final evaluation purpose.

Dynamic fusion stage is used to integrate the obtained pixel-wise image saliency and depth weighting factor, w_{mn}, as defined by

$$w_{mn} = \left(k_a \cdot p_a(m,n) + k_d \cdot p_d(m,n)\right) \tag{8}$$

where k_a and k_d are constant values, empirically chosen as 1.2 and 0.8, respectively, by taking into account visual importance. $p_a(m,n)$ denotes the attention probability (AP) and $p_d(m,n)$ is defined as the depth variation probability (DVP) between original and distorted images (**Fig. 4**). Assuming that the larger the original depth and the differential depth, the more influence there is on visual perception, the DVP is defined as

$$\begin{aligned} p_d(m,n) &= p_{d,org}(m,n) \cdot \left(1 + p_{diff}(m,n)\right) \\ p_{diff}(m,n) &= \left| p_{d,org}(m,n) - p_{d,dst}(m,n) \right|, \end{aligned} \tag{9}$$

where $p_{d,org}(m,n)$ and $p_{d,dst}(m,n)$ denote depth probabilities in the original and the distorted image pairs, respectively.

Finally, the SDP index, Q_{mn}, is obtained by *visual weighting* to the visible distortion probability in Eq. (8) as

$$Q_{mn} = w_{mn} \cdot p_{vdp}(m,n) \tag{10}$$

where $p_{vdp}(m,n)$ denotes the visible distortion probability at pixel (m,n), calculated by the CVDP distortion map. Note that all these operations are performed for both left and right images separately and the final SDP map is given to both left and right image pair.

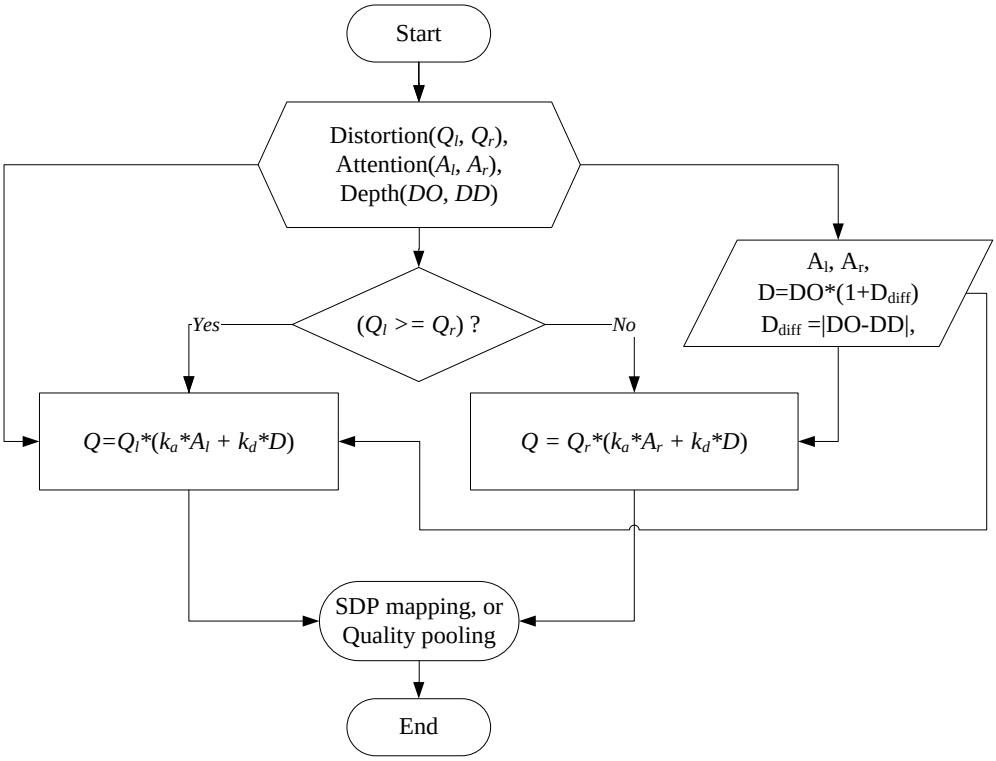


Fig. 4. Float chart of proposed stereoscopic attention based model. (DO: Original Disparity, DD: Distorted Disparity, Q_l: Quality for left image, Q_r: Quality for right image, A_l: Attention for left image, A_r: Attention for right image, k_a and k_d: weighting factors for attention and depth, respectively.)

Despite the pixel-wise SDP mapping, single number quality metrics are often useful for quantitative evaluation and comparison. The Minkowski sum [46] of the final SDP indices in the *quality pooling* stage (**Fig. 5**) is calculated as defined by

$$Q = \left(\sum_{m=0}^{M} \sum_{n=0}^{N} Q_{mn}{}^{\beta} \right)^{1/\beta} \tag{11}$$

where the parameter $\beta = 2.4$. The SDP mapping provides an indication of the location of visible distortions, while the single number provides a simple way to represent image quality.

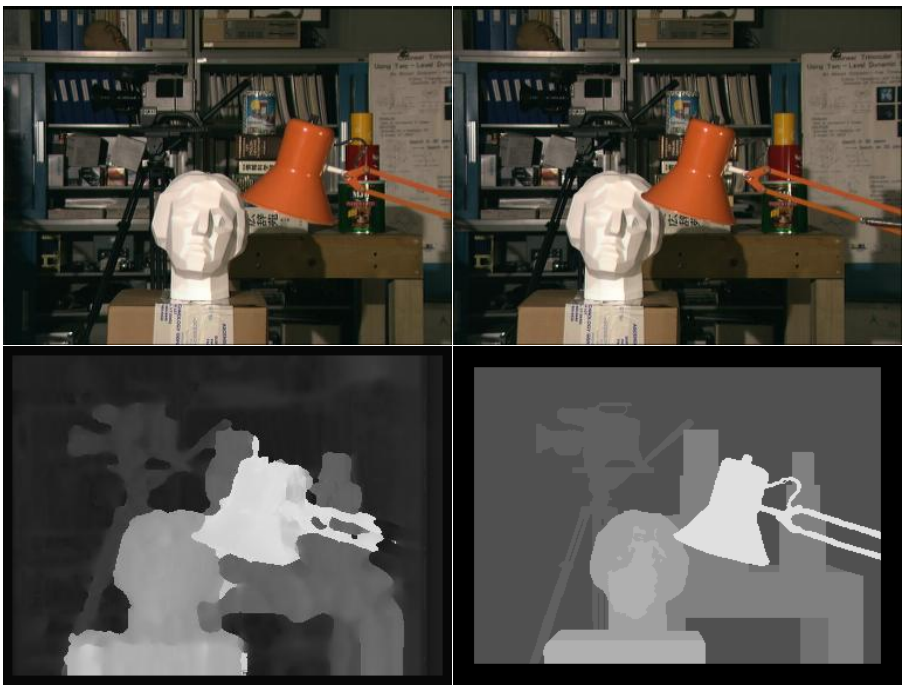


Fig.5. Stereo image and disparity map: (*Top-left*) Left-view image "tsukuba_r.bmp", (courtesy of U. of Tsukuba), (*Top-right*) Right-view image, (*Bottom-left*) Disparity map obtained by Zitnick-Kanade algorithm, and (*Bottom-right*) Ground-truth depth map used to evaluate the results.

4. Simulation and Results

A stereo image pair, "tsukuba.bmp", that consists of foreground objects (head sculpture and stand light) has been tested. First, disparity values are measured at each pixel as shown in **Fig. 5**. The nearest object in **Fig. 5** (top-left and -right) is the stand light whose depth is represented by white pixels in the disparity map, while the background (book-shelf) is drawn by dark pixels, meaning that they are far away from the foreground.

Stereo images are compressed by JPEG coding as shown in **Fig. 6**, resulting in blocking and ringing artifacts as well as various other well-known picture coding distortions. Quantization level zero (Q0) of twelve levels provides the worst quality in the Photoshop CS4 toolkit [47]. Both left and right images are encoded with Q0 and Q2, respectively, resulting in slightly different PSNR performance due to small amount of disparity and independent encoding for two channels. The disparity maps in **Fig. 7** are also affected by compression, although some results suggested that the JPEG encoding had no effect on perceived depth [28].

Color images captured by a camera or other input devices are normally represented in the RGB color space. It is first necessary to convert RGB images to other color formats, e.g., the opponent color space [48], since the opponent color theory suggests that there are three visual pathways in the human color vision system. One pathway is sensitive mainly to light-dark (W-Bk) variations. The other two pathways are sensitive to red-green (R-G) and blue-yellow (B-Ye) variation. We decided to use the YCbCr color model in this work, which is a color

space in digital TV broadcasting [49]. The Y contains the luminance information, while Cb and Cr contain chromatic difference information.

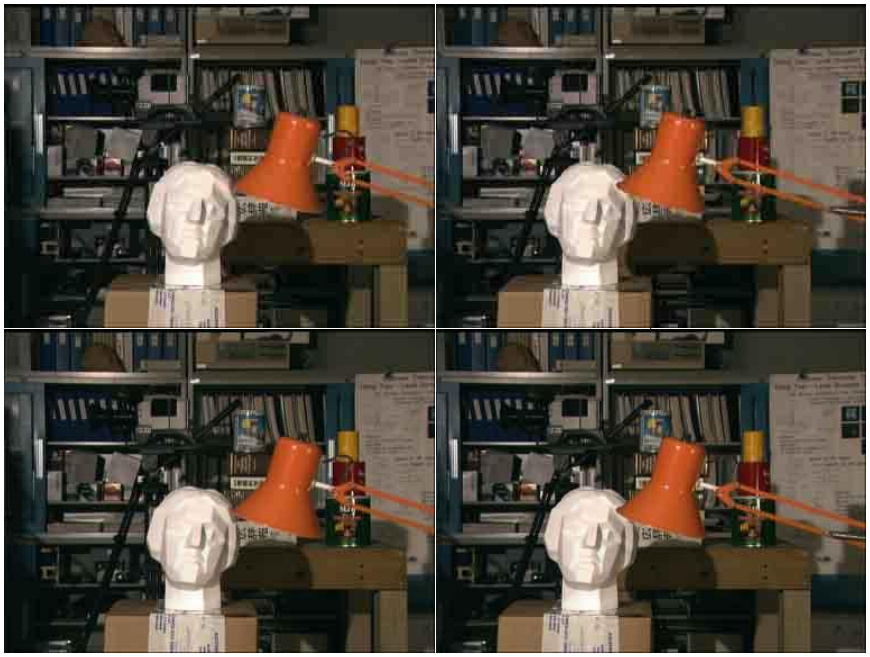


Fig. 6. Stereo images distorted by JPEG codec: (*Top-left*) Left-view image (PSNR 29.91 dB, Q_l2), (*Top-right*) Right-view image (PSNR 30.01 dB, Q_r2), (*Bottom-left*) Left-view image (PSNR 31.92 dB, Q_l0), and (*Bottom-right*) Right-view image (PSNR 32.08 dB, Q_r0). Blocking artifacts are detectable.

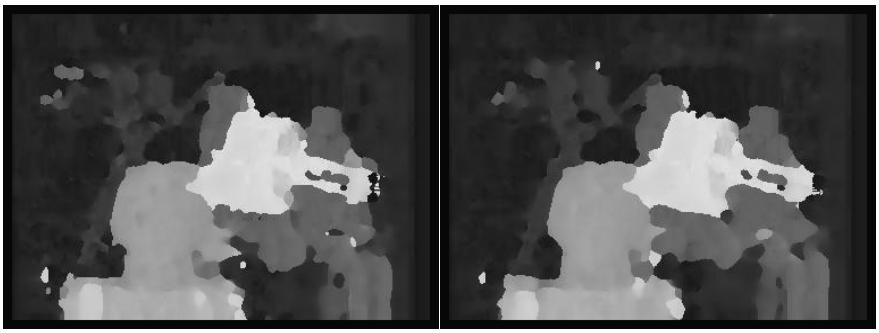


Fig. 7. Disparity maps of stereo images distorted by JPEG codec: (*Left*) Q0, (*Right*) Q2.

Fig. 8 shows attention maps for color difference channels, obtained by using algorithms described in Section 2. The entropy and inverse contrast-based attention model is applied to three color difference signals. The results show higher attention values for the sculpture region in luminance component and for the stand light region in color difference components. The attention maps for color difference components are combined to obtain an attention map for color component, which will be used as an input to overall attention model. The orientation map is the third component to be combined to obtain an overall attention map for left and right channels, as shown in **Fig. 9**.

The CVDP outputs (**Fig. 10**) show all distortions in terms of visual sensitivity, since the visual attention has not been taken into account. More distorted images (top-row, Q0) result in

higher number of white pixels where distortions are visible. However, objects with larger depths (e.g., sculpture region) are evaluated with less visible distortions, which are not normal in terms of visual attention. Such results are caused by the fact that larger object (lower spatial frequency) is less perceptible by spatial frequency sensitivity of human eye. This is a reason that we have to integrate visual attention model.

The proposed model gives more emphasis on visually more attended regions or objects (**Fig. 11**) by considering attention and disparity information. As an evaluation of performance of the VAD-SIQAM, scatter plots are drawn, showing the perceived distortion level versus the ground-truth attention in **Fig. 12**(a)-(b) and versus depth values in **Fig. 12**(c)-(d). The CVDP detects the visible distortions in the left image (**Fig. 12**(a)) and the right image (**Fig. 12**(b)), respectively. The VAD-SIQAM predicts the significant distortions in the left and right images by combining attention and depth information, resulting in more correlated performance in terms of visual attention, which is derived based on assumption that the sculpture and the stand light are two most attentive objects, called *ground-truth attention* as shown in **Fig. 9**. In the same reason, there is no absolute criteria to detect depth information unless the actual distance is measured, called *ground-truth depth* as shown in **Fig. 5**. Objects with higher depth are used to have higher attention. However they are not always in this case, showing that the sculpture object is the most attentive.

The linear correlation coefficients are widely used to compare a pair of data set and to evaluate the performance of quality model. In general, Pearson's correlation coefficient (CC) and Spearman rank-order correlation coefficient (ROCC) are used to measure the prediction performance [50]. First, the CC r_p is defined as:

$$r_p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{12}$$

where r_p is the Pearson's correlation coefficient for pairs of (x_i, y_i), $i = 0,...,N-1$. $\bar{x}$ and $\bar{y}$ denote means of data set x_i's and y_i's, respectively. The value of r_p are in the range of -1 and 1. It takes a value of 1, meaning "complete positive correlation", while -1 meaning "complete negative correlation". If it is near zero, the variables **x** and **y** are almost uncorrelated. The ROCC r_s is similarly defined as:

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}} \tag{13}$$

where R_i is the rank of x_i among the other *x*'s and S_i is the rank of y_i among the other *y*'s.

Fig. 8. Attention maps for left and right image by entropy and inverse contrast model: (*Top-left*) Luminance (Opponent color W-Bk) component of left reference image, (*Top-middle*) *Cb* (B-Ye) chromatic component of left reference image, (*Top-right*) *Cr* (R-G) chromatic component of left reference image, (*Bottom-left*) Luminance component of right reference image, (*Bottom-middle*) *Cb* chromatic component of right reference image, (*Bottom-right*) *Cr* chromatic component of right reference image.

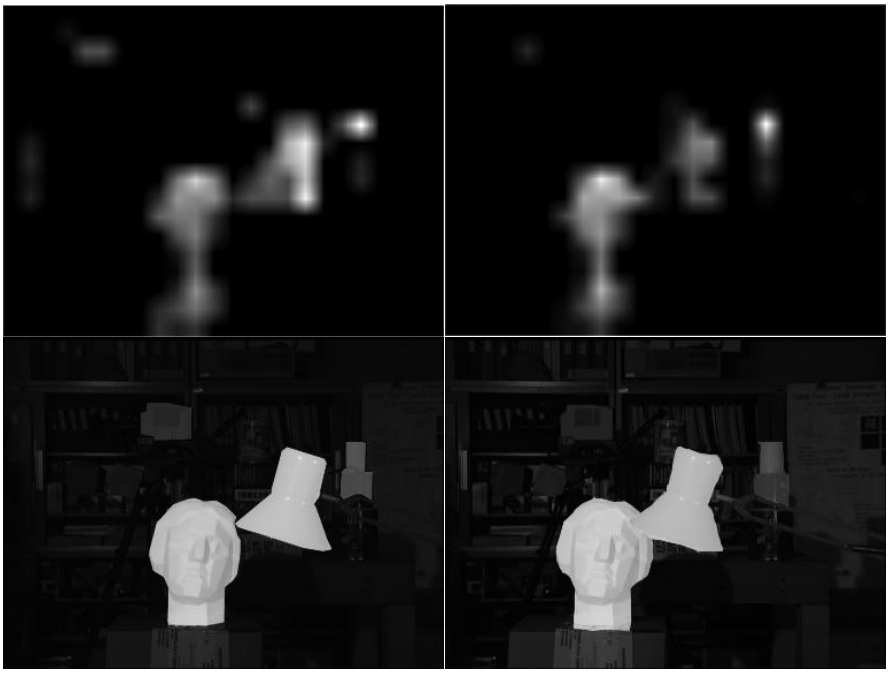


Fig. 9. Attention maps for original "tsukuba.bmp": (*Top-row*) Derived left and right attention that intensity, color, and orientation features are combined by Itti's winner-take-all algorithm, (*Bottom-row*) Left and right ground-truth attention used to evaluate the results.

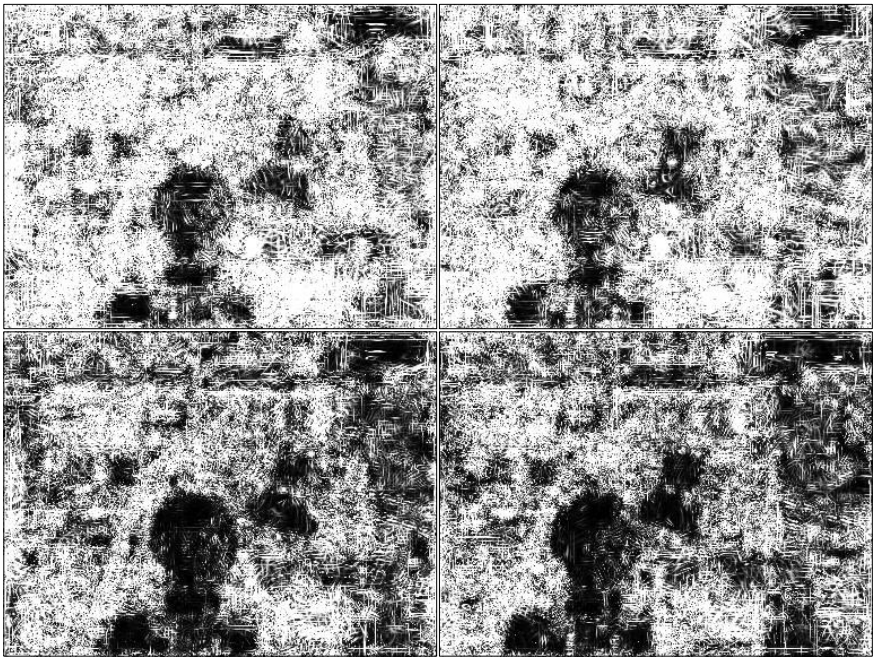


Fig. 10. Color VDP result images: (*Top-left*) Left image (Q0), (*Top-right*) Right image (Q0), (*Bottom-left*) Left image (Q2), (*Bottom-right*) Right image (Q2). White level represents the most visible distortion in the pixel basis, while black level represents no discrimination of distortion.

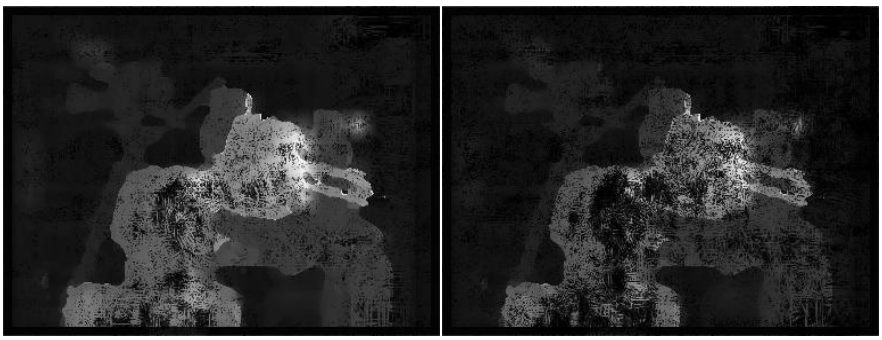


Fig. 11. The final perceptual distortion map of the proposed VAD-SIQAM for stereo data set at (*Left*) Q0 and (*Right*) Q2 quantization level.

We consider the CC and ROCC as very important factors in comparing the ground-truth attention and depth with the proposed VAD-SIQAM and the CVDP model. Notice that the prediction monotonicity of the performance is given by the ROCC, while the prediction accuracy is measured by the CC. The negative correlation of the CVDP turns out to be positive correlation by means of attention and depth information in our model. Correlation coefficients are up to 0.668 on average versus attention and up to 0.949 versus depth as shown in **Table 2**. Although the final results are obtained by combining the visible distortions in the left and right images, we show the single-channel correlation properties for quantization level zero (Q0) and level two (Q2) images.

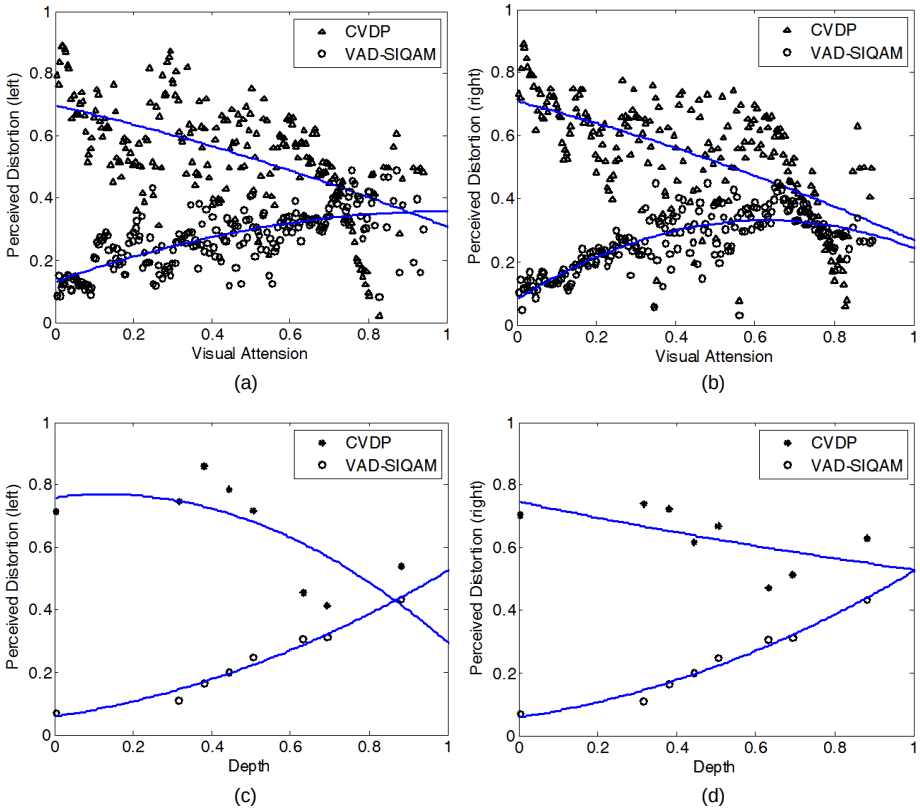


Fig. 12. Scatter plots of perceived distortion versus ground-truth visual attention and depth. (a) and (c) for left image; (b) and (d) for right image. The VAD-SIQAM shows positive correlation, while the CVDP shows negative correlation.

Table 2. Comparison of correlation coefficients (CC) versus attention and depth for compressed stereo images.

			Pearson CC		Spearman CC	
			CVDP	VAD-SIQAM	CVDP	VAD-SIQAM
vs. Attention Map	Q0	L	-0.586	0.686	-0.587	0.706
		R	-0.616	0.683	-0.632	0.690
	Q2	L	-0.435	0.659	-0.448	0.673
		R	-0.628	0.532	-0.617	0.603
	Average		-0.566	0.640	-0.571	0.668
vs. Depth Map	Q0	L	-0.642	0.967	-0.598	0.970
		R	-0.608	0.967	-0.667	0.970
	Q2	L	-0.641	0.931	-0.653	0.901
		R	-0.539	0.931	-0.695	0.901
	Average		-0.608	0.949	-0.653	0.936

5. Conclusion

A visual attention and depth assisted stereo image quality assessment model, i.e., the VAD-SIQAM, is introduced based on color VDP quality map, visual attention map, and differential disparity map. First, disparity for each pixel in stereo image is measured using Zitnick-Kanade algorithm to decide visual attention map in 3-D images. This approach is useful because regions or objects of interest are dependent on their depth, i.e., attention to an object usually decreases as its depth increases. When the disparity map is used to assess distorted stereo images, the disparity information is also subject to distortion and the depth variation is more important than the original depth information. Thus, a stereo attention predictor combined with the amount of depth variation is devised based on the principle that more distorted depth information results in higher attention values.

Second, visual attention is measured using the low-level image features such as intensity, color, and orientation. Saliency value is closely related to the rarity information, since it affects the final point(s) of attention. An entropy and inverse contrast based saliency measurement model is formulated for intensity and color components which are further combined with orientation component.

Third, an attention probability map is generated which is comparable to the distortion probability of the CVDP. The CVDP provides a quality map, indicating which part of an image is more sensitive to human eye. Motivation of using stereo attention to the CVDP is that distortions in more attended regions are easily detected. The attention probability is weighted by the HVS-based distortion probability to derive a quality index of stereo images. The results show that the VAD-SIQAM is able to detect visually significant distortions which correspond to visually more attended regions from the rest of distortions based on human visual properties. The performance of the VAD-SIQAM is evaluated in terms of the Pearson linear correlation coefficient and the Spearman rank order correlation coefficient, demonstrating significant improvement in visual attention and depth measurements compared with the CVDP.

References

[1] A. Kubota, et al. “Multiview Imaging and 3DTV,” *IEEE Signal Processing* Mag., vol. 24, no. 6, pp. 10-21, Nov. 2007. Article (CrossRef Link)

[2] L.M.J. Meesters, W.A. IJsselsteijn, P.J.H. Seuntiens, “A Survey of Perceptual Evaluations and Requirements of Three-dimensional TV,” *IEEE Trans. on Circuits and Systems for Video Technol.*, vol. 14, no. 3, pp. 381-391, Mar. 2004. Article (CrossRef Link)

[3] ICIP2010 Special Session, WP.L1, 3D Video Quality Assessments, Hong Kong, Sept. 26-29, 2010.

[4] ITU-R, Recommendation BT.500-11, “Methodology for the Subjective Assessment of the Quality of Television Pictures,” 2002.

[5] ITU-T, Recommendation P.910, “Subjective Video Quality Assessment Methods for Multimedia Applications,” Apr. 2008.

[6] ITU-T, Recommendation J.144, “Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference,” Geneva, Mar. 2004.

[7] J. Caviedes, F. Oberti, “No-reference Quality Metric for Degraded and Enhanced Video,” in *Proc. SPIE*, *vol.5150*, pp. 621-632, July 2003. Article (CrossRef Link)

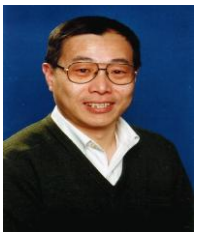
[8] M.H. Pinson, S. Wolf, “A New Standardized Method for Objectively Measuring Video Quality,” *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312-322, Sep. 2004. Article (CrossRef Link)

[9] F. Yang, S. Wan, Q. Xie, H.R. Wu, “No-reference Quality Assessment for Networked Video via Primary Analysis of Bit Stream,” *IEEE Trans. on Circuits and Sys. for Video Tech.*., vol. 20, no. 11, pp. 1544-1554, Nov. 2010. Article (CrossRef Link)

[10] S. Narkhede, F. Golshani, "Stereoscopic imaging: a real-time, in depth look," *IEEE Potentials*, vol. 23, no. 1, pp. 38-42, Feb.-Mar. 2004. Article (CrossRef Link)
[11] G. Sun, N.S. Holliman, "Evaluating Methods for Controlling Depth Perception in Stereoscopic Cinematography," *Stereoscopic Displays and Virtual Reality Systems*, SPIE, vol. 7237, Jan. 2009. Article (CrossRef Link)
[12] A. Awawdeh, G. Fan, "Pseudocepstrum for Assessing Stereo Quality of Retinal Images," in *Proc. of Asilomar Conf. on* Signals, Systems and Computers, vol. 2, pp. 1953-1957, Nov. 2003.
[13] M. Ferre, R. Aracil, M. Sanchez-Uran, "Stereoscopic human interfaces," *IEEE Robotics & Automation Mag.*, vol. 15, no. 4, pp. 50-57, Dec. 2008. Article (CrossRef Link)
[14] W.A. IJsselsteijn, H. de Ridder, J. Vliegen, "Subjective Evaluation of Stereoscopic Images: Effects of Camera Parameters and Display Duration," *IEEE Trans. on Circuits and Systems for Video Technol.*, vol. 10, no. 2, pp. 225-233, Mar. 2000. Article (CrossRef Link)
[15] J. You, L. Xing, A. Perkis, X. Wang, "Perceptual Quality Assessment for Stereoscopic Images based on 2D Image Quality Metrics and Disparity Analysis," in *Proc. of 15th Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Jan. 13-15, 2010.
[16] A. Benoit, et al., "Quality Assessment of Stereoscopic Images," *EURASIP J. on Image and Video Process., Special issue on 3D Image and Video Process.*, vol. 2008, pp. 1-13, 2008.
[17] Z. Wang, et al., "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004. Article (CrossRef Link)
[18] M. Carnec, P. Le Callet, D. Barba, "An Image Quality Assessment Method based on Perception of Structural Information," in *Proc. of the IEEE Int. Conf. on Image Processing (ICIP '03)*, vol. 2, pp. 185-188, Sept. 2003. Article (CrossRef Link)
[19] C.T.E.R. Hewage, et al., "Quality Evaluation of Color Plus Depth Map-based Stereoscopic Video," *IEEE J. of Sel. Topics in Signal Process.*, vol. 3, no. 2, pp. 304-318, Apr. 2009. Article (CrossRef Link)
[20] L. Shen, J. Yang, Z. Zhang, "Quality Assessment of Stereo Images with Stereo Vision," *Int. Congress on Image and Signal Processing*, pp. 1-4, 2009. Article (CrossRef Link)
[21] J. Yang et al., "Objective Quality Assessment Method of Stereo Images," in *Proc. of 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 1-4, 2009. Article (CrossRef Link)
[22] P. Campisi, P.L. Callet, E. Marini, "Stereoscopic Images Quality Assessment," in *Proc. of 15th European Signal Process. Conf. (EUSIPCO)*, pp. 2110-2114, Sep., 2007.
[23] H. Shao, X. Cao, G. Er, "Objective Quality Assessment of Depth Image based Rendering in 3DTV System," in *Proc. of 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 1-4, 2009. Article (CrossRef Link)
[24] Z.M.P. Sazzad, et al., "Stereoscopic Image Quality Prediction," in *Proc. of Int. Work. on Quality of Multimedia Experience*, pp. 180-185, July 2009. Article (CrossRef Link)
[25] D. Huang, M. Yu, Y. Yang, "Image Evaluation Algorithm for Right View of Stereoscopic Video," in *Proc. of Int. Conf. on Signal Processing*, pp. 1051-1054, Oct. 2008. Article (CrossRef Link)
[26] A. Smolic, et al., "Coding algorithms for 3DTV- A Survey," *IEEE Trans. on Circuits and Systems for Video Technol.*, vol. 17, no. 11, pp. 1606–1620, 2007. Article (CrossRef Link)
[27] F. Lu, et al., "Quality Assessment of 3D Asymmetric View Coding using Spatial Frequency Dominance Model," in *Proc. of 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 1-4, 2009. Article (CrossRef Link)
[28] P. Seuntiënsa, L. Meestersa, W. IJsselsteijna, "Perceptual Evaluation of JPEG Coded Stereoscopic Images," *SPIE Stereoscopic displays and virtual reality systems*, vol. 5006, pp. 215-226, Jan. 2003. Article (CrossRef Link)
[29] M.G. Perkins, "Data Compression of Stereopairs," *IEEE Trans. on Communication*, vol. 40, no. 4, pp. 684-696, 1992. Article (CrossRef Link)
[30] P.W. Gorley, N.S. Holliman, "Stereoscopic image quality metrics and compression," in *Proc. of SPIE-IS&T Electronic Imaging, Stereoscopic Displays and Virtual Reality Systems*, vol. 6803, Jan. 2008. Article (CrossRef Link)
[31] E. Peli, "Contrast in complex images," *J. Opt. Soc. Am. A*, vol. 7, no. 10, pp. 2032-2040, Oct. 1990.

Article (CrossRef Link)
[32] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," in *Proc. of Int. Conf. on Computer Vision*, vol. 2, pp. 1150-1157, 1999. Article (CrossRef Link)
[33] M.A. Fischler, R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. of the ACM*, vol. 24, pp. 381-395, June 1981. Article (CrossRef Link)
[34] A. Boev, et al., "Modelling of the stereoscopic HVS," *MOBILE 3DTV Technical Report D5.3*, Apr. 2009.
[35] Y. Zhang, et al., "Stereoscopic Visual Attention Model for 3D Video," *Lecture Notes in Computer Sci.*, vol. 5916, pp. 314-324, Dec. 2009. Article (CrossRef Link)
[36] L. Itti, C. Koch, "Computational Modeling of Visual Attention," *Nature Rev. Neuroscience*, vol. 2, no. 11, pp. 194-203, Mar. 2001. Article (CrossRef Link)
[37] A.M. Treisman, G. Gelade, ''A feature-integration theory of attention,'' *Cognitive Psychology.*, vol. 12, pp. 97-136, 1980. Article (CrossRef Link)
[38] J.W. Crabtree, et al., "Contributions of Y- and W-cell Pathways to Response Properties of Cat Superior Colliculus Neurons: Comparison of Antibody- and Deprivation-induced Alterations," *J. Neurophysiol.*, vol. 56, no. 4, pp. 1157-1173, 1986.
[39] M. Mancas, B. Gosselin, B. Macq, "A Three-level Computational Attention model," in *Proc. of ICVS Workshop on Comput. Attention & Appl.*, 2007.
[40] A.K. Jain, *Fundamentals of digital image processing*, Prentice Hall, 1989.
[41] Y. Zhai, M. Shah, "Visual Attention Detection in Video Sequences using Spatiotemporal Cues," in *Proc. the 14th ACM Int. Conf. on Multimedia*, pp. 815-824, Dec. 2006.
[42] S. Daly, "The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity," *Digital Image and Human Vision*, Cambridge, MIT press, pp. 179-206, 1993.
[43] B.A. Wandell, *Foundations of vision*, Sinauer Associates, Inc. Pub., 1995.
[44] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis", *IEEE Trans. PAMI*, vol. 20, no. 11, pp. 1254-1259, 1998. Article (CrossRef Link)
[45] C. Zitnick, T. Kanade, "A Cooperative Algorithm for Stereo Matching and Occlusion Detection, *Robotics Institute Tech. Report*, CMU-RI-TR-99-35, Carnegie Mellon University, Oct. 1999.
[46] A.B. Watson, J. Hu, J.F. McGowan, "Digital Video Quality Metric based on Human Vision", *J. of Electronic Imaging*, vol. 10, no. 1, pp. 20-29, 2001. Article (CrossRef Link)
[47] Adobe Creative Team, "Adobe Photoshop CS4 classroom in a book," Adobe Press, 2008.
[48] A.B. Poirson, B.A. Wandell, "Appearance of Colored Patterns: Pattern-Color Separability," *J. Opt. Soc. Am. A*, vol. 10, no. 12, pp. 2458-2470, Dec. 1993. Article (CrossRef Link)
[49] CCIR, "Encoding Parameters of Digital Television for Studios," *CCIR Recommendation 601-2*, Int. Radio Consult. Committee, Geneva, 1990.
[50] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, "Numerical recipes : the art of scientific computing," Ch. 14, Cambridge University Press, 2007.

Jae-Jeong Hwang received the B.S., M.S., and Ph.D. degrees in electronic engineering from the Chonbuk National University in 1983, 1986, and 1992, respectively. He is currently full-professor at the Kunsan National University, Korea, and adjunct professor of RMIT University, Australia. His research interests are digital image/video coding & processing, information theory, object segmentation and tracking, and 2D/3D visual quality assessment and evaluation. He is the coauthor of *Techniques and standards for image, video and audio coding* (Prentice Hall, 1996) and *Fast Fourier transform – Algorithms and applications* (Springer, 2010).



Hong Ren Wu was born in Beijing, China, in 1956. He received the BEng. degree and the MEng. degree from University of Science and Technology, Beijing, P.R. China, in 1982 and 1985, and the PhD degree in Electrical and Computer Engineering from the University of Wollongong, N.S.W. Australia, in 1990, respectively. He is a full-professor at Royal Melbourne Institute of Technology, Australia. His research interests include fast DSP algorithms, digital picture compression and quality assessment, video processing and enhancment, embedded DSP systems and their industrial applications. He is a co-editor of the book, *Digital Video Image Quality and Perceptual Coding*, CRC Press, 2006 (ISBN: 0-8247-2777-0).