# Quality Assessment of Perceptual Crosstalk on Two-View Auto-Stereoscopic Displays

Jongyoo Kim, *Member, IEEE*, Taewan Kim, *Member, IEEE*, Sanghoon Lee, *Senior Member, IEEE*, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*— Crosstalk is one of the most severe factors affecting the perceived quality of stereoscopic 3D images. It arises from a leakage of light intensity between multiple views, as in auto-stereoscopic displays. Well-known determinants of crosstalk include the co-location contrast and disparity of the left and right images, which have been dealt with in prior studies. However, when a natural stereo image that contains complex naturalistic spatial characteristics is viewed on an auto-stereoscopic display, other factors may also play an important role in the perception of crosstalk. Here, we describe a new way of predicting the perceived severity of crosstalk, which we call the Binocular Perceptual Crosstalk Predictor (BPCP). BPCP uses measurements of three complementary 3D image properties (texture, structural duplication, and binocular summation) in combination with two well-known factors (co-location contrast and disparity) to make predictions of crosstalk on two-view auto-stereoscopic displays. The new BPCP model includes two masking algorithms and a binocular pooling method. We explore a new masking phenomenon that we call duplicated structure masking, which arises from structural correlations between the original and distorted objects. We also utilize an advanced binocular summation model to develop a binocular pooling algorithm. Our experimental results indicate that BPCP achieves high correlations against subjective test results, improving upon those delivered by previous crosstalk prediction models.**

***Index Terms*— Auto-stereoscopic display, crosstalk, quality assessment, human visual system.**

## I. Introduction

MOTIVATED by the great success of 3D action movies, the 3D-related cinematic industry has recently experienced tremendous growth. Nowadays it is possible to view 3D content on a wide variety of media, including cinematic displays, 3D TVs and personal mobile devices. At the same time, the availability of 3D content for all types of 3D displays has grown rapidly. Contemporary 3D display devices can be broadly categorized into two classes: stereoscopic and auto-stereoscopic. Stereoscopic displays are popular in the market and are able to present high resolution, realistic 3D visual experiences. However, these devices require the user to don cumbersome polarized or shutter spectacles that separate the left and right views, thereby presenting the corresponding image to each eye. Auto-stereoscopic displays do not require the use of any spectacles, hence are often termed 'glasses-free'. Commercial flat-panel auto-stereoscopic systems generally use a parallax barrier or a lenticular lens system to present distinct images to the two eyes. In both stereoscopic and auto-stereoscopic displays, crosstalk is defined as the leakage of either or both of the left/right views to the other view. While current technology has sufficiently advanced that crosstalk may be regarded as effectively negligible in most commercial two-view stereoscopic systems, it is still a significant design element in auto-stereoscopic systems. The presence of crosstalk significantly affects 3D picture quality, hence remains an important concern. The severity of crosstalk generally depends on the display geometry and on the location of the viewer. The perception of crosstalk may be of a ghostly double-image, greatly reducing the quality of the 3D experience [1].

We seek to develop methods to automatically predict the degree to which the quality of experience, when viewing a given 3D content, is impaired by the presence of crosstalk. This implies measurement of the degree of crosstalk that is present as well as its perceptual impact. Previous research on measuring the visual effects induced by crosstalk can be divided into two ways of thinking: hardware-oriented and perception-oriented approaches. Huang *et al.* defined two types of crosstalk: system crosstalk and perceived crosstalk [2]. The former is the degree of displayed crosstalk arising from imperfections in the hardware system independent of the displayed 3D content, while the latter measures the perceived degree of crosstalk. System crosstalk can be estimated by measuring displayed luminances using simplistic visual test signals [3], [4]. Some studies have modeled the 'viewing zone' as a function of viewing angle and distance [5]–[7]. In [8], Woods surveyed a wide range of system crosstalk related issues that are observed on stereoscopic displays.

The major drawback of only measuring system crosstalk is that this approach does not account for those aspects of visual perception that may, or may not be affected by occurrences of system crosstalk. Importantly, perceived crosstalk is affected

Manuscript received September 9, 2015; revised May 23, 2016, November 5, 2016, and February 25, 2017; accepted June 7, 2017. Date of publication June 19, 2017; date of current version July 25, 2017. This work was supported by the Institute for Information & communications Technology Promotion through the Korea Government (MSIP) (development of mobile GPU hardware for photo-realistic realtime virtual reality) under Grant 2016-0-00204. The work of A. C. Bovik was supported by the U.S. National Science Foundation under Grant 11S-0917175 and Grant 11S-1116656. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jens-Rainer Ohm. *(Corresponding author: Sanghoon Lee.)*

J. Kim, T. Kim, and S. Lee are with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea (e-mail: jongky@yonsei.ac.kr; top.kim@sk.com; slee@yonsei.ac.kr).

A. C. Bovik is with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA (e-mail: bovik@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2717180

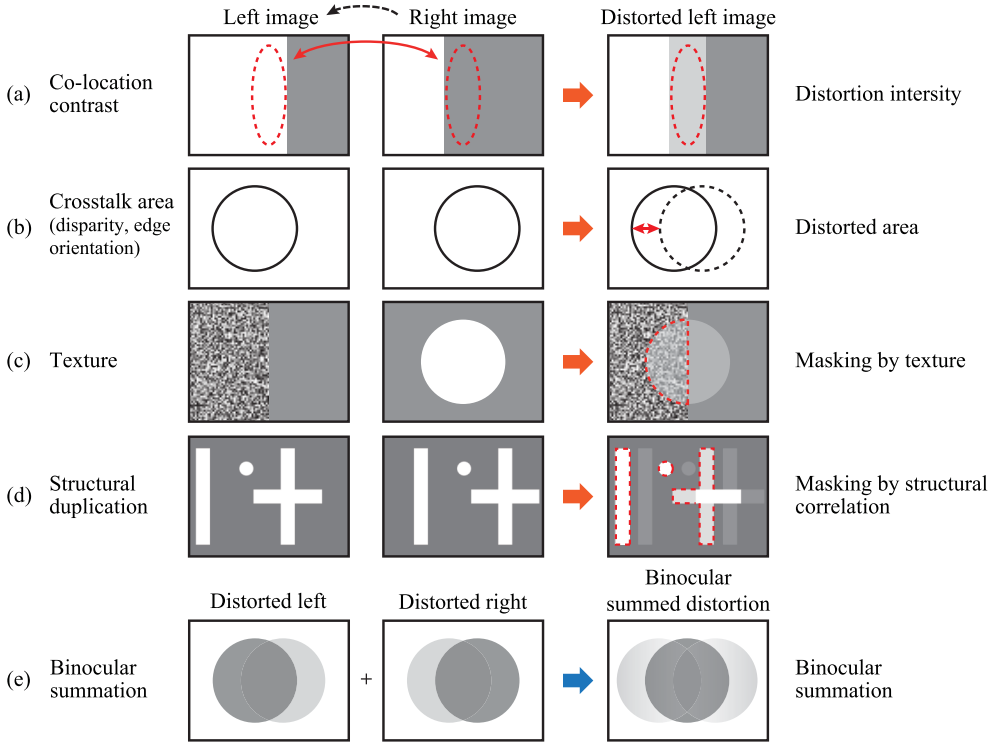



Fig. 1. Major factors affecting perceptual crosstalk. (a) Co-location contrast. (b) Crosstalk area (disparity, edge orientation). (c) Texture. (d) Structural duplication. (e) Binocular summation.

not only by the viewing geometry but also by the image content, which is not describable from measurements on a display device. Huang *et al.* studied image contrast features related to perceived crosstalk. Wang *et al.* investigated perceived crosstalk issues related to contrast and binocular disparity using synthetic images [9]. Thresholds of crosstalk visibility and acceptability were measured by performing subjective experiments. Pastoor also measured crosstalk visibility while controlling disparity and contrast [10], and found that crosstalk must be quite small not to be noticeable. Seuntiens found a similar crosstalk threshold visibility [11], as did Tsirlin *et al.* using two naturalistic color images [1]. These authors also found that visible crosstalk adversely affects the perception of depth.

A number of researchers have studied the use of features other than contrast and disparity to predict crosstalk-induced annoyance. Lipton proposed measures of texture complexity and edge ghosting along with contrast and parallax as factors affecting perceived crosstalk [12]. Huang *et al.* [2] proposed a model of perceived crosstalk as a product of contrast and system crosstalk. He concluded that perceived crosstalk is proportional to system crosstalk. Wang *et al.* [13] surveyed the effects of system crosstalk, shadow and linear perspective on subjective 3D image quality. More recently, Xing *et al.* [14] proposed a quality assessment metric which predicts perceptual crosstalk using 2D and 3D attributes of shadow degree, separation distance and spatial position. In [14], 2D attributes are measured using the Structural Similarity index (SSIM) [15], while 3D depth information is used to formulate regional weights.

Here we develop a new model for predicting the degree of perceived crosstalk that we dub the Binocular Perceptual Crosstalk Predictor (BPCP). BPCP embodies five critical determinants that affect perceptual crosstalk, as depicted in Fig. 1, including both monocular (Figs. 1(a)-(d)) and binocular (Fig. 1(e)) factors.

*(a) Co-location contrast and (b) Crosstalk area:* Prior studies [1], [2], [9]–[14] have shown that co-location contrast and crosstalk area (disparity) are important factors that affect perceptual crosstalk. Co-location contrast affects the intensity of crosstalk distortion (Fig. 1(a)). The area of distortion is affected by disparity and edge orientation (Fig. 1(b)).

*(c) Texture:*Local image texture determines the local spatial complexity and energy of the image. The presence of texture obscures other spatially coincident image changes, which is a well-known phenomena called contrast masking [16]. The contrast masking phenomenon has a strong influence on the perception of crosstalk as well. This is conceptually illustrated in Fig. 1(c), where the crosstalk distortion would be less visible on the heavily textured random noise region (red dotted region).

*(d) Structural duplication:* Structural correlations between crosstalk distortion and the original image is also a critical determinant of perceived crosstalk. This masking phenomenon, which we call duplicated structure masking, is a perceptual factor different from those used in previous studies on image quality assessment, although it is similar to the mutual masking concept used in the MOVIE index [17]. If a distorted image is compared to a pristine original version of it, the distorted image may contain duplicated or shifted

local structures, such as edges, as depicted in Fig. 1(d), where the red dotted lines delineate erroneously duplicated structures. The distortion may be hardly noticeable by a human observer. Section III-B describes how this new process is modeled.

*(e) Binocular summation:* The human visual system (HVS) perceives a single 3D image via collective processing of the signals from the two eyes. The result of this processing is either a local state of image fusion at and around the point of fixation, which involves a process of binocular summation, or a state of binocular rivalry, as shown in Fig. 1(e) [18]. By exploiting an accepted model of binocular summation, we develop a computational model of binocular crosstalk distortions and an algorithm that computes a cyclopean distortion map from a stereo image pair.

In Table I, relevant aspects of perceptual crosstalk are summarized in the context of previous research on the problem. Measurements of co-location contrast and disparity are common factors used in all of the studies. The use of hard edges [12] and linear perspective models [13] relate to the concept of structural duplication as explained in Fig. 1. Shadows [13] relate to both co-location contrast and disparity. Although crosstalk only arises in 3D displays, most previous studies have only studied a single view scenario. The authors of the study in [14] assume that the HVS is more sensitive to crosstalk on nearer objects, which may not be true. Here, we do not assume any preferences on depth. Instead, we deploy a binocular summation model derived from psychophysical experiments to model binocular perception of the cyclopean image [19].

The rest of the paper is organized as follows. Section II analyzes crosstalk as an image distortion in the context of other common distortions such as noise and blocking artifacts. Aspects of visual masking phenomena relevant to crosstalk distortion are also discussed. Section III develops contrast masking and duplicated structure masking models, and a binocular pooling model that derives from psychophysical studies of binocular summation. Section IV is devoted to evaluation of the overall crosstalk prediction engine on subjective data, and a statistical evaluation of the results.

## II. Analysis of Crosstalk Distortion

### A. Crosstalk as an Image Distortion

Modern algorithms that conduct full reference quality assessment (FR QA) of images require that both a reference image and a distorted image be available [20], [21]. For the problem of crosstalk distortion assessment, the reference and distorted images must be particularly defined in regards to the presumed observer. Fig. 2 depicts the conceptual process of perceiving a 3D visual signal on an auto-stereoscopic display. Each cell on the display panel is assigned to display a pixel from either the left or the right view. The parallax barrier plays the roll of blocking the view of cells to the unintended eye. If the left view is perfectly separated from the right, then the luminance perceived by the left eye will range from 0 to 255. However, if crosstalk occurs at a level of $p\%$,[1] then the maximum luminance will increase to $255(1+p)$. Unlike common distortions such as image compression or channel errors, the digital image signal is not actually modified or intrinsically distorted by crosstalk. Instead, crosstalk arises from a failure of spatio-temporal synchronization of the signals delivered to the two eye. To measure this, it is necessary to define the 'reference' (i.e., ideally synchronized) and distorted images in order to be able to numerically analyze crosstalk.

We now describe the concepts of an 'expected' image, which plays a similar role as a 'reference' image in classical image quality models, and an 'observed' image possibly suffering from crosstalk. The (luminance, for simplicity) expected image $I_e$ is observed when the left and right views exactly coincide. By contrast, a distorted image $I_o$ is observed when the left and right views are different, as shown in Fig. 2. Given a level $p$ of system crosstalk on a 3D display, the two types of images are defined as

$$I_e^L(x,y) = (1+p)\cdot I^L(x,y) \tag{1}$$

$$I_o^L(x,y) = I^L(x,y) + p\cdot I^R(x,y) \tag{2}$$

at coordinate $(x,y)$. The superscripts $L$ and $R$ denote the image content intended for viewing by the respective left and right eyes. Note that $I_e$ and $I_o$ are assumed to be represented as unbounded float values which are not necessarily bound to the range $[0, 255]$. The distorted image may be viewed as the sum of a reference image and an error signal, in which case

$$I_o^L = I_e^L + p\left(I^R - I^L\right) \tag{3}$$

where $p\left(I^R - I^L\right)$ is the error signal. The error signal is utilized to define the concept of difference map and to model duplicated structure masking later, in Section III.

### B. Perceptual Issues of Crosstalk Distortion

Beginning with the ideas of the expected and observed images, an FR metric can be developed to quantify the degree of experienced viewer crosstalk. In the right half of Fig. 2 (representing visual processing), the left and right eyes receive a potentially crosstalk-distorted visual signal. For simplicity of understanding, separate the visual processing steps into independent monocular and binocular processes. Monocular processing includes visual masking, which explains content-dependent decreases in visual sensitivity. The binocular stage includes a process of binocular summation which accounts for stereoscopic fusion into a single image, which weights the perceptual importance of local information from the left and right views.

Visual masking is a fundamental aspect of the visual perception of distortions. We deploy two forms of visual masking modes: contrast masking and duplicated structure masking The former is widely used in visual quality assesment models which describe the phenomenon that the visibility of distortion may be decreased by the presence of textured content [22]. This is important in the current context, since crosstalk distortions may be rendered less visible on textured regions.

The latter type, which we call duplicated structure masking, may be viewed as a form of dichoptic masking phenomena that

[1] $p\%$ indicates the percentage of leaked signal from an unintended channel to the intended channel.

TABLE I

PREVIOUS STUDIES OF PERCEPTUAL CROSSTALK. THE SYMBOL OF '○' INDICATES THAT THEY CONSIDERED THE NATURAL IMAGE OR PROVIDED THE METRIC, AND THE SYMBOL OF '-' INDICATES THAT THEY DID NOT CONSIDERED THE FACTORS

| | Considering attributes | | | |
|---|---|---|---|---|
| Researches | Monocular factors | Binocular factors | Natural Image | Metric |
| [1], [2], [9]–[11] | Co-location contrast, Disparity | - | - | - |
| [12] | Co-location contrast, Disparity, Hard edges | - | - | - |
| [13] | Co-location contrast, Disparity, Linear perspective, Shadows | - | - | - |
| [14] | Co-location contrast, Disparity | Spatial position in 3D space | ○ | ○ |
| **Proposed (BPCP)** | Co-location contrast, Disparity, Texture, Structural duplication | Binocular summation | ○ | ○ |

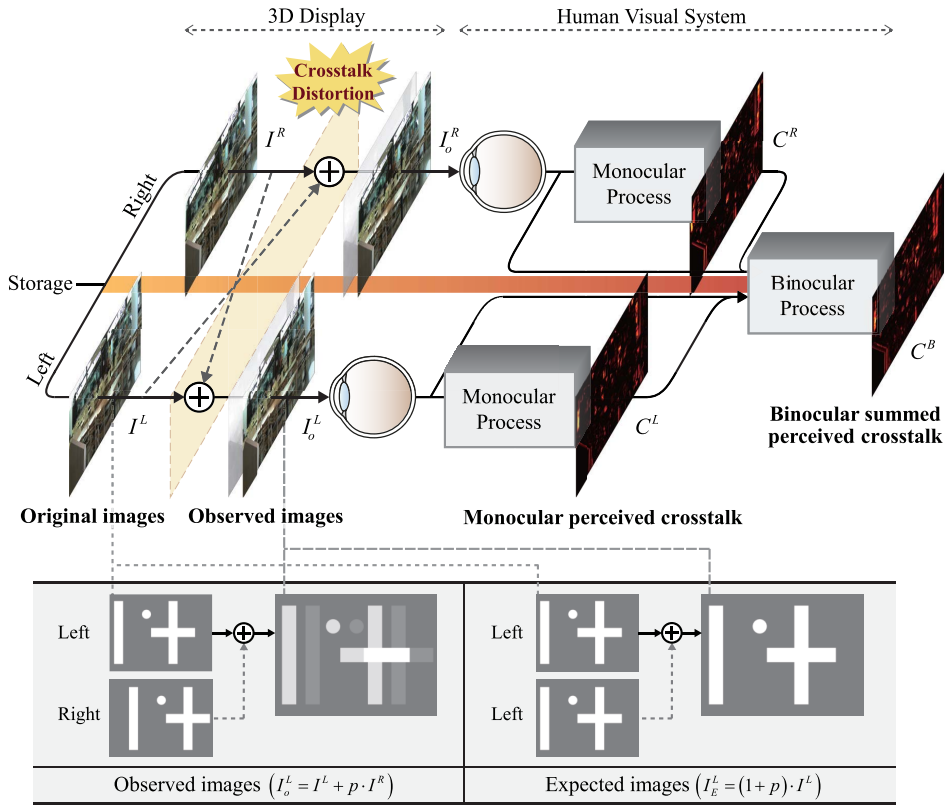



Fig. 2. Pathway from the display to the viewer for the perception of crosstalk distortion and conceptual definition of an observed image and an expected image.

describes the interaction of content with crosstalk in our model of crosstalk distortion visibility. Crosstalk distortion presents as an induced diplopia, viz., structural duplication between the two views. The degree of apparent structural duplication depends heavily on the local image contrast and on the local disparity distribution. When perception of the crosstalk signal coincides with perception of original signal structure in space and time, the perceived 'duplicated structure' distortion may be moderated by a masking effect.

Fig. 3 illustrates the duplicated structure masking effect over an image patch where there occurs a small disparity between the left and right views. The left observed image $I_o^L$ (a) is generated by a linear summation of the left and right images $I^L$ and $I^R$ with the crosstalk signal as described by (2). The observed image (a) contains the error signal (c), where the regions labeled A, B and C depict crosstalk distortion. All three regions have the same magnitude of luminance difference relative to the expected image (b). When the observed image (a) is shown to a viewer, region A will be perceived as suffering from severe crosstalk distortion. By contrast, the viewer is much less likely to notice the crosstalk distortion within region C, since there are no induced artifacts or structural distortions. If no significant structural differences occur between corresponding regions of each view, the distortion is more likely to be suppressed by dichoptic masking.

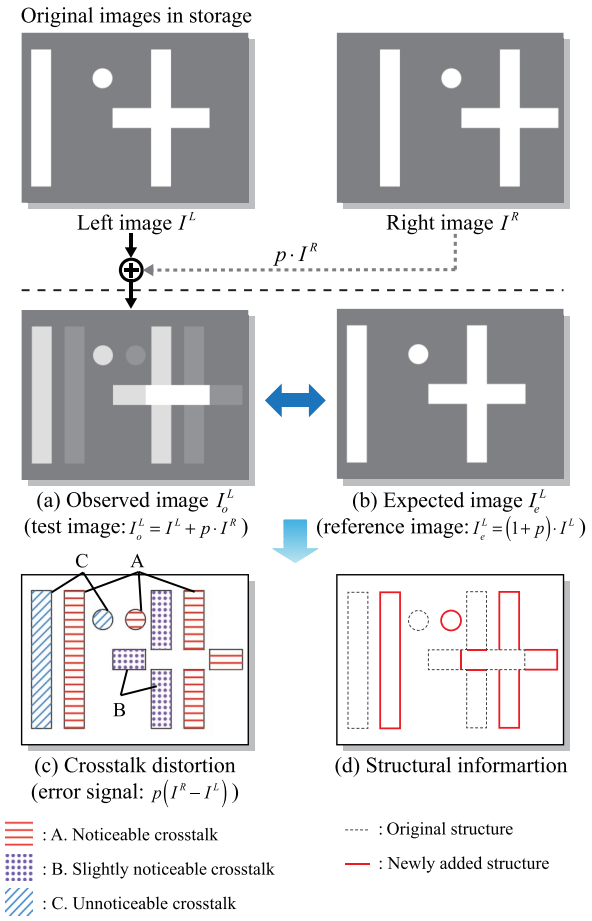



Fig. 3. Example of duplicated structure masking. Filled areas labeled by A, B and C have distorted pixel values. Region A is recognized as crosstalk distortion. Perception of error in region B is reduced. The error in region C is hardly noticeable. (a) Observed image $I_o^L$(test image:$I_o^L = I^L + p.I^R$). (b) Expected image $I_e^L$(reference image: $I_e^L = (1 + p).I^L$). (c) Crosstalk distortion (error signal:p($I^R - I^L$)). (d) Structural informartion.

High-contrast, edge-like structures strongly contribute to the perception of crosstalk. For example, in regions A, superfluous edges will be present in the perceived **stereoscopic 3D (S3D)** image, making crosstalk easily noticeable. However, regions C contribute no supernumerary edges to the perceived S3D image, leading to less visibility of crosstalk. When the crosstalk distortion is highly similar to the original local image structure depicted in region B, it is difficult to determine whether the region is corrupted or not. In such instances, the most important cues are edges. If there are visible edge artifacts, then it is likely that there is noticealbe crosstalk occuring around them.

## III. Objective Metric Development

The objective quality index which we describe next predicts the degree of perceptual crosstalk experienced when viewing natural **S3D** images. As shown in Fig. 4, the flow of image feature analysis follows two paths: monocular and binocular processing. The stages of processing shown in the figure correspond to the processes shown in Fig. 2 consisting of monocular and binocular element. At the monocular stage, each left and right image is processed independently. This monocular process consists of two processes of luminance comparison and two types of masking.

The luminance comparison process extracts co-located contrasts between the input S3D stereo pair. Using the computed difference map, texture and duplicated structure masking nonlinearities are applied to produce a pair of crosstalk maps. In the binocular process, the results of the left and right monocular processes are combined to derive an overall cyclopean quality score. The perceptual weights are computed from the contrast energies of the left and right images. One cyclopean crosstalk map is then generated from the left and right crosstalk maps.

Preliminary to the objective metric, the expected images, $I_e^L$ and $I_e^R$, and the observed images, $I_o^L$ and $I_o^R$, are derived using (1) and (2), and subsequently used as the reference and test images respectively. It is well-known that the strength of crosstalk is significantly determined by co-located luminance and contrast differences between the left and right images [1], [2], [9]–[14]. The error signal in (3) is used to define an absolute difference map, which contains the co-located contrast information:

$$D = p\left|I^L - I^R\right|, \tag{4}$$

To capture the effects of neighboring pixels and any structural correlation between the original and crosstalk structures, two visual masking nonlinearities are applied to the difference map.

### A. Contrast Masking

We model the contrast masking effect by adapting Daly's masking function with a small modification [23]. The first step in the contrast masking model is a Gabor decomposition of the expected image $I_e$ [24], [25], followed by a contrast sensitivity function (CSF) based weighting of each frequency band. The particular CSF model that we use here captures both the reduced spectral sensitivity at low and high spatial frequencies as well as away from the cardinal orientations. We follow the standard CSF model proposed by Watson and Ahumada [26], which has been shown to closely fit the ModelFest stimuli [27]. The decline in visibility along oblique orientations as compared to the horizontal and vertical directions is modeled using the 'oblique effect filter' (OEF) [28], [29], which is modeled as [26]:

$$oef\,(f,\theta) = \begin{cases} 1 - \left(1 - \exp\left(-\dfrac{f-\zeta}{\lambda}\right)\right)\sin^2(2\theta) & \text{for } f > \zeta \\ 1 & \text{for } f \le \zeta \end{cases} \tag{5}$$

where $\lambda = 13.57$ cycles/degree and $\zeta = 3.48$ cycles/degree. The overall CSF/OEF model is constructed as the product of CSF model with (5) as follows [26]

$$CSF\,(f,\theta) = csf\,(f) \cdot oef\,(f,\theta) \tag{6}$$

$$csf\,(f) = \operatorname{sech}\left[(f/f_0)^q\right] - a \cdot \operatorname{sech}\left[f/f_1\right] \tag{7}$$

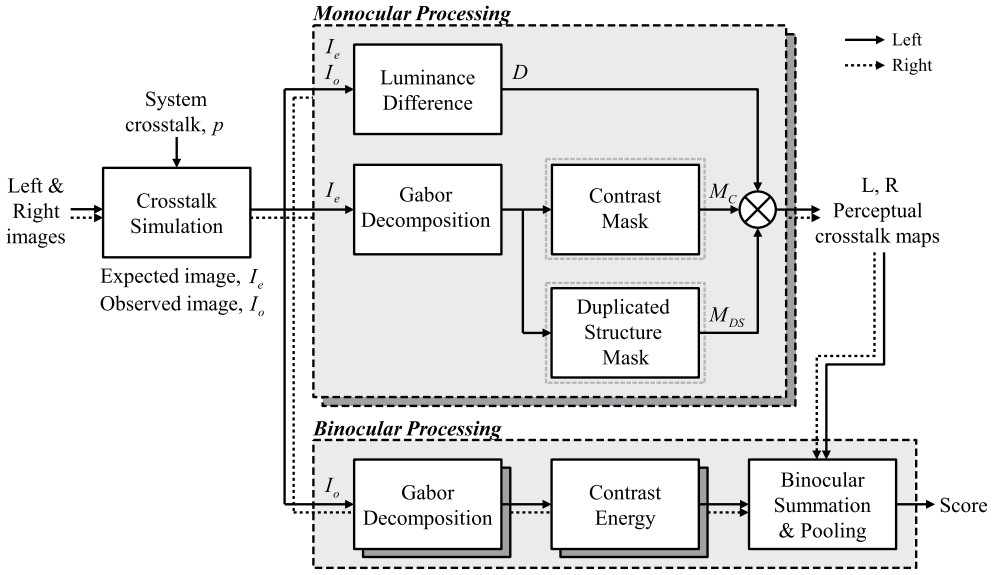



Fig. 4. Block diagram of the perceptual crosstalk predictor.

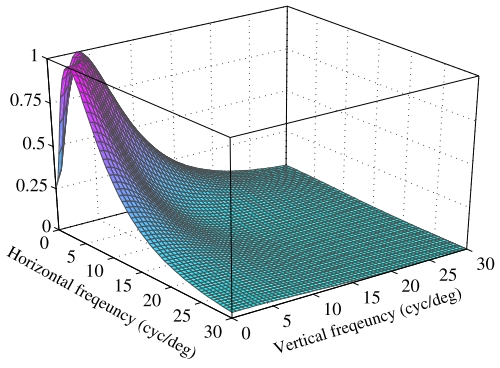



Fig. 5. Combined CSF/OEF model plotted against spatial frequency.

where $f_0 = 4.3469$, $f_1 = 1.4476$, $a = 0.8514$ and $q = 0.7929$. Fig. 5 plots the CSF/OEF as a function of horizontal and vertical spatial frequencies.

The neurons in the primary visual cortex are bandpass spatial frequency and orientation selective with a gaussian shaped passband. To model this, we decompose each image into discrete frequency bands using a 2-D Gabor filter bank. Gabor filters accurately model the receptive fields of V1 simple cells [30]. A Gabor filter kernel is given by

$$g(x, y, f, \theta) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2}\left(\left(\frac{x'}{\sigma}\right)^2+\left(\frac{y'}{\sigma}\right)^2\right)\right]\exp\left(i2\pi f x'\right) \quad (8)$$

$$x' = x\cos\theta + y\sin\theta \quad (9)$$

$$y' = -x\cos\theta + y\sin\theta \quad (10)$$

where $\sigma$ scales the Gaussian envelope and $f$ is the radial center frequency. In our model, six frequency bands having radial center frequencies 13.33, 9.43, 6.67, 4.71, 3.33 and 2.36 cycle per degree, and four orientations 0°, 45°, 90° and 135° were implemented. The bandwidth of each filter was fixed at 1 octave. which approximates the bandwidth of simple cells.

The Gabor responses of the expected image $I_e$ are normalized using the CSF/OEF model. Daly's model of contrast threshold elevation [23] is adopted as follows:

$$T_{f,\theta} = \left(1 + \left(k_1\left(k_2\left|c_{f,\theta}\cdot CSF(f,\theta)\right|\right)^s\right)^b\right)^{1/b} \quad (11)$$

where $c_{f,\theta}$ is a Gabor coefficient of the expected image $I_e$, $CSF(f,\theta)$ is the CSF evaluated at spatial frequency $f$, and the orientation is $\theta$. Further, $s$ is the slope of the contrast masking asymptote which, as modeled in [23], falls in the range $0.65 \leq s \leq 1$ depending on the subject's level of naÃ¯vity with respect to the stimulus, $b$ determines how closely the curve will follow the asymptote ($2 \leq b \leq 4$), and $k_1$ and $k_2$ determine the pivot point. The visibility threshold of the image becomes large when it has high contrast and a large CSF/OEF value. In our experiments, we assumed that the subjects had little experience viewing this kind of test stimuli, hence we set $s = 1.0$. For the other parameters, we followed the reasoning used in [31], fixing $b = 4.0$, $k_1 = 0.0153$, and $k_2 = 392.5$.

After calculating the threshold elevation for each frequency and orientation band, a contrast masking map is obtained by averaging the thresholds over all of the sub-bands:

$$M_C = 1/\left(\frac{1}{N_f\cdot N_\theta}\sum_f\sum_\theta T_{f,\theta}\right) \quad (12)$$

where $N_f$ and $N_\theta$ are the number of frequency levels and orientations in the Gabor filter bank, respectively.

### B. Duplicated Structure Masking

Duplicated structural masking occurs when the structures within a distorted region caused by crosstalk resemble the original signal structure. A reference image $I_e^L$ and the error

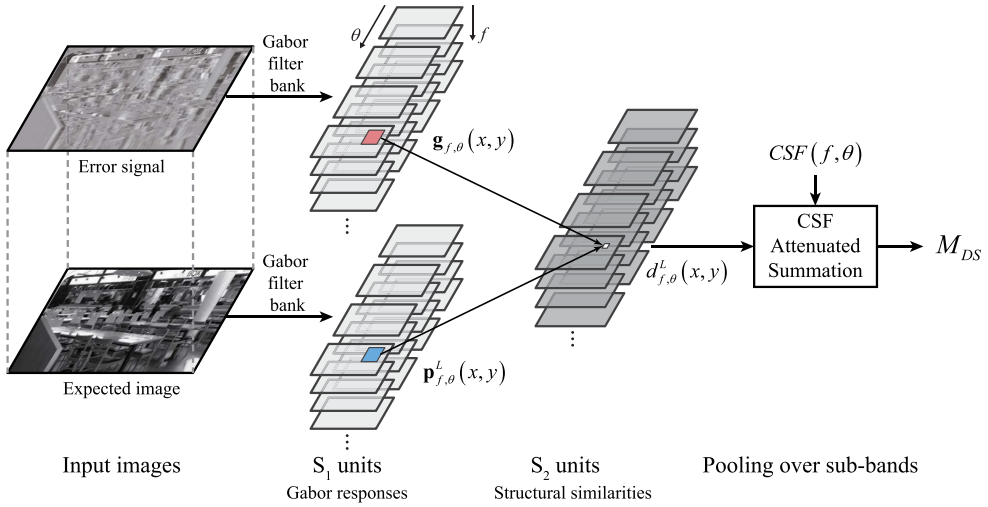



Fig. 6. Conceptual process of duplicated structure masking.

signal $p\left(I^R - I^L\right)$ defined in (3) are compared to derive the masking weight.

To quantify structural duplication, we adopted a processing model that is reminiscent of the Gabor-based object recognition scheme in [32], where features for object recognition are based on models of simple and complex cell responses in human visual cortex. As in [32], we model two types of simple units $S_1$ and $S_2$ having selectivity for specific patterns. In their model, simple units alternate with complex units in the processing flow. We model only the simple units $S_1$ using a Gabor filter bank, which decomposes the input into frequency- and orientation-selective subbands. In units $S_2$, the delivered signals are tuned to more specific structures. The distances between the input and stored prototypes that are obtained by a training process are measured using a Gaussian radial basis function (RBF). Since the prototypes are obtained from random image patches, they contain more complex pattern than $S_1$ units.

The aim of the duplicated structure masking model is to account for structural duplications between the error signal $p\left(I^R - I^L\right)$ and the expected image $I_e^L$. Fig. 6 shows the overall flow of the duplicated structure masking process, which is described in detail next.

As in [32], the tuning responses of $S_1$ units are obtained as the outputs of a Gabor filter bank. The same Gabor filter bank is used as in the previously-described contrast masking model. Let the Gabor responses to the left and right expected images and the error signal be denoted $c^L_{f,\theta}$, $c^R_{f,\theta}$. The absolute Gabor response to the error signal is

$$d_{f,\theta} = p \cdot \left| c^L_{f,\theta} - c^R_{f,\theta} \right|. \tag{13}$$

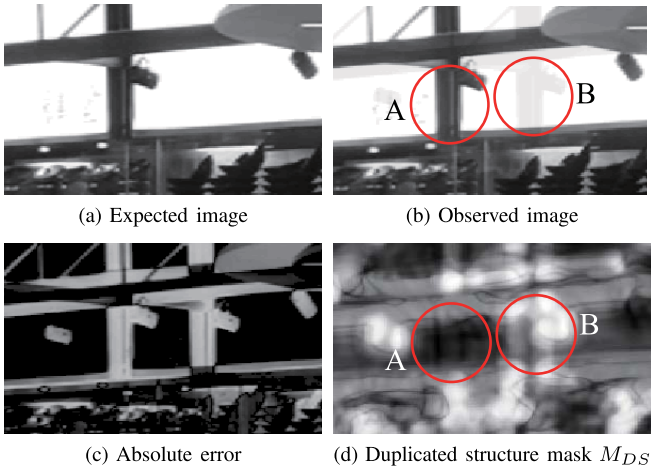
Using the model of $S_2$ units, a structural similarity measurement is then derived. Unlike [32], we use image patches from an expected image $I_e^L$ as tuning prototypes instead of randomly sampled patches. Thus a strong response by a unit $S_2$ occurs when there is a high correlation between the error signal and the expected image patch. First, an image patch $\mathbf{g}_{f,\theta}(x,y)$ centered at $(x,y)$ is extracted from the Gabor response to the error signal $d_{f,\theta}$. An image patch $\mathbf{p}^L_{f,\theta}(x,y)$ of like dimensions is then extracted at the same location from the Gabor responses to the left expected image $c^L_{f,\theta}$. An RBF is then used to compare the structures of the error signal and the expected image:

$$ds^L_{f,\theta}(x,y) = \exp\left(-\left\|\mathbf{g}_{f,\theta}(x,y) - \left|\mathbf{p}^L_{f,\theta}(x,y)\right|\right\|^2\right). \tag{14}$$

Thus $ds^L_{f,\theta}(x,y)$ is the structural similarity between the error signal and the left expected image on $(x,y)$, which is computed over the entire error signal to create a structural similarity map.

The structural similarities over all sub-bands are then linearly combined using the CSF/OEF weighting model (5) - (7) yielding

$$DS = \frac{1}{N_f \cdot N_\theta} \sum_f \sum_\theta ds_{f,\theta} \cdot CSF(f,\theta) \tag{15}$$




(a) Expected image (b) Observed image

(c) Absolute error (d) Duplicated structure mask $M_{DS}$

Fig. 7. Examples of duplicated structure masking. Although regions A and B have similar absolute errors as shown in (c), object A has a much lower masking value (d). (a) Expected image. (b) Observed image. (c) Absolute error. (d) Duplicated structure mask $M_{DS}$.

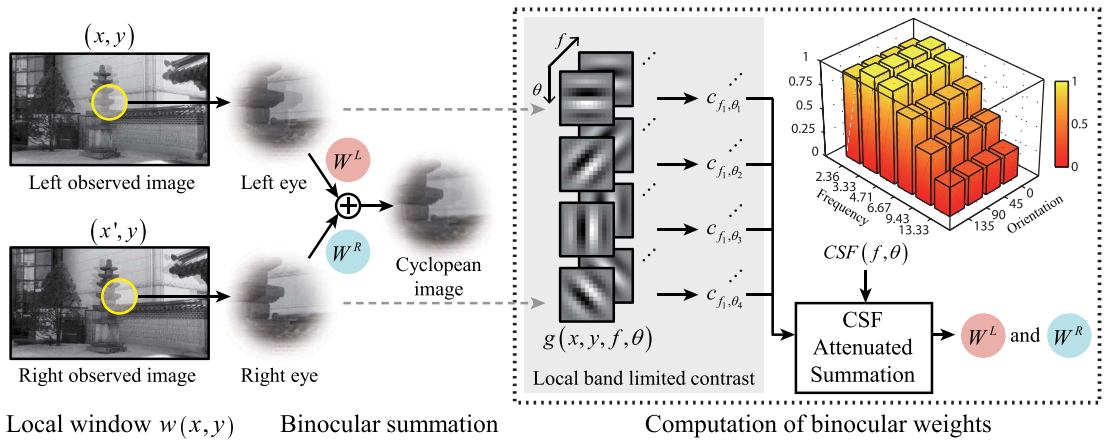



Fig. 8. Flow diagram of the binocular pooling method used in the crosstalk prediction model.

where $N_f$ and $N_\theta$ are the number of frequency levels and orientations in the Gabor filter bank. Since the values of both $ds_{f,\theta}$ and $CSF$ fall within $[0, 1]$, define the final masking factor:

$$M_{DS} = 1 - DS \tag{16}$$

Fig. 7 shows experimental examples of the duplicated structure masking model. Although both objects A and B result in similar absolute errors (Fig. 7(c)), the crosstalk distortion on object B is likely to be much more noticeable than that on object A under the duplicated structure masking model. The final masking map $M_DS$ is shown in Fig. 7(d); note that the masking value on object A is much smaller than that on object B.

Finally, the monocular crosstalk map is derived by multiplying the two masks, $M_C$ and $M_{DS}$, by the difference map $D$

$$C = D \cdot (M_C)^\alpha \cdot (M_{DS})^\beta \tag{17}$$

where $\alpha > 0$ and $\beta > 0$ balance the relative importances of the contrast and duplicated structure masking terms.

### C. Binocular Pooling

An essential ingredient of the binocular perception of crosstalk is the manner by which the signals from the two eyes are combined into a single cyclopean image. Binocular pooling makes an important role that discriminates quality of experience assessment of S3D contents from the 2D contents [33]–[35]. We deploy an effective binocular pooling model which we apply to the left and right crosstalk maps towards arriving at a single cyclopean crosstalk annoyance score. The binocular summation model of [19] provides a modern, comprehensive computational approach for generating a perceived cyclopean image.

Fig. 8 diagrams the processing flow of the binocular pooling method. The pooling process deploys visual weights $W^L$ and $W^R$ that correspond to the left and right images. In our implementation of the method of [19], a Gabor filter bank is used to estimate the local band-limited contrast energies of both left and right images which are used to define a set of visual weights. Finally, the CSF/OEF model described earlier is used to normalize the local band-limited contrast energies [36], which are then utilized to obtain the final visual weights.

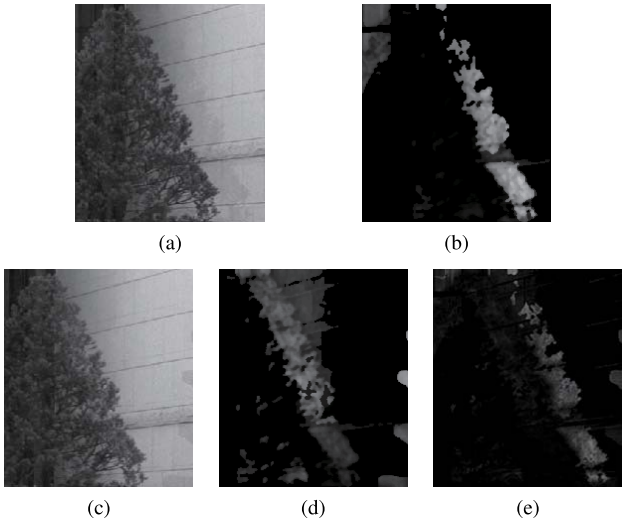



Fig. 9. Examples of the binocular pooling stage. (a) Left observed image. (b) Left crosstalk map. (c) Right observed image. (d) Right crosstalk map. (e) BPCP.

The sub-band responses (left and right) have the form

$$s_{f,\theta}(x, y) = I_o(x, y) * g(x, y, f, \theta) \tag{18}$$

where $I_o$ is the observed image and $g(x, y, f, \theta)$ is a Gabor filter indexed by radial frequency $f$ and orientation $\theta$. Then form the weighted responses

$$ce_{f,\theta}(x, y) = \frac{\sum_{x', y' \in P(x,y)} w(x', y')\, s_{f,\theta}(x', y')}{L_m(x, y)} \tag{19}$$

where

$$L_m(x, y) = \sum_{x', y' \in P(x,y)} w(x', y')\, I(x', y') \tag{20}$$

is a weighted average over the window

$$w(x, y) = A \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right). \tag{21}$$

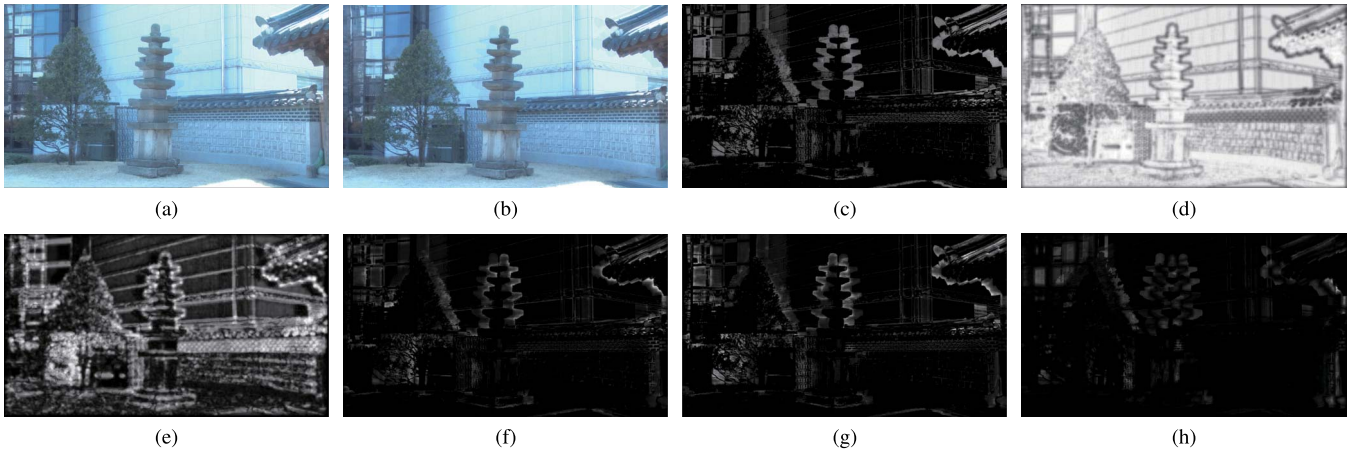



Fig. 10. Results of the proposed crosstalk prediction model. (a), (b) Left and right crosstalk simulated images. (c) Difference map. (d) Left texture mask. (e) Left duplicated structure mask. (f) Left crosstalk visibility map. (g) Right crosstalk visibility map. (h) Binocular summed crosstalk visibility map.

The visual weight at each coordinate $(x, y)$ is a CSF-weighted sum of the contrast energies:

$$W(x, y) = \sum_{f} \sum_{\theta} ce_{f,\theta}(x, y) \cdot CSF(f, \theta). \tag{22}$$

Finally, the crosstalk map is generated using the pair of visual weights $W^L$ and $W^R$:

$$C^B(x, y) = \frac{W^L(x, y)\, C^L(x, y) + W^R(x', y)\, C^R(x', y)}{W^L(x, y) + W^R(x', y)} \tag{23}$$

where the monocular crosstalk maps $C^L$ and $C^R$ are derived as in (17).

The final binocular perceptual crosstalk score is calculated via Minkowski summation of the elements of the crosstalk map $C^B$:

$$BPCP = \left( \frac{1}{N_{pixel}} \sum_{x,y} \left( C^B(x, y) \right)^{\gamma} \right)^{1/\gamma}, \tag{24}$$

where $N_{pixel}$ is the number of pixels in the binocular summed crosstalk map.

Fig. 9 shows examples of the binocular pooling process. Figs. 9(a) and (b) are the left observed image and the crosstalk map following the monocular processing stage. Figs. 9(c) and (d) show the corresponding right observed image and the associated crosstalk map. In Fig. 9(a), the crosstalk distortion falls outside the tree region, while in Fig. 9(c), the distortion falls within the tree region. The distortions in (a), (b) have higher contrast energy than those in (c). By contrast, the distortions in (c), (d) yield almost the same contrast energy as the leaves in the tree in (a). As a result, the error in (a) is emphasized, rather than the error in (c), as shown in Fig. 9(e).

Fig. 10 shows the result of the crosstalk visibility prediction process on an S3D sequence from the IEEE 3D database [37]. Figs. 10(a) and (b) show observed images from the left and right views, while Fig. 10(c) shows the pixel-wise difference map between the expected and the observed images. Contrast masking is determined using only the expected image (Fig. 10(d)). The duplicated structure masking process depends on the degree of correlation between the error signal and the expected image, as shown in Fig. 10(e). The monocular crosstalk visibility maps computed from the left and right images are shown in Figs. 10(f) and (g) respectively, which are arrived at by taking the products of the two masking maps with the difference map. The final example perceptual crosstalk map is shown in Fig. 10(h).

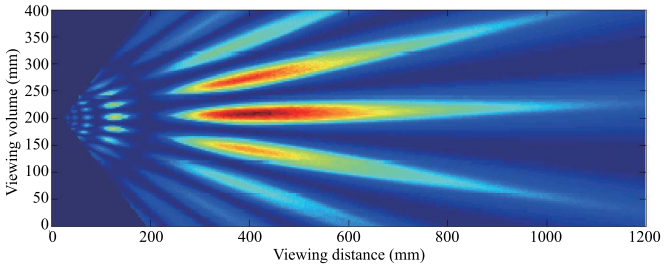



Fig. 11. Measured system crosstalk of the auto-stereoscopic display using ELDIM VCMaster3D.

## IV. Experiment and Analysis

### A. Subjective Experiment

We conducted a subjective experiment to collect data descriptive of the relationship between a viewer's perception of crosstalk and the spatial characteristics of S3D images.

*1) Test Environment:* A $10.1''$ auto-stereoscopic display was used in the subjective experiment, with display resolution $1280 \times 800$. The auto-stereoscopy principle used in the display is the parallax barrier. Using an ELDIM VCMaster3D [38] measurement tool, the degree of system crosstalk in the display was measured as a function of distance and angle, as shown in Fig. 11.

The viewing distance of the subjects was about $400mm$ from the center of the display. The system crosstalk level under this condition was about 3% ($p = 0.03$). To guarantee a consistent crosstalk level during each session, each subject was continuously monitored during the subjective experiment

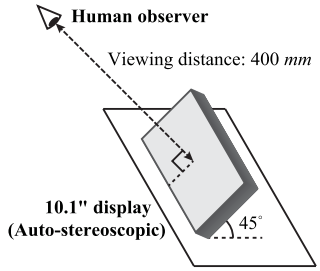



Fig. 12. Viewing environment used in the subjective experiments.

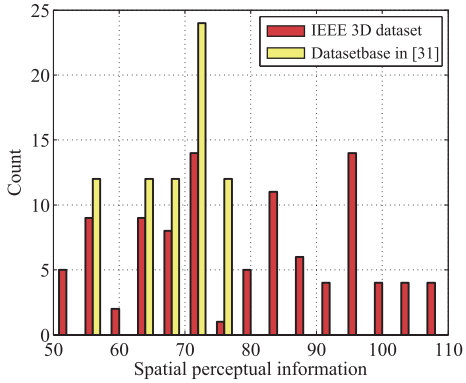



Fig. 13. Spatial information distribution in the two datasets. The yellow bars (Database [39]) are distributed in a narrower range than the red bars (IEEE 3D Database [37]).

to ensure that he or she maintained a proper viewing position. The overall viewing environment is described by Fig. 12. The auto-stereoscopic display was tilted backwards by about 45°, so that the subjects' eyes were approximately perpendicular to the center of the screen while in the sitting position.

*2) Test Stimuli:* We utilized two kinds of datasets in the experiments. The first dataset is drawn from the IEEE 3D database [37], while the second is drawn from the stereoscopic quality dataset described in [39]. Since the second dataset already provides mean opinion score (MOS) data, the subjective test was only conducted on the first dataset.

Twenty five color sequences were chosen from the IEEE 3D database [37], which contains stereo images captured using a twin-lens 3D camcorder (Panasonic AG-3DA1). The original images are of full HD resolution, but were down-scaled to $1280 \times 720$ to match the resolution of the display. For each of the 25 scenes, four different disparity settings were used by adjusting the zero parallax distance setting of the stereo camera (the zero parallax distances were set to $5.89m$, $9.58m$, $13.26m$, and $16.90m$). The database contains both indoor and outdoor scenes.

From the second database, we utilized only the category of images labeled *Crosstalk Stereoscopic* [39]. The dataset includes 6 different scene contents of resolution of $1280 \times 960$. For each scene, three different camera baseline distances were used. Four different crosstalk levels of 0%, 5%, 10% and 15% were used to simulate virtual crosstalk distortion in images. Therefore, the total number of stimuli is 72.

To compare the spatial characteristics of the two datasets, we analyzed the spatial information (SI) described in ITU recommendation P.910 [40]. The SI is a measure of the amount of spatial detail in an image. Fig. 13 shows the distribution of SIs in the two datasets. In the IEEE 3D database, the distribution of SI ranges from 50 to 110. By contrast, Xing's dataset ranges from 55 to 75. This suggests that the IEEE 3D database contains more complicated scenes than does Xing's dataset. However, Xing's dataset contains a greater diversity of crosstalk levels and camera baselines. Therefore, we viewed the IEEE 3D database as having greater value for investigating the effects of spatial characteristics, while the dataset [39] is more useful for evaluating the effect of diverse crosstalk levels on the perception of crosstalk.

*3) Test Procedure:* 25 subjects participated in the subjective test. The participants' ages ranged between 21 and 31. All had normal or corrected visual acuities of no less than 1 as measured using the Landolt C-test. All had stereo acuities of 60 arcseconds or better as measured with the Randot stereo test.

We controlled the viewing conditions in accordance with ITU-R BT.2022 [41]. The luminance levels of the viewing environment were measured using a KONICA MINOLTA CA-310. The ratio of the luminance of the inactive screen to its peak luminance was less than 0.02. The ratio of the luminance of the black screen to peak white was about 0.01. The ratio of the luminance of the background behind the display to the peak luminance of the screen was about 0.15.

The single-stimulus method recommended in [42]–[44] was employed as the test methodology. Each stimulus was shown to the subjects for 10 seconds with a 5 second interval between them. During each interval, a gray screen was displayed while the participants recorded a subjective score on the previous stimulus. Participants were asked to rate the annoyance level of perceived crosstalk when they viewed each image using a Likert scale: 5 = imperceptible (cannot see any crosstalk), 4 = perceptible but not annoying, 3 = slightly annoying, 2 = annoying, 1 = very annoying.

We divided the subjective experiment into three sessions: one training session and two testing sessions. In the training session, each participant was instructed with regards to the process of the test methodology and familiarized with the general range of crosstalk annoyance levels by showing them 10 exemplar stimuli. For the training samples, we selected a variety of diverse scenes and disparity settings representing a reasonable range of rating levels. Using an objective crosstalk measure (the summed absolute difference between the left and right images), and the 3D experts subjective opinions, the training samples were selected to adequately encompass the rating scales. In each testing session, 50 randomly shuffled S3D image pairs were evaluated. The stimuli were randomly shuffled once, and the order was fixed over all the subjects. The testing sessions were separated by a break period of 10 minutes.

### *B. Experimental Results and Statistical Analysis*

We evaluated the performance of the proposed algorithm using three different standard measures [45]; the Spearman

TABLE II

PERFORMANCE COMPARISON OF THE CROSSTALK PREDICTION MODEL ON THE IEEE 3D DATABASE [37] FOR DIFFERENT COMBINATIONS OF MODULES

| | Metrics | SROCC | LCC | RMSE | OR |
|---|---|---|---|---|---|
| | Diff | 0.556 | 0.510 | 0.796 | 0.410 |
| Left only | Diff + CM | 0.812 | 0.752 | 0.574 | 0.360 |
| | Diff + CM + DSM | 0.846 | 0.852 | 0.491 | 0.310 |
| Right only | Diff + CM | 0.801 | 0.743 | 0.619 | 0.340 |
| | Diff + CM + DSM | 0.821 | 0.815 | 0.534 | 0.240 |
| Bino sum | Diff + CM | 0.823 | 0.758 | 0.546 | 0.260 |
| | Diff + CM + DSM | 0.850 | 0.858 | 0.476 | 0.190 |

rank order correlation coefficient (SROCC), the Pearson linear correlation coefficient (LCC) and the root mean square error (RMSE). To evaluate the LCC and RMSE, nonlinear regression using the logistic function was used to regress BPCP on the MOS, following [46]. The predicted MOS ($MOS_p$) is then

$$MOS_p = \frac{b_1}{1 + \exp\left(-b_2\left(BPCP - b_3\right)\right)}, \tag{25}$$

where $b_1$, $b_2$ and $b_3$ are the regression parameters.

As shown earlier in Fig. 4, the proposed crosstalk predictor consists of independent monocular stages for each view and a binocular stage which integrates the outputs of the monocular stages. The effect of each of these stages was evaluated using the IEEE 3D database. In Table II, the Bino sum (binocular summation) is the result of the BPCP model, including binocular summation. Left only and Right only indicate the results using only monocular processing on each respective view. From the correlation scores, it may be observed that binocular processing achieved a slightly higher accuracy than when only the left image was considered. On the other hand, Bino sum significantly outperformed versus the situation where only the right image was used. This strongly suggests that binocular processing yields an advantage in accuracy relative to monocular processing. In addition, to identify the relative contributions of the contrast masking (CM) and duplicated structure masking (DSM) units, we measured the performance of BPCP when only CM is included as compared to including both CM and DSM. As shown in Table II, the highest correlation scores are attained when both CM and DSM are used.

To capture the relative contributions to performance of CM and DSM, we deployed weights $\alpha$ and $\beta$ to balance the contrast and duplicated structure masking terms in (17). We studied the variation in performance using various combinations of $\alpha$ and $\beta$ ranging from 0 to 1, where $\alpha+\beta = 1$. Also, we examined each combination using five different Minkowski exponents $\gamma$ ranging from 1 to 5. A total of 320 combinations were tested.

The results are shown in Fig. 17. Fig. 17(a) is the result obtained on the IEEE 3D database. With regard to the effect of the Minkowski exponent on perceptual crosstalk prediction, the performances were similar except when $\gamma = 1$. The highest performance was achieved for $\gamma = 3$ or 4. When

TABLE III

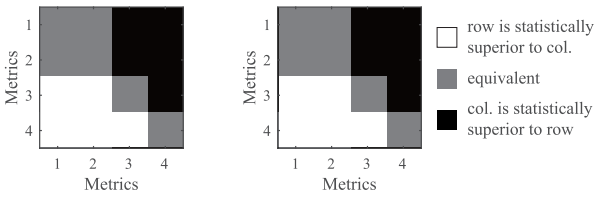
PERFORMANCE COMPARISON ON THE IEEE 3D DATABASE [37]

| Metrics | SROCC | LCC | RMSE | OR |
|---|---|---|---|---|
| PSNR | 0.556 | 0.510 | 0.796 | 0.410 |
| SSIM | 0.566 | 0.523 | 0.789 | 0.430 |
| $V_{pdep}$ [14] | 0.645 | 0.482 | 0.784 | 0.420 |
| **BPCP** | 0.850 | 0.858 | 0.476 | 0.190 |

TABLE IV

PERFORMANCE COMPARISON ON THE DATASET [39]

| Metrics | SROCC | LCC | RMSE | OR |
|---|---|---|---|---|
| PSNR | 0.763 | 0.822 | 0.463 | 0.287 |
| SSIM | 0.760 | 0.792 | 0.497 | 0.347 |
| $V_{pdep}$ [14] | 0.859 | 0.884 | 0.382 | 0.306 |
| **BPCP** | 0.923 | 0.888 | 0.376 | 0.306 |




(a) IEEE 3D database [37] (b) Dataset [39]

Fig. 14. Results of F-test performed on the model residuals between predicted scores and MOS values at the significant level of 95%. The metric numbers refer to each metric: PSNR (1), SSIM (2), $V_{pdep}$ (3), and BPCP (4). (a) IEEE 3D database [37]. (b) Dataset [39].

$\alpha = 0.4$ or 0.5, the SROCC was approximately maximized. The results on the database [39] follows the same trends as those on the IEEE 3D database. The correlation was maximized near $\alpha = \beta = 0.5$. These observations suggest that the proposed crosstalk prediction model BPCP does not have database-determined performance and that the two modes of masking are about equally important.

To compare performance against other algorithms, PSNR, SSIM [15] and Xing's algorithm [14] were also tested on the same datasets. We used the default parameters for SSIM, where the dynamic range was set to $255 \cdot (1 + p)$ (hence, $c1 = 6.90$ and $c_2 = 62.09$). Four different measures were used to evaluate the performances [45]; the Spearman rank order correlation coefficient (SROCC), the Pearson linear correlation coefficient (LCC), the root mean square error (RMSE) and the outlier ratio (OR). Samples that exceeded the 95% significance level of MOS were deemed to be outliers. For PSNR and SSIM, the expected and observed images from (1) and (2) were used as reference and test images respectively. To evaluate the models on dataset [39], we adaptively modified the crosstalk level as a function of the simulated crosstalk level. The crosstalk level was chosen as the summation of the simulated crosstalk level and the system crosstalk level. Tables III and IV show results on the IEEE 3D database and on dataset [39] respectively. On both datests, BPCP delivered the best score among the four algorithms. Xing's algorithm,

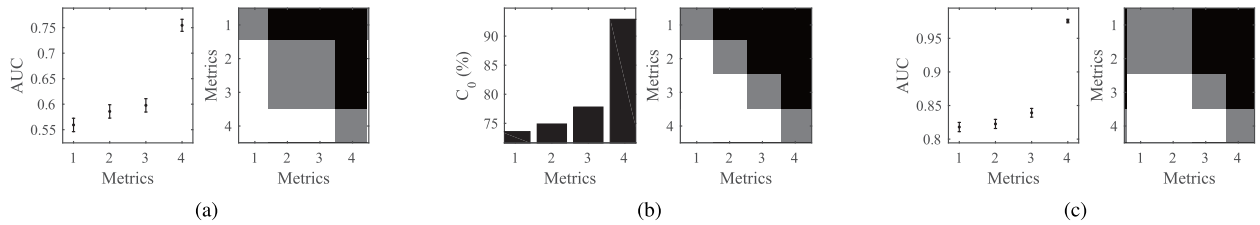



Fig. 15. Results and statistical analysis on the IEEE 3D database [37]. The metric numbers refer to each metric: PSNR (1), SSIM (2), $V_{pdep}$ (3), and BPCP (4). (a) Different vs. similar (AUC). (b) Better vs. worse ($C_0$). (c) Better vs. worse (AUC).

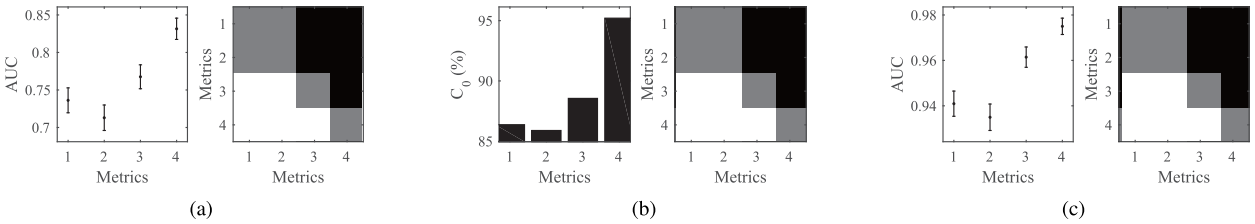



Fig. 16. Results and statistical analysis on the dataset [39]. The metric numbers refer to each metric: PSNR (1), SSIM (2), $V_{pdep}$ (3), and BPCP (4). (a) Different vs. similar (AUC). (b) Better vs. worse ($C_0$). (c) Better vs. worse (AUC).

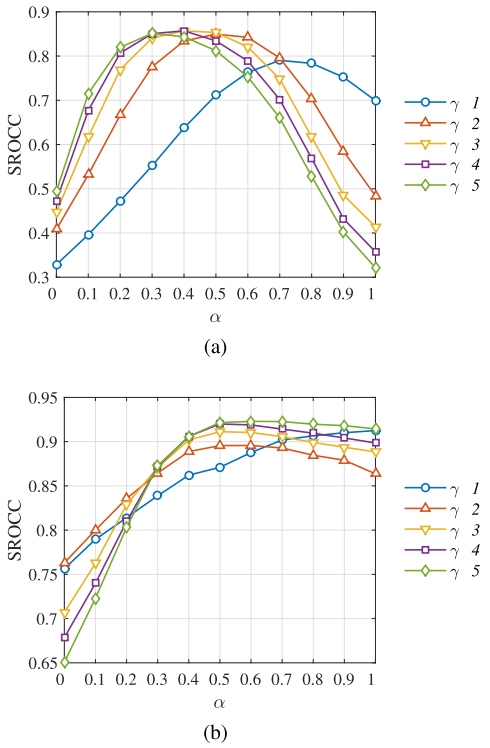



Fig. 17. SROCC variation with different weights and $\gamma$ . The x-axis represents the weight of $\alpha$ where $\beta = 1.0 - \alpha$. (a) SROCC variation (IEEE 3D database [37]). (b) SROCC variation (Database in [39]).

$V_{pdep}$ was ranked next. Scatter plots for each model when applied on the IEEE 3D database and dataset [39] are shown in Figs. 18 and 19. The black dotted line is the fitting curve from (25). Data points corresponding to the same scene are marked by the same color. In Fig. 19, each mark has three connected data points, and each line refers to the same crosstalk level but with different disparity settings. It is obvious that the clusters with lower MOS indicate higher crosstalk levels. In Figs. 18(b) and 19(b), it is interesting that the SSIM values are very high (more than 0.9). This is because the system crosstalk was set to be very small, and the original structure was not noticeably distorted due to crosstalk.

To analyze the statistical significance of the relative performances of our proposed model and the other algorithms, we conducted F-tests on the errors between the predicted scores and the MOS on the IEEE 3D database and on Xing's dataset. The F-test was used to compare one model against another model at the 95% significance level, as shown in Table 14. White (or black) box indicates that the performance of the model in the row was statistically superior (inferior) to that in the column, while gray box indicates equivalence of the models. The results confirm the superior performance of BPCP with statistical significance against other models on both datasets.

We further conducted a recently proposed statistical test whereby the classification capabilities of the models were determined in two scenarios [47]. First, the pairs in the dataset were divided into two groups: significantly different and similar. The detailed process is described in [47]. In the first scenario, we determined how well the model could distinguish between the two groups. In the second scenario, we determined whether the model was able to correctly recognize the stimulus of higher quality in the pair. The results of these two analyses, viz., the areas under the curves (AUCs) and the percentage of correct classification ($C_0$) with statistical significance are respectively shown in Figs. 15 and 16. The meanings of the colors in the significance plots are same as in Fig. 14. It may be seen that the BPCP model significantly outperformed the other metrics in all of the analyses on both datasets.

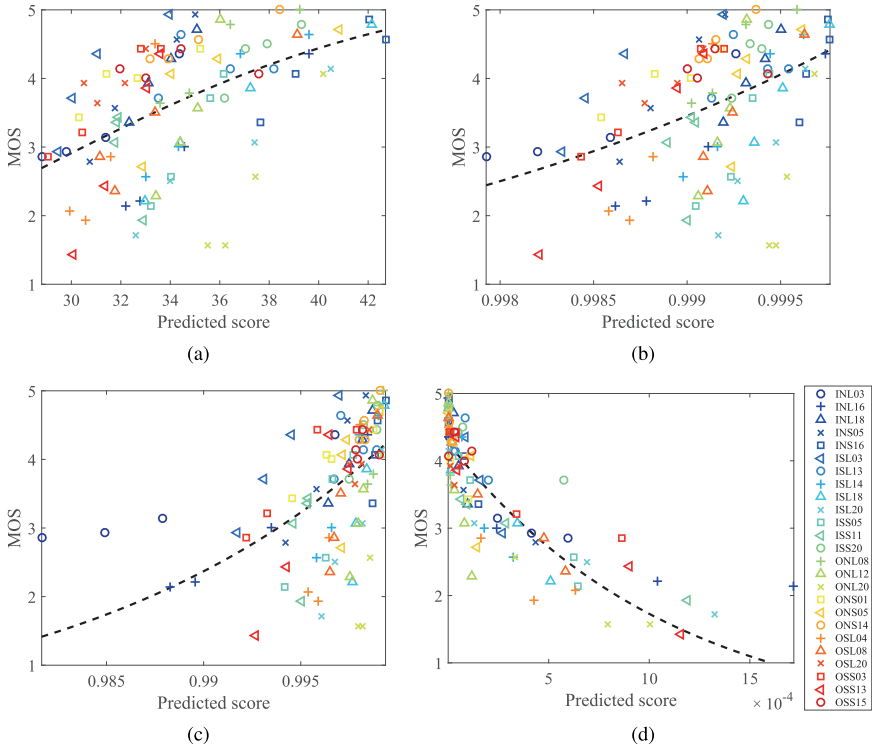



Fig. 18. Scatter plots on the IEEE 3D database [37]. (a) PSNR. (b) SSIM. (c) $V_{pdep}$. (d) Proposed BPCP.

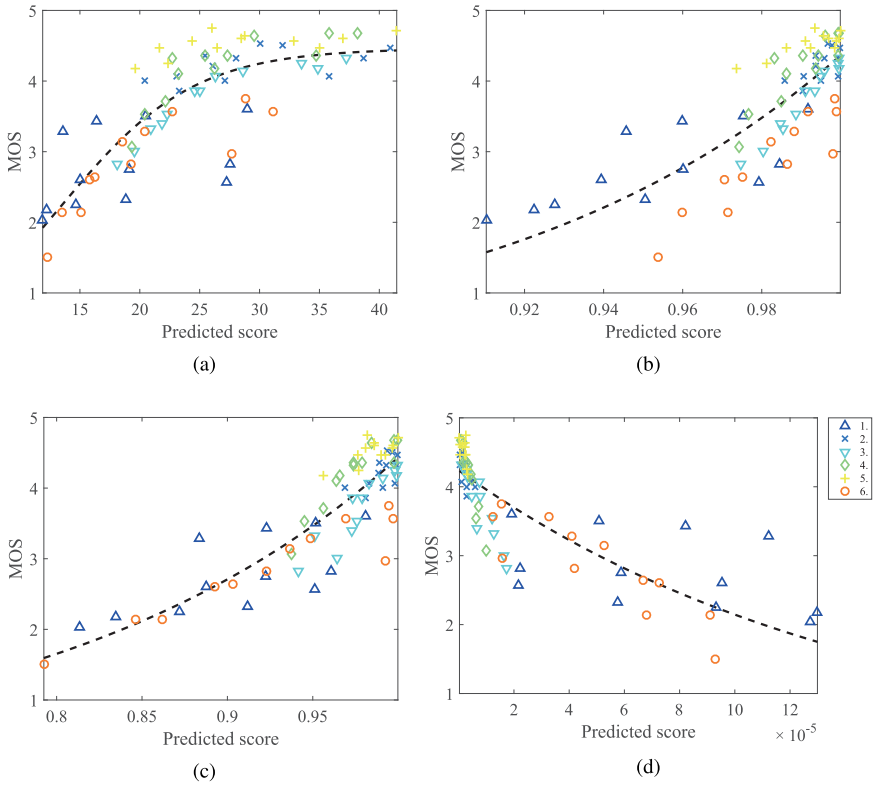



Fig. 19. Scatter plots on the dataset [39]. (a) PSNR. (b) SSIM. (c) $V_{pdep}$. (d) Proposed BPCP.

## V. Conclusion

In the future, we plan to extend the proposed framework to be able to apply to multi-view auto-stereoscopic displays. Multi-view auto-stereoscopic displays can provide a motion parallax experience as well as stereo parallax, allowing viewers to receive an ever more immersive and natural 3D experience. However, crosstalk is inevitable in

slanted lenticular multi-view systems, since a viewer will usually perceive three or more adjacent views simultaneously. Modeling perceptual quality in multi-view auto-stereoscopic displays is challenging, since the image source in the display provides multiple views. All possible pairs of stereo views should be considered to decide the overall perceived quality. Crosstalk distortion is a significant factor since it can occur during transitions between the different views as the viewer moves his head. Modeling viewer experience in multi-view auto-stereoscopic displays would be valuable since the technology continues to advance and perceptual quality of these images have not yet been studied in depth. In addition, we plan to use a Wheatstone stereoscope arrangement to enable zero system crosstalk. By applying various levels of simulated system crosstalk on images, we plan to further determine the relationship between perceptual crosstalk and system crosstalk. In addition, applying a deep learning technique would increase the prediction accuracy further [48], [49].

## REFERENCES


[1] I. Tsirlin, L. M. Wilcox, and R. S. Allison, "The effect of crosstalk on the perceived depth from disparity and monocular occlusions," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 445–453, Jun. 2011.

[2] K. C. Huang, C.-H. Tsai, K. Lee, and W.-J. Hsueh, "Measurement of contrast ratios for 3D display," *Proc. SPIE*, vol. 4080, pp. 78–86, Jun. 2000.

[3] J.-C. Liou, K. Lee, F.-G. Tseng, J.-F. Huang, W.-T. Yen, and W.-L. Hsu, "Shutter glasses stereo LCD with a dynamic backlight," *Proc. SPIE*, vol. 7237, p. 72370X, Jan. 2009.

[4] S. Pala, R. Stevens, and P. Surman, "Optical cross-talk and visual comfort of a stereoscopic display used in a real-time application," *Proc. SPIE*, vol. 6490, pp. 649011-1–649011-12, Mar. 2007.

[5] P. Boher, T. Leroux, V. C. Patton, T. Bignon, and D. Glinel, "A common approach to characterizing autostereoscopic and polarization-based 3-D displays," *J. Soc. Inf. Display*, vol. 18, no. 4, pp. 293–300, 2010.

[6] N. A. Dodgson, "Analysis of the viewing zone of multiview autostereoscopic displays," *Proc. SPIE*, vol. 4660, pp. 254–265, May 2002.

[7] D. J. Montgomery, G. J. Woodgate, A. M. S. Jacobs, J. Harrold, and D. Ezra, "Performance of a flat-panel display system convertible between 2D and autostereoscopic 3D modes," *Proc. SPIE*, vol. 4297, pp. 148–159, Jun. 2001.

[8] A. Woods, "Understanding crosstalk in stereoscopic displays," in *Proc. D, Syst. Appl. Conf. (DSA)*, 2010, pp. 19–21.

[9] L. Wang *et al.*, "Crosstalk evaluation in stereoscopic displays," *IEEE J. Display Technol.*, vol. 7, no. 4, pp. 208–214, Apr. 2011.

[10] S. Pastoor, "Human factors of 3D imaging: Results of recent research at Heinrich-Hertz-Institut Berlin," in *Proc. IDW*, vol. 95. 1995, pp. 69–72.

[11] P. J. H. Seuntiëns, L. M. J. Meesters, and W. A. Ijsselsteijn, "Perceptual attributes of crosstalk in 3D images," *Displays*, vol. 26, nos. 4–5, pp. 177–183, 2005.

[12] L. Lipton, "Factors affecting 'Ghosting' in time-multiplexed piano-stereoscopic crt display systems," *Proc. SPIE*, vol. 0761, pp. 75–78, Jun. 1987.

[13] P.-C. Wang, S.-L. Hwang, H.-Y. Huang, and C.-F. Chuang, "System cross-talk and three-dimensional cue issues in autostereoscopic displays," *J. Electron. Imag.*, vol. 22, no. 1, p. 013032, 2013.

[14] T. Ebrahimi, L. Xing, J. You, and A. Perkis, "Assessment of stereoscopic crosstalk perception," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 326–337, Apr. 2012.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[16] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer.*, vol. 70, no. 12, pp. 1458–1471, Dec. 1980.

[17] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[18] K. Lee and S. Lee, "3D perception based quality pooling: Stereopsis, binocular rivalry, and binocular suppression," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 3, pp. 533–545, Apr. 2015.

[19] J. Ding and G. Sperling, "A gain-control theory of binocular combination," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 4, pp. 1141–1146, 2006.

[20] A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms: What does the future hold?" *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 675–696, 2011.

[21] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[22] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[23] S. J. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, pp. 2–15, Jun. 1992.

[24] M. Clark and A. C. Bovik, "Experiments in segmenting texton patterns using localized spatial filters," *Pattern Recognit.*, vol. 22, no. 6, pp. 707–717, 1989.

[25] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[26] A. B. Watson and A. J. Ahumada, "A standard model for foveal detection of spatial contrast," *J. Vis.*, vol. 5, no. 9, p. 6, 2005.

[27] T. Carney *et al.*, "Modelfest: Year one results and plans for future years," *Proc. SPIE*, vol. 3959, pp. 140–151, Jun. 2000.

[28] F. W. Campbell, J. J. Kulikowski, and J. Levinson, "The effect of orientation on the visual resolution of gratings," *J. Physiol.*, vol. 187, no. 2, pp. 427–436, 1966.

[29] M. J. McMahon and D. I. A. MacLeod, "The origin of the oblique effect examined with pattern adaptation and masking," *J. Vis.*, vol. 3, no. 3, p. 4, 2003.

[30] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer. A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[31] X. Fei, L. Xiao, Y. Sun, and Z. Wei, "Perceptual image quality assessment based on structural similarity and visual masking," *Signal Process. Image Commun.*, vol. 27, no. 7, pp. 772–783, 2012.

[32] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[33] H. Kim, S. Lee, and A. C. Bovik, "Saliency prediction on stereoscopic videos," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1476–1490, Apr. 2014.

[34] J. Park, S. Lee, and A. C. Bovik, "3D visual discomfort prediction: Vergence, foveation, and the physiological optics of accommodation," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 3, pp. 415–427, Jun. 2014.

[35] H. Oh, S. Lee, and A. C. Bovik, "Stereoscopic 3D visual discomfort prediction: A dynamic accommodation and vergence interaction model," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 615–629, Feb. 2016.

[36] D. H. Baker and E. W. Graf, "Natural images dominate in binocular rivalry," *Proc. PNAS*, vol. 106, no. 13, pp. 5436–5441, 2009.

[37] (2015). *IEEE-SA Stereoscopic (3D Imaging) Database*. [Online]. Available: http://grouper.ieee.org/groups/3dhf

[38] T. Leroux, P. Boher, T. Bignon, D. Glinel, and S. I. Uehara, "VCMaster3D: A new Fourier optics viewing angle instrument for characterization of autostereoscopic 3D displays," in *SID Symp. Dig. Tech. Papers*, vol. 40. 2009, pp. 115–118.

[39] L. Xing, J. You, T. Ebrahimi, and A. Perkis, "Stereoscopic quality datasets under various test conditions," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, 2013, pp. 136–141.

[40] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T P910, 1999.

[41] *General Viewing Conditions for Subjective Assessment of Quality of SDTV and HDTV Television Pictures on Flat Panel Displays*, document ITU-R BT2022, 2012.

[42] *IEEE Standard for Quality of Experience (QoE) and Visual-Comfort Assessments of Three-Dimensional (3D) Contents Based on Psychophysical Studies*, IEEE Standard 333311-2015, pp. 1–46, 2015.

[43] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT500-13, 2012.

[44] *Subjective Methods for the Assessment of Stereoscopic 3DTV Systems*, document ITU-R BT2021-1, 2015.

[45] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document ITU-T P1401, 2012.
[46] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II*, Video Quality Experts Group (VQEG), Sweden, 2003.
[47] L. Krasula, K. Fliegel, P. Le Callet, and M. Klima, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Proc. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.
[48] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 1, pp. 206–220, Jan. 2017.
[49] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

**Jongyoo Kim** received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree. His research interests include 2D/3D image and video processing based on human visual system, quality assessment of 2D/3D image and video, 3D computer vision, and deep learning.

**Taewan Kim** received the B.S., M.S., and Ph.D. degrees from Yonsei University, Seoul, South Korea, in 2008, 2010, and 2015, respectively, all in electrical and electronic engineering. He is currently with the Video Technology Laboratory, SK Telecom, Seoul. His research interests include quality assessment of 2D and 3D image and video, 3D video coding, cross-layer optimization, and wireless multimedia communications. He has participated in the IEEE Standard Working Group for 3D quality assessment (IEEE P3333.1). He received the Samsung Humantech Thesis Prize in 2013.

**Sanghoon Lee** (M'05–SM'12) received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 1989, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology in 1991, and the Ph.D. degree in electrical engineering from The University of Texas at Austin in 2000. From 1991 to 1996, he was with Korea Telecom. From 1999 to 2002, he was with Lucent Technologies on 3G wireless and multimedia networks. In 2003, he joined the Faculty of the Department of Electrical and Electronics Engineering, Yonsei University, where he is currently a Full Professor. His research interests include image/video quality assessment, computer vision, graphics, cloud computing, and multimedia communications and wireless networks. He received the 2015 Yonsei Academic Award from Yonsei University, the 2012 Special Service Award from the IEEE Broadcast Technology Society, and the 2013 Special Service Award from the IEEE Signal Processing Society. He has been serving on the Technical Committee of the IEEE Multimedia Signal Processing since 2016 and the IEEE IVMSP Technical Committee since 2014. He was a Technical Program Co-Chair of the International Conference on Information Networking in 2014 and the Global 3D Forum in 2012 and 2013. He was the General Chair of the 2013 IEEE IVMSP Workshop. He has been the Chair of the IEEE P3333.1 Quality Assessment Working Group since 2011. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2010 to 2014. He served as a special issue Guest Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING in 2013, and an Editor of the *Journal of Communications and Networks* from 2009 to 2015. He has been an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS since 2014 and the *Journal of Electronic Imaging* since 2015.

**Alan Conrad Bovik** (F'96) is currently the Curry/Cullen Trust Endowed Chair Professor with The University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering. He is also a Faculty Member with the Department of Electrical and Computer Engineering and the Center for Perceptual Systems, Institute for Neuroscience. He has authored over 650 technical articles in these areas and holds two U.S. patents. His several books include the recent companion volumes *The Essential Guides to Image and Video Processing* (Academic Press, 2009). His research interests include image and video processing, computational vision, and visual perception. He has received a number of major awards from the IEEE Signal Processing Society, including the Best Paper Award (2009), the Education Award (2007), the Technical Achievement Award (2005), and the Meritorious Service Award (1998). He was a recipient of the IS&T/Society of Photo-Optical and Instrumentation Engineers (SPIE) Imaging Scientist of the Year in 2011, the SPIE Technology Achievement Award in 2012, and the Honorary Member Award of the Society for Imaging Science and Technology in 2013. He received the Hocott Award for the Distinguished Engineering Research from The University of Texas at Austin, the Distinguished Alumni Award from the University of Illinois at Champaign–Urbana (2008), the IEEE Third Millennium Medal (2000), and two journal paper awards from the international Pattern Recognition Society (1988 and 1993). He is a fellow of the Optical Society of America, SPIE, and the American Institute of Medical and Biomedical Engineering. He was involved in numerous professional society activities, including Board of Governors and the IEEE Signal Processing Society from 1996 to 1998. He was a Founding General Chairman of the First IEEE International Conference on Image Processing, held in Austin, TX, in 1994. He is a registered Professional Engineer in the State of Texas and a frequent Consultant to legal, industrial, and academic Institutions. He was a Co-Founder and an Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1996 to 2002. He served on the Editorial Board of the Proceedings of the IEEE from 1998 to 2004. He has been a Series Editor of *Image, Video, and Multimedia Processing*, Morgan and Claypool Publishing Company, since 2003.