Optimal Region Selection for Stereoscopic Video Subtitle Insertion

Guanghui Yue, Chunping Hou, Jianjun Lei, *Member, IEEE*, Yuming Fang, *Senior Member, IEEE*, and Weisi Lin, *Fellow, IEEE*

***Abstract*— Stereoscopic subtitle insertion is a fundamental and essential element in stereoscopic film and TV industry. However, little work has been dedicated to the optimal region selection for stereoscopic subtitle insertion. In addition, there is no public database reported for the performance evaluation of it. In this paper, we build the first large-scale video database (TJU3D) for stereoscopic video subtitle insertion, which includes 50 video sequences with rich screen scenes. Compared with 2D subtitle region selection, there are several problems we have to consider in stereoscopic subtitle region selection: 1) the subtitle should avoid depth cue collision and occlusion from objects in stereoscopic video sequences; 2) the disparity value of the subtitle must be minimized to reduce visual discomfort; and 3) the temporal coherence constraint must be considered during region selection for subtitles in video sequences. By considering these constraints, we propose an optimal region selection algorithm for stereoscopic subtitle insertion. First, we compute the disparity map of each video frame in video sequences. For each frame, the optimal position and disparity value of the subtitle are determined by a subtitle region selection algorithm, which contains two parts (i.e., the coarse selection and fine selection). After that, by considering the temporal consistency between adjacent frames, the position and disparity value of each frame are further classified and processed in order to avoid the subtitle jitter. We evaluate the proposed method on TJU3D video database through two visual discomfort prediction metrics and one subjective experiment. To further verify the effectiveness of the proposed method, we also validate the performance of the proposed method on video comfort assessment database, i.e., IEEE-SA Stereo Database. Experimental results demonstrate that the visual discomfort is greatly reduced when using the proposed method compared with the basic method.**

***Index Terms*— Stereoscopic subtitle, visual discomfort, subtitle insertion, subjective evaluation.**

I. Introduction

THREE dimensional (3D) stereoscopic multimedia services greatly increase the immersive experience by enabling depth perception. During recent years, 3D display technology has received much attention and been widely used in various areas [1]–[3]. Meanwhile, safety and health problems occur during 3D content viewing for some observers, e.g., eyestrain, nausea, and headache. The visual discomfort from 3D stereoscopic images can be caused by various factors, such as the excessive disparity, accommodation and vergence (AV) conflict, windows violation, geometrical distortions, and rapid disparity changes [4]–[9].

Since any content in stereoscopic video sequences might induce visual discomfort, we have to consider the visual discomfort from subtitle insertion in stereoscopic video sequences. In video sequences, the subtitles serve as linguistic approximation of visual content. They are used for observers to understand the video content better with additional information or a textural translation of the audio track [10]. However, the characteristics of subtitles, including character size, font and position, have great influence on the quality of experience (QoE) during video sequence viewing [11], [12]. Currently, researchers have begun to investigate the influence of subtitle's characteristic on QoE. Yeh and Lee [13] explored the influence of the character size, font and position on QoE in watching 2D videos. Davoudi *et al*. [14] proposed a fuzzy system to improve the QoE by coloring the subtitle. Compared with 2D video, more attention should be paid during the subtitle insertion for 3D stereoscopic video. On the one hand, there is depth perception for the objects in 3D visual scenes, and thus, the subtitle should avoid depth cue collision and occlusion. On the other hand, the distance between subtitle and screen plane requires to be minimized to reduce visual discomfort. Lambooij *et al*. [12] explored the impact of subtitles on visual discomfort of 3D video sequence. The subjective experimental results in that study demonstrate that the subtitles regarded as the secondary salient region by viewers can reduce the visual discomfort. Wan *et al*. [15] conducted a subjective experiment to explore the approximation depth value for subtitle insertion in 3D video sequence. The experimental results in that study reveal that the disparity of subtitles should not be larger than the disparity of the nearest object to the viewer by two pixels in the 3D visual scene for a good viewing experience.

Although researchers have begun to realize the significance of the 3D subtitles in stereoscopic display. However, these attempts are far from sufficient to construct perfect stereoscopic display systems. Horizontal disparity is regarded as the main factor that causes visual discomfort as well as the stereo vision. Hence, from the perspective of discomfort reduction,

Manuscript received November 24, 2016; revised March 5, 2017 and June 24, 2017; accepted August 3, 2017. Date of publication August 15, 2017; date of current version November 5, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61471262, Grant 61520106002, and Grant 61571212. This paper was recommended by Associate Editor D. Mukherjee. *(Corresponding author: Yuming Fang.)*

G. Yue, C. Hou, and J. Lei are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yueguanghui@tju.edu.cn; hcp@tju.edu.cn; jjlei@tju.edu.cn).

Y. Fang is with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330032, China (e-mail: fa0001ng@e.ntu.edu.sg).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

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/TCSVT.2017.2739756

3D subtitles should be inserted in a relatively comfortable region. Currently, subtitles are always placed in the front of the foremost objects to obtain exciting visual experience. Unfortunately, it may cause visual discomfort for observers. Besides, manual insertion costs too much time and is not conductive to industrialized production. Thus, it is much desired to design an effective algorithm for optimal region selection to reduce visual discomfort.

In this paper, we propose an optimal region selection algorithm for 3D subtitles by considering visual discomfort. First, we optimize the position and disparity value of subtitle for each frame in video sequences by an automatic subtitle region selection method. Then, considering the temporal consistency between adjacent frames, the position and disparity value of each frame are further classified and processed to avoid the subtitle jitter. The visual discomfort degree of inserted subtitle is evaluated by two objective visual discomfort assessment schemes. Furthermore, we have conducted a subjective test to validate the performance of the proposed method. Experimental results show the promising performance of the proposed method. In sum, the major contributions of this study are listed as follows.

1) We build a 3D video database (namely TJU3D Video Database), composed of 50 video sequences with various visual content, for optimal region selection of stereoscopic subtitles. In this database, each video sequence contains both negative and positive disparities. In particular, the disparities of most video sequences are beyond comfort threshold to study where to put the subtitles in these cases. We also conduct three subjective experiments to explore the influence of subtitle positions on viewing experience. To the best of our knowledge, this is the first database for performance evaluation of 3D subtitle insertion in this research area.
2) We propose an automatic subtitle region selection algorithm for subtitle insertion in stereoscopic video sequences. First, the operations of coarse selection and fine selection are implemented for each video frame. In addition, the subtitle positions are further classified and the disparity values are processed to avoid jitter with the consideration of temporal consistency.
3) We apply two different visual discomfort assessment schemes, i.e., Neural 3D-VDP [16] and comfort function [17], as well as a subjective experiment to evaluate the visual discomfort of inserted subtitles. Through comprehensive validations, we show that the optimal region selection algorithm for subtitle insertion indeed reduces the visual discomfort. The proposed method provides a reference for stereoscopic film or television program production.

The rest of this paper is organized as follows: Section II introduces the background and analysis for the research work. Section III presents the subjective experiments in details. Section IV shows the proposed optimal region selection algorithm for subtitle insertion in stereoscopic video sequences. The experimental results and analysis of the proposed method are described in section V. The final section provides the concluding remarks for this research work.

II. Related Work

This section contains two subsections. In the first subsection, we give a brief review of recent works in region/position selection for 3D subtitles. The problems that should be considered during the region selection are highlighted and the limitations of previous works are also analyzed. In the second subsection, we describe the visual discomfort assessment methods and their feasibility of applications in the evaluation of subtitle discomfort.

A. Region/Position Selection for 3D Subtitle Insertion

In the literature, there is little investigation into region/position selection for 3D subtitle insertion. Wan *et al*. [15] conducted a subjective experiment to explore the optimal depth distance between the foremost object and subtitle. Sixty 3D subtitled sequences (15 videos $\times$ 4 subtitles) without in-depth motion were used in the experiment. Twenty-five observers participated in the experiment and marked the score from 1 to 5, corresponding bad depth position to excellent depth position. The experimental results reflect that subjective score decreases as the increment of disparity range in depth direction. Meanwhile, the subtitle disparity should be shorter than two pixels for better viewing experience. Lambooij *et al*. [12] selected two 3D video sequences to study the influence of subtitle on visual discomfort. Forty participants were involved in the subjective experiments. The experimental results show that, when the depth discontinuities exist in video sequences, the subtitle could be treated as the second salient region. Human beings are more interested in salient object [18], [19]. As much attention is paid to the subtitle, the visual discomfort caused by scene change reduces.

In [15], the disparity range of the selected 3D video sequences was produced as small as possible within the comfort zone. Meanwhile, there is no depth motion in all the used video sequences, which is inconsistent with the natural scenes. However, in order to pursue a better visual effect, the movie producer always sets the objects outside the screen. If subtitles are placed in the front of the foremost object, their disparities may exceed the comfort threshold. As a result, it may cause visual discomfort as well as occlusion of object of interest. Taking Fig. 1 (a) as an example, the disparity of bottom subtitle region (marked by red box) is $-2.12°$. If subtitles are placed in the front of the foremost object, the disparity of subtitles is larger than $-2.12°$ which leads to serious visual discomfort. In addition, if only a portion of the disparity map is beyond the discomfort threshold in all subtitle region, the visual discomfort would also be induced by using Wan's metric [15]. Inspired by the fact that the subtitle region can be regarded as the second salient region, the subtitle should be presented on relatively comfortable region. When the subtitle is placed on the comfort region where the disparity value is in the comfort zone, the visual discomfort would greatly reduce, as shown in Fig. 1 (b). Besides, the subtitle can also be presented on the top subtitle region in the case that there is no suitable region in the bottom subtitle region, as shown in Fig. 1 (c).

(a) (b) (c)

Fig. 1. Examples for subtitle insertion: (a) on the bottom with large disparity value; (b) on the comfort region with small disparity value; (c) on the top with small disparity value.

As analyzed above, the discomfort assessment of 3D subtitles is still limited to place the subtitle on the bottom of screen. Moreover, there is no study investigating to automatically estimate the optimal region for 3D subtitle insertion in stereoscopic video sequences.

B. Visual Discomfort Assessment and Analysis

Recently, there are many 3D visual discomfort metrics proposed in the literature [20]–[23]. However, most of them are developed for evaluating the visual discomfort of stereoscopic images. They are not designed specifically for visual discomfort evaluation of subtitle insertion in stereoscopic video sequences. The success of these algorithms is greatly dependent on the disparity richness for discomfort representation, such as statistical features [24], [25] and stimulus width [26]. On the contrary, all the letters in 3D subtitle have the same disparity values. Hence, these metrics are not suitable for 3D subtitle discomfort evaluation. Since the 3D subtitle visual discomfort is mainly induced by AV conflicts, we can reasonably draw lessons from the image discomfort assessment metrics, dependent on single disparity, to 3D subtitle discomfort assessment.

Since the stereo vision is mainly based on binocular disparity and processing of special visual path, it can achieve good performance via stimulating the procedure of stereo vision processing [27]. The detailed neural mechanism of the binocular disparity processing can be referred to [28] and [29]. Compared with other cortical areas, such as areas V1 and V4 [30], the area MT plays a major role in the subsequent horizontal disparity processing and depth perception in the visual path [31]. As the MT is related to the guidance of convergence eye movement, it can be used in the depth perception. Inspired by these perceptual theories, Park *et al.* [16] proposed an effective visual discomfort predictor (named Neural 3D-VDP) by using the neural activity of area MT during the processing of binocular disparity as visual discomfort prediction feature. Formally, the neural activity features used in the predictor can be described as:

$$f_i = \frac{E[r_i]}{R_{max}}, \quad i = \{1, 2, 3, \cdots, 13\} \tag{1}$$

where i is the neuron number, R_{max} is the maximum MT neuron response, $E[r_i]$ is the expected mean firing rate. The MT neuron response tuning functions can be modeled as:

$$R_i(d) = R_0^i + A_i \cdot e^{\frac{-0.5((d-d_0)^2}{\sigma_i^2})} \cdot cos(2\pi f_i(d - d_0^i) + \Phi_i) \tag{2}$$

where d is the horizontal disparity; R_0^i is the baseline response, A_i is the amplitude of the Gaussian kernel; σ_i is the width of the Gaussian; d_0^i is the center of the Gaussian; f_i is the frequency and Φ_i is the phase. The expected mean firing rates, $E[r_i]$, can be obtained by horizontal disparity probability distribution, $P[d]$, and the tuning function:

$$E[r_i] = \sum_d P[d] \cdot R_i(d) \tag{3}$$

Given a stereoscopic pair, it can be transformed into feature vector (f_i, $i = \{1, 2, \cdots, 13\}$) by feeding its disparity values into Eq. (1). Then, the Neural 3D-VDP model is trained by support vector regression (SVR), which builds the relationship between the feature vector and subjective opinion scores [32]. Given that all the letters in 3D subtitle have the same disparity value, we can take the disparity value as the input of Neural 3D-VDP model. Then the degree of visual discomfort is predicted as the output of the trained model. Therefore, we can reasonably employ Neural 3D-VDP model as a tool for visual discomfort evaluation of 3D subtitles. In addition to disparity, the visual discomfort of stereo image can also be induced by other factors, such as spatial Fourier energy [33], crosstalk [34] and vertical disparity [35], [36]. Perrin [17] designed a comfort function based on spatial frequency and horizontal disparity. It can be formulated as:

$$C(d, s) = \alpha(d - d_0 - ks^{k'}) \tag{4}$$

where s is the spatial frequency of images. α, d_0, k and k' are the constants with values of -0.01, 18.9, 221.1 and 0.74, respectively [17]. As can be seen from Eq. 4, it is not difficult to find that Perrin's metric relies on two variables. Compared to 3D image, the 3D subtitle has simpler characteristics, such as small spatial frequency, same disparity value. Moreover, the visual discomfort of 3D subtitle is mainly induced by horizontal disparity. Therefore, Perrin's method is a good choice for discomfort evaluation of 3D subtitles. Specifically, since the subtitles have the same texture and similar spatial frequency, their spatial frequency can be set to a constant. In such case, the visual discomfort score of Perrin's method mainly relies on the horizontal disparity. Based on the above discussion, both two metrics (i.e., Neural 3D-VDP and Perrin's method) can be used for discomfort evaluation of 3D subtitles.

In this paper, we propose a novel optimal region selection algorithm for 3D subtitle insertion by taking the visual discomfort into account. First, a 3D video database with various visual scenes and large range of disparities is constructed to evaluate the visual discomfort of inserted subtitles. Then, three subjective experiments are conducted to explore the approximate subtitle region. Subsequently, an automatic subtitle insertion algorithm is designed to select the most comfortable region for the subtitle insertion on each video frame. Finally, the subtitle positions of all video frames are analyzed and reorganized to form the video subtitle. We use two metrics of visual discomfort assessment to conduct the comparison experiments as well as a subjective evaluation. Experimental results demonstrate that the proposed automatic optimal region selection algorithm greatly reduces the visual discomfort of inserted subtitles.

Fig. 2. Example images from the TJU3D Video Database. To better present in the text, only the left view of the first frame in video is given.

TABLE I
DESCRIPTION OF TJU3D VIDEO DATABASE

Video Number	Time	Scene Scale	Video Number	Time	Scene Scale
001	10s	Small	026	10s	Small
002	10s	Large	027	10s	Small
003	20s	Small	028	15s	Small
004	20s	Large	029	10s	Small
005	10s	Large	030	10s	Small
006	10s	Large	031	15s	Small
007	20s	Large	032	10s	Large
008	15s	Large	033	15s	Small
009	15s	Large	034	15s	Small
010	15s	Small	035	15s	Small
011	10s	Small	036	10s	Large
012	10s	Large	037	15s	Small
013	15s	Large	038	15s	Small
014	10s	Small	039	15s	Large
015	10s	Small	040	15s	Small
016	20s	Small	041	10s	Large
017	15s	Large	042	10s	Large
018	20s	Small	043	10s	Large
019	15s	Large	044	20s	Large
020	8s	Large	045	10s	Small
021	10s	Small	046	10s	Large
022	10s	Small	047	10s	Small
023	10s	Small	048	10s	Small
024	10s	Large	049	10s	Small
025	15s	Small	050	10s	Small

III. SUBJECTIVE EVALUATION: A BENCHMARK

In this section, we introduce the constructed 3D database, subjective experiment of subtitle insertion, and data analysis of the subjective experiment. Specifically, the subjective experiment of subtitle insertion is conducted to explore the suitable region for subtitle insertion, while the subjective data is analyzed to summarize the optional region for subtitle insertion in terms of top, bottom and depth directions.

A. TJU3D Database

In order to create a benchmarking for 3D subtitle insertion, we build a 3D video database, namely, TJU3D video database. Totally, it includes 50 stereo video pairs with a high-definition (HD) resolution (1920×1080 pixels), captured by JVC GS-TD1BAC stereo camera. The video sequences in this database contain various natural visual scenes with highly diverse disparity values. During the video acquisition, the stereo camera was placed horizontally to reduce vertical disparity. Some example images chosen in video sequences are shown in Fig. 2. The optical flow [37] is used to compute the disparity under the consideration that the horizontal disparity can be estimated as the 'motion vector' between the left and right images. Table I lists the detailed information about the video sequences in this database.

Currently, there are two frequently used 3D video databases, namely IEEE-SA Stereo Database [38] and NAMA3DS1_COSPAD1 [39]. Compared to these existing public databases, there are two unique features for the built database in this study. First, NAMA3DS1_COSPAD1, whose quality is degraded with various levels, is mainly used for quality assessment rather than visual discomfort assessment. On the contrary, the video sequences in TJU3D video database, encoded in MPEG4 format with the frame rate of 30, are built specifically for visual discomfort assessment. Second, compared to IEEE-SA Stereo Database, there is no camera shake, which might affect the viewing comfort, for the video sequences in TJU3D video database. In addition, it contains complex scenes with many objects out of the screen beyond the comfort threshold.

B. Subjective Experiment of Subtitle Insertion

In 2D video sequences, the subtitles are usually inserted on the bottom of the screen. However, this method is inadequate for 3D subtitle insertion. As mentioned in Section II, the AV conflict in 3D video sequence is the main reason that causes visual discomfort. If subtitles are inserted in a fixed region (e.g., the bottom of the screen), the disparity value might be large to keep the subtitles in the front of the foremost object. In this case, it induces large visual discomfort. Therefore, we should place the 3D subtitles in optional regions rather than only one fixated region in order to reduce visual discomfort. Nevertheless, no work has been conducted to explore the optional regions for subtitle insertion. To the best of our knowledge, this is the first work to investigate optimal regions for 3D subtitle insertion and explore the regularity of subtitle's position on affection degree.

In this experiment, we first consider two cases for optional region selection: top subtitle region selection and bottom subtitle region selection, as shown in Fig. 5(a). Top subtitle region selection means the case of selecting the optional region on the top of the screen, while bottom subtitle region selection means the case of selecting the optional region on the bottom of the screen. Then, we design another experiment for subtitle disparity value selection compared to the foremost object in 3D scene. The depth subtitle region selection is used to explore the favorite disparity value of 3D subtitle compared to

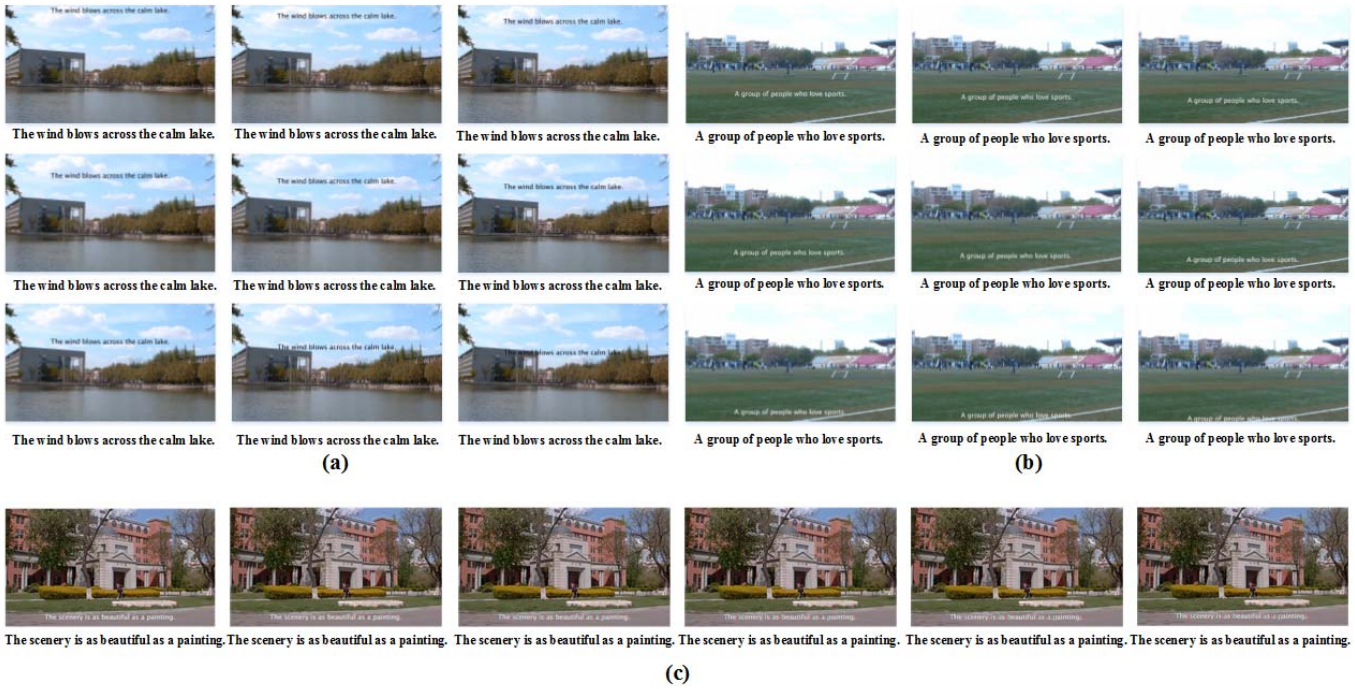

Fig. 3. Example images from the first frame of experiment stimuli, (a) top region selection, (b) bottom region selection and (c) depth region selection.

the foremost object in depth direction. Ten 3D video sequences with the duration time of 10s selected from TJU3D Video database were chosen as the experimental stimuli. All the disparity values are uniformly distributed without drastic change in each video. As both the cases of top and bottom subtitle region selection can be implemented in 2D videos, only the left views of 3D video sequences are used to conduct the experiments. For top subtitle region selection, the position changes from $\frac{1}{36}$ to $\frac{9}{36}$ of image height, H_I, in vertical direction with the step of $\frac{1}{36}$ H_I. For bottom subtitle region selection, the position changes from $\frac{27}{36}$ to $\frac{35}{36}$ H_I in vertical direction with the step of $\frac{1}{36}$ H_I. In subtitle disparity value selection, the initial disparity value of subtitle is the same with the minimum disparity value of the video frames. The other disparity values decrease in the step of 0.071° in depth direction. Note that, all the subtitles are centered in the horizontal direction (i.e., placed on the $\frac{1}{2}$ of image width, W_I, in the horizontal direction). Fig. 3 visualizes the example images from the first frame of experiment stimuli. In order to facilitate the viewing of the sentence in the image, we rewrite the sentence below the image.

The subjective experiments were conducted in the experiment environment with ordinary illumination level. The viewing distance was set as three times of the screen height according to the guideline of ITU-R BT.500 [40] and ITU-R 1438 [41]. All video sequences were displayed on a 23 inch LG2343P monitor with the resolution of 1920×1080 pixels. Twenty-two naive observers (10 female and 12 male) aged from 20 to 26 participated in the experiments. Before the depth subtitle region selection experiment, 3D vision test was conducted to verify the subjective abilities to view stereoscopic content. Single-stimulus (SS) method was employed to conduct the subjective experiment and all the stimuli were randomly presented by E-prime software [42]. Each test video sequence was presented for 10 seconds, and then participants were given 5 seconds to assess the QoE of the subtitle positions with a mild-grey image. The rating scores range from 1 to 100, where 1 denotes the worst perceptual experience and 100 denotes the best. To reduce the accumulated visual fatigue, participants were encouraged to take a break every 30 video sequences. The three experiments were arranged at three different days to reduce the learning effect.

C. Subjective Data Analysis

After the subjective experiment, we eliminate the abnormal data and analyze the experimental data. In this experiment, the Grubbs algorithm is adopted to eliminate the abnormal data. The standard deviation was first calculated as:

$$S = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2} \tag{5}$$

where n is the number of sample data, x_j is the j-th value after sorting, $\overline{x}$ is the mean value. The Grubbs operators can be expressed as:

$$G_j = \begin{cases} (x_j - \overline{x})/S, & if \ \ j > 1 \\ (\overline{x} - x_0)/S, & if \ \ j = 1 \end{cases} \tag{6}$$

The $G_{1-\alpha}$, where α equals to 0.05 in this study, can be checked in Grubbs test table. If $G_j > G_{1-\alpha}$, x_j is labeled as the abnormal value. The subjective data is identified as the outlier when more than 30 percent of the data is labeled as abnormal value. After abnormal data elimination, the final quality for each video sequence is computed as the average of subjective

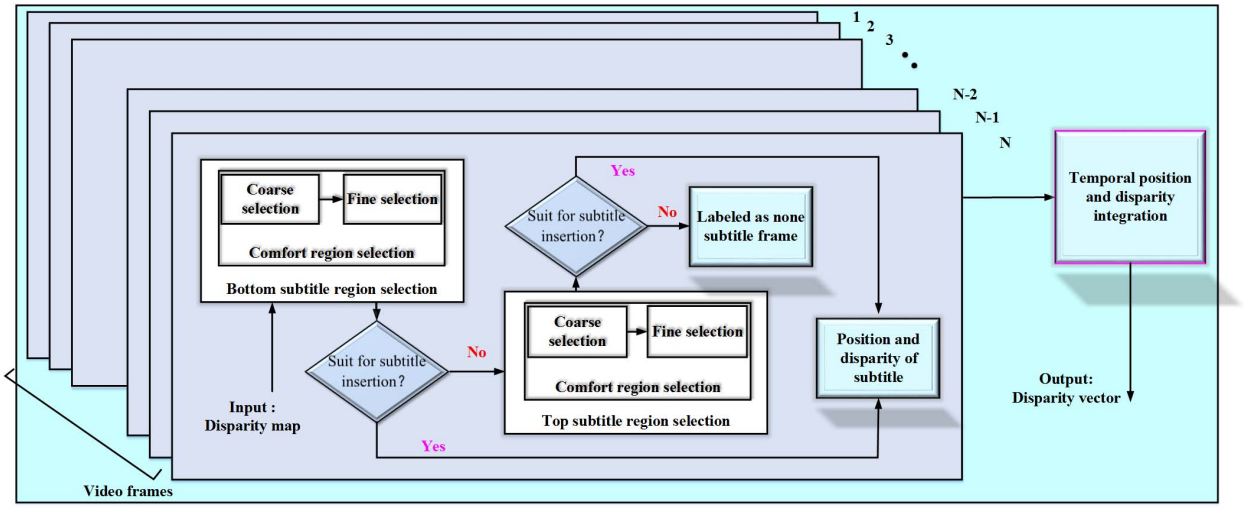

Fig. 4. Diagram of the proposed optimal region selection method.

values, namely the mean opinion score (MOS). The performance of individual subject can be evaluated by computing the correlation coefficients between subjective ratings and MOS values for each video set. Two commonly used performance measures (i.e., Pearson Linear Correlation Coefficient (PLCC) and Spearmans Rand-order Correlation Coefficient (SRCC)) are employed as the evaluation criteria. The higher PLCC and SROCC values, the better performance of prediction is [43], [44]. Through analysis, we find that the subjects perform quite consistently with relatively lower variations for different video sequences.

D. Subjective Experiment Summation

After the abnormal value elimination by the method in Part C, the remaining values with the same video are further analyzed. First, we compare the MOS values among different positions in the same subtitle region (e.g., the top subtitle region). Second, we also calculate the statistical significance among the MOS values of video sequences with different subtitle positions. For this purpose, analysis of variance (ANOVA) is used to investigate the main effects of subtitle position. The test is conducted at 5% significance level, where $p < 0.05$ means that MOS values are not significantly equivalent. The ANOVA test was conducted separately on each region selection result, i.e., top region, bottom region and depth region.

As expected, the subtitle position has influence on the degree of viewing experience. To be specific, the MOS exhibits the maximum value on the position of $\frac{1}{12}$ H_I and $\frac{11}{12}$ H_I in top and bottom subtitle region selection, respectively. Then, the MOS value slowly gets smaller on both sides of the maximum value. In addition, the MOS exhibits higher value in depth region selection experiment when the disparity value is 0.071° or 0.141° smaller than that of the foremost object. Moreover, based on the statistical results, there is significant influence for QoE with the subtitle position in the top, bottom and depth region selection with $p < 0.01$. Therefore, considering the viewing experience, we stipulate that the top (bottom) subtitle region ranges from 5.6% (88.9%) to 11.1% (94.4%) in the vertical direction and contain all the regions in the horizontal direction. Besides, the favorite value of subtitle disparity is 0.106° smaller than that of the foremost object in the depth direction.

IV. Proposed Optimal Region Selection Method

The proposed optimal region selection method for stereoscopic subtitle insertion is depicted as Fig. 4. It mainly includes comfort region selection, temporal position selection and disparity value integration. Particularly, the comfort region selection is composed of coarse region selection and fine region selection. Considering the temporal coherence, the temporal position and disparity value integration are further separated into two portions, region classification and position integration.

A. Comfort Region Selection

1) Coarse Selection: The main goal of the coarse selection is to preselect the disparity value and position of subtitle from the optional subtitle region (Ω). In this paper, we stipulate both bottom and top subtitle regions, determined by subjective experiments in Section III-D, as the optional regions. Since the disparity values vary greatly in the optional region, we divide the optional region into many small candidate subtitle regions (O), whose sizes are the same as that of subtitle text, and analyze each candidate subtitle region individually to validate whether this region is suitable for subtitle insertion. Fig. 5 depicts the schematic diagram of coarse region selection. For each candidate subtitle region, we firstly judge whether the region meets the conditions and suits for subtitle insertion. Formally, for the candidate subtitle region, the mean value of the disparity values is first computed as:

$$\overline{\mu} = \frac{1}{K}\sum_{k=1}^{K}\mu_k \tag{7}$$

where μ_k is the disparity value of k-th pixel in the candidate subtitle region, K is the number of pixels. Then the standard

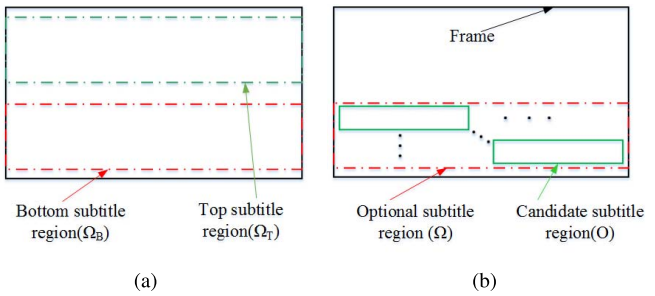

Fig. 5. Schematic diagram of subtitle region, (a) optional subtitle regions and (b) procedure of coarse region selection.

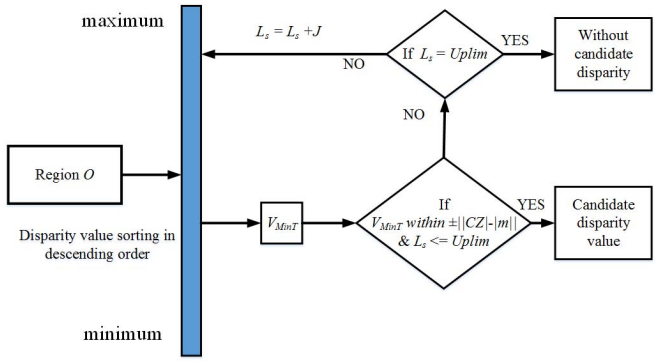

Fig. 6. Flowchart of candidate disparity value selection.

deviation is calculated as:

$$\delta = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\mu_k - \overline{\mu})} \tag{8}$$

The disparity value is treated as an outlier if it is not in the range $[\overline{\mu} - \delta, \overline{\mu} + \delta]$. A region can be declared as a qualified region only if the ratio between the numbers of outlier and all the pixels are less than 0.1 under the basis that the mean value is not beyond the comfort threshold.

If the region is preliminary judged as the suitable region, then, a further inspection is required to determine the candidate disparity for subtitle insertion. Fig. 6 gives the flowchart of candidate disparity value selection for candidate region. It can be depicted as following steps:

1) We sort all the disparity values in O in ascending order.
2) The disparity values, T, which are larger than the disparity value in the L_s-th place of the resort disparity values, are selected to further inspection. Specifically, if the minimum value (V_{MinT}) of T is within comfort zone, CZ (i.e., $\pm 1°$ [45]), it is selected as candidate disparity value. Then perform the Step 4. Otherwise, perform the Step 3. Note that, as the subtitle should be inserted in the front of object, its comfort zone should be modified as $[-|-1-|m||, 1-|m|]$.
3) The L_s is updated by adding J. Then, we execute Step 2 again. The region O is abandoned when no candidate disparity value is obtained during the L_s changing from $Lowlim$ to $Uplim$.
4) According whether the candidate disparity value is obtained, region O is determined whether to insert subtitle. Moreover, each qualified candidate region selects one candidate disparity value for subtitle insertion. In this paper, the values of $Lowlim$, $Uplim$ and J, are set to 0.3, 0.6 and 0.1 empirically; m is set to be 0.106 according to the subtitle disparity selection experiment, described in Section III-B.

All the candidate subtitle regions (O) are processed using the above procedure in the step of 20 pixels from the upper left to the lower right (as shown in Fig. 5(b)) with two considerations, computation complexity reduction and candidate disparity value reduction. Then, the candidate disparity values from the qualified regions are sorted in ascending order. In order to reduce the AV conflicts, candidate disparity values with smaller distances from zero disparity are selected preferentially as the final candidate disparity value, f_m. In addition, f_m is with the high frequency among all the candidate disparity values.

2) Fine Selection: The purpose of fine selection is to determine the position and disparity value of subtitles more accurately on the basis of the coarse selection. First, all the coordinates with the final candidate disparity value are pricked down. Then, all the subtitle regions centered on the chosen coordinates with width W and height H are selected. W and H are the height and width of the inserted subtitle, respectively. For the selected subtitle region, the statistical procedure mentioned in coarse selection phase is conducted to determine the disparity value. As a result, we can obtain an optimal disparity value for each selected subtitle region. Given that observers are more likely to view the subtitles placed in the middle of the horizontal direction of screen, the coordinate (x_{co}, y_{co}) of subtitle is determined by the following two conditions:

$$x_{co} = \underset{p \varepsilon X}{argmin} \|p - X_c\|_2 \tag{9}$$

$$y_{co} = \underset{q \varepsilon Y}{argmin} \|q - Y_c\|_2 \tag{10}$$

where X_c and Y_c are the optimal locations in the video frame which are ascertained by subjective experiments with values $\frac{33}{36}\ H_I$ and $\frac{1}{2}\ W_I$, (discussed in Section V-C), respectively. X and Y are the coordinate sets with the optimal disparity value for each selected subtitle region. Meanwhile, the disparity of subtitle is defined as $(f_d - 0.106°)$, where f_d is the optimal disparity in the selected region of coordinate (x_{co}, y_{co}). In such case, the subtitle region satisfies the demands of discomfort reduction and basic viewing habits.

B. Temporal Consistency for Subtitle Insertion

Due to the similarity of adjacent frames, we have to consider temporal coherence between consecutive frames when integrating the subtitle position and disparity value. This section contains two parts, region classification and position integration.

1) Region Classification: As discussed in Section IV, the frame, which is not suitable for subtitles insertion, is labeled as none subtitle frame. Since the consecutive frames has consistence characteristic, the subtitles disparity of adjacent frames may also be close to discomfort threshold.

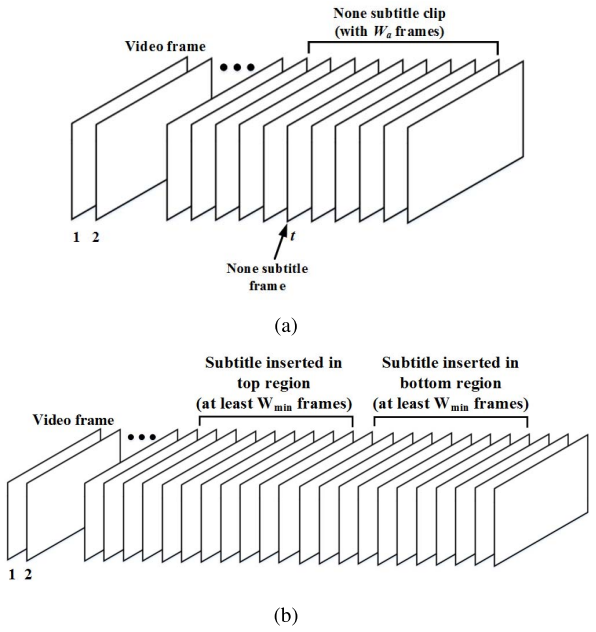

Fig. 7. Two operations in *Region Classification*, (a) transitional window arrangement and (b) classification of top and bottom regions.

To further reduce visual discomfort and avoid the subtitle disappearing suddenly, a transitional window W_a is arranged on both sides of the none subtitle frame, as shown in Fig. 7(a). Specifically, for the t-th video frame without subtitle in the video sequence, the clip centered on the t-th video frame with number of W_a frames are also demarcated as none subtitle clip. Besides, the subtitle can be inserted both on the top and bottom of the frame using the proposed method. Since the disparity information changes quickly in the video sequence, ($f_d - 0.106°$), the subtitle positions might change frequently in bottom and top regions. To address this problem, we classify the regions as top and bottom regions and then analyze them individually for better presentation. In practice, the subtitles with the same contents must be lasted for more than the minimum value W_{min} for better viewing experience in video sequences. Generally, for each subtitle region, the consecutive frames with the same region category should be extended to minimum value if it is less than W_{min}, as shown in Fig. 7(b). In order to obtain good viewing experience, the values of W_a and W_{min} are set as 30 and 30 empirically, respectively.

2) Position Integration: Although the region selection indeed reduces the visual discomfort, some new problems arise unfortunately. In the proposed method, we aim to reduce the visual discomfort by finding the optimal region with smaller AV conflicts. As the disparity contents are different, the subtitle position changes from frame to frame using the proposed method. That may induce the visual discomfort. How to unify the positions of subtitles between consecutive frames is indeed a serious problem. If the subtitle positions for all frames in a video sequence are set to be a constant, the similar problem will occur as the basic method. Fortunately, with the improvement of the film production technology, the scene duration has declined from ca. 15s to ca. 3s [46], leading to highly similar characteristics in the similar visual scene. Therefore, the subtitle positions and disparity values are similar in one scene. In this study, we reorganize the subtitle positions in unit of visual scene, where the video frames have similar content and disparity distribution. For a scene, the position of subtitle is further analyzed in unit of category if it has two categories defined as in *Region Classification*. Generally speaking, for a certain range area with the same category, the final subtitle position, (x_s, y_s), is determined by the median values of both vertical and horizontal coordinates.

$$(x_s, y_s) = median(X_f) \tag{11}$$

where X_f is the subtitle coordinate sets in the same region category, e.g., top subtitle region. Meanwhile, the disparity value, V_{ds}, is denoted as the minimum subtitle value of all the frames in this area with the same category.

$$V_{ds} = argmin(V_f) \tag{12}$$

where V_f is the disparity value set in the same region category. Since we aim to design an optimal subtitle region selection algorithm in this study, how to determine the duration of scene is not discussed here. It can be manually marked or automatically marked by certain algorithm [47], [48].

V. Experimental Results and Analysis

In this section, the experiment setup of 3D subtitle insertion is firstly introduced. Then, the performance of the proposed region selection algorithm on the aspect of reducing the visual discomfort is analyzed by two existing objective metrics as well as a subjective experiment. For this task, we test the proposed method on IEEE-SA Stereo Database [38] and TJU3D Video database.

A. Experimental Setup

The IEEE-SA Stereo Database contains two parts, namely Vesion1 and Vesion2. There are 17 indoor scenes and 24 outdoor scenes in Vesion1 and 9 indoor scenes and 14 outdoor scenes in Vesion2. Throughout analysis, we observe that this database also contains visual scenes that easily induce visual discomfort. It seems meaningful to verify the effectiveness of the proposed method on it even though it is specifically built for video comfort assessment. According to the disparity distribution, we divided it into three types, screen-in scenes, half screen-in scenes, and screen-out scenes. The screen-in scenes are the ones where most image contents are presented in the screen and no disparity exceeds the comfort zone. The half screen-in scene is defined as that where most image content is presented on both sides of the screen with smaller disparity values and no disparity exceeds the comfort zone. The screen-out scene is defined as that where the disparity exceeds the comfort zone. 40 video sequences are chosen from three types in total. Similarly, the TJU3D Video database was also divided into these three categories.

To the best of our knowledge, it is the first attempt to select optimal region for subtitle insertion in 3D video sequences and there is no other specialized visual discomfort prediction metric for it. Here, we regard subtitle insertion in front of

TABLE II
PERFORMANCE EVALUATION OF PROPOSED METHOD AND COMPLETING METRIC ON TJU3D VIDEO DATABASE

Name	Scene	Neural 3D-VDP		Comfort function		Name	Scene	Neural 3D-VDP		Comfort function	
		Basic	Pro	Basic	Pro			Basic	Pro	Basic	Pro
001	HSI	2.28	2.34	2.19	2.19	026	SO	1.90	1.93	2.06	2.07
002	SI	2.37	2.37	2.21	2.21	027	HSI	2.35	2.35	2.20	2.20
003	HSI	1.45	1.45	1.91	1.91	028	HSI	2.38	2.38	2.22	2.22
004	HSI	2.29	2.29	2.16	2.16	029	HSI	2.15	2.15	2.10	2.10
005	SI	1.98	2.03	2.08	2.10	030	SO	1.32	1.85	1.82	2.00
006	SI	1.24	1.24	1.74	1.81	031	SI	2.40	2.40	2.26	2.26
007	HSI	2.33	2.34	2.18	2.19	032	HSI	2.18	2.18	2.14	2.14
008	SI	2.36	2.37	2.20	2.21	033	SO	1.90	1.97	2.02	2.06
009	SI	2.39	2.39	2.24	2.25	034	SO	1.40	1.85	1.41	2.07
010	HSI	2.26	2.31	2.15	2.17	035	SO	2.00	2.04	2.07	2.08
011	SO	1.29	2.35	1.04	1.09	036	SI	2.22	2.22	2.13	2.13
012	SI	2.36	2.36	2.20	2.20	037	HSI	2.10	2.10	2.09	2.09
013	HSI	2.33	2.33	2.18	2.18	038	SO	1.80	1.90	1.94	2.05
014	HSI	1.81	1.81	2.01	2.01	039	HSI	2.35	2.35	2.19	2.19
015	HSI	2.15	2.19	2.11	2.12	040	HSI	1.74	1.74	1.99	1.99
016	SO	1.65	1.99	1.85	2.12	041	SI	2.36	2.36	2.20	2.20
017	SI	2.32	2.33	2.18	2.18	042	HSI	2.18	2.20	2.13	2.13
018	HSI	1.44	1.44	1.90	1.90	043	SI	2.40	2.40	2.28	2.28
019	SI	2.40	2.40	2.27	2.27	044	HSI	2.37	2.38	2.22	2.23
020	SI	2.40	2.40	2.27	2.27	045	SO	2.01	2.24	1.22	2.06
021	SO	1.45	1.65	1.76	1.99	046	HSI	2.19	2.19	2.12	2.12
022	SO	1.52	1.55	1.76	1.92	047	SO	1.24	2.37	1.20	2.40
023	SI	2.27	2.27	2.15	2.15	048	SO	1.16	2.37	1.61	2.40
024	HSI	2.32	2.32	2.14	2.14	049	SO	1.24	2.39	1.04	2.32
025	SO	1.75	1.77	1.91	2.00	050	SO	1.16	1.35	1.68	1.84

the foremost objects as the basic method, which is used as the baseline for comparison experiment. The visual discomfort of inserted subtitles is evaluated using Neural 3D-VDP and comfort function. Since the Neural 3D-VDP was obtained by training IEEE-SA Image Database, its performance may be changed as the training data changes. In order to preserve the same performance, the model was trained by using IEEE-SA Image Database [49] with the same operation in [16].

In order to further validate the effectiveness of the proposed method, we conduct another subjective test. Twenty subjects, from twenty to thirty years old, participated in the subjective evaluation. The experimental environment and preparation work are the same as Section III-B. For a video, it was inserted with subtitles using both basic and the proposed methods. Then, the generated video pairs with subtitles by these two methods were presented to the subjects successively in random order. The subjects need to assess the uncomfortable degree of inserted subtitles between two video sequences: "1: better", "0: equal" and "−1: worse". Since the experiment comparison is only conducted on each generated video pair with different subtitles by these two methods, the experimental results would not be affected by the duration time of video sequences.

B. Visual Discomfort Prediction

Since the disparity values of subtitles have been determined, the degree of discomfort is predicted using the two metrics mentioned above. Specifically, for Neural 3D-VDP method, the model is firstly trained using IEEE-SA Image database. Then, the disparity values of video sequence are converted into angle unit and treated as the input of the model. As there are two variables in the comfort function, the influence of spatial frequency should be eliminated. Since the subtitle has the same texture, the spatial frequency is low. Thus, we set it with value of 1 cpd. The whole discomfort of subtitles in video sequence is ultimately made up of the discomfort degree of each frame. As larger disparity value induces more discomfort degree, it should occupy a larger weight in summation. Hence, the overall visual discomfort (VD) can be formulated as:

$$VD = \sum_{l=1}^{L} \omega_l \cdot s_l \tag{13}$$

where ω_l is the weight of the l-th frame in the video sequence and it is determined by the absolute value of the disparity value of subtitles of this video frame; s_l is the predicted score of l-th video frame using the discomfort assessment metric; L is the number of video frame in the video sequence. For simplicity, we renamed basic method and the proposed method as 'Basic' and 'Pro', respectively.

C. Experimental Results

Tables II and III provide the performance of two visual discomfort evaluation metrics on both 3D databases. For simple expression, the screen-in scene, half screen-in scene and screen-out scene are simplified as SI, HIS and SO, respectively. Note that the higher predicted value corresponds to the higher visual comfort.

From these two Tables, it can be observed that there is little difference between the performance of the basic method and the proposed method in terms of screen-in scenes and half screen-in scenes. The reasons are explained as follows. On the one hand, the image contents of screen-in scenes are in the screen or near the screen plane where is comfortable. The subtitles could be arranged in a position with smaller

TABLE III
PERFORMANCE EVALUATION OF PROPOSED METHOD AND COMPLETING METRIC ON IEEE-SA STEREO DATABASE

Name	Scene	Neural 3D-VDP		Comfort function		Name	Scene	Neural 3D-VDP		Comfort function	
		Basic	Pro	Basic	Pro			Basic	Pro	Basic	Pro
Walking-person1	SI	2.37	2.37	2.40	2.40	Library6	SO	1.31	1.42	1.78	1.88
Toy-train1	SI	2.37	2.37	2.40	2.40	Flume-ride1	SO	1.24	1.74	0.88	1.98
Swing-tree2	SO	1.92	1.94	1.76	1.80	Flower1	SO	1.33	1.65	1.85	1.96
Street-market7	SO	1.81	2.05	1.36	1.92	Crosswalk2	SO	1.20	1.61	1.32	2.00
Street-market6	SO	1.83	2.05	1.47	1.90	Car1	SI	2.24	2.25	2.14	2.14
Street-market2	SO	1.75	2.13	0.95	2.06	Bungee-drop1	SO	1.48	1.48	1.89	1.90
Statue1	HSI	2.28	2.29	2.10	2.12	Basketball1	SO	1.24	2.35	1.11	2.33
Statue2	HSI	2.29	2.31	2.11	2.13	Amusementpark1	SI	2.12	2.16	2.10	2.11
Spin-ride1	HSI	2.06	2.06	1.93	1.93	University1	SI	1.36	1.38	1.76	1.87
Parade1	SO	1.90	1.97	1.64	1.83	Concert2	SI	2.35	2.35	2.23	2.23
Metro1	HSI	2.30	2.31	2.15	2.18	Concert4	SI	2.38	2.38	2.35	2.35
Metro2	SO	1.91	2.05	1.63	1.92	Excavator1	SI	2.37	2.37	2.40	2.40
Metro3	SO	1.87	1.99	1.53	1.84	Roller-coaster1	SO	1.24	2.22	1.07	2.14
Market1	HSI	2.20	2.20	2.05	2.06	Marathon2	SO	1.80	1.98	1.90	2.04
Marathon1	SO	1.49	1.51	1.56	1.89	Statue3	SO	1.26	1.29	1.78	1.83
Marathon4	SI	1.53	1.53	1.93	1.93	Street2	SO	1.27	2.07	0.69	2.20
Library2	SO	1.75	1.83	1.95	2.02	Walking-person7	HSI	1.99	1.99	2.06	2.06
Library3	HSI	2.25	2.25	2.14	2.14	Walking-person8	HSI	2.02	2.02	2.07	2.07
Library4	SO	1.81	1.85	2.01	2.02	Car2	SI	2.31	2.31	2.17	2.17
Library5	HSI	2.18	2.19	2.15	2.16	Concert3	SI	2.37	2.37	2.37	2.37

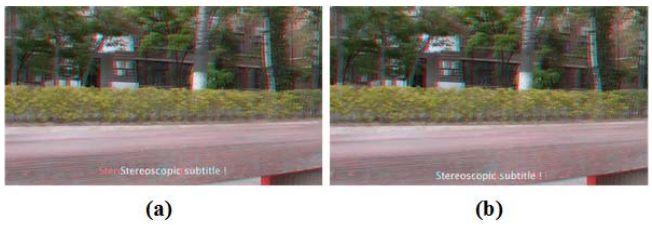

Fig. 8. Comfort score of different subtitle insertion regions, (a) basic method with 1.71 and (b) the proposed method with 2.01.

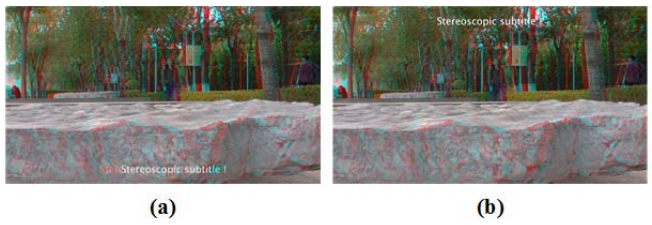

Fig. 9. Subtitle insertion region selection using different methods, (a) basic method and (b) the proposed method.

disparity value. Thus, no automatic region search is required to reduce visual discomfort, leading to the subtitle position similar with the basic method. On the other hand, for the half screen-in scenes, the automatic region search is also unnecessary. As there is no disparity beyond the comfort threshold, the subtitle disparity value also does not exceed the threshold. In order to meet the viewing habits, the subtitles are presented in the middle of the bottom subtitle region as the same as the basic method.

However, the results have changed dramatically with respect to the screen-out scenes. It can be intuitively observed that the visual discomfort of the automatically inserted 3D subtitles is greatly reduced compared to the basic method. The rational explanation of this can be summarized as follows. Firstly, adaptive selection region for subtitle insertion, to a great extent, reduces the AV conflicts. Fig. 8 depicts the differences between the basic method and the proposed adaptive method. Fig. 8 (a) is the frame inserted by traditional method with comfort score 1.71, while Fig. 8 (b) is processed by the proposed method with comfort score 2.01. In basic method, the subtitle is inserted directly in the front of the foremost objects, which may exceed the comfort threshold and cause visual discomfort. On the contrary, the proposed method selects the optimal region with less AV conflicts. Therefore, it reduces the visual discomfort to a certain. Secondly, two optional subtitle regions (top and bottom subtitle regions) are provided for subtitle insertion. The proposed method tests the feasibility of inserting subtitles in the two regions successfully. If the bottom subtitle region is not suitable for subtitle insertion, then the top region is considered. In Fig. 9(a), the disparity value in bottom region is sufficiently large to insert the subtitle. Thus, the subtitle should be inserted on the top region for sematic expression and visual discomfort reduction, as shown in Fig. 9(b). Thirdly, the proposed method can automatically determine whether to insert subtitles or not. If both optional regions are not suitable for subtitle insertion, then, the video frame is defined as none subtitle frame. As a result, the visual discomfort would not be induced in this video frame in view of no subtitle insertion. In this case, the visual discomfort is avoided. Last but not the least, the interpolation of interim window for none subtitle frame further reduces the visual discomfort for these video frames. Since more video frames do not require subtitle insertion, the total visual discomfort score is determined by weighted summation of the remaining frames. On the contrary, the basic method needs to summary all the video frames together.

Fig. 10 shows the subjective comparison results for visual discomfort on two databases. In these figures, the x-axis represents the video sequence index and the y-axis represents their corresponding mean values of all the subjective scores. From Fig. 10, we can intuitively observe that both the basic and proposed methods obtain the similar performance

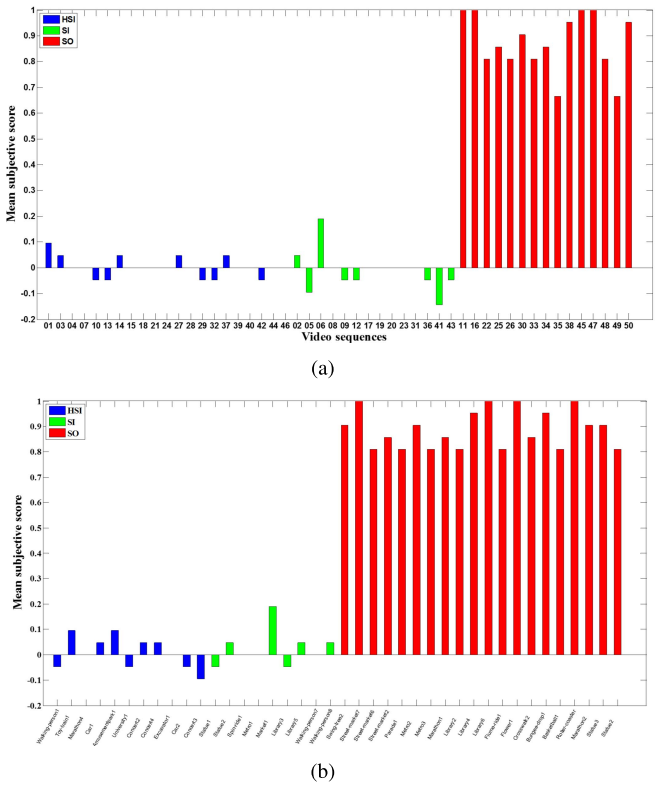

Fig. 10. Mean subjective scores of the inserted subtitles on (a) TJU3D database and (b) IEEE-SA Stereo Database.

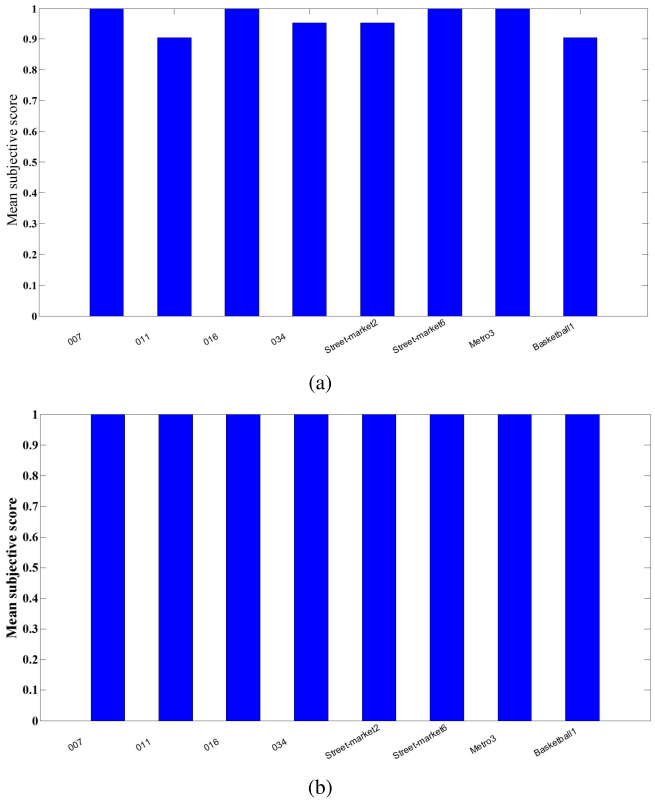

Fig. 11. Mean subjective scores of the proposed method with/without certain operation (a) *Region Classification* (b) *Position Integration*.

on HSI and SI scenes, while the performance changes dramatically in terms of SO scenes. Specifically, the proposed method reduces the discomfort of the inserted subtitles. Moreover, we also find some detailed information from these figures. First, small fluctuations exist in terms of HSI and SI scenes. This phenomenon may be attributed to that subjects might not clearly distinguish the difference between the two subtitle-inserted video sequences if the difference is sufficient small. Second, parts of the SO scenes receive low scores (e.g., the video 36 obtains 0.6). The reason is that some judgement mistakes are easily caused when the task becomes difficult. Specifically, according to the parameter setting in this article (the video duration is mostly longer than 10s and the video shot duration is 2s), one video can be divided into 5 video shorts (obtained as the merchant between video duration and video shot duration). For the simple scene, most video shots have the same or similar subtitle positions. On the contrary, the position of subtitle changes frequently when the video content is complex. Therefore, it increases the difficulty of task, leading to the judgement mistakes. This phenomenon would be reduced when it meets the true film scene, whose video shot ranges from ca. 15s to ca. 3s [46].

D. Impact of Temporal Coherence

As discussed in the manuscript, the operation of temporal coherence helps to reduce the visual discomfort of 3D subtitle. In this section, we conduct an additional subjective experiment to investigate the impact of the temporal coherence. Since the temporal coherence contains two components, i.e., region classification and position integration, we separately test their impact on visual discomfort reduction. Moreover, the validation of region classification's impact requires video sequences, which is inserted with both top and bottom subtitle by the proposed method. For this purpose, the video sequences (007, 011, 016, 034, Street-market2, Street-market6, Basketball1, Metro3) are chosen for the subjective experiments. To test the impact of region classification, each video sequence is processed twice (with and without the operation of region classification) using the proposed method. Because the subtitle is usually inserted on the bottom of image, we only insert subtitle on the bottom of video frame in the case without region classification. That is, the subtitle inserted on the top region is changed to bottom region. With this operation, the subtitle disparity value becomes smaller and induce larger AV conflict. Similarly, to test the impact of position integration, each video was processed twice (with and without the operation of position integration) using the proposed method. The generated video pairs processed in the case with/without certain operation (i.e., region classification or position integration) were presented to the subjects in random order. The experiment procedure is described as bellow.

Twenty subjects were involved in the subjective experiment, whose setup is the same as that in Section III-B. The subjects were asked to assess the uncomfortable degree of inserted subtitles for these two video sequences with the following ratings: "1: better", "0: equal" and "−1: worse". Note that,

for exploring any components impact, the other component is normally processed based on the operation in Section IV-B. Fig. 11 depicts the experimental results. In these figures, the x-axis represents the video sequence index and the y-axis represents their corresponding mean values of all the subjective scores. Theoretically, the mean value of subjective scores ranges from -1 to 1, where -1 (1) denotes the visual discomfort increasing (declining) when the video sequence is processed with the specific operation (e.g., region classification and position integration). During the experiment, most subjects perceive better viewing experience when the video sequence is processed with region classification. As can be seen from Fig. 11(a), the visual discomfort is declined with the region classification operation. The reason might be that, by changing the top subtitle to bottom subtitle, the AV conflict caused by subtitle is reduced. Therefore, the visual discomfort is declined. Fig. 11(b) shows the subjective results in cases with & without position integration operation. During the experiment, all the subjects agree that the viewing experience is better when the video sequence is processed with position integration. Because without position integration operation, the subtitle position of each video frame changes and causes the subtitle jitter between video frames, which seriously affects the subtitle reading. In summary, both components of temporal coherence help to reduce visual discomfort of 3D subtitle.

VI. Conclusion

In this paper, we have presented an optimal region selection algorithm for 3D video subtitle insertion by considering visual discomfort. To the best of our knowledge, it is the first attempt for automatically inserting subtitles in 3D video sequence. We first build the TJU3D Video database to test the degree of visual discomfort of inserted subtitles. Then, subjective experiments are conducted to explore the favorite degree of subtitle region both in plane and depth. Next, a 3D subtitle insertion region selection algorithm is operated on each frame for subtitle insertion. Finally, a region classification and disparity value integration algorithm are applied to improve the performance of video subtitle insertion. The effectiveness of the proposed method on discomfort reduction is validated via the comparison experiment between the proposed method and basic 3D subtitle insertion method on IEEE-SA Stereo database and TJU3D Video database. Experimental results demonstrate that the proposed method greatly reduces the visual discomfort compared with the basic method.

References

[1] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett, "Three-dimensional displays: A review and applications analysis," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 362–371, Jun. 2011.

[2] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 2971–2983, Oct. 2015.

[3] F. Shao, W. Lin, Z. Li, G. Jiang, and Q. Dai, "Toward simultaneous visual comfort and depth sensation optimization for stereoscopic 3-D experience," *IEEE Trans. Cybern.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/document/7604111/

[4] Y. Jung, H. Sohn, S.-I. Lee, and Y. Ro, "Visual comfort improvement in stereoscopic 3D displays using perceptually plausible assessment metric of visual comfort," *IEEE Trans. Consum. Electron.*, vol. 60, no. 1, pp. 1–9, Feb. 2014.

[5] K. Masaoka, A. Hanazato, M. Emoto, H. Yamanoue, Y. Nojiri, and F. Okano, "Spatial distortion prediction system for stereoscopic images," *J. Electron. Imag.*, vol. 15, no. 1, p. 013002, 2006.

[6] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence–accommodation conflicts hinder visual performance and cause visual fatigue," *J. Vis.*, vol. 8, no. 3, p. 33, 2008.

[7] S.-P. Du, B. Masia, S. M. Hu, and D. Gutierrez, "A metric of visual comfort for stereoscopic motion," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 222:1–222:9, 2013.

[8] K. Templin, P. Didyk, K. Myszkowski, M. M. Hefeeda, H.-P. Seidel, and W. Matusik, "Modeling and optimizing eye vergence response to stereoscopic cuts," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 145:1–145:8, 2014.

[9] T. Bando, A. Iijima, and S. Yano, "Visual fatigue caused by stereoscopic images and the search for the requirement to prevent them: A review," *Displays*, vol. 33, no. 2, pp. 76–83, 2012.

[10] B. Zafarifar, J. Cao, and P. H. N. de With, "Instantaneously responsive subtitle localization and classification for TV applications," in *Proc. IEEE Int. Conf. Consum. Electron.*, Jan. 2011, pp. 274–282.

[11] D. Cai, C.-F. Chi, and M. You, "The legibility threshold of chinese characters in three-type styles," *Int. J. Ind. Ergonom.*, vol. 27, no. 1, pp. 9–17, 2001.

[12] M. Lambooij, M. J. Murdoch, W. A. IJsselsteijn, and I. Heynderickx, "The impact of video characteristics and subtitles on visual comfort of 3D TV," *Displays*, vol. 34, no. 1, pp. 8–16, Jan. 2013.

[13] Y.-Y. Yeh and D.-S. Lee, "Characteristics of subtitle on preferred viewing distance and subjective preference of liquid crystal display high-definition television," in *Proc. Int. Congr. Image Signal Process.*, 2012, pp. 1734–1737.

[14] M. Davoudi, M. B. Menhaj, N. S. Naraghi, A. Aref, M. Davoodi, and M. Davoudi, "A fuzzy logic-based video subtitle and caption coloring system," *Adv. Fuzzy Syst.*, vol. 2012, Jan. 2012.

[15] S. Wan, B. Chang, and F. Yang, "Viewing experience of 3D movie with subtitles where to put subtitles in a 3D movie?" in *Proc. Int. Workshop Qual. Multimedia Exper.*, Jul. 2013, pp. 170–175.

[16] J. Park, H. Oh, S. Lee, and A. C. Bovik, "3D visual discomfort predictor: Analysis of disparity and neural activity statistics," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1101–1114, Mar. 2015.

[17] J. Perrin, P. Fuchs, C. Roumes, and F. Perret, "Improvement of stereoscopic comfort through control of the disparity and of the spatial frequency content," *Proc. SPIE*, vol. 3387, pp. 124–134, Jul. 1998.

[18] J. Lei *et al.*, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.

[19] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2625–2636, Jun. 2014.

[20] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "On predicting visual comfort of stereoscopic images: A learning to rank based approach," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 302–306, Feb. 2016.

[21] Z. Chen, J. Shi, X. Huang, L. Yun, and Y. Tai, "Visual comfort modeling for disparity in 3D contents based on Weber–Fechner's law," *J. Display Technol.*, vol. 10, no. 12, pp. 1001–1009, 2014.

[22] Q. Jiang, F. Shao, G. Jiang, M. Yu, and Z. Peng, "Three-dimensional visual comfort assessment via preference learning," *J. Electron. Imag.*, vol. 24, no. 4, p. 043002, 2015.

[23] H. Ren, Z. Su, C. Lv, and F. Zou, "Effect of region contrast on visual comfort of stereoscopic images," *Electron. Lett.*, vol. 51, no. 13, pp. 983–985, 2015.

[24] C. Jung and S. Wang, "Visual comfort assessment in stereoscopic 3D images using salient object disparity," *Electron. Lett.*, vol. 51, no. 6, pp. 482–484, 2015.

[25] H. Sohn, Y. J. Jung, S.-I. Lee, and Y. M. Ro, "Predicting visual discomfort using object size and disparity information in stereoscopic images," *IEEE Trans. Broadcast.*, vol. 59, no. 1, pp. 28–37, Mar. 2013.

[26] S.-I. Lee, Y. J. Jung, H. Sohn, F. Speranza, and Y. M. Ro, "Effect of stimulus width on the perceived visual discomfort in viewing stereoscopic 3-D-TV," *IEEE Trans. Broadcast.*, vol. 59, no. 4, pp. 580–590, Dec. 2013.

[27] 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.

[28] T. B. Czuba, A. C. Huk, L. K. Cormack, and A. Kohn, "Area MT encodes three-dimensional motion," *J. Neurosci.*, vol. 34, no. 47, pp. 15522–15533, 2014.
[29] T. M. Sanada and G. C. DeAngelis, "Neural representation of motion-in-depth in area MT," *J. Neurosci.*, vol. 34, no. 47, pp. 15508–15521, 2014.
[30] J. Read, "Early computational processing in binocular vision and depth perception," *Prog. Biophys. Mol. Biol.*, vol. 87, no. 1, pp. 77–108, 2005.
[31] G. C. DeAngelis, B. G. Cumming, and W. T. Newsome, "Cortical area MT and the perception of stereoscopic depth," *Nature*, vol. 394, no. 6694, pp. 677–680, 1998.
[32] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2007, Art. no. 27.
[33] O. Penacchio and A. J. Wilkins, "Visual discomfort and the spatial distribution of Fourier energy," *Vis. Res.*, vol. 108, pp. 1–7, Mar. 2015.
[34] 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.
[35] X. Meng, J. Jin, L. Shen, K. Fan, S. Zhang, and Y. Huang, "Objective evaluation of vertical parallax using EEG," in *Proc. Int. Conf. Comput. Sci. Edu.*, 2013, pp. 407–410.
[36] Z. Di, S. Xinzhu, W. Peng, and C. Duo, "Comparative visual tolerance to vertical disparity on 3D projector versus lenticular autostereoscopic TV," *J. Display Technol.*, vol. 12, no. 2, pp. 178–184, Feb. 2016.
[37] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2432–2439.
[38] *Stereoscopic (3D Imaging) Database*. Accessed on Jun. 2015. [Online]. Available: http://grouper.ieee.org/groups/3dhf/ and ftp://165.132.126.47/
[39] M. Urvoy *et al.*, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *Proc. Int. Workshop Qual. Multimedia Exper.*, 2012, pp. 109–114.
[40] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT-500.11, Geneva, Switzerland, 2002.
[41] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.1438, Geneva, Switzerland, 2000.
[42] X. Y. Zeng, "Techniques of experimental design for E-prime," (in Chinese), 2014.
[43] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3218–3231, Oct. 2015.
[44] K. Gu *et al.*, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.
[45] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks, "The zone of comfort: Predicting visual discomfort with stereo displays," *J. Vis.*, vol. 11, no. 8, pp. 74–76, 2011.
[46] J. E. Cutting, K. L. Brunick, J. E. DeLong, C. Iricinschi, and A. Candan, "Quicker, faster, darker: Changes in Hollywood film over 75 years," *i-Perception*, vol. 2, no. 6, pp. 569–576, 2011.
[47] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot detection and condensed representation. A review," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 28–37, Mar. 2006.
[48] P. P. Mohanta, S. K. Saha, and B. Chanda, "A model-based shot boundary detection technique using frame transition parameters," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 223–233, Feb. 2012.
[49] J. Park, H. Oh, and S. Lee. *IEEE-SA Stereo Image Database*. Accessed on Jun. 2015. [Online]. Available: http://grouper.ieee.org/groups/3dhf/

Guanghui Yue received the B.S. degree in communication engineering from Tianjin University, Tianjin, China, in 2014, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering. His research interests include bioelectrical signal processing, image quality assessment, and 3D image visual discomfort prediction.

Chunping Hou received the M.Eng. and Ph.D. degrees in electronic engineering from Tianjin University, Tianjin, China, in 1986 and 1998, respectively.

Since 1986, she has been the Faculty of the School of Electronic and Information Engineering, Tianjin University, where she is currently a Full Professor and the Director of the Broadband Wireless Communications and 3D Imaging Institute. Her current research interests include 3D image processing, 3D display, wireless communication, and the design and applications of communication systems.

Jianjun Lei (M'11) received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2007.

He was a Visiting Researcher with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, from 2012 to 2013. He is currently a Professor with Tianjin University, Tianjin, China. His research interests include 3D video processing, 3D display, and computer vision.

Yuming Fang (M'13–SM'17) received the B.E. degree from Sichuan University, Chengdu, China; the M.S. degree from Beijing University of Technology, Beijing, China; and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently a Professor with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He has authored or co-authored over 90 academic papers in international journals and conferences in multimedia processing. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, and 3D image/video processing. He serves as an Associate Editor of IEEE ACCESS. He is on the Editorial Board of Signal Processing: Image Communication.

Weisi Lin (F'16) received the Ph.D. degree from King's College London, London, U.K. He is currently an Associate Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include image processing, visual quality evaluation, and perception-inspired signal modeling, with over 340 refereed papers published in international journals and conferences. He is a fellow of the Institution of Engineering Technology and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He has been elected as an APSIPA Distinguished Lecturer from 2012 to 2013. He served as a Technical-Program Chair at the Pacific-Rim Conference on Multimedia 2012, the IEEE International Conference on Multimedia and Expo 2013, and the International Workshop on Quality of Multimedia Experience 2014. He has been on the Editorial Board of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA from 2011 to 2013, IEEE SIGNAL PROCESSING LETTERS, and *Journal of Visual Communication and Image Representation*.