Subjective evaluation of two stereoscopic imaging systems exploiting visual attention to improve 3D quality of experience

Philippe Hanhart and Touradj Ebrahimi

Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

ABSTRACT

Crosstalk and vergence-accommodation rivalry negatively impact the quality of experience (QoE) provided by stereoscopic displays. However, exploiting visual attention and adapting the 3D rendering process on the fly can reduce these drawbacks. In this paper, we propose and evaluate two different approaches that exploit visual attention to improve 3D QoE on stereoscopic displays: an offline system, which uses a saliency map to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions. The gaze points were used in conjunction with the disparity map to extract the disparity of the object-of-interest. Horizontal image translation was performed to bring the fixated object on the screen plane. The user preference between standard 3D mode and the two proposed systems was evaluated through a subjective evaluation. Results show that exploiting visual attention significantly improves image quality and visual comfort, with a slight advantage for real time gaze determination. Depth quality is also improved, but the difference is not significant.

Keywords: 3D, stereoscopic display, visual attention, saliency map, eye tracking, quality of experience, subjective quality assessment

1. INTRODUCTION

With the recent advances in 3D imaging technologies, 3D television (3DTV) has gained in popularity. 3DTV is expected to provide an enhanced multimedia experience when compared to 2D television. In particular, 3D displays support binocular depth cues, which provide a more immersive and realistic viewing experience. However, visual impairments might be produced by the 3D display, which reduce the added value of the depth dimension, decrease picture quality, and cause adverse problems, such as visual discomfort. One of the most annoying distortions in stereoscopic displays is due to imperfect separation between the left and right views of the stereo pair.[1] This impairment is referred to as crosstalk and is one of the main perceptual factors contributing to image quality and visual comfort.[2] Another issue is the unnatural decoupling of vergence and accommodation: when watching a stereoscopic display, the eyes converge to the location of the virtual object while the accommodation is always set for the screen surface. This effect is referred to as vergence-accommodation rivalry and is believed to increase visual discomfort.[3]

To improve the quality of experience (QoE) provided by stereoscopic displays, researchers have proposed to exploit visual attention.[4] Since two decades, researchers have investigated different solutions based on visual attention to improve the rendering of stereoscopic displays. Several systems have been developed using gaze tracking to determine the virtual object fixated by the user. The proposed solutions have been implemented either by hardware means, to control the stereoscopic display,[5,6] or by software means, to adapt the rendering of the 3D content.[7,8]

To reduce the vergence-accommodation rivalry, the 3D content should be reconverged such that the fixated object lies on the screen plane. This can be achieved by performing view synthesis to generate a new stereo pair, which requires depth information and typically introduces visible artifacts due to imperfect depth data, or simply by applying a horizontal image translation, i.e., shifting horizontally the left and right views of the original stereo pair.[9] By bringing the fixated object on the screen plane, perceived crosstalk is also reduced

Further author information: (Send correspondence to Philippe Hanhart) E-mail: {firstname.lastname}@epfl.ch

as the virtual object is translated to the zero disparity plane (ZDP). Xu *et al.*[10] have shown that horizontal parallax adjustment significantly increases the overall 3D QoE.

Chamaret *et al.*[11] have proposed a system to adapt the 3D rendering based on the region-of-interest. To predict the most salient region, a visual attention model was used to compute the saliency map considering spatial, temporal, and depth features. Then, the most relevant disparity of the region-of-interest was extracted. Finally, a shift was applied to the left and right views of the stereo pair, based on filtered disparity values, to translate the region-of-interest to the screen plane. They have reported that the proposed system was more pleasant to watch than the original video sequence, but no results of a proper subjective evaluation are reported to support their claim.

In this paper, we propose and evaluate two different approaches that exploit visual attention to improve 3D QoE on stereoscopic displays: an offline system, which uses a saliency map to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions. The saliency map based system relies on a 3D visual attention model proposed by Wang *et al.*[12] From the computed saliency map, the region-of-interest and its disparity were extracted similarly to the approach used by Chamaret *et al.*[11] The eye tracking based system relies on a Smart Eye Pro 5.8 remote eye tracking system to determine the gaze points of the left and right eyes independently. The gaze points were filtered taking into account the time required to fuse two images. The filtered gaze points were used in conjunction with the disparity map to extract the disparity of the object-of-interest. In both systems, horizontal image translation was performed to bring the fixated object on the ZDP. The shift was determined based on the extracted disparity values and filtered in time to have smooth transitions that do no create visual discomfort. The user preference between standard 3D mode and the two proposed systems was evaluated in terms of image quality, depth quality, and visual discomfort through a subjective evaluation. The paired comparison methodology was selected for its improved simplicity and reliability in assessing 3D quality.[13] Eight stereoscopic video sequences with various characteristics were used in the evaluation. A total of 21 subjects took part in the experiments. Results show that exploiting visual attention significantly improves image quality and visual comfort, with a slight advantage for real time gaze determination. Depth quality is also improved, but the difference is not significant.

The remainder of the paper is organized as follows. The saliency map and eye tracking based systems are described in Sec. 2. In Sec. 3, the methodology used to evaluate the performance of the different systems is described. Results are presented and analyzed in Sec. 4. Finally, concluding remarks are given in Sec. 5.

2. SYSTEM DESCRIPTION AND IMPLEMENTATION

This section describes the saliency map computation and most salient disparity extraction used in the saliency map based system, as well as the gaze points filtering and disparity extraction used in the eye tracking based system. Details are provided regarding the shift filtering, horizontal image translation used to reconverge the 3D scene, and implementation of the application.

2.1 Saliency map based system

One of the approaches considered in this paper to improve 3D QoE relies on a visual attention model to determine the most salient region-of-interest and its corresponding disparity. Since we are considering stereoscopic video sequences, the following features should be exploited to compute the saliency map: spatial, temporal, and depth features.

2.1.1 Saliency map computation

To compute the spatial and temporal saliency maps, the Graph-Based Visual Saliency[14] algorithm was used as it is reported to be one of the best algorithms according to Ref. 15, it considers temporal features, and a MATLAB implementation is publicly available.* The algorithm was applied twice on the left view of the stereo pair: once to compute the spatial saliency map using Derrington-Krauskopf-Lennie (DKL) color space, intensity, and orientation features; once to compute the temporal saliency map using motion features only.

*MATLAB implementation available at `http://www.klab.caltech.edu/~harel/share/gbvs.php`

To compute depth saliency, a few models have been proposed in the recent years, but there is no publicly available implementation. Readers can refer to Ref. 12 for a recent review of some of these models. Most visual attention models considering depth features perform a simple weighting of the 2D saliency map with the depth map, based on the assumption that pixels located closer to the observers and in front of the screen are more salient. However, Wang *et al.*[12] have shown that combining 2D saliency and depth contrast provides better results. Therefore, in this paper, depth contrast was used to compute the depth saliency map. To compute depth contrast, a difference of gaussians (DoG) filter was applied to the left depth map.[12]

First, the perceived depth map, which represents the distance between the observer and the virtual object, was computed from the left disparity map considering viewing conditions.[12] The relationship between the perceived depth D in meters and the disparity d in pixel is given by

$$D = \frac{V}{1 + \frac{d \cdot W}{I \cdot R_x}} \tag{1}$$

where I is the interocular distance, V is the viewing distance, and W and R_x are the width and horizontal resolution of the screen, respectively. In our experiments, the interocular distance was set to 65 mm and the screen property parameters were set according to the setup of the subjective evaluation (see Sec. 3.2).

Then, the depth contrast was computed by filtering the perceived depth map with the DoG filter defined as

$$f(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) - \frac{1}{2\pi K^2\sigma^2} \exp\left(-\frac{x^2+y^2}{2K^2\sigma^2}\right) \tag{2}$$

where (x, y) is the location in the filter, σ and K are used to control the scales of DoG and the ratio between the "center" area and "surround" area, respectively. According to Ref. 12, only one scale of DoG was applied, with $\sigma = 32$ pixels and $K = 1.6$. Finally, the depth saliency map was computed as the absolute value of the depth contrast, normalized by its maximum value.

Similarly to the straightforward approach used by Wang *et al.*,[12] the final saliency map SM was computed as a weighted sum of the spatial saliency map SM_s, temporal saliency map SM_t, and depth saliency map SM_d

$$SM(i,j) = \omega_1 SM_s + \omega_2 SM_t + \omega_3 SM_d \tag{3}$$

where $\omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$.

2.1.2 Most salient disparity extraction

To extract the most salient region-of-interest and to determine its most relevant disparity, a similar approach to that presented by Chamaret *et al.*[11] was used. First, the N most salient pixels were determined using the saliency map computed as described in Sec. 2.1.1. In our experiments, N was empirically set to 2% of the total number of pixels. Then, a connected-component analysis, considering an 8-connected neighborhood, was performed to determine the M regions formed from the N most salient pixels. Finally, the region containing the highest number of most salient pixels was considered as the most salient region-of-interest.

Once the most salient region-of-interest was determined, the most relevant disparity of this area was extracted. First, a histogram was constructed from the disparity values inside the most salient region-of-interest to estimate the spreading of disparity values. A step of 5 disparity units was used to construct the histogram. Then, the bin with the highest frequency count was extracted from the histogram. Finally, the median disparity of this bin was considered as the most salient disparity.

2.2 Eye tracking based system

The second approach considered in this paper to improve 3D QoE relies on an eye tracking system to determine the gaze positions on the screen, and therefore, the object-of-interest watched by the viewer and its corresponding disparity.

The left and right gaze points were computed independently using the gaze origin and gaze direction measured by the eye tracking system and the position of the eye tracker with respect to the screen. Additionally, the gaze direction quality estimated by the eye tracking system was used to discard low quality measurements. In our experiments, if the gaze direction quality was below 0.1, the gaze point was discarded. This process was performed independently for the left and right eyes.

The eye tracking system used in our experiments (see Sec. 3.2) provides both unfiltered and filtered versions of the gaze origin and gaze direction measures. In our experiments, the unfiltered measurements were used as no details are provided regarding the filtering technique applied by the eye tracking system. However, to discard erroneous measurements, a median filter was applied separately on the x- and y-components of the left and right gaze points. To provide robust but reactive gaze positions, a tradeoff has to be made on the size of the kernel. The delay introduced by the filter should be lower than the minimum time required to fuse stereoscopic stimuli, which is around 400 to 500 ms.[3] In our experiments, a kernel of size $L = 13$ was used, which corresponds to a delay of approximately $\frac{1}{r} \cdot \frac{L+1}{2} \approx 120$ ms, where r is the sample rate of the eye tracking system (60 fps in our experiments).

If the gaze points of the left and right eyes could be determined with 100% accuracy, the disparity of the virtual object would be simply given by the difference of the x-coordinates of the left and right gaze points. However, due to the limited precision of the eye tracking system, such a trivial approach would not provide good results in practice. Therefore, the left and right disparity maps of the stereoscopic video sequence were used in conjunction with the filtered gaze points. For the same reason, the shift cannot simply be determined by using the disparity values at the gaze points. For example, if the gaze position is near the boundary between foreground and background, the measured gaze point might be located on the background, whereas the actual gaze position was located on the foreground. Based on the assumption that foreground objects are more salient than background objects, the disparity of the object-of-interest was determined as the maximum disparity in a neighborhood of $N \times N$ pixels around the filtered gaze point. In our experiments, N was empirically set to 15 pixels. This process was performed independently for the left and right eyes and the most salient disparity was computed as the maximum value of the disparity values extracted from the left and right disparity maps.

2.3 Shifting

To bring the fixated object on the screen plane, horizontal image translation was performed by shifting horizontally the left and right views of the original stereo pair.[9] The shift parameter is given by the disparity value of the fixated object, which was determined by the saliency map or eye tracking system. However, if the shift difference between two successive frames is too large or oscillates at a high frequency, the process may become visible and induce visual discomfort. Therefore, filtering of the shift values was performed.

2.3.1 Shift filtering

Chamaret *et al.*[11] have determined that a maximum shift difference of 1.5 pixel can be applied without any visual notification. However, a non-integer shift requires interpolation of the pixel values, which is time consuming and not suitable for a real time application. Therefore, in our experiments, only integer shifts were applied and the maximum shift difference between two successive frames was set to 1 pixel.

To improve robustness, the disparity values determined by the saliency map and eye tracking system were filtered using a median filter

$$d_f(n) = \text{median}\left(d(n-K), \ldots, d(n)\right) \tag{4}$$

where d and d_f are the raw and filtered disparity values, respectively, and n is the frame index. In our experiments, K was empirically set to 4.

The target shift value s_t is given by the filtered disparity value

$$s_t(n) = d_f(n) \tag{5}$$

Then, the current and future shift values are determined according to Algorithm 1, where s is the shift, F is the number of frames of the video sequence, and n is the frame index. To avoid flickering due to the shifting process, a threshold δ was considered before updating the current and future shift values. In our experiments, δ was empirically set to 1.

Algorithm 1 Determination of current and future shift values

```
if $|s_t(n) - s(F)| > \delta$ then
  if $|s_t(n) - s(n-1)| \leq \delta$ then
    $s(n, \ldots, F) \leftarrow s(n-1)$
  else
    $k \leftarrow 0$
    $shift \leftarrow s(n-1)$
    if $s_t(n) > s(n-1)$ then
      while $shift < s_t(n)$ do
        $shift \leftarrow shift + 1$
        $s(n+k) \leftarrow shift$
        $k \leftarrow k + 1$
      end while
    else
      while $shift > s_t(n)$ do
        $shift \leftarrow shift - 1$
        $s(n+k) \leftarrow shift$
        $k \leftarrow k + 1$
      end while
    end if
    $s(n+k, \ldots, F) \leftarrow s_t(n)$
  end if
end if
```

2.3.2 Horizontal image translation

Both views were shifted half way

$$s_l = \left\lceil \frac{s}{2} \right\rceil, s_r = \left\lfloor \frac{s}{2} \right\rfloor \tag{6}$$

where s_l and s_r are the shift parameters for the left and right views, respectively. The horizontal image translation was performed by adding black borders on the left or right side of the picture and cropping the other side by the same amount to preserve the size of the picture.[9] To reduce potential stereoscopic window violation that might occur due to horizontal image translation, the floating window[9] technique was applied as follows: the pixel positions corresponding to the black border in the left view were also set to black in the right view, and vice versa. Therefore, both pictures had black borders on both sides.

2.4 Implementation

For our experiments, a dedicated video player was implemented in C++ using the OpenCV library. The application implements the saliency map and eye tracking systems described in Sec. 2.1 and 2.2, respectively. The video player displays a stereoscopic video sequence and performs horizontal image translation on the fly. To ensure real time processing, the application uses multithreading, which was implemented using the Boost library. In particular, one thread per input data, i.e., left view, right view, left disparity map, and right disparity map, was launched to load the data. One thread was dedicated to the eye tracking system, to collect the measurements that were sent by UDP from a remote computer and to compute the gaze points. Finally, one thread was dedicated to the remaining processes, such as horizontal image translation, interlacing (the stereoscopic monitor used in the experiments was line-interleaved), synchronization, filtering, logging, etc.

For each stereoscopic video sequence, the left and right views were converted to RGB 4:4:4, downsampled vertically by two (due to the line-interleaved monitor), and stored as a stack of bitmap image files. The conversion to RGB 4:4:4 allows shifting by odd values and is necessary for the interleaving process. The downsampling process avoids aliasing, which could occur when the interleaving is done by taking every other line without any pre-filtering. The depth maps were stored in half resolution as a stack of monochrome bitmap image files. This solution was found to provide the best results in terms of loading time and quality when compared to other

Table 1: Stereoscopic video sequences used in the experiments.

Sequence	Frames	Views	$d_{\min}$	$d_{\max}$
Boxers	1-250	0-1	-14	29
Hall	1-250	0-1	-15	20
Lab	151-400	0-1	-100	44
News report	1-250	0-1	-45	71
Phone call	151-400	0-1	-35	39
Musicians	1-250	0-1	0	176
Poker	1-250	0-1	0	176
Poznan Hall2	1-200	7-6	16	118

alternatives, such as using compressed video sequences for each stream. With this solution, a frame rate of 30 fps was achieved with the developed video player.

3. SUBJECTIVE EVALUATION

This section presents the details of the subjective evaluation conducted to assess user preference between standard 3D mode and the two proposed systems.

3.1 Dataset

Eight stereoscopic video sequences with associated depth maps were used in the experiments (see Table 1). Sequences *Boxers*, *Hall*, *Lab*, *News report*, and *Phone call* were obtained from the NAMA3DS1 database†.[16] These sequences are available as raw video files, progressively scanned, with YCbCr 4:2:2 color sampling, and 8 bit per sample. Additionally, the left disparity map is available as 16 bit floating values, half resolution. The right disparity map was generated by warping the left disparity map to the right view, filling holes using background propagation, and applying 3 × 3 median filtering. Both disparity maps were converted to 8 bit and scaled to the range [0, 255] using the minimum and maximum disparity values observed in the whole video (see Table 1). Sequences *Musicians* and *Poker* were obtained from the European FP7 Research Project MUSCADE‡.[17] Sequence *Poznan Hall2* was obtained from the Poznań multiview video database.[18] These video sequences and associated depth maps are available as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bit per sample. The camera parameters are provided to convert the disparity values scaled in the range [0, 255] to real disparity values in pixels.

Since the cameras used for recording sequences *Musicians*, *Poker*, and *Poznan Hall2* were set in a parallel direction, they are assumed to converge at an infinite point. This setup leads to stereoscopic window violation[9] and does not sufficiently exploits the depth range, as the 3D content appears only in front of the screen plane. Therefore, when displaying the original version of these three sequences, horizontal image translation was applied with a shift defined as[12]

$$s = \frac{d_{\min} - d_{\max}}{2} \tag{7}$$

where $d_{\min}$ and $d_{\max}$ are the minimum and maximum disparity values computed from the camera parameters (see Table 1), respectively. Note that no shift was applied to the sequences from the NAMA3DS1 database, as the cameras were already converged during recording.

3.2 Test environment

The experiments were conducted at the MMSPG quality test laboratory, which fulfills the recommendations for the subjective evaluation of visual data issued by ITU-R.[19] The test room was equipped with a controlled lighting system with a 6500K color temperature and an ambient luminance at 15% of the maximum screen luminance, whereas the color of all the background walls and curtains present in the test area were in mid grey. The test room was separated in two by a curtain to isolate the subject and equipment from the test operator, which was

† `http://www.irccyn.ec-nantes.fr/spip.php?article954`

‡ MUltimedia SCAlable 3D for Europe, `http://www.muscade.eu`, grant agreement n°247010

present during the test session to supervise the recording of the eye tracking data. The laboratory setup was intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors.

To display the test stimuli, a full HD 46” Hyundai S465D polarized stereoscopic monitor was used. The monitor was calibrated using an X-Rite i1Display Pro color calibration device according to the following profile: sRGB gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level.

A Smart Eye Pro 5.8 remote eye tracking system was used to determine the gaze position on the screen of the left and right eyes independently. The system was equipped with three Sony HR-50 cameras at a frame rate of 60 fps and two infrared flashes. All measurements were recorded on a separate computer and sent by UDP on a local network.

The experiment involved one subject per test session. The subjects were seated in line with the center of the monitor, at a distance of 3.2 times the picture height, corresponding to roughly 1.8 meters from the stereoscopic monitor, as suggested in recommendation ITU-R BT.2021.[20] The eye tracking system was placed at 1.28 meters from the stereoscopic monitor such that the face was well captured by the cameras.

3.3 Test methodology

The paired comparison methodology[20] was chosen as judging the quality of different imaging systems individually may be quite difficult. Pairs of video sequences, “A” and “B”, which resulted from different imaging systems, were presented in succession order on the same display. Subjects were asked to judge which video sequence in a pair (“A” or “B”) is preferred in terms of picture quality, depth quality, and visual comfort.[20] The option “same” was also included to avoid random preference selections. For each of the 8 test sequences, all the possible combinations of the 3 test conditions (original, saliency map, and eye tracking) were considered, leading to a total of $8 \times \binom{3}{2} = 24$ paired comparisons.

3.4 Training session

Before the experiment, oral instructions were provided to the subjects to explain their tasks. Additionally, a training session was organized to allow subjects to familiarize with the assessment procedure. The training materials shown in the training session were selected by expert viewers to include examples of all evaluated aspects. More particularly, the *Umbrella* and *Basket* sequences[16] were used with different horizontal image translations to illustrate picture quality. To illustrate depth quality, the *Soccer* sequence[16] was shown in 2D and 3D viewing conditions. Finally, from the EPFL 3D video database[§],[21] the *Notebook* sequence, with 10 cm and 20 cm baselines, was used to illustrate visual comfort. The training materials were presented to subjects exactly as for the test materials.

3.5 Test session

Before the test session, the aperture and focus settings of the eye tracker cameras were adjusted for optimal conditions and a full camera calibration was performed to maximize the accuracy of the measurements. For each subject, a personal profile was created by recording several head poses and gaze calibration using four calibration points close to the screen corners. Subjects were instructed to hold their head still while watching the video sequences to ensure good tracking of their gaze.

A trial was initiated by the presentation of a message showing the letter “A”, at zero disparity and with a mid-grey background, for 2 s. Then, the sequences to be compared were presented. The sequences were temporally separated by the presentation of a message showing the letter “B” for 2 s. The trial ended with a message showing the words “Vote now” without any restriction in time. Votes were collected by the test operator such that the subject was always seated optimally with respect to the eye tracking system.

Two dummy pairs, whose scores were not included in the results, were included at the beginning of the session to stabilize the subjects’ ratings. To reduce contextual effects, the stimuli orders of display, both within and between trials, were randomized applying different permutation for each subject, whereas the same content was never shown consecutively. In total, the test session lasted for about 16 minutes.

[§] `http://mmspg.epfl.ch/3dvqa`

Twenty-one naive subjects (five females and sixteen males) took part in the experiment. They were between 18 and 31 years old with an average of 21.8 years of age. All subjects were screened for correct visual acuity, color vision, and stereo vision using Snellen chart, Ishiara chart, and Randot test, respectively.

3.6 Analysis of the results

To analyze user preference for the different imaging systems, statistical tools were applied to the individual ratings. First, the winning frequency w_{ij} of stimulus i against stimulus j and tie frequency t_{ij} between the two stimuli were computed from the individual ratings. Note that $t_{ij} = t_{ji}$ and $w_{ij} + w_{ji} + t_{ij} = N$, where N is the number of subjects. This can be done individually for each test sequence or jointly over all test sequences. Then, a preference matrix was constructed from the individual paired comparisons. To compute the preference matrix, only wins were taken into account, whereas ties were discarded. The preference matrix provides the empirical probability P_{ij} of choosing stimulus i over stimulus j, where $P_{ij} = \frac{w_{ij}}{N}$, and i and j are the row and column of the matrix, respectively.

Finally, the Bradley-Terry-Luce model[22] was used to convert the winning frequencies to continuous-scale quality scores, which are equivalent to mean opinion scores (MOS). In this model, the empirical probability P_{ij} of choosing stimulus i is defined as

$$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \tag{8}$$

where π_i satisfying $\pi_i \geq 0$ and $\sum\limits_i \pi_i = 1$ can be considered as the quality score for stimulus i and can be obtained via maximum likelihood estimation. Ties were considered as half way between the two preference options and equally distributed between P_{ij} and P_{ji}.[22] The confidence intervals (CI) for the maximum likelihood estimates of the scores were obtained using the Hessian matrix of the log-likelihood function. The results were normalized to the range $[0, 100]$ for a better representation.

4. RESULTS AND DISCUSSION

Figures 1 and 2 show the preference probabilities and preference matrices over all test sequences, respectively. Results for picture quality, depth quality, and visual comfort are provided individually. As it can be observed, the visual attention based systems significantly improve picture quality when compared to the original video sequence, as they are preferred in about 65% of the test stimuli, whereas the original video sequence is preferred in only about 20% of the test stimuli. The eye tracking and saliency map based systems are quite competitive, with a preference probability of 40% and 33%, respectively, which shows a slight advantage for real time gaze determination.

Similar results can be observed for visual comfort. However, the preference probabilities for the visual attention based systems over the original video sequence have decreased by about 6%, whereas the ties have increased, especially for the original versus saliency map paired comparison. Regarding the eye tracking versus saliency map paired comparison, the probability of choosing the saliency map based system has decreased to 25%, whereas the probability of choosing the eye tracking based system is about 42%, which means that the difference between the two systems is slightly bigger in terms of visual comfort than in terms of picture quality.

Regarding depth quality, the preference between the different conditions is not obvious, as the preference probabilities are between 29% and 47%, which is close to the random 33%. However, the results show a slight preference for the two visual attention based systems, with a preference probability for the eye tracking and saliency map based systems over the original video of 43% and 47%, respectively.

Results for picture quality and visual comfort show similar behavior, which was expected as the individual ratings for these two aspects were usually highly correlated. The link between picture quality and visual comfort is quite intuitive, since objects located far in front or behind the screen plane may induce vergence-accommodation rivalry, which creates visual discomfort, and are perceived with double edges due to imperfect separation of the left and right views, which reduces picture quality. However, for some of the subjects, the depth quality ratings

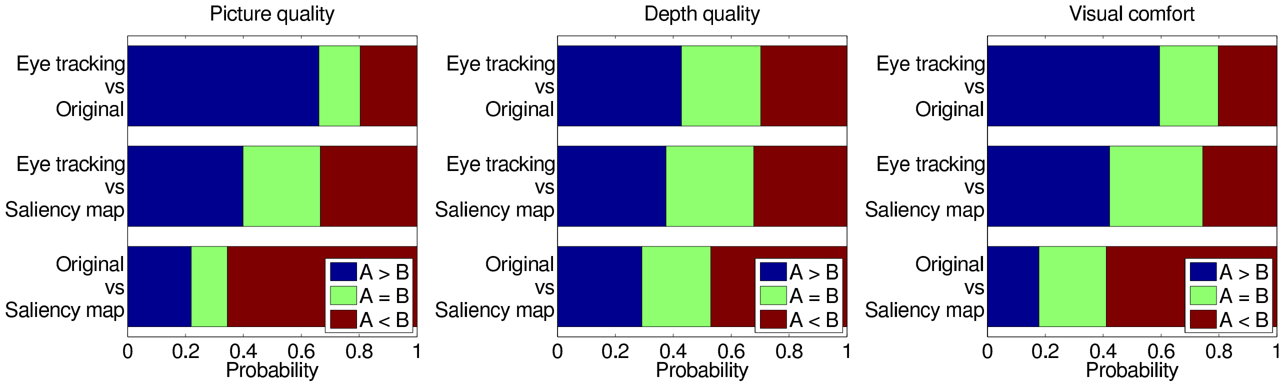

Figure 1: Preference and tie probabilities.

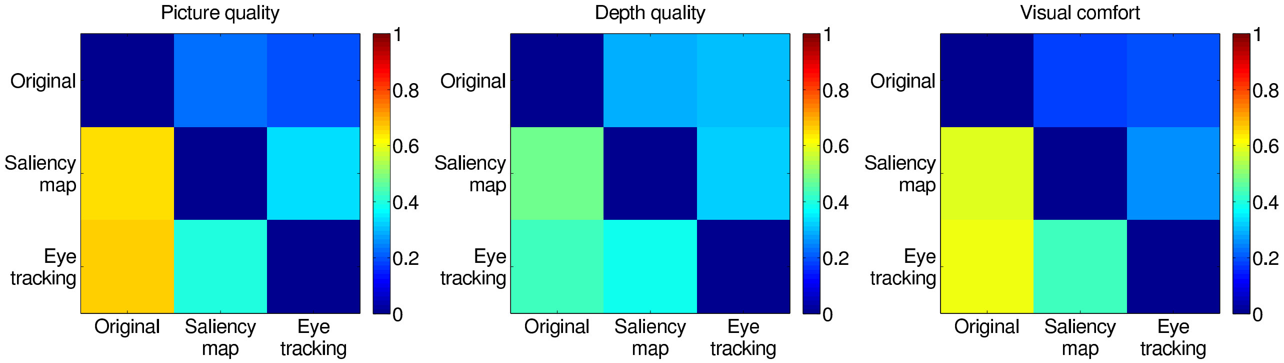

Figure 2: Preference probability of choosing stimulus i (row) over stimulus j (column).

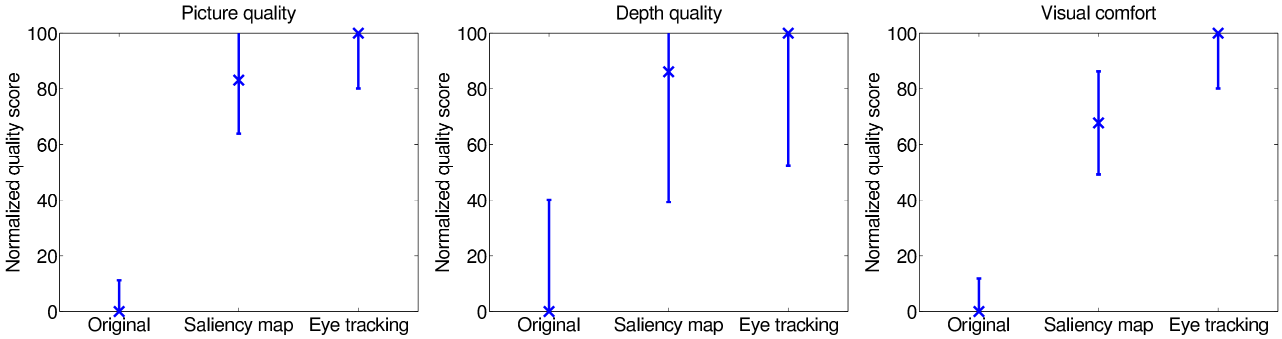

Figure 3: Normalized quality scores (MOS and CI).

usually agreed with the picture quality and visual comfort ratings, whereas they were opposed for the other subjects. This means that horizontal image translation had a positive impact on depth quality for some of the subjects, whereas it had a negative impact for others. As the 3D picture was shifted to reconverge the scene, not recalculated, this process translated the whole 3D scene along the z-axis, which could induce a scale-down effect[9] and negatively impact depth perception. To overcome this drawback, a new stereo pair should be synthesized, with different camera parameters, but this process cannot be performed in real time. Ideally, depth blur should be considered as well to strengthen monocular depth cues.

Figure 3 shows the obtained MOS and CI over all test sequences. Results for picture quality, depth quality, and visual comfort are provided individually. As it can be observed, the visual attention based systems significantly enhance picture quality and visual comfort. However, the difference between the saliency map and eye tracking based systems is not significant, as expected by largely overlapping CIs. For some of the test sequences, the most salient region-of-interest predicted using the saliency map highly correlated with the object-of-interest watched by the subject. However, assuming that the eye tracking system can accurately measure real time gaze position, this system is expected to perform better than the saliency map system, which relies on an algorithm to predict visual saliency. Considering that the saliency map system does not require a specific hardware setup and can be performed off line, which would allow more advanced techniques than horizontal image translation to reconverge the 3D picture, this system shows promising results. In a future study, we plan to benchmark the performance of different 3D saliency models using the ground truth eye tracking data recorded in our experiments.

Regarding depth quality, the visual attention based systems provide slightly better results when compared to the original video sequence, but the difference is not as significant as for picture quality and visual comfort. However, the difference between the eye tracking based system and original sequence can be expected as statistically significant, as the CIs do not overlap. To further investigate the influence on depth quality, additional statistical analysis should be performed, considering ties to determine the confidence intervals.

5. CONCLUSION

In this paper, we proposed and evaluated two different approaches that exploit visual attention to improve 3D QoE on stereoscopic displays: an offline system, which uses a saliency map to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions. From the saliency map, which was computed using a 3D visual attention model, the region-of-interest and its disparity were extracted. From the eye tracking measurements, filtered gaze points were used in conjunction with the disparity map to extract the disparity of the object-of-interest. Horizontal image translation was performed to bring the fixated object on the screen plane. The shift was determined based on the extracted disparity values and filtered in time to have smooth transitions that do no create visual discomfort. The user preference between standard 3D mode and the two proposed systems was evaluated in terms of image quality, depth quality, and visual discomfort through a subjective evaluation. Results show that exploiting visual attention significantly improves image quality and visual comfort, with a slight advantage for real time gaze determination. Depth quality is also improved, but the difference is not significant.

ACKNOWLEDGMENTS

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1) and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET. Special thanks to Martin Rerabek for interesting discussions on the concept of using an eye tracking system to improve 3D QoE.

REFERENCES

[1] Seuntiëns, P., Meesters, L., and IJsselsteijn, W., "Perceptual attributes of crosstalk in 3D images," *Displays* **26**, 177–183 (October 2005).

[2] Meesters, L., IJsselsteijn, W., and Seuntiens, P., "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology* **14**, 381–391 (March 2004).

[3] Hoffman, D. M., Girshick, A. R., Akeley, K., and Banks, M. S., "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision* **8** (March 2008).

[4] Huynh-Thu, Q., Barkowsky, M., and Le Callet, P., "The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives," *IEEE Transactions on Broadcasting* **57**, 421–431 (June 2011).

[5] Shiwa, S., Omura, K., and Kishino, F., "Proposal for a 3-D display with accommodative compensation: 3DDAC," *Journal of the Society for Information Display* **4**, 255–261 (December 1996).

[6] Talmi, K. and Liu, J., "Eye and gaze tracking for visually controlled interactive stereoscopic displays," *Signal Processing: Image Communication* **14**, 799–810 (August 1999).

[7] Peli, E., Hedges, T. R., Tang, J., and Landmann, D., "A Binocular Stereoscopic Display System with Coupled Convergence and Accommodation Demands," *SID Symposium Digest of Technical Papers* **32**, 1296–1299 (June 2001).

[8] Yang, R. and Zhang, Z., "Eye gaze correction with stereovision for video-teleconferencing," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 956–960 (July 2004).

[9] Mendiburu, B., [*3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*], Focal Press/Elsevier (2009).

[10] Xu, D., Coria, L., and Nasiopoulos, P., "Quality of experience for the horizontal pixel parallax adjustment of stereoscopic 3D videos," in [*IEEE International Conference on Consumer Electronics (ICCE)*], 394–395 (January 2012).

[11] Chamaret, C., Godeffroy, S., Lopez, P., and Le Meur, O., "Adaptive 3D rendering based on region-of-interest," in [*Stereoscopic Displays and Applications XXI*], *Proc. SPIE* **7524** (February 2010).

[12] Wang, J., Da Silva, M., Le Callet, P., and Ricordel, V., "Computational Model of Stereoscopic 3D Visual Saliency," *IEEE Transactions on Image Processing* **22**, 2151–2165 (June 2013).

[13] Lee, J.-S., Goldmann, L., and Ebrahimi, T., "Paired comparison-based subjective quality assessment of stereoscopic images," *Multimedia Tools and Applications* **67**, 31–48 (November 2013).

[14] Harel, J., Koch, C., and Perona, P., "Graph-Based Visual Saliency," in [*Proceedings of Neural Information Processing Systems (NIPS)*], (2006).

[15] Judd, T., Durand, F., and Torralba, A., "A Benchmark of Computational Models of Saliency to Predict Human Fixations," Tech. Rep. MIT-CSAIL-TR-2012-001, MIT Computer Science and Artificial Intelligence Laboratory (January 2012).

[16] Urvoy, M., Barkowsky, M., Cousseau, R., Koudota, Y., Ricorde, V., Le Callet, P., Gutierrez, J., and Garcia, N., "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in [*Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*], 109–114 (July 2012).

[17] ISO/IEC JTC1/SC29/WG11, "Proposed Stereo Test Sequences for 3D Video Coding." Doc. M23703, San Jose, USA (February 2012).

[18] ISO/IEC JTC1/SC29/WG11, "Poznań Multiview Video Test Sequences and Camera Parameters." Doc. M17050, Xian, Chine (October 2009).

[19] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures." International Telecommunication Union (January 2012).

[20] ITU-R BT.2021, "Subjective methods for the assessment of stereoscopic 3DTV systems." International Telecommunication Union (August 2012).

[21] Goldmann, L., De Simone, F., and Ebrahimi, T., "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in [*Three-Dimensional Image Processing (3DIP) and Applications*], *Proc. SPIE* **7526** (February 2010).

[22] Glickman, M. E., "Parameter estimation in large dynamic paired comparison experiments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **48**(3), 377–394 (1999).