INFLUENCE OF AUDIO-VISUAL ATTENTION ON PERCEIVED QUALITY OF STANDARD DEFINITION MULTIMEDIA CONTENT[1]

Jong-Seok Lee, Francesca De Simone, and Touradj Ebrahimi

Multimedia Signal Processing Group
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015, Lausanne, Switzerland
{jong-seok.lee, francesca.desimone, touradj.ebrahimi}@epfl.ch

ABSTRACT

When human subjects assess the quality of multimedia data, high level perceptual processes such as Focus of Attention (FoA) and eye movements are believed to play an important role in such tasks. While prior art reports incorporation of visual FoA into objective quality metrics, audio-visual FoA has been rarely addressed and utilized in spite of the importance and presence of both audio and video information in many multimedia systems. This paper explores the influence of audio-visual FoA in the perceived quality of standard definition audio-visual sequences. Results of a subjective quality assessment study are reported, where it is shown that the sound source attracts visual attention and thereby the visual degradation in the regions far from the source is less perceived when compared to sound-emitting regions.

Index Terms— quality assessment, audio-visual focus of attention, cross-modal interaction, perceived quality

1. INTRODUCTION

Research in objective visual quality assessment aims at developing quantitative measures which can automatically and reliably predict the quality of still images or image sequences, as perceived by a human observer. The objective assessment can be used to monitor visual quality in order to dynamically adjust it, to benchmark image/video processing systems and algorithms, or to optimize algorithms and their parameters setting. Although human beings can rather easily make judgments about quality of multimedia content, objective visual quality assessment is a very hard task and challenging research topic. This is due to the lack of a good understanding of underlying mechanisms in subjective evaluation of quality by human observers. Additionally, user's perception of quality varies depending on *a priori* expectations, and the specific application under consideration.

In order to develop objective models able to fulfill this difficult task, the study of the subjective perception of quality is fundamental. Many models of the early vision properties of the Human Visual System (HVS), like sensitivity to contrast changes rather than to luminance changes, varying sensitivity to stimuli at different spatial frequencies, and visual masking, have been developed in literature and included in the so-called HVS-based quality metrics [1]. On the other hand, subjective assessment of quality requires taking into account higher level perceptual processes, such as human visual Focus of Attention (FoA) and eye movements. It is known that the HVS is space-variant in that only a small region, the fovea, around a point of fixation is captured at a high spatial resolution and that the resolution for the peripheral regions of retina dramatically decreases with eccentricity. The location of the scene being projected into fovea is therefore updated by rapid eye movements, called saccades, and their positions are usually driven by the FoA [2].

The visual FoA has been utilized in various fields. In video coding, coding efficiency can be improved by discarding redundant information outside small expected fixation regions without degradation of perceived quality. The regions of attention can be identified based on conspicuity in terms of low level features, detection of moving objects, face detection, and so on [3,4]. The visual attention has been also exploited in computer graphics. Carter *et al.* demonstrated that, when subjects are concentrated on the assigned tasks, they consistently fail to notice quality degradations of image details even when these degradations occur within the subjects' gaze [5]. This "inattentional blindness" can be used for perceptual rendering with reduced computational complexity without degradation of perceived quality. Similar work can be found in [6].

Attempts have already been made to take into account the visual attention in the design of quality metrics. Lee *et al.*

[1] The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia), and the Swiss NCCR Interactive Multimodal Information Management (IM2).

developed an objective quality criterion, called foveal signal-to-noise ratio, by using the non-uniform resolution of the HVS [7]. Osberger *et al.* proposed a quality metric for the assessment of still pictures, which includes an attention model to weight the influence of visible errors produced by an early vision model of the HVS [8]. In [9], a quality metric which is based on object tracking and segmentation has been introduced and its validity was tested in the case of face segmentation. In [10], a quality metric is proposed for monitoring the quality of video sequences transferred over mobile networks for sign language conversations; since in this application face and hands are the areas on which the user's attention is mostly focused, the distortions in these regions are optimally weighted to create an objective intelligibility score for a distorted sequence.

However, the visual aspect of the stimuli in a video sequence is not the only modality which affects the humans' attention and perceived quality of the content. The acoustic modality and the audio-visual cross-modal interaction also play important roles in attention and quality perception.

In the field of rendering for virtual reality or computer graphics, it has been demonstrated that the characteristics of the audio-visual (AV) FoA can be used for high-fidelity rendering with reduced computational complexity. In [11], the authors have shown that rendering at high quality only the sound emitting object and lowering the frame rates for unattended regions can significantly reduce rendering time, without loss of perceived quality. A similar result is also presented in [12].

This paper investigates the effect of the AV FoA in the assessment of multimedia quality of experience, which has been rarely addressed previously. We assume that the auditory stimulus can guide the visual attention, which in turn affects the perceived quality of multimedia content. First, we give a review of the studies related to AV FoA and perceived quality. Then, we report the results of a subjective test performed with compressed video streams containing blurring artifacts. These results show that, except for extreme cases, the influence of the visual degradation outside the sound source is insignificant.

The rest of the paper is organized as follows. The following section presents backgrounds from diverse fields related to the present work. In Section 3 the methodology of the subjective test is explained. Section 4 shows the test results along with discussions. Finally, concluding remarks are given in Section 5.

2. AUDIO-VISUAL ATTENTION

In humans' attention both auditory and visual sensory modalities are often involved simultaneously and may influence each other. In particular, there are two distinct forms of attention: The "endogenous" attention is stimulus driven and captured reflexively by events in a bottom-up manner; the "exogenous" attention requires a conscious decision and is directed voluntarily in a top-down manner [13].

In exogenous spatial attention, it has been proved that a spatially non-predictive cue in one modality can attract covert attention toward the location of the cue in the other modality [14], which is called the "cross-modal facilitatory effect". For example, an abrupt sound draws visual attention to the spatial location of the sound source. In the experiments by Spence and Driver [14], subjects were asked to judge the elevation (up or down) of peripheral visual targets which follow uninformative auditory cues on either their left or their right side; the judgments were faster and/or more accurate for the visual targets preceded by the auditory cues occurring on the same side.

Similar cross-modal facilitatory effects are also observed in endogenous attention. Spence and Driver conducted similar experiments to those for exogenous attention [15]: The elevation judgments for either auditory or visual targets were faster when the subjects expected a stimulus of the other modality in that side. This proves that attending to stimuli of one modality at a given location enhances processing of stimuli of the other modality at the same spatial location.

In [16], it was shown that, even when people are performing a visual task, a novel auditory stimulus can capture their visual attention. Such cross-modal orienting is automatic in that it occurs even when detailed information about the target is given to the subjects in order to prevent uninformative auditory cues from orienting attention [17].

There are some other interesting phenomena related to the AV perception and attention. The "ventriloquist effect" refers to the illusion that, when synchronous auditory and visual information is presented in slightly separate locations, the perceived location of the sound is biased to the direction of the visual stimulus [18]. The "freezing phenomenon" shows that, when a rapidly changing visual stimulus is shown, an abrupt sound may freeze the display with which the sound is synchronized, i.e., it is perceived as if the display is shown for a longer time [19].

AV speech perception is an example of the advantage of exploiting the link between AV attention and perception. If one has a problem in listening to the spoken language due to the environmental noise, it is useful to observe the lip movements or the gesture to understand the speech better via integration of the acoustic and the visual stimuli [20]. On the contrary, discrepancy of the two modalities may result in an illusion due to their confliction in AV speech perception, as demonstrated by the McGurk effect [21].

Finally, a neurological analysis of the human brain also shows evidences of multimodal information processing. When different senses reach the brain, the sensory signals converge to the same area in the superior colliculus and a large portion of the neurons leaving the superior colliculus are multisensory [22]. Additionally, neuroimaging studies have shown that not only sensory-specific cortices

anatomically converge in to multisensory brain areas, but also multisensory spatial interactions conversely affect unimodal brains [23].

3. PROPOSED SUBJECTIVE STUDY

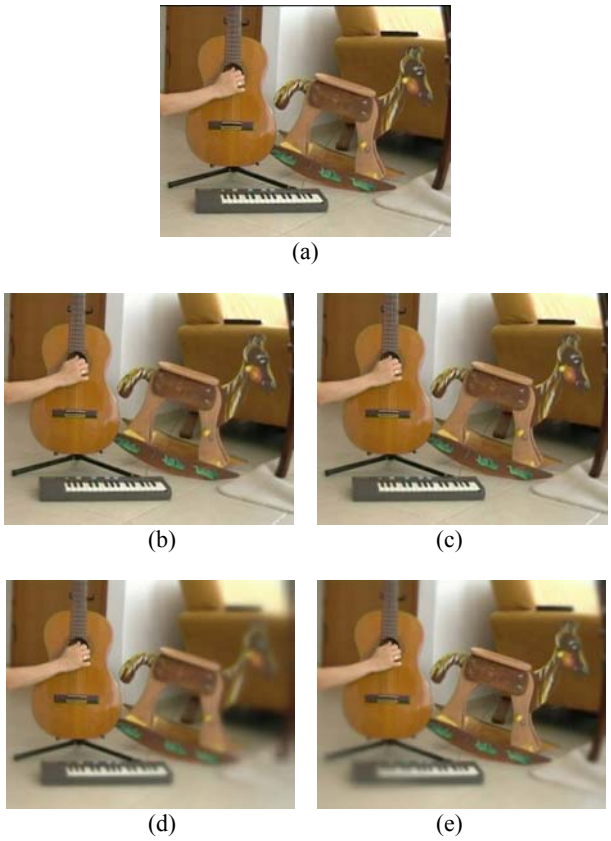
In order to test the influence of the AV FoA on the subjective perception of quality for a typical multimedia experience, for instance when viewing a video sequence including an audio channel, we designed and performed a subjective quality test with AV test material. In our experiment video sequences containing artifacts in different areas of the scene are shown to subjects. At the same time, the audio signal is listened to through headphones. The subject is asked to rate the overall quality of the test material. Since we assume that the presence of an auditory source guides the visual attention, a "masking effect" should be present for artifacts in areas far from the audio source. We want to prove whether this effect relevantly affects the perceived quality of the multimedia content.

3.1. Test material

We used standard resolution video sequences with a synchronous mono-channel audio. Artifacts have been introduced only on the visual channel by coding the video sequence using H.264/AVC at constant bitrate. The set of test sequences has been produced by coding three different versions of each content:

- the original content;
- a degraded version of the content produced by localizing the sound source in the scene and then blurring the remaining part;
- a degraded version of the content produced by localizing all the moving parts (including the sound source) in the scene and then blurring the remaining part.

In particular, the applied blur degradation varies according to a priority map. This map is based on the distance between each pixel and the nearest highest priority region which is the sound-emitting region in the first case and the moving objects in the second case. The blur degradation of the visual signal was performed through a Gaussian pyramid decomposition with L levels. A stronger blurring is obtained with a larger value of L. Each level of the pyramid is linearly assigned to one priority level, i.e., the highest level (the original image) is assigned to the highest priority and the lowest level to the lowest priority. Two different values of L are considered. Thus, a complete dataset of five realizations for each coding bitrate have been produced: no blurring (NB), blurring according to the identified sound source with L=2 (S2) and L=6 (S6), and blurring according to the identified moving objects with L=2 (M2) and L=6 (M6). Example frames for these conditions are shown in Fig. 1.



(a) (b) (c) (d) (e)

Fig. 1. Example frames of the compressed stream (500 kbps) of Data #1 for different conditions used in the test: (a) NB, (b) S2, (c) M2, (d) S6 and (e) M6.

Four different original video sequences have been used for the tests. The first two sequences (Data #1 and #2) are from [24]. Their visual component was recorded at 25 fps with a resolution of 720×576 pixels and the acoustic component was recorded at the rate of 44.1 kHz. These sequences include 240 and 263 frames, respectively. In Data #1, a hand plays a guitar and then a synthesizer, while a wooden horse is rocking at the same time. In Data #2, a talking head and a rocking wooden horse appears at the same time. Data #3 and #4 are from the "group" section of the CUAVE database [25]. The first 300 frames of each sequence were used in the test. They are recorded at the visual frame rate of 29.97 fps with a resolution of 720×480 pixels and the acoustic frequency of 44.1 kHz. They contain two and three people pronouncing continuous English digits in turn, respectively. While a person speaks, the other persons also move their heads and mouths.

The x264 tool was used for H.264/AVC encoding of test sequences [26]. The constant bitrate encoding mode was used to produce sequences at low bitrate (100 kbps) and high bitrate (500 kbps). The audio part was encoded as the MP3 format.

In total, a dataset of 40 AV stimuli (5 test conditions × 4 contents × 2 bitrates) has been used.

Table 1. Details of the test room conditions.

LCD monitor	Eizo CG301W (2560×1600 pixels)
Monitor calibration using EyeOne Display2 calibration device	Gamut: sRGB; white point: D65; brightness: 120 cd/m^2; minimum black level
Ambient lighting	Neon lamps with 6500 K color temperature

3.2. Test methodology

The AV subjective quality test has been performed according to the guidelines provided by standards [27]. The Double Stimulus Continuous Quality Scale (DSCQS) method has been selected as test methodology [27]: two AV sequences are consecutively shown to the subject and he/she is asked to enter a quality rate for each of them. One of the stimuli is the reference signal (i.e. the uncompressed AV sequence) while the other is the test signal (i.e. the coded AV sequence). The subject is not told about the presence of the reference signal and he/she is asked to rate the overall perceived quality of the data, using the ITU continuous quality scale ranging from 0 to 100. To limit the duration of the test, two separate test sessions were designed, each of which are for only data compressed at either 100 kbps or 500 kbps, respectively. Fifteen naïve assessors took part in each session. Before each session, instructions regarding the subject's task and the goal of the test are provided and a training session takes place, where the experimenter explains the usage of the rating scale to the subject and some examples of conditions which can be found in the test material are shown. The contents used for the training were different from those used in the test sessions.

3.3. Test environment

The test room environment is intended to assure the reproducibility of the subjective test activity by avoiding the involuntary influence of any controllable external factors. Thus, it is important to fix some features of the viewing environment, regarding general viewing conditions and some crucial features of the used monitor. The information regarding our laboratory environment is detailed in Table 1.

4. RESULTS

4.1. Subjective data processing

The raw subjective scores have been processed in order to obtain the final Differential Mean Opinion Scores (DMOS). First, for each pair of stimuli, the differential scores have been computed as the difference between the scores assigned by the subject to the reference stimulus (i.e. the uncompressed video) and the processed stimulus (i.e. the compressed video). Second, an ANalysis Of VAriance (ANOVA) has been performed to check whether an intra-subject normalization of the scores would be needed. The results of the ANOVA have shown that an offset normalization was needed [28]. Finally, the screening of possible outlier subjects has been performed according to the guidelines described in [27]. The DMOS has been computed for each test condition together with the 95% confidence interval, as shown in Figs. 2 and 3. It is assumed that the overlap of 95% confidence intervals provides indication of absence of statistical differences between DMOS values.

4.2. Analysis of results

For both bitrate conditions, it is observed that, when the blurring effect is not strong (i.e., L=2), the difference in terms of perceived quality among the three processing conditions is statistically insignificant. When the bitrate is low (i.e., 100 kbps), even the strong blurring condition usually does not affect the quality much. For the high bitrate condition, the degradation by the strong blur may be noticeable and even attract the viewer's attention, which could result in worse quality of the processed sequences in comparison to those without blurring.

In most cases, regardless of the bitrate, the two prioritization schemes (i.e., high priorities for either sound sources or moving objects) do not show significant differences in quality. In other words, even if moving regions which do not produce sound are blurred in S2 and S6, the perceived quality is not significantly degraded in comparison to M2 and M6, respectively. This implies that all moving objects do not receive attention equally since sound-emitting objects tend to attract more attention than others.

It is noticed that, for Data #4 with 100 kbps encoding, the quality of the S6 case is worse than the other four cases. This is because, when a person stops speaking and another starts speaking, introduction of blur to the first speaker's face and disappearance of blur from the second one occurs abruptly, which may be unpleasant to the subjects' eyes. On the other hand, for the high bitrate condition, the perceived quality for S6 and M6 is statistically similar because the blurring effect is clearly visible even in the case of M6.

Finally, a small variability in the perceived quality is noticed depending on the original content of the sequence. Data #3 and #4 contain faces which are blurred sometimes while Data #1 and #2 do not. Since the humans' faces often attract visual attention more than other objects, perceived quality degradation by blur is more prominent for Data #3 and #4.

5. CONCLUDING REMARKS

We have investigated the influence of AV FoA on perceived quality of standard definition multimedia content. Imposing higher spatial resolutions on the sound-emitting regions in image sequences resulted in the same quality

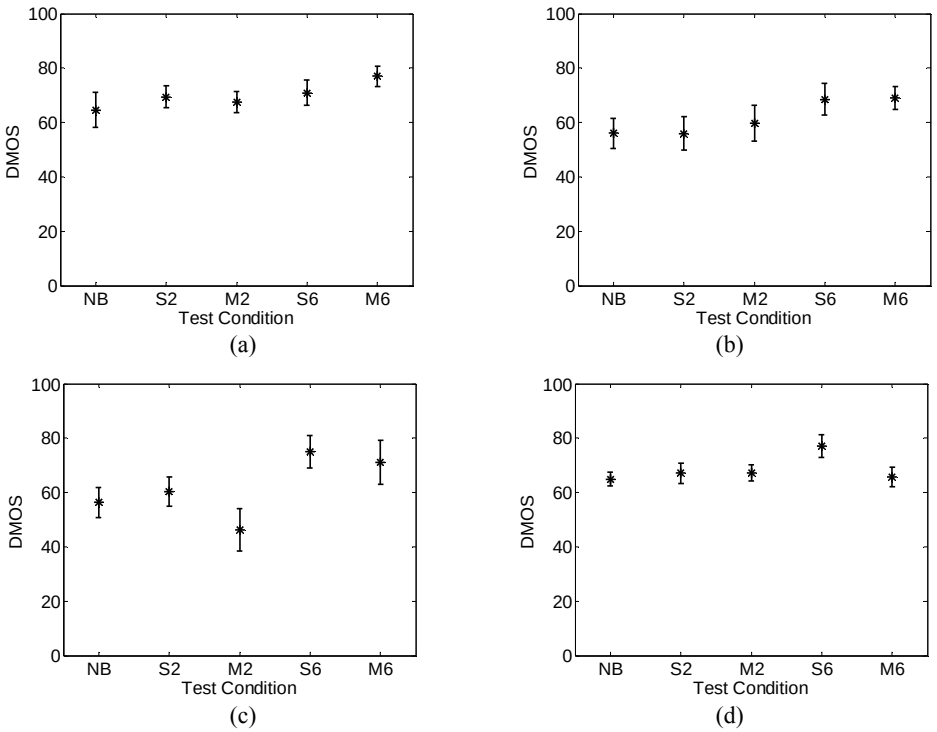


Fig. 2. Results of the subjective test for the bit rate of 100 kbps. The differential mean opinion score (DMOS) values and confidence intervals are shown for (a) Data #1, (b) Data #2, (c) Data #3 and (d) Data #4.

when compared to the case where all moving objects receive high priorities for the spatial resolution and even to the cases without blurring, unless the blur effect is too strong.

The reported result can be used in various applications. For example, in video coding one can obtain high coding efficiency by removing high frequency components or using coarse quantization steps outside the sound-emitting regions without significant perceived quality degradation [29].

In our future work we plan to perform further subjective tests for deeper understanding about the AV FoA and perceived quality. Quality assessment for high definition test material will also be considered. One of the interesting aspects would be instantaneous perceived quality, which would enable us to analyze the relationship between the quality and the content in detail. At the same time, we will work on developing objective quality metrics which are close to the human perception by exploiting complete FoA models.

6. REFERENCES

[1] W. Osberger, A. J. Maeder, and D. McLean, "A computational model of the human visual system for image quality assessment," in *Proc. Digital Image Computing: Techniques and Application*, New Zealand, Dec. 1997, pp. 337-342.

[2] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image," in *Proc. Int. Conf. Pattern Recognition*, Brisbane, Australia, Aug. 1998, pp. 701-704.

[3] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1200-1209, Oct. 2005.

[4] G. Boccignone, A. Marcelli, P. Napoletano, G. D. Fiore, G. Iacovoni, and S. Morsa, "Bayesian integration of face and low-level cues for foveated video coding," *IEEE Trans. Circuits, Syst. Video Technol.*, vol. 18, no. 12, pp. 1727-1740, Dec. 2008.

[5] K. Cater, A. Chalmers, and G. Ward, "Detail to attention: exploiting visual tasks for selective rendering," in *Proc. Eurographics Symposium on Rendering*, Aire-la-Ville, Switzerland, 2003, pp. 270-280.

[6] V. Sundstedt, D. Gutierrez, O. Anson, F. Banterle, and A. Chalmers, "Perceptual rendering of participating media," *ACM Trans. Applied Perception*, vol. 4, no. 3, article 15, Nov. 2007.

[7] S. Lee, M. S. Pattichis, and A. C. Bovic, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129-132, 2002.

[8] W. Osberger, A. J. Maeder, and N. Bergmann, "A technique for image quality assessment based on human visual system model," in *Proc. European Signal Processing Conf.*, Rhodes, Greece, Sep. 1998, pp. 1049-1052.

[9] A. Cavallaro and S. Winkler, "Segmentation-driven perceptual quality metrics," in *Proc. Int. Conf. Image Processing*, Singapore, 2004, pp. 3543-3546.

[10] F. M. Ciaramello and S. S. Hemami, "'Can you see me now?' An objective metric for predicting intelligibility of compressed American sign language video," in *Proc. Human Vision and Electronic Imaging*, San Jose, CA, USA, Jan. 2007, pp. 64920M.

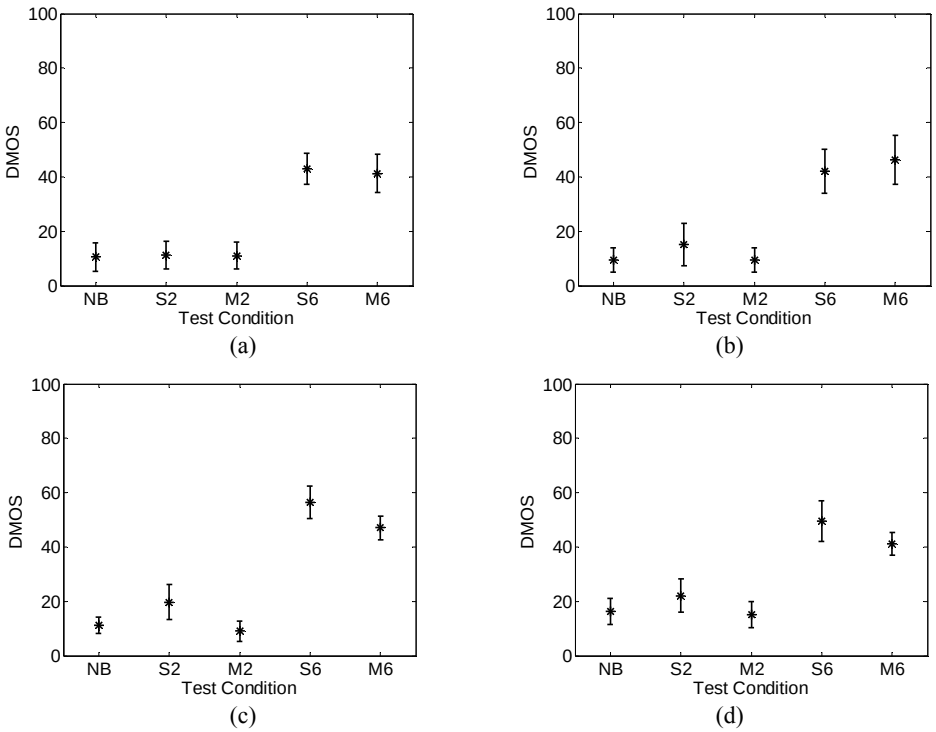


Fig. 3. Results of the subjective test for the bit rate of 500 kbps. The differential mean opinion score (DMOS) values and confidence intervals are shown for (a) Data #1, (b) Data #2, (c) Data #3 and (d) Data #4.

[11] G. Mastoropoulou, K. Debattista, A. Chalmers, and T. Troscianko, "Auditory bias of visual attention for perceptually-guided selective rendering of animations," in *Proc. Int. Conf. Computer Graphics and Interactive Techniques in Australasia and South East Asia*, New York, NY, USA, 2005, pp. 363-369.

[12] V. Hulusic, M. Aranha, and A. Chalmers, "The influence of cross-modal interaction on perceived rendering quality thresholds," in *Proc. Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision*, Plzen, Czech Republic, 2008, pp. 41-48.

[13] J. Driver and C. Spence, "Attention and the crossmodal construction of space," *Trends in Cognitive Sciences*, vol. 2, no. 7, pp. 254-262, 1998.

[14] C. Spence and J. Driver, "Audiovisual links in exogenous covert spatial orienting," *Perception and Psychophysics*, vol. 59, pp. 1-22, 1997.

[15] C. Spence and J. Driver, "Audiovisual links in endogenous covert spatial attention," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, pp. 1005-1030, 1996.

[16] D. J. Tellinghuisen and E. J. Nowak, "The inability to ignore auditory distractors as a function of visual task perceptual load," *Perception and Psychophysics*, vol. 65, no. 5, pp. 817-828, 2003.

[17] V. Mazza, M. Turatto, M. Rossi, and C. Umiltà, "How automatic are audiovisual links in exogenous spatial attention?" *Neuropsychologia*, vol. 45, pp. 514-522, 2007.

[18] J. Vroomen and B. de Gelder, "Perceptual effects of cross-modal stimulation: ventriloquism and the freezing phenomenon," in *Handbook of Multisensory Processes*, MIT Press, 2004.

[19] J. Vroomen and B. de Gelder, "Sound enhances visual perception: cross-modal effects of auditory organisation on vision," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 5, pp. 1583-1590, Oct. 2000.

[20] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe, "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," *Cerebral Cortex*, vol. 17, no. 5, pp. 1147-1153, 2007.

[21] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, Dec. 1976.

[22] R. Sharma, V. I. Pavlović, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853-869, May 1998.

[23] E. Macaluso and J. Driver, "Multisensory spatial interactions: a window onto functional integration in the human brain," *Trends in Neuroscience*, vol. 28, no. 5, pp. 264-271, May 2005.

[24] E. Kidron, Y. Y. Schechner, and M. Eland, "Cross-modal localization via sparsity," *IEEE Trans. Signal Processing*, vol. 55, no. 4, pp. 1390-1404, Apr. 2007.

[25] E. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: a new audio-visual database for multimodal human-computer interface research," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, 2002, pp. 2017-2020.

[26] http://www.videolan.org/developers/x264.html

[27] Rec. ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," Geneva, Switzerland, 2002.

[28] E. D. Gelasca, "Full-reference objective quality metrics for video watermarking, video segmentation and 3D model watermarking," Ph.D. Thesis, EPFL, Switzerland, 2005.

[29] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Video coding based on audio-visual attention," in *Proc. Int. Conf. Multimedia and Expo*, New York City, USA, Jun. 2009. (accepted)