Audiovisual focus of attention and its application to Ultra High Definition video compression

Martin Rerabek[a], Hiromi Nemoto[a], Jong-Seok Lee[b], and Touradj Ebrahimi[a]

[a]Multimedia Signal Processing Group (MMSPG), Institute of Electrical Engineering (IEE), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[b]School of Integrated Technology, Yonsei University, Incheon, Republic of Korea

ABSTRACT

Using Focus of Attention (FoA) as a perceptual process in image and video compression belongs to well-known approaches to increase coding efficiency. It has been shown that foveated coding, when compression quality varies across the image according to region of interest, is more efficient than the alternative coding, when all region are compressed in a similar way. However, widespread use of such foveated compression has been prevented due to two main conflicting causes, namely, the complexity and the efficiency of algorithms for FoA detection. One way around these is to use as much information as possible from the scene. Since most video sequences have an associated audio, and moreover, in many cases there is a correlation between the audio and the visual content, audiovisual FoA can improve efficiency of the detection algorithm while remaining of low complexity. This paper discusses a simple yet efficient audiovisual FoA algorithm based on correlation of dynamics between audio and video signal components. Results of audiovisual FoA detection algorithm are subsequently taken into account for foveated coding and compression. This approach is implemented into H.265/HEVC encoder producing a bitstream which is fully compliant to any H.265/HEVC decoder. The influence of audiovisual FoA in the perceived quality of high and ultra-high definition audiovisual sequences is explored and the amount of gain in compression efficiency is analyzed.

Keywords: Quality assessment, Video coding, Foveated coding, Audiovisual source localization, Audiovisual focus of attention, H.265/HEVC, Ultra High Definition

1. INTRODUCTION

In the past few decades, FoA mechanisms have been drawing intense research interest because of their potential applications in efficient image and video coding, objective quality metrics and scene analysis. It is well known that while observing the scene in a video sequence, the human visual system captures only small regions around fixation points at high resolution and the resolution for the peripheral area decreases. Thus, the degradation of visual quality contained in peripheral areas might not be noticed by human observers. Exploiting this fact in video coding allows to remove or suppress the imperceptible information outside of the small fixation regions without significant impact on perceived quality. This results in a process commonly referred to as foveated video coding. The first step in foveated video coding is to create the spatial prioritization scheme which determines the priorities in different scene regions by considering the human FoA mechanisms. Then, the encoding is performed according to those priority maps.

Several computational models of FoA exploiting bottom-up saliency detection, face detection, moving object detection, etc., have been proposed.[1–3] Although these attention-based techniques have been proven to be beneficial for the coding efficiency and quality metrics accuracy through subjective and objective experiments, visual attention guided by the acoustic modality has been rarely taken into account.[4] Our previous works exploit

Further author information: (Send correspondence to Martin Rerabek)
Martin Rerabek: E-mail: martin.rerabek@epfl.ch
Hiromi Nemoto: E-mail: hiromi.nemoto@epfl.ch
Jong-Seok Lee: E-mail: jong-seok.lee@yonsei.ac.kr
Touradj Ebrahimi: E-mail: touradj.ebrahimi@epfl.ch

the correlation between audio and visual content, and present a simple audiovisual source localization method improving the efficiency of FoA detection.[5]

With the rapid progress of computational FoA algorithms, various models of foveated video coding have been proposed by taking advantage of properties of human visual system.[6,7] Wang et al.[2] proposed a perceptually scalable video coding framework based on the human foveation model by assuming that face regions are the points of fixation. The bits containing the details of expected salient regions are sent first, whereas bits for other regions may be discarded according to the given bit rate condition in the encoding system. Itti[1] uses a bottom-up visual attention model to detect salient regions and employs foveation filter to the video based on the saliency values of each pixel before coding. Tang[8] builds a visual attention priority map by incorporating image features from both a spatio-velocity visual sensitivity model and visual masking model. The coding efficiency is improved without perceptual quality degradation by varying the quantization parameter (QP) value for the salient region and the background region.

However, the acoustic modality, as an important aspect in visual attention and perceived quality, has been rarely considered in the aforementioned coding methods. An efficient video coding method using audiovisual FoA has been previously reported in recent studies.[9,10] The proposed coding was implemented in the framework of H.264/AVC by assigning different QPs for different regions. It was shown that significant coding gain in comparison to the constant quantization mode of H.264/AVC can be achieved without deterioration of perceived image quality on both standard and high definition sequences. In higher resolution video sequence, such as 4K and 8K, which are expected to be next standard video format resolutions, more peripheral vision is expected to be used due to a more immersive environment. Therefore, in such an immersive environment, FoA plays more important role for efficient video coding or image quality objective metrics.

This paper investigates the effect of foveated coding algorithm, first described in,[5] on the perceived quality of high and ultra high definition video sequences. We assume that the aural stimuli correlated to the visual information can drive visual attention, and therefore affects the perceived quality of multimedia content. The effectiveness and usefulness of the proposed foveated coding method implementing into H.265/HEVC encoder are discussed and analyzed through the results of a subjective quality assessment. The results show that the effect of visual degradation outside the FoA fixation point is, apart from extreme cases, insignificant.

The rest of the paper is organized as follows. The next section explains the background of the audiovisual source localization algorithm. In Section 3, the details about subjective assessment experiment are presented and its results are discussed. Finally, concluding remarks are given in Section 4.

2. FOVEATED VIDEO CODING USING AUDIOVISUAL FOCUS OF ATTENTION

Finding the location of the sound source in the visual scene is a challenging task in that, among multiple objects or parts showing visual motion in the scene, we need to identify which one is responsible for generating the audio signal. In our work, this is accomplished by exploiting the correlation structure residing in the audio and video signals. Below in this section, the audiovisual source localization algorithm[10] used in this paper is described.

First, it is necessary to extract features from the raw audio and video signals. The difference of the luminance component of consecutive frames is used as visual features. For the audio signal, the energy within a moving window is obtained and its temporal difference is used as audio features. The window moves at the rate corresponding to visual frame rate in order to obtain temporally synchronized features from the two modalities.

The localization algorithm basically uses the canonical correlation analysis (CCA) to find the pixel location showing the maximum correlation with the audio signal. The objective of CCA is to find a pair of projection vectors for the audio and visual data, noted as $\mathbf{w}_a$ and $\mathbf{w}_v$, respectively, which maximize the linear correlation of the projected data. It can be shown that solving the CCA problem becomes equivalent to solving[11]

$$\mathbf{V}\mathbf{w}_a = \mathbf{A}\mathbf{w}_a, \tag{1}$$

where A and V are collections of audio and visual feature vectors over a certain time period. Note that, when the audio feature dimension is one as in our case, $\mathbf{w}_a$ can be omitted, i.e,

$$\mathbf{V}\mathbf{w}_a = \mathbf{A}. \tag{2}$$

(a) (b)

Figure 1: Original image frame (a), and the results of source localization, region partitioning, and blurring (b).

Two principles are employed on top of the above formulation for effective sound source localization. The first is the principle of spatial sparsity, meaning that the sound source is localized in a small region rather than scattered over the entire scene. This can be stated as a l^1-norm minimization problem together with (2) as a constraint, i.e.,

$$\min ||\mathbf{w}_v||_1 = \sum_{i=1}^{n} |w_{vi}| \quad \text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A}, \tag{3}$$

where w_{vi} is the i-th component of the n-dimensional feature vector $\mathbf{w}_v$. The second principle is the spatio-temporal consistency, i.e., the sound source tends to move smoothly over time, which modifies (3) as

$$\min \sum_{i=1}^{n} |f_i w_{vi}| \quad \text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A}. \tag{4}$$

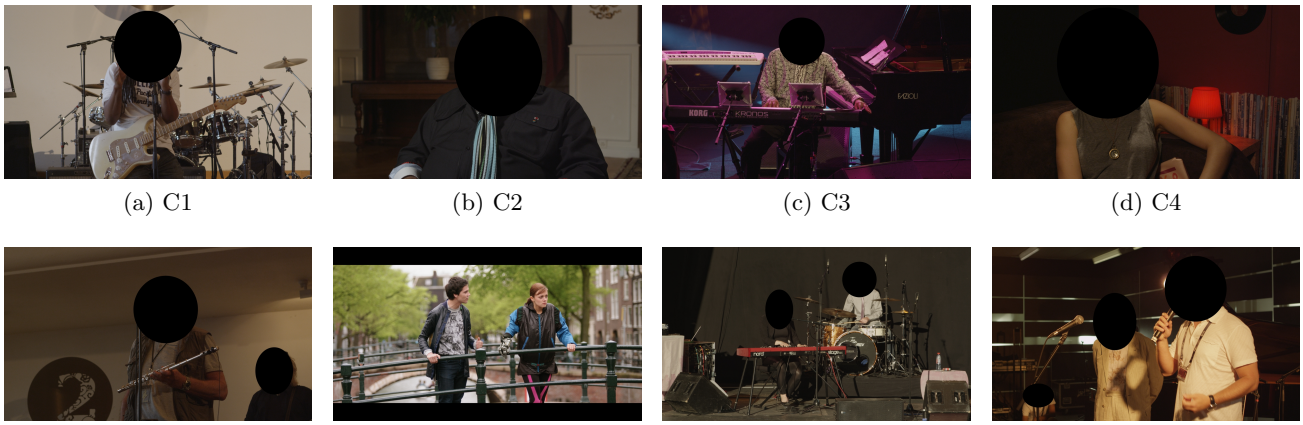
The weighting factor f_i suppressing abrupt motion is given by

$$f_i = \max_{1 \le j \le n} w_{vj}^{old} - w_{vi}^{old} + 1, \tag{5}$$

where w_{vi}^{old} is the i-th component of the spatially smoothed version of the solution for the previous temporal window. Here, a Gaussian filter is applied to the image representation of the solution for smoothing. Thus, the weight is small for the region near the sound source for the previous temporal window in order to force the localization result to stay near the previous source localization. The problem (4) can be solved by linear programming, which is repeated over time for tracking the sound source. The solution $\mathbf{w}_v$ can be viewed as a cross-modal energy concentrated on the visual features that are highly correlated to the audio signal. Thus, the pixel location corresponding to the feature showing a high energy is regarded as a part of the sound source.

Once the sound source is localized in the scene, spatially uneven quality degradation is performed by using Gaussian blurring as a preprocessing step before video coding. For each image frame, a priority map is produced, which represents the weighted distance between each pixel and the nearest localized energy location. When there are more than one energy source identified by the localization algorithm, the weighting is calculated in such a way that a pixel near a smaller energy receives a larger distance than one near a larger energy location, just as in a contour map. Then, blurring is performed with a Gaussian pyramid, i.e., stronger blurring is applied to low priority regions. Each level of the pyramid is assigned to the linearly spaced values within the range of the priority values. For the priority values between two levels, trilinear interpolation is applied. Figure 1(a) shows the original frame of content *C6* and the Figure 1(b) illustrates the results of source localization, partitioning of the image frame into $L = 8$ regions, and application of uneven blurring. Finally, the blurred image frames are encoded with a conventional encoder (H.265/HEVC in our case), which produces the final video bit stream. Higher compression ratios are obtained for the smoothed regions since high frequency components are eliminated via smoothing before coding.

Note that it is also possible to embed a process conducting spatially uneven quality assignment in a video encoder (e.g., by applying different quantization step sizes for each of the partitioned regions[5]), which may show



(a) C1 (b) C2 (c) C3 (d) C4

(e) C5 (f) C6 (g) C7 (h) Training

Figure 2: Sample frames of individual contents considered in the subjective test. The MJF content is censored because of a request from copyright owners.

better coding efficiency. However, the preprocessing-based approach has an advantage in that any existing video encoder can be easily utilized without modification and the produced bit stream is fully compatible with any existing decoder corresponding to the used encoder.

3. EXPERIMENT

In order to test the influence of audiovisual FoA on the perceived quality of HD and UHD audiovisual sequences, and to explore the amount of gain in compression efficiency, a subjective quality assessment for diverse contents, coding, and rendering conditions was conducted. In the experiment, audiovisual content containing the spatially variant degradation is presented to subjects and subjective rating of the overall perceived quality of test material is subsequently collected. This section presents the details and results of the subjective experiment evaluating the foveated coding method described in the previous section.

3.1 Dataset description

The dataset consists of eight ten-second audiovisual sequences of different contents. Seven of them (*C1*, *C2*, *C3*, *C4*, *C5*, *C7*, and *training*) were shot during the 2012 edition of the Montreux Jazz Festival (MJF) (copyright protected), with a RED SCARLET-X camera in REDCODE RAW (R3D) format, DCI 4K resolution (4096×2160), 25 fps. The last sequence (*C6*) was downloaded as a part of the Tears of Steel movie * in 4096×1714 16-bit sRGB tiff files format at 24 fps, specifically frames number 608-848 were chosen for the test purposes. Figure 2 shows the sample frame of each content. Since the MJF test material is copyright protected and the sample frames of MJF content are edited/censored, more details and closer description of each content and its characteristics are given in Table 1. For better understanding of the content, the spatial information (SI) and temporal information (TI) indexes were computed on the luminance component of each content according to[12] (see Figure 3).

The recorded video sequences were first cropped and padded to 4K UHD resolution (3840×2160 pixels) and stored as raw video files, progressively scanned, with YUV 4:2:0 color sampling, and 8 bits per sample. Then, a spatially variant quality degradation was performed on each video sequence as described in Section 2. More specifically, Gaussian pyramids with various levels were applied in order to produce different versions of blurred data. Thus, for each content, five versions of blur levels were considered: *L0*, *L2*, *L4*, *L6*, *L8*. Note that the level *L0* corresponds to the reference or unblurred data. UHD sequences were consequently downsampled to full HD resolution (1920×1080 pixels) using bilinear interpolation. All sequences were then compressed using

*Tears of Steel is a computer generated movie produced by the Blender Institute using the open source computer graphics software Blender and released under the Creative Commons Attribution license (`http://mango.blender.org/`)

Content	*Description and characteristics*
C1	An artist talking to the audience. No distraction except his moving hands, fixation point is his mouth which is covered by microphone.
C2	Interview with an artist. No distraction, low movement, mouth as a fixation point.
C3	An artist singing while playing keyboard on the stage. No distraction, low movement, mouth as a fixation point.
C4	Interview with a singer. Low movement, no distraction, mouth as a fixation point.
C5	An artist holding flute, talking to the audience. One static person in background, low movement, shot slightly from side, mouth as a fixation point covered by microphone.
C6	Two persons staying on bridge, discussing. Medium shot focused on them while background is blurred, some movement, fixation point changes position once in the middle of the sequence from guy's to girl's face.
C7	Two artists at the stage, one playing keyboard and singing, the other playing drums. Attention changes from singer to drummer once at the end of the sequence, some movement, mouth of the singer covered by microphone most of the time.
Training	Two persons in the studio, one talking to the audience, second one nodding while listening the first one. Third person moving in background, higher level of movement, fixation point (mouth of the talking person) is visible from profile and partially covered by microphone.

Table 1: Characteristics of the contents used in our experiments.

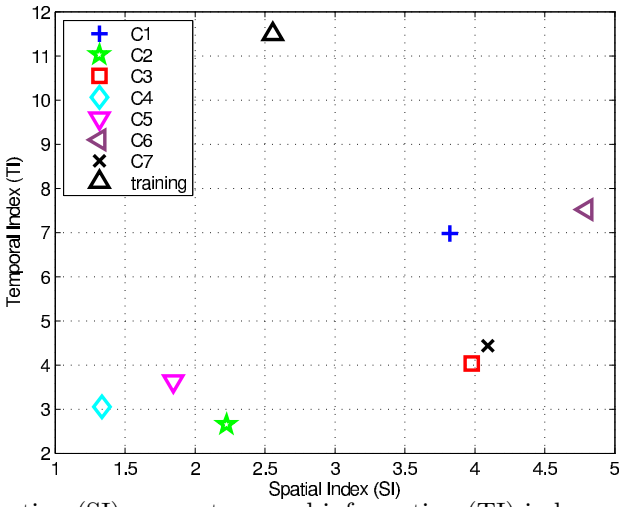


Figure 3: Spatial information (SI) versus temporal information (TI) indexes of the selected contents.

H.265/HEVC with different QPs for each resolution. After encoding, the HD sequences were padded to UHD resolution using the mid gray color. For all sequences, mono audio PCM format, sampled at 48 kHz 24 bits was used.

The video sequences were compressed with HEVC using HM 12.1. main profile and level 6.2. The Random Access (RA) configuration was selected for this study. The configuration parameters were selected according the configuration template accessible at HEVC software repository [†]. The only change in the configuration template was that the Intra Period parameter was set to 1s. Furthermore, to obtain sequences with different quality level, two QP values were selected based on expert screening for each resolution: QP=20 for high quality (HQ) of both HD and UHD, and QP=30 and QP=33 for low quality (LQ) of HD and UHD, respectively. Various content, resolution, blur level and QP lead to a total of 140 audiovisual sequences (70 for UHD and 70 for HD).

[†] `https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/trunk/cfg/encoder_randomaccess_main.cfg`

3.2 Test methodology

The audiovisual subjective quality assessment was conducted according to the guidelines provided by the ITU recommendations.[13] The Single-Stimulus (SS) evaluation scheme[13] was selected as the test methodology in order to replicate the home viewing condition. The audiovisual sequences were consecutively presented to subjects in a way that they usually watch the material without a source reference, and they were asked to enter the quality score for each. More specifically, subjects were instructed to rate the overall perceived quality of the presented sequences using the ITU continuous quality scale ranging from 0 (bad quality) to 100 (excellent quality). The length of the test sequences was 10s and the time window for voting was set to 5s.

In order to retain the concentration of the subjects, the test was separated into four different sessions, each approximately 10 minutes long followed by 10 minutes of resting phase. To prevent the inter-resolution comparison, first two sessions were dedicated to UHD content, whereas within the second two session the HD content was evaluated. Furthermore, to avoid a possible effect of the presentation order, the stimuli were randomized in a way that the same content was never shown consecutively. Also, dummy sequences, whose scores are not included in the results but the observer was not told about, were inserted at the beginning of the first and the third session to stabilize observers' rating after training and UHD sessions, respectively. Overall, 3 dummy presentations were included at the beginning of the first and the third session.

Nineteen naive subjects (4 females, 15 males) took part in our experiments. They were between 18 and 27 years old with an average of 21.6 years of age. All subjects were screened for correct visual acuity (no errors on 20/30 line) and color vision using Snellen and Ishiara charts, respectively. They all provided written consent forms.

Before the first session, the oral instructions were provided to participants to explain their tasks and a training session was conducted to allow participants to familiarize with the assessment procedure. The content shown in the training session was selected in order to show to participant the high and the low quality of encoded material without blurring. The participants were not informed about the presence of blurring in the test sequences and were specifically instructed not to search the audiovisual content for distortions but to watch it in a normal way as they usually do when watching TV at home.

To play the test audiovisual sequences, a 56-inch professional high-performance 4K/QFHD LCD reference monitor Sony Trimaster SRM-L560 [‡] and two PSI A14-M professional studio full range speakers [§] were used. To assure the reproducibility of results by avoiding involuntary influence of external factors, the laboratory for subjective video quality assessment was set up according to.[13] The monitor was calibrated using an EyeOne Display2 color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level. The room was equipped with a controlled lighting system that consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid gray. The illumination level measured on the screens was 20 lux and the ambient black level was 0.2 cd/m^2. The test area was controlled by an indoor video security system to keep track of all the test activities and of possible unexpected events, which could influence the test results.

It is known that a person with a normal or corrected to normal vision can see the maximum of details of full HD content without distinguishing two adjacent lines when the visual angle between two adjacent lines equals one arcminute. Considering this perceptual criteria, the viewing distance for subjective quality assessment was set to 1.6 and 3.2 times the picture height for UHD and HD sequences, respectively.[14]

3.3 Data processing and results

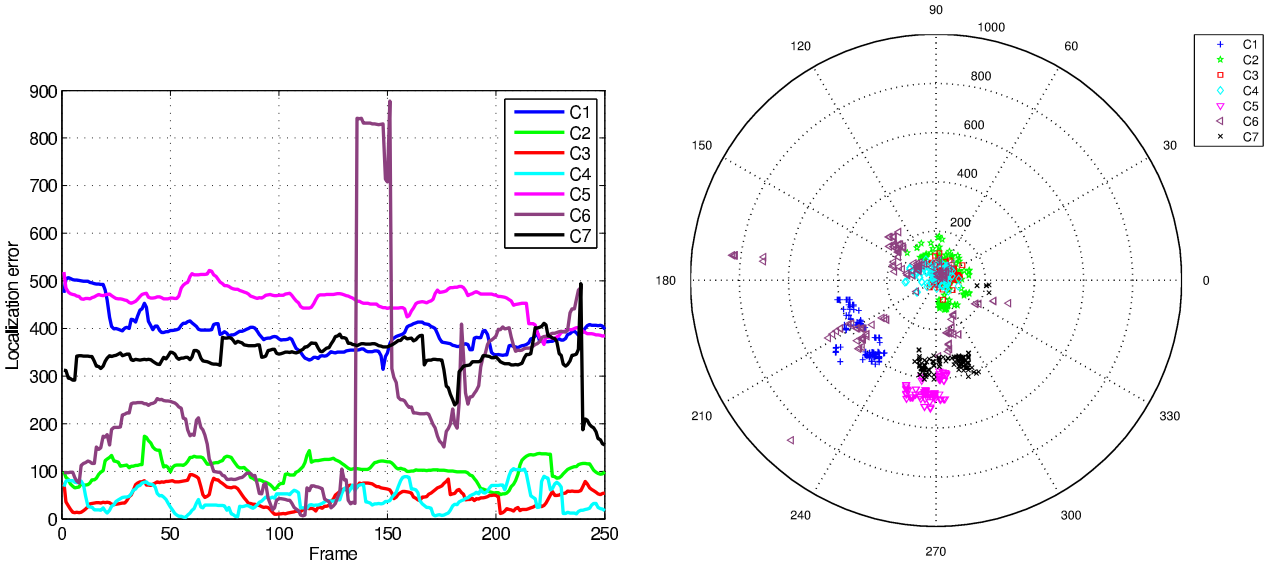
In this section, the results of source localization and subjective evaluation are shown and analysed, as well as the coding efficiency of foveated coding for HD and UHD. First, the performance results of the proposed source localization algorithm in comparison to defined ground truth are discussed. Then, the usefulness and effectiveness of the proposed method are evaluated in terms of perceived quality degradation and coding efficiency, respectively.

[‡] http://pro.sony.com/bbsccms/assets/files/cat/mondisp/brochures/di0195_srm1560.pdf

[§] http://www.psiaudio.com/product/active-monitors/a14-m

Table 2: Source localization performance - average localization error and its standard deviation values over time for each content.

Content	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*
Error[px]	390.6±39.3	104.9±22.7	48.0±21.9	43.4±24.5	457.5±32.1	232.7±198.2	340.2±43.4



(a) (b)

Figure 4: Source localization error per frame for each content in radial (a) and (b) angular direction.

3.3.1 Source localization

The results of the source localization appear as the cross-modal energies located in the pixel locations of the estimated source. The performance of the localization algorithm is evaluated based on the localization error computed as a pixel distance between the maximum localized energy and the sound-emitting region. In order to compute the performance measure, the sound-emitting region in each sequence was identified manually and used as the ground truth. Figure 4(a) shows the temporal change of the localization error for each content in radial direction, and Figure 4(b) shows its fluctuation in the angular direction. Table 2 shows the localization performance of proposed algorithm in terms of the average and standard deviation values of the localization error in pixels over time.

The results of source localization vary depending on the content. The localization works well for content where one person and no sound source occlusion appear (see results for *C2*, *C3* and *C4*). For more challenging content with either sound source occlusion (*C1, C5*) or sound source change over time (*C7*), the localization error in radial direction is bigger, however the fluctuation in angular direction over the time is still relatively small. The results for content *C6* exhibit both, the larger fluctuation over the time and the radial distance error, especially within frames number 135-152. This means that the estimated source location moves more around the image frame. Thus, if these results are used for coding, the quality of each region in the scene will significantly change over time, which may degrade the perceived quality of the resultant sequence.

3.3.2 Subjective quality evaluation

In order to judge the usefulness of the proposed foveated coding method, the subjective scores, in terms of the degradation of the perceived quality are evaluated.

To detect and remove subjects whose scores appear to deviate strongly from the other scores in a session, the outlier detection was performed. In each set of scores assigned to a test sequence, a score by subject j and test

condition i, s_{ij}, was considered as outlier if $s_{ij} > q_3 + 1.5(q_3 - q_1) \vee s_{ij} < q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles of the scores distribution for test condition i, respectively.[15] This range corresponds to approximately ± 2.7 the standard deviation or 99.3% coverage if the data is normally distributed. A subject was considered as an outlier, and thus all her/his scores were removed from the results of the session, if more than 20% of her/his scores over the session were outliers.[15] In this study, no outlier subjects were detected.

Statistical measures were computed to describe the score distribution across the subjects for each of the test conditions (combination of content, resolution and encoding quality). For the used methodology SS, the mean opinion score (MOS) is computed as

$$MOS_i = \frac{\sum_{j=1}^{N} s_{ij}}{N} \tag{6}$$

where N is the number of valid subjects and s_{ij} is the score by subject j for the test condition i. The relationship between the estimated mean values based on a sample of the population (i.e., the subjects who took part in our experiments) and the true mean values of the entire population is given by the confidence interval of the estimated mean. The $100 \times (1-\alpha)\%$ confidence intervals (CI) for MOS values were computed using the Student's t-distribution according the following equation

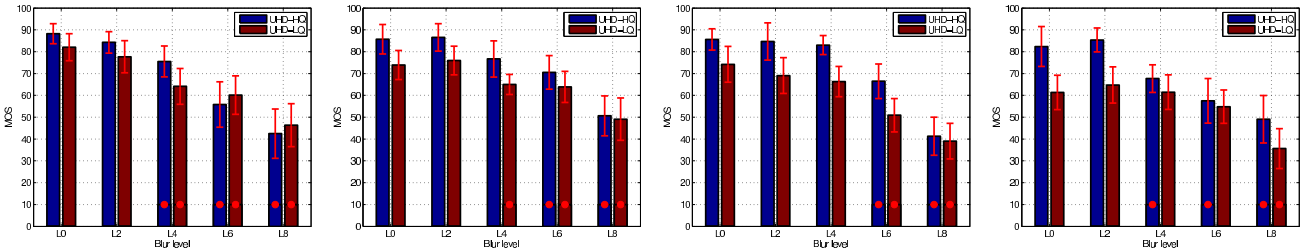
$$CI_i = t(1 - \alpha/2, N) \cdot \frac{\sigma_i}{\sqrt{N}} \tag{7}$$

where $t(1 - \alpha/2, N)$ is the t-value corresponding to a two-tailed t-Student distribution with $N - 1$ degrees of freedom and a desired significance level α (equal to 1-degree of confidence). N corresponds to the number of valid subjects, and σ_i is the standard deviation of a single test condition i across the subjects j. The confidence intervals were computed for an α equal to 0.05, which corresponds to a degree of significance of 95%. In order to examine the statistical significance of the quality difference between the reference (*L0*) and foveated (*L2, L4, L6, L8*) sequences, two-tailed t-tests were performed under the null hypothesis that the two rating scores are independent random samples from normal distributions with equal means, against the alternative that they do not have equal means.

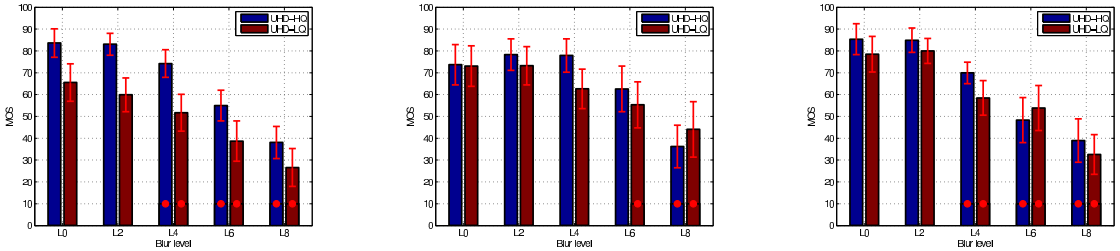
Figure 5 and Figure 6 show the MOS and CI values of the ten coding conditions (five blur levels, two quality levels) for each UHD and HD content, respectively. The results of the t-test between the reference and foveated sequences are shown with the red dots in the bar plots. Bars with a red dot for the MOS of a foveated coding case indicate that the ratings for the corresponding foveated sequence are significantly different from those for the reference, whereas bars without the red dot imply that the difference of the two MOS values is not significant.

Overall, the MOS values presented in the plots are above 50 for all HQ sequences of both resolutions and for blur level up to *L6*. In most of the cases, it is observed that unfoveated sequences have the best quality and, as expected, the quality decreases as the value of blur level increases. This observation is valid for both quality levels *HQ, LQ*. Furthermore, the *HQ* sequences outperform *LQ* sequences in all cases except in few exceptions for *L8* blur level. This means that the coding artifacts can mask the foveated blurring and decrease the variance of the spatial degradation. Relatively small variability in the perceived quality is noticed for blur level up to *L4*, whereas after this blur level, the perceived quality drops faster within each content. In some cases (UHD: *C2, C4, C6, C7*; HD: *C1, C4, C5, C7*), the quality of the foveated sequence with blur level *L2* and/or *L4* is equal or even higher than the reference.

In all contents, it is observed that foveated blurring with blur level *L2* can be used without degradation of perceived quality. Moreover, in several cases of high quality content, such as (UHD: *C2, C3, C6*; HD: *C4, C5*), the *L4* blur level of foveated coding can provide more coding gain without statistically significant perceived quality degradation. For UHD *C6-HQ*, even using *L6* does not lead to statistically significant quality degradation in comparison to the unfoveated mode, which is interesting taking into account the results of the source localization for this content. This means that, even for a higher localization distance error in this content, the amount of blur at the fixation point is not too high and doesn't decrease the perceived quality. For more detailed analysis of the results for this content, the effect of its characteristics on the perceived quality must be taken into account. The background region of the content *C6* is already blurred for artistic effect, whereas the foreground (the fixation

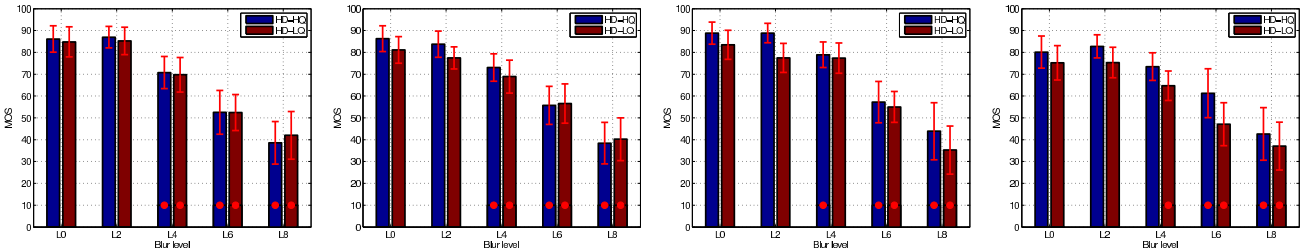


(a) C1 (b) C2 (c) C3 (d) C4

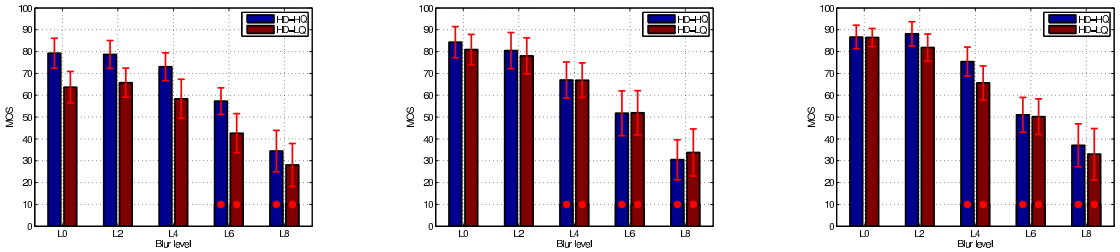


(e) C5 (f) C6 (g) C7

Figure 5: Results of subjective test comparing overall quality of foveated blurring with different amount of blur for UHD resolution. Red points in each bar indicates the case where the quality degradation caused by certain blur lever is statistically significant at a significance level of 95%.



(a) C1 (b) C2 (c) C3 (d) C4



(e) C5 (f) C6 (g) C7

Figure 6: Results of subjective test comparing overall quality of foveated blurring with different amount of blur for HD resolution. Red points in each bar indicates the case where the quality degradation caused by certain blur lever is statistically significant at a significance level of 95%.

area) contains a stronger attention attractor (i.e., conversation of a man and a woman). Conversation, or speech in general, usually attracts attention more strongly than, for instance, musical instruments, and blurring in the background in this content further helps to focus the attention around the two people. These facts are probably the main reasons why even applying the higher level of additional blurring doesn't deteriorate the perceived quality of this content.

In general, the effect of FoA mechanisms, as well as the decreased resolution of peripheral vision in the UHD sequences in comparison to the HD, is demonstrated.

3.3.3 Coding gain

The effectiveness of the proposed method, in terms of the coding efficiency, is investigated in comparison to the reference modes for both quality levels (*L0-HQ, L0-LQ*) of each content. Table 3 shows the relative coding gain in bit rate for all coding conditions.
For the UHD content: when the QP is small (i.e., better quality), the advantage of the proposed method in terms of coding efficiency is clearly visible even for low level of blurring *L2*. On the other side, the bigger QP value producing the lower quality stream brings much lower coding gain for all blurring levels. It can be explained by the fact that for the lower QP the encoder tries to preserve the high frequency components, thus any blurring caused by foveated coding brings a significant coding gain. The special attention belongs to the computer generated content *C6*, where the coding efficiency range from 68% to 89%, and from 9% to 39% for the low and high QP, respectively. Such a high coding efficiency can be explained by nature of the content, which differs from the rest of the test material in various aspects. Scenes in the content are much brighter with better contrast (i.e., higher dynamic range) and contain more complex background and diverse colors in comparison to the scenes of other contents. Furthermore, the focus is on the two people having much details and covering a large portion of the scene. In summary, *C6* contains much more details (i.e., high frequency components) than other content which leads to higher coding gain. Moreover, this content could be basically considered as a more realistic content that we can think of when we talk about video in general. Thus, the results for this content, in terms of coding gain, could be expected more likely for real applications.
For the HD content: foveated coding of high quality content exhibits the coding efficiency at least as twice smaller in comparison to UHD, and it gets even lower for low quality content. The significant amount of the coding gain (more than 20%) can be achieved with blur level *L4* and *L6* for high and low quality content, respectively. To prepare the HD sequences, the bilinear interpolation is applied to UHD content, which can a reason of the lower coding gain of HD sequences.

The maximum amount of the coding gain which can be achieved without degradation of the perceived quality is presented as bold in Table 3. These values correspond to the level of blur for which the degradation of perceived quality is still not statistically significant at a significance level of 95%. Although the subjects were instructed to feel like being at home and to freely watch the stimuli without excessive focus on the quality evaluation task, the viewing conditions might not be the same as a normal free-viewing. In fact, it has been shown that given task demands can affect viewing patterns of observers significantly because sensory-driven bottom-up saliency features are immediately overridden by task demands.[16] It was demonstrated that the pattern of eye movement is clearly dependent on the instructions given to the observers in viewing a painting.[17] Therefore, in the real free-viewing scenario the effectiveness of the foveated coding may be even more significant in comparison to what was measured in our experiments.

4. CONCLUSION

A preprocessing-based approach to video coding using the audiovisual information to determine the importance of each image frame area for efficient encoding has been presented in this paper. Furthermore, the influence of audiovisual FoA mechanisms on perceived quality of high and ultra high definition multimedia content was investigated through extensive subjective assessment. Exploiting the audiovisual FoA principles, a significant efficiency improvement of video coding without perceived quality degradation can be achieved especially for UHD multimedia content. Moreover, the results of the subjective evaluation and the coding gain showed that

	L2	L4	L6	L8
C1	**13.69**%	31.86%	41.45%	47.92 %
C2	14.22%	**37.29**%	47.13%	53.74 %
C3	17.99%	**40.62**%	52.80%	60.81 %
C4	**16.92**%	39.67%	51.64%	60.23 %
C5	**18.31**%	40.23%	50.39%	58.57 %
C6	67.86%	83.68%	**87.07**%	88.59 %
C7	**12.21**%	30.72%	41.14%	48.07 %

(a) UHD resolution - High Quality

	L2	L4	L6	L8
C1	**5.25**%	15.35%	26.06%	35.47 %
C2	**7.71**%	21.10%	31.47%	39.21 %
C3	8.84%	**23.86**%	37.54%	48.46 %
C4	6.51%	18.97%	**30.39**%	40.22 %
C5	**3.73**%	11.74%	20.09%	27.97 %
C6	8.63%	**20.36**%	30.64%	39.60 %
C7	**9.73**%	26.28%	39.97%	50.06 %

(b) UHD resolution - Low Quality

	L2	L4	L6	L8
C1	**5.47**%	16.21%	25.65%	33.20 %
C2	**5.75**%	18.26%	26.98%	32.23 %
C3	**8.84**%	23.78%	36.17%	45.76 %
C4	7.01%	**20.98**%	32.24%	42.03 %
C5	4.80%	**14.54**%	22.36%	30.33 %
C6	**8.81**%	21.34%	31.48%	38.62 %
C7	**7.63**%	22.33%	34.78%	43.79 %

(c) HD resolution - High Quality

	L2	L4	L6	L8
C1	**4.14**%	13.25%	24.31%	34.41 %
C2	**5.40**%	15.34%	25.36%	34.09 %
C3	6.94%	**19.38**%	32.76%	44.43 %
C4	**4.94**%	14.96%	25.58%	35.55 %
C5	3.36%	**9.77**%	17.56%	25.78 %
C6	**5.31**%	14.43%	24.59%	34.65 %
C7	**7.02**%	20.55%	34.67%	45.79 %

(d) HD resolution - Low Quality

Table 3: A relative coding gains by the given blur level for each content.

due to the size of the peripheral vision area, UHD is more robust to uneven quality degradation by blurring, and therefore, foveated coding is more beneficial for UHD.

In the future, the foveation methods, combining other FoA mechanisms with audiovisual FoA will be developed. Then, the different foveated coding algorithms, such as Flexible Macroblock Ordering (FMO) scheme for H.265/HEVC, and their impact to diverse viewing conditions (resolution, display size, environment, and context) will be investigated.

ACKNOWLEDGMENTS

This work has been performed in the framework of the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services - QUALINET and the Eurostars-Eureka Project E! 8307 - Transcoders Of the Future TeleVision (TOFuTV).

REFERENCES

[1] Itti, L., "Automatic foveation for video compression using a neurobiological model of visual attention," *Image Processing, IEEE Transactions on* **13**(10), 1304–1318 (2004).

[2] Wang, Z., Lu, L., and Bovik, A. C., "Foveation scalable video coding with automatic fixation selection," *Image Processing, IEEE Transactions on* **12**(2), 243–254 (2003).

[3] Boccignone, G., Marcelli, A., Napoletano, P., Di Fiore, G., Iacovoni, G., and Morsa, S., "Bayesian integration of face and low-level cues for foveated video coding," *Circuits and Systems for Video Technology, IEEE Transactions on* **18**(12), 1727–1740 (2008).

[4] Lee, J.-S. and Ebrahimi, T., "Efficient video coding in H.264/AVC by using audio-visual information," in [*Proc. Int. Conf. Multimedia Signal Processing*], 1–6 (Oct. 2009).

[5] Lee, J.-S., De Simone, F., and Ebrahimi, T., "Video coding based on audio-visual attention," in [*Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*], 57–60 (2009).

[6] Chen, Z., Lin, W., and Ngan, K. N., "Perceptual video coding: Challenges and approaches," in [*Multimedia and Expo (ICME), 2010 IEEE International Conference on*], 784–789 (2010).

[7] Lee, J.-S. and Ebrahimi, T., "Perceptual video compression: A survey," *Selected Topics in Signal Processing, IEEE Journal of* **6**(6), 684–697 (2012).

[8] Tang, C.-W., "Spatiotemporal visual considerations for video coding," *Multimedia, IEEE Transactions on* **9**(2), 231–238 (2007).

[9] Lee, J.-S., Simone, F. D., and Ebrahimi, T., "Efficient video coding based on audio-visual focus of attention," *J. Vis. Commun. Image R.* **22**(8), 704–711 (2011).

[10] Lee, J.-S., De Simone, F., and Ebrahimi, T., "Subjective quality evaluation of foveated video coding using audio-visual focus of attention," *Selected Topics in Signal Processing, IEEE Journal of* **5**(7), 1322–1331 (2011).

[11] Kidron, E., Schechner, Y. Y., and Eland, M., "Cross-modal localization via sparsity," *IEEE Trans. Signal Processing* **55**, 1390–1404 (Apr. 2007).

[12] ITU-R, "P.910: Subjective video quality assessment methods for multimedia applications," (1992).

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

[14] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays." International Telecommunication Union (August 2012).

[15] De Simone, F., Goldmann, L., Lee, J.-S., and Ebrahimi, T., "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *Journal of Visual Communication and Image Representation* **22**(8), 734 – 748 (2011).

[16] Einhäuser, W., Rutishauser, U., and Koch, C., "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," *Journal of Vision* **8**, 1–19 (Feb. 2008).

[17] Yarbus, A. L., [*Eye Movements and Vision*], Plenum Press, New York (1976).