# Subjective Quality Evaluation of Foveated Video Coding Using Audio-Visual Focus of Attention

Jong-Seok Lee, *Member, IEEE*, Francesca De Simone, and Touradj Ebrahimi, *Member, IEEE*

***Abstract*—This paper presents a foveated coding method using audio-visual focus of attention and its evaluation through extensive subjective experiments on both standard-definition and high-definition sequences. Regarding a sound-emitting region as the location drawing the human attention, the method applies varying quality levels in an image frame according to the distance of a pixel to the identified sound source. Two experiments are presented to prove the efficiency of the method. Experiment 1 examines the validity and effectiveness of the method in comparison to the constant quality coding for high-quality conditions. In Experiment 2, the method is compared to the fixed bit rate coding for low quality conditions where coding artifacts are noticeable. The results demonstrate that the foveated coding method provides considerable coding gain without significant quality degradation, but uneven distributions of the coding artifacts (blockiness) by the method are often less preferred than the uniform distribution of the artifacts. Additional interesting findings are also discussed, such as content dependence of the performance of the method, the memory effect in multiple viewings, and the difference in the quality perception for frame size variations.**

***Index Terms*—Audio-visual focus of attention, content dependence, foveated coding, H.264/AVC, memory effect, quality of experience, subjective quality assessment.**

## I. Introduction

In the video coding community, there has been much effort to improve coding efficiency by reducing the number of bits in an encoded video sequence and minimizing quality degradation due to the information loss. One way to achieve this goal is to exploit the focus of attention mechanisms of the human visual system. It is known that, when a human observer watches a video sequence, only a small region around the point of fixation is captured at a high resolution, whereas the resolution for the peripheral regions significantly decreases with eccentricity. Thus, if the imperceptible information outside the small fixation region is removed in an efficient way, it is possible to allocate less bits for the scene without significant impact on perceived quality. This process is often called *foveation* and results in *foveated* video sequences [1].

A key issue in foveated video coding is to determine the spatial priority of a scene by considering human focus of attention mechanisms. A bottom-up attention mechanism has been considered in [2], where salient regions containing conspicuity of intensity, color or motion are identified in a scene and spatial low-pass filtering is applied based on the saliency value of each pixel. In [3], it was assumed that moving objects draw visual attention and the static background regions were blurred before encoding. The work presented in [4] considered multiple visual attention models such as a motion attention model, a spatiovelocity visual sensitivity model and a visual masking model. A scalable video coding method was presented in [1], where the bits containing the details of face regions are placed first in an encoded bit stream so that those for the other regions may be skipped under insufficient resource conditions. Bottom-up saliency and face cues were used together in a Bayesian framework in [5]. These two information sources were also considered together in [6] to develop a foveated just-noticeable-distortion model. In [7], bottom-up cues and top-down cues such as faces and captions were considered in a scalable visual sensitivity profile generating a hierarchy of saliency maps. Visual rhythm analysis was performed in [8], from which the region-of-interest (ROI) was outlined and used for foveated video coding in H.264/AVC.

While the aforementioned work focused on visually driven attention mechanisms, the effect of an additional modality, i.e., the acoustic modality, to the visual attention has been only recently considered to determine ROIs for foveated video coding [9]. Based on the evidence that a sound source tends to draw visual attention and enhances the visual processing ability in the attended region in both top-down and bottom-up attention mechanisms [10], [11], the sound source in a scene is localized by examining the correlation between the acoustic and visual signals in the given audio-visual sequence. Then, the identified sound-emitting region is considered as the most salient area for foveated video coding. The preliminary experimental results on a small set of data showed that such an audio-visual focus of attention mechanism can be successfully exploited for producing foveated video sequences to improve coding efficiency without significant degradation of perceived quality when the algorithm parameters are adjusted properly [9].

This paper presents an extensive subjective quality evaluation study to evaluate the foveated coding method first de-

Manuscript received February 01, 2011; revised July 11, 2011 and August 08, 2011; accepted August 09, 2011. Date of publication August 18, 2011; date of current version October 19, 2011. This work was supported in part by the European Community's Seventh Framework Programme (FP7/2007-2011) under Grant 216444 (PetaMedia), and in part by the Swiss NCCR Interactive Multimodal Information Management (IM2). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Edward J. Delp.

J.-S. Lee was with the Multimedia Signal Processing Group, Institute of Electrical Engineering, Swiss Federal Institute of Technology Lausanne (EPFL), 1015 Lausanne, Switzerland. He is now with the School of Integrated Technology, Yonsei University, 406-840 Incheon, Korea (e-mail: jong-seok.lee@yonsei.ac.kr).

F. De Simone and T. Ebrahimi are with the Multimedia Signal Processing Group, Institute of Electrical Engineering, Swiss Federal Institute of Technology Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: francesca.desimone@epfl.ch; touradj.ebrahimi@epfl.ch).

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

Digital Object Identifier 10.1109/JSTSP.2011.2165199

scribed in [9], [12] thoroughly for diverse contents and viewing/coding conditions. In particular, distinguished contributions of the present work in comparison to the prior works in [9], [12] can be summarized as follows: considered coding conditions in this paper spans both low and high bit rate conditions, while only high bit rate conditions were considered in [9], [12]. Especially, since the visibility of coding artifacts in coded video sequences and consequent quality perception by observers are different for different bit rate conditions, we design an appropriate experimental approach for each condition by considering the most useful scenario of the foveated coding. In other words, for the high bit rate condition, efficient transmission of high-quality video content to users is considered and the experiment was designed so that improvement coding efficiency by foveation can be measured. For the low bit rate condition, however, the visual quality is low due to clearly visible coding artifacts even in the unfoveated coding, and a non-uniform distribution of artifacts in the scene by the foveated coding is more interesting to investigate than measuring bit rate reduction due to introduction of more artifacts; thus, fixed bit rate conditions were considered to examine the effect of the different artifact distributions in unfoveated and foveated coding to quality perception. In addition, an extensive database containing 12 standard-definition (SD) and high-definition (HD) contents is used in this work, while only two simple SD contents were used in [9] and six contents in [12]. Most of the contents used in this paper are extracted from professionally created contents (e.g., movie, music concert, and interview), which enables investigation of the validity and limitations of the foveated coding method in real applications. Overall, we perform a complete, rigorous subjective analysis of the effect of audio-visual focus of attention in the context of video coding in various viewpoints including content, bit rate, frame size, multiple viewing, and distribution of coding artifacts, and compare our results and findings with those presented in existing studies, which has not been reported previously.

First, subjective evaluation of foveated video sequences in high-quality conditions is conducted to prove the effectiveness of the foveation method for diverse contents. Here, the high-quality conditions mean that the visual quality of the coded stimuli is "acceptable," i.e., it remains "good" or even "excellent" for most subjects. In addition, the effects of various aspects such as the spatial resolution and content are analyzed to understand their impact on the effectiveness of the foveated coding method. Second, perceived quality of unfoveated and foveated video sequences encoded in low quality is compared, where the coding artifacts are clearly visible in the stimuli, in order to examine the effects of the spatial distribution of the coding artifacts under fixed bit rate conditions.

The rest of the paper is organized as follows. In the following section, theoretical background about three related topics, namely, audio-visual focus of attention, foveated video coding, and subjective quality assessment, is provided. Section III describes the foveated video coding algorithm based on audio-visual focus of attention. Section IV presents the results and analysis of the two subjective evaluation experiments. Finally, in-depth discussion and conclusion are given in Section V.

## II. Background

### A. Audio-Visual Focus of Attention

Visual attention is a cognitive process allocating resources of the human visual system. It can be classified into two categories with respect to its driving factors: *bottom-up* (or *exogenous*) and *top-down* (or *endogenous*) attention. The former is automatically induced by low-level salient features such as abrupt change or prominent appearance of color, shape, motion, orientation, contrast, and size, whereas conscious cognitive control drives the latter, e.g., road signs attracting car drivers' attention.

Often, multiple sensory modalities are involved simultaneously and interact with each other in humans' attention mechanisms. In particular, we consider the influence of auditory modality on the visual focus of attention in this paper. Such cross-modal interaction is observed in both top-down and bottom-up attention [10], [11]. In an "orthogonal cueing" experiment described in [13], each subject was asked to judge the locations (up or down) of visual targets appearing in his/her peripheral vision (left or right side of the subjects), which aimed at invoking the bottom-up attention of the subject. Shortly before the visual targets, uninformative auditory cues were presented on either the subject's left or right side, i.e., the side where an auditory cue was presented was chosen randomly so that it did not provide any information about on which side the visual target would appear. The judgments were faster and more accurate when the auditory cues occurred on the same side to that of the visual targets. This result proves that an abrupt sound draws visual attention to the spatial location of the sound source so that the subjects' visual processing capability is enhanced on the attended region, which is called *cross-modal facilitatory effect*. This effect is also observed in top-down attention, i.e., when subjects strongly expect a sound on one side in the above elevation discrimination task, visual judgments on the same side are improved [14].

Even when people are performing a visual task, a novel audio stimulus tends to capture their visual attention [15]. In [16], it was shown that such cross-modal orienting occurs even when the detailed information about the visual target is provided in order to prevent uninformative audio cues from orienting attention.

### B. Foveated Video Coding

*Fovea* is a circular region of about 1.5 mm in diameter on the retina, which has the highest density of sensor cells and takes up approximately 50% of the visual cortex in the brain. It captures the scene projected onto it at a high resolution, which covers only a small visual angle of about 2° around the center of gaze. The resolution in the peripheral region outside the fovea decreases logarithmically with eccentricity.

Motivated by this uneven visual processing capability of the retina, foveated video coding aims at maintaining a high quality only in the image region projected on the fovea and reducing the quality in the peripheral region (called *foveation*), which will improve coding efficiency without significant perceived quality degradation.

An important issue of foveated video coding is how to determine the attended area in a given visual scene. In some cases,

the content provider and viewers have common agreement on the attended area, which can be determined *a priori*; for example, in medical applications, specialists can select diagnostically important regions that will be encoded with a relatively high quality [17]. In some other applications, it may be allowed for users to determine the attended area via explicit inputs (e.g., mouse) or implicit inputs (e.g., eye-tracking), which can be used for interactive video transmission systems [18].

Opposed to these approaches, automatic selection of the attended area for foveated video coding has been extensively investigated in literature, for which various human attention models have been used. Some examples based on bottom-up attention modeling are as follows. In [2], a neurobiologically motivated attention model was proposed in order to mimic humans' bottom-up attention and implement foveated coding based on it. The model performs nonlinear integration of low level cues of conspicuity in terms of intensity, color, motion, flicker, and orientation. A spatiotemporal saliency detection model was proposed in [19], which considers the phase spectrum information of transformed intensity, color, and motion features. Top-down attention has been more popularly employed because of its intuitiveness. Frequently used top-down cues include faces [20], [21], skin regions [22], [23], moving objects [3], [24], etc. Bottom-up and top-down attention models can be combined for more accurate attention modeling. In [7], the bottom-up saliency map and top-down attention maps (face and captions) are multiplied to obtain an integrated attention map. In [5], probabilistically modeled low-level and face cues are combined in a Bayesian framework.

Although it has been shown that the aforementioned methods are effective, audio-visual interaction in attention has been rarely considered. The preliminary results reported in [9] showed that audio-visual focus of attention described in the previous subsection can be used for effective foveated video coding.

Foveation can be implemented either through pre-processing or as an embedded process in encoding. In the former approach, called offline foveation, the image frames to be encoded are processed in such a way that the quality of the peripheral region is degraded (e.g., low-pass filtering), which are inputted to an existing encoder (e.g., [2], [25]). The latter approach applies different encoding parameters for the foveated and peripheral regions (e.g., [4]), or allocates more bits on the foveated region in a rate-control scheme (e.g., [26]). The foveated coding method used in this paper is based on the scheme existing in H.264/AVC, which allows us to use different quantization parameters (QPs) for different regions.

### C. Subjective Quality Assessment

Research on quality assessment investigates how humans perceive quality of given stimuli and how automatic algorithms can imitate such perceptual processes. Quality assessment can be performed either subjectively or objectively. In subjective quality assessment, a number of subjects are asked to rate given stimuli by following a predefined procedure. Considering that the ultimate receivers of the stimuli are human subjects, this is the most accurate and reliable way to quantify of the quality of the given stimulus. Objective quality assessment tries to predict subjective quality assessment results automatically in order to reduce the complexity of the assessment process and allow the assessment to be used in real-time operations.

In order to prove the effectiveness of a video coding algorithm over a conventional one, it is necessary to show that the quality of the encoded video by the former is at least as good as or better than that of the latter. This requires quality assessment of the encoded video sequences by the two algorithms. In many cases, this is done objectively by using the peak signal-to-noise ratio (PSNR) values due to its simplicity. However, foveated video coding tries to remove perceptual redundancy in the peripheral region, which is difficult to be accounted for by PSNR measured over the whole scene. The PSNR value of the foveated region in foveated coding would be similar to or better than that in conventional coding, whereas the PSNR measure for the peripheral region would be worse in the case of foveated coding. Thus, the PSNR over the whole scene does not correctly represent the perceived quality of video sequences produced by foveated coding. Moreover, it is well-known that PSNR is not well-correlated with perceived quality [27], whereas foveated video coding is based on perceptual mechanisms of the human visual system.

Therefore, subjective quality assessment is a desirable approach for evaluation of foveated video coding. In fact, it is largely agreed that the subjective quality assessment is the ultimate way to evaluate different video coding techniques, as can be seen from the recent video coding standardization activity employing subjective assessment for identification of promising coding technologies [28].

There are a lot of environmental and contextual factors influencing results of a subjective quality assessment experiment. Thus, it is important to carefully design the experiment, including test material selection, test procedure design, environmental setup, subject screening, and subjective data processing, in order to exclude unwanted external factors and obtain reliable results. For this, there has been effort to standardize subjective test activities, e.g., [29].

The goal of the present work lies in this context, i.e., we present extensive, rigorous subjective quality assessment results in order to validate the effectiveness of the foveated video coding algorithm using the audio-visual focus of attention mechanism.

## III. Foveated Video Coding Based on Audio-Visual Focus of Attention

The procedure of the foveated video coding method is described below (Fig. 1).

The first important step is to identify where the sound signal comes from in the visual scene, i.e., to solve the audio-visual source localization problem. This is a challenging problem particularly when multiple moving objects appear in the scene but only one of them is responsible for the sound signal, since conventional motion detection approaches using only the visual information are not able to deal with this situation.

Setups with multiple microphones are frequently used to detect the direction of arrival (e.g., [30] and [31]), which are not applicable for multimedia content containing already recorded (possibly mono) audio signal. There are a few methods proposed

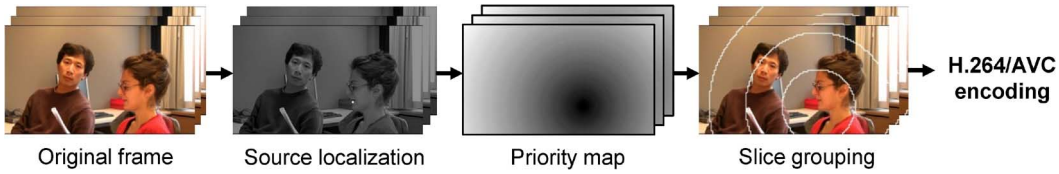


Fig. 1. Procedure of the foveated video coding method based on audio-visual focus of attention.

to solve the audio-visual source localization problem for multimedia [32], [33], but they mostly assume that the sound source is a human speaker's mouth. Thus, they cannot be applied to general multimedia content that contain non-speech audio signal. The method used in this paper does not have any assumption on the sound-emitting region and therefore a training phase requiring possibly manually labeled training data is not needed.

The audio-visual source localization method is based on the canonical correlation analysis (CCA) [34] to find the pixel location that shows the maximum correlation with the audio signal [35]. The objective of CCA is to find a pair of projection vectors $\mathbf{w}_a$ and $\mathbf{w}_v$ for the audio and visual modalities, respectively, which maximize the correlation of the projected data, i.e.,

$$\mathbf{w}_a, \mathbf{w}_v = \arg\max \frac{E\left[\left(\mathbf{w}_a^T\mathbf{a}\right)\left(\mathbf{w}_v^T\mathbf{v}\right)\right]}{\sqrt{E\left[\left(\mathbf{w}_a^T\mathbf{a}\right)^2\right]E\left[\left(\mathbf{w}_v^T\mathbf{v}\right)^2\right]}} \quad (1)$$

where $\mathbf{a}$ and $\mathbf{v}$ are the audio and visual features, respectively. If we let $\mathbf{A} = [\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+T-1}]$ and $\mathbf{V} = [\mathbf{v}_t, \mathbf{v}_{t+1}, \ldots, \mathbf{v}_{t+T-1}]$ be the collections of the acoustic and the visual feature vectors over $T$ frames, the above equation can be written as

$$\mathbf{w}_a, \mathbf{w}_v = \arg\max \frac{\mathbf{w}_a^T(\mathbf{A}^T\mathbf{V})\mathbf{w}_v}{\sqrt{\mathbf{w}_a^T(\mathbf{A}^T\mathbf{A})\mathbf{w}_a\mathbf{w}_v^T(\mathbf{V}^T\mathbf{V})\mathbf{w}_v}}. \quad (2)$$

It can be shown that solving the above problem is equivalent to solving the following [36]:

$$\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a. \quad (3)$$

It is further necessary to consider two important principles in order to effectively solve the audio-visual localization problem. The first one is the spatial sparsity, i.e., it is usually expected that the sound source is not distributed over the visual scene but spatially localized in a small region. It can be shown that, after formulation of this principle as a $l^1$-norm minimization objective, the problem in (3) becomes a constraint minimization problem [36]:

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

for $\mathbf{a} \in R^1$ and

$$\min \sum_{i=1}^{n} |w_{vi}|$$
$$\text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a,\ \mathbf{h}_k^T\mathbf{w}_a = 1 \text{ and } \mathbf{H}_k\mathbf{w}_a \geq 0 \quad (5)$$

for $\mathbf{a} \in R^m (m > 1)$, where $w_{vi}$ is the $i$th component of $\mathbf{w}_v \in R^n$, the elements of $\mathbf{h}_k$ are the binary representation of $k$ with $+1$ and $-1$, and $\mathbf{H}_k$ is a diagonal matrix whose diagonal is $\mathbf{h}_k$.

The second principle is the spatio-temporal consistency, i.e., the sound source tends to make smooth spatial movement over time. Then, the problems in (4) and (5) can be written as

$$\min \sum_{i=1}^{n} |f_i w_{vi}| \quad \text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A} \quad (6)$$

and

$$\min \sum_{i=1}^{n} |f_i w_{vi}|$$
$$\text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a,\ \mathbf{h}_k^T\mathbf{w}_a = 1 \text{ and } \mathbf{H}_k\mathbf{w}_a \geq 0 \quad (7)$$

respectively. The weighting value $f_i$ is given by

$$f_i = \max_{1 \leq j \leq n} w_{vj}^{\text{old}} - w_{vi}^{\text{old}} + 1 \quad (8)$$

where $w_{vi}^{\text{old}}$ is the $i$th component of the spatially smoothed version of the solution for the previous temporal window. A Gaussian filter is applied to the image representation of the solution for smoothing. Thus, the weighting value 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 location. Adding 1 in (8) is to ensure that all weights are greater than zero. The problems (6) and (7) can be solved by linear programming. Tracking of the sound source is performed by repeating this over time by using a moving temporal window.

The solution $\mathbf{w}_v$ can be interpreted as "cross-modal energy" concentrated on the visual features that are highly correlated to the audio signal. Therefore, the pixel location corresponding to the feature showing a high cross-modal energy is regarded as a part of the sound-emitting region.

After the sound source is localized, a priority map is generated for each frame based on the localization result. The map basically represents the Euclidean distance between each pixel and the nearest localized energy location. When there are multiple energy locations, a pixel near a smaller energy location is assigned with a larger distance than one near a larger energy source, as in a geographical contour map.

Then, the image frame is divided into $L$ partitions called slices according to the priority map. A slice is a group of macroblocks to be encoded together. Then, each slice can be decoded independently. The whole range of the priority values are linearly divided into $L$ levels, which become the boundaries of the partitions.

TABLE I
SUMMARY OF THE FORMATS AND THE CONTENTS OF THE TEST SEQUENCES. ALL SPEECH WAS IN ENGLISH

| | Format | Sound type | Sound source | Other motions |
|---|---|---|---|---|
| SD1 | V: 720×576@25 fps A: 48 kHz | Speech | A talking face | Two listening people; walking people |
| SD2 | V: 720×576@25 fps A: 48 kHz | Speech | A talking face | Moving hands |
| SD3 | V: 720×576@25 fps A: 48 kHz | Speech | A talking face | A listening person; walking people outside the window |
| SD4 | V: 720×576@25 fps A: 48 kHz | Speech | A talking face | A listening person |
| SD5 | V: 720×480@30 fps A: 48 kHz | Music | A guitar player | Two other instrument players; a camera man |
| SD6 | V: 720×480@30 fps A: 48 kHz | Music | Two hands of a piano player | The player's body; severe camera motion |
| SD7 | V: 720×576@25 fps A: 48 kHz | Speech | A talking face | Three listening people; running cars outside the window |
| SD8 | V:720×576@25 fps A: 48 kHz | Speech | A talking face | A listening person; falling snow |
| SD9 | V: 720 ×576@25 fps A: 48 kHz | Speech | A talking face | A listening person |
| HD1 | V: 1920×1080@25 fps A: 48 kHz | Speech | A talking face | A silent person walking around |
| HD2 | V: 1920×1080@25 fps A: 48 kHz | Bumping sound | A pen beating a desk | A silent person walking around |
| HD3 | V: 1920×1080@30 fps A: 48 kHz | Speech | A talking face | A listening person |

Finally, the image sequence is encoded with H.264/AVC by using the flexible macroblock ordering (FMO) scheme (Type 6) in the baseline profile to assign different QPs to different slices. The QP value for slice $j$ is given by

$$QP_j = \min\{QP_0 + j \cdot \Delta QP, 51\}, \quad j = 0, \ldots, L-1 \quad (9)$$

where $QP_0$ is the QP value for the highest priority region (i.e., sound-emitting region) and $\Delta QP$ is the incremental value of QP between each slice.

The information of slice grouping of macroblocks needs to be added in the encoded bit stream whenever the partitioning is changed. In order to minimize the overhead for such additional bits, the slice groups are updated only when more than 10% of macroblocks in a frame are assigned differently from those in the previous frame.

It is worth mentioning that, in some prior work, offline foveation (e.g., blurring) was performed and then the foveated image frames were encoded, where resulting coding gains should be considered as lower bounds of the expected gains [2], [35], [37]. In contrast, the foveation process of this method is directly embedded in the H.264/AVC encoding by using the FMO scheme.

## IV. SUBJECTIVE EVALUATION

This section presents two subjective experiments evaluating the foveated video coding method described in the previous section.

The main goal of Experiment 1 is to verify the effectiveness of the method for diverse contents and to examine to which extent it is effective without perceived quality degradation. Thus, sequences produced by the foveated coding and the constant QP mode of H.264/AVC are compared for high-quality conditions. For the sound-emitting region, the quality obtained by using the two methods is the same. However, the background region was encoded with higher QP values in the foveated coding, so that the quality for the region was degraded and a further coding gain could be obtained when compared to the constant QP mode.

Experiment 2 was designed to investigate the relationship between the attention and the spatial distribution of coding artifacts. In the foveated coding method, low bit rate conditions were considered by using large values of $QP_0$ and $\Delta QP$ in (9). The constant bit rate mode of H.264/AVC was employed for comparison, in which the target bit rates were set to be the same to those of the foveated sequences. Therefore, an uneven distribution of coding artifacts by the foveated coding (i.e., more artifacts in the background region) is compared with the uniform distribution of the artifacts by the constant bit rate mode.

Below, the database that was commonly used for the two experiments is described and then, the two experiments are explained in detail.

### A. Database

Table I lists the contents used in our work and summarizes their characteristics. Eight SD contents and three HD contents are included. It can be seen that they span a wide range of content in terms of sound types, sound sources, silent motions, etc. It contains not only speech content but also other sound sources and types such as music and bumping sound. Sources, sizes and numbers of the moving objects without sound also vary significantly, e.g., walking people, faces, moving hands, and running cars. Camera motion is included in some cases (e.g., SD6).

TABLE II
BIT RATES OF THE UNFOVEATED SEQUENCES USED IN EXPERIMENT 1

| Content | Bit rate | Content | Bit rate |
|---|---|---|---|
| SD3 | 596 kbps | HD1 | 2503 kbps |
| SD6 | 2459 kbps | HD2 | 6109 kbps |
| SD7 | 285 kbps | HD3 | 4947 kbps |

### B. Experiment 1

*1) Stimuli:* For this experiment, 3 representative SD sequences (SD3, SD6, and SD7) and 3 HD sequences (HD1, HD2, and HD3) were chosen among those in Table I.

Four different coding conditions were considered, namely, the unfoveated constant QP mode and the foveated coding with three different $\Delta QP$ values:

- the constant QP mode with a QP value of 26;
- the foveated coding with $QP_0 = 26$, $L = 4$ and $\Delta QP = 1$;
- the foveated coding with $QP_0 = 26$, $L = 4$ and $\Delta QP = 2$;
- the foveated coding with $QP_0 = 26$, $L = 4$ and $\Delta QP = 4$.

Therefore, the quality for the regions outside the sound-emitting region in the foveated coding was degraded when compared to the constant QP mode, which led to reduced bit rates.

We used the JM Reference Software [38] for H.264/AVC video coding. The rate-distortion optimization scheme was enabled. The search range of full search motion estimation was set to 32. The context adaptive variable length coding (CAVLC) was used.

The bit rates of the sequences encoded with the constant QP mode are given in Table II.

*2) Test Environment:* The test environment is intended to assure the reproducibility of the subjective test results by avoiding unwanted influences of external factors. Thus, it is important to fix features of the viewing environment such as general viewing conditions and crucial features of the used monitor.

The tests were performed in a space dedicated to professional subjective evaluations. The test room was equipped with an Eizo CG301W 30-inch LCD monitor having a response time of 6 ms and a native resolution of 2560 $\times$ 1600, which was calibrated by using EyeOne Display 2. The color of the desktop window background and the wall color were gray 128, as recommended in [29]. The ambient lighting consisted of neon lamps with 6500 K color temperature. Each subject sat in front of the monitor at a distance of 2–3 times the height of stimuli.

*3) Subjects:* Fifteen subjects (nine males and six females) participated in the tests. Their ages ranged between 20 and 35 with a mean of 28. They reported normal or corrected-to-normal vision.

*4) Procedure:* The single stimulus continuous quality scale (SSCQS) methodology was adopted for the tests [29].

When a subject sat in front of the monitor, a training session was held, where the test methodology was described by using training stimuli whose contents were different from those of test stimuli. During the test session, the subject watched each audio-visual stimulus and had 5 seconds to provide a visual quality score on a score sheet. A continuous rating scale between 0 and 100 was used. Five adjective descriptions of the ranges of the scale were also provided next to numeric scores, which were "excellent," "good," "fair," "poor" and "bad." The presentation order of the stimuli was randomized and care was taken not to show the same content consecutively. The SD and HD sequences were presented in separate sessions.

Each stimulus was shown only once except for the most degraded sequences obtained by the foveated coding, i.e., those with $\Delta QP = 4$, which were played again at the end of the test. This was to examine the memory effect, i.e., to compare the perceived quality before and after the content became familiar to subjects.

*5) Data Processing:* Screening of subjects was conducted by following the guideline described in [29], from which no subject was found as an outlier.

The mean opinion score (MOS) was computed for each stimulus by averaging the scores of all subjects for the stimulus. The confidence interval (CI) for a stimulus was obtained by assuming a Student's t-distribution of the scores as

$$\mathrm{CI} = \frac{\sigma}{M} \cdot t(1 - \alpha/2, M - 1) \tag{10}$$

where $\sigma$ is the standard deviation of the scores for the stimulus over all subjects, $M$ is the number of subjects and $t(1 - \alpha/2, M-1)$ is the t-value associated with the significance level $\alpha$ for a two-tailed test with $M - 1$ degree of freedom. We set $\alpha = 0.05$ to obtain 95% CI values.

In order to examine the statistical significance of the quality difference between the unfoveated and foveated 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.

*6) Results:* Fig. 2 shows the MOS and CI values of the four coding conditions for each content. The results of the t-test between the unfoveated and foveated sequences are shown with bars in the plots. A dark gray bar for the MOS of a foveated coding case indicates that the ratings for the corresponding foveated sequence were significantly different from those for the unfoveated, whereas a light gray bar implies the difference of the two MOS values is not significant. The relative coding gains (%) by the foveated coding are also shown below the $x$-axis.

Overall, the MOS values shown in the plots are always above 50, which shows that the stimuli of this experiment can be regarded as for high-quality conditions.

It is observed that the unfoveated sequences have the best quality in most cases and, as expected, the quality decreases as the value of $\Delta QP$ increases in the foveated coding. However, the quality difference between the two coding methods is not statistically significant when $\Delta QP$ is small except for SD7. This indicates that the foveated coding can lead improved coding efficiency without significant quality degradation. The maximum coding gain without quality degradation is 69.8% for HD3.

Content dependence of the maximum value of $\Delta QP$ corresponding to insignificant quality degradation is clearly observed. For SD7, the quality degradation for $\Delta QP = 1$ is already significant, which is mainly because in this content the audio source is the talking person appearing small in a corner of the scene.

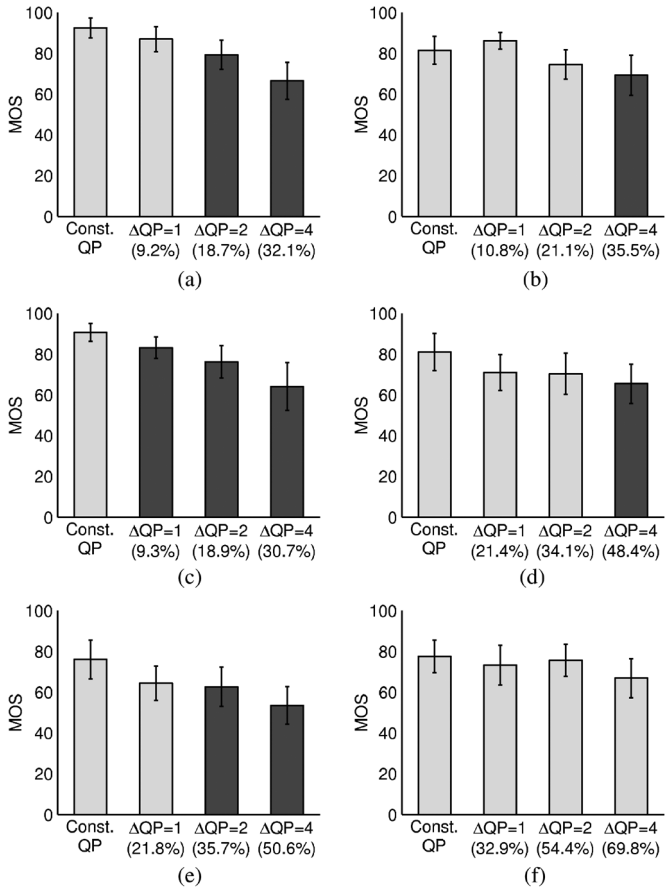


Fig. 2. Subjective test results comparing the constant QP mode and the foveated coding. A dark gray bar indicates that the MOS value of the corresponding case is significantly different from that of the constant QP coding. (a) SD3. (b) SD6. (c) SD7. (d) HD1. (e) HD2. (f) HD3.

When the SD and HD cases are compared, the foveated coding brings higher coding gains without significant quality degradation for HD sequences (only up to 21.1% in SD versus 21.8% to 69.8% in HD), even though the motion that does not produce sound is sometimes more severe in the HD content. This can be explained by a more prominent effect of the focus of attention mechanism, i.e., decreased resolution in the peripheral vision, in the HD cases when compared to SD. Especially, the HD3 sequence foveated by using even $\Delta QP = 4$ did not show significant quality deterioration when compared to the constant QP mode. This is due to the fact that this sequence is HD and, additionally, the silent motion in the background is not severe in comparison to the other contents.

Table III compares the MOS and CI values for the first and second viewings of the foveated sequences when $\Delta QP = 4$. Two-tailed t-tests were performed in order to check if the score difference is significant, and the cases with significant differences are indicated by bold faces. Although lower scores were recorded for the second viewing when compared to the first viewing, significance of the score lowering was found only in two contents (i.e., SD6 and HD2). Interestingly, these two cases contain non-speech contents, whereas the contents containing talking faces and resultant speech did not show significant differences, which implies that a talking face generally acts as a stronger attractor of visual attention. Therefore, we can conclude that, although there exists a memory effect by which subjects tend to attend regions initially unattended in the additional viewing and thus more likely to notice the coding artifacts in the regions, it is highly content-dependent in a way that the effect is prominent for speech-related contents.

TABLE III
MOS AND CI VALUES FOR THE FIRST AND SECOND VIEWINGS OF THE SEQUENCES FOVEATED BY USING $\Delta QP = 4$. THE CASES WHERE THE DIFFERENCE BETWEEN THE TWO VIEWINGS IS SIGNIFICANT ARE HIGHLIGHTED BY BOLD FACES

| | First viewing | Second viewing |
|---|---|---|
| SD3 | 66.5±9.0 | 60.5±10.1 |
| SD6 | 69.2±9.8 | **54.7±10.9** |
| SD7 | 64.1±11.8 | 55.6±8.4 |
| HD1 | 65.4±9.6 | 56.7±10.7 |
| HD2 | 53.5±9.2 | **45.0±8.6** |
| HD3 | 66.8±9.6 | 56.1±10.3 |

TABLE IV
BIT RATES OF THE FOVEATED AND UNFOVEATED SEQUENCES USED IN EXPERIMENT 2

| Content | Bit rate | Content | Bit rate |
|---|---|---|---|
| SD1 | 133 kbps | SD7 | 58 kbps |
| SD2 | 179 kbps | SD8 | 400 kbps |
| SD3 | 119 kbps | SD9 | 106 kbps |
| SD4 | 75 kbps | HD1 | 256 kbps |
| SD5 | 212 kbps | HD2 | 702 kbps |
| SD6 | 462 kbps | HD3 | 319 kbps |

### C. Experiment 2

*1) Stimuli:* In this experiment, all of the 12 contents in Table I were used. For the foveated coding, we set $QP_0 = 38$, $L = 4$ and $\Delta QP = 8$, so that coding artifacts are clearly visible. The constant bit rate mode of H.264/AVC was used to produce video sequences with a uniform QP in each frame. The target bit rates were set to the same as those of the corresponding foveated sequences, which are shown in Table IV. It is observed that the bit rate range is much lower than that shown in Table I.

*2) Test Environment:* The same test environment as that in Experiment 1 was used.

*3) Subjects:* Eleven subjects (seven males and four females) participated in the tests, whose ages were between 20 to 35 with a mean of 27. They reported normal or corrected-to-normal vision.

*4) Procedure:* Each pair of a foveated and unfoveated sequences having the same bit rate were played one after the other as stimulus A and stimulus B in a random order, along with a 3-second-long gray 128 image between the two stimuli. Then, the subject had 5 seconds to provide the preference between the two stimuli in terms of visual quality on a score sheet. Three preference options were given, namely, "A," "B," and "same."

In order to examine the reliability of the subject's voting, the two presentation orders of a stimuli pair (i.e., A versus B and B versus A) were both included in the test session, which was used for subject screening. As in Experiment 1, a training session took place before the test session of each subject.

*5) Data Processing:* Outlier detection in paired comparison was conducted differently from Experiment 1. For each subject, the voting results for the two cases with different orders of a

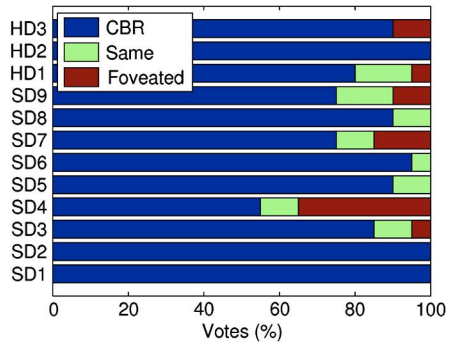


Fig. 3. Results of paired comparison in term of percentages of the three preference options, i.e., the constant bit rate (CBR) mode, tie, and foveated coding.

stimulus pair were compared and the cases where the two did not agree were counted. The subjects showing such "inconsistency" for more than 15% of the total ratings were considered as outliers. As a result, one subject was identified as an outlier and his ratings were discarded. The counts of the three preferences across the entire panel of subjects for each stimulus are reported in the following.

*6) Results:* In Fig. 3, the results of the paired comparison test are summarized as the percentages of the three preferences. It can be noticed that in most cases the constant bit rate mode, where the coding artifacts are visible over the whole scene, is preferable to the foveated coding showing uneven quality distribution in the image frames. One exception is SD4, where a significant percentage of subjects preferred the foveated sequences. A reason behind this could be that the scene of SD4 contains a talking face appearing in the foreground, which is large enough to fix the attention on the face area and to overlook the artifacts in the background. In this experiment, no clear difference in SD and HD sequences was observed.

## V. Conclusion and Discussion

We have presented an extensive study for subjective evaluation of the fovieated coding based on audio-visual focus of attention for diverse contents and coding conditions, and analyzed the results in various viewpoints. It was shown that the method can be effectively used to improve coding efficiency without significant degradation of perceived quality when the quality of the coded sequences is acceptable or higher.

The extent to which the foveation can be introduced without perceived quality degradation was investigated in Experiment 1 by varying the value of $\Delta QP$. Since the value of $L$ was fixed, the limitation in the coding gain is not definitive and higher coding gains could be obtained by adjusting both $L$ and $\Delta QP$. It was shown that the maximal coding gain is highly dependent on the content type and image frame size. Especially, several factors in the content are directly or indirectly involved in affecting viewing patterns of the subjects, as confirmed by previous studies as well. In [39], it was shown that selection of fixation points for multimodal conditions depends on the saliencies of both auditory and visual stimuli, because different unimodal saliencies are integrated before the subsequent fixation point is determined. Therefore, it would be necessary to consider such content dependence of audio-visual focus of attention in order to improve the focus of attention model and thereby the usefulness of the foveated video coding.

It is also worth mentioning that, 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 condition might not be the same to normal free-viewing. In fact, it has been shown that given task demands can affect viewing patterns of observers significantly. In [40], it was demonstrated that the pattern of eye movement is clearly dependent on the instructions given to the observers in viewing a painting. Sometimes, sensory-driven bottom-up saliency features are immediately overridden by task demands [41]. Therefore, it can be reasoned that in the free-viewing scenario, the effectiveness of the foveation would be more significant in comparison to what was measured in our experiments.

As in the results of Experiment 1, some previous studies confirmed that the memory effect exists in multiple viewings of the same content. In [37], repeated viewings of the same stimuli cause changes in gaze patterns of observers, so that the background region having poor quality tends to be attended after multiple viewing. The work in [42] also showed that perceptual memory leads to increases in the impact of top-down influences on attentional selection during natural vision. These observations may raise a question about the effectiveness of the foveated coding. However, it should be noted that the results in [37], [42] and ours were drawn from different conditions, i.e., absence and existence of the sound signal. In our experiment, the memory effect was not clearly observed for the speech contents, which might be because a talking face accompanied with speech was interesting enough even in the second viewing. Moreover, in multimedia experience in real life, repeated viewing of the same content in a short time period is much less usual than in laboratory-based subjective quality experiments. Therefore, considering its significant coding gain, the foveated coding method can be used effectively for real-time applications (e.g., videoconferencing) where the replay is not required and thus no memory effect would be involved in.

In Experiment 2, it was shown that for low-quality conditions the subjects usually prefer the uniform distribution of coding artifacts to the uneven distribution, even though the quality for the sound-emitting regions is better in the latter case. This could be due to the fact that the strong artifacts in the background region of the foveated image frames drew the attention of the observers. In [37], it was shown that foveation does not change significantly the gaze pattern of human observers. Also, the eye tracking experiment presented in [43] showed that gaze fixation locations are not significantly altered by visible coding artifacts. Nevertheless, the impairment in the background region of the foveated sequences used in Experiment 2 are extremely severe in comparison to those used in [37], [43]. Additionally, the artifacts in [37] and [43] were uniformly distributed, while in our case they were not. Further analysis such as eye-tracking experiments for a wide range of the temporal and spatial distributions of foveation artifacts would allow to validate our interpretation of the results.

As shown in our experiment, the display size, viewing environment and viewing context are important for determining

the viewing pattern of foveated video sequences. Therefore, it would be interesting in the future to investigate quality of experience of foveated sequences under various viewing conditions including mobile, immersive, task-free and unattended conditions. In addition, the content dependence of the effectiveness of the foveated coding would need to be explored by considering more diverse audio-visual contents. In our future work, other focus of attention mechanisms such as bottom-up visual saliency models and top-down attention models (e.g., faces, text) will be combined with the audio-visual focus of attention in order to develop more sophisticated and effective foveation methods.

## REFERENCES


[1] Z. Wang, L. Lu, and A. C. Bovic, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[2] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[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] C.-W. Tang, "Spatiotemporal visual considerations for video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231–238, Feb. 2007.

[5] 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.

[6] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.

[7] Q. Chen, G. Zhai, X. Yang, and W. Zhang, "Application of scalable visual sensitivity profile in image and video coding," in *Proc. IEEE Int. Symp. Circuits Syst.*, Seattle, WA, May 2008, pp. 268–271.

[8] M.-C. Chi, C.-H. Yeh, and M.-J. Chen, "Robust region-of-interest determination based on user attention model through visual rhythm analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 1025–1038, Jul. 2010.

[9] J.-S. Lee and T. Ebrahimi, "Efficient video coding in H.264/AVC by using audio-visual information," in *Proc. Int. Conf. Multimedia Signal Process.*, Rio de Janeiro, Brazil, Oct. 2009, pp. 1–6.

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

[11] C. Spence, "Crossmodal spatial attention," *Ann. New York Acad. Sci.*, vol. 1191, pp. 182–200, Mar. 2010.

[12] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Efficient video coding based on audio-visual focus of attention," *J. Vis. Commun. Image R.*, 2010, doi: 10.1016/j.jvcir.2010.11.002.

[13] C. Spence and J. Driver, "Audiovisual links in exogenous covert spatial orienting," *Percept. Psychophys.*, vol. 59, no. 1, pp. 1–22, Jan. 1997.

[14] C. Spence and J. Driver, "Audiovisual links in endogenous covert spatial attention," *J. Exper. Psychol.: Human Percept. Perf.*, vol. 22, no. 4, pp. 1005–1030, Aug. 1996.

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

[16] V. Mazza, M. Turatto, M. Rossi, and C. Umilta, "How automatic are audiovisual links in exogenous spatial attention?," *Neuropsychologia*, vol. 45, no. 3, pp. 514–522, 2007.

[17] M. G. Martini and C. T. E. R. Hewage, "Flexible macroblock ordering for context-aware ultrasound video transmission over mobile WiMAX," *Int. J. Telemed. Appl.*, vol. 2010, pp. 1–14, 2010.

[18] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," in *Proc. SPIE*, San Jose, CA, Jan. 1998, vol. 3299, pp. 294–305.

[19] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[20] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 134–139, Jan. 2008.

[21] M.-C. Chi, M.-J. Chen, C.-H. Yeh, and J.-A. Jhu, "Region-of-interest video coding based on rate and distortion variations for H.263 +," *Signal Process.: Image Commun.*, vol. 23, pp. 127–142, 2008.

[22] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 551–934, Jun. 1999.

[23] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit-rate coding of image sequences using adaptive regions of interest," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 8, pp. 928–934, Dec. 1998.

[24] Y. Sun, I. Ahmad, D. Li, and Y.-Q. Zhang, "Region-based rate control and bit allocation for wireless video transmission," *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 1–10, Feb. 2006.

[25] H. R. Sheikh, B. L. Evans, and A. C. Bovik, "Real-time foveation techniques for low bit rate video coding," *Real-Time Imaging*, vol. 9, pp. 27–40, 2003.

[26] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, Jan. 2011.

[27] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[28] F. D. Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi, "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *J. Vis. Commun. Image R.*, 2011, doi: 10.1016/j.jvcir.2011.01.008.

[29] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Rec. ITU-R BT.500-11, 2002, Geneva, Switzerland.

[30] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 601–616, Feb. 2007.

[31] P. Perez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proc. IEEE*, vol. 92, no. 3, pp. 495–513, Mar. 2004.

[32] V. Pavlovi\'c, A. Garg, J. M. Rehg, and T. S. Huang, "Multimodal speaker detection using error feedback dynamic Bayesian networks," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Hilton Head Island, SC, 2000, pp. 34–41.

[33] P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 63–73, Jan. 2008.

[34] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[35] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Video coding based on audio-visual attention," in *Proc. Int. Conf. Multimedia Expo*, New York, Jun. 2009, pp. 57–60.

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

[37] M. Nyström and K. Holmqvist, "Effect of compressed offline foveated video on viewing behavior and subjective quality," *ACM Trans. Multimedia Comput., Commun., Applicat.*, vol. 6, no. 1, pp. 1–14, Feb. 2010.

[38] H.264/AVC JM Reference Software, 2008 [Online]. Available: http://iphome.hhi.de/suehring/tml/

[39] S. Onat, K. Libertus, and P. König, "Integrating audiovisual information for the control of overt attention," *J. Vis.*, vol. 7, no. 10, pp. 1–16, Jul. 2007.

[40] A. L. Yarbus, *Eye Movements and Vision*. New York: Plenum, 1976.

[41] W. Einhäuser, U. Rutishauser, and C. Koch, "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," *J. Vis.*, vol. 8, no. 2, pp. 1–19, Feb. 2008.

[42] R. Carmi and L. Itti, "The role of memory in guiding attention during natural vision," *J. Vis.*, vol. 6, no. 9, pp. 898–914, Aug. 2006.

[43] O. L. Meur, A. Ninassi, P. L. Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?," *Signal Process.: Image Commun.*, vol. 25, no. 8, pp. 597–609, Sep. 2010.

**Jong-Seok Lee** (M'06) received the Ph.D. degree in electrical engineering and computer science from KAIST, Daejeon, Korea, in 2006.

From 2006 to 2008, he was a Postdoctoral Researcher and an Adjunct Professor at KAIST. From 2008 to 2011, he was a Research Scientist in the Multimedia Signal Processing Group, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. Currently, he is an Assistant Professor at the School of Integrated Technology, Yonsei University, Incheon, Korea. His research interests include audio-visual signal processing, multimedia quality assessment and multimodal human–computer interaction. He is author or coauthor of over 40 publications.

Prof. Lee was the chair of the First Spring School on Social Media Retrieval held in 2010 and an organizing committee member of its second edition in 2011. He is a member of Multimedia Communication Technical Committee of the IEEE Communication Society.

**Francesca De Simone** was born in Italy in 1983. She received the B.S. and M.S. degrees in electronic engineering from Università degli Studi Roma Tre, Rome, Italy, in 2004 and 2006, respectively. She is currently pursuing the Ph.D. degree at the Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland.

From August 2006 until November 2006, she was a Research Assistant at the Institute of Signal Processing, Tampere University of Technology, Tampere, Finland. Since May 2007, she has been a Research Assistant at the Multimedia Signal Processing Group, EPFL. Her current research interests include subjective and objective multimedia quality assessment.

**Touradj Ebrahimi** (M'92) received the M.Sc. and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland, in 1989 and 1992, respectively.

From 1989 to 1992, he was a Research Assistant at the Signal Processing Laboratory, EPFL. During the summer 1990, he was a Visiting Researcher at the Signal and Image Processing Institute, University of Southern California, Los Angeles. In 1993, he was a Research Engineer at the Corporate Research Laboratories, Sony Corporation, Tokyo, Japan. In 1994, he served as a research consultant at AT&T Bell Laboratories. He is currently a Professor heading the Multimedia Signal Processing Group at EPFL, where he is involved with various aspects of digital video and multimedia applications. He is author or coauthor of over 100 papers and holds 10 patents.