J. Vis. Commun. Image R. 22 (2011) 704–711

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

# Efficient video coding based on audio-visual focus of attention

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

*Multimedia Signal Processing Group (MMSPG), Institute of Electrical Engineering (IEL), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

ARTICLE INFO

*Article history:*
Available online 19 November 2010

*Keywords:*
Video coding
Audio-visual focus of attention
Quality of experience
Audio-visual source localization
H.264/AVC
Flexible macroblock ordering (FMO)
Canonical correlation analysis
Subjective quality assessment

ABSTRACT

This paper proposes an efficient video coding method using audio-visual focus of attention, which is based on the observation that sound-emitting regions in an audio-visual sequence draw viewers' attention. First, an audio-visual source localization algorithm is presented, where the sound source is identified by using the correlation between the sound signal and the visual motion information. The localization result is then used to encode different regions in the scene with different quality in such a way that regions close to the source are encoded with higher quality than those far from the source. This is implemented in the framework of H.264/AVC by assigning different quantization parameters for different regions. Through experiments with both standard and high definition sequences, it is demonstrated that the proposed method can yield considerable coding gains over the constant quantization mode of H.264/AVC without noticeable degradation of perceived quality.



## 1. Introduction

It is well known that the human attention mechanisms play an important role in visual perception. When humans observe a scene in a video sequence, only a small region around the point of fixation is captured at a high resolution, while the resolutions for the peripheral regions drastically decrease with eccentricity. Therefore, even if visual quality degradations appear in peripheral regions, they may not be noticeable to human viewers. This implies that it may not be necessary to encode the whole scene with a uniform quality; compression efficiency can be achieved by discarding imperceptible information outside the small fixation region without significant impact on perceived quality. Several techniques have been proposed for such visual attention-based coding, where spatial prioritization schemes determine the priorities of different regions in the scene and encoding is performed according to those priorities. The priority can be determined through bottom-up saliency detection, face detection, moving object detection, and so on [1–6].

Although previous attention-based coding approaches have been shown to be effective through subjective and objective experiments, they are still far from fully exploiting the actual attention mechanism of the human visual system. One of the interesting and important aspects of the human visual system is the visual attention guided by the acoustic modality, which has been rarely addressed previously. One can imagine that audio events around us often attract our visual attention, e.g. making us turn our heads or eyes toward the sound source. Psychological studies have demonstrated the importance of the acoustic cues in the visual attention due to their cross-modal interaction. An auditory stimulus in a particular location tends to draw visual attention occurring at the same spatial location [7]. Moreover, such orientation of attention can improve perception of the subsequent visual stimulus [8].

Motivated by these observations, we propose a novel video coding method based on audio-visual focus of attention. It is assumed that the sound source and its neighboring region in multimedia content is the region-of-interest (ROI) that attracts human observers' visual attention. First, we present an audio-visual source localization algorithm in which, for a given video sequence accompanied with an audio channel, the region emitting the sound is localized by analyzing the correlation between the acoustic and the visual signals. The algorithm has advantages in that it does not impose any assumption on the sound source and can be applied to conventional audio–video sequences containing only one audio channel. Then, we propose an optimized video coding scheme using the audio-visual source localization result. Each frame of the visual sequence is divided into several regions according to their spatial distance from the sound source. By utilizing the flexible macroblock ordering (FMO) scheme in H.264/AVC, each region is mapped into a slice that is encoded with a different quantization parameter (QP). For a slice far from the sound source, a large QP value is assigned so that a small number of bits are allocated for encoding of the slice in comparison to a slice near the sound source. The experimental results on various standard definition (SD) and high definition (HD) contents show that the proposed method can improve coding efficiency in comparison to the

* Corresponding author.
*E-mail addresses:* jong-seok.lee@epfl.ch (J.-S. Lee), francesca.desimone@epfl.ch (F. De Simone), touradj.ebrahimi@epfl.ch (T. Ebrahimi)

1047-3203/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2010.11.002

conventional constant QP mode of H.264/AVC. Especially, thorough subjective quality evaluation tests were conducted in order to investigate the relationship between the obtained coding gain and the perceived quality degradation, which confirms that considerable coding gains can be achieved without perceptible degradation of quality.

The rest of the paper is organized as follows. The next section introduces previous work related to the proposed method. In Section 3, an audio-visual source localization algorithm is described. Section 4 presents the proposed efficient video coding strategy that exploits the source localization results. Section 5 demonstrates the effectiveness of the proposed method through objective and subjective experiments on both SD and HD sequences. Finally, concluding remarks are given in Section 6.

## 2. Related work

### 2.1. Audio-visual focus of attention

In humans' attention, both auditory and visual sensory modalities are often involved simultaneously and may influence each other. In particular, there exist two distinct forms of attention. The "exogenous" attention is stimulus-driven and captured reflexively by events in a bottom-up manner; the "endogenous" attention requires a conscious decision and is directed voluntarily in a top-down manner [7].

In exogenous spatial attention, it has been proved that a spatially non-predictive cue in one modality can attract covert attention toward the location of the cue in the other modality, which is called the "cross-modal facilitatory effect" [9]. For example, an abrupt sound draws visual attention to the spatial location of the sound source. In the experiments by Spence and Driver [9], subjects were asked to judge the elevation (up or down) of peripheral visual targets that follow uninformative auditory cues on either their left or right side; the judgments were faster and/or more accurate for the visual targets preceded by the auditory cues occurring on the same side.

Similar cross-modal facilitatory effects are also observed in endogenous attention. Spence and Driver conducted similar experiments to those for exogenous attention [10]. The elevation judgments for either auditory or visual targets were faster when the subjects expected a stimulus of the other modality in that side. This proves that attending to stimuli of one modality at a given location enhances processing of stimuli of the other modality at the same spatial location.

In [11], it was shown that, even when people are performing a visual task, a novel auditory stimulus can capture their visual attention. Such cross-modal orienting is automatic in a sense that it occurs even when detailed information about the visual target is given to the subjects in order to prevent uninformative auditory cues from orienting attention [12].

There are some other interesting phenomena related to the audio-visual perception and attention. The "ventriloquist effect" refers to the illusion that, when synchronous auditory and visual information is presented in slightly separate locations, the perceived location of the sound is biased to the direction of the visual stimulus [13]. The "freezing phenomenon" shows that, when a rapidly changing visual stimulus is shown, an abrupt sound may freeze the visual scene with which the sound is synchronized, i.e. it is perceived as if the scene is shown for a longer time [14].

Audio-visual speech perception is an example of the advantage of exploiting the link between audio-visual attention and perception. If one has a problem in listening to the spoken language due to the environmental noise, it is useful to observe the lip movements or the gesture to understand the speech better via integration of the acoustic and the visual stimuli [15]. On the contrary, discrepancy of the two modalities may result in an illusion due to their conflict in audio-visual speech perception, as demonstrated by the McGurk effect [16].

Finally, neurological analysis of the human brain also shows evidences of multimodal information processing. When different senses reach the brain, the sensory signals converge to the same area in the superior colliculus and a large portion of the neurons leaving the superior colliculus are multisensory [17]. Additionally, neuroimaging studies have shown that not only sensory-specific cortices anatomically converge into multisensory brain areas, but also multisensory spatial interactions conversely affect unimodal brains [18].

### 2.2. Attention-based video coding

As mentioned in the introduction, the attention mechanisms of the human visual system can be exploited to improve efficiency of video compression schemes. When human subjects observe a scene, only a small region around a point of fixation is captured at a high spatial resolution while resolutions for the peripheral regions dramatically decrease with eccentricity. Therefore, if information in a region far from the fixation point is appropriately discarded, compression efficiency can be achieved without significant perceived quality degradations.

Various models have been proposed in literature to define the attended region and perform efficient video coding. Itti [1] used a neurobiologically motivated bottom-up model to identify the salient regions in an image. The model identifies conspicuity in terms of intensity, color, motion, etc. in the scene and smoothing is performed based on the saliency values of each pixel. The work in [2] combined the bottom-up cues and the top-down information (i.e. human faces) in a Bayesian framework to improve the selection of the fixation model. In [3], a scalable video coding method was presented based on the human foveation model by assuming that face regions are the points of fixation. The encoded bit stream is organized in a way that the bits containing the details of expected attended regions are sent first while those for the other regions may be dropped according to the given bitrate condition. Chen et al. [4] proposed a scalable visual sensitivity profile that generates a hierarchy of saliency maps prioritizing both the bottom-up cues as in [1] and the top-down cues such as faces and captions, and used it for scalable video coding. Cavallaro et al. [6] assumed that moving objects draw attention. The moving objects in the scene were detected and then a Gaussian low-pass filter was applied to the background region to improve compression efficiency. Tang [5] combined different visual factors of attention, i.e. a motion attention model, a spatiovelocity visual sensitivity model and a visual masking model, and coding efficiency was achieved by varying the QP value for the attended region and the background region.

Although the aforementioned methods have been shown to be effective, multimodal (i.e. audio-visual) interaction in attention has been rarely considered. Our preliminary results showed that such interaction can be used for efficient video coding [19], which is extended in this paper with an improved video coding algorithm and a more extensive and thorough performance study.

## 3. Audio-visual source localization

Audio-visual source localization aims at identifying the spatial location of the region producing sound in each frame of a video sequence. It is useful when the scene contains multiple moving objects but only one object is responsible for the sound and thus a conventional motion detection approach with only the visual information may not work satisfactorily.

Frequently, multiple microphones such as stereo microphones or microphone arrays are used for solving the problem of

audio-visual source localization, so that the time delay of arrival can be used to estimate the direction of sound sources. In [20], a stereo audio and cycloptic vision sensor was used to track speaking people by using a Kalman filter technique. Gatica-Perez et al. [21] used an array of eight microphones and three cameras for tracking speakers in meeting scenarios. Perez et al. [22] presented a visual tracking method based on the particle filter; the color information is dominantly used for tracking, while the motion information and the acoustic signal acquired by stereo microphones are additionally used according to necessity.

It also becomes important to perform source localization for multimedia content which is prevalent on the Internet. In such cases, only the recorded acoustic and visual data are available in the sequences, and thus the aforementioned methods cannot be used since they require special hardware setups. Pavlović et al. [23] presented a method in which a person is tracked by using modules such as a face detector, a skin color detector, a mouth motion detector and a face texture detector in a casino scenario. And, the dynamic Bayesian model was used for joint audio-visual modeling. Cutler and Davis [24] modeled audio-visual correlations with a time-delayed neural network that receives acoustic and visual features as its inputs and produces an output indicating whether there is a speaking activity or not. Besson et al. [25] proposed a method that finds the optimal linear transformation of the acoustic features so that the mutual information between the transformed acoustic and the visual features becomes the maximum, which is for detecting a speaker among multiple persons in a scene.

As seen in the above examples, a popular application of audio-visual source localization is tracking of a speaking person. In this case, models or *a priori* knowledge about the appearance of humans or faces are frequently utilized [21,23,25]. In more general scenarios, however, a sounding object may not be a speaking person. Then, it is impossible to have assumptions on the objects to be localized (e.g. the shape or color of faces). While Perez et al. [22] claimed that they alleviated necessity of models for tracked objects, a less restrictive object model (i.e. reference color model) was still needed.

The audio-visual source localization method presented in this paper does not require any special setup with multiple microphones, but only one channel acoustic signal is used. Moreover, it does not have any assumption on the region or object to be localized. As a result, it does not require a training step that may need manually processed or labeled training data.

Our method uses the canonical correlation analysis (CCA) [26] and finds the pixel location which shows the maximum correlation with the acoustic signal [27]. It is based on the method presented in [28], but an important improvement has been introduced, as described below.

The first step of the method is to extract features for the acoustic and the visual modalities. Then, the correlation between the two feature vectors is exploited in the following way. Let $\mathbf{a} \in R^m$ and $\mathbf{v} \in R^n$ be the $m$-dimensional acoustic and $n$-dimensional visual feature vectors, respectively. The objective of CCA is to find a pair of projection vectors $\mathbf{w}_a$ and $\mathbf{w}_v$ that maximize the correlation of the projected features. The projection vectors are obtained by

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

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 within a temporal window, respectively. Then, the above equation is written as

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

The above problem can be solved as an eigenvalue problem. Alternatively, it can be shown that solving the above maximization problem is equivalent to solving the following equation [28]:

$$\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a \tag{3}$$

If $\mathbf{V}$ is full rank, there are an infinite number of solutions for the above equation because the dimension of $\mathbf{w}_v$ is usually much larger than $T$. In order to alleviate this, a spatial sparsity criterion is imposed for finding a unique solution [28]:

$$\min \|\mathbf{w}_v\|_1 \quad \text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A} \tag{4}$$

for $m = 1$, where $\|\cdot\|_1$ is the $l^1$-norm. If $m > 1$, the following optimization is solved for each of $k = 1, 2, \ldots, 2^{m-1}$:

$$\min \|\mathbf{w}_v\|_1 \quad \text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a,$$
$$\mathbf{h}_k^T\mathbf{w}_a = 1 \quad \text{and} \quad \mathbf{H}_k\mathbf{w}_a \geqslant 0 \tag{5}$$

where 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 $\mathbf{h}_k$. Then, the one giving the smallest objective value is chosen for the final solution. The linear programming can solve the above constrained optimization problems. The solution of (4) or (5) can be interpreted as the "cross-modal energy" concentrated on the visual features responsible for the acoustic signal. Therefore, the pixel location corresponding to the visual feature having a high cross-modal energy is identified as (a part of) the sound-emitting region. The sound source is tracked over time by applying the above procedure repetitively through a moving temporal window.

A limitation of the above algorithm is the lack of consideration of spatio-temporal consistency of the solutions over multiple frames. In order to consider such consistency and improve the tracking performance, a weighting scheme is incorporated in the above formulation. Then, the problems (4) and (5) are modified as

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

and

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

respectively. Here, $w_{vi}$ is the $i$th component of $\mathbf{w}_v$ and $f_i$ is the weighting factor given by

$$f_i = \max_{1 \leqslant j \leqslant n} w_{vj}^{old} - w_{vi}^{old} + 1 \tag{8}$$

where $w_{vi}^{old}$ is the $i$th component of the spatially smoothed version of the solution for the previous temporal step. Smoothing is done by applying a Gaussian filter to the image representation of the solution. In other words, if a pixel location has received a large energy value in the solution for the previous temporal step, the weight values for the location and its neighboring pixels are set to be small so that large values are obtained for these locations in the solution for the current temporal step. The spatial smoothing allows some margin in consistency by assigning low weights not only to the identified sound source location in the previous step but also to its neighboring region. Adding 1 in (8) is to ensure that all weights are greater than zero. The problems (6) and (7) are also solved by linear programming.

An example of localization is shown in Fig. 1(a) and (b): in the scene, the hand plays the guitar producing sound, while the wooden horse in the right part of the image is rocking back and forth. The white dot indicates the location detected by the method described above.

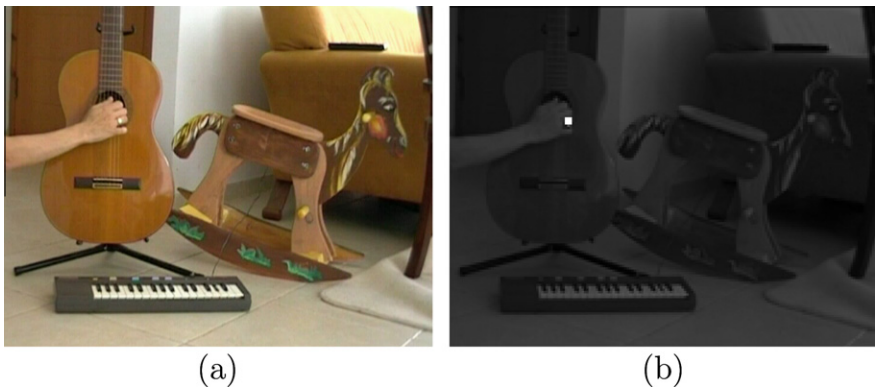


**Fig. 1.** Example image frame where a hand produces the guitar sound while a wooden horse is rocking at the same time. (a) Original frame. (b) Source localization result (indicated by the white dot) overlaid on the image.

## 4. Video coding based on audio-visual attention

The result of the audio-visual source localization explained in the previous section is used to determine which region is to be encoded in higher and which in lower quality. The quality of a region is controlled by varying the QP value in H.264/AVC. By assuming that the sound-emitting region is attended more than the other regions, a large QP is assigned for the region far from the sound source, thereby gain in terms of coding efficiency can be achieved without noticeable degradation of perceived quality.

First, a priority map for each frame is generated based on the localization result, which represents the weighted distance between each pixel and the nearest localized cross-modal energy location (Fig. 2(a)). When there are multiple energy locations, a pixel near a smaller energy location is assigned a larger distance than one near a larger energy location, just as in a contour map. The Euclidean distance is used to measure the distance between pixels. It is possible to monotonically scale the priority values, e.g. by applying logarithm or exponentiation. However, such scaling was not necessary in our experiments. Then, the image frame is divided into $L$ partitions (called slices) according to the priority map. The linearly spaced values within the range of the priority values are assigned as the boundaries of the partitions.

The FMO scheme in H.264/AVC is used for assigning different QPs to different slices. A slice is a group of macroblocks to be encoded together. Each slice can be decoded independently. In H.264/AVC, there are six different pre-defined types of grouping patterns (Types 0–5). However, in our case, a slice can have an arbitrary shape depending on the priority map, which does not correspond to pre-defined Types 0–5 but Type 6 in FMO. Fig. 2(b) and (c) show examples of slice grouping with different values of $L$.

The QP for slice $j$, $QP_j$, is determined by

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

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

In order to minimize the overhead for sending additional bits containing the information of partitioning macroblocks into slices, the slice groups are updated only when more than 10% of macroblocks in a particular frame are assigned differently from those in the previous frame.

## 5. Experiments

### 5.1. Setup

In our experiments, six test audio-visual sequences having a length of 10 s were used: three SD and three HD sequences. Their characteristics are summarized in Table 1. It can be seen that they span a wide range of content in terms of sound types, sound sources, silent motions, etc. No scene change was included in the sequences by assuming that scene changes are detected and then the source localization algorithm is applied to each scene segment.

For source localization, we used the difference of the luminance component of consecutive frames as visual features. The energy of audio samples within a moving window was extracted and its temporal difference was used as acoustic features. The window moved at the rate of the visual frame rate so that the acoustic and visual features are temporally synchronous. The value of $T$, i.e. the number of frames to be analyzed, was set to twice the visual frame rate.

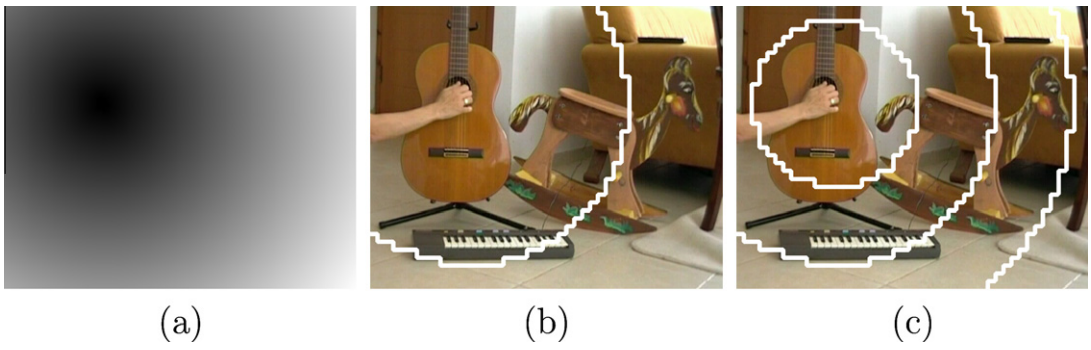


**Fig. 2.** Video coding based on the localization result shown in Fig. 1(b). (a) Priority map (shown with small brightness values for high priorities). (b) Slice boundary for $L = 2$. (c) Slice boundaries for $L = 4$.

**Table 1**
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<br>*A*: 48 kHz | Speech | A talking face | A listening person; walking people outside the window |
| SD2 | *V*: 720 × 480@30 fps<br>*A*: 48 kHz | Music | Two hands of a piano player | The player's body; severe camera motion |
| SD3 | *V*: 720 × 576@25 fps<br>*A*: 48 kHz | Speech | A talking face | Three listening people; running cars outside the window |
| HD1 | *V*: 1920 × 1080@25 fps<br>*A*: 48 kHz | Speech | A talking face | A silent person walking around |
| HD2 | *V*: 1920 × 1080@25 fps<br>*A*: 48 kHz | Bumping sound | A pen beating a desk | A silent person walking around |
| HD3 | *V*: 1920 × 1080@30 fps<br>*A*: 48 kHz | Speech | A talking face | A listening person |

For H.264/AVC video coding, we used the JM Reference Software version 15.1 [29]. The constant QP mode and the proposed method were compared. The former implies the case where the whole image frame is encoded by using a single QP without the FMO scheme. The QP value in this case corresponds to that for the region of the highest priority (i.e. the region including the sound source) in the proposed method. The encoder used the baseline profile in order to use the FMO scheme. 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.

### 5.2. Source localization results

First, the performance of our localization algorithm is shown in comparison to that in [28]. As mentioned in Section 3, the result of the source localization appears as the cross-modal energies located in the pixel locations of the estimated source. The locations and the magnitudes of these energies are evaluated by using two performance measures. First, the accuracy measure is defined by the ratio of the energies concentrated in the sound-emitting region to those in the whole image frame. Thus, it ranges from 0 (i.e. failure of localization) to 1 (i.e. successful localization without error). Its value is 0 when the maximum energy is located inside the sound-emitting region, and its large value indicates a large error of localization. Second, the pixel distance between the maximum localized energy and the sound-emitting region is used. In order to compute the two performance measures, the sound-emitting region in each sequence was identified manually and used as the ground truth.

Table 2 shows the localization performance of the two methods in terms of the average and standard deviation values of the two measures over time. From the table, it is observed that the localization results in terms of the accuracy and the error distance are significantly improved by considering spatio-temporal consistency in our method. The proposed method shows successful localization results even when there exists a moderate global camera motion in the case of SD2. This content contains translational and rotational movements around the piano player's hands. However, the analysis considering spatio-temporal consistency for each of the short temporal segments was able to localize the sound source at a high accuracy.

**Table 2**
Source localization performance of the conventional and proposed methods. The average and standard deviation values over time are shown.

| | Accuracy | | Distance | |
|---|---|---|---|---|
| | Conventional | Proposed | Conventional | Proposed |
| SD1 | 0.44 ± 0.34 | 0.99 ± 0.00 | 80.4 ± 90.8 | 0.1 ± 1.6 |
| SD2 | 0.22 ± 0.26 | 0.95 ± 0.13 | 100.6 ± 77.6 | 7.6 ± 34.3 |
| SD3 | 0.19 ± 0.19 | 0.98 ± 0.04 | 177.9 ± 150.8 | 0.9 ± 2.9 |
| HD1 | 0.41 ± 0.34 | 0.99 ± 0.00 | 213.0 ± 262.1 | 0.0 ± 0.0 |
| HD2 | 0.30 ± 0.32 | 0.93 ± 0.20 | 131.2 ± 150.6 | 10.9 ± 39.3 |
| HD3 | 0.09 ± 0.15 | 0.96 ± 0.14 | 330.5 ± 266.1 | 17.4 ± 89.0 |

Fig. 3 shows examples of the temporal change of the distance measure. The results of the conventional method not only have large errors, but also show large fluctuations over time. This means the estimated source location by the conventional method moves all around in 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 will degrade the perceived quality of the resultant sequence.

### 5.3. Coding efficiency

The coding efficiency of the proposed method is investigated in comparison to the constant QP mode. Fig. 4 shows the relative reduction of bitrate of the SD1 sequences produced by the proposed method for various $QP_0$ values when compared with those produced by JM with fixed QPs. The results for the other contents are omitted here because they are very similar to those in this figure. When $QP_0$ is small, the advantage of the proposed method in terms of the coding gain is clearly visible. On the other hand, when a large $QP_0$ value is used, the gain becomes small and even negative (i.e. a larger bitrate by the proposed method compared to the constant QP mode) because the amount of bits for sending the slice grouping information is not negligible any more when compared to that for the encoded image frames. In addition, one can observe the effects of the number of slices ($L$) and the incremental QP between slices ($\Delta QP$). By using large values of $L$ or $\Delta QP$, large QP values are assigned to the background regions that are far from the sound source, which in turn yields large coding gains.

### 5.4. Perceived quality evaluation

While the results presented above show the effectiveness of the proposed method in terms of coding efficiency, it is also necessary to verify that the gain is obtained without degradation of perceived quality in order to judge the usefulness of the method. Therefore, in this section we present the results of our subjective quality evaluation tests for the produced sequences.

The following four coding conditions were considered in the tests:

- JM with constant QP = 26.
- Proposed method with $QP_0 = 26$, $L = 4$, $\Delta QP = 1$.
- Proposed method with $QP_0 = 26$, $L = 4$, $\Delta QP = 2$.
- Proposed method with $QP_0 = 26$, $L = 4$, $\Delta QP = 4$.

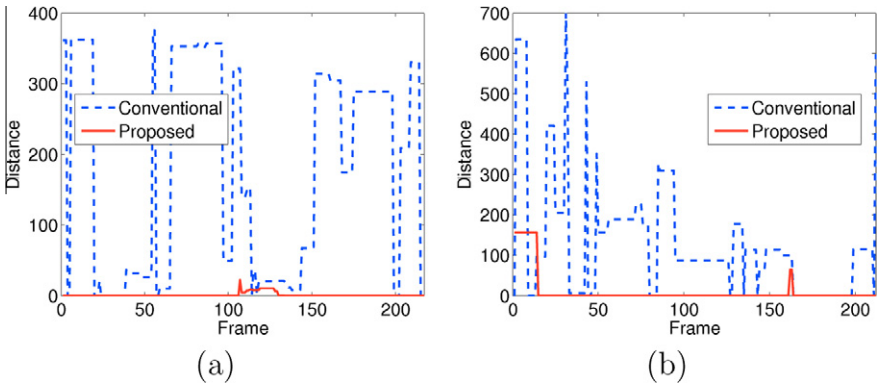


**Fig. 3.** Examples of the temporal variations of the distance measure by the conventional and proposed methods. (a) SD3; (b) HD2.

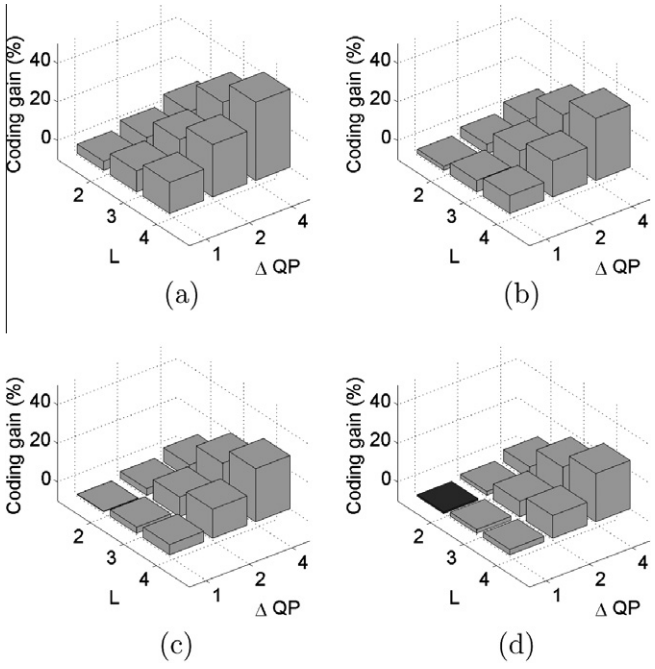


**Fig. 4.** Coding gain (%) by the proposed method in comparison to JM (fixed QP) for SD1 when (a) $QP_0 = 22$, (b) $QP_0 = 26$, (c) $QP_0 = 30$, and (d) $QP_0 = 34$. The dark bar indicates a negative value, i.e. higher bitrates by the proposed method compared to the fixed QP condition.

The test environment was intended to assure the reproducibility of the subjective test results by avoiding any involuntary influences of external factors. Thus, it is important to fix some features of the viewing environments, such as general viewing conditions and some crucial features of the used monitor. The information regarding our test environment is detailed in Table 3.

The tests have been performed according to the guidelines provided by the standards [30]. The single stimulus continuous quality scale (SSCQS) method was selected as the test methodology. Each audio-visual stimulus was shown to the subject and he/she was asked to enter a visual quality score for it. Before the test, instructions regarding the subject's task were provided. Then, a training session was held, where the test methodology was described to the subject by using a set of training stimuli that were different from the test stimuli. The subject watched each stimulus and had 5 s for voting on the score sheet. The rating scale used was a continuous scale ranging from 0 to 100, which accompanied adjective descriptions of each range of the scale ('excellent', 'good', 'fair', 'poor', and 'bad'). The presentation order of stimuli was randomized and it was ensured not to play the same content consecutively.

Fifteen subjects (9 males and 6 females) participated in the test. Their ages ranged between 20 and 35 with a mean of 27.9. They reported normal or corrected-to-normal vision. Screening of the subjects was performed by following guidelines described in [30], from which no outlier was detected.

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

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

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

Fig. 5 shows the MOS and CI values for the four coding conditions of each content. A general observation that can be easily made from the MOS values is that the quality of JM (the constant QP mode) is usually the best and a larger $\Delta$QP yields a lower quality level due to the quality degradation in the slices that do not contain the sound source. In order to examine the statistical significance of such quality degradation, two-tailed *t*-tests were performed under the null hypothesis that the two rating scores (one for JM and the other for the proposed method) are independent random samples from normal distributions with equal means, against the alternative that they do not have equal means. The cases where the quality degradation in the proposed method is found to be significant by the tests are indicated by the 'x'-marks on the corresponding bars in Fig. 5. It is observed that the difference of the perceived quality between the constant QP mode and the proposed method is statistically insignificant when the value of $\Delta$QP is small, while the limit of $\Delta$QP keeping the perceived quality degradation insignificant varies according to the content. For SD3, $\Delta QP = 1$ already causes significant quality deterioration in comparison to JM, which is mainly because the sound-emitting re-

**Table 3**
Details of the test room conditions.

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

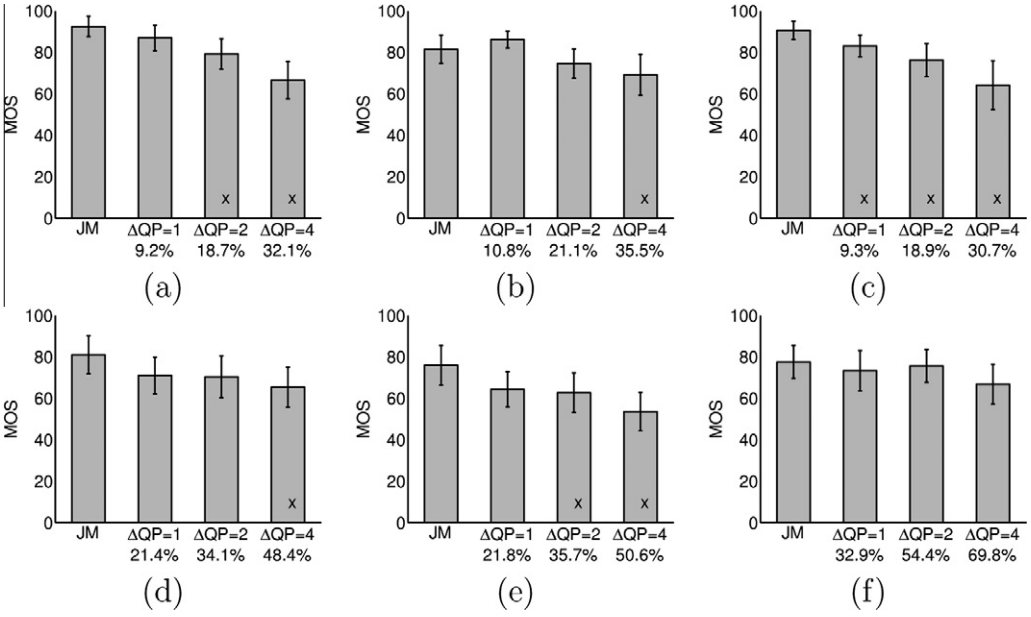


**Fig. 5.** Subjective test results comparing JM (QP = 26) and the proposed method ($QP_0$ = 26 and $L$ = 4) with three different values of ΔQP. The MOS values and the confidence intervals are shown for (a) SD1, (b) SD2, (c) SD3, (d) HD1, (e) HD2, and (f) HD3. The relative coding gains by the proposed method (%) are also mentioned below the *x*-axis. An '×' marks indicates the case where the quality degradation by the proposed method in comparison to JM is statistically significant at a significance level of 95%.

gion (the talking person) appeared small in a corner of the scene.[1] Interestingly, the HD sequences are much less vulnerable to quality degradation in comparison to the SD sequences. For HD3, even using ΔQP = 4 does not lead to statistically significant quality degradation in comparison to the constant QP mode, which is thought because the soundless motion by the listening person is not severe when compared to HD1 and HD2. This demonstrates more prominent effects of the focus of attention mechanism and the decreased resolutions of the peripheral vision in the HD sequences than in the SD ones. Except for SD3, the coding gains that can be obtained by the proposed method with $L$ = 4 without significant quality degradation are from 9.2% (for SD1) up to 69.8% (for HD3).

## 6. Discussion and conclusion

In this paper, we have proposed an audio-visual focus of attention-based video coding technique in the framework of H.264/AVC, which uses an audio-visual source localization method to identify the sound source in a scene by using the correlation between the one channel audio signal and pixel value information. Through the experimental results, we demonstrated that the audio-visual focus of attention mechanism can be exploited for improving efficiency of video coding without significant subjective quality degradation. In addition, it was shown that the effectiveness of the proposed method is more prominent in HD sequences.

In our method, two encoding parameters were introduced: the number of slices ($L$) and the incremental QP value between slices (ΔQP). In the previous section, we showed the results for different combinations of their values, from which it was observed that there exists a trade-off relationship between the coding gain and the quality. Particularly, the higher either of these parameters is, the larger the coding gain is, but accordingly the worse the quality of the background region is. Difficulty in determining these values lies in the fact that their optimal values are highly dependent on the content. For example, we found that the subjects often feel more annoyed when the coding artifact is located on the silent face region than when it is on the other background regions. In this case, the values of $L$ and ΔQP should be kept small compared to other cases having no face in the background.

Such additional coding parameters also exist in other existing methods for attention-based video coding. In [1], the author tried to use different values of the 'depth' parameter that determines the level of the spatial resolution reduction in different regions and reported its effect on the quality. The depth parameter can be considered to be similar to $L$ in our case. In the work presented in [5], an image frame is divided into three regions which are encoded with different QPs and ΔQP for these regions is simply set to 1, 2 and 3, respectively, as a conservative policy. In future work, it would be desirable to reach useful guidelines for determining the parameters by extensive experiments with diverse content having different types of sound sources, silent motions and temporal/spatial complexity.

One thing to be noted is that the audio-visual focus of attention is one of the attention mechanisms of the human visual system. Most of the other attention models mentioned in Section 2.2 can be complementarily used to be augmented to the proposed method. On the other hand, the assumption on the attention in the previous work [6], i.e. moving objects in a scene draws visual attention, is closely related to ours about the audio-visual focus of attention. In fact, the sound-emitting object in our analysis is a subset of the moving objects in a scene because it is the moving object that is responsible for the acoustic signal. In [31], the importance of the sound source and the other moving objects in the scene was compared in terms of perceived quality by considering the following two cases. In one case, all the moving objects were identified and, according to the distance of each pixel from the nearest moving object, blurring was applied through a Gaussian pyramid decomposition so that a region far from a moving object is strongly blurred. The other case is the same to this except that only the sound-emitting object is considered instead of the moving objects. The subjective quality evaluation results showed that, even if the moving objects that do not emit sound are blurred in the latter case, the perceived quality is not significantly degraded

[1] This does not mean that it is impossible to obtain coding gain for this content by using the proposed method because the value of $L$ was fixed to 4 in our subjective experiments and a smaller $L$ may produce sequences without significant quality degradation.

in comparison to the former case. This implies that all moving objects do not receive visual attention equally, since sound-emitting objects tend to attract more attention than others, which supports the assumption of our work in this paper. Nevertheless, our method could be improved by considering both the sound-emitting region and the other moving objects. In our work, we did not consider the moving objects as fixation points of visual attention, but they may also receive a certain amount of attention that is presumably less than that for the sound source. Therefore, it would be possible to develop a hybrid method in which the sound source and the other moving objects are all considered but differently weighted for prioritization in coding. In our future work, we plan to investigate further the relative importance of various sources drawing human attention in a scene by analyzing gaze patterns collected from human observers during multimedia consumption.

## Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2011) under Grant Agreement No. 21644 (PetaMedia) and the Swiss National Foundation for Scientific Research in the framework of the NCCR Interactive Multimodal Information Management (IM2).

## References

[1] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. Image Process. 13 (2004) 1304–1318.
[2] G. Boccignone, A. Marcelli, P. Napoletano, G.D. Fiore, G. Iacovoni, S. Morsa, Bayesian integration of face and low-level cues for foveated video coding, IEEE Trans. Circuits Syst. Video Technol. 18 (2008) 1727–1740.
[3] Z. Wang, L. Lu, A.C. Bovic, Foveation scalable video coding with automatic fixation selection, IEEE Trans. Image Process. 12 (2003) 243–254.
[4] Q. Chen, G. Zhai, X. Yang, W. Zhang, Application of scalable visual sensitivity profile in image and video coding, in: Proc. IEEE Int. Symposium on Circuits and Systems, Seattle, USA, 2008, pp. 268–271.
[5] C.-W. Tang, Spatiotemporal visual considerations for video coding, IEEE Trans. Multimedia 9 (2007) 231–238.
[6] A. Cavallaro, O. Steiger, T. Ebrahimi, Semantic video analysis for adaptive content delivery and automatic description, IEEE Trans. Circuits Syst. Video Technol. 15 (2005) 1200–1209.
[7] J. Driver, C. Spence, Attention and the crossmodal construction of space, Trends Cogn. Sci. 2 (1998) 254–262.
[8] J.J. McDonald, W.A. Teder-Salejarvi, F.D. Russo, S.A. Hillyard, Neural substrates of perceptual enhancement by cross-modal spatial attention, J. Cogn. Neurosci. 15 (2003) 10–19.
[9] C. Spence, J. Driver, Audiovisual links in exogenous covert spatial orienting, Percept. Psychophys. 59 (1997) 1–22.
[10] C. Spence, J. Driver, Audiovisual links in endogenous covert spatial attention, J. Exp. Psychol.: Human Percept. Perform. 22 (1996) 1005–1030.
[11] D.J. Tellinghuisen, E.J. Nowak, The inability to ignore auditory distractors as a function of visual task perceptual load, Percept. Psychophys. 65 (2003) 817–828.
[12] V. Mazza, M. Turatto, M. Rossi, C. Umilta, How automatic are audiovisual links in exogenous spatial attention?, Neuropsychologia 45 (2007) 514–522
[13] J. Vroomen, B. de Gelder, Perceptual effects of cross-modal stimulation: ventriloquism and the freezing phenomenon, in: Handbook of Multisensory Processes, MIT Press, 2004, pp. 141–150.
[14] J. Vroomen, B. de Gelder, Sound enhances visual perception: cross-modal effects of auditory organisation on vision, J. Exp. Psychol.: Human Percept. Perform. 26 (2000) 1583–1590.
[15] L.A. Ross, D. Saint-Amour, V.M. Leavitt, D.C. Javitt, J.J. Foxe, Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments, Cereb. Cortex 17 (2007) 1147–1153.
[16] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (1976) 746–748.
[17] R. Sharma, V.I. Pavlović, T.S. Huang, Toward multimodal human–computer interface, Proc. IEEE 86 (1998) 853–869.
[18] E. Macaluso, J. Driver, Multisensory spatial interactions: a window onto functional integration in the human brain, Trends Neurosci. 28 (2005) 264–271.
[19] J.-S. Lee, T. Ebrahimi, Efficient video coding in H.264/AVC by using audio-visual information, in: Proc. Int. Conf. Multimedia Signal Processing, Rio de Janeiro, Brazil, 2009, pp. 1–6.
[20] H. Zhou, M. Taj, A. Cavallaro, Target detection and tracking with heterogeneous sensors, IEEE J. Sel. Top. Sign. Proc. 2 (2008) 503–513.
[21] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, I. McCowan, Audiovisual probabilistic tracking of multiple speakers in meetings, IEEE Trans. Audio Speech Lang. Proc. 15 (2007) 601–616.
[22] P. Perez, J. Vermaak, A. Blake, Data fusion for visual tracking with particles, Proc. IEEE 92 (2004) 495–513.
[23] V. Pavlović, A. Garg, J.M. Rehg, T.S. Huang, Multimodal speaker detection using error feedback dynamic Bayesian networks, in: Proc. Int. Conf. Computer Vision and Pattern Recognition, Hilton Head Island, SC, USA, 2000, pp. 34–41.
[24] R. Cutler, L. Davis, Look who's talking: speaker detection using video and audio correlation, in: Proc. Int. Conf. Multimedia and Expo, New York, NY, USA, 2000, pp. 1589–1592.
[25] P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran, M. Kunt, Extraction of audio features specific to speech production for multimodal speaker detection, IEEE Trans. Multimedia 10 (2008) 63–73.
[26] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (2004) 2639–2664.
[27] J.-S. Lee, F. De Simone, T. Ebrahimi, Video coding based on audio-visual attention, in: Proc. Int. Conf. Multimedia and Expo, New York, NY, USA, 2009, pp. 57–60.
[28] E. Kidron, Y.Y. Schechner, M. Eland, Cross-modal localization via sparsity, IEEE Trans. Signal Process. 55 (2007) 1390–1404.
[29] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, H.264/AVC JM Reference Software, 2008, <http://iphome.hhi.de/suehring/tml/>.
[30] Recommendation ITU-R BT.500-11, Methodology for the Subjective Assessment of the Quality of Television Pictures, 2002.
[31] J.-S. Lee, F. De Simone, T. Ebrahimi, Influence of audio-visual attention on perceived quality of standard definition multimedia content, in: Proc. Int. Workshop on Quality of Multimedia Experience, San Diego, CA, USA, 2009, pp. 13–19.