VIDEO CODING BASED ON AUDIO-VISUAL ATTENTION[1]

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

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

ABSTRACT

This paper proposes an efficient video coding method based on audio-visual attention, which is motivated by the fact that cross-modal interaction significantly affects humans' perception of multimedia content. First, we propose an audio-visual source localization method to locate the sound source in a video sequence. Then, its result is used for applying spatial blurring to video frames in order to reduce redundant high-frequency information and achieve coding efficiency. We demonstrate the effectiveness of the proposed method for H.264/AVC coding along with the results of a subjective evaluation.

Index Terms— video coding, audio-visual attention, cross-modal interaction, source localization, H.264/AVC, perceived audio-visual quality

1. INTRODUCTION

Techniques based on the characteristics of the human visual system for efficient video coding have recently received much interest. When humans 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. This implies that it may not be necessary to encode the whole scene with a uniform quality. Compression efficiency can be achieved by discarding redundant information outside small fixation regions without significant degradation of perceived quality. Several techniques have been proposed for such visual attention-based coding, in which spatial prioritization schemes determine the priorities of different regions in the scene and encoding is performed according to those priorities [1,2].

An important aspect which has been rarely considered in attention-based coding is the acoustic modality. Cross-modal interaction of auditory and visual modalities plays an important role in spatial attention. An auditory stimulus in a particular location attracts visual attention occurring at the same spatial location, for both exogenous attention (i.e., stimulus driven) and endogenous attention (i.e., directed voluntarily) [3]. Even when people are performing a visual task, a novel auditory stimulus can automatically capture their visual attention [4]. Directing attention to an auditory stimulus improves perception of the subsequent visual stimulus [5]. As for motion perception, simultaneous auditory motion information introduces a bias or affects the sensitivity in a visual motion detection task [6].

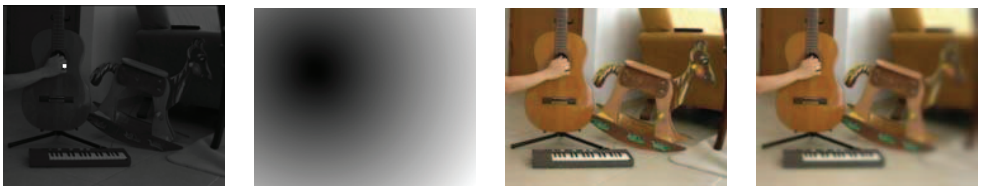
Motivated by these observations, we propose a new attention-based video coding technique which considers both the acoustic and the visual modalities. For a given video sequence containing an audio channel, the region emitting sound is localized by analyzing the correlation between the acoustic and the visual signals. Then, a priority map is generated based on the distance of each pixel to the localized region; the priority is the highest for the localized region and decreases as the distance increases. Spatial blurring is applied so that a low priority region is strongly blurred. The introduction of blur at the locations far from the sound source attenuates high frequency information in those areas which are likely to be less attended. Thus, coding efficiency can be increased without significant perceived quality degradation. We demonstrate the effectiveness of the proposed localization method and the consequent coding method through experimentations. Moreover, the results of a subjective quality test with the produced video are reported.

The following section presents the proposed audio-visual source localization method. In Section 3, we describe the encoding scheme based on the audio-visual localization. The experimental results are shown in Section 4 and, finally, concluding remarks are given in Section 5.

2. AUDIO-VISUAL SOURCE LOCALIZATION

Audio-visual source localization is to specify the location of the region producing sound in a video sequence. Especially,

[1] The research leading to these results has received partial funding from the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia).



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

Fig. 1. Illustration of the proposed method for a frame of Data #1 (see Section 4 for description of the data). (a) Localization result overlaid on the image (marked with a white dot). (b) Priority map (shown with small brightness values for high priorities). (c) Blurred image for L=2. (d) Blurred image for L=6.

we are interested in the case where multiple moving objects appear simultaneously but only one object is responsible for the sound, and thus a conventional motion detection with only the visual information may not be satisfactory.

Our source localization method is based on the one proposed in [7], to which an important improvement has been introduced. The method does not have any assumption on the object or region to be localized but tries to pinpoint image pixels associated with the one-channel acoustic source. Moreover, the method does not require a training step which may need manually processed training data.

Basically, the method utilizes the canonical correlation analysis (CCA) technique. Let $\mathbf{a}$ and $\mathbf{v}$ be the temporally synchronuous n_a-dimensional acoustic feature vector and the n_v-dimensional visual vector for a frame. In this paper, pixel values are used for $\mathbf{v}$. The objective of CCA is to find a pair of vectors $\mathbf{w}_a$ and $\mathbf{w}_v$ which maximize the correlation of the projected features, i.e., $\mathbf{w}_a^T\mathbf{a}$ and $\mathbf{w}_v^T\mathbf{v}$, respectively. Let $\mathbf{A}$ and $\mathbf{V}$ be the collections of the features over multiple frames, i.e., $\mathbf{A}=[\mathbf{a}_t,\mathbf{a}_{t+1},\ldots,\mathbf{a}_{t+T-1}]$ and $\mathbf{V}=[\mathbf{v}_t,\mathbf{v}_{t+1},\ldots,\mathbf{v}_{t+T-1}]$, where T is the number of frames. Then, projection vectors are obtained by solving

$$\mathbf{w}_a,\mathbf{w}_v = \arg\max_{\mathbf{w}_a,\mathbf{w}_v} \frac{\mathbf{w}_a^T(\mathbf{V}^T\mathbf{A})\mathbf{w}_v^T}{\sqrt{\mathbf{w}_a^T(\mathbf{A}^T\mathbf{A})\mathbf{w}_a\mathbf{w}_v^T(\mathbf{V}^T\mathbf{V})\mathbf{w}_v}}. \quad (1)$$

It can be shown that solving the above problem is equivalent to resolving the following equation [7]:

$$\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a. \quad (2)$$

If $\mathbf{V}$ is full rank, there exist an infinite number of solutions for (2), because the dimension of $\mathbf{w}_v$ is usually much larger than T.

In order to alleviate this problem, a spatial sparsity criterion is imposed to find a unique solution:

$$\min \|\mathbf{w}_v\|_1 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \quad (3)$$

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

$$\min \|\mathbf{w}_v\|_1 \quad \text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a, \mathbf{h}_k^T\mathbf{w}_a = 1, \text{ and } \mathbf{H}_k\mathbf{w}_a \geq 0, \quad (4)$$

where the elements of $\mathbf{h}_k$ are the binary representation of k with +1 and -1, and $\mathbf{H}_k$ is the 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 (3) and (4), $\mathbf{w}_v$, can be interpreted as the *cross-modal energy* concentrated on the visual features which are responsible for the acoustic signal.

A limitation of the above algorithm is the lack of consideration of temporal and spatial consistency. Therefore, we improve it by incorporating such consistency in the method, which will result in improved tracking performance. For this, the problems in (3) and (4) are modified as:

$$\min \sum_{i=1}^{n_v} |f_i w_{vi}| \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \quad (5)$$

and

$$\min \sum_{i=1}^{n_v} |f_i w_{vi}| \quad \text{subject to } \mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a, \mathbf{h}_k^T\mathbf{w}_a = 1, \text{ and } \mathbf{H}_k\mathbf{w}_a \geq 0, \quad (6)$$

respectively, where w_{vi} is the i-th component of $\mathbf{w}_v$. Here, the weighting factor $\{f_i\}_{i=1}^{n_v}$ is to consider the localization result of the previous frame for obtaining the result of the current frame. If a high value has been obtained for a pixel at a certain frame, we assign low weight values to the pixel and its spatial neighbors so that high energy values are assigned for those pixels in the next frame. This weighting scheme promotes consistency in the localization results in consecutive frames. Again, the problem (5) or (6) is solved by using linear programming.

The weights are obtained as follows: Once we have the localization result for the previous frame, we apply a smoothing filter to its image representation. For this, we use a Gaussian filter. Then, the weights are obtained by

$$f_i = \max_j w_{vj}^{old} - w_{vi}^{old} + 1, \quad (7)$$

where w_{vi}^{old} is the i-th component of the smoothed image representation of the localization result for the previous frame. Adding 1 in (7) is to ensure that all weights are nonzero. Using a smoothing filter is to assign low weights not only to the localized pixel but also to its neighboring region, which allows some margin in consistency.

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

3. ATTENTION-BASED CODING

To exploit the localization result for video coding, we apply a spatially variable blur to each image frame of the video sequence according to the result for the frame.

Once we have obtained the source localization result for each frame, a priority map is produced, which represents the weighted distance between each pixel and the nearest localized energy location (Fig. 1(b)). When there are more than one energy locations, the weighting is calculated in such a way that a pixel near a smaller energy receives a larger distance than one near a larger energy location, just as in a contour map. It is possible to monotonically scale the priority, e.g. by applying logarithm or exponentiation. However, such scaling was not necessary in our experiments.

Higher compression ratios are obtained for the smoothed regions due to elimination of high frequency components via smoothing. The compressed streams produced with this approach are fully compatible with existing decoders.

Blurring is performed with a Gaussian pyramid with L levels. An image with stronger blurring at low priority locations is obtained with a larger value of L. Each level of the pyramid is assigned to the linearly spaced values within the range of the priority values; the highest level (the original image) is assigned to the highest priority (at the sound-emitting location with the largest energy) and the lowest level to the lowest priority. For the priority values between two levels, trilinear interpolation is applied. Fig. 1(c) and (d) depict examples of the blurred images for two different values of L.

The blurred frames are encoded by a conventional encoder (e.g. MPEG-2, MPEG-4 or H.264/AVC) to produce the final video stream.

4. EXPERIMENTS

4.1. Setup

We used four video data for our experiments, namely, Data #1 and #2) from [7], and Data #3 and #4 selected from the "groups" section of the CUAVE database [8]. Their lengths are about 10 seconds each. In Data #1 a hand plays a guitar and then a synthesizer, while a wooden horse is rocking. In Data #2 a talking head and a rocking wooden horse appears at the same time. Data #3 and #4 contain two and three people pronouncing continuous English digits in turn, respectively. While a person speaks, the other persons act as distractors by moving their heads and mouths.

For source localization, we use the difference of the luminance component of consecutive frames for the visual features. For the acoustic features, the energy of audio samples within a moving window is extracted for each frame. We set T=32 for Data #1 and #2 as in [7], and T=16 for Data #3 and #4 considering faster speech rates in them compared to the formers.

We use the x264 implementation of H.264/AVC [9] for creating compressed video sequences. The constant quantization parameter (QP) and the constant bitrate encoding modes are used. The audio part is encoded by MP3.

4.2. Localization performance

First, we show the performance of the proposed localization method by comparing with that of the method in [7]. Fig. 2 shows the localization performance over frames for Data #1 and #2. The performance is defined as the ratio of the energy concentrated in the sound-emitting region to that in all moving parts which have been identified manually. It ranges from 0 (failure of localization) to 1 (localization without error). It is observed that the localization results are significantly improved by considering consistency in our method.

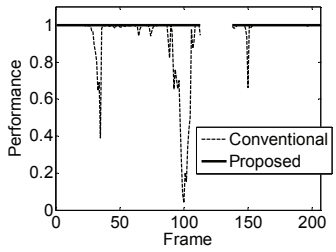
4.3. Coding efficiency

To evaluate the efficiency of the proposed coding approach, we compare the file size of the original and the processed (i.e., spatially blurred) video sequences after compression when the constant QP mode is used. Table 1 shows the file size of the processed sequences when compared to the originals for different values of L and QP. The gain is larger for larger values of L, but blurring artifacts would be more clearly observed. Smaller values of QP (i.e., a better quality) result in larger gains because the encoder tries to keep high frequency components. The result in the first column is worth paying attention to, because setting QP=26 produces high quality streams and it is quite difficult to notice blurring artifacts for L=2. Therefore, we can estimate from this example that the proposed method typically yields compression gains of about 24-35% while producing the test video sequences of good quality.

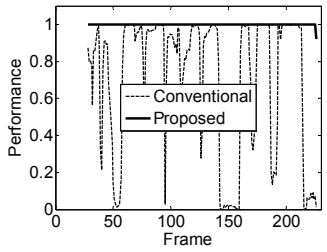
4.4. Subjective quality assessment

The audio-visual subjective quality test aims at providing justification of the proposed coding method: We want to show that discarding high frequency information outside the sound-emitting region does not degrade perceived quality significantly. This would be an indirect support of the validity of the proposed method.

Five processing conditions are compared to each other: no blurring (NB), the case where all the moving objects are identified and receive high priority with L=2 (M2) and L=6 (M6), the proposed method with L=2 (P2) and L=6 (P6). The processed frames are compressed at 100 kbps and 500 kbps by using the constant bitrate mode. The double stimulus continuous quality scale (DSCQS) method was selected as the test methodology [10]. Fifteen naïve assessors participated in evaluations. Each subject had two



(a)



(b)

Fig. 2. Comparison of localization performance for (a) Data #1 and (b) Data #2. The discontinuous regions are silent periods.

separate sessions for each bitrate value. The guidelines provided by standards [10] were used to determine the test environments, the duration of each session, the training session and the method of processing the subjective data.

Tables 2 and 3 show the differential mean opinion score (DMOS) values and the confidence interval obtained for Data #1 and #4 in the test sessions for 100 kbps and 500 kbps, respectively. It is observed that, assuming that the overlap of confidence intervals indicates absence of statistical differences between DMOS values, the difference of perceived quality among the three methods for L=2 and that between the proposed method and the method prioritizing all the moving regions for L=6 are usually small and statistically irrelevant. Also, the results vary depending on the original content of the sequence; the perceived quality is sensitive to face regions, as it could have been expected.

5. CONCLUSION

We have presented a new pre-processing method in video coding, in which audio-visual information is utilized to determine importance of each part in image frames for efficient video coding without introducing a perceived quality degradation. It was demonstrated that considering spatio-temporal consistency improves source localization performance, and spatial blurring based on the priority map obtained from the localization leads to better coding efficiency. The subjective test results also reveal the validity of the proposed method.

For future work, experiments with higher definition content will be conducted. Also, we will consider using visual saliency information or user-provided tags in the present method to improve perceived quality.

Table 1. Compressed file size of processed videos compared to that of originals (%).

Data	L=2		L=6	
	QP=26	QP=46	QP=26	QP=46
#1	76.0	96.1	48.9	82.4
#2	69.7	96.6	43.6	84.7
#3	64.6	93.9	42.8	84.3
#4	64.6	93.5	37.3	80.5

Table 2. DMOS and confidence interval (CI) values for the compression rate of 100 kbps.

Method	Data #1		Data #4	
	DMOS	CI	DMOS	CI
NB	64	6.4	64	2.9
P2	69	3.9	67	4.1
M2	67	3.8	67	3.6
P6	70	4.6	77	4.1
M6	77	3.8	65	3.6

Table 3. DMOS and confidence interval (CI) values for the compression rate of 500 kbps.

Method	Data #1		Data #4	
	DMOS	CI	DMOS	CI
NB	10	5.1	16	4.8
P2	11	5.2	22	6.1
M2	11	5.0	15	4.8
P6	42	5.7	49	7.5
M6	41	7.0	41	4.0

6. REFERENCES

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

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

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

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

[5] J. J. McDonald, W. A. Teder-Sälejärvi, F. D. Russo, and S. A. Hillyard, "Neural substrates of perceptual enhancement by cross-modal spatial attention," *Journal of Cognitive Neuroscience*, vol. 15, no. 1, pp. 10-19, 2003.

[6] G. F. Meyer and S. M. Wuerger, "Cross-modal integration of auditory and visual motion signals," *NeuroReport*, vol. 12, no. 11, pp. 2557-2560, Aug. 2001.

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

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

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

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