

Lossless Audio Hiding Method for Synchronous Audio-Video Coding

^aWeiwei Chen, ^aJin Li ^aMoncef Gabbouj ^bJarmo Takala

^aDepartment of Signal Processing, ^bDepartment of Computer Systems
Tampere University of Technology, Tampere, Finland
FirstName.LastName@tut.fi

ABSTRACT

The absence of audio-video synchronization has been one of the most annoying effects in multimedia application. This paper proposes a lossless information hiding method for synchronous audio-video coding. The binary bits of an encoded audio signal are embedded into the DCT domain of the video data to produce synchronous hybrid signal which is further encoded. The decoder can extract the audio bits from the hybrid bit stream, and thus, is able to reconstruct the separate signals. This scheme enables synchronous coding, transmission, storage and playback of a video signal and the associated audio signal. Experimental results show that the proposed method outperforms competing techniques in terms of both the reconstructed video and audio quality.

1. INTRODUCTION

Currently, audio and video signals are compressed with separate encoders in standard systems. Time stamps and clock references can be used to obtain correct playback time for the video signal and the associated audio signal at the decoder side. However, the synchronization can be disturbed, e.g., by jitter in network transmission or variation in storage access times. Phase-lock-loop [1] can be used to obtain some tolerance against an occurring jitter in the decoder with a limited buffer size, but the absence of audiovisual synchronization is still unavoidable in practice. Recently, a new proposal for solving the synchronous decoding was developed by synchronizing audio with lip movement [2]. But the application is only for the video containing the movements of human mouths and the implementation is complicated.

Several techniques have been developed to embed the audio bits into its associated video for synchronous audio-video coding. These algorithms can be classified into spatial domain algorithms [3]-[5] and transform domain algorithms [6]. In [3]-[5], the host video sequence is first transformed into frequency domain through 4×4 DCT. Then, the DCT coefficients in each block are modified to hide one audio bit, i.e., 1 or 0. Finally, the new DCT blocks are transformed back into its spatial domain to obtain the stego-video sequence which is further to be processed by video encoder. However, it is a lossy technique for both video signal and audio signal due to the following compression operation. On the other hand, an information hiding method is designed in [6] to embed the audio bits into the quantized DCT coefficients in the process of video compression. Compared to approaches in spatial domain, this algorithm is a lossless hiding operation for audio. However, since each DCT block only accommodates one bit, the capacity is limited.

Although different hiding algorithms are developed in [3]-[6], the synchronization of audio-video cannot be reached only if the audio bits be properly embedded into the corresponding video frames. That means that the length of audio bit stream for different frames is variable. Therefore, it is important to ensure that the decoder can identify the embedded blocks and correctly extract the audio bits. However, solutions have not been proposed to deal with such situations.

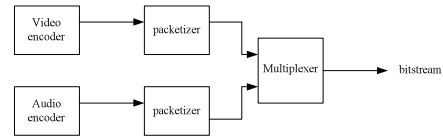


Fig.1 Audio-Video encoding systems in MPEG-2

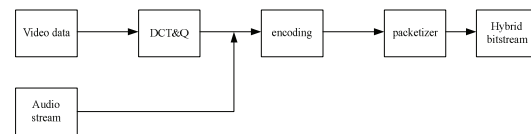


Fig.2. Proposed audio-video encoding system

In this paper, we propose a lossless audio hiding method for synchronous audio-video coding. The audio bits are embedded into the DCT coefficients of the host video frames after quantization. The magnitudes of the non-zero quantized DCT coefficients are modified into odd or even values to correspond the audio bits, 0 or 1. Additional bits are encoded for each frame to indicate the number of the audio bits and their positions so that the decoder can properly extract the audio from the hybrid audio-video signal. Compared to competing techniques, the proposed method can exactly retrieve the audio signal and achieves better video quality. Moreover, the host video frames have a more flexible accommodation capacity for audio bits.

The rest of this paper is organized as follows. The lossless audio information hiding algorithm is proposed in Section 2. The experimental results are presented in section 3. Finally, we conclude the paper in section 4.

2. PROPOSED AUDIO HIDING METHODS

2.1 Framework of Proposed Method

The key idea of the proposed method is to embed the bit stream of an audio signal into the quantized DCT coefficients of the video and encode them jointly. At the receiver, the decoder is able to extract the audio bits from the hybrid signal and thus to fractionate the hybrid signal back into two separate signals, i.e., the audio and the video.

Fig.1 shows the conventional audio/video encoding system in MPEG-2 and the proposed hybrid encoding system for audiovisual synchronization is illustrated in Fig.2. The corresponding decoding systems just operate in the reverse order to get the reconstructed synchronous audiovisual signal.

Compared to the conventional separate coding systems, the proposed synchronous audio-video coding scheme mainly have several advantages. First, both the audio signal and the video sequence are guaranteed to be played back properly. Second, the synchronization between the audio and the video can be reliably ensured and will not be affected by packet loss during transmission. Third, since the audio bits are embedded into the video data, no additional communication channel is

TABLE I MAIN STEPS FOR HIDING ALGORITHM ($u, v = 0, 1, \dots, N - 1$)

Condition	Operation	
audio bit =0, $F_i^Q(u, v)$ is even	$F_i^Q(u, v) = F_i^Q(u, v)$	
audio bit =0, $F_i^Q(u, v)$ is odd	if $F_i^Q(u, v) = 1$	$F_i^Q(u, v) = F_i^Q(u, v) + 1$
	if $F_i^Q(u, v) = -1$	$F_i^Q(u, v) = F_i^Q(u, v) - 1$
	if $ F_i^Q(u, v) > 1, F_i^Q(u, v) \geq F_i(u, v)$	$F_i^Q(u, v) = F_i^Q(u, v) - 1$
	if $ F_i^Q(u, v) > 1, F_i^Q(u, v) < F_i(u, v)$	$F_i^Q(u, v) = F_i^Q(u, v) + 1$
audio bit =1, $F_i^Q(u, v)$ is odd	$F_i^Q(u, v) = F_i^Q(u, v)$	
audio bit =1, $F_i^Q(u, v)$ is even	if $F_i^Q(u, v) \geq F_i(u, v)$	$F_i^Q(u, v) = F_i^Q(u, v) - 1$
	if $F_i^Q(u, v) < F_i(u, v)$	$F_i^Q(u, v) = F_i^Q(u, v) + 1$

needed for audio signal transmission. Finally, the complex tasks of multiplexing, de-multiplexing, and the head bits concerning of time stamps and clock references in MPEG systems have been avoided.

2.2 Analysis of PSNR in DCT Domain

Since the audio bits are embedded into the DCT coefficients after quantization and the following encoding block is a lossless process, the audio bits can be properly retrieved at the decoder. On the other hand, hiding of audio information can result in additional video quality degradation. Therefore, it is important to ensure a minimal PSNR degradation when embedding the audio bits into the video.

According to Parseval’s theorem, the mean square error (MSE) in the pixel domain is equivalent to the mean square error (MSQE) in the DCT domain, because DCT is a normalized orthogonal transformation. Therefore, it is possible to measure the quantization error in the DCT domain, using the following equation in [7]

$$PSNR = 20 \log_{10} \left(\frac{255}{\sqrt{MSQE}} \right) \tag{1}$$

where

$$MSQE = \frac{1}{NN} \sum_{v=0}^{N-1} \sum_{u=0}^{N-1} (F_i^Q(u, v) - F_i(u, v))^2$$

and $F_i^Q(u, v), F_i(u, v)$ are the DCT coefficients of the source signal and the corresponding decoded signal in the i th $N \times N$ DCT block, $0 \leq u, v \leq N - 1$, and M defines the number of the DCT blocks. The reconstructed DCT coefficient $F_i^Q(u, v)$ is defined as

$$F_{i(u,v)}^Q = F_i^Q(u, v) \alpha Q_p \tag{2}$$

where

$$F_i^Q(u, v) = \left[\frac{F_i(u, v)}{\alpha Q_p} \right]$$

and $[x]$ denotes the nearest integer to x , Q_p is the quantization parameter and α represents the mapping relationship between the quantization parameter and the quantization step. Theoretically, the error $E_i(u, v)$ between the decoded DCT coefficient $F_i^Q(u, v)$ and the original coefficient $F_i(u, v)$ can be estimated from (2) as

$$E_i(u, v) = |F_i^Q(u, v) - F_i(u, v)| \leq \frac{\alpha Q_p}{2} \tag{3}$$

where $|x|$ denotes magnitude of x .

Now, if the quantized DCT coefficient $F_i^Q(u, v)$ has to be modified to its nearest integer, i.e., $F_i^Q(u, v) + 1$ or $F_i^Q(u, v) - 1$, it is desirable to first investigate how the modification affect the quantization error. Since smaller errors

usually lead to better PSNR performance, high quantization errors should be always avoided.

If $F_i^Q(u, v) \geq F_i(u, v)$, which means $0 \leq F_i^Q(u, v) - F_i(u, v) \leq \frac{\alpha Q_p}{2}$, the error $E_i(u, v)$ between the reconstructed coefficient of $F_i^Q(u, v) + 1$ and $F_i(u, v)$ falls

$$\alpha Q_p \leq E_i(u, v) \leq \frac{3\alpha Q_p}{2} \tag{4}$$

While the distance between the reconstructed coefficient of $F_i^Q(u, v) - 1$ and $F_i(u, v)$ is

$$\frac{\alpha Q_p}{2} \leq E_i(u, v) \leq \alpha Q_p \tag{5}$$

In this case, the modified coefficient $F_i^Q(u, v) - 1$ tends to have smaller errors than $F_i^Q(u, v) + 1$, although they both generate larger errors compared to $F_i^Q(u, v)$ as shown in (2) and (3).

Similarly, if $F_i^Q(u, v) < F_i(u, v)$, modifying $F_i^Q(u, v)$ to the nearest smaller integer always have better PSNR performance than to the larger integer.

2.3 Audio Bit Hiding Algorithm

The proposed algorithm is to embed the audio bits into the DCT coefficient of the corresponding video frames after quantization. For a non zero-quantized DCT coefficient selected as host for audio, the magnitude is changed to an even or odd value according the audio bit, i.e., 0 or 1. At the decoder, the hidden audio bits can be easily extracted by identify those embedded video coefficients. The main steps can be described as follows

- 1) A non-zero quantized DCT coefficient $F_i^Q(u, v)$, i.e., $F_i^Q(u, v) \neq 0$, is picked up to hide one audio bit;
- 2) The reconstructed DCT coefficient $F_i^Q(u, v)$ is calculated from $F_i^Q(u, v)$ according to (2);
- 3) One audio bit is embedded by modifying the magnitude of the coefficient $F_i^Q(u, v)$ as shown in Table I.

At the receiver, the audio bits can be easily retrieved by identifying the magnitudes of the quantized DCT coefficients as long as the decoder knows what coefficients have been selected.

Therefore, additional bits are required to indicate the number of the embedded audio bits at the encoder side. In this paper, 11 bits are defined for each frame, among which the first 10 bits are used to include the information of the number of audio bits in this frame and the last bit defines the number of audio bits in each DCT block if there are enough non zero-quantized coefficients. For instance, “0” indicates only the

first non zero-quantized DCT coefficient is embedded by one audio bit if there is at least non zero-quantized DCT coefficient in the block; and “1” is denoted as first two non-zero quantized DCT coefficients are selected. However, if there is only one non-zero quantized DCT coefficient in the block, only one audio bit is hidden. But the decoder can still properly extract the audio bits since it has known the total number of the audio bits for each frame.

3. EXPERIMENTAL RESULTS

To evaluate the proposed model, a series of experiments were carried out, using C-code, based on XVID codec [8] against competing methods. Since XVID codec is designed for MPEG-4 Visual where the 8×8 DCT is employed, we take $N = 8$ throughout the experiments. Several 4CIF video sequences (Crew, City, Harbour and Soccer) are tested against with different quantization. The codec was compiled with Microsoft Visual Studio 2008.

3.1 Channel Capacity of Video

Hiding algorithms in spatial domain usually have a better channel capacity than their counterpart in DCT domain. According to [3]-[5], for a 4CIF sequence the maximum number of audio bits for each frame is

$$704 \times 576 \times 1.5 \div (4 \times 4) = 38,016 \text{ bits}$$

which significantly outperforms the capacity in DCT domain. However, the more the audio bits are embedded, the more errors happen to the audio and video signal at the decoder due to the following coding process.

Compared to the techniques in spatial domain, channel capacity has been a bottleneck for embedding algorithms in DCT domain of the video. Since most of the DCT coefficients become zeros after quantization, the number of non-all-zero quantized DCT blocks is very limited, particularly at low bitrate. Therefore, it is important to evaluate the channel capacity of the proposed algorithm.

The capacity of the proposed algorithm is compared with [6] at different quantization Q_p . The average results on the four 4CIF sequences are shown in Fig.3. Since hiding algorithm in [6] is implemented on the 4-D DCT based video coding. In this paper, it is applied to the 8×8 DCT based XVID codec. That is, for intra frame two non-zero quantized DCT coefficients are selected among the middle frequencies so as not to significantly change the values of low frequency coefficients. For inter frame the non zero-quantized DCT coefficients can be at any frequency position since they only concern high frequency information. It is shown in Fig.3 that the proposed method significantly outperforms the reference in terms of channel capacity. For [6], only those blocks containing two non zero-quantized DCT coefficients can be selected for embedding. However, according to the proposed algorithm the audio bits can be embedded into all the non-all-zero-quantized DCT blocks. This advantage is in particular obvious at low bitrate where a lot of DCT blocks only contain one non zero-valued coefficient.

3.2 Video Quality Comparison

The audio bits are embedded into the video signal by modifying the magnitudes of the DCT coefficients. At the receiver, the decoder can correctly extract the audio by identifying the embedded coefficients of the video. However, the modified DCT coefficients cannot be retrieved any more. Thus, video quality degradation is inevitable. A lot of experiments have been done to compare the video quality among the proposed algorithm, the references and the standard. Audio bitstreams of 16Kb/s are selected in the experiment.

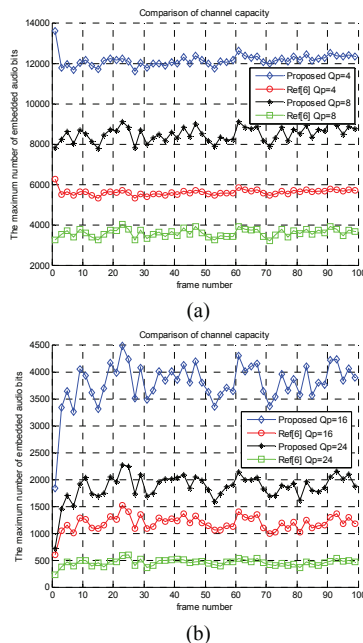


Fig. 3 Comparison of channel capacity between the propose algorithm and [6], (a) $Q_p = 4$ and $Q_p = 8$, (b) $Q_p = 16$ and $Q_p = 24$.

In [3]-[5], thresholds are required to embed the audio bits into the video. However, it is not clearly explained how the thresholds are selected. Practically, higher thresholds preserve better audio quality at the decoder, but lead to higher video quality degradation. Therefore, the thresholds should be compromised between the decoded video quality and audio quality. In our experiments, the thresholds are empirically selected as $2Q_p$ in connection to the quantization Q_p .

It is worth to point out that the PSNR is measured between the reconstructed video after extracting the audio and the pure video without hiding information. The results are illustrated in Fig.4. For [6], since there are not enough DCT blocks for an audio of 16Kb/s at $Q_p = 24$, only the R-D performances at $Q_p = 4, 8, 16$ are shown for City, Crew and Soccer.

According to the experimental results, several conclusions can be drawn. First, it is obvious that the proposed algorithm achieves better video quality in all cases than the references. In addition, the degradation is really negligible compared to the standard codec. Second, although embedding in spatial domain usually has advantage in channel capacity, the video quality can be seriously deteriorated because of the following lossy coding process. Third, the PSNR degradation caused by the proposed algorithm becomes larger with the increase of quantization. At the very high bitrate, the proposed algorithm has almost no impact on the PSNR compared to the standard codec. This is because at low quantization, the errors tend to fall in a very small interval as shown in (5) and has less effect on the PSNR measurement.

3.2 Audio Quality Comparison

As stated above, the proposed algorithm and [6] are both lossless hiding methods for audio signal. At the decoder, all the audio bits can be correctly retrieved. Therefore, the audio quality remains the same quality as the original, which has been verified in the experiments. In addition, the video quality in [3] is evaluated in the measurement of error rate (ER), which is defined as

$$ER = \frac{E_A}{T_A} \times 100\% \quad (6)$$

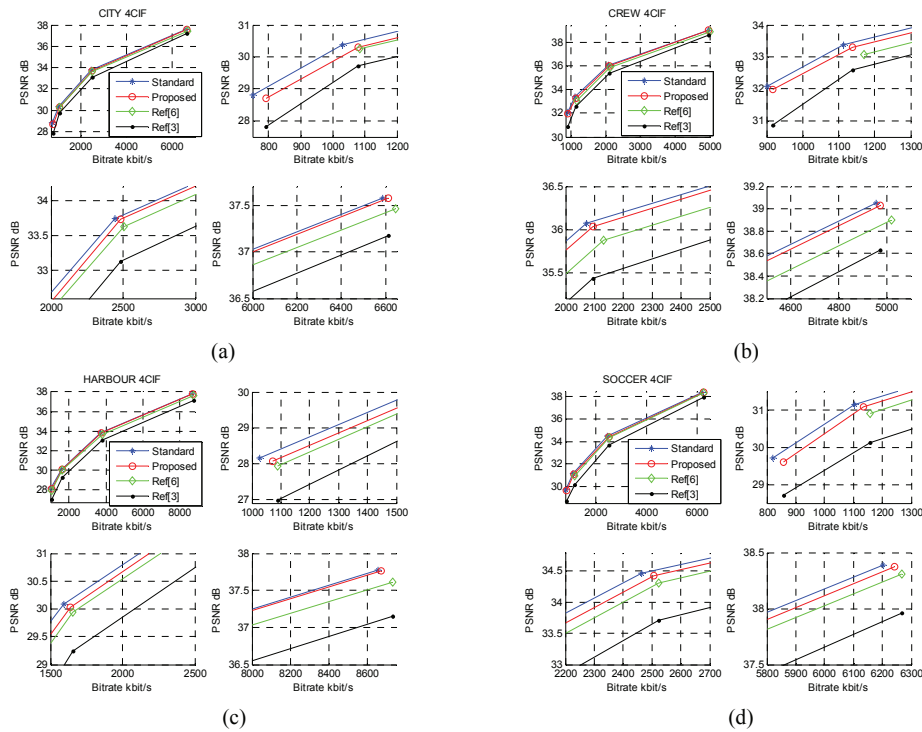


Fig. 4 Rate-Distortion performance regarding luminance component, (a) City, (b) Crew, (c) Harbour and (d) Soccer. Each sequence consists of an overall comparison and three local comparisons. Results for each sequence consist of an overall R-D performance and three locally magnified comparisons.

where E_A denotes the number of errors when extracting the audio bits and T_A is the total number of embedded audio bits. Fig.5 shows the results for the four sequences at different quantization. Experiments show that the error rate ranges between 1.56 and 7.96 for different quantization. A high threshold for vector quantization in [3] tends to have a small error rate for audio bits. However, the video quality would be seriously degraded. Therefore, the thresholds should be empirically determined and is just a compromise between the decoded video quality and the audio quality.

4. CONCLUSIONS

A lossless hiding method is proposed for synchronous audio-video coding. The binary bits of an encoded audio signal are embedded into the DCT domain of the video to produce synchronous hybrid signal. The decoder can retrieve the audio bits from the hybrid stream, and thus, is able to reconstruct the separate signals. This scheme enables synchronous coding, transmission, storage and playback of a video signal and the associated audio signal. Experimental results show that the proposed method outperforms competing techniques in terms of both the reconstructed video quality and audio quality.

5. ACKNOWLEDGMENT

This work was supported by the Academy of Finland of project No. 213462, the Academy of Finland Grant 117065 and by the National Natural Science Foundation of China Grant 609111301281.

6. REFERENCES

- [1] W. F. Egan, "Frequency synthesis by phase lock (2nd ed.)," John Wiley and Sons, 2000.
- [2] S. Aggarwal and A. Jindal, "Comprehensive overview of various lip synchronization techniques," *Int. Symp. on Biometrics and Security Technology*, pp.1-6, April 2008

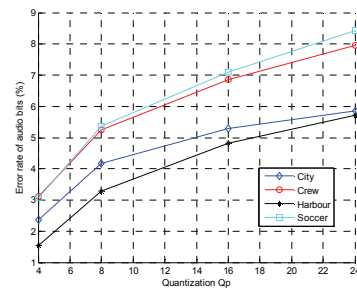


Fig.5 Error rate of audio bits for different sequences

- [3] M. Yang, N. Bourbakis, Z. Chen and M. Trifas, "An efficient audio-video synchronization methodology," *IEEE Int. Conf. on Multimedia and Expo*, pp.767-770, 2007.
- [4] M. Yang and N. Bourbakis, "A high bitrate information hiding algorithm for digital video content under H.264/AVC compression," *Midwest Symp. on Circuits and Syst.*, pp.935-938 Vol. 2, 2005.
- [5] S. D. Wang, C. B. Xiao and Y. Lin, "A high bitrate information hiding algorithm for video in video," *World Academy of Science, Engineer and Technology*, pp. 413-418, 2009.
- [6] L. F. Qi, H. X. Chen and Y. Zhao, "New Synchronization Scheme Between Audio and Video," *Int. Conf. on Software Eng., Artificial Intelligence, Networking, and Parallel/Distributed Computing*, pp.26-29, 2007.
- [7] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu, "A method of estimating coding PSNR using quantized DCT coefficients," *IEEE Trans. on Circuits and Syst. for Video Tech.*, vol.16, no.2, pp. 251- 259, 2006.
- [8] [online] <http://www.xvid.org/>