AN EFFICIENT AUDIO-VIDEO SYNCHRONIZATION METHODOLOGY

Ming Yang[1], Nikolaos Bourbakis[2], Zizhong Chen[1], and Monica Trifas[1]

[1]Math and Computer Science Department
Jacksonville State University
Jacksonville, AL 36265, USA
{myang, zchen, atrifas}@jsu.edu

[2]Information Technology Research Institute
Wright State University
Dayton, OH 45435, USA
nikolaos.bourbakis@wright.edu

ABSTRACT

In a multimedia information system, different types of digital media (such as video, audio, etc.) are stored, transmitted, and presented. During presentation time, the synchronization between audio and video data has to be preserved in order to offer the best perceptual quality. However, the timestamp-based synchronization methodology in MPEG/System layer suffers data packet loss during transmission, and the resulted absence of synchronization will be unacceptable for users. In order to address this issue, an information hiding based synchronization methodology has been proposed. In this methodology, audio data is embedded within the corresponding video frames by means of high bitrate information hiding techniques. At the receiver, the embedded audio data is extracted and played with the host video frames to achieve the synchronization. With this approach, significant advantages have been obtained: (1) the communication channel for audio data transmission is avoided; (2) the synchronization between audio and video data is robust to packet loss.

1. INTRODUCTION

In a multimedia information system, different types of digital media, such as video, audio, and etc, are stored, transmitted, and presented. These different types of digital media possess very different and diverse properties while being temporally dependent on each other. The temporal relationships have to be preserved in order to offer the best perceptual quality during presentation [1][2]. These issues lead to one of the major challenges in multimedia communication system – synchronization [3].

In the MPEG standard, the synchronization between audio and video data is ensured with the timestamp-based approach. The video stream and the audio stream will be multiplexed at the encoder to form a single stream. At the decoder, the single stream will be de-multiplexed into separate video stream and audio stream. Both steams compare their own embedded timestamps with the system clock to ensure synchronization. The timestamp mechanism is easy to implement and has low computational complexity. However, this approach suffers packet loss during transmission [4][5], and thus the video and audio signal will fail to synchronize with each other (Fig. 1).

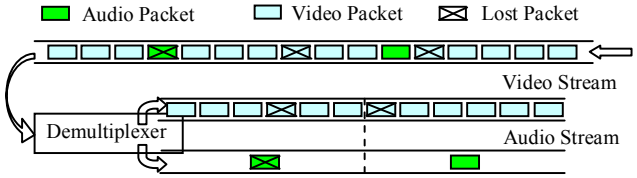


Fig. 1: Audio and Video Packet Loss in MPEG Transmission

In order to address this issue, an information hiding based synchronization methodology has been proposed. According to modern information hiding theory, the embedding capacity of the host video frames provides a communication channel with a certain capacity. Thus, it is possible to make use of this additional communication channel to transmit the associated audio data. The basic idea of the proposed methodology is that audio data is embedded within the corresponding video frames by means of high bitrate information hiding (Fig. 2), without visually degrading the original video frames [6][7].

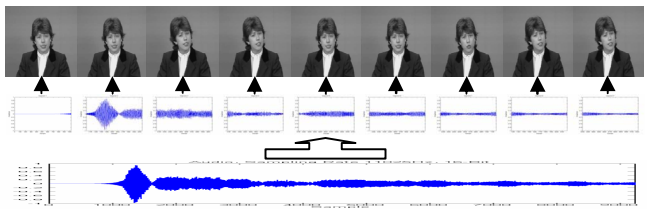


Fig. 2 Audio Segmentation and Information Hiding within Video Frames

The proposed algorithm is very robust to lossy video codec, and thus embedded audio data can be reliably transmitted [8]. During playback, the embedded audio data will be extracted from the video frames and played with the corresponding video data at the same time to achieve synchronization (Fig. 3). In the case of packet loss, the video frame and audio segment will be lost together, and this has no negative effect to the synchronization between video and audio data. With the proposed methodology, the following advantages have been achieved:

(1) no communication channel is needed for audio data transmission, and thus bandwidth is saved;
(2) the video/audio data are guaranteed to be played back properly;
(3) the synchronization between audio and video data can be reliably ensured and will not be affected by packet loss;
(4) the complex tasks of multiplexing, de-multiplexing, and synchronization in MPEG/System have been avoided.

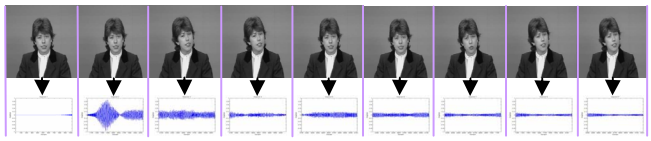


Fig. 3 Audio Data Extraction and Synchronized Playback

2. INFORMATION HIDING STRATEGY

In the proposed methodology, audio data need to be embedded within the video frames without visually degrading the host video data. In order to achieve this goal, the locations of hidden information need to be carefully selected.

2.1 Temporal Location of Hidden Information

In the Group-Of-Pictures (GOP) structures defined in H.26X and MPEG-X, there are three types of frames: I-frame, P-frame, and B-frame, each of which is compressed with different coding modes. It is expected that data embedded within I-frames have the best chance to survive the lossy video codec, because I-frames have the lowest compression ratio. However, if higher channel capacity is desired, P-frames and B-frames will also be used for data hiding.

2.2 Spatial Location in YUV Domain

In video compression, the source video data is comprised of three sample arrays: one luma (Y) sample array and two chroma (U & V) sample arrays. It is observed in human vision experiments that human eyes are more sensitive to changes in chroma component than to those in luma component. As such, only the luma array will be modified to host hidden information to avoid any color distortion.

2.3 Spatial Location in Discrete Cosine Transformation (DCT) Domain

Existing DCT-domain information hiding algorithms try to embed information by perturbing the whole DCT block. However, there are two drawbacks: (1) too much visual distortion; (2) modification of high-frequency coefficients will degrade the performance of run-length coding in JPEG compression.

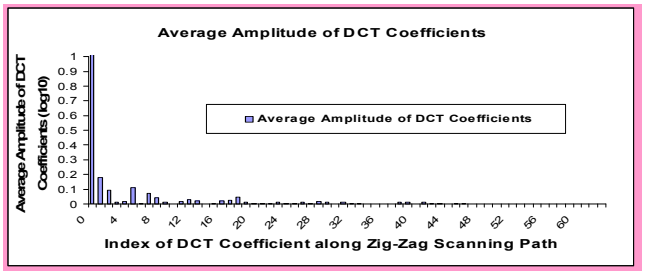


Fig. 4 Average DCT Coefficient Amplitude along Zig-Zag Scanning Path ('Lena' Image)

It is observed that low-frequency coefficients have much higher average amplitude compared to high-frequency coefficients (Fig. 4). In the proposed algorithm, the DCT block is divided into sub-blocks and only the coefficients within as few as only one sub-block will be modified. Experimental results also verify that the proposed algorithm works best at low-frequency coefficients due to their high amplitudes. As such, the choice of sub-band for information hiding is biased to low-frequency DCT coefficients.

3. INFORMATION HIDING ALGORITHM

3.1 Algorithm Overview [9][10]

In high bitrate information hiding, channel capacity (i.e. how many bits can be efficiently embedded within the host video) is mainly concerned. Since digital video contents are stored and transmitted in compressed formats, the robustness of hidden information against lossy video codec is also of great importance. The proposed DCT-based high bitrate information hiding algorithm is graphically illustrated in Fig. 5. The original host video frame is first transformed into frequency domain through 4x4 DCT. After that, DCT coefficients within the 4x4 block will be modified to hide 1 bit data through vector quantization. The coefficient matrix will finally be transformed back to spatial domain to obtain the stego-video.

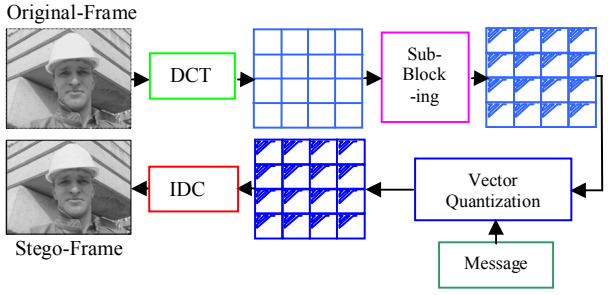


Fig. 5 Information Hiding Algorithm

3.2 Information Hiding Algorithm

The embedding algorithm is described in the following:

(1) DCT (4x4) transform the original video frame;
(2) Scan the 4x4 DCT block along Zig-Zag scanning path;

(3) Convert the 8 low-frequency coefficients to an 8-D vector;

(4) V: the 8-D vector $V = (c_0, c_1, c_2, ..., c_6, c_7)$

T: the threshold for vector quantization

$|V|$: the length of vector V

[]: round-off operation

$$l = |V| = \sqrt{\sum_{i=0}^{15} c_i^2}$$

$$l_T = \left[\frac{|V|}{T}\right] = \left[\frac{\sqrt{\sum_{i=0}^{15} c_i^2}}{T}\right]$$

(5) One bit is embedded by modifying l_T : $l_T' = l_T \pm 0.25$ (+0.25 to embed 1, -0.25 to embed 0);

(6) $l' = l_T' * T$, $V' = \frac{l'}{l} * V$;

(7) Put the vector V' back to its original location in the 4x4 DCT block;

(8) Repeat the same operation for each 4x4 DCT block until all the information bits have been embedded.

3.3 Blind Information Retrieval

The information retrieval algorithm is the following:

(1) DCT transform the stego-frame (video frame with hidden information);

(2) For each 4x4 DCT block, scan the coefficients along Zig-Zag scanning path;

(3) Pick up the 8 lowest frequency coefficients and convert them to an 8-D vector V'';

(4) $l'' = |V''|$

(5) $l_T'' = \frac{l''}{T} = \frac{|V''|}{T}$

(6) $I = l_T'' - [l_T'']$

If $I>=0$, then 1 is extracted as the information bit;

If $I<0$, then 0 is extracted as the information bit.

(7) Repeat the same operation to each 4x4 DCT block until all the information bits have been extracted.

With this approach, the hidden information can be extracted without the presence of the original image. This feature is extremely important in many application scenarios.

3.4 Performance Analysis

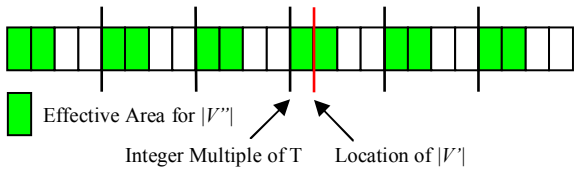


Fig. 6 Survival Regions of the Hidden Information

In order to reliably transmit the embedded audio data, the algorithm has to be robust to DCT-based lossy video codec such as MPEG-X and H.26X. As shown in Fig. 6, the hidden information bit ("1" in this example) will be extracted correctly as long as $|V''|$ is located in the green shaded areas. The Bit Error Rate (BER) can be modeled as:

$$P_e = 1 - p\left(\left|\frac{l_S''}{T} - \left(\frac{l_S'}{T} + m\right)\right| \le \frac{1}{4}\right)$$

$$= 1 - \sum_{m=-\infty}^{\infty} \left(\int_{2(m-\frac{1}{4})T\sqrt{\sum_{i=0}^{n-1}(c_i')^2}+(m-\frac{1}{4})^2T^2}^{2(m+\frac{1}{4})T\sqrt{\sum_{i=0}^{n-1}(c_i')^2}+(m+\frac{1}{4})^2T^2} f_Q(x)\right)$$

where $f_Q(x)$ is the joint probability density function of quantization-factor vector Q in the quantization matrix of video codec. In order to minimize BER, $\boldsymbol{T}$ (threshold for vector quantization) needs to be maximized under the constraint that no visual distortion is resulted.

4. EXPERIMENTAL RESULTS

4.1 Channel Capacity of Video Frames

Here audio data at the sampling rate of 8 KHz and dynamic range of 8-bit has been generated by η-law, and thus the data rate of the audio stream is 64 Kbps. For a QCIF video content at 30 frame/second, the channel capacity is:

(176x144) / (4x4) x 1.5 x 30 = 71.28 Kbps

As shown above, the channel capacity provided by the video frames is higher than the data rate of audio signal. According to Shannon's theory, the audio data can be reliably transmitted in the communication channel provided by the information hiding capacity of the video frames.

4.2 Video Information Hiding

(a) Original Video Frames

(b) Stego Video Frames

Fig. 7 Video Frames Comparison

As shown in Fig. 7, no perceptual degradation can be detected by comparing the original video frames and the stego-video frames (with embedded audio data). The robustness of the proposed algorithm against H.264/AVC codec under different Quantization Parameters (QP) is measured by (1.0-BER) (i.e. the percentage of the hidden information bits that survived the video codec). A GOP structure of "IBPB" is adopted in the video codec. The periodical spikes in Fig. 8 correspond to the I-frames in each GOP. As expected, information within I-frame has the best chance to survive H.264/AVC codec. The smaller spike between two big spikes corresponds to the P-frame (between

two B-frames). The experimental results clearly support the notion that the hidden audio data is able to survive the lossy video codec, and thus can be reliably transmitted.

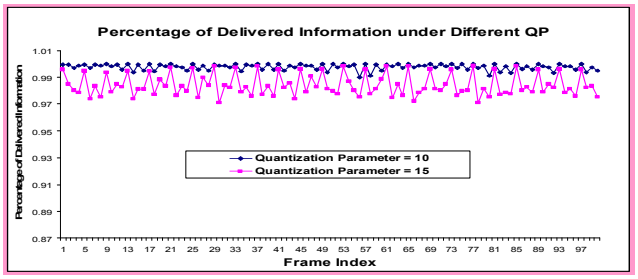
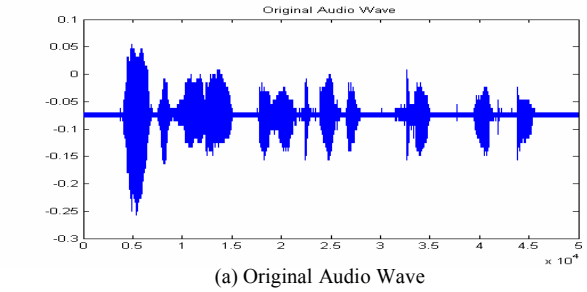


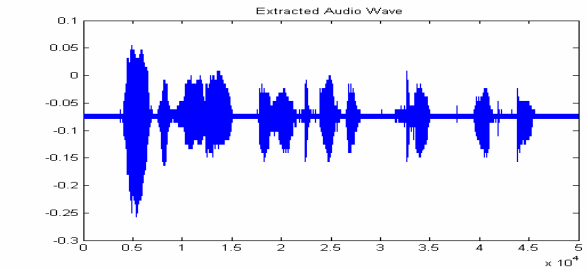
Fig. 8 Performance under Different Quantization Parameters (QP)

4.3 Embedded Audio Signal

The experimental results in section 4.2 strongly verify the conclusion that embedded audio data can be reliably transmitted. With bit error localization approach and filtering, the error of the audio signal can be controlled to a very low level. As shown in Fig. 9, the audio wave extracted from the compressed video frames is almost identical to the original audio wave (8 KHz, 8-bit range) except for some samples corrupted by noise. The bit error and noise can be further suppressed by applying error correction code.



(a) Original Audio Wave



(b) Extracted Audio Wave

Fig. 9 Original Audio Wave and Extracted Audio Wave

5. CONCLUSIONS

In this research, a novel information hiding based synchronization methodology for Video-over-IP has been developed. The key idea is making use of the communication channel provided by the information hiding capabilities of video frames to transmit the associated audio data. The audio data is embedded within the video frames without visually degrading the quality of the host video data. The proposed algorithm is very robust to lossy video codec such as H.26X and MPEG-X, and thus embedded audio data can be reliably transmitted. At the receiver end, the embedded audio data is extracted (without the presence of original video data) and played with the host video frames. With the proposed methodology, the synchronization between audio and video data can be easily and reliably ensured, and the complex tasks of de-multiplexing and synchronization are avoided. The proposed methodology is also robust to packet loss, because the synchronization between audio and video data will not be negatively affected by packet loss. The proposed methodology can be also used for the reliable transmission of caption data and other types of auxiliary data in real-time multimedia applications.

6. REFERENCES

[1] T.D.C. Little and A. Ghafoor, "Synchronization and Storage Models for Multimedia Objects", IEEE Journal on Selected Areas in Communication, Vol. 8, No. 3, pp. 413-427, Apr. 1990.

[2] B.K. Schmidt, J.D Northcutt, and M.S. Lam, "A Method and Apparatus for Measuring Media Synchronization", Proc. of the 5th International Workshop on Network and Operating System Support for Digital Audio and Video, Durham, NH, pp. 130-141, April 1995.

[3] N. Bourbakis, and A. Dollas, "A SCAN based Method for Multimedia on Demand", IEEE Multimedia Magazine, pp. 79-87, July-Sept. 2003.

[4] M. Yang and N. Bourbakis, "A Prototyping Tool for Analysis and Modeling of Video Transmission Traces over IP Networks", Proceeding of IEEE International Workshop on Rapid System Prototyping (RSP 2006), Chania, Crete, Greece, June 2006.

[5] S.G. Aly and A. Youssef, "Real-time Motion-based Frame Estimation in Video Lossy Transmission", Proc. of the 2001 IEEE Symposium on Applications and the Internet, pp. 139-147, 2001.

[6] M.D. Swanson, B. Zhu, and A.H. Tewfik, "Data Hiding for Video-in-Video", Proc. of the 1997 IEEE International Conference on Image Processing (ICIP 1997), pp. 676-679, 1997.

[7] J.J. Chae, and B.S. Manjunath, "Data Hiding in Video", Proceeding of the 6th IEEE International Conference on Image Processing (ICIP 99), Kobe, Japan, vol. 1, pp. 311-315, Oct. 1999.

[8] M. Ramkumar and A.N. Akansu, "Theoretical Capacity Measures for Data Hiding in Compressed Images", Voice, Video and Data Communications, Vol. 3528, pp. 482-492, Nov. 1998.

[9] M. Yang and N. Bourbakis, "A High Bitrate Multimedia Information Hiding Algorithm in DCT Domain", Proceeding of World Conference of Integrated Design and Process Technology (IDPT 2005), Beijing, China, Jun. 13th-17th, 2005.

[10] M. Yang and N. Bourbakis, "A High Bitrate Information Hiding Algorithm for Digital Video Content under H.264/AVC Compression", Proceeding of IEEE International. Midwest Symposium on Circuits and Systems, Cincinnati, OH, Aug., 2005.