EMBEDDING SIGN LANGUAGE INTERPRETATION OF A VIDEO IN THE WAVELET TRANSFORM DOMAIN

Mohsen Ashourian

Azad University of Iran, Majlesi Branch; Isfahan, P.O. Box 86315-111, IRAN
Email: mohsena@iaumajlesi.ac.ir

ABSTRACT

In this paper we propose a low complexity scheme for hiding sign language video signal in another video signal with higher resolution. We decompose the sign language video and the host video using three-dimensional wavelet transform, and embed the information of the sign language video based on their importance in different subbands of the host video. At the receiver the system reconstructs the sign language video by recovering the embedded data in various subbands. The developed system shows robustness in recovering the embedded video even when the host video undergone various type of signal processing operation.

1. INTRODUCTION

Embedding a multimedia signal into another multimedia signal has applications in data hiding and digital watermarking. In digital watermarking applications, emphasis is put on authentication rather than quantity and quality of the recovered data. It is necessary and satisfactory for the watermarking scheme to be able to prove the ownership, even though the host signal undergone various signal processing or geometrical attacks. During the last few years, much progress has been made in developing watermarking techniques that are robust to signal processing operations, such as compression [1, 2].

Data hiding has some other types of applications, such as broadcasting, in which the goal is to use an already established multimedia transmission channel for transmission of another multimedia signal. In this case, we need to recover the embedded information with high quality [3]. However, in these applications, unlike watermarking systems, the host signal usually does not face active and sever attacks. The host multimedia data may only face some signal processing operations, such as compression and addition of noise during transmission.

In this paper, we introduce a data hiding scheme that can be used to convey supplementary information in digital video streams. The hidden data include compressed information of sign language interpretation of the original video signal. Upon receiving video, the receiver may extract the information from the video stream automatically.

The main problem of hiding video in video is the large amount of data that requires a special data embedding method with high capacity as well as transparency and robustness.

There have been few reports on large capacity data embedding [4, 5, 6]. Chae and Manjunath used the discrete wavelet transform (DWT) and lattice code for embedding a signature image into another image [4]. They further improved their system by using a joint source-channel coding scheme and employing the human visual system (HVS) model in the process of information embedding [5]; however, exact adjustment of the HVS model is not easy in many applications, and, the channel optimized quantizer is not suitable in image hiding applications, where intentional or non-intentional manipulations, are variable and not known in advance. In another approach Swanson et. al. [6] designed a method for embedding video in video based on linear projection in the spatial domain. Also the method is not explained well in detail, but in general data embedding in the spatial domain is very sensitive to various types of signal processing operations.

In this paper, we use three-dimensional wavelet decomposition to split the information of the guest and the host video based on their importance. We use over-sampling and data partitioning for further protection of the important portion of guest video when the host video faces various type of signal processing and compression operations. In the following sections, at first we explain the method of sign language and host video decomposition. In Section 3 we explain the method of hiding and extraction of video. In Section 4 we provide the experimental results, and finally in Section 5 we conclude the paper.

2. SIGN LANGUAGE ENCODING

Studies [7,8] show that effective deaf sign language communication requires temporal resolution of 8 to 10 frames per seconds, though the size of pictures can be 80*60 pixels per frame or lower. In the early 1980's, several researcher proposed system for transmitting sign language using low resolution image over low bandwidth channels. Most of these used two-level images (cartoons) generated automatically with an edge or line finding algorithm. One such system was developed into special-purpose hardware [9] which ran over conventional telephone lines for deaf people's home. The equipment was expensive and the project was terminated in the early 1990's partly on the basis of increasing penetration of ISDN networks.

The key information for survive of the sign language messages are the motion of the body and some features of the face. We use a 3-D wavelet (or subband)

decomposition for splitting the information of the video signal and proper selection of them. A 3-D subband coder uses a unique approach for encoding intra-frame and inter-frame redundancy in a video sequence. The video signal passed through a 3-D filter bank and then different subbands are encoded based on their visual importance [10].

Fig. 1 shows the structure of three-dimensional wavelet decomposition used for both video signals. The terms HP and LP refer to high-pass filtering and low-pass filtering, where the subscripts *t, h,* and *v* refer to temporal, horizontal, and vertical filtering respectively. The selected subband framework consists of 11 spatio-temporal frequency bands. The temporal frequency decomposition is restricted to only two subbands due to potential delay problems in a practical implementation and reducing dependency in coding consecutive frames. The image frames are filtered temporally using the two-tap Harr basis functions, Temporal decomposition is followed by horizontal spatial filtering and vertical spatial filtering using use Daubechies'6 (db6).

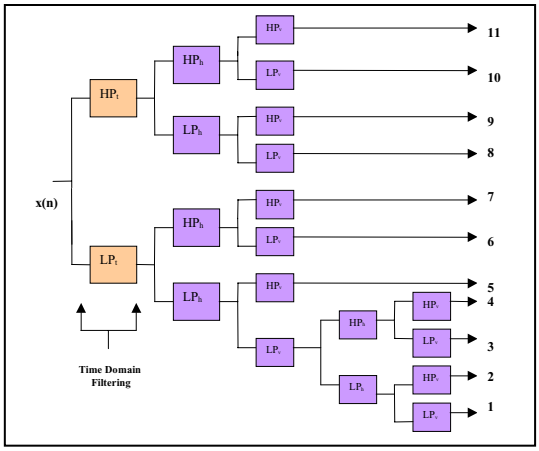

Fig. 1: 3-D Filter Bank Structure

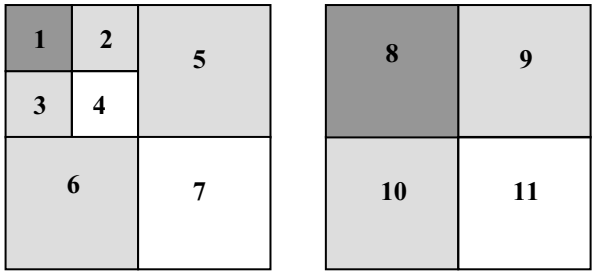

Fig.2: 3-D Filter Bank Frequency Map

Since the information of sign language video signal are embedded in another video signal that might face various types of signal processing operation, we need to protect the key information of the signal properly. Figure 2 shows the frequency map of the 11 video subbands. Based on the energy and visual quality importance, we classify the subbands to three groups. Bands 1 and 8 are the visually most important subbands. Band 1, the lowest spatio-temporal frequency subband, contains the main skeleton of the video scene, and Band 8, the lowest spatial frequency band of high temporal band, shows the amount of change consecutive frames. We use over-sampling to add redundancy to these subbands. Figure 3 and 4 show the method of over-sampling by zero-padding in the DCT domain. Similar to multiple description coding scheme [11, 12] the added redundancy helps to increase the chance of signal recovery in error prune communication channel. As we explain in the following section, the pixels of the over-sampled subbands are split into two groups in different spatio-temporal portions of the host video, so that it can be protected more.

The second and third groups of subbands are visually less important, as they contain the texture information or high frequency components of the signal. We drops the bands 4,7 and 11 as they have very low energy and only keep the other six subbands (2,3,5,6,8,9), without addition of any redundancy or protection scheme.

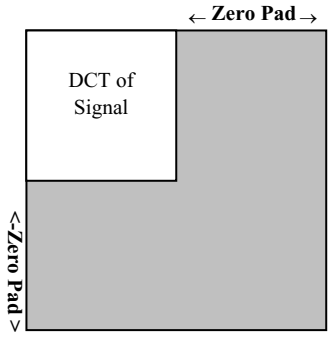

Fig. 3: Zero-padding Signal

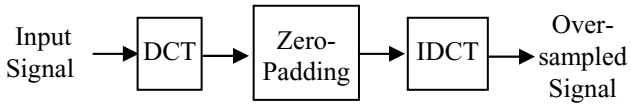

Fig. 4: Over-sampling in the DCT Domain

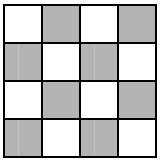

Fig. 5: Splitting the Over-sampled Subband

3. DATA EMBEDDING AND EXTRACTION

The data embedding in the host image could be in the spatial or transformed domains. The spatial domain embedding schemes are simpler to implement, but their capacity for data embedding is lower than transformed domain schemes [1]. Various transformation domain methods were suggested for data embedding [2,3]. The DCT and wavelet domains were more popularly selected due to their compatibility with compression methods.

For our experiments, we embed the sign language video with frame size of 88×72 pixel and rate of 10 frames per second into the host video with CIF format (352×288 pixels and 30 frames per second). Both video sequences are monochrome.

We decompose the host video using similar three-dimensional decomposition depicted in Figure 2.

Since the sign language video frame rate is 1/3 of the host video and the its spatial size is 1/4 of it, we split and distribute the information of its subbands among the host video subbands, in a way to reduce the visibility distortion in the host video, and at the same time maintain high quality recovery of the embedded video. Here are the details of the embedding policy for each band of the sign language video:

- Band 1 is over-sampled four times. The over-sampled subband has 44*72 pixels. With proper weighting, we split this band, and place them in the position of some pixels in Bands 2 and 4 of the original host video in consecutive frame decomposition.
- Band 8 is also over-sampled four times. The over-sampled subband has 176*144 pixels. We split these information into three portions and placed them in the position of Band 8 of the host video among 3 consecutive frames.
- Bands 2, 3, 5 and 6 of the sign language video sequence, which are totally have 3872 wavelet coefficients, are distributed among the bands 5 and 6 of the host video sequence. Each of these subbands has 176*144=25344 pixels. Since the rate of host video is three times of the guest video, only 1/4 of the host video sequence wavelet coefficients in these subbands should be replaced, that would not have great perceptual distortion.

At the receiver, we first reconstruct the lowest frequency subband of the sign language video from the extracted indices in each portion of the host video, and recombine the two over-sampled information. As the lowest frequency band is a blurred version of the original video, we can estimate the corrupted pixels from the over-sampled subband using various error detections and concealment methods [13]. In developed system, we follow a simple scheme based on comparison of each pixel with its neighboring pixel average value. As Fig. 5 shows each pixel has four neighboring pixels from its own description and four neighboring pixels from the other description.

For the pixel with intensity value $x_{i,j}$, first we calculate the average value of neighboring pixels in the first descriptions: m_1, and the second description: m_2; and then we calculate

$$\lambda_1 = \left| \frac{x_{i,j} - m_1}{m_1} \right| \quad (1)$$

$$\lambda_2 = \left| \frac{x_{i,j} - m_2}{m_2} \right| \quad (2)$$

High value of λ_1 or λ_1 suggests possibility of corruption of the pixel. In this case we can replace the pixel with m_2 or m_1. We set the threshold for these parameters (λ_1 and λ_1) based on the amount of distortion the host video faces.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The objective of project is to transmit sign language interpretation of a video sequence using the same communication channel. We arranged a video sequence that speaker tells 30 independent sentences. We embed the sign language interpretation of the sentence in it.

The embedding scheme did not result to high visual distortion and the average PSNR of the host video sequence stays above 40 dB. In order to evaluate the system performance in transmission of sign language, three deaf persons familiar with sign language interpret each sentence after data recovery, and we report the average percentage of correct sentence recovery (PCSR). In the normal situation that host image does not facing any modification or signal processing operation, the PCSR is above 95%. We test the system resistance to MPEG compression. The MPEG-II compression algorithm with various compression ratios (CR) is tested. Table 1 shows the percentage of correct sentence recovery (PCSR) for each case. In another experiments, additive noise with various standard deviations was added to the video sequences. The PCSR of recovered video based on PSNR of the host video are shown in Table 2.

5. CONCLUSION

We have presented a new scheme for embedding a sign language video sequence into a host video sequence. We used three-dimensional subband decomposition and over-sampling of the key information for robust encoding of the sign language video. We use three-dimensional subband decomposition for the host video and embed the information of sign language signal with very low visible distortion in the host video. The results show that the system is able to transmit the sign language video even when the host video face operations like compression and addition of noise.

REFERENCES

[1] F.A.P. Petitcolas., R.J. Anderson, and M.G. Kuhn, "Information Hiding-a Survey," *Proceedings of the IEEE*, Vol. 87, No.7, 1999 pp. 1062-1078.

[2] F. Hartung, and M. Kutter, "Multimedia Watermarking Techniques," Proceedings of the IEEE , Vol.87, No.7 , 1990, pp.1079-1107.

[3] R.B. Wolfgang, C.I. Podilchuk, and E.J. Delp "Perceptual watermarks for digital images and video," *Proceedings of the IEEE*, Vol. 87, No. 7, 1999, pp. 1108-1126.

[4] J.J. Chae, , and B.S. Manjunath, "A Robust Embedded Data from Wavelet Coefficients," *Proc. of SPIE, Storage and Retrieval for Image and Video Databases VI*, 1998, pp. 308-317.

[5] D. Mukherjee, J.J. Chae, S.K. Mitra, and B.S. Manjunath, "A Source and Channel-Coding

Framework for Vector-Based Data Hiding in Video," *IEEE Transaction on Circuits and System for Video Technology*. Vol. 10, No. 6 , 2000, 630-645.

[6] M. D. Swanson, B. Zhu, A. H. Tewfik, "Video Data Hiding for Video-in-Video and Other Applications," *Proc. of SPIE Multimedia Storage and Archiving Systems*, vol. 3229, pp. 32-43, Dallas, TX, November, 1997.

[7] G. Sperling, "Video Transmission of American sign language and finger spelling: Present and projected bandwidth requirement," I*EEE Transaction Communication*, Vol. 20, pp. 993-1002, Dec. 1981.

[8] G. Hellstrom, "Quality measurement on Video Communication for Sign Language," *Proc. Of 16th International Symposium on Human Factor in Telecommunications*, Norway, May 1997, pp. 2127-2224.

[9] P. Letellier, M. Nadler and J.F.Abramatic, "The telesign Project," *Proceeding of IEEE* , vol.37, pp. 813-827, April 1985.

[10] C. Podilchuk, N. Jayant and N. Farvardin, "Three-dimensional subband coding of video." *IEEE Transaction on Image Processing* . vol. 4, No. 2, pp. 125-139., 1995.

[11] V. K. Goyal, "Multiple Description Coding: Compression Meets the Network," *IEEE Signal Processing Magazine*, Vo.18, Issue 5, pp.74-93.,2001.

[12] M. Ashourian, Y.-S. Ho. "Multiple Description Coding for Image Data Hiding Jointly in the Spatial and DCT Domains," *Lecture Notes in Computer Science (LNCS)* , Vol. 2836, 2003.

[13] S. Shirani, F. Kossentini, and R. Ward, "Error concealment methods, a comparative study," *IEEE Canadian Conference on Electrical and Computer Engineering*, Vol.2, pp.835-840 , 1999.

Table 1. PCSR of the recovered sign language video after MPEG compression of the host video

CR	3	6	12	24
PCSR %	92.4	86.1	80.6	73.0

Table 2. PCSR of the recovered sign language video after addition of noise to the host video

PSNR (dB)	35	30	27	25
PCSR %	93.8	87.6	74.3	68.4