A VIDEO MULTIPLEXING SCHEME USING DATA EMBEDDING

Mohsen Ashourian
Department of Information Technology, Faculty of Engineering, Isfahan University, Isfahan, IRAN
mohsena@ieee.org

ABSTRACT
In this paper we propose a video multiplexing scheme using data embedding techniques for transmission of a low resolution video signal with another video signal through the same communication channel. Our proposed system can be implemented by addition of a pre- and post-processing block. We encode the low resolution video and embed the encoded bit-stream in the main host video. We develop a low complexity and robust encoding scheme for low resolution video by applying three-dimensional wavelet decomposition on it, and encoding the important subbands using multiple description scalar quantizers. We also apply three-dimensional wavelet decomposition on the host video and embed the encoded bit-stream of the low resolution video in the host signal subbands. At the receiver the system reconstructs the low resolution video by recovering the embedded data in subbands. We examine the system performance by evaluating the quality of recovered video after host video faces various types of processing. At the receiver we evaluate the percentage of correct sign language messages recovery. We further evaluate the system performance when the host video undergone compression and addition of noise during transmission.

1. INTRODUCTION
One potential demand is addition of a low resolution video transmission system to an already established multimedia transmission system. In our application, we assume that there is an established high quality digital video transmission system, and we would like to use it for transmission of another video with lower resolution without interrupting the main transmission system. As depicted in Figure 1, we propose a multiplexing system by encoding the low resolution video and embedding the encoded bit-stream in another digital video signal.

Embedding a multimedia signal into another multimedia signal has been popularly used in digital watermarking [1]. 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.

In data hiding applications, we need to recover the embedded information with high quality. However; in these applications, unlike watermarking systems, the host signal usually does not face active and severe attacks by unauthorized people for destroying the embedded data. The host multimedia data may only face signal processing operations, such as compression and addition of noise during transmission.

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 [2,3,4]. Chae and Manjunath used the discrete wavelet transform (DWT) and lattice code for embedding a signature image/video into another image/video [2,3]. They further improved their system by using joint source-channel quantizers. However; the channel-optimized quantizer is not suitable in data hiding applications, where intentional or non-intentional manipulations are variable and not known in advance.

In another approach Swanson et. al. [4] designed a method for embedding video in video based on linear projection in the spatial domain. The method is flexible and simple to implement, but like other spatial domain embedding techniques, it is not robust to compression [1].

In [5] we propose an image data hiding scheme that can hide a gray-scale image in another gray scale image with bigger size. The key advantage of the developed system is using multiple description coding which is a joint source-channel coding method [6,7,8]. In this paper, we extend the image hiding scheme for embedding a video with lower resolution in another video with higher resolution (host video). In the following sections, at first we explain sign language encoding scheme in Section 2. 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 give conclusion and suggestion for further research.

Fig. 1. Overview of the Proposed System
filtering using 9/7 biorthogonal filters. Horizontal-spatial filtering and vertical-spatial filtering respectively. The selected subscripts t, h, and v refer to temporal, horizontal, and vertical filtering. The amplitude distribution of Bands 1 and 8 does not follow any fixed probability distribution function (PDF) [7]. We use phase scrambling operation to change the PDF of these bands to a nearly Gaussian shape [8]. The added random phase could be an additional secret key between the transmitter and the registered receiver. Since the information of the 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 subbands properly. We use multiple description PDF-optimized scalar quantizer for Bands 1 and 8, as they follow well a Gaussian distribution after phase-scrambling. Multiple Description Coding (MDC) is a joint source-channel coding technique where the source is encoded by multiple descriptions (MDC) is a joint source-channel coding technique where the source is encoded by multiple descriptions. The amplitude distribution of Bands 1 and 8 does not follow any fixed probability distribution function (PDF) [7]. We use phase scrambling operation to change the PDF of these bands to a nearly Gaussian shape [8]. The added random phase could be an additional secret key between the transmitter and the registered receiver. Since the information of the 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 subbands properly. We use multiple description PDF-optimized scalar quantizer for Bands 1 and 8, as they follow well a Gaussian distribution after phase-scrambling. Multiple Description Coding (MDC) is a joint source-channel coding technique where the source is encoded by multiple descriptions and there is some redundancy among descriptions. These descriptions are transmitted over separate channels for error protection. Figure 5 shows the block diagram of MDC. At the receiver, the multiple description decoder combines the information of the two descriptions and reconstructs the original signal. It can decode only one channel, when data on the other channel is highly corrupted; otherwise it can combine the received information from both channels.

Selection of optimum quantizer for different subbands based on their statistical characteristics and visual importance is the key factor for developing subband coder. Figure 4 shows the frequency map of the 11 video subbands which can be classified as below:

1. Band 1, the low temporal and spatial frequency band, is a blurred version of the original video frame. It has much higher energy compared to other subbands and has the most visual importance. However while all the subbands histogram follow well a generalized Gaussian distribution, this subband does not follow any fixed distribution [8].

2. Bands 2-7, the low temporal and high spatial frequency bands, include information of texture and sharpness of video signal in the spatial domain. Depends on amount of these information in scene, the energy of these bands could be higher or lower. Among these bands, bands 4 and 7 have much lower due to two times highpass filtering (vertical and horizontal).

3. Band 8, the high temporal and low spatial frequency band, has higher average energy compared to other high temporal bands, and it shows the major changes in consecutive video frames.

Bands 9-11, the high temporal and high spatial frequency bands (Bands 9-11) have low energy, but high variation in time. They represent sharp and fast objects movements in the video scene.

In our proposed application we embed the bit-stream of encoded video in another host video that might face various signal processing operations like compression and addition of noise. Therefore, we need to use a robust video coding scheme. The hybrid coders have high encoding efficiency, but they are susceptible to noise due to using variable length code, and using motion estimation/compensation that increases the possibility of error propagation in group of frames. With knowledge of communication channel, it is possible to use various channel coding schemes for protection of the encoded bit-stream in a hybrid coder. However, in data hiding application usually there is not enough information about the type of processing that host signal faces, and the effect of this operation on the embedded information can not be easily estimated. Therefore, we preferred to use a robust source coding scheme, despite its lower coding efficiency.

We use a three-dimensional (3-D) subband coder with fixed rate quantizers [8]. Splitting the spatio-frequency information among subbands, and the fixed rate quantizers increase the encoder resistance to loss of information that can be happened due to various processing on carrier (or host) video.

Figure 3 shows the structure of 3-D subband decomposition used for both video signals. A three-dimensional 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 [9]. 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 [9]. Temporal decomposition is followed by horizontal-spatial filtering and vertical-spatial filtering using 9/7 biorthogonal filters.

2. ENCODING THE LOW RESOLUTION VIDEO

Figure 2. Multiple description source coding

Encoder

MD

Channel 1

Channel 2

Decoder 1

Decoder 2

Decoder 0

Fig. 2. Multiple description source coding
The other high frequency subbands contain the texture information of the video scene which are not visually important in recognizing the sign language video. We drop Bands 4, 7 and 11, as they have very low energy and only quantize the other six subbands (2, 3, 5, 6, 8, 9) with PDF-optimized quantizer assuming Laplacian PDF for them. We split the quantizers indices into two groups and embed them in different spatio/temporal position or frequency bands of the host video to have higher protection against compression and noise addition. The splitting and scrambling of the encoded bit-stream reduce the effect of loss of information on the reconstructed video.

3. DATA EMBEDDING AND EXTRACTION

The data embedding in the host video 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 transform domain methods were suggested for data embedding [1]. The DCT and wavelet domains were more popularly selected due to their compatibility with popular compression methods.

We embed the low resolution video with QCIF frame size of (176×144) pixel and rate of 15 frames per second into the host video with CIF format (352×288 pixels and 30 frames per second). The sign language video is monochrome, but the host video is a colored video in YUV (4:1:1) format. We decompose the Y component of the host video using the same 3-D subband decomposition depicted in Figure 3. The chroma components are decomposed only in 8 subbands as depicted in Figure 6. The high temporal subbands usually have low energy, and they can be easily affected by various types of compression schemes, therefore, we embed the date only in the medium frequency and low temporal subbands. The selected subbands are: 2, 3, 5 and 6 of luminance signal and bands 2, 3, 6 and 7 of chrominance signal.

The host video is in 4:1:1 format, it means for each group of four luminance pixels, there is only one pixel with two chrominance information. However the visual system is less sensitive to color distortion. We split and distribute the encoded bit-stream among the selected host video subbands blocks. Since the sign language video has much lower tempo/spatial resolution compared to the host video (1/4 in spatial and 1/2 temporal resolution), therefore we have enough flexibility in distributing the embedded bits. We embed the bit-stream of signature information in the host video frames in the area with high texture content to be less visible. In order to evaluate the texture content of a (4×4) block, we define a normalized measure for the energy of high frequency bands

$$\mu_k = \frac{\mathbf{e}_H}{\mathbf{e}_L}$$

where \(\mathbf{e}_H\) is the average of the absolute value of the high frequency bands (2, 3, 5, 6), and \(\mathbf{e}_L\) is the absolute value of the lowest frequency band (Band 1) of the corresponding block. \(\mu_k\) characterizes the given block texture energy. Higher value of \(\mu_k\) shows the block has strong high frequency component or high texture. We consider these blocks good candidates for data embedding, and replace wavelet coefficients with embedded data [9].

We split and distribute the encoded bit-stream among the selected host video subbands blocks. Since the sign language video has much lower tempo/spatial resolution compared to the host video (1/6 in spatial and 1/2 temporal resolution), therefore we have enough flexibility in distributing the embedded bits. We embed the significant bits of quantization indices in higher frequency subbands and less significant bits in lower frequency subbands (Bands 5 and 6). This will increase the robustness of our data embedding scheme to compression methods which are usually affecting high frequency bands.

At the receiver, we first reconstruct the two important frequency subbands (Band 1 and 8) of the sign language video from the extracted indices in each portion of the host video, and recombine the indices. If the two indices are close together we can recombine
them and use the multiple description scalar quantizer table to have higher resolution [6]. On the other hand if the two indices were far from each other, we assume that one of them is highly corrupted. In order to choose the less corrupted index, we compare the difference between the block containing those indices with former blocks in time-domain.

3. EXPERIMENTAL RESULTS AND ANALYSIS
In our developed system, we do not want the host video to face any visual distortion due to data embedding. At the same time, we would like to recover sign language video with reasonable quality. In all of our experiments we set the embedding factor so that average PSNR of the host video sequence stays above 35 dB.

There are various ways to evaluate the recovered video depends on the application. We use two color video sequences of "Akio" and "Salesman" as the host image with QCIF format, and embed the low resolution of the same video sequences, but monochrome, with QCIF and monochrome are embedded. We report the average percentage PSNR of the low resolution video after recovery. In the normal situation that host video does not face any modification or signal processing operation, the PSNR is above 25dB We test the system resistance to MPEG-2 compression. The MPEG-2 compression scheme with various compression ratios (CR) is tested. Table 1 shows the value of recovered video PSNR for each case. In another experiment, additive noise with various standard deviations was added to the host video sequences. The PSNR of recovered video based on PSNR of the host video are shown in Table 2.

4. CONCLUSION
We have presented a new scheme for embedding a low resolution video sequence into a host video sequence. We used 3-D subband decomposition and multiple description coding of the key information for robust encoding of the low resolution video. We also use subband decomposition for the host video and embed the information of the low resolution signal with very low visible distortion in the host video. The results show that the system is able to transmit the low resolution video even when the host video face operations like compression, addition of noise and down-sampling. The system has very low complexity and it can be used for transmission of low resolution video in an already established video communication system. It is possible to use the low resolution video for error concealment of the original video.

5. REFERENCES

Table 1. PSNR of the recovered low resolution video after MPEG compression of the host video

<table>
<thead>
<tr>
<th>Compression Ratio</th>
<th>6</th>
<th>12</th>
<th>24</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR of The Recovered Video</td>
<td>24.2</td>
<td>20.8</td>
<td>17.9</td>
<td>16.8</td>
</tr>
</tbody>
</table>

Table 2. PSNR of the recovered low resolution video after addition of noise to the host video

<table>
<thead>
<tr>
<th>PSNR (dB) of the Host Video</th>
<th>35</th>
<th>30</th>
<th>27</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR of The Recovered Video</td>
<td>25.4</td>
<td>21.1</td>
<td>20.2</td>
<td>18.1</td>
</tr>
</tbody>
</table>

Table 3. PSNR of the recovered sign language video after down-sampling the host video

<table>
<thead>
<tr>
<th>Down-Sampling</th>
<th>1/2</th>
<th>1/4</th>
<th>1/8</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR of The Recovered Video</td>
<td>20.3</td>
<td>18.0</td>
<td>16.5</td>
</tr>
</tbody>
</table>