Raw Speech Based Digital Watermarking Using Zerotrees of DWT

Sataporn Schwindt and Thumrongrat Amornraksa
Multimedia Communications Laboratory, Department of Computer Engineering,
Faculty of Engineering, King Mongkut's University of Technology Thonburi,
91 Pracha-Uthit Road, Bangkok 10140, Thailand
Tel: +66-2-470-9083, Fax: +66-2-872-5050
E-mail: sataporn_schwindt@hotmail.com, t.amornraksa@cpe.eng.kmutt.ac.th

Abstract: In this paper, the zerotrees of DWT is proposed to be used in a speech based digital watermarking for digital images. Since in this research work the raw speech and its content are used as a watermark signal, in the watermarking scheme, the PCM coded speech signal is embedded into a sequence of images. The performance of the scheme is evaluated by the PSNR obtained from the watermarked images and the strength of attacks the embedded speech signal can survive. Moreover, since in this research work the contents contained in the speech is used to identify the specific information hidden in the embedded signal. The speech signal after being extracted from the watermarked images is played back to the listeners to determine whether its content is intelligible or not. The experimental results show impressive performance of the scheme implementing our proposed technique, judged by the higher robustness obtained form the embedded signal against various types of attack, including brightness/ contrast enhancement, Twirling, highpass filtering and JPEG compression standard.

1. Introduction

With remarkable characteristics of digital representation and distribution such as convenience in manipulation and duplication, the growth of digital information technology has been the growth of digital information technology has been rapidly improved. On the contrary, possible tendency of redistributing digitized data, without permission from the original owner, has become a common issue due to its extraordinary characteristics. There are presently many approaches, which enable preventing this kind of problem. Digital watermarking is one of the techniques used to solve such a problem by securely embedding some information into multimedia data, in such a way that it is invisible to the Human Visual System (HVS). The watermark data to be embedded can be any information such as copyright owner, authorized recipient or purchasing information. This information can later be used in some ways to identify the original owner or the traitor etc. Generally, the basic requirements of an efficient digital watermarking technique should fulfil the following: invisible, undetectable, unalterable and unambiguous. Moreover, the embedded watermark signal should survive all possible attacks, including common signal processing based attacks.

In this paper, a concept of embedding raw speech as watermark signal into a sequence of images is presented. The idea of using numerous redundancies contained within the raw speech in recognition process to enhance the robustness of the embedded watermark signal is considered. As long as the raw speech, extracted from the modified or attacked watermarked image, contains enough important information, its contents can still be intelligible. Moreover, since several compression schemes nowadays employ wavelet-based techniques for reducing the redundancies contained inside the data. The raw speech is therefore embedded into the sequence of images by a wavelet-based technique i.e. zerotrees of discrete wavelet transform. The results from experiments show the impressive quality of the watermarked image. Furthermore, when the embedded speech is extracted, after being attacked by some common signal processing, the listening test is performed to evaluate the intelligibility contained within the extracted speech. The experimental results also show that the impression of the human perceptual system, as it tends to adjust and learn quickly to recognize the repeated speech, enhances the probability of recognizing the contents of the extracted raw speech precisely.

2. Related work

Various efficient techniques designed for watermarking purposes have been proposed in the past few years. For example, in [1], a DCT-based watermarking technique was proposed, where the watermark signal is embedded into the mid-frequency range of DCT coefficients of the original image. However, when the watermarked image is compressed with a high ratio, the watermark embedded in the mid-frequency range of DCT coefficients will be destroyed. An efficient watermarking technique based on zerotrees of DCT was proposed in [2]. The technique embedded the watermark signal by changing the threshold value of coefficients derived by the DCT transform. Doing this can change an amount of zerotrees an image contains, and the number of zerotrees (odd or even) was used to identify if the watermark bit is 1 or 0. This technique does not need the original image in the retrieval process, but needs to carefully consider the suitable reference threshold value instead.

A similar technique, which employed the wavelet transform, was proposed by [3]. In this technique, the positions of zerotree defined in the embedded zerotree wavelet (EZW) and threshold value were used to detect the embedded data. In [4], the authors implemented a scheme for hiding 8KHz speech sampled at 16 bits/sample in a 30 frames/s QCIF video. In the scheme, the successive samples of speech were vector-quantized, and the indices were embedded into the LL-HH subband coefficient. Then watermarked video was piped through a H.263 encoder. The speech and video extracted from the compressed video at high compression ratio were found to be intelligible, and acceptable for visual quality, respectively.

3. Description of the proposed technique

3.1 Wavelet transform

Wavelets are mathematical functions that divide data into different frequency component, and then study each component with a resolution matched to its scale [5]. Each transformation of wavelets involves correlating the wavelet with the given signal (derived from image), where the coefficient value depend on how closely correlated the wavelet is with the given part of the signal. The wavelet is stretched after the entire signal is covered and the process is executed in this fashion repeatedly for all scales. The arrangement of coefficients in the wavelet transform is showed in Figure 1.

1	2	5	6	17	18	21	22
3	4	7	8	19	20	23	24
9	10	13	14	25	26	29	30
11	12	15	16	27	28	31	32
33	34	37	38	49	50	53	54
35	36	39	40	51	52	55	56
41	42	45	46	57	58	61	62
43	44	47	48	59	60	63	64

Figurel: Arrangement of coefficients in 8x8 pixels block after taking wavelet transform 3 times

Basically, the spatial coefficient tree is considered as the set of coefficients from different bands that correspond to the same spatial region in an image. The lowest frequency band is the root node of the tree and the highest frequency band is the leaf nodes of the tree. There is a parent-child relation between the higher and the lower frequency coefficients in the same spatial location. Each parent node represents a lower frequency component than its children. Figure 2 illustrates the parent-child relationships of the spatial coefficient of the DWT block. Note that the arrows identify the parent-child dependencies.

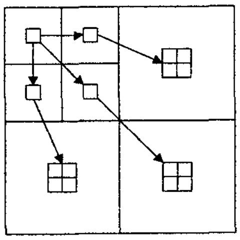

Figure 2: Parent-child relationships

3.2 Watermark embedding/extraction technique

The block diagram watermark embedding process is shown in Figure 3. The original sequence of images is first divided into individual image frames. Each frame is then segmented into 8×8 non-overlapping blocks, and each block is transformed by using the DCT twice.

The transformed coefficients are then compared with the pre-defined threshold value to classify all coefficients into two groups; one and zero. In this research work, the value '0' is used as the threshold to determine the zerotree. That is, if the value of coefficient is negative (or positive), after thresholding process, its value is set to zero (or one respectively).

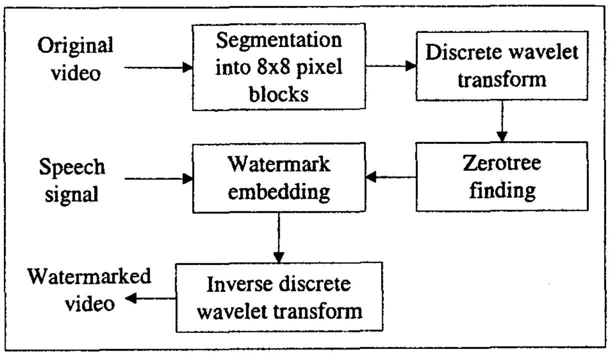
Original video
Segmentation into 8x8 pixel blocks
Discrete wavelet transform
Speech signal
Watermark embedding
Zerotree finding
Watermarked video
Inverse discrete wavelet transform

Figure 3: Block diagram of the watermark embedding process

In the watermark retrieval process, the watermark bits can be directly extracted from the watermarked image frames without the need of original image sequence. Firstly, the watermarked frame is divided into 8x8 pixel blocks. Then, the DWT is taken to each block, and the resultant coefficients are compared with the pre-defined threshold to obtain the zerotree. By considering the number of zerotree contained within each block, the watermark bits can be recovered. Figure 4 illustrate the block diagram of the watermark extraction process.

The watermark extraction process is shown in Figure 5. The watermarked image is divided into 8x8 pixel blocks. After that, the wavelet transform are performed to obtain the zerotree in each block.

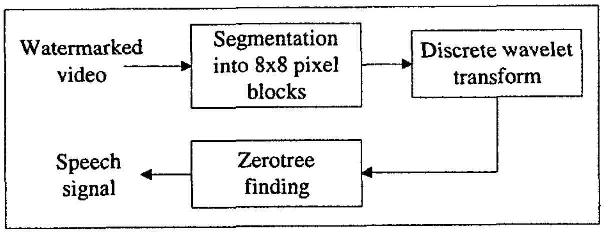
Watermarked video
Segmentation into 8x8 pixel blocks
Discrete wavelet transform
Speech signal
Zerotree finding

Figure 4: Block diagram of the watermark extraction process

Figure 5 illustrates an example of finding the zerotree of the DWT coefficients from the DWT transformed blocks. From the given example, three zerotrees are found e.g. at the coordinates {(2,4), (3,7), (3,8), (4,7), (4,8)}. The technique for embedding the watermark bit depends on the number of zerotrees contained within each block. That is, there will be no change if the number of zerotrees is odd and the watermark bit we want to embed is one. However, if the number of zerotrees is even, a change must be made in some coefficients until the number of zerotrees in that block is odd. Nevertheless, the change must be performed in a proper area in which its change should not degrade the image quality too much, at the same time, this change should be robust against any signal processing based attacks. In this research work, mid frequency area is considered and used to make a change.

194.75	349.25	-5.75	-39.75	-1.5	-5.5	-13	-17.5
242.75	507	-16.25	-28	-2	-4.5	-7	5
-8.25	-76.25	4.25	4.75	6.5	-4	-13	-10
-51.25	-61	-11.75	-12	-10	-12	-10	-3
-1.5	0.5	-14	16.5	0.5	0.5	2	1.5
-2	-5.5	-19	-17	0.5	1.5	0	-1
5.5	-14	-22	-12	-1.5	-3	0	-2
-9	-21	-19	5	-1	-0.5	-2	-2

(a)

1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	1
0	0	1	1	1	0	0	0
0	0	0	0	0	0	0	0
0	1	0	1	1	1	1	1
0	0	0	0	1	1	1	0
1	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0

(b)

0	0	1	0	1	1	0	0
0	0	0	1	1	1	0	0
0	0	0	0	0	0	1	1
0	0	1	0	0	0	1	1
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0
0	0	0	0	1	1	0	0

(c)

Figure 5: (a) The DWT transformed coefficient block (b) the coefficients after thresholding process (c) the zerotrees left in the block

4. Experimental setting

To prepare the speech signal, Pulse Code Modulation (PCM) technique with 8 KHz sampling rate was used to digitize the speech input from the microphone. The PCM encoded words "*Sataporn Schwindt*" was used as the watermark signal, while the sequence of images "*Miss America*" was used as the original multimedia data. The watermark signal was then embedded into the images using the process as described in Section 2. It should be noticed that since the size of the watermark signal is larger than the total numbers of zerotrees within one image frame, the zerotrees from several image frames were used in stead.

The next step, common image processing based attacks were applied to the watermarked image frame including brightness enhancement, contrast enhancement, twirling, highpass filtering and JPEG compression standard. After being attacked, the embedded speech signal was extracted and played back to the listeners to hear whether they could recognize the spoken words in the corrupted speech signal or not. The same format of the standard listening test, suggested in [6] and also used in [7], was finally used to evaluate the efficiency of the proposed techniques.

5. Experimental results

In the experiments, the sequence "*Miss America*" was first decoded into individual image frames with the size of 352×288 pixels. Then the wavelet transform was applied twice, and the zerotree in each transformed block was determined for embedding the watermark signal. Note that with this image size, the watermark signal up to 1584 bits, can be embedded into one image frame. The original image frame no. 0 from the sequence "*Miss America* and its watermarked version using zerotree of DWT with the PSNR of 52.498 dB are shown in Figure 6 (a) and (b), respectively. The original speech waveform "*Sataporn Schwindt*" and its extracted version without being attacked are illustrated in Figure 6 (c) and (d), respectively.

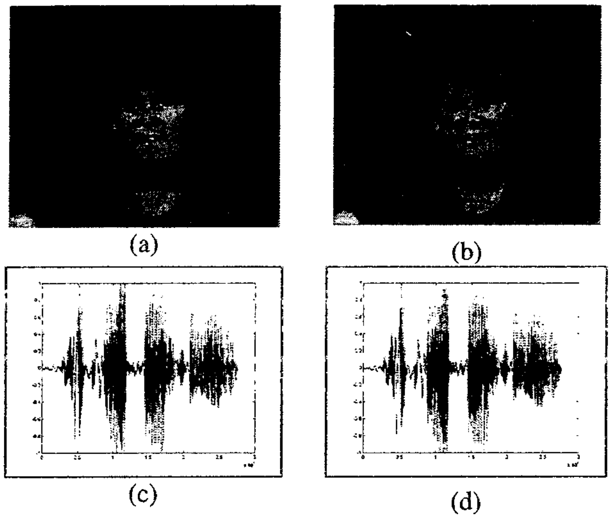
(a) (b)
(c) (d)

Figure 6: (a) The original image frame "*Miss America*" (b) its watermarked version (c) the speech waveform and (d) its extracted version without being attacked

As shown in Figure 6 shows, the quality degradation of the watermarked images could not be seen by the human eye, and its PSNR was very high, compared to most modified images processed by various signal processing (30-40 dB). It should be noted that the PSNR obtained from different image frames might be varied depending on the texture contained in such images. The next step, after the speech signal was embedded in the image frame, various types of attacks at different level of strength were applied to the watermarked images. Table 1 shows the highest strength of attacks that the contents within the extracted speech can still be recognized by all listeners.

Table 1. The highest strength of attacks that all listeners could recognize what was said in the extracted speech

Types of attack	*Resultant PSNR*	*Strength of attacks*
Brightness enhancement	10.0698 dB	80 %
Contrast enhancement	34.3949 dB	7 %
Twirling	31.5625 dB	10 degrees
Highpass filtering	11.5684 dB	10-pixel radius
JPEG compression	51.0048 dB	97 % of quality

Examples of the attacked watermarked image and the extracted speech waveform are illustrated in Figure 7.

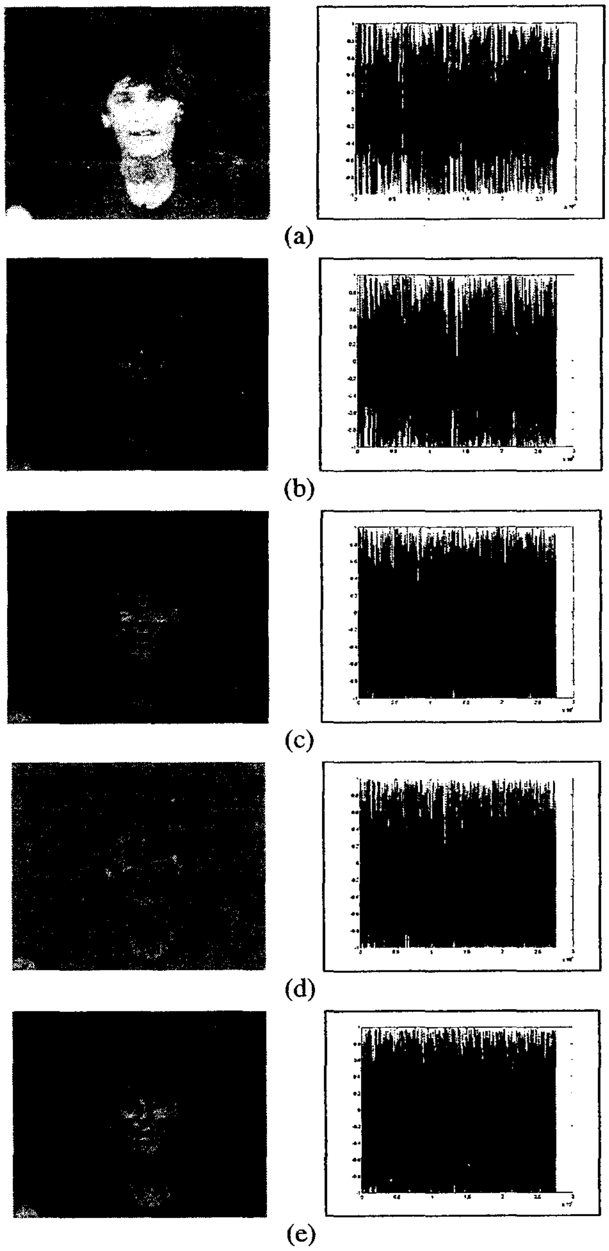

Figure 7: The watermarked image frame "*Miss America*" and the extracted speech waveform after being attacked by (a) brightness enhancement by 80% (b) contrast enhancement by 7 % (c) twirling by 10 degrees and (d) highpass filtering with radius of 10 pixels (e) JPEG compression at 97% quality

By looking at the corrupted waveform of the extracted speech in Figure 7 and the strength level of attacks shown in Table 1, it was quite impressive that all listeners could recognize the word "*Sataporn Schwindt*", even if its waveform was mostly damaged. This is because of large amount of redundancy contained in the raw speech.

6. Conclusions

In this paper, the speech based digital watermarking using zerotree of DWT coefficients has been presented. The raw speech encoded by the PCM encoder was used as the watermark signal, which was embedded into a sequence of images. According to the experimental results, the watermarking scheme implementing our proposed technique obtained the impressive results in term of quality degradation and robustness against a number of attacks.

7. Acknowledgement

This research work is supported by the National Research Council of Thailand (NRCT).

8. References

[1] I. Racocevic, B. Reljin and I. Reljin, "A method for providing digital image authenticity", *Proceedings of the 4th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services*, Vol. 1, 1999, pp. 173-176.

[2] M. N. Xia, M. L. Zhe and H. S. Sheng, "Digital watermarking based on zerotrees of DCT coefficients", *Proceedings of ICIP*, Vol. 2, pp. 237-240, 1996.

[3] H. Inoue, A. Miyazaki, A. Yamamoto and T. Katsura, "A digital watermark based on image compression", *Proceedings of International Conference on Image Processing ICIP98*, Vol. 2 pp. 391-395, 1998.

[4] 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 Transactions on Circuits and Systems for Video Technology*, Vol. 10, no. 4, June 2000.

[5] G. S. Burrus, A. R. Gopinath and H. Guo, *Introduction to Wavelet and Wavelet Transform*, Houston, Rice University, 1996.

[6] M. H. Segal, "Speech intelligibility in the space shuttle mid-deck noise environment", *The Effect of Active Noise Reduction Technology*, Nov. 1999, http://ergo.human.cornell.edu/

[7] H. K. Dunn and S. D. White, "Statistical measurements on conversational speech", *Journal of Acoustic Society of America*, January 1940, pp. 278-288.