JOINT WATERMARKING OF AUDIO-VISUAL DATA

Jana Dittmann, Martin Steinebach
German National Research Center for Information Technology
Institute IPSI - Dept. Mobile Interactive Media
Darmstadt, Germany

Abstract –Both audio and video watermarking enable copyright protection with owner or customer authentication and the detection of media manipulations. The available watermarking technology concentrates on single media like audio or video. But the typical multimedia stream consists of both video and audio data.

Our goal is to provide a solution with robust and fragile aspects to guarantee authentication and integrity by using watermarks in combination with content information. To achieve this, video stream parsing capabilities have to be added to existing watermarking algorithms. We propose to extract the audio and video content, called feature, and embed the content features with a robust watermarking scheme into the audio and video data. These features of audio and video data have to be identified. Watermarking payload and feature data requirements have to be compared.

In this paper, we describe our current state of work in audio and video processing. We introduce a new approach for video content security. Our new approach is to embed watermarks both in video and audio channels of a MPEG System stream to increase security against attacks on content integrity. We discuss features of multimedia data suited for mutual integrity protection and watermarking algorithms, describe a MPEG System stream parsing system and finally show how this components work together to provide security for multimedia data.

1. INTRODUCTION

Audio and video watermarking enable copyright protection with owner or customer information and detection of media manipulations. Most watermarking approaches are related to copyright protection and currently available watermarking technology concentrates on single media like audio or video. But multimedia streams usually consists at least of video and audio data. Therefore solutions for multimedia watermarking are necessary to provide security for the complete media stream with one technology.

For audio-visual copyright protection we would need a robust watermarking scheme for both media. Existing techniques for each media could be used to embed the copyright information into audio and video data separately to have increased protection.

If we look at the authentication to ensure manipulation recognition solutions for audio-visual data are much harder to design. In general we would need fragile watermarking techniques to detect manipulations. For audio-visual data two design goals are important: first the recognition of manipulations in each media type (time

independent) and second the recognition of manipulations in the synchronization of both streams (protection of the sequences, time dependent).

In comparison to audio-visual data copyright protection, where we could use existing robust techniques for single protection, it is not sufficient to watermark each part of the multimedia stream with a fragile video and audio watermark, because we would not protect the relation of both streams to each other. For example to protect each video frame we could use existing approaches for single images, like the approaches from Fridrich et al [K] or Delp et al. [I]. To protect the video sequence we could use the approach of Mobasseri et al [J] about "Content-video authentication by self-watermarking in color space" to address alterations of sequence of events in the video. With a redesign we can protect frame and sequence manipulations but still we are not able with both techniques to protect the complete audio-visual data . We also need fragile watermarks for audio. A possible audio attack against the integrity of an audio recording would be the removal of words, so that a sentence like "I am not guilty" could be changed into "I am guilty". With the design of a fragile audio and video frame and sequence watermark it is additionally necessary to bind the audio and video stream together. Otherwise an attacker could substitute an existing audio stream with another audio stream (same applies for video). Therefore a synchronized combination of audio and video watermarking referring to each other is needed.

We can summarize the following design goals for audio-visual authentication watermarks:

- the watermark has to recognize manipulations in each audio and video frame
- the watermark has to recognize manipulations of alterations of audio and video sequences
- the watermark has to recognize manipulations or replacements of corresponding audio and video sequences.

Another problem of the design is the necessary robustness against allowed manipulations not changing then content likelike format conversion, time re-synchronizations due to network relays or scaling and compression. One possible design is called content fragile watermarking to be robust against content-preserving operations and to detect real content changes [Jana-Martin-SAPIE2000].

In the following chapter we introduce a solution with robust and fragile aspects to guarantee authentication and integrity by using watermarks in combination with content information. We introduce a linked structure with mutual information about the content of the linked parts. Only then it will be possible to achieve a way to prove the authentication and integrity of multimedia, the combination of video and audio. We show two solutions for the protection of audio and video data with a combined robust and fragile watermarking approach: a seed-based and a direct embedding approach.

To achieve robustness against content preserving manipulations and fragileness against content changing manipulations, we need feature descriptions of the protected media. In section two we will introduce audio features for this propose, video features have already been discussed in [B].

2. Combined Video and Audio Watermarking for Manipulation Recognition

Our goal is to provide a solution with robust and fragile aspects to guarantee authenticity and integrity by using secret key watermarks in combination with content information. We combine audio and video watermarking, with the media referring to each other, building a linked structure with mutual information about the content of the linked parts and embed the watermarks with a user key to ensure authenticity. The combination of feature extraction and robust watermarking has been introduced by us as "content-fragile watermarking" in [A].
Multimedia data consist of at least an audio and a video stream. Taking MPEG as a typical multimedia example, both are combined into a system stream. This stream provides audio and video data and additional information for synchronization. To use both media types as carrier signals for the watermarking information, one either needs a watermarking algorithm capable of working on a system stream or the media types have to be marked individually. As there are numerous watermarking algorithms for audio and video data and none for combined data, we propose the following sequence:

1. Split the system stream – Separate audio and video streams and preserve the synchronization information for later use
2. Extract content fragile features of the audio and the video stream
3. Link both media features to provide mutual protection
4. Individually reduce the content-describing bitrate to the payload of the applied watermarking algorithm
5. Embed the reduced content description into both media
6. Merge the marked media streams into a system stream using the synchronization information

Watermarking algorithms not changing the bitrate of the carrier signal are preferred as the system streams are easier to handle. The new system stream then can be created by replacing its media contents, e.g. the audio frames. If this is not possible, a new system stream must be completely recreated.

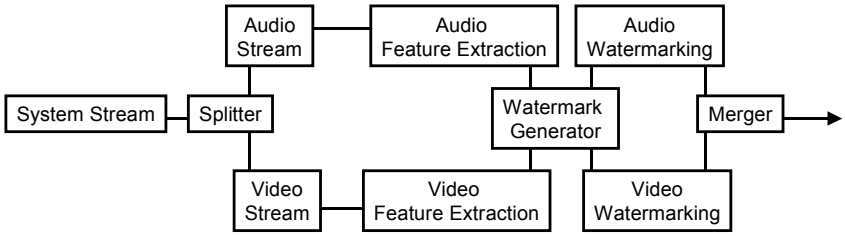

Figure 1: Mutual content fragile system stream watermarking

2.1 Concept: Embedding Content Information in Multimedia Data

In this section we discuss two different approaches to content embedding: Direct embedding and seed based embedding. With the first approach, a complete feature based content description in embedded in the cover signal. The second approach uses the content description to generate information packages of smaller size based on the extracted features.

2.1.1 Direct Embedding

In direct embedding the extracted features are embedded bit by bit into the corresponding media data. The feature description has to be coded as a bit vector to be embedded in this way. The methods of embedding differ for every watermarking algorithm. What they have in common is that the feature bit vector is the embedded watermarking information.
The problem with direct embedding is the payload of the watermarking technology: To embed a complete and sufficiently exact content description, very high bitrates would be necessary most watermarking algorithms can not provide.

2.1.2 Seed-based approach

Features are used to achieve robustness against allowed media manipulations while still being able to detect content manipulations. The amount of data for the describing features is much less then the described media. But usually even this reduced data can not be embedded into the media as a watermark. The maximum payload of today's watermarking algorithms is still too small. Therefore to embed some content description, we have to use summaries of features or very global features – like the RMS of one second of audio. This leads to security problems: As we only have information about a complete second, parts smaller than a second could be changed or removed without being noticed.
A possible solution is to use a seed-based approach. Here we use the extracted features as an addition to the embedding key. The embedding process of the watermark now depends on the secret key and the extracted features.
The idea is that only if the features have not been changed, the watermark can be extracted correctly. If the features are changed, the retrieval process can not be initialized to read the watermark.

2.2 Audio Features

To produce a robust description of sound data, we need to examine which features of sound data can be extracted and described. Some ideas about the description of audio data by features can be found in [E], [F]. We worked with Root Mean Squares (RMS), Zero Crossing Rate (ZCR) and the spectrum of the data.

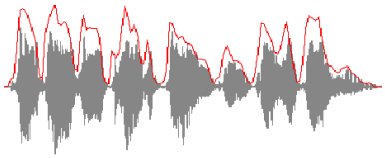

Figure 2: RMS curve of a speech sample

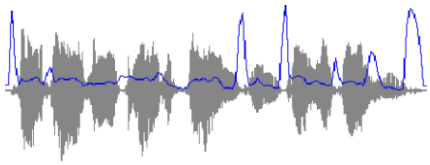

Figure 3: : ZCR curve of a speech sample

RMS provides information about the energy of a number of samples of an audio file. It can be interpreted as loudness. If we can embed RMS information in a file and compare it after some attack, we can recognize muted parts or changes in the sequence.

ZCR provides information about the amount of high frequencies in a window of sound data. It is calculated by counting the time the sign of the samples changes. The brightness of the sound data is described by it. Parts with small volume often have a high ZCR as they consist of noise or are similar to it.

Combined with RMS, ZCR can be used to decide if audio material consists of speech or music.

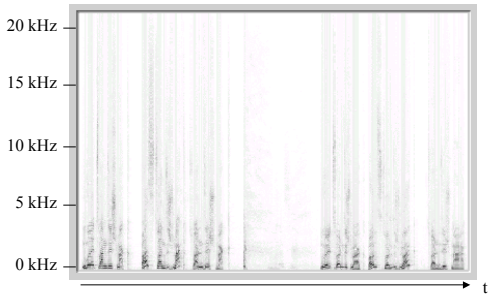

Figure 4: Spectrum of 8 seconds of speech

Transformation from time domain to frequency domain provides the spectrum information of audio data. Pitch information can be retrieved from the spectrum.
The amount of spectral information data is similar to the original sample data. Therefore concepts for data reduction a necessary, like combining frequencies into sub bands or using quantization.

To protect the semantic integrity of audio data, only a part of the full range of the spectrum is necessary. We have chosen a range similar to the frequency band transmitted with analogue telephones, from 500 Hz to 4000 Hz. Thereby all information to detect changes in the content of spoken language is kept while other frequencies are ignored.

2.3 Embedding strategy

There are numerous algorithms for video and audio watermarking, a number of them are described in [C], [H], . Most of them are designed as copyright protection mechanisms. This usually means the robustness, security and transparency are most important, while payload and complexity come second. In the case of content fragile watermarking the importance of the parameters has changed: Payload and security are most important.
A high payload is necessary to embed sufficient data. Security is important as the whole idea of content fragile watermarking is about providing security, and a weak watermarking algorithm security would mean a weak overall systems as embedded information could be forged. Security can be further increased by using cryptography while embedding the data, a asymmetric system could be used the ensure the authenticity of the embedded content descriptions.
Robustness is not as important as security. If due to media manipulations a certain loss of quality is reached and the content is changed or not recognizable any more, the watermark can be destroyed. Transparency is usually less important also, as content protected by this scheme is usually not to be used for entertainment with high end quality demands. Complexity can become relevant if the system is to work in real time, which is the case if it is applied directly into recording equipment like cameras.

2.3.1 Protection by single references

One possible strategy to embed content information would be to insert the features of one stream into a second corresponding stream. For example audio features could be inserted in the video stream. A change of the audio content then would be detectable by comparing the actual audio features with the embedded ones. This protection scheme could be used if there is only a watermarking algorithm for one of the media types providing a high enough payload for the features to embed. But in this case the third design goal can not be achieved: The watermark can not recognize manipulations or replacements of corresponding audio and video sequences in both directions but only in one direction. In our example, by using the information embedded in the video stream changes in the audio stream can be detected. But the audio stream does not include any information about the video stream.

2.3.2 Protection by mutual references

The best security can be achieved by mutual references. Here the watermarking information embedded in one media stream includes references to each other media stream. Thereby most types of attacks can be detected. Manipulations in each marked stream are recognized by the embedded watermark of the stream and the watermark in the other corresponding stream. When parts of the streams are replaced, the correspondences of the watermarks between the streams are destroyed: The streams cannot detect changes in their own content, but the stream they refer to has been changed and a manipulation is detected.
So by mutual references all three design goals given in section 1 can be achieved.

3 CONCLUSIONS AND FURTHER WORK

With content-fragile watermarking using features and mutual references we can achieve all design goals given in section 1:

- the watermark can recognize manipulations in each audio and video frame
- the watermark can recognize manipulations of alterations of audio and video sequences
- the watermark can recognize manipulations or replacements of corresponding audio and video sequences.

Therefore all types of media manipulations can be detected and integrity of the protected audio/video stream can be ensured.
The complete paper will include a short summary and conclusion and a perspective for future work. We will briefly discuss different attacks on the protection scheme introduced in this paper.

REFERENCES

[A] J. Dittmann, M. Steinebach, I. Rimac, S. Fischer,R. Steinmetz: Combined video and audio watermarking: Embedding content information in multimedia data, in , In Proc. of the SPIE Conference on Electronic Imaging '99, Security and Watermarking of Multimedia Contents II, 24-26 January 2000, San Jose USA, Proceedings of SPIE Vol. 3971, pp. 176-185, 2000

[B] J. Dittmann, "Content-fragile Watermarking for Image Authentication", to appear in Proceedings of SPIE: Security and Watermarking of Multimedia Contents III, 21 - 26 January, San Jose, California, USA, Vol. 4314, 2001

[C] J. Dittmann, M. Steinebach, R. Steinmetz, Digital Watermarking for MPEG Audio Layer 2, Proceedings of ACM Multimedia'99

[D] S. Pfeiffer. Information Retrieval aus digitalisierten Audiospuren von Filmen, Shaker Verlag, Aachen, 1999.

[E] G. Eska. Schall & Klang, Wie und was wir hören, Birkhäuser Verlag, Basel, 1997

[F] Z. Liu, J. Huang, Y. Wang, T. Chen: Audio feature extraction & analysis for scene classification, in IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing

[G] E. Lin, E. Delp: A review of fragile image watermarks, in J. Dittmann, K. Nahrstedt, P. Wohlmacher (Eds.), Multimedia and Security, Workshop at ACM Multimedia'99, Orlando, Florida, USA, Oct. 30 – Nov 5 1999, pp. 35 - 40, 1999

[H] J. Dittmann, *Digitale Wasserzeichen*, Springer Verlag, ISBN 3 - 540 - 66661 - 3, 2000

[I] A.M. Eskicioglu, E.J. Delp, „Overview of multimedia content protection in cosumer electronics devices", In Proc. of the SPIE Conference on Electronic Imaging '00, Security and Watermarking of Multimedia Contents II, 24-26 January 2000, San Jose USA, Proceedings of SPIE Vol. 3971, pp. 246-263, 2000

[J] B.G. Mobasseri, A. Evans, “Content-depedent video authentication by self-watermarking in color space”, to appear in Electronic Imaging 2001, Proceedings of SPIE Vol. 4314

[K] J. Frierich, M. Goljan, R. Du, “Invertible authentication”, to appear in Electronic Imaging 2001, Proceedings of SPIE Vol. 4314