Chapter 2

Background: Video Fingerprinting

In this chapter the basic idea of fingerprinting will be stated and the important aspects of a fingerprinting system will be discussed. The main idea of a few existing video fingerprinting techniques will then be summarized. At the end of the chapter the most important uses for video fingerprinting, namely copyright management, advertisement tracking (broadcast monitoring) and media tagging will be mentioned.

2.1 Basic Idea

The basic idea or use of video fingerprinting is to create fingerprints for a set of videos, based on the content of those videos, and saving it to a database. If an unknown video is then seen, it can also be fingerprinted and matched to the database to determine the unknown video's name, description and other details. To create content-based fingerprints, the algorithm that is used to create the fingerprints, look at certain characteristics or structures in the video to derive a fingerprint.

2.2 Uses for Video Fingerprinting

As video fingerprinting is still a rather new technology, its full potential probably hasn't been realised yet, but the two main uses are copyright management and advertisement tracking. Later video fingerprinting may be used to manage personal multimedia galleries or to identify videos that are not named correctly on user's computers.

2.2.1 Copyright management

Video fingerprinting can be used for copyright management in large internet companies or in a peer-2-peer situation. Companies can keep track of videos by monitoring it automatically over the internet with programs that use video fingerprinting.

One of the first large scale implementations was done by a very well known internet company, named YouTube, to prevent their users from uploading copyrighted videos, thus protecting them from law suits [7].

2.2.2 Advertisement tracking

Advertisements are broadcasted on television channels every day in between other programmes. Companies pay a lot of money to place their advertisements on certain channels and in certain time slots. Video fingerprinting can be used to automatically monitor television stations to determine whether the correct advertisements were played at the correct times. This way companies can confirm that their advertisements were broadcast, and if it wasn't, follow the needed steps to solve the problem.

2.2.3 Media tagging

Video fingerprinting can be used as a plug-in or stand-alone program to create tags and get more information on video files even if the file name is incorrect or there is some

distortion present. This will make it much easier for users to rename and sort their video files on their computer.

2.3 Video Background

A video is a set of still images, called frames, that follow each other in sequential order and represents scenes in motion. Digital videos are usually compressed and saved as a certain file type (wrapper). The basic properties of videos along with the use of codecs and wrappers are shortly discussed in this section.

2.3.1 Basic properties

- **Frames per Second:** This is the number of frames that should be displayed per second, when playing the video.
- **Resolution:** This refers to the number of pixels of a video frame and is shown in the format: width $\times$ height.
- **Colour depth:** The number of bits used to save the colour data for each bit. Normally 24 bits are used for videos using the Red Green Blue (RGB) colour space and 8 bits for gray scale.
- **Bit rate:** This is the number of bits for a second of video data. The compression used in the video greatly influences this value.
- **Duration:** The length of the video in time.
- **Video size:** This is size of the video in memory, for example 700 Mb. This value is a function of the bit rate and duration.

2.3.2 Wrappers and codecs

Video can be saved in a variety of different wrappers, sometimes referred to as file types. A wrapper or container format describes how the information of the video and meta-data is stored and the specific wrapper used for a video is indicated by the file type of the video, for example .avi, .mp4, .flv, .mpg, .mov, .wmv, .mkv, .ogg, etc. A wrapper contains all the information needed to present a video, including the video and audio streams, meta-data, subtitles and so on.

A codec (compressor/decompressor), on the other hand, determines how the video information is to be compressed and decompressed. The video stream is run through the compression algorithm to create the video file using less memory and the decompression algorithm is run when the video is played back. Codecs can be lossy or lossless. Lossless codecs keep all the original video data, while lossy algorithms sacrifice some of the original data to create files that are much smaller than the lossless files. A few examples of codecs are: XviD, DivX, H.264, H.263, MP4, MPEG-1 and FLV.

Videos can make use of different wrapper and codec combinations, but the most important aspect that influences video quality and size is the codec that is used.

2.4 Important Aspects of Video Fingerprinting

There are certain requirements for a video fingerprinting system to be a good fingerprinting system. These requirements include: robustness, accuracy, speed and efficiency [8]. These aspects will be discussed below. Normally a trade-off has to be done between the aspects until a proper balance is found for the specific application.

2.4.1 Robustness

Robustness is the ability of the system to detect videos even when distortion is present. Distortions in **frames** include uniform and non-uniform scaling, rotation, blur, noise, compression, artefacts, missing parts and so on. These distortions can occur simultaneously and the level of image distortion that a fingerprinting system should be able to handle depends on the application.

There can also be distortion in **videos**, like increases or decreases in the frames per second and missing frames. The way these distortions affect the system, depend on the use of frame or sequence fingerprinting, that will be discussed later in this chapter. If the video is fingerprinted using single frames, the video distortions won't affect it that much, but if a section of a video is used to create a fingerprint, these distortions may play a bigger part.

The features that are used to characterize the video and create the fingerprints are very important when it comes to the level of distortion a fingerprinting system can handle. If the features are chosen and used wisely, the system will be able to handle it very effectively.

2.4.2 Accuracy

Accuracy is the ability of the fingerprinting algorithm to generate unique fingerprints for unique videos to accurately detect them. Thus, if the content of the videos are totally different, the algorithm should generate totally different fingerprints, but if the videos are similar, the algorithm should generate fingerprints that are similar. The level of distortion that a system should be able to handle affects the accuracy of the system. A higher robustness to distortions means a lower accuracy, because more variations of the video give almost the same fingerprint.

If the accuracy of a system is poor a lot of false positives may be detected, but if it is too high, on the other hand, it may allow a lot of false negatives (missed detections). Thus,

a good balance has to be found in the accuracy of the fingerprints created for the videos.

2.4.3 Speed

This is the ability of a system to fingerprint and detect videos quickly. No one wants to wait long periods for a computer to finish processing and therefore the program must execute as fast as possible. This means that the faster the system can detect a video, the better, but a fast system has less time to do the processing. Normally there will be a trade-off between the speed and the robustness of a system, because a system can normally be more robust if it uses an algorithm that can process the frame or video thoroughly.

The speed is also affected by the time it takes to save fingerprints to the database or to match an unknown fingerprint to the database. The speed of a system depends heavily on the fingerprinting algorithm and the structure that is used to save the fingerprints. The algorithm influences the fingerprinting time and the structure influences the database saving and matching times.

2.4.4 Efficiency

This aspect has to do with the data management in the system. The efficiency of the system is measured based on how much memory space a fingerprint uses in comparison to the video and how the fingerprints are saved and retrieved from the database.

Efficient systems don't use unnecessary amounts of space to save fingerprints and the structures are well designed to improve database search speeds.

2.5 Existing Algorithms and Techniques

There have been a lot of attempts to implement video fingerprinting over the last few years. The techniques can be divided into two groups, namely those techniques that use

single frames to match one video to another and the techniques that use the whole video or sections of a video to create a fingerprint for the video. These two types of video fingerprinting techniques are compared at the end of the section.

Research papers explaining some of the techniques include Radial Projection of Key Frames [9], Centroids of Gradient Orientations [10] [11], Visual Attention Areas [12], Average Luminance over Time [8], Colour Histograms [13], Visual Digest of Local Fingerprints [14] and Spatio-temporal Sequence Matching [15]. There are also a few frame or image fingerprinting systems that look promising for use in video fingerprinting, like the one created by Sun et al. in [16], using Linear Local Embedding (LLE).

Most of the existing techniques focus on the robustness and accuracy of the video fingerprinting, but seem to neglect the speed of the video fingerprint matches. This research aims to expand what has been done in the field by developing a new technique that focuses on speed, while maintaining robustness, accuracy and efficiency.

A few existing video fingerprinting algorithms are shortly discussed below, with positive and negative traits mentioned. There has been a lot of attempts to create video fingerprinting algorithms, so there may be many algorithms that are not mentioned here. These algorithms are only mentioned to give a quick idea of what people have tried and the diversity of the methods used to try and solve this video fingerprinting problem.

2.5.1 Radial Projection of Key Frames

Cédric De Roover wrote a paper in 2005 explaining a new frame fingerprinting algorithm based on the RASH (**R**adial h**ASH**ing) algorithm [9]. RASH is said to be robust against common image distortions such as rotation, scaling, blurring and compression.

The algorithm is executed in two steps: First, radial projection of image pixels are used to calculate a vector called the RAV (**RA**dial **V**ariance) vector. This RAV vector is calculated by determining the variance of pixel intensity values of pixels that lie near a line that goes through the center of the image. 180 lines are used per vector, meaning the

consecutive lines and 1° apart. The next step is to create the fingerprint by taking the 40 lowest Discreet Cosine Transform (DCT) coefficients of the RAV vector and quantising them by using 8 bits per DCT coefficient value, thus resulting in a final fingerprint size of 320 bits.

In the paper, the author goes on to show the robustness of the novel technique, and then uses key frames to fingerprint videos. The technique is very robust and interesting, and they mention that is seems to work in real time (doesn't specify exact matching times), but their technique uses cross-correlation to determine the distance between frame fingerprints and may result in long search times if the database is large. This said, the technique seems to be one of the best currently existing techniques.

2.5.2 Centroids of Gradient Orientations (CGO)

The CGO method was introduced by Lee et al. in [10] and [11]. The technique proposed in these papers sample a video's frame at 10 frames per second and then it normalizes the frames before calculating the fingerprints. To fingerprint a frame, it is divided into a fixed number of blocks and the CGO is calculated for each block. The vectors that were calculated for each block are the frame's fingerprint. To fingerprint a frame sequence (video), a number of consecutive frame's fingerprints are combined. To match a video fingerprint to the database, the Euclidean distance is calculated between the frame fingerprint sequence.

CGO is quite robust to common video distortions. The speed of detection and fingerprinting is never mentioned, but it can be assumed that search times will become longer as the database increase, as the technique does a range search with a single frame's fingerprint and then has to match the fingerprint sequence at the promising locations. The technique also fingerprints every frame extracted (10 frames per second) and to match videos, fingerprint sequences have to match, thus the sequences have to begin at the same point in time, otherwise fingerprint mismatches will occur.

2.5.3 Visual Attention Areas (VAA)

In [12], Xing Su brings forth a novel technique to fingerprint videos using VAA proposed by Itti et al. in [17]. This technique uses color, intensity and orientation maps of a frame to create a saliency map. The saliency map is then divided into blocks and a saliency value is calculated for each block and the values are then used to create a frame fingerprint.

The VAA technique has a lot of similarities with CGO, in the sense that it also samples the video at low frames per second and then extracts fingerprints by splitting the frame into blocks. A video fingerprint is then created by creating a sequence of frame fingerprints. In Su's paper VAA is compared with CGO and it is shown to have a much lower false rejection rate.

VAA is also quite robust to common video distortion, but may be slow due to the searching characteristic to match fingerprints. VAA implements a link-list to increase fingerprint matching times. To solve the problem of fingerprint sequence start time mismatch, VAA uses a sliding window in the unknown video.

2.5.4 Average Luminance over Time

The luminance histogram method of fingerprinting videos was proposed by Pereira et al. in [8]. This method uses the average luminance of the images in a video to create a luminance histogram, that is matched with the database to find a match.

The average luminance of each frame is calculated. The luminance level is then saved in a database with a begin time. While the luminance stays within a certain threshold, no new information is added. Once the luminance goes outside the threshold a new entry is added to the database. To get a match the luminance-time sequence of the different videos have to be correlated.

This technique has a very fast database search time, because the fingerprint information is very little compared to the video data. It is also robust to color changes as it only

makes use of the luminance.

A certain length of video sequence is needed to ensure pairwise independence and the technique is sensitive to luminance changes, because that is the characteristic it uses to fingerprint the videos. Cropping also influence the fingerprint matching as it influences the average luminance.

2.5.5 Colour Histograms

A technique is proposed in [13] that fingerprints videos by extracting a Group of Frames (GoF) from the video. The GoF model share a lot of features with MPEG-7 segment structures. A cumulative histogram is then calculated for the GoF. This histogram is an alpha-trimmed average histogram which serves as a fingerprint for the GoF.

The technique does a lot of processing, as it has to calculate alpha values for the GoF and compare histograms within the GoF to calculate the alpha-trimmed average histogram. The database search also consists of comparing the query histogram to the database by calculating the distance between histograms, which can take a considerate amount of time, but the exact search times aren't mentioned.

2.5.6 Linear Local Embedding (LLE)

LLE is a non-linear dimensionality reduction technique that was introduced by Roweis et al. in [18]. In [16], Rui Sun and his colleagues used it to create a robust image hash to identify images. This is therefore not a video fingerprinting algorithm, but an image fingerprinting algorithm. Image fingerprinting algorithms are also important and can be expanded to detect videos.

To create a descriptor for an image, the image is normalized to a fixed size, converted to gray scale and filtered to remove high frequency noise. A certain number of patches are then extracted from the image, pseudo-randomly. The $M \times M$ image patches are then

converted to $M^2 \times 1$ vectors and concatenated in a sequence to form a $M^2 \times p$ matrix, where p is the number of image patches extracted. LLE is applied to the matrix and the LLE weights are calculated. Each cell in the weight matrix is then converted to a binary value, depending on whether the weight is bigger or smaller than the weight in the previous column. The binary matrix is then converted to a binary string that is the robust image hash.

The robust image hash functions as the fingerprint for an image and is used to compare the similarity of images by using the normalized Hamming distance. The results in the paper show that the technique is very robust to common image distortions and can be very promising if it is used in a video fingerprinting technique.

2.5.7 Frame vs. Sequence fingerprinting

One of the main aspects of any video fingerprinting system is whether it makes use of frame or sequence fingerprinting. In frame fingerprinting individual frames from videos are fingerprinted and the fingerprints matched to a database of frame fingerprints to get video matches. Specific frames can be extracted from the video using a Key Frame Detector (KFD) that can be used for the fingerprinting process. Sequence fingerprinting makes use of a section of a video, or the whole video to create a fingerprint and match it to the fingerprints in the database.

The different techniques have different strengths and weaknesses, although it depends a lot on the specific video fingerprinting technique being used, because the problems may be addressed by the system:

- **Minimum video length:** For frame fingerprinting only a few frames may be needed to detect the video, while sequence fingerprinting needs a whole video or a section of the video to detect it.
- **Fingerprint size:** Sequence fingerprinting may use less space to create a fingerprint for a video, as it summarizes the whole video, but this is very subjective and depends

a lot on the specific fingerprinting technique used. Frame fingerprinting techniques can reduce the video fingerprint size a lot by using a KFD, as mentioned in Section 3.4.

- **Speed:** Frame fingerprinting techniques can process each frame as it comes in, while a sequence fingerprinting technique has to process the unknown video sequence as a whole. This means frame fingerprinting can detect the video playing before it is finished.
- **Embedded videos:** Frame fingerprinting can easily detect videos that are embedded within another stream, as it handles every frame separately. For sequence fingerprinting an algorithm will have to be created or the fingerprinting system designed in such a way that it can detect videos that are embedded in a video.

For advertisement tracking the frame fingerprinting technique would work better, because the advertisements are embedded in the TV broadcast stream and need to be detected while the broadcast is being played in real-time.