Image and Video Fingerprinting: Forensic Applications

Frédéric Lefèbvre, Bertrand Chupeau, Ayoub Massoudi and Eric Diehl
Thomson R&D France, 1 avenue Belle Fontaine, 35510 Rennes
{first_name.last_name}@thomson.net

ABSTRACT

Fighting movie piracy often requires automatic content identification. The most common technique to achieve this uses watermarking, but not all copyrighted content is watermarked. Video fingerprinting is an efficient alternative solution to identify content, to manage multimedia files in UGC sites or P2P networks and to register pirated copies with master content. When registering by matching copy fingerprints with master ones, a model of distortion can be estimated. In case of in-theater piracy, the model of geometric distortion allows the estimation of the capture location. A step even further is to determine, from passive image analysis only, whether different pirated versions were captured with the same camcorder. In this paper we present three such fingerprinting-based forensic applications: UGC filtering, estimation of capture location and source identification.

Keywords: fingerprint, perceptual hash, forensics, filtering, content identification

1. INTRODUCTION

According to [26], digital forensics is the use of analytical and investigative techniques to identify, collect, examine and preserve evidence/information which is magnetically stored or encoded, usually to provide digital evidence of a specific or general activity. The first question to answer when analyzing a suspect media file is to determine whether it is a copy of a copyrighted title, and if so which title it is. In such a copy identification context, watermarking is commonly used but not all copyrighted content is watermarked. A crude method compares two different bit streams using bit-to-bit distance. Whenever one bit changes, the whole hash code changes. Due to voluntary or natural content manipulations, such as analog to digital conversion (camcorder capture), compression, frame removing or adding (advertising), this strategy based on bit-to-bit or hash code comparison is doomed to fail. A multimedia identification/authentication process has to withstand natural distortions. Fingerprinting solves this problem, by extracting discriminating features, called fingerprints or multimedia DNA, representative and unique for each multimedia content. Section 2 reviews existing works in image and video fingerprinting. And section 3.1 illustrates content identification with fingerprinting for User Generated Content (UGC) filtering application.

Forensic tracking aims at identifying the traitor (the dishonest user) when an illegal copy is found, to put pressure on him, and possibly to sue him in court. Watermarking techniques handle that purpose: identification marks are embedded. They either identify the customer ID (in pay-per-view or video-on-demand application, or on DVD screeners) or the theater [23]. Sometimes pirated videos exhibit strong signal distortions that hamper the decoding of forensic marks. Thus, registration with original content is a mandatory pre-processing step to subsequent forensic analysis. And ahead tracking, pirate profiling is a broad application task which comprises the extraction of any other useful information on the pirate habits and trends: for example, analysis of geometric distortions can help deriving the recording location in a theater. Recording location estimation in a theater is tackled in section 3.2.

Source identification (or sensor forensics) aims at identifying the acquisition device that captured an image (e.g., digital camera, cell-phone or scanner) [21]. This means associating the image with a class of devices with common characteristics (i.e., device brand and model) and eventually matching the image to an individual device. Main clues for image source identification are sensor noise and artifacts in CCD array, lens distortions, demosaicing artifacts or sensor dust. This work has recently been extended to video, to determine whether two cams came from the same camcorder or not [3]. Section 3.3 details sensor forensics.

2. IMAGE AND VIDEO FINGERPRINTING

Image fingerprinting extracts discriminating features, called multimedia DNA [2] or fingerprints, specific to each image. Several techniques exist to find these unique features. Some of them are semantic (high level analysis) and other ones are more signal based (low level analysis) (Figure 1).

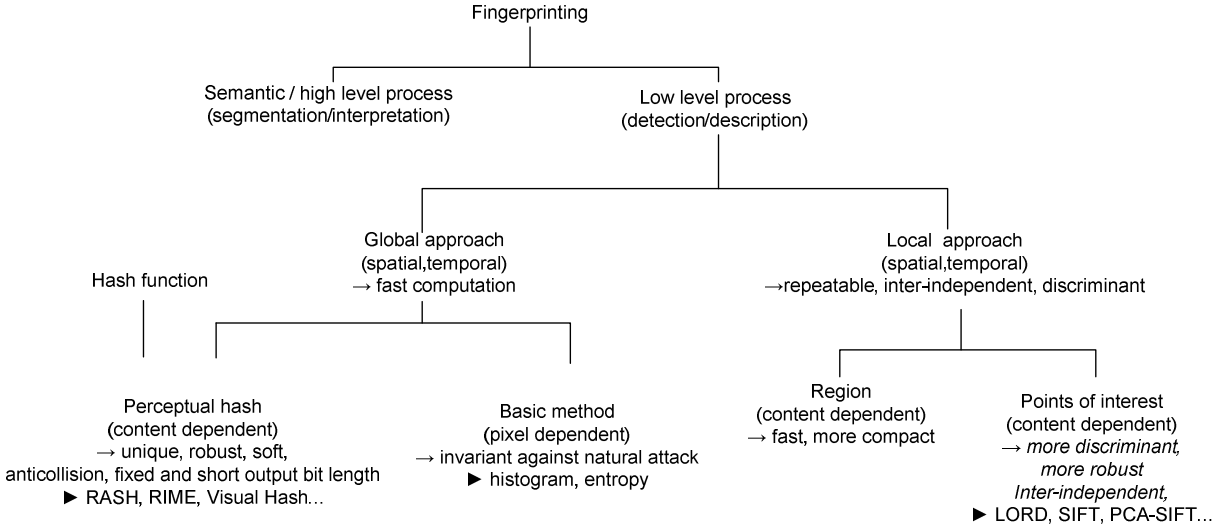

Figure 1: Classification of fingerprinting techniques

In this paper we focus on signal-processing based "low level" processes. Low level processes are divided in two categories: global description and local description. The global approach considers an image as a global content such color/luminance histogram, texture based. The local approach considers an image as a multitude of spatially localized characteristics [16]. The global approach is faster but suffers from collision and lack of robustness against strong attacks such as cropping, occlusion, and addition. The local approach is more resistant against strong attacks and provides interesting properties to identify the geometrical distortion. The extracted features are repeatable, inter-dependent and more discriminant than global features. Among local approaches, the techniques based on points of interest solve the spatial synchronization problem, as a model of distortion can be estimated from a mapping of such points in original and copy images. Local image fingerprints are mainly used in object recognition [16] and biometry. In a biometric context, the "finger print" highlights local key points, called minutiae [4]. Ross *et al* [20], however, show that it is possible to reconstruct fingerprints from minutiae points. They demonstrate that the minutiae template of a given user can be used to synthesize a fingerprint. Security and output bit length are thus the main weaknesses of fingerprinting. Perceptual hash algorithms are designed to overcome these security issues.

Digital signature is largely used in cryptography to guaranty a document's authentication. But Voyatziz *et al* [24] explained that this solution is not adapted to multimedia applications due to the length of the output hash function bit stream. A hash function, used in signature generation, computes a condensed version of data – a bit stream summary –, called message digest. A perceptual hash is a hash function designed for multimedia contents. Perceptual hash functions use fingerprinting techniques with cryptosystem-like constraints [10]. The properties of a perceptual hash function are: easiness of computation, weak collision resistance, and a fixed output bit length called perceptual digest. The perceptual digest must be resistant and robust, in other words it shall remain the same before and after attacks that do not alter the perceived contents. A small content change yields a small change of the perceptual digest. A large content change leads to a large change of the perceptual digest.

Image fingerprinting is an established key technology for biometry and image indexing, but some recent works only have extended image description to video description. A video fingerprint can be a global description of the video (number of scene cuts, size of the video, etc.), a set of image fingerprints of still video frames, a set of video key frames fingerprints [18] or the description of the motion vectors. For instance in [11], the authors combine a few local

fingerprints, measurement of a global similarity and fingerprint database management. The measurement of the global similarity is done after a correct fingerprint matching. It is the first scientific work which combines efficient fingerprint and efficient search in a fingerprint database. In [8], the authors identify a candidate movie by its MPEG motion vectors. This method is well integrated in the broadcast workflow but is encoder dependent. If motion vector fields change too much due to different encoders, the movie description will change enough to make the detection process erroneous. In [19], authors propose efficient and fast video fingerprint. The video is divided into frames. Each frame is also divided into blocks. For each block of each frame, the mean luminance (gray level histogram) is computed and compared to adjacent blocks and to adjacent frames (spatial-temporal 2×2 Haar filter). In practice, the process is done for N blocks per frame, called hash words, and for M consecutive frames, called hash blocks. This visual hash for video content generates a signature of $N \times M$ bits per fragment of M frames. The full length size of this signature depends on the video size. In [18], the authors combine a visual hash function and a local fingerprinting to describe a video content. The perceptual digest of each frame captures the video content variation and detects key frames. A local image fingerprint technique characterizes the detected key frames. The set of local fingerprints for the whole video summarizes the video or fragments of the video. The algorithm proposed in [18] is more detailed in section 3.1 as a solution to identify contents over community sites such as UGC sites (YouTube, DailyMotion).

3. FORENSIC APPLICATIONS

3.1 Filtering of UGC sites

The User Generated Content (UGC) context provides an example of automatic content identification: “What is the title of this video? Am I in copyright violation if I distribute this content?” UGC sites such as YouTube or DailyMotion are becoming the platforms of choice to exchange and view videos. Unfortunately, indelicate users also post copyrighted contents and UGC sites become a new channel of distribution of illegal contents. The limitation of the size of uploaded content is not a deterrent mechanism. According to Digital Ethnography [27], 200 000 videos were uploaded per day in March 2008. The average video length was 2 minutes 46.17 seconds. It means that 384 days of contents, more than one year of contents, were uploaded every day in March 2008. Digital Ethnography estimates that 80.3% are amateur contents (unambiguously user-generated) while 14.7% are professional and 4.7% are commercial contents. As explained in [27], the percentage of video that are probably in violation of copyright is 12%. If we consider that some uploaded videos are removed immediately by YouTube, how many copyrighted contents are really uploaded every day? Of course, content owners request UGC sites to banish their contents. Thus UGC sites need methods to detect copyrighted content.

There are four main approaches. The first method is manual reviewing. The second method automatically analyses the filenames. The third method uses watermark to detect possibly copyrighted content. The fourth method, also the most efficient one, is based on fingerprints. But the system can only identify opus that it knows; therefore, a new opus has first to be registered into the system. In a content identification context, two symmetrical processes are therefore mandatory: registration (or indexing) and detection. To register a new opus, we submit the original file to the fingerprint extractor, which generates the visual hashes (or fingerprints). The visual hashes are stored in a database together with the name of the opus and optional metadata. To identify the origin of a sample, we submit the test file to the fingerprint extractor to generate its visual hash(es). The decision engine retrieves them and searches the database for the potential candidate; the response is either the name of the opus, or a failure statement (Figure 2).

The decision engine does not need the complete content: just a short fragment is sufficient to identify the content. Thus, a submitted sample may start anywhere in the opus. The authors of [18] have developed a two-step adaptive video identification process for the UGC context. The proposed algorithm was tuned to provide a good speed versus accuracy tradeoff for detection. The technique combines a visual hash function and a local fingerprinting. The first step is a global analysis of the test file. Frames that present a high degree of change are detected as shot boundaries. The most significant frame per each shot is selected. And for each selected frame, a small, fixed-length, pixel dependent descriptor is extracted. This descriptor characterizes the global picture. The sequence of these global descriptors forms the first level of content description. This global descriptor is resistant to many transformations such as rotation, compression, temporal cropping, etc. The second step is a local analysis of the test file to cope with more serious transformations. For each frame selected by the first step, the system extracts a set of salient points also called points of interest. For each point of interest, parameters describing its neighborhood are extracted. The set of descriptions forms the local descriptor

that is region dependent. The sequence of these local descriptors forms the second level of content description. Obviously the local descriptors are longer than the global ones.

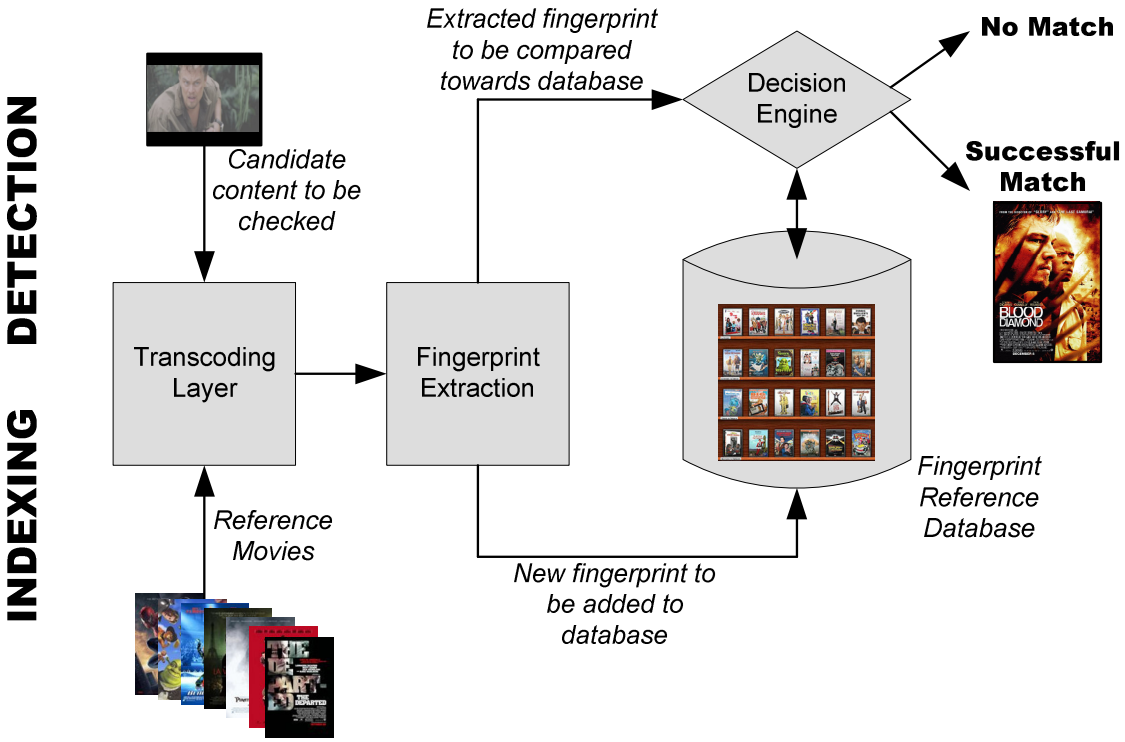

Figure 2: Indexing and detection process for UGC filtering

Content owners provide both types of above described global and local descriptors (fingerprints), of the content they want to be filtered by the UGC site. This constitutes a reference database. Database is a generic term for data organization. Conventional databases such as Oracle, MySQL are said relational. The similarity between two contents is usually based on binary distance between two elements. In case of a fingerprinting application, two contents can have two similar fingerprints with a very important binary distance. Binary distance or Hamming distance are thus not adapted to fingerprint applications. In the literature we find fingerprints of dimension 128 [16], 144 [18], 180 [14], etc. It means that the fingerprint matching application has to search an element of dimension 128, 144 or 180 in a database populated by N elements of dimension 128, 144 or 180. With the algorithm described in [18], 315 hours of master contents generate $N=140$ millions of points of interest with descriptors of dimension 144. And the database engine has to search a candidate fingerprint among all elements or master fingerprints which are not binary stable. To solve this issue, the fingerprint matching process is based on nearest neighbor search techniques [1]. It means that we do not search the index reference with the same binary representation as the candidate, but the index reference which provides the closest distance between candidate and reference fingerprints. The main challenge is to efficiently address the tradeoff between detection speed, database size and detection precision. A fingerprint architecture applied to UGC filtering performs efficiently if both the fingerprint generation and the fingerprint database perform efficiently. We cannot therefore dissociate fingerprint generation from the fingerprint database. Once the reference database is populated, detection of copyrighted content can start on UGC sites. Every uploaded content on UGC sites is fingerprinted. The decision engine starts with the global descriptors and, in case of non convergence, automatically switches to the local descriptors (Figure 3).

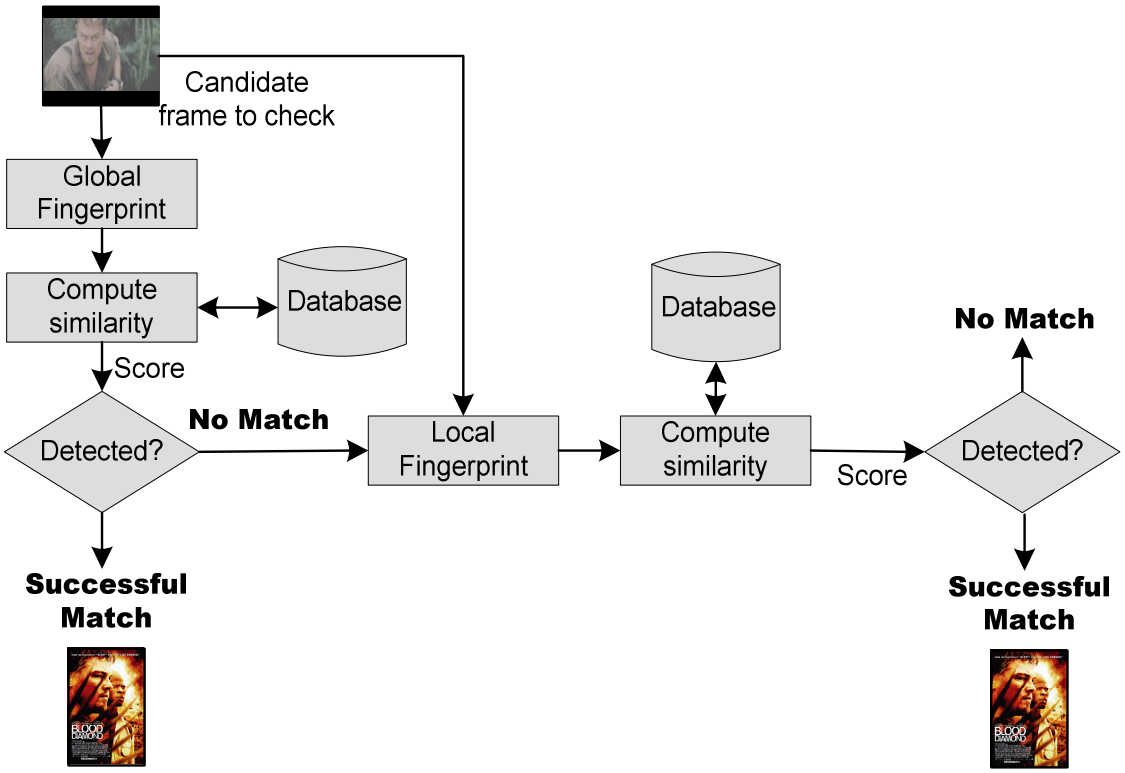

Figure 3: Detection process

This two-step strategy provides a good compromise between speed and robustness. In case of matching, there is a high probability of upload of copyrighted content. Some companies announce 99% of hit detection. But such accuracy figure does not mean anything if the whole context is not specified: duration of the original sequence, duration of the copy, manipulations (attacks) applied on the original… False positive and false negative rates increase with the strength of the attacks and the database size. To evaluate relevant detection rates for content filtering application, the following information are expected:

- **Size of the database (hours):** the larger the database, the higher the false positive and false negative rates. Database size has also usually an impact on detection speed.
- **Definition of the attack(s):** camcorder, spatial stretching, frame rate changes, transcoding, compression… The stronger the attack (camcorder), the more difficult to correctly identify a copy.
- **Duration of the candidate(s):** the shorter the candidate, the more difficult the detection and the more false negatives. We can imagine adapting the decision module depending on the candidate duration.
- **Speed of the detection:** fast detection reduces the number of required machines and allows live events filtering application.
- **Size of the fingerprints**: the longer the fingerprints, the smaller the false positive rate but the higher the database size.
- **Size of the database:** it affects fingerprint architecture. Some of them are based on RAM memory or hard disk drive.

MPAA and MovieLabs conducted such objective measurements. They took in consideration realistic attacks (camcorder, scaling, frame rate changes, DivX compression) and usage scenario (P2P: full length video and online file sharing: snippets of video).

Once a copyright content identified, the UGC site has several possible actions. The most obvious one (and most current) is filtering copyrighted content. Content is not allowed to be posted. Some UGC sites are starting some contractual and commercial agreements with some content providers. In this case, the reaction of the UGC site may be richer. It may allow the posting and will share the corresponding advertising revenues with content owner. It may even replace an identified copyrighted upload by the official version that has a guaranteed quality.

3.2 In-theater piracy forensics

Video fingerprints based on local points of interest can be used for more in-depth forensic analysis than simply identifications. For example, camcorder recording in cinemas often leads to geometrically distorted images, usually through the trapezoidal effect (also known as 'keystoning'). Wang and Farid recently proposed a method to automatically detect camcorder captures of projected movies [25]. We proposed [5] to analyze the keystoning distortions, using the derived camcorder viewing angle to estimate the approximate position in the cinema from which the movie was captured. This can be done in three steps, using the fingerprints from the pirate copy and the corresponding master. The first step identifies pirate copies in peer-to-peer networks or on UGC sites by matching digital fingerprints to the master database. Second, the cinema in which the movie was recorded illegally is tracked down by decoding a forensic cinema-identification tag. The final step involves registration and seat localization. Temporal synchronization provides pairs of reference and copy frames [7]. Spatial registration applied to pairs of points of interest from pirate and master copies allows construction of an eight-parameter homographic model [6] (Equation 1) of the keystoning effect from which the camcorder viewing direction can be determined.

$$\begin{cases} x' = \dfrac{h_{00}x + h_{01}y + h_{02}}{h_{20}x + h_{21}y + 1} \\ y' = \dfrac{h_{10}x + h_{11}y + h_{12}}{h_{20}x + h_{21}y + 1} \end{cases} \quad (1)$$

The system of equations and its resolution is detailed in [5]. Intersection of the viewing axis and a 3D model of the cinema's seating plane eventually lead to the pirate's seat (Figure 4).

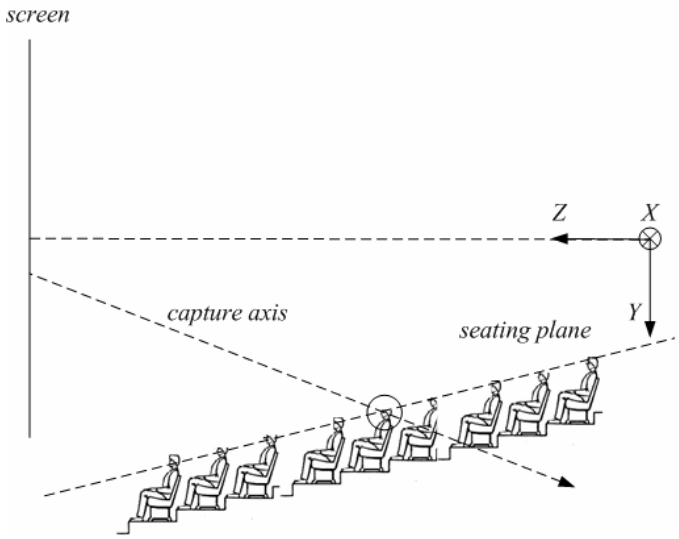

Figure 4: Capture location estimation

The accuracy of the proposed estimation method was assessed through a couple of validation experiments with ground truth data. A first experiment was performed with a camcorder located at nine predefined seats in the audience room of a real theater. A graphical representation of the viewing axis estimates and their intersection with a modeling of the seating plane is depicted below (Figure 5). The numerical analysis of the results shows an average location error ranging from 4 cm to 61 cm parallel to the screen (width error), from 3 cm to 178 cm perpendicular to the screen (depth error). This accuracy is quite acceptable, as in this theater the distance between two seats in a row is 50 cm, and the distance between two seating rows is 100 cm. A second experiment in a smaller theater yielded a similar error range: from 0 to 30 cm for width error, from 2 to 78 cm for depth error.

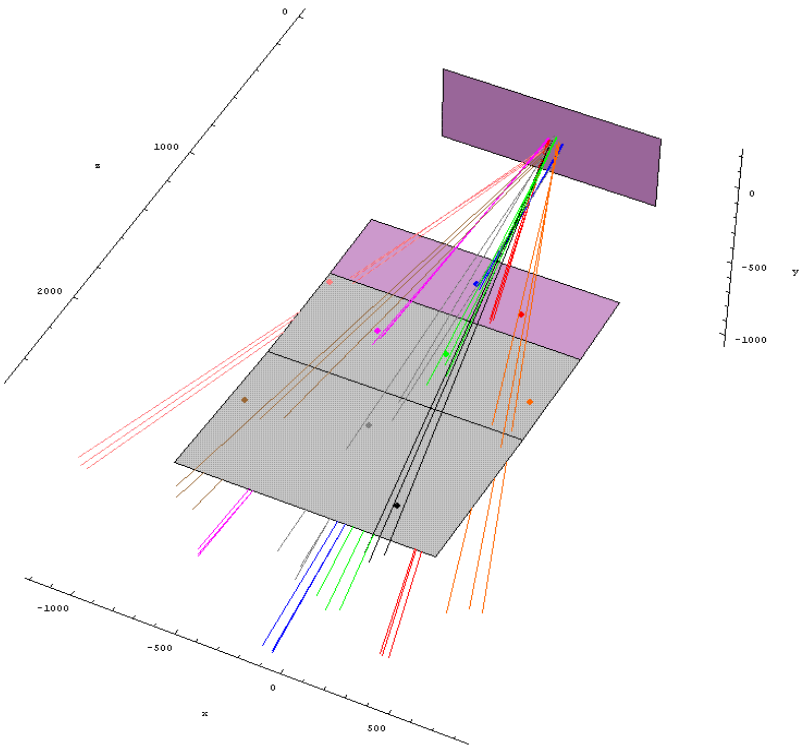

Figure 5: Capture axis estimates

The experimental results show that the approximate capture location within the theater can be extracted from the pirate copy, with an acceptable accuracy. A first result would be to automatically determine whether the copy was captured from the projection booth or from the seating area. Additionally, with regard to the measured standard deviation, it seems attainable to divide this seating area into about a dozen of zones and to automatically assign the capture position to one of them. Some of our initial simplifying assumptions may however be wrong when working with real pirate cams. Future study could thus consist in integrating in the image formation model more precise information about the theater geometry, such as the screen curvature, in order to increase the localization accuracy.

3.3 Sensor forensics

Source identification forensics looks for the device (e.g., digital camera or camcorder) that captured an image or video. Forensic methods for digital cameras may use metadata in image header, watermarking or fingerprinting. Metadata stored in image header may be the date, serial number of the camera, camera references, etc. But the metadata are not persistent in case of post-processing and file format conversions. Watermarking is another possibility but it requires embedding a watermarking algorithm inside the camera device. Only few manufacturers only are ready to include such a feature in their products. Tracing a digital camera device only based on the output patterns generated by this device was the purpose of the pioneering work by Lukas [17]. It is based on the observation that each sensor generates a specific noise, also called sensor fingerprint.

The sensor noise is divided into Fixed Pattern Noise (FPN) and Photo Response Non Uniformity noise (PRNU). The FPN is observed in dark frames and is sensitive to exposure and temperature. PRNU is present in all frames and is generated by the non uniformity of the noise due to different sensitivities of pixels to the light, optical lens characteristics and light refraction on dust particles. Optical lens and light refractions on dust particles are of low frequency and are not representative of the sensor. The authors focus their research on pixel non-uniformity which is defined as sensitivity of pixels to the light. This noise is mainly due to the heterogeneity of silicon wafers and imperfections during the manufacturing process. Much of this information is captured using a wavelet-based de-noising filter and subtracting a de-noised version of the image from the original image. Identification is done by inter-correlation between the fingerprint of the candidate sensor with the fingerprint of the reference sensor. Experiments with nine sensors of different sizes and technologies highlighted a high discriminance and robustness against classical attacks: JPEG compression, gamma correction, re-sampling and stability along the time. Methods using alternative features to noise pattern were proposed for camera model identification. For example Kharrazi feeds 34 image features into a multi-class classifier [12][13] and Swaminathan estimates the interpolation filter coefficients of the color filter array (CFA) pattern and identifies the valid CFA for the analyzed image [22].

This sensor fingerprint technology suffers however from the same security weaknesses against "copy attack" [15] than some watermark algorithms. If anybody can create a sensor fingerprint from 50 images (according to the authors) taken by a digital camera A, an adversary can copy and paste the sensor fingerprint from the camera A into a new image. A malicious adversary can also remove sensor fingerprint from an image taken by a camera A and paste a new sensor fingerprint from a camera B into this image. Manufacturers make also efforts to reduce sensor noise. When sensor noise is removed, does sensor fingerprinting still make sense?

4. CONCLUSION

Fingerprinting is a powerful tool for media forensics. Not only does it provide fast and accurate copy identification, robust to signal distortions, without the embedding constraints of watermarking, but it also enables many other forensic tasks. In this paper we first reviewed the most popular image and video fingerprinting algorithms, focusing on our own method which was initially described in [18]. Based on a two-step, global and local, description of the key-frames of a video, it provides a good compromise between detection speed and accuracy. Filtering of copyrighted contents on UGC sites is the application in the spotlight, but they are many other possible uses of fingerprinting. As an example, we showed that an image description based on points of interest and associated local features enables the modeling of the geometric distortion. In case of in-theater camcorder capture, the capture location can be retrieved from this distortion model. Other specific fingerprints keep the tracks of the image sensor and can be used to identify the capture device. By combining all those various passive, non-intrusive, image and video analysis techniques, into a comprehensive forensic analysis scheme, one can learn a lot from an anonymous video file.

We must keep in mind, however, that as digital media forensics is a new emerging field, most, if not all, proposed schemes have not so far really considered serious attackers [9]. Possible goals of such attackers could be fooling copy identification, hiding content tampering or erasing (or even forging) content origin. Taking this in consideration opens the door to a new research area and a never ending arm races with the pirates.

REFERENCES

[1] Amsaleg, L., Gros, P. and Berrani, S.-A., "Robust object recognition in images and the related database problems", Special Issue of the Journal of Multimedia Tools and Applications, 23(3), 221-235 (2004).

[2] Batle, E., Neuschmied, H., Uray, P. and Ackermann, G., "Recognition and analysis of audio for copyright protection: the RAA Project", Journal of the American Society for Information Science and Technology, 55(12), 1084-1091 (2004).

[3] Chen, M., Fridrich, J., Goljan, M. and Lukas, J., "Source digital camcorder identification using sensor photo response non-uniformity", Proc. SPIE 6505 (2007).

[4] Chikkerur, S. and Ratha, N., "Impact of singular point detection on fingerprint matching performance", Proc. Fourth IEEE Workshop on Automatic Identification Advanced Technologies, 207-212 (2005).

[5] Chupeau, B., Massoudi, A. and Lefèbvre, F, “In-theater piracy: Finding where the pirate was”, Proc. SPIE 6819 (2008).
[6] Chupeau, B., Massoudi, A. and Lefèbvre, F, “Automatic estimation and compensation of geometric distortions in video copies”, Proc. SPIE 6508 (2007).
[7] Chupeau, B., Oisel, L. and Jouet, P., “Temporal video registration for watermark detection”, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (2006).
[8] Coudray, R. and Besserer B., “Global motion estimation for MPEG-encoded streams”, Proc. IEEE Int. Conf. on Image Processing (2004).
[9] Gloe, T., Kirchner, M., Winkler, A. and Böhme, R., “Can we trust digital image forensics?”, Proc. ACM Int. Conf. on Multimedia, 78-86 (2007).
[10] Haitsma, J., Kalker, T. and Oostveen, J., “Robust audio hashing for content identification”, Proc. Content-Based Multimedia Indexing (2001).
[11] Joly, A., Frelicot, C. and Buisson, O., “Content-based video fingerprint statistical similarity search approach”, Proc. IEEE Int. Conf. on Image Processing (2005).
[12] Kharrazi, M., Sencar, H. T. and Memon, N., “Blind source camera identification”, Proc. IEEE Int. Conf. on Image Processing, 709-712 (2004).
[13] Kharrazi, M., Sencar, H. T. and Memon, N., “Blind camera identification based on CFA interpolation”, Proc. IEEE Int. Conf. on Image Processing (2005).
[14] Lefèbvre, F., “Message digests for photographic images and video contents”, Thesis report, UCL (2004).
[15] Kutter, M., Voloshynovsky, S. and Herrigel, A., “The watermark copy attack”, Proc. SPIE 3971, 371-380 (2000).
[16] Lowe, D. G., “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, 60(2), 91-110 (2004).
[17] Lukas, J., Fridrich, J. and Goljan, M., “Digital camera identification from sensor pattern noise”, IEEE Trans. on Information Forensics and Security, 1(2), 205-214 (2006).
[18] Massoudi, A., Lefebvre, F., Demarty, C.-H., Oisel, L. and Chupeau, B., “A video fingerprint based on visual digest and local fingerprints”, Proc. IEEE Int. Conf. on Image Processing, 2297-23000 (2006).
[19] Oostven, J., Kalker, T. and Haitsma, J., “Visual hashing of digital video: applications and techniques”, Proc. SPIE 4472, 121-131 (2001).
[20] Ross, A., Shah, J. and Jain, A. K., “Towards reconstructing fingerprints from minutiae points”, Proc. SPIE 5779, 68-80 (2005).
[21] Sencar, H. T. and Memon, N., “Overview of state-of-the-art in digital image forensics”, World Scientific Press (2008).
[22] Swaminathan, A., Wu, M. and Liu, K. J. R., “Nonintrusive component forensics of visual sensors using output images”, IEEE Trans. on Information Forensics and Security, 2(1), 91-106 (2007).
[23] Van Leest, A., Haitsma, J. and Kalker, T., “On digital cinema and watermarking”, Proc. SPIE 5020 (2003).
[24] Voyatziz, G. and I. Pitas, “The use of watermarks in the production of digital multimedia products”, Proc. of IEEE, Special Issue on Identification and Protection of Multimedia, 1197-1207 (1999).
[25] Wang, W. and Farid, H., “Detecting re-projected video”, Proc. Int. Workshop on Information Hiding (2008).
[26] Computer Forensics World: http://www.computerforensicsworld.com/
[27] http://ksudigg.wetpaint.com/page/YouTube+Statistics?t=anon