

Towards camcorder recording robust video fingerprinting

Adriana Garboan

▶ **To cite this version:**

Adriana Garboan. Towards camcorder recording robust video fingerprinting. Other [cs.OH]. Ecole Nationale Supérieure des Mines de Paris, 2012. English. NNT : 2012ENMP0097 . pastel-00871762

HAL Id: pastel-00871762
https://pastel.hal.science/pastel-00871762

Submitted on 10 Oct 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale 432 : SMI - Sciences des Métiers de l'Ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité "Informatique temps réel, robotique et automatique - Paris "

présentée et soutenue publiquement par

Adriana GARBOAN

13 Décembre 2012

Towards camcorder recording robust video fingerprinting

Traçage de contenu vidéo : une méthode robuste à l'enregistrement en salle de cinéma

Directeur de thèse : **Françoise PRETEUX**

Co-directeur de thèse : **Mihai Petru MITREA**

Jury

Mme Christine GRAFFIGNE, Professeur, Université Paris Descartes	Président
Mme Adriana VLAD, Professeur, Université POLITEHNICA de Bucarest	Rapporteur
M. Arnaud REICHART, Professeur, ENSTA ParisTech	Rapporteur
M. Claude DELPHA, HDR, Université ParisSud	Examinateur
Benoit MAUJEAN, Responsable R&D, MIKROS Image	Examinateur
Mme Françoise PRETEUX, Professeur, Institut Mines-Télécom; Mines ParisTech	Examinateur
M. Mihai Petru MITREA, HDR, Institut Mines-Télécom; Télécom SudParis	Examinateur

T H È S E

MINES ParisTech
Mathématiques et Systèmes, CAOR - Centre de CAO et Robotique
MINES ParisTech, 60 Boulevard Saint-Michel 75006 Paris, France

The video corpus created under the framework of the HD3D-IIO project is processed in the present research study according to the HD3D-IIO and HD3D2 intellectual property agreements.

To my husband and to my parents

Acknowledgment

Three years ago, as a young engineer in search of the great knowledge I was starting my PhD program. This work could not have been accomplished and I couldn't have learned all that I have learned without the help and collaboration of numerous people.

My deep gratitude goes to my thesis director, Professor Françoise Prêteux for the warm welcome she gave me for the first time in the ARTEMIS department at Institut Telecom, Telecom SudParis and for the second time in the CAOR department of MINES ParisTech. I would like to thank her for her trust and for seeing in me the future PhD. She taught me the lesson of hard work and precision in research activities and therefore I would like to express my appreciation for her help in broadening my technical skills, for her guidance and constant support.

Equally, my deep gratitude goes to my thesis co-director, HDR Mihai Mitrea for introducing me in the fascinating world of information theory while a student at POLITEHNICA University of Bucharest and then for granting me the chance to start the research work at ARTEMIS on the novel and exciting topic of video fingerprinting. I would like to express my profound appreciation and special thanks for his step by step guidance of the thesis, for his expertise, energy and enthusiasm that he invested in working with me, in providing me with the theoretical research tools as well as with the necessary logistic and administrative resources. I would also like to express my admiration and respect for Mr. Mitrea's passion and hard work for his research activity and for his dedication to his co-workers which highly inspired me and contributed to the success of this thesis.

I would like to express my gratitude and special thanks to Professor Adriana Vlad from POLITEHNICA University of Bucharest for granting me the honor of being a reviewer of my thesis, for her constant help throughout my student activity, her genuine interest in my work and the perspectives she envisioned for the thesis.
My gratitude and special thanks go to Professor Arnaud Reichart, deputy director at ENSTA ParisTech for granting me the honor and accepting the difficult task of reviewing the thesis, for the time and attention he dedicated to my thesis in the meeting that preceded the defense and for his insightful advice on the perspectives of the work.
I would like to thank Professor Christine Graffigne from the René Descartes University Paris for the honor of chairing the jury, for the time, patience and expertise she invested in evaluating my work.
I would like to thank HDR Claude Delpha from Paris-Sud University for the time he spent in evaluating the work, for his thoughtful comments and kind advice for the perspectives of the thesis.
I would like to express my appreciation to Mr. Benoît Maujean, R&D Head responsible at Mikros Image for bringing the industry and strategic point of view in the evaluation of the thesis and during the development of the HD3D2 project.

I would like to thank Mr. Arnaud de La Fortelle, Director of the CAOR Department and Mr. François, Goulette, coordinator of the thesis program at CAOR, who welcomed me at the

CAOR Department of MINES ParisTech and for their interest they showed for my work during the Doctorades.

I would like to express my appreciation to all the partners of the HD3D2 project for their help and support during the project meetings and for the availability of the HD3D-IIO video corpus (the testing corpus of the present thesis). Especially I would like to thank the two industrial partners who were particularly involved on the applicative side of the work, namely Mikros Image, through Mr. Benoît Maujean and Mr. Guillaume Chatelet and the CST (Commission Supérieure Technique de l'Image et du Son) through Mr. Rip O'Neil.

I would like to thank Mrs. Evelyne Taroni for her proactive attitude and valuable help with the administrative matters at the ARTEMIS department, and for the kind and informal French lessons I had when I went in her office.
I would also like to thank Mrs. Christine Vignaux and Mrs. Sylvie Barizzi-Loisel from the CAOR department for their valuable help in the administrative matters.

Special thanks: To Raluca Sambra for her generous help and very useful discussions on many image processing matters and beyond. To Thomas Laquet for his efficient help with the demos. To Afef Chammem and Sameh Hamrouni for their friendship and optimism day by day at the office. To Bojan Joveski, for being an exemple of patience and positive attitude towards work and people and for sharing his good spirit with me and with the colleagues.

I would also like to thank the entire ARTEMIS team, former and present members that I have met and who have contributed sometimes from the ARTEMIS offices, sometimes from far away to my work and who have all taught me something.

Contents

ABSTRACT i

PART I: VIDEO FINGERPRINTING OVERVIEW ..... - 1 -

I.1 Introduction ..... - 5 -

I.2 Definition ..... - 7 -

I.3 Theoretical properties and requirements ..... - 10 -

I.3.1 Uniqueness ..... - 11 -
I.3.2 Robustness ..... - 11 -
I.3.3 Database search efficiency ..... - 11 -
I.3.4 Evaluation framework ..... - 12 -
I.3.5 Video fingerprinting requirements ..... - 14 -

I.4 Applicative and industrial panorama ..... - 21 -

I.4.1 Video identification and retrieval ..... - 21 -
I.4.2 Authentication of multimedia content ..... - 23 -
I.4.3 Copyright infringement prevention ..... - 23 -
I.4.4 Digital watermarking ..... - 24 -
I.4.5 Broadcast monitoring ..... - 25 -
I.4.6 Business analytics ..... - 25 -

I.5 State of the art ..... - 26 -

I.5.1 Industrial solutions ..... - 26 -
I.5.2 Academic state of the art ..... - 27 -

I.6 Conclusion ..... - 45 -

References ..... - 47 -

PART II: VIDEO FINGERPRINTING AT WORK: TRACKART ..... - 53 -

II.1 TrackART: Synopsis ..... - 57 -

II.2 Offline phase ..... - 58 -

II.2.1 Pre-processing ..... - 58 -
II.2.2 Offline localization ..... - 59 -

II.3 Online phase ..... - 78 -

II.3.1 Pre-processing ..... - 78 -
II.3.2 Online localization ..... - 79 -
II.3.3 Fingerprint ..... - 84 -
II.3.4 Reduced fingerprint ..... - 91 -

II.4 TrackART possible configurations ..... - 94 -

II.5 Conclusion ..... **- 95 -**

References ..... **- 98 -**

PART III: TRACKART – EXPERIMENTAL RESULTS ..... **- 103 -**

III.1 Context ..... **- 107 -**

III.2 Testing corpus ..... **- 107 -**

III.3 Video retrieval use-case ..... **- 111 -**

III.3.1 TrackART Full Fingerprint evaluation ..... *- 112 -*
III.3.2 TrackART Reduced Fingerprint evaluation ..... *- 114 -*

III.4 Live camcorder recording use-case ..... **- 117 -**

III.4.1 TrackART Full Fingerprint evaluation ..... *- 119 -*
III.4.2 TrackART Reduced Fingerprint evaluation ..... *- 119 -*

III.5 Computational cost ..... **- 120 -**

III.6 Video fingerprint demonstrator ..... **- 122 -**

III.7 Conclusion ..... **- 125 -**

Reference ..... **- 126 -**

PART IV: FINAL CONCLUSIONS AND PERSPECTIVES ..... **- 127 -**

Conclusions ..... *- 131 -*
Perspectives ..... *- 132 -*

Appendix ..... **- 135 -**

A.1. Online localization illustrations ..... *- 135 -*
A.2. Publications ..... *- 138 -*
A.3. Selection of publications ..... *- 139 -*

ABSTRACT

Context

With the advent of affordable devices (capturing, processing, storage) and with the wide spread of broadband Internet access, massive amount of video content is produced and disseminated instantaneously. Hence, efficient tools for searching, retrieving and tracking video content in very large video databases (*e.g.* YouTube) have to be deployed in order to serve the purposes of applications like copyright protection, parental control, *etc*. Moreover, augmented reality turns live camcorder recorded video into a challenging research topic. Live camcorder recording (or, live camcording) is the process through which some video content displayed on a screen (in theaters, on a TV set, on an advertising display, …) is captured with an external camera.

A potential solution intensively considered in research studies is video fingerprinting. Video fingerprints are compact and salient video features computed from the video itself and which can uniquely identify it.

Video fingerprints can be best defined in relation to the human fingerprints, as illustrated in Figure 1. While the human fingerprint can be seen as a human summary (a signature) that is unique for every person, the video fingerprint can be seen as some short video feature (*e.g.* a string of bits with no particular format constraint) which can uniquely identify that video.

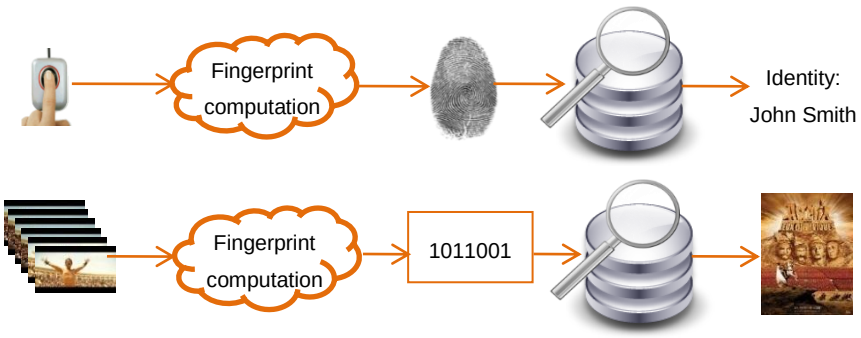


Figure 1: Human *versus* video fingerprinting

Scientific and technical challenges

Fingerprinting methods have three main characteristics:

- *Robustness to distortions*: fingerprints extracted from a video subjected to content-preserving distortions (attacked video) should be similar to the fingerprints extracted from the original video. Such attacks may include gray-scale conversion, linear or non-linear filtering, geometric transformations, *etc*. The robustness property is also quantified by two objective evaluation criteria, namely the *probability of missed detection* (P_{md}) and the *recall rate* (*Rec*).
- *Uniqueness*: fingerprints extracted from different video clips should be considerably different. This property is assessed by two objective evaluation criteria: the *probability of false alarm* (P_{fa}) and the *precision rate* (*Prec*).
- *Database search efficiency*: for applications with a large scale database, fingerprints should be conducive to efficient database search (fast fingerprint computation and matching, compact form, …), resulting in scalable solutions.

Current limitations

The fingerprinting state-of-the-art covers a large area of methodological tools from pixel difference of consecutive frames or RGB histograms to transform domain based fingerprinting approaches. However, despite the wide range of such methods, open research topics are still connected to each of the three above mentioned scientific challenges, see Table 1.

First, concerning the uniqueness, the state-of-the-art methods are generally constructed on heuristic basis and, with singular exceptions, tested on limited databases. Consequently, their mathematical basis (if any) comes rather as an *a posteriori* validation than as a true demonstration of the results. Secondly, to our best knowledge, no fingerprinting method was yet reported to withstand the live in-theater camcorder recording. Finally, the scalability issue, *sine qua non* for the content distribution on Internet can currently be obtained only at the expense of the uniqueness / robustness proprieties.

Moreover, these methods are generally tested on TV content data sets and don't take into account the particularities of the cinema content characterized by very high quality and presenting a high dynamics of the visual content, outdoor/indoor scenes and arbitrarily changing lighting conditions.

Methodological contributions and achievements

The present thesis takes a different approach and advances a novel DWT (discrete wavelet transform)-based video fingerprinting method involving a mathematical decision rule for the detection of replicas.

The fingerprint *per-se* is represented by a set of 2D-DWT coefficients of frames sampled from the video sequence. An in-depth statistical investigation on the 2D-DWT coefficients demonstrated not only the stationarity of such coefficients but also the stationarity of their modifications under the computer-simulated camcorder attacks.

Through its accurate representation of visual content, the wavelet transform grants the fingerprints the uniqueness property and limits the occurrences of false alarms (*i.e.* fingerprints extracted from different video content have to be different). The fingerprint matching is done based on a repeated Rho test on correlation which allows the detection of replicas, hence ensuring the robustness property (*i.e.* fingerprints extracted from an original video sequence and its replicas should be similar in the sense of the considered similarity metric).

In order to make the method efficient in the case of large scale databases, a localization algorithm is employed. Consequently, the replica sequence is not matched to the entire reference video collection but only with a few candidates determined based on a bag of visual words representation (concept introduced by Sivic and Zisserman in 2003) of the video keyframes. An additional synchronization mechanism able to address the strong distortions from difficult use-cases such as camcorder recording in cinema was also designed.

The method scalability is granted by the localization and synchronization procedures and by its low complexity which is kept under the $O(n \log n)$ limit.

Summarizing, the contributions of the thesis are threefold:

- a novel fingerprinting feature with a new mathematical matching procedure;
- a dynamic synchronization block addressing for the first time the live camcorder recording;
- a bag of visual words algorithm employed for granting the fingerprinting system scalability to large scale databases;

Functional evaluation

This method is evaluated in industrial partnership with professional players in cinematography special effects (Mikros Image) and with the French Cinematography Authority (CST - Commission Supérieure Technique de l'Image et du Son).

Two use cases have been incrementally considered: (1) computer generated replica video retrieval and (2) live camcorder recorded video retrieval. The reference dataset was composed of 14 hours of video content from different movies produced in Ile de France (*e.g.* Asterix), under the framework of the HD3D-IIO and HD3D2 CapDigital Competitiveness Cluster Projects. The query dataset was organized differently for each use case. For computer generated replica video retrieval, the query dataset consists of 24 hours of replica video content generated obtained by applying eight types of distortions (*i.e.* brightness increase/decrease, contrast decrease, conversion to grayscale, Gaussian filtering, sharpening, rotations with 2°, stirMark) on 3 hours of original video content from the reference dataset. For the live camcorder recording, the query corpus consisted of 1 hour of live camcorder recorded video content from the reference dataset.

The inner 2D-DWT properties with respect to content preserving attacks (such as linear filtering, sharpening, geometric, conversion to grayscale, small rotations, contrast changes, brightness changes, live camcorder recording), ensure the following results: in the first use case the probability of false alarm reached its null ideal value whereas the missed detection was lower than 0.025, precision and recall were higher than 0.97; in the second use case, the probability of false alarm was 0.000016, the probability of missed detection was lower than 0.041, precision and recall were equal to 0.93

In the absence of a clear benchmarking between state-of-the-art video fingerprinting methods (different testing data sets), the performances of the proposed fingerprinting system have been set by the industrial partners Mikros and CST to lower than 5% for the probability of false alarm and missed detection and higher than 95% for the precision and recall.

Considering the first case the performance limits have been successfully met by the proposed method, whereas considering the second use case, the precision and recall performances although feature satisfactory results, still need to be improved with 2%.

Thesis structure

The present manuscript is structured in four main parts related to the video fingerprint overview, the proposed video fingerprinting method, the evaluation of the proposed method and to the conclusions which can be formulated, respectively.

Part I (Video fingerprinting overview) is composed by six sections (numbered from I.1 to I.6) and covering an introduction to the video fingerprinting: the main underlying definitions, a theoretical properties and requirements, a general panorama of the applicative and industrial use cases as well as state of the art on the research studies. The concluding section summarizes the open research challenges versus the current day methodological limitations

Part II is devoted to the specification of the TrackART, the new fingerprinting method advanced in the thesis. Its synoptic presentation (Section II.1) is followed by the detailed definitions of its building blocks, structured according to the offline (Section II.2) and online (Section II.3) blocks. Section II.4 considers two possible TrackART functional configurations: TrackART Full Fingerprint and TrackART Reduced Fingerprint. These two configurations are considered as a solution for reaching a potential trade-off among not only uniqueness, robustness and scalability but also fingerprint length. The TrackART key features are summarized in the concluding Section II.5.

Part III describes the experimental validation. The context of the study (the HD3D2 competitiveness cluster project in Ile de France) and the processed corpus are presented in Section III.1. On this occasion, two fingerprinting use cases are stated by the two industrial partners, namely the retrieval of video sequences under computer generated distortions by Mikros Image and the live camcorder recording use case by CST. The uniqueness and robustness experimental results corresponding to the two use cases and to the two TrackART configurations are presented and discussed in Section III.3-4. The computational cost (invariant with respect to the use case) is analyzed in Section III.5. Section III.6 briefly introduces a software demonstrator meant to accustom a novice user with the video fingerprinting basic concepts. The conclusions on the quantitative results are drawn in Section III.7.

Although each part of the thesis contains detailed conclusions, Part IV gives a retrospective view on the thesis main contribution and presents the direction for future work.

The thesis has three Appendices which contain visual illustrations of the online localization block (procedure included in the TrackART method, detailed in Section II.3.2), the list of publications co-authored by the PhD candidate and a selection of these publications.

Constraints	Challenge	Current limitation	Thesis contributions
Uniqueness	Accurate representation of the video content	Heuristic procedures	Fingerprint computation independent of random, time-variant conditions: ❑ stationary/ergodic fingerprints ❑ 2D-wavelet coefficients
Robustness	Mathematical ground In-theater live camcorder recording	Heuristic procedures No related method reported in the state-of-the-art	Mathematical decision rule in fingerprint matching: ❑ method based on a repeated statistical test ❑ statistical error control
Search efficiency	Scalability	Very few full scalable mono-modal methods reported in the state-of-the-art	Scalable method ❑ automatic retrieval procedure ❑ O(n) complexity for fingerprint computation ❑ O(nlog(n)) complexity for fingerprint matching with respect to the fingerprint size

Table 1. Camcorder recording robust video fingerprinting: constraints, challenges, state of the art limitations and thesis contributions.

PART I: VIDEO FINGERPRINTING OVERVIEW

Abstract

This part incrementally presents the main definitions and the state of the art limitations for video fingerprinting.

Video fingerprints are compact and salient video features computed from the video itself and which can uniquely identify it. Fingerprinting methods have three main characteristics. The first is the uniqueness, *i.e.* fingerprints extracted from different video clips should be considerably different. The second is robustness to distortions, *i.e.* fingerprints extracted from a video subjected to content-preserving distortions should be similar to the fingerprints extracted from the original video. Such attacks may include gray-scale conversion, linear or non-linear filtering, geometric transformations, live camcorder recording, etc. The third is database search efficiency, *i.e.* for applications with a large scale database, fingerprints should be conducive to efficient database search (fast fingerprint computation and matching, compact form, …), resulting in scalable solutions.

The fingerprinting state-of-the-art analysis brings to light that research challenges are still taken for each of the above mentioned properties. First, concerning the uniqueness, the state-of-the-art methods are generally constructed on heuristic basis and, with singular exceptions, tested on limited databases. Consequently, their mathematical basis (if any) comes rather as an *a posteriori* validation than as a true demonstration of the results. Secondly, to our best knowledge, no fingerprinting method was yet reported to withstand the live in-theater camcorder recording. Finally, the scalability issue, *sine qua non* for the content distribution on Internet can currently be obtained only at the expense of the uniqueness / robustness proprieties.

Keywords

Video fingerprints, uniqueness, robustness, database search efficiency, distortions, gray-scale conversion, linear or non-linear filtering, geometric transformations, live camcorder recording.

Resumé

Ce chapitre regroupe les principales définitions et limitations de l'état de l'art pour le traçage du contenu vidéo.

Le traçage du contenu vidéo est réalisé à partir des empreintes numériques qui sont des caractéristiques compacts et saillantes extraites à partir du vidéo contenu lui même, et qui peuvent identifier une séquence vidéo de manière unique.

Les méthodes de traçage ont trois propriétés principales. La première propriété est l'unicité, c'est-à-dire les empreintes numériques extraites de différents clips vidéo doivent être considérablement différentes. La seconde propriété est la robustesse aux distorsions, c'est-à-dire, les empreintes numériques extraites d'une vidéo soumise à différentes distorsions préservant le contenu visuel doivent être similaires aux empreintes d'origine. Des telles distordions peuvent inclure la conversion en niveaux de gris, le filtrage linéaire ou non linéaire, les transformations géométriques, ou bien l'enregistrement en salle de

cinéma. La troisième propriété est la scalabilité, c'est-à-dire pour les applications vouées au traitement des bases des données à grande échelle, les empreintes numériques doivent être propices à une recherche efficace dans ces bases (calcul rapide des empreintes numériques, appariement rapide, forme compacte, ...).

L'analyse de l'état de l'art pour le traçage de la vidéo met en exergue qu'il y a encore des défis à adresser pour chacune des propriétés mentionnées ci-dessus. Tout d'abord, concernant l'unicité, les méthodes de l'état de l'art sont généralement construites sur des bases heuristiques et, à quelques exceptions singulières, testés sur des bases de données limitées en contenu. Par conséquent, leur support mathématique vient plutôt comme une validation *a posteriori* au lieu d'une véritable démonstration des résultats. Deuxièmement, à notre connaissance, aucune méthode de traçage du contenu video n'a été encore signalée à résister aux distordions introduites par l'enregistrement en salle de cinéma. Finalement, la scalabilité, est actuellement obtenue au détriment des propriétés d'unicité et de robustesse.

Mots clés

Empreintes digitales, unicité, robustesse, scalabilité, distorsions, conversion en niveaux de gris, filtrage linéaire ou non- linéaire, transformations géométriques, enregistrement en salle de cinéma.

I.1 Introduction

With Shannon's discovery of information theory and with the breakthrough brought in the hardware technology by the transistor's invention, the digital technology developed at a tremendous pace. Major technical achievements, discoveries and inventions exploded year after year, such as the microprocessor, the cell phone, the PC, the operating system, the Internet, the smart phone as illustrated on the timeline in Fig.I.1.

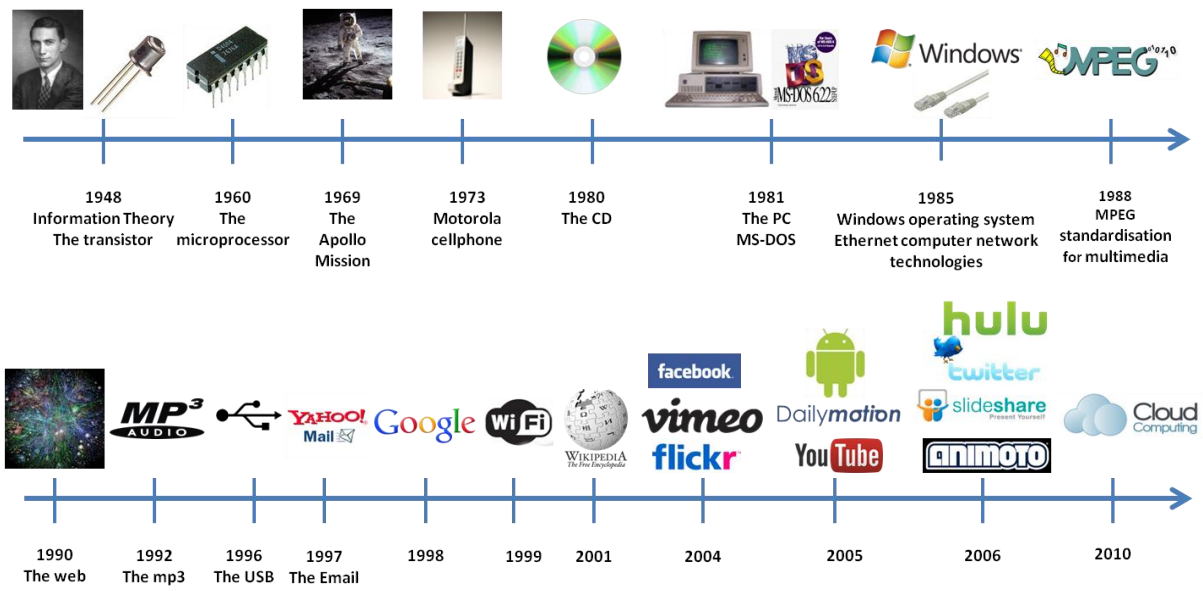


Fig.I.1: Timeline of memorable inventions and advancements of the digital revolution

On the one hand, the theoretical and technological progress of the digital revolution fostered the progress in all domains of activity (telecommunications – cell phones, Internet, medical – imagistic, industry – robotics, education – e-learning, research – NASA, CERN, defense – surveillance, drones, transportations, commerce – online shopping, entertainment – movie industry). On the other hand, in the context of worldwide economic growth, mass production of devices (*i.e.* PCs, cameras, cell phones as illustrated by the increasing sales depicted Fig.I.2) and large spread of broadband Internet access, technology became an essential attribute of people's lives. The customary cell phone, the mandatory PC with Internet connection (2.1 billion Internet users worldwide by the end of 2011 [ROY 12]), the personal music and video collection, the ubiquitous social networks (2.4 billion social networking accounts by the end of 2011 [ROY 12]) are a few examples of technologies considered as necessary by the majority of people. Such a state of mind combined with the available technology enables people to reach and use information and knowledge in a few mouse clicks and empowers them to build and distribute their own creations, be them ideas, text, software, multimedia, etc.

Among the sectors which were influenced the most by the user becoming interactive with technology is the multimedia industry which saw a wide range of applications, opportunities and challenges

coming up. With 72 hours of video content (and increasing as illustrated in Fig.I.3) being uploaded every minute and with 1 trillion playbacks on YouTube (*i.e*. 140 playbacks per person) on Earth in 2011 [ROY 12] the multimedia content becomes a serious and profitable resource.

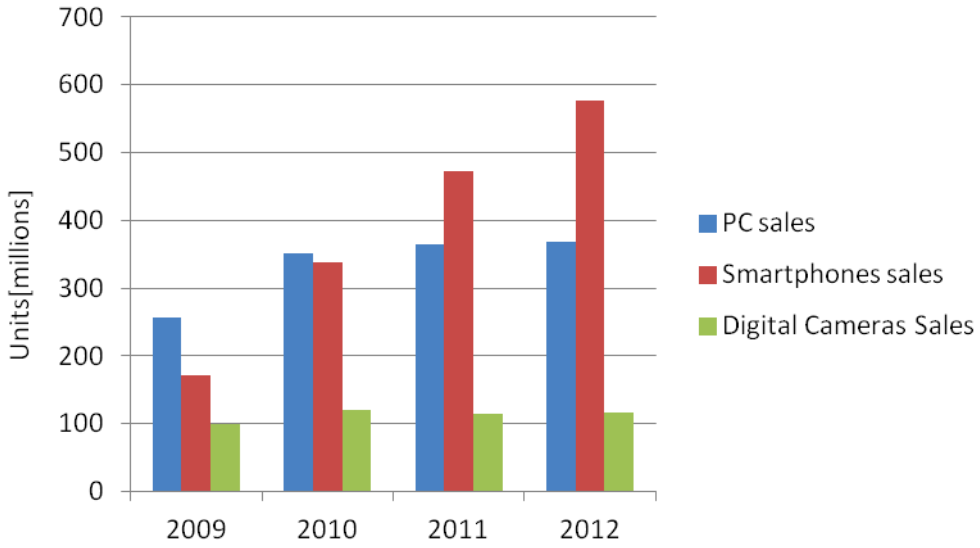


Fig. I.2: Sales in PCs, smart phones and digital cameras 2009-2012 [GAR 12]

Moreover, the Google sites, of which YouTube is the largest, hold 43% of the video views worldwide, the rest of 57% being assured by Vimeo, DailyMotion, Flickr, Facebook, Animoto, SlideShare and others [ROY 12].

Jointly with the increase in multimedia content, new viewing environments and delivery options have become available beyond the traditional TV: video on demand systems hosted by cable, telephone or satellite providers stream content through a connected TV, traditional set-top boxes, mobile phone, tablet, car entertainment system or PC allowing users to choose from a wide menu of programs and watch them at their convenience [AUD 12b].

Driven by the booming multimedia industry, the viewing devices sector has also seen a growth in smart gadgets, smart TVs (*i.e.* television set with integrated Internet capabilities, operating systems), PCs, handheld Internet phones and table devices (Apple iPhone, iPod, iPad, Motorola Droid, HP TouchPad, Samsung Galaxy, Motorola Xoom).

Currently, in order to make their businesses profitable and sustainable, multimedia stakeholders such as video sharing platforms (*i.e.* YouTube, Vimeo), television networks (*i.e.* BBC, TF1), national audio visual agencies (*i.e.* INA [INA 12], Beeld en Geluid [BEE 12]), smart TV providers, news portals, film studios (*i.e.* DreamWorks, Gaumont Film Company, Pixar), filmmakers, comedians, market analysis (*i.e.* Xerfi [XER 12]) monitoring agencies and advertising agencies (*i.e.* Auditoire [AUD 12a]) have to create added value for their content and to keep it protected from copyright infringements. These two key prerequisites become incrementally challenging with the continuously increasing volume of produced and consumed multimedia content and with the ease of the user interfering in the content creation and consumption phases.

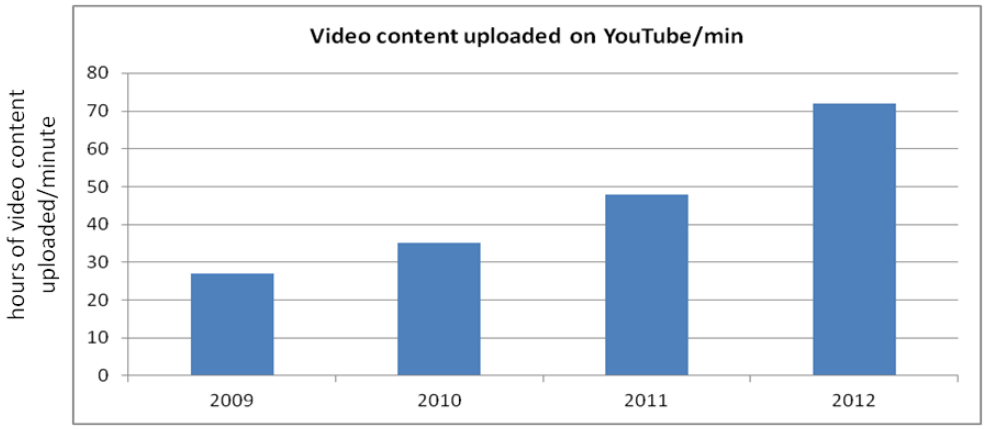


Fig.I.3: Video content uploaded on YouTube every minute

The solution that is intensively considered and researched is *multimedia digital fingerprinting*, commonly denoted as *multimedia content-based copy detection (CBCD)* or *near duplicate detection*. These terms were coined in order to designate technologies able to uniquely identify the multimedia content by means of the content's features (*e.g.* colors, shapes, textures, …) and not by its name of other metadata such as user annotations. In order to enhance the applicability and use of such technologies, two additional requirements are necessary. Firstly, multimedia content should be identified even if mundane or malicious transforms were applied to the content. Secondly, this type of video identification should be scalable with respect to the database size, so as to be successfully deployed even foe very large databases.

I.2 Definition

Video fingerprints can be best defined in relation with the human fingerprints [OOS 02] as illustrated in Fig.I.4. The patterns of dermal ridges on the human fingertips are natural identifiers for humans as discovered by Sir Francis Galton in 1893. Although they convey very little information compared to the entire human, human fingerprints are sufficient to uniquely identify a person even if the person changes haircut, clothes, or wears a wig or a disguise.

Analogously, video fingerprints are intended to be video identifiers. The video fingerprints have to be able to uniquely identify videos even if the video content goes under a predefined, application dependent set of transformations. The *transformations* a video can undergo will be further referred to as *modifications*, *distortions*, or *attacks*, be them malicious or mundane. The video which is transformed, modified, distorted or attacked will be denoted as a *copy* or a *replica* video.

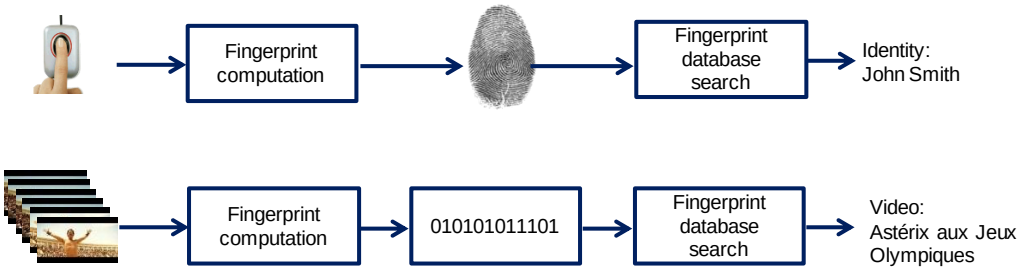


Fig.I.4: Human versus video fingerprinting

Content-based copy detection systems (CBCD) should not be confused with *Content Based Video Retrieval (CBVR)* system.

On the one hand, the CBVR systems aim at retrieving visually similar videos, *i.e.* from the same genre of category, for instance, soccer games, or episodes of soap operas.

On the other hand, the CBCD systems aim at retrieving the original version of a query sequence, and have to be able to discriminate between different content belonging to the same genre as illustrated in Fig.I.5, [LAW 06].

	
The same content, in color and grayscale version	Different content – different ties, a pin on the suit

Fig.I.5. Content based copy detection system *vs.* Content based video retrieval system requirements

Content-based copy detection systems should also not be confused with the watermarking systems.

A watermarking system inserts imperceptibly and persistently some additional information into a digital content (*e.g.* image, audio, video) [COX 08]. The additional information generally consists of some copyright information (*e.g.* owner, seller, *etc*). Imperceptibility refers to the property of the watermark to be invisible for a human observer while the persistency refers to the property of the mark to be detected even when strong malicious operations were applied to the marked content.

Watermarking schemes can address the technical challenges related to rights management, content management (*e.g.* filtering, classification), broadcast monitoring under the condition that the content is *a priori* watermarked, *i.e.* the additional information is inserted in the multimedia content before its distribution.

Although similar in terms of their applicative field, watermarking and fingerprinting differ in one essential aspect. Watermarking is an active technique: it inserts a mark prior to the multimedia content's distribution and then extracts the watermark in order to obtain the owner's information. Multimedia fingerprinting is a passive technique: it computes the fingerprints from the content itself and matches them to the reference fingerprints thus establishing the ownership.

Under the framework of a fingerprinting system, a *query* is the name given throughout this paper to a video whose identity is inquired, whereas *reference* is the name given to the video sequences belonging to the database of known identity videos. Consequently, the fingerprint of a query video sequence will be denoted as *query fingerprint* and the fingerprint of a reference video sequence will be denoted as a *reference fingerprint*.

Analogous to the human fingerprinting system, the video fingerprinting system consists of two steps: 1 - query video *fingerprint computation* and 2 - query *fingerprint matching* with reference fingerprints.

Giving more theoretical basis to this analogy, the design of a video fingerprinting system leads to: 1 - finding *features* from the video able to concisely represent and summarize video content and 2 - using as fingerprint matching strategy a *similarity metric,* which can assure the retrieval of replica videos in the context of various transformations.

The peculiarity of the video fingerprinting system compared to the human fingerprinting system arises from the peculiarities of the video content and from the variety of transformations that a query video can subsist. For instance the video fingerprint and its matching have to be designed in order to address the common situation of videos sequences with different lengths and the situation in which the query video is a fragment of a reference video. Therefore a *localization* procedure of the query in the reference video sequence has to be developed as an integrating part of the video fingerprinting system.

A video fingerprinting system comports a general principle as illustrated in the schema in Fig. I.6.

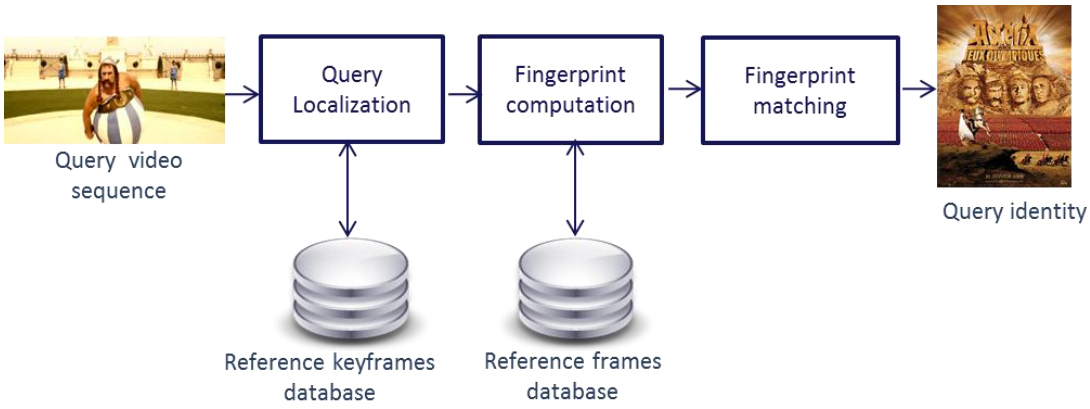


Fig.I.6: Video fingerprinting system schema

Firstly, a query video sequence whose identity is inquired is given as input to the fingerprinting system. Secondly the query sequence is localized within the collection of reference videos sequences. Thirdly, the fingerprints of the query and reference sequences are computed. Fourthly, the matching operation establishes the identity of the query video sequence.

As illustrated in Fig.I.6 the query localization and the fingerprint computation stages have to be connected to the reference database. In this way, the video sequences and relevant information derived from them is made available: in the query localization stage, the system needs to find the position of the query sequence within a reference video sequence while in the fingerprint computation stage, the system computes the fingerprints of the query and reference video sequences.

It can be intuitively noticed that, in a video fingerprinting system, while some computation has to be done when inquiring for a query, other operations can be performed before that moment in order to speed up the process of query identity retrieval. Therefore, the computation can be split in two parts, as illustrated in Fig.I.7: an online phase, when a user or a system is interested in the identity of the query video, and an offline phase which is performed before the query inquire and which computes all the relevant necessary information needed in the query localization and fingerprint matching stages.

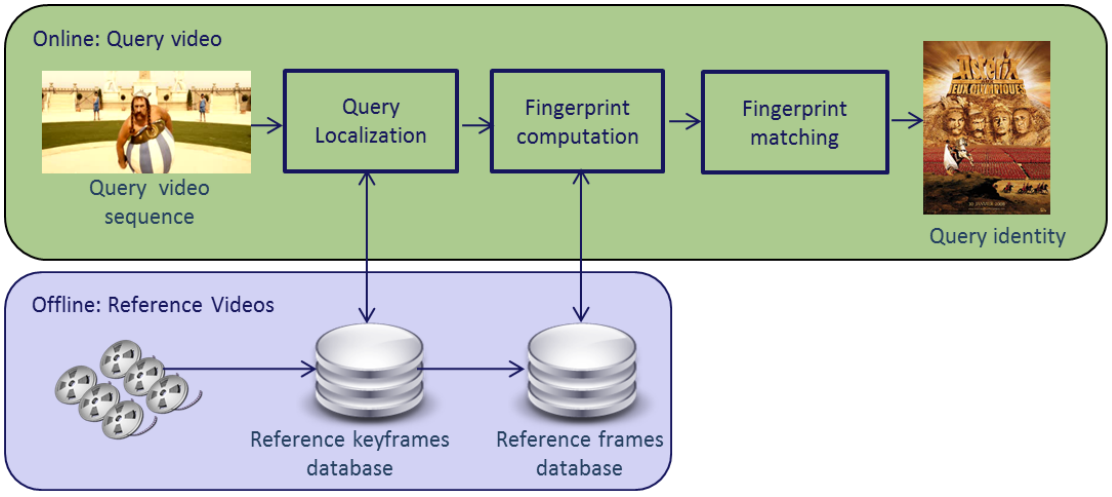


Fig.I.7: Video fingerprinting system: online and offline phase

I.3 Theoretical properties and requirements

The main properties a fingerprinting method features are robustness, uniqueness and database search efficiency.

I.3.1 Uniqueness

A fingerprinting method is said to feature *uniqueness* if the fingerprints computed from two video sequences with different content are different in the sense of the considered similarity metric. Fig.I.8.a illustrates how two different video contents are identified as having different identities based on the matching of their fingerprints.

The uniqueness property is assessed by the incidence of false alarms. A false alarm is encountered when the video fingerprinting system retrieves a video sequences which is neither the query not its replicas. Consequently, the uniqueness property is evaluated by the probability of false alarm or alternatively by the precision rate, as detailed in Section I.3.4.

I.3.2 Robustness

A fingerprinting method is said to feature *robustness* to a particular distortion if the fingerprint computed from an original video sequence and its replicas with respect to the considered distortion, are similar in the sense of the considered similarity metric. Fig.I.8.b illustrates the robustness property in the case of an original video content and its grayscale replica.

The robustness property is assessed by the incidence of missed detections. A missed detection is encountered when the video fingerprinting system does not retrieve a replica video sequence of the query video. Consequently, the uniqueness property is evaluated by the probability of missed detections or alternatively by the recall rate, as detailed in Section I.3.4.



Fig.I.8: (a) The uniqueness property; (b) The robustness property

I.3.3 Database search efficiency

A fingerprinting method is said to feature *database search efficiency* if the computation of the fingerprints and the matching procedure ensure low, application dependent computation time for the video's identity retrieval. The database search efficiency is assessed by the mean computation

time needed to retrieve the identity of a query in the context of a considered video fingerprinting use case.

I.3.4 Evaluation framework

The performances of a video fingerprinting system can be objectively assessed by evaluating its properties: the uniqueness, the robustness and the database search efficiency.

The evaluation of the uniqueness and the robustness properties can be synoptically achieved by using the schema in Table I.1.

Considering a query sequence whose identity is looked up in a reference database with the help of a video fingerprinting system. The two statistical hypotheses are H_0: the query is a replica of a video sequence and H_1: the query is not a replica of a video sequence. The output of the system can be of two types: (1) - positive when the query is identified as replica of a video sequence and (2) - negative when the query is not a replica of a video sequence.

When a user is examining the results outputted by the system or when these results are compared with the ground truth, the correctness/rightness of the results is established: if the result provided by the system is correct – the attribute given to the results is true and if the result is incorrect – the attribute given is false.

The above principle yields four types of situations arising at the output of a fingerprinting system:

- False positive: the system erroneously retrieved a reference video sequence as a copy of the query.
- False negative: the system erroneously did not retrieve a reference sequence which is a copy of the query.
- True positive: the system correctly retrieved a reference video sequence which was a copy of the query.
- True negative: the system correctly did not retrieve a reference video sequence which was not a copy of the query.

	D_0	D_1
H_0	True Positive	False Positive
H_1	False Negative	True Negative

Table I.1 Decision matrix

The false positive results are also referred to in the literature as false alarms and will be denoted as *fp*. The false negative are also referred to as missed detections and will be denoted as *fn*. The true positives will be denoted as *tp* and the true negatives, as *tn*.

Having a fingerprinting system, a reference database and being given a query video sequence, the system can output several false positives, false negatives, true positives and true negatives.

In order to objectively evaluate a video fingerprinting system, the measures above have to be formalized into some performance indicators, as follows.

In order to evaluate the uniqueness property two measures are considered in the literature: the *probability of false alarm* (P_{fa}) and the *precision rate* (*Prec*), defined by the following formulas, [SU 09], [LEE 08]:

$$P_{fa} = \frac{fp}{tp + fn + fp + tn} \quad \text{(I.1)} \qquad Prec = \frac{tp}{tp + fp} \quad \text{(I.2)}$$

In order to evaluate the robustness to distortions property is also quantified by two objective evaluation criteria, namely the probability of missed detection (P_{md}) and the recall rate (*Rec*), as defined below:

$$P_{md} = \frac{fn}{tp + fn + fp + tn} \quad \text{(I.3)} \qquad Rec = \frac{tp}{tp + fn} \quad \text{(I.4)}$$

On the one hand, an efficient fingerprinting method should ensure a low probability of false alarm (*i.e.* low probability of retrieving video sequences which are neither the query nor its replicas) and low probability of missed detection (*i.e.* a low probability of not retrieving replica video sequences of the query). On the other hand, high values for precision (*i.e.* a high probability of retrieving replica video sequences for a given query out of all the retrieved video sequences) and recall (*i.e.* a high probability in retrieving all the replica video sequences existing in a database for a given query) should also be obtained.

The probability of false alarm and miss detection probabilities in their classical format cannot be applied to a video fingerprinting system, unless the query and reference sequences have the same lengths, *i.e.* when a query is individually compared to each sequence in the database. However such a situation is not corresponding to the reality when the query and reference videos sequences can have various lengths and when the query can be a part of a reference sequence, at an unknown position.

Assuming the video fingerprinting system is time-invariant (which is always the case) the probabilities of false alarm and miss detection can be temporally estimated, as follows.

$$P_{fal} = \frac{fp_l}{T_{refdata}} \quad \text{(I.5)} \qquad P_{mdl} = \frac{fn_l}{T_{target}} \quad \text{(I.6)}$$

Where fp_l is the total length in (minutes) or the false alarms and fn_l is the total length in (minutes) or the missed detections, $T_{refdata}$ is the total length (in minutes) of the entire reference database and T_{target} is the total length (in minutes) of the video replicas in the reference database. Hence $T_{refdata}$ and T_{target} are fixed values, known by pre-processing the database, while *fpl* and fn_l are random results (experiment dependent) outputted by the system)

In order to properly evaluate a system, the precision and recall have to be jointly used with the probabilities of false alarm and missed detection.

Precision and recall are two measures very commonly used in the evaluation of information retrieval systems. However they are not statistical measures as they are not taking into account the true negative results. In order to take into account the true negative results and present the properties of a system comprehensively, the probabilities of false alarm and missed detection have to be taken into account. The probability of false alarm is a type II statistical error (*i.e.* wrong data are taken as good), while the probability of miss detection is a type I statistical error (*i.e.* good data are refuted by the test); hence they grant statistical relevance to the obtained results.

The database search efficiency property can be objectively assessed by the average processing time required by the video fingerprinting system to identify the query within the reference database and to output the result for a query video sequence. The average processing time can be obtained by averaging the processing time required by the system for the considered collection of queries.

I.3.5 Video fingerprinting requirements

With the social, economic and technical context beneficial to video fingerprinting applications, a large set of distortions and modifications can be envisioned to affect the video content. Leveraging the robustness, uniqueness and database efficiency performances for different applications, hence for different modifications encountered, is the innovation playground for video fingerprinting systems.

The modifications a video sequence can be subject to in order to become a replica can be classified in three major categories depending on the video features they modify, namely the video format, the frame aspect and the video content; they are synoptically presented in Table I.2.

The limit of the applicative field of video fingerprinting is given by the commercial (or entertainment) value of the altered video, hence the transformations which render the video unusable are not considered in the sequel.

I.3.5.1 Video format modifications

With the wide range of applications, manufacturers and devices a variety of formats have been developed for the video content (*e.g.* the MPEG-4 Part 2, H.264/MPEG-1 AVC standards giving rise to the Blu-ray, HD DVD Digital Video Broadcasting, iPod Video, Apple TV implementations) imposing a mandatory requirement on video fingerprinting systems, robustness to different encoding and successive transcoding. Encoding refers to the process of converting the source video into digital code

symbols followed by compression in order to make the video easier to distribute. Transcoding refers to the process of converting the video to another encoding format which is usually necessary when the target device has limited storage capacity or when the device does not support the initial format. Other video format modifications can occur when some encoding parameters are changed, *i.e.* frame rate changes or changes in the compression rate which yields bitrate changes and which degrade the visual quality of the video.

On the one hand encoding, transcoding and other parameter changes can occur in mundane and automatic video manipulations such as video upload on the Internet or video serving applications on thin clients. On the other hand, these modifications can be intentionally induced in videos by malicious users in order to render the video untraceable and to avoid the copyright policies. Software which serve transcoding and parameters modifications are open source software such as ffmpg or SUPER (Simplified Universal Player Encoder & Renderer), Mencoder, Mplayer, x264, *etc*.

Another important distortion that a video can subsist is the analog to digital conversion which occurs when a video projected on a screen is captured with a digital camera, *e.g.* camcording in theatre. Due to the inherent quantization, although sometimes not visible to human observers, video information is lost.

While the previous modifications could be mundane or malicious, the frame addition, frame dropping and frame substitution modifications are largely malicious aiming at desynchronizing the video and to render it undetectable by fingerprinting systems.

Frame addition refers to inserting white/black frames in the beginning or the end of the video, or to inserting a certain amount of copy frames between the original frames. Fade-over is a particular case of frame addition consisting in a transition effect in which the content of a frame fades away and leaves place to new content. Frame dropping is the opposite of frame addition and consists in removing from the video sequence a certain amount of frames. The frames can be added or removed from the original video sequence at random positions or uniformly through the entire video or through parts of it. Frame substitution consists in replacing a certain amount of frames at particular/random chosen location with frames from other videos or from the video itself. Depending on the amount and on the type of frames added/dropped/substituted, these modifications can be noticed by the user, *i.e.* for 1-3 copy frames added/dropped per second at a frame rate of 25 fps, most users will not notice disturbing effects, while when adding 1 black/white frame in the same conditions would decrease the user experience.

Distortions	Examples
Video format	➢ encoding, transcoding ➢ compression ➢ frame rate changes ➢ bitrate changes ➢ D/A and A/D conversions ➢ frame dropping, frame addition (*e.g.* fade-over), frames substitution
Frame aspect	➢ color modifications: conversion to grayscale; conversion to sepia ➢ color filtering or corrections ➢ decrease of color depth
	➢ photometric changes: brightness, contrast, saturation ➢ gamma correction ➢ histogram equalization
	➢ filtering: linear (Gaussian, sharpening), non-linear (median filter)
	➢ noise addition
Frame content	▪ *affine transformations* ▪ geometric modifications: ○ uniform or non-uniform scaling, rotations ○ reflection ○ aspect ratio changes ○ dilations ○ contractions ○ shear ▪ similarity transforms (spiral similarity) ▪ translations ▪ *cropping* ▪ letterbox removal ▪ row or columns removal ▪ *insertion*: text, caption, pattern, letter-box insertion ➢ *picture in picture* ▪ *shifting* ▪ *StirMark*
Mixed	▪ combinations of all the above modifications

Table I.2: Types of computer or camcorder generated video modification

I.3.5.2 Frame aspect modifications

The modifications changing the aspect of the video frames refer to the following categories of distortions: color, photometric, filtering and noise addition.

The color modifications consist in changing the composition of the color balance in the video frames (*i.e.* modifying the values of the pixels' colors), changing the color depth (*i.e.* changing the numbers of bits used to represent the color of an image pixel: 1-bit color, monochrome; 8-bit color, 256 colors; 24-bit color true color more than 16 million colors; 30-48-bit color, deep color), converting the image in grayscale, filtering a certain color channel (R, G, B) or swapping colors (*i.e.* RGB to BGR, replacing the red color channel with the blue one, or other configurations).

Adjustments in the brightness (*i.e.* in the RGB color space, brightness is the arithmetic mean of the red, green and blue color coordinates), contrast (*i.e.* the difference between the black and white levels in images), saturation (*i.e.* the dominance of hue in the color), gamma corrections (*i.e.* an nonlinear operation $\mathrm{Vout} = \mathrm{AV_{in}}^{\gamma}$, where A is a constant and V is the value of a pixel, which changes the brightness of an image) or histogram equalizations (*i.e.* enhancement of the contrast of the images) of a frame are the photometric distortions a video often subsists.

Image filtering is an operation which consists in removing some unwanted components of the 2D signal which is the frame. The Gaussian filtering or blurring is a type of linear filtering which passes the low frequencies and attenuates the the high frequencies, *i.e.* attenuating the contours of the shapes in the video frames. Sharpening is a type of linear filtering, which attenuates the low frequencies but passes the high frequencies, hence keeping the details in the images. Median filtering is a non-linear filtering operation used to remove noise from images and which is usually used in pre-processing steps in order to enhance the results of further processing, *e.g.* edge detection.

Image noise addition consists of adding a noise signal (Gaussian noise, white noise, salt and pepper noise) to the video frame in order to decrease its quality. The noise can be added by image processing operations or can be produced by the sensors and circuitry of digital cameras when camcording.

The color and photometric modifications as well as filtering and noise addition can be induced in videos by image processing operations or intrinsically by capturing the video with external devices which implicitly change the colors and the values of the photometric parameters due to the device dependent sensors, circuitry and transducers' parameters.

I.3.5.3 Frame content modifications

The distortions which modify the content of the frame itself can be of the following types: affine transformations, cropping, insertion, picture in picture, rows or columns shifting. By changing the intrinsic content of the frames, these modifications are difficult to handle by fingerprinting systems and generally require dedicated pre-processing blocks before the fingerprinting solution is deployed, *e.g.* letterbox removal block, detection and removal of caption, text or pattern, detection and extraction of the videos of interest from the background or the foreground.

The affine transforms are the transforms which preserve the collinearity of points (*i.e.* all points lying on a line initially still lie on a line after the transformation) and ratios of distances (*i.e.* the midpoint of a line segment remains the midpoint after transformation).

The affine transformations for videos (applied at frame level) include the following types of modifications: geometric contraction, expansion, dilation, reflection, rotation, shear, translations, and their combinations. In general, the affine transformations are a combination of rotations, translations, dilations and shears.

An example of affine transformation is the rotation-enlargement transformation which combines a rotation and an expansion and can be mathematically written as in (I.7):

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = s \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{(I.7)}$$

Where (x', y') are the coordinates of the rotated point, (x, y) are the coordinates of the original point, s is the scale factor and α is the rotation angle.

Such distortions are induced in videos either by using image processing software *e.g.* the Adobe Photoshop or by means of camcording, Fig. I.10.

Affine transformations can greatly modify the content, hence the limit of applicability of fingerprinting solutions.

Scaling or resizing consists in changing the dimensions of the video frames, *e.g.* dilations or contractions of the height and width. Scaling can be done with the same scale factor for both height and width of frames (*i.e.* uniform/isotropic scaling) with different scale factor *i.e.* non-uniform, anisotropic scaling. The advantage of using uniform scaling is the fact that it preserves the shapes of objects inside the frames whereas non-uniform scaling changes these shapes. However in practice both scaling are intensively used in all types of applications, hence the modifications they induce in videos have to be addressed by video fingerprinting systems.



Original version	Camcorded replicas

Fig.I.10: Affine transforms induced by camcording

Small rotations (with angles ranging from ± 1° to ± 5°) often combined with cropping and scaling are efficient attacks as they generally do not modify the commercial value of the video. However they affect the frame content itself by removing the cropped parts and therefore can make a video fingerprinting system to mistakenly take the rotated, cropped and resized content for a new content and not a replica. In Fig. I.11 an original frame is trigonometrically rotated with 2°, 3° 5° and 10° in column (a) and cropped and resized in column (b). It can be noticed that up to 5° rotation the video content is visually similar to the original while the 10° rotation and cropping removes a large part of the initial content.

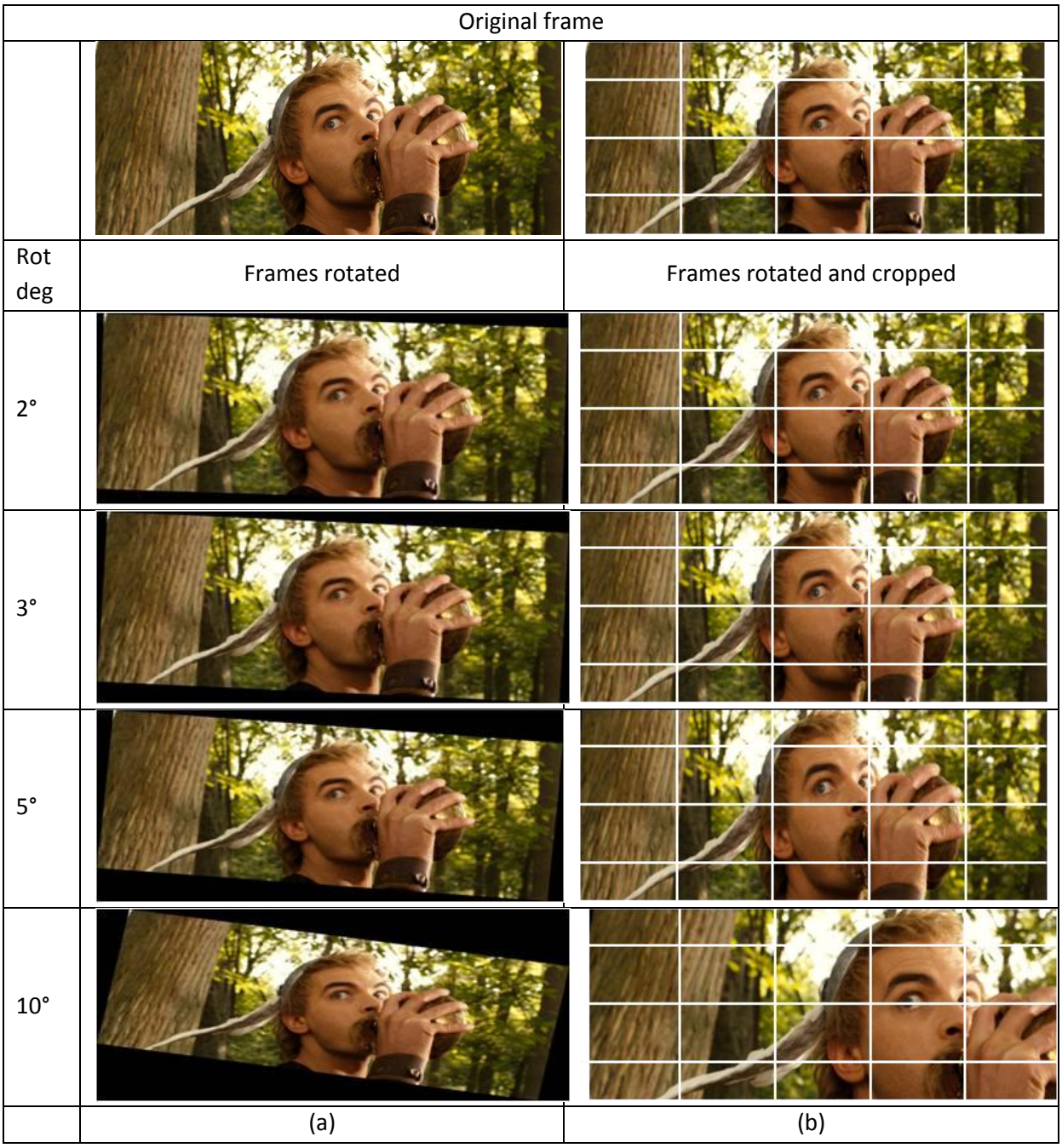


Fig. I.11 : Frames rotated with 2°, 3°, 5° and 10° in (a) column and frames rotated and cropped in (b) column

Reflection or vertical flipping consists of generating a replica frame by mirror-reversal of an original frame as illustrated in Fig. I.12. While the commercial value of the video is not altered, disturbing artifacts can appear when the scenes change.



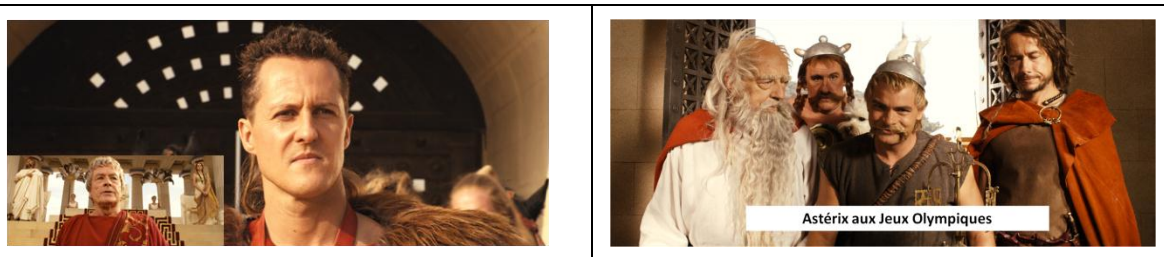
Original frame	Vertical flipped frame

Fig. I.12: Vertical flipping

Image aspect ratio is the proportional relationship between its width and height. The diverse video standards deployed in various applications *e.g.* the HD 16:9, the standard television 4:3, the widescreen cinema standard 39:1 demand from video fingerprinting systems to cope with aspect ratio changes.

Cropping consists in removing certain parts of the frames content such as letter boxes, rows or columns depending on the application. Insertion of content does the reverse of cropping, which is inserting other visual content in the video frames such as text, captions, patters, letter-boxes.

Picture in picture consists in displaying two videos in the same time and on the same frame, one video being the foreground and one video being the background as illustrated in Fig. 13.a.



(a) Picture in picture	(b) Insertion of caption

Fig. I.13: Television specific modifications

Cropping, insertion and picture in picture modifications are widely used in post-production and television processing of the video when several videos are needed to be displayed at the same time on the screen or other information relevant for the broadcast program, news or other announcements are necessary, Fig. 13.b.

Frame shifting consists in moving to the right, to the left, up or down a certain amount of columns or rows of the video frames. The amounts of columns or rows shifted can vary between 1% to 5% of the frame's width or height as it easily affects the visual quality of the video.

StirMark is a software package developed by Fabien Petitcolas [PET 00] which is generally used for benchmarking watermarking schemes. The software package contains several attacks such as cropping, rotation, rotation-scale, sharpening, Gaussian filtering, aspect ratio modifications and the StirMark random bending attack. The most well known is the StirMark random bending attack (which will be further referred as the StirMark attack), or random geometric distortion which applies a combination of minor geometric distortions *i.e.* the image is slightly stretched, sheared, shifted and/or rotated by an unnoticeable random amount and finally re-sampled. The stirMark attack simulates camcording in cinema and is one of the strongest attacks for fingerprinting and watermarking schemes.

Once the applicative ground for video fingerprinting is defined and its concepts established, the state of the art comes with a wide palette of methods and approaches.

I.4 Applicative and industrial panorama

Video fingerprinting is the tool that enables a system to manage video content according to some predefined rules, by using the video content itself. Therefore, in the booming video industry its applications have a wide range and are summarized below.

The performance requirements for video fingerprinting system can slightly vary across use cases but in general the missed detections and false alarms have to feature very low values, and the computational time has to stay reasonable low in order to comply with the time requirements of in the use case.

I.4.1 Video identification and retrieval

Video identification and retrieval is at the heart of all systems dealing with video. The ability to identify and retrieve video even under distortions is a powerful tool for increasingly many applications.

Given a very large database of videos (*e.g.* TV broadcast archive) and a query video sequence (*e.g.* a segment of a film), the identification of such a query can pose complex challenges (*e.g.* time requirements, human observes). A video fingerprinting system enables the identification of a particular video sequence by computing its fingerprint and by efficiently querying it among the reference fingerprints without using human observers, Fig I.14.

A possible use case for identification of multimedia in large databases is interactive advertising. In Fig.I.15, an agency has created a digital fingerprint for their specific TV commercial. When the fingerprint of the content playing on the screen is detected, a pop-up overlay dialog box is triggered on top of the advertisement asking the viewers if they want to take advantage of the coupon being offered on screen. By pressing their TV remote select button the viewers confirm they would like the coupon offered. Using the LAN connection, a coupon request is sent via the Internet to the retail web site. The requested coupon is sent by the retailer to the viewer's smart phone [AUD 12c].

The video fingerprinting scheme employed in the identification and retrieval of videos from large databases has to be adapted to the use case.

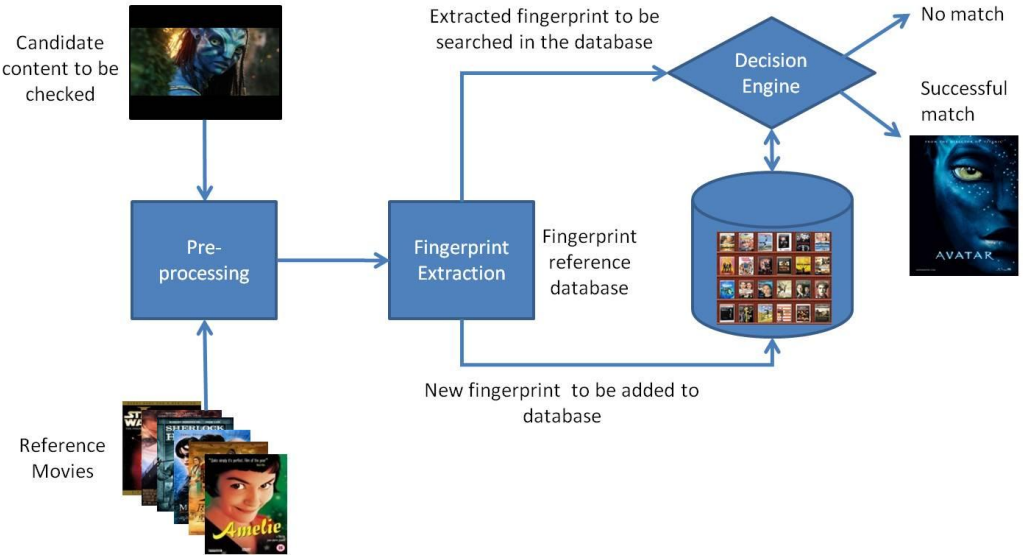


Fig.I.14: Video Identification and retrieval of video sequences

Considering the interactive advertising, the false alarms have a slightly greater impact than the miss detections due to the fact that the user is involved. A missed detection is preferable, *i.e.* the pop-up overlay dialog box with the promotional coupon does not appear and hence the user does not see the offer. A false alarm means that the promotional coupon appears when another video sequence is running at TV, making it unpleasant for the user. However, the miss detections have to be sufficiently low so as to promote the offer. In this use case, the distortions are also related to the changes which can appear in the video format during the broadcast.



Fig.I.15: Interactive advertising

I.4.2 Authentication of multimedia content

Due to powerful software (*e.g.* Photoshop, Windows Movie Maker, Pinnacle) for multimedia manipulation, content became very easy to manipulate and alter (*e.g.* change of hair color of one of the characters), therefore in many cases the originality of the content might need to be checked. An authentication system based on fingerprinting verifies the originality of the content and aims at detecting the malicious transformation. This is achieved by designing a fingerprint and a similarity metric able to detect any minor transformation in the query compared to the original version.

In general, in content authentication applications the distortions that can appear are related to image aspect and content modification, as detailed in Sections I.3.4.2-3. Moreover, for such applications, the miss detections have a critical impact on the performances of the system, whereas the false alarms can be easier accepted, therefore the fingerprint and the similarity metric between the fingerprints have to be designed accordingly.

I.4.3 Copyright infringement prevention

Web 2.0 services like YouTube, Vimeo, DailyMotion offer platforms for users to view and exchange videos. On numerous occasions YouTube was accused of being an illegal distribution channel and trials such as Viacom [WIR 12] pushed Google Inc. (the YouTube holder) to implement technology able to detect copyright infringement in their video database. Such technology relies on video fingerprinting principles and in the case of YouTube it is named the Content ID system. In order to achieve copyright infringement-free video database by means of video fingerprinting, content owners would have to provide reference fingerprints to user generated content (UGC) sites, which would allow through the matching procedure the identification of the video content. According to this identification and to the business or copyright rules established for each video, action can be taken, *e.g.* allow, filter, notify as illustrated in Fig.I.16.

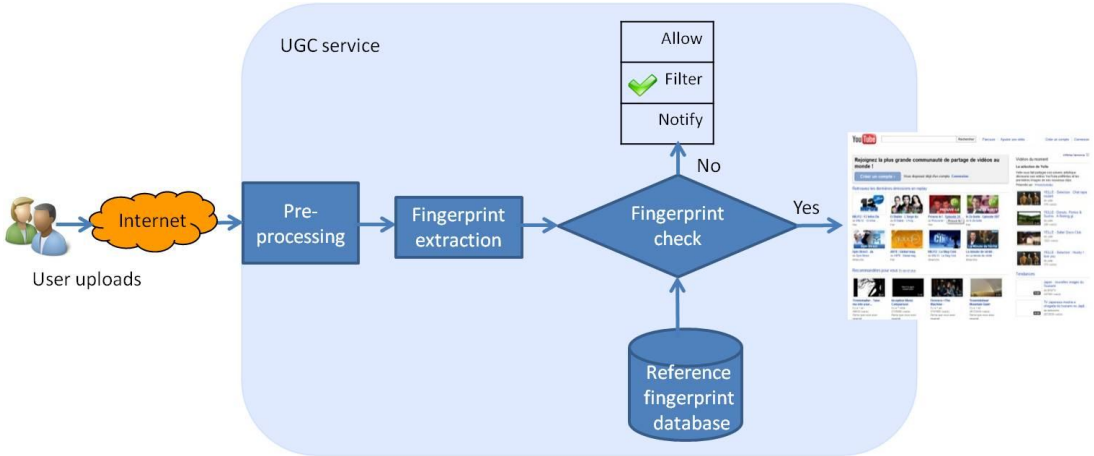


Fig.I.16: Video filtering in UGC platforms

The traffic of multimedia content in the P2P networks (*e.g.* BitTorrent, Cyberblocks, Gnutella) augmented to 50-60% of the total Internet traffic nowadays [VOB 08]. In Envisional's 2011 report [ENV 11] it was calculated that a minimum of 23.76% of all Internet bandwidth is devoted to the transfer of infringing content. In this context, detecting and tracking the copyright infringing traffic is an interesting application for content producers and owners. Video fingerprinting can be a solution for monitoring and tracking the copyright infringing video traffic. The system consists of a web crawler engine which discovers and downloads monitored videos from the P2P systems as illustrated in Fig.I.17.

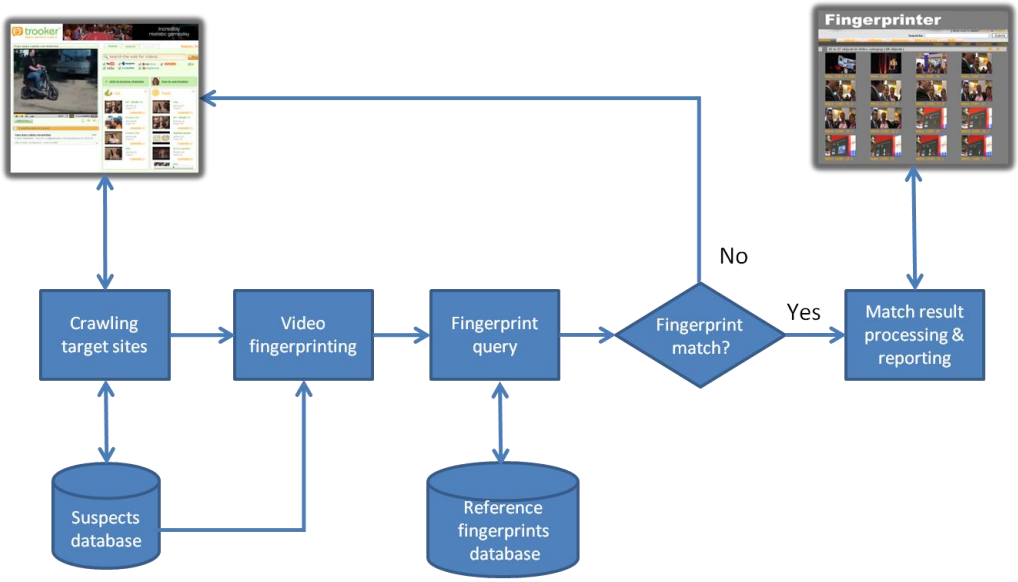


Fig. I.17: Video content tracking scenario

By matching the fingerprints of the monitored videos with those from the reference video database, the copyright infringement can be detected and legal action can be taken.

In video filtering and video tracking scenarios, the miss detections are highly costly for the copyright owners whereas the false alarms are less disturbing as they can be discarded by a human observer. However, the very low values for both false alarms and missed detections are very important.

Considering the distortions that can appear in these use cases they cover all the possibilities: video format, frame aspect and frame content modifications as detailed in Section I.3.5.

I.4.4 Digital watermarking

Fingerprinting can be used to prevent certain attacks against watermarking schemes. A well-known attack is the "copy attack" [OOS 01]: from the watermarked content an estimate of the embedded watermark is obtained. This estimate is subsequently embedded in another video content. Consequently, unauthorized users can create watermarked content. A method to prevent this is to embed content-dependent watermarks. The integrity of the watermark will be decided by matching

the fingerprint computed from the content with the watermark (which is the content dependent fingerprint).

In such a scenario, the miss detections have a higher negative impact on the system, than the false alarms, as when missing the replica, a pirated content can pass for a watermarked one. Therefore the fingerprinting system designed for such application has to feature a very strong robustness. Moreover the system has to be robust to the copy-attack and to all types of modifications detailed in Section I.3.5.

I.4.5 Broadcast monitoring

A broadcast monitoring application consists in tracking television or web video broadcasts [SEO 03]. In a broadcast monitoring application, video fingerprinting consists in computing the fingerprints of an interest broadcast channel and matching them to the reference database hence obtaining its playlist. Such an application enables program verification, ensures monetization of advertisements air-runs, and can provide audience measurement statistics.

Considering for instance, the video archives of a TV station, and the counting of a particular aired commercial in a month, the missed detections as well as the false alarms are equally important. The miss detections cause losing revenue to the TV station while the false alarms cause losing revenue for the commercial provider. Regarding the distortions for such a case, they are mostly related to the video format modification (detailed in Section I.3.4.3) and which are due to the storage requirements, *e.g.* compression.

I.4.6 Business analytics

With the enormous volume of multimedia content comes the great challenge of making it profitable through added value services. For instance, with the fast changing and diverse mix of broadcast platforms, accurate and reliable audience measurement services have become vital. Evaluating the media consumption, the user behavior and social reach can help understanding the multimedia market and therefore lead to successful business planning, decision making or brand management [VOB 12], [AUD 12].

Using the principles of video fingerprinting, multimedia can be identified and tracked during its consumption on the Internet or at TV. Consequently, related analytics information can be obtained and different bussiness models for multimedia monetization can be enabled.

Examples of existing analytics services are YouTube analytics - which provides the number of viewers, their location, their age segment, the engagement in the viewing experience, the popularity and Audible's CopySense - which tracks, audits and reports usage across the Internet, radio, TV, cable and satellite transmission. Regarding the monetization services, YouTube developed a content ID technology which manages to monetize a third of the interest video playbacks and advertisement on their portal [YOU 12].

I.5 State of the art

Due to its large applicability in various existing domains and its enabling potential for monetization and business intelligence applications, video fingerprinting got an increasing interest from both industry and academia. Consequently, the state of the art for video fingerprinting presents a dichotomy, on the one hand the industrial approach and on the other the academia approaches as presented n the sequel.

I.5.1 Industrial solutions

Companies such as Vobile [VOB 12] developed multimedia fingerprinting solutions like the Vobile Video Tracker (a SaaS - Software as a Service, *i.e.* the software and the associated data is centrally hosted on the cloud - application which allows content owners to monitor online sharing sites, to identify their content and to decide whether to allow it to remain on the site or to send a notice to the site operator asking for the copy to be taken down), vCloud9 (Cloud Based Content Identification and Management an application for content identification and management which enables the file-hosting services to eliminate unauthorized content, to assure storage efficiency by identifying duplicate content and to generate revenues by identifying premium content that can be legally monetized), the Media Tracker Analytics (an application which provides metrics on online audience viewing behavior for specific content).

Civolution [CIV 12] provides television (Teletrax Television Monitoring) and Internet **(**Teletrax Internet Monitoring) multimedia monitoring solutions which and are based on a combination of watermarking and fingerprinting technologies. Teletrax Television Monitoring enables clients such as entertainment studios, news and sport organizations, TV syndicators, and advertisers to determine when, where and how their video content is being used around the world (*e.g.* confirmation and prove of airing content). Civolution is currently monitoring over 1,500 television channels in more than 50 countries. Teletrax Internet Monitoring identifies controls and monetizes content as it travels around the Internet (peer-to-peer file sharing networks, video sharing and social media websites, live streaming sites, usenet newsgroups, chat rooms, forums and blogs).

With clients such as 20th Century Fox Studio, Disney NBC Universal, RTL Group, Canal Plus, Viacom, DailyMotion, Facebook and with more than 12 million music, movies and television fingerprints in their Global Rights Registry™ database, Audible Magic [AUD 12b] proposes a broad range of solutions based on fingerprinting. From broadcasting monitoring services for music and advertising content to recognition technology for cloud service operators, from social TV services (*e.g.* interactive advertising, social engagement) to audience measurement and from copyright compliance to HEOA compliance (*i.e.* Higher Education Opportunity Act is a law passed by the American Congress on August 14, 2008, one of the policies of the Act requires colleges and universities to mitigate the use of P2P networks to illegally upload or download copyrighted materials across campus networks).

IPharro, a Fraunhoher Institut spinoff ,proposes iPharro MediaSeeker Core Platform a solution for multimedia search applications, archive versioning, content future-proofing, media redundancy prevention, broadcast monitoring.

Zeitera offers cloud services for smartTV (allows SmartTV apps to interact with TV content, interactive advertising applications, targeted ads, coupon capabilities commercial monitoring/localization/replacement) and synchronous mobile applications (smartphone two screen interaction with video content immersive social networking applications, 2nd screen applications, direct check-in, program guide correlation with rich meta-data).

Technicolor [TEC 12] provides solutions that serve content enrichment purposes (*i.e.* the video fingerprint is used as an index to retrieve relevant information in a relational database *e.g.* the title of the movie, the names of the actors playing in it, *etc.*), localization of copyright content (*i.e.* a crawler browses the Internet and retrieves movie files which are then inspected with video fingerprinting to check whether they are copyrighted or not), data loss prevention (the video fingerprint are used to find out which and where movies are stored and processed in order to prevent unauthorized operations), copyright infringement prevention.

ZiuZ [ZIU 12] developed Twin Match, a video fingerprinting-based software which compares videos from confiscated material to previously classified material in order to eliminate from sharing sites videos featuring child pornography.

Vercury [VER 12], Advestigo [ADV 12], GraceNote [GRA 12] are other companies that provide similar with the above video fingerprinting solutions for diverse purposes.

I.5.2 Academic state of the art

The academic state of the art for video fingerprinting exhibits a large variety of methods. While in the case of the industrial video fingerprinting solutions their methodology is not available for investigation, the academic solutions are publicly available and can be consulted and compared.

Considering the three main blocks of a video fingerprinting system, presented in Fig.I.6, namely the localization, the fingerprint computation and the fingerprint matching a classification can be made on three criteria: the type of localization strategies, the type of features chosen as fingerprints and on the type of similarity metric employed between fingerprints.

In the sequel, Section I.5.2.1 – Section I.5.2.3, such a classification will be made in order to structure and discuss the existing approaches to video fingerprinting and to provide a global overview. After the synoptic classification, seven reference video fingerprinting methods were selected, briefly described and analytically compared in Section I.5.2.4.

I.5.2.1 State of the art for video fingerprinting: localization strategies

The first key component of a fingerprinting method is the localization of the query sequence in the reference video sequence. This aspect is very important because in the majority of the applications the typical query video sequence is a part of a reference video sequence.

The localization strategy can be strongly related to the video feature selected as fingerprint (*e.g.* a sliding window is a usual strategy for binary fingerprints) or it can be totally different (*e.g.* obtaining a few video candidates through an independent localization strategy and just then applying the fingerprinting algorithm).

[OOS 02] proposes an indexing look-up table for binary fingerprints containing all the possible fingerprints and the position in the reference videos where they occur. When a new query video sequence is searched, its fingerprint is computed and identified in the look-up table, hence localized in the reference video collection. However, such a strategy is just a pseudo-localization approach as the queries are considered of equal length as the sequences in the reference database. Note that the localization of different lengths query sequences within a reference sequence is not addressed.

In [COS 06] the search is enabled by sliding a frame window, with the same size as the query video clip throughout the reference sequences. The fingerprint of the sequence covered by the window is computed and matched to the reference fingerprints. A value-position matching strategy is advanced by [MUK 10] and consists in moving a sliding window with same length as the query sequence over the reference video sequence and counting the matching fingerprints at the corresponding positions in the window.

[SU 09] builds up on the sliding window idea and designs a coarse to fine sliding window, but improves it with a look up table and voting strategy. [RAD 08] proposes the division of video sequences into chunks and associates fingerprints with each chunk.

A temporal pyramid matching is proposed by [JIA 11] and consists in partitioning the videos into increasing finer temporal segments and in computing similarities over each granularity.

Through a k-nearest neighbor matching algorithm for interest points, [LAW 06] localizes similar frames from the query and the reference video sequence and then trough a voting function based on a label description of interest point motion, the matching video sequences are identified.

Another approach used by [IND 09] and [JIA 11] is the Locality Sensitive Hashing (LSH) which consists in indexing binary fingerprints in a high dimensional indexing data structure, [GIO 99]. The LSH indexes a bit string representing points in a high dimensional space. Given a query bit string example and some distance threshold *m*, the LSH returns a list of stored bit strings within Hamming distance *m* of query bit string.

A nearest neighbor search and mapping of each query frame to the closest reference frame is proposed by [FOU 11].

A different approach comes from [HIL 10] which proposes as localization strategy the linear fit filtering (RANSAC [FIS 81]) and the Bi-partite match filtering which filters out a list of candidate video sequences yielded from the fingerprint matching algorithm.

As a side note, in the video fingerprinting methods which use the audio component, [JIA 11], [MUK 10], the localization strategy generally starts with a localization based on audio: *e.g.* the WASF audio descriptor is computed and then searched for in the database in [JIA 11], or the spectrogram of the audio signal is computed, divided into small regions and then query regions are localized in the reference sequence.

As a conclusion, a myriad of localization strategies can be envisioned, depending on the feature chosen as fingerprint, on its mathematical formalization and its matching procedure. The most desirable localization strategies are those which have a low overall computational time and whose computation time in the online phase is independent with respect to the size of the testing database.

I.5.2.2 State of the art for video fingerprinting: features

The second key component of a video fingerprinting system is the fingerprint. The quality of the fingerprint and its properties depend on the fingerprint features selected from the video sequences. The types of features which were used in the state of the art as video fingerprints are presented in Table I.3.

The state of the art presents a dichotomy for the types of features: they can be computed only from the visual content (*i.e*. the case of *mono-modal methods*) or from visual and audio content (*i.e*. the case of *multi-modal methods*). Independently with respect to its type, the video fingerprint can be computed at different *granularity* levels, *e.g*. frames, keyframes, blocks or regions of frames, group of frames, points of interest.

According to the domain in which the fingerprints are computed, the group of mono-modal methods can be of four types: spatial, temporal, transform and color.

The *spatial fingerprints* computed on blocks, regions of frames or whole frames are robust to non-geometric distortions, but they lack in robustness against geometric modifications (*e.g*. cropping, rotations). The interest points based features have a high robustness against the geometric distortions and transcoding transformations but lack in resilience against changes in color, illumination and filtering. Moreover, this type of features poses problems of uniqueness in the case of very similar video sequences, (*e.g*. TV news) therefore needs to be used in combination with other features.

The category of *temporal fingerprints* is generally robust to global changes in the quality of the video like non-geometric modifications of the frame aspect and they can resist several encoding (*e.g*. MPEG compression), but they are generally sensitive to distortions affecting the video format (*e.g*. frame-rate changes, frame-dropping, transcoding) and to geometric modifications.

Transform based fingerprints ensure robustness to geometric and non-geometric frame aspect modifications and to video format modification but are sensitive to modifications of video content such as cropping and content addition.

The *color based* category of fingerprints lacks resilience to global variations in color and illumination but can be used along with other features in order to enhance discriminability.

Types of fingerprints		Granularity	Fingerprint examples
Mono-modal methods (Video content features)	Spatial	Blocks, regions of frames, frames, keyframes	▪ visual attention regions, [SU 09] ▪ ordinal ranking of average gray level of frame blocks, [HAM 02], [KIM 09] ▪ quantized block motion vectors of frames, [HAM 02] ▪ invariant moments of frames edge representation, [HU 62] ▪ centroid of gradient orientations, [LEE 08] ▪ dominant edge orientation, [HAM 01]
		Points of interest	▪ signal description of motion of interest points (corner features, Harris points), across videos, [LAW 06], [LAW 07], [JOL 05] ▪ scale-space features (*e.g.* SIFT), [SAR 08] ▪ descriptors of interest points, [MAS 06]
	Temporal	Group of frames	▪ differential block luminance features between consecutive frames, [OOS 02] ▪ temporal ordinal measure (ordering of intensity blocks in successive frames depending on their average intensity), [CHE 08], [HUA 04], [KIM 05], [HAM 01] ▪ ordinal histogram over the frames of the entire video, [SAR 08], [YUA 04] ▪ pixel differences between consecutive frames, [HAM 01] ▪ shot duration sequence, [IND 99]
		Down-sampled frames	
		Keyframes	
		Every frame	
	2D/3D Transform	GOP	▪ quantized compact Fourier-Mellin transform coefficients of keyframes, [SAR 08] ▪ subspace embedding using the singular value decomposition [RAD 08] ▪ the signs of DCT coefficients of keyframes, [ARN 09] ▪ the averages of DC coefficients blocks of I frames [YAN 08] ▪ 3D DCT coefficients of sub-sampled keyframes, [COS 06] ▪ DCT coefficients of the radial projection vector of the keyframes pixels, [ROO 05] ▪ 2D wavelet transform, [GAR 11a], [GAR 11b], [GAR 12], [DUT 10]
		Re-sampled video	
		Frame transform	

	Color	Histogram based	▪ YUV histograms of the DC sequence of MPEG videos [NAP 00], [HAM 07] ▪ YCbCr histogram of a group of frames, [SAR 08] ▪ color moment representation, [GAU 01] ▪ RGB, HSV histogram of frames, [HAM 01] ▪ the principal component of the color histograms of keyframes, [SAN 99]
Multi-modal methods (Video and Audio features)	Combined	Combined approaches	▪ SIFT, GIST and color correlogram features for keyframes, [HIL 10] ▪ global visual feature (DCT), local visual feature (SIFT, SURF), audio feature (WASF, modified MPEG-7 descriptor ASF), [GAO 10] ▪ visual feature: center-symmetric local binary pattern (CS-LBP), hamming embedding; audio feature: filter banks, [JEG 10], [AYA 11] ▪ coarsely quantized area matching – visual feature, divide and locate – audio feature, [FOU 11], [MUK 11] ▪ cascade of multimodal features (Dense Color SIFT, Bag of Words, DCT, WASF) and temporal pyramid matching, [JIA 11]

Table I.3: Types of video fingerprints

As explained above, the mono-modal methods employ a reduced number of visual features as fingerprints in order to identify the limitations that they pose and their possible applications. The multi-modal types of fingerprints combine the advantages of video and audio features of videos can achieve better results with faster computation time than the mono-modal methods.

The frequent disadvantage of the multi-modal types of fingerprints is their excessive number of computed features, which leads to redundant video information used as fingerprint (*e.g.* [GAO 10] using SIFT and SURF features simultaneously). As the computational resources increase steadily due to technological development, extra computation is not considered a prohibitive factor. However, a clear mathematical ground for video fingerprinting should not be ignored.

I.5.2.3 State of the art for video fingerprinting: similarity measures

The third key aspect of a video fingerprinting system is the matching between the fingerprints. The matching can be achieved by employing a similarity metric adapted to the feature chosen as fingerprint and to the distortions envisioned.

According to the similarity distance employed for matching, the fingerprinting methods can be divided in two categories, distance based and probability based, as illustrated in Table I.5.

Types of similarity measures	Similarity measure	Applicability
Distance based	L1 distance (Manhattan)	▪ non-binary fingerprints, [HAM 01]
	L2 (Euclidian) distances	▪ non-binary fingerprints, [LEE 08]
	Hamming distance	▪ binary fingerprints [COS 06], [SU 09], [OOS 02]
	Haussdorff distance	▪ edge points based fingerprints [HAM 01]
	Normalized histogram intersection	▪ histogram based fingerprints [HAM 01]
	Normalized correlation coefficient	▪ histogram of block motion vectors, [HAM 02]
	k-nn, voting function	▪ interest point-based fingerprints, [LAW 06], [LAW 07], [JOL 07]
Probability based	Based on statistical tests	▪ hypothesis testing, multivariate Wald-Wolforwitz, [DUT 10] ▪ Rho test on correlation, [GAR 11a], [GAR 11b]

Table I.5: Types of similarity measures

As it can be observed, a multitude of similarity measures are available, depending on the selected feature. The distance-based group of methods has the advantage of allowing a decision based on an experimentally determined threshold. While they are easier to use due to their immediate empiric observation, they don't permit in the majority of cases a decision based on a mathematical ground. Therefore the alternative is the probability-based similarity measures which can grant a statistical rule for decision.

The desideratum for a similarity measure under a fingerprinting framework is that it does not depend on an empirical threshold but on a rigorous mathematical decision rule which can handle any content, distortion or use case particularity.

I.5.2.4 State of the art for video fingerprinting: representative methods

In this section, seven reference video fingerprinting methods developed in university labs have been searched, briefly presented and their performances analytically compared. Note that in the next Sections 1.5.2.4.a - 1.5.2.4.f the performances of the presented methods are not discussed individually but comparatively at the end of Section I.5.2, in Table.I.6.

I.5.2.4.a The 3D-DCT method

In [COS 06], Coskun *et al* propose as fingerprints 64 quantized low-pass coefficients resulting from a 3D Discrete Cosine Transformation (DCT) applied on the luminance component of the spatio-temporal normalized video sequence, Fig.I.16.

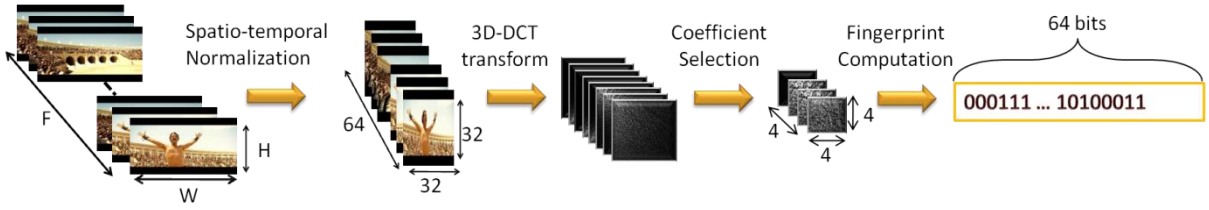


Fig.I.16: 3D transform fingerprinting principle

The 3D-DCT transformation is employed for its high energy compaction property and its low frequency coefficients' insensitivity to minor spatial and temporal perturbations. As an alternative, for enforced security, the 3D Random Bases Transform, (*i.e.* the calculation of the coefficients is made secret by involving a secret key within the cosine transforms bases) was proposed. Fingerprints are matched by using the Hamming distance. If the Hamming distance between two fingerprints is below a certain threshold, the videos are declared as identical; if the distance is above the threshold, the videos are declared as different. The threshold is computed based on the statistical properties of the Hamming distance and depending on the length of the fingerprint. Concerning the query localization procedure, the paper proposes a sliding window of the same size as the query's length, which moves throughout the longer reference sequence and matches the query hashes and the reference hashes under the sliding window.

I.5.2.4.b The visual attention regions method

In [SU 09], Su *et al.* propose a fingerprint extracted from the visual attention regions of sampled frames in the video. According to [KOCH 85], attention is implemented in the form of a spatially circumscribed region of the visual field, the so called focus of attention. This focus scans the image in two ways: a rapid bottom-up, saliency driven manner and afterwards in a task dependent, object driven manner. The saliency based visual attention model proposes an algorithm for computing the master saliency map which implements the bottom-up mechanism of attention by topographically coding the local conspicuity over the entire visual scene. The saliency map represents the saliency of every location in the visual field by a scalar quantity and guides the selection of attended locations based on the spatial distribution of the saliency. The robustness of the method relies on the fact that visual attention regions are invariant if the distorted video is content-preserving, even under heavy distortions.

The fingerprint consists of a sequence of bits obtained by quantizing a saliency map which holds the visual attention regions; the saliency map is computed by combining intensity, color and orientation maps for sampled and normalized frames, Fig.I18.

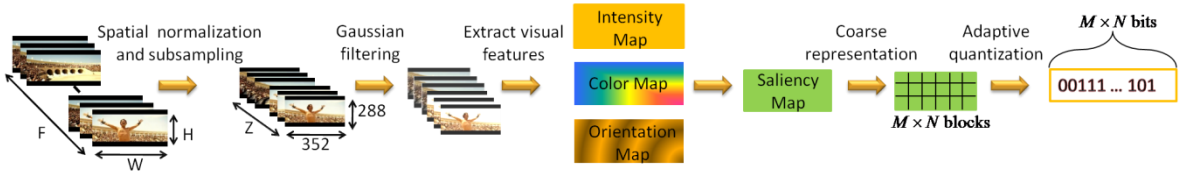


Fig.I.18: Visual attention regions fingerprint computation

The matching procedure follows the same concept as 3D-DCT [COS 06] method, namely the Hamming distance as a distance measure and the statistically obtained threshold for establishing the perceptual similarity of two videos.

I.5.2.4.c The differential block luminance method

In [OOS 02], Oostveen *et al.* propose a 32-bit fingerprint obtained by quantizing the values obtained by differencing the mean values of neighboring blocks of luminance inside a frame and by differencing the mean values luminance blocks in subsequent frame, Fig.I.19.

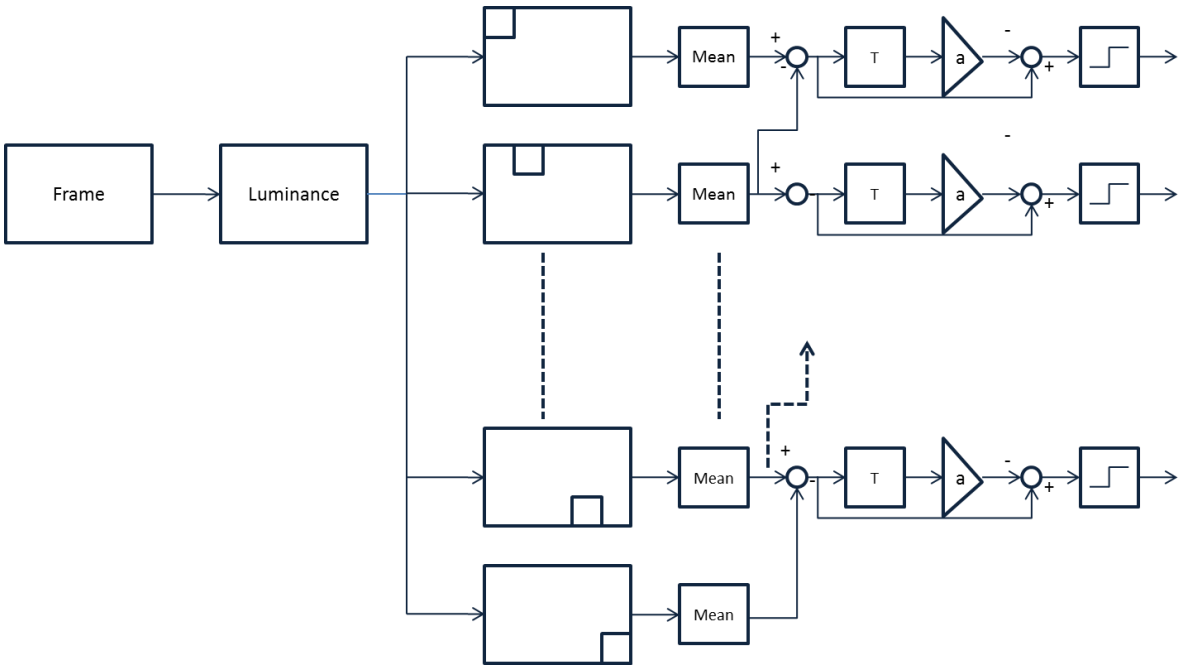


Fig.I.19: Block diagram of the differential block luminance algorithm

The matching of fingerprints is done based on the Hamming distance between the fingerprint of an original video and the fingerprint of its processed version. For localization, the authors propose a look-up table for all possible 32-bit fingerprints, Fig.I.20 The entries in the table point to the video clip and to the the positions within that clip where this 32-bit word occurs.

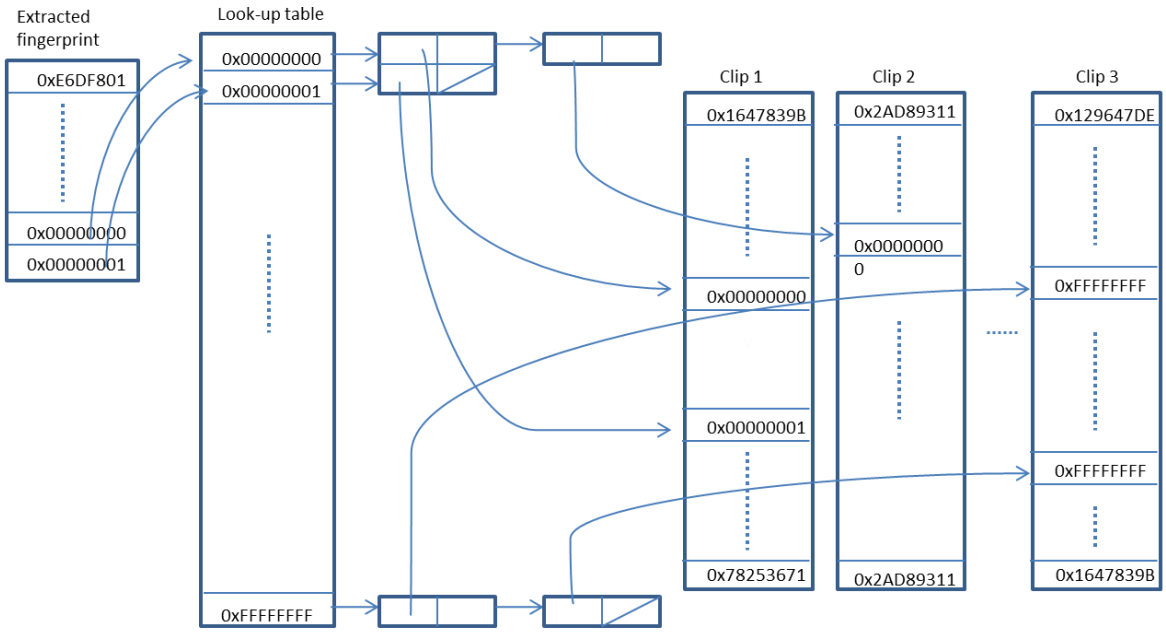


Fig.I.20: Look up table for database efficiency

I.5.2.4.d The point of interest behavior method

In [LAW 06] and [LAW 07], Law-To *et. al* propose a fingerprinting scheme based on the description of motion of interest points across videos. The method is concept wise similar to [JOLY 03] with the differences that it is faster and developed in such a manner that it allows changing its parameters depending on the constraints of the desired application.

The fingerprint consists of a description at three levels of the video content: the low level consists 20-dimensional descriptors of Harris interest points in every frame; the mid level description consists of the trajectories of the Harris interest points across the video; the high level description) consists of the labels attached to the trajectories defined.

The Harris interest points were used due to their local uniqueness and their high information content. Every Harris point is assigned a descriptor: $\vec{S} = (\frac{\vec{s}_1}{\|\vec{s}_1\|}, \frac{\vec{s}_2}{\|\vec{s}_2\|}, \frac{\vec{s}_3}{\|\vec{s}_3\|}, \frac{\vec{s}_4}{\|\vec{s}_4\|})$ where $\vec{s}_i$ is a differential decomposition of the gray level signal $I(x,y)$, $\vec{s}_i = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial^2 I}{\partial^2 x \partial y}, \frac{\partial^2 I}{\partial x^2}, \frac{\partial^2 I}{\partial y^2})$. By thresholding the L_2 distances between descriptors of consecutive and subsequent frames the motion

trajectories of the interest points are built as illustrated in Fig.I.21. For each trajectory, the following salient characteristics are considered:

- the mean descriptor (defined as the average of each component of the local descriptors) is associated;
- the average position along the trajectory μ_x, μ_y ;
- the time codes of the beginning and the end tc_{in}, tc_{out} ;
- the variation of the position $\left[x^{\min}, x^{\max}\right]$, $\left[y^{\min}, y^{\max}\right]$ are retained.

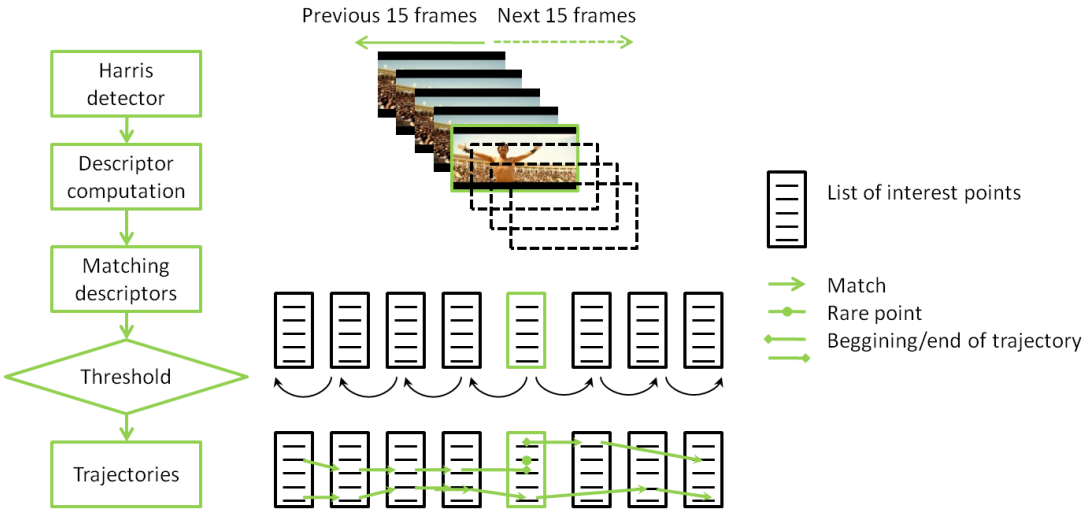


Fig.I.21: Algorithm for points of interest trajectory estimation

Disposing of the previous description for the interest points, it is possible to describe their behavior and to attach a label accordingly: moving points/motionless points, persistent points/rare points, fast motion/low motion points, horizontal motion/ vertical motion.

Due to the underlying properties of the proposed fingerprints the matching and the localization procedure is done in the same time, at three levels, Fig.I.22. First a k-nearest neighbor search is performed for the interest points of the query video frames which are described by their descriptors; potential matches are found among the interest points of the reference video frames. Second, distances between the trajectories of interest points in the query and reference video sequences are computed and most similar trajectories are selected. Third and last, the final decision on the similarity of a query video with a reference video is taken by comparing the associated labels.

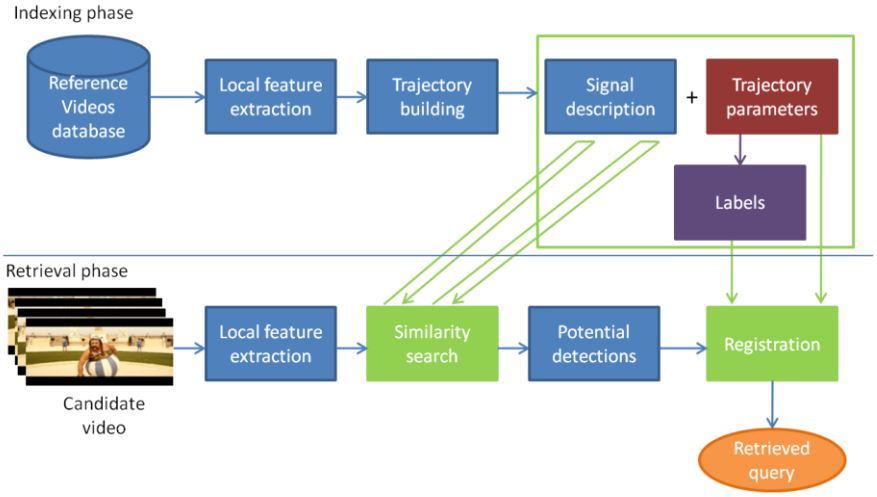


Fig.I.22: Video copy detection framework in [LAW 07]

I.5.2.4.e The centroid of gradient orientations method

In [LEE 08], Lee *et. al* provide a fingerprinting method based on the centroid of gradient orientation (CGO) detailed further in Fig.I.23. The CGO provides a measure of variation of the luminance throughout the frame. The CGO is computed based on the gradients related to the distribution of edges in the frame which provide relevant information about the visual content in the frame. As the gradients are not based on the pixel values but on their differences, the proposed fingerprint is automatically robust against global changes in pixel intensities such as brightness, color, contrast.

The video is sampled at a fixed frame rate in order to cope with the frame rate change attacks. The sampled frames are converted to grayscale to make the method robust against color variation and applicable to black and white films as well. Each grayscale frame is resized to a fixed format in order to assure the robustness in case of a resizing processing. The fingerprint of the videos consists of the CGO computed on blocks of the sampled frames.

The CGO is obtained as a sum of the gradient of the luminance of every pixel in the block, according to the following formula:

$$c[n,m,k]=\frac{\sum_{(x,y)\in B_{n,m,k}} r[x,y,k]\theta[x,y,k]}{\sum_{(x,y)\in B_{n,m,k}} r[x,y,k]}$$

where $B_{n,m,k}$ is the block in the nth row and mth column of the kth frame and $c[n,m,k]$ is the centroid obtained from the $B_{n,m,k}$ $(1\le n\le N, 1\le m\le M)$ block. $r[x,y,k]$ and $\theta[x,y,k]$ are respectively the magnitude and orientation of the gradient vector $\nabla f=[G_x\ G_y]=\left[\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right]$ which is the luminance value at location (x,y).

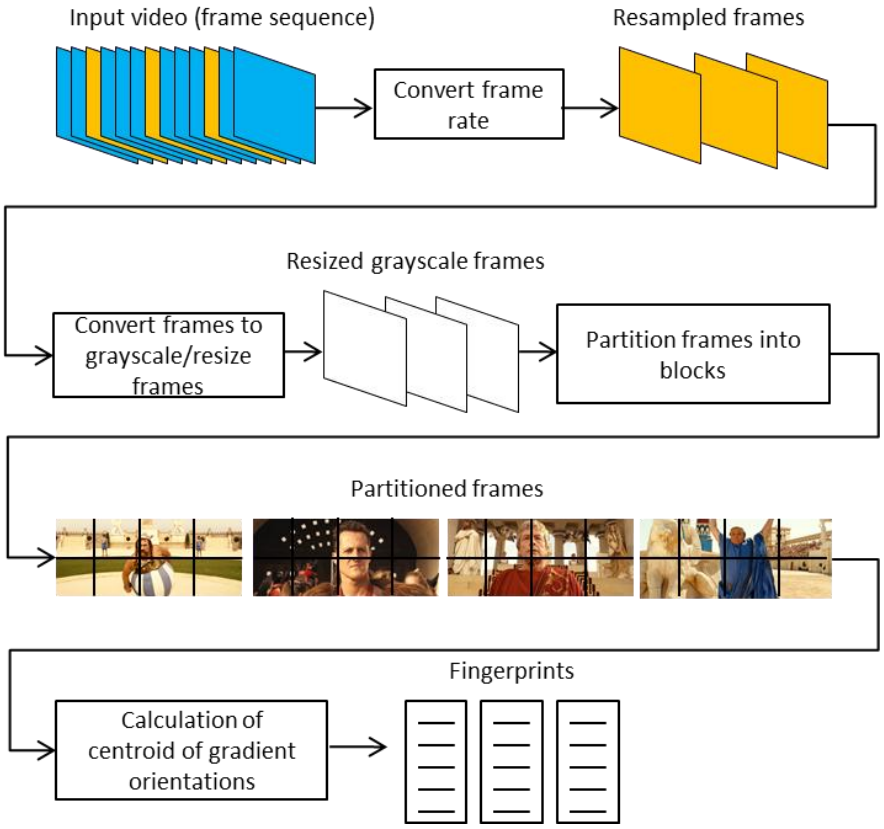


Fig.I.23: Centroid of gradient orientations fingerprint computation

The fingerprint matching is done by using the squared Euclidian distance as similarity measure. The decision threshold is statistically established by considering the fingerprints as realizations of stationary ergodic processes and by considering the central limit theory in order to minimize the false alarm probability.

Concerning the query localization, the proposed approach is the range search: the fingerprint of a query frame is searched in the database of reference frames fingerprints. Upon matching of frame fingerprints, the fingerprints for the videos sequences are further matched in order to ensure a correct decision.

I.5.2.4.f Shot duration method

In [IND 99], Indyk *et al.* exploit the temporal dimension of the videos and propose a fingerprinting method based on shot durations *i.e.* the fingerprint consists of the timing patterns of when shots change in videos. As illustrated in Fig.I.24, a shot transition algorithm [GAR 98] is run on a video v and its timing sequences $T(v)=[t_1,t_2,...,t_n]$ is obtained; t_i $(1\le i\le n)$ denotes the time (in seconds) at which the i^{th} shot transitions occurs in v. The timing sequence is further divided into segments of the timing sequence, of k seconds each, in order to be able to identify smaller parts of the video.

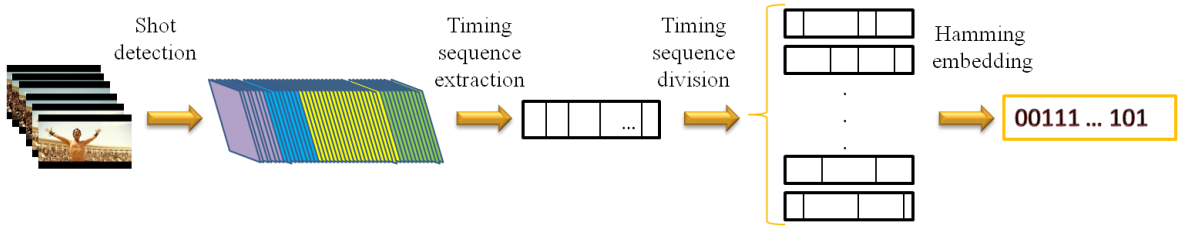


Fig.I.24: Basic principle of timing segments extraction

The similarity between the segments is computed based on a defined fuzzy distance measure which is a hybrid between the L_1 measure and the Hamming distance notions, for a certain integer $a \geq 1$.

$$dist_a(v,w) = \left(\sum_{t \in T(v)} \min_{t' \in T(v)} E_a(t,t') \right) + \left(\sum_{t' \in T(w)} \min_{t \in T()} E_a(t,t') \right)$$

where $E_a(t,t') = \begin{cases} |t-t'|/a & if\ |t-t'| \leq a \\ 1 & otherwise \end{cases}$ and t and t' are the transitions in the timing sequence of videos v and w.

The integer a, called the fuzzyfication window, together with other parameters such as the time division interval for the timing segments and the C threshold (*i.e.* the number of timing segments that need to be matching in order to declare that two videos are similar) are parameters proposed and tested by the authors depending on the application desired.

The similarity of two videos is finally decided based on the number C of matching timing segments.
As search strategy, a high dimensional indexing data structure, namely the LSH (Locality Sensitive Hashing), [GIO 99] is used for the construction of the fingerprints database. The LSH structure indexes a bit string representing points in a high dimensional space. Given a query bit string example and some distance threshold *m*, the LSH returns a list of stired bit strings within Hamming distance *m* of query bit string.

I.5.2.4.g Cascade of multimodal features with temporal pyramid matching

In order to provide a comprehensive overview on the state of the art for video fingerprinting methods, a multi modal fingerprinting method is also presented in the sequel.

The study in [JIA 11] exploits the audio and video components of a video sequence and proposes the use of several features: a local visual feature – the Dense Color SIFT [LOW 04], a global visual feature - the DCT and an audio feature - the WASP.
The architecture of their system is presented in Fig.I.25.

In the pre-processing part, the query and reference audio and video components are split: the video is uniformly sampled, and the audio content is divided in audio frames. Additional blocks for picture in picture and flipping distortions are also included. The multimodal features used as fingerprints are: a local visual feature of Dense Color SIFT [LOW 04], a global visual feature based on the DCT and an audio feature, WASP [CHE 08].

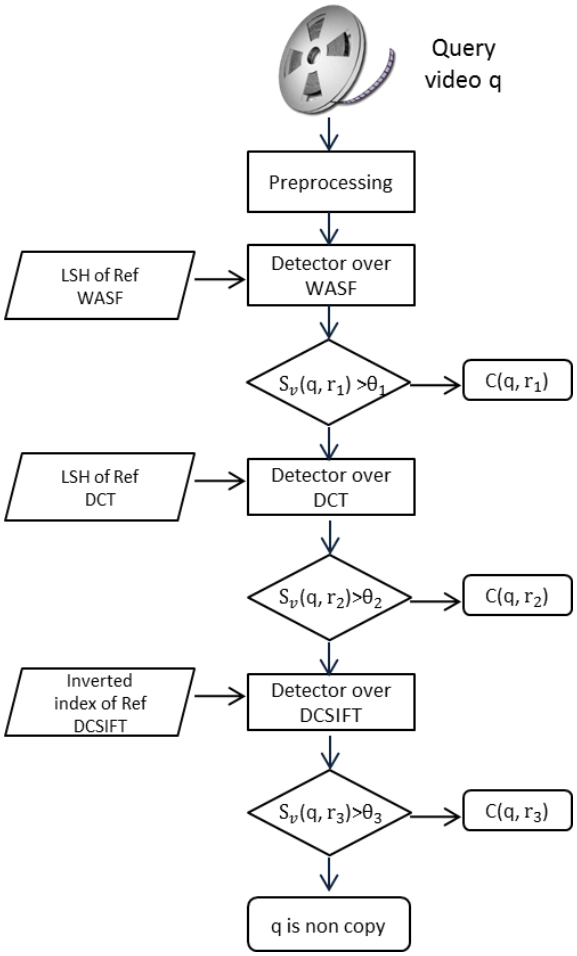


Fig.I.25: Video copy detection approach proposed in [JIA 11]

The dense color version of the SIFT descriptor is employed to cope with spatial content altering such as simulated camcording, picture in picture, pattern insertion, and postproduction. DCSIFT differs from SIFT in that there is no interest point/keypoint detection and localization, instead regular grids with overlapping are used for the 216 - dimensional descriptor construction. The Bag of Words framework proposed by [SIV 03] is used for transforming the reference DCSIFT feature vectors in visual words, Fig.I.26. An 800 words, visual vocabulary is created by using a k-means algorithm then quantized and stored in an inverted index together with the position of the keypoints.

The DCT global image feature is based on the relationship between the DCT coefficients of adjacent image blocks, of the Y component in the YUV color space. Such a feature was used due to its robustness to content-preserving transformations such as transcoding, change of gamma, decrease of quality (blurr, frame dropping, contrast, compression, noise). In order to speed up feature matching, the DCT features are indexed by Locality Sensitive Hashing.

For the audio feature the Weighted Audio Spectrum Flatness (WASF) [CHE 08] was used and a 14-D single WASF feature was considered; WASF extends the MPEG-7 Audio Spectrum Flatness (ASF) audio descriptor by introducing Human Auditory System functions to weight audio data. For the audio

fingerprints matching, the Euclidian distance is adopted and the LSH is used for efficient feature matching.

Concerning the matching of the video sequences, a cascade architecture is used: the query is first processed by the WASF detector; if the matching is positive (*i.e*. the query contains a copy clip) it leads to immediate acceptance, while a negative result triggers the evaluation of the second DCT detector. If the copy is asserted as a non-copy again by the DCT detector, it will be passed to the last DCSIFT detector.

Although the frames of two matched video sequences should have consistent timestamps, a certain extent of freedom is required due to the temporal distortions therefore a temporal pyramid matching algorithm (TPM), [LAZ 06], [GRA 05], [LIU 10] (TPM) is used in the query localization part. The TPM consist in the partitioning of the videos into increasing finer temporal segments and compute video similarities over each granularity.

The performances of the method have been tested under the TRECVID platform, [TRE 12] in the Content-Copy Detection 2011 and proved to achieve the best results compared to other participating systems.

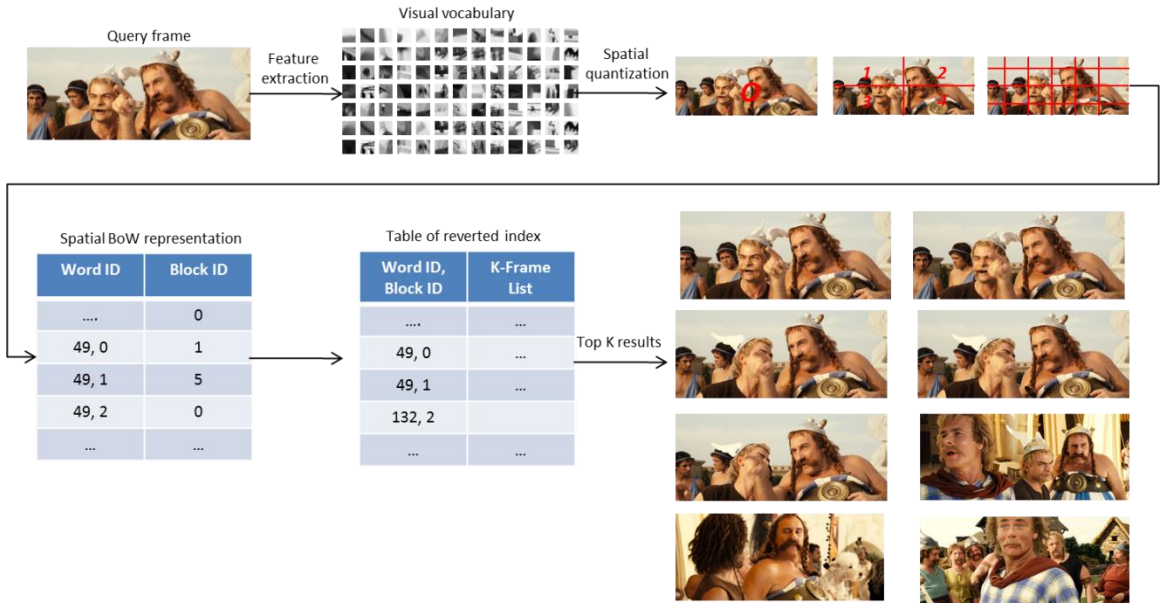


Fig.I.25: Keyframe retrieval using the inverted index of DCSIFT visual words and spatial information in [JIA 11]

In the context of the TRECVID platform the performance evaluation is done with three indicators, the normalized detection cost rate (NDCR) which is a cost function taking into account the incidence of false alarms and missed detection by assigning corresponding costs, the F1 measure combining precision and recall and the mean processing time per query.

I.5.2.4.h Comparative view of the academic state of the art methods

Table I.6 offers a comparative view of the academic state of the art for video fingerprinting methods detailed in Sections I.5.2.4 a –f.

Note that the video fingerprinting method presented in Section I.5.2.4.g has not been included in the critical review in Table I.6 due to the fact that it has a multi-modal approach and therefore does not fit the comparison.

Method	Corpus	Evaluation metrics and performances	Conclusions
3D-DCT [COS 06]	**Reference database**: 244 video clip database, with lengths between 12-300 seconds **Queries**: 61 sequences of 14 seconds of video	• Ratio of correct detections to total tests between 83% and 100%, when the $P_{fa} = P_{md}$	• Robustness to: blurring, white Gaussian noise addition, brightness increase/decrease, MPEG-4 compression, lossy channel, fade-over, time clipping, frame rotation (up to 3°), frame dropping, frame rate changes • No scalable localization procedure
VAR [SU 09]	**Reference database**: 100 hours consisting of video clips from the Internet, TV, broadcasting, DVD **Queries**: N/A	• Probability of false alarm = probability of missed detection = 0.086	• Robustness to: resolution reduction to CIF, QCIF format, logo overlay (by 5%-30%), frame-rate changes conversion to gray • No scalable localization procedure
Differential block luminance [OOS 02]	**Reference database**: 10 seconds video clips from movies and television broadcasts, at a resolution 480 x 720 **Queries**: the replica versions of reference video sequences	• Bit error rate between binary fingerprints 2.9% and 30%	• Robustness to: MPEG compression, median filtering, histogram equalization; not robust to scaling or shifting operations • No localization procedure

Table I.6: A comparative view on the performances of the state of the art video fingerprinting systems

Method	Corpus	Evaluation metrics and performances	Conclusions
Interest points behavior [LAW 07]	**Reference database**: 3 hours of TV content stored as MPEG-1, 25fps, at a resolution 352 x 288 **Queries**: 69 replica versions of the reference video sequences with lengths between 1 second – 30 minutes	• The recall = 0.82 for a precision = 0.95 • Computational time= 27 seconds for a query of 15 mins to be searched in a 3 hours reference database	• Robustness to: geometric modifications: cropping, resizing, shifting, and encrusting, image translation and to affine illumination changes • A basic localization is available but its computational complexity is not assesed
CGO [LEE 08]	**Reference database**: 1) 300 movies totalizing 590 hours (for testing the uniqueness) 2) 50 videos totalizing 100 hours (for testing the robustness) **Queries:** replica version of 10 seconds video sequences selected from the reference sequences 2)	The probability of missed detection = 0.0246	• Robustness to: non geometric distortions: lossy compression, global change in color, brightness, gamma, Gaussian blur, additive noise and combinations of the above; frame rate changes, frame dropping up to 70% of frames lost, robust to rotations up to 1° and cropping which retains 80% of the central portion of the frame • No localization procedure
Shot duration [IND 99]	**Reference database**: 2000 MPEG clips from Internet with lengths between 2-5 minutes and 5 movie trailers **Queries**: a selection of a few sequences from the reference database and their frame-rate was changes	Processing time per query: between 10 seconds and 25 minutes	• Robustness to: translations in time and jittered video sequences • Limited applicability for news programs as they have predefined formats which leads to similar timing sequences. • No localization procedure

Table I.6 (continued): A comparative view on the performances of the state of the art video fingerprinting systems

I.6 Conclusion

As it can be observed from the synoptic review presented in Section I.4 the applicability field of video fingerprinting grew steadily in the last decade. To answer the need for efficient video fingerprinting systems, a lot of research has been done in the industrial and university sector as presented in Section I.5. In Table I.6, a comparative analysis between the performances of state of the art video fingerprinting methods is presented.

Despite the wide range of methods that have been investigated, limitations are identified and challenges are still to be taken considering video fingerprinting systems as formulated next and synthetically organized in Table I.7

- The uniqueness property of fingerprints is not granted by a mathematical ground *i.e*. the features which represent the visual content are not selected according to a comprehensive mathematical approach.
- The robustness property of fingerprints is based on partial mathematical models without a general framework able to address the wide variety of existing distortions (*i.e.* the methods presented do not feature robustness to video format, frame aspect and frame content distortions at the same time). Secondly, the academic state of the art methods presented in Section I.5.2 have not addressed yet, at our best knowledge, the challenging case of live camcorder recording. Thirdly, the methods are generally tested on TV content data sets and don't take into account the particularities of the cinema content. These particularities are twofold and refer to the types of visual content and to the types of distortions that need to be addressed by the fingerprinting method.
- The uniqueness and the robustness properties are never object to a joint optimization strategy. Consequently, the trade-off among the probability of false alarm, the probability of missed detection, the precision, the recall and the computational time required by such a use case has not yet been investigated.
- In general the state of the art video fingerprinting methods do not have query localization support able to result in scalable solutions for large scale databases.
- The video fingerprinting methods are experimentally validated on relatively reduced video collections.

Constraints	Challenge	Current limitation
Uniqueness	Accurate representation of visual content	Heuristic procedures
Robustness	Mathematical ground In-theater live camcorder recording	Heuristic procedures No related method reported in the state-of-the-art
Search efficiency	Scalability	Very few, full scalable, mono-modal methods reported in the state-of-the-art

Table I.7 The constraints, challenges and current limitations for state of the art video fingerprinting systems

References

[AUD 12a] http://www.auditoire.com/web/

[AUD 12b] www.audiblemagic.com

[AUD 12c] http://www.audiblemagic.com/white-papers/Digital_Fingerprinting_Enables_New_Forms_of_Interactive_Advertising.pdf

[ADV 12] http://www.advestigo.com/

[ARN 09] Arnia, F., Munadi, K., Fujiyoshi, M., Kiya, H., "Efficient content-based copy detection using signs of DCT coefficient," *IEEE Symposium on Industrial Electronics & Applications*, Issue 4-6 Oct. 2009, pp. 494 – 499, 2009

[AYA 11] Ayari, M., Delhumeau, J., Douze, M., Jégou, H., Potapov, D., Revaud, J., Yuan, J., Schmid, C., "INRIA@TRECVID'2011:Copy DetectionMultimedia event Detection," *in Proceedings of TRECVID 2011*

[BAR 10] Barrios, J. M., Bustos, B., "Content-Based Video Copy Detection: PRISMA at TR*ECVID 2010,*" *in Proceedings* of *TRECVID 2010*

[BHA 96] Bhat, D. N., Nayar, S. K., "Ordinal Measures for Visual Correspondence," *in Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1996.

[BEE 12] http://www.beeldengeluid.nl/

[BUC 99] Buccigrossi, R., Simoncelli ,E., "Image Compression via Joint Statistical Characterization in the Wavelet Domain," *IEEE Transactions. on Image Processing*, Vol.8, No.12, 1999, pp. 1688 - 1700.

[COX 08] Cox, J. I., Miller, L. M, Bloom, A. J., Fridrich, J., Kalker, T., "Digital Watermarking and Steganography", *Second edition, Morgan Kaufmann*, Burlington, MA, 2008.

[COS 06] Coskun, B., Sankur, B., Memon, N., "Spatio-temporal transform based video hashing," *IEEE Transactions on Multimedia*, Vol. 8, No. 6, 2006.

[CHE 08] Chen, L., Stentiford,F. W. M., "Video sequence matching based on temporal ordinal measurement," *Pattern Recognition Letters*, Vol. 29 , Issue 13, 2008, pp. 1824 – 1831.

[CHE 08] Chen, J., Huang, T. "A Robust Feature Extraction Algorithm for Audio Fingerprinting," *PCM'08*, pp. 887-890, December 9-13, 2008

[CHU 08] Chupeau, B., Massoudi, A., Lefèbvre F., "In-theater piracy: Finding where the pirate was", *SPIE'08, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, 2008.

[CIV 12] http://www.civolution.com/

[DUM 08] Dumitru, O., Mitrea, M., Prêteux, F., "Video Modelling in the DWT domain," *Proceedings*. *SPIE*, Vol. 7000, 2008, Strasbourg, pp. 7000 OP: 1-12.

[DUM 10] Dumitru, O., Mitrea, M., Prêteux, F "Noise sources in robust uncompressed video watermarking", PhD thesis, Universite Pierre et Marie Curie, Paris, 2010

[DUT 10] Dutta, D, Saha, S.K., Chanda, B., "A hypothesis test based robust technique for video sequence matching," *International Journal of Future Generation Communication and Networking*, Vol. 3, No.3, 2010.

[ENV 11] http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf

[FFM 12] http://ffmpeg.org/

[FIS 81] Fischler, M., Bolles, R., "Random sample consensus… ," *Communications of the ACM*, 24(6), 1981.

[FOU 11] Foucher, S., Lalonde, M. Gupta, V., Darvish, P., Gagnon L., Boulianne, G., "CRIM Notebook Paper - TRECVID 2011 Surveillance Event Detection", *Proceedings of TRECVID 2011*.

[GAO 10] Gao, W., Huang, T., Tian, Y., Wang Y., Li, Y., Mou, L., Su, C., Jiang, M., Fang, X., Qian, M., "PKU-IDM@TRECVID-CCD 2010: Copy Detection with Visual-Audio Feature Fusion and Sequential Pyramid Matching", Proceedings of *TRECVID 2010*.

[GAR 12] http://www.gartner.com/it/page.jsp?id=1924314

[GAR 11a] Garboan, A., Mitrea, M., Prêteux, F., "DWT-based Robust Video Fingerprinting", Proceedings for the "*3rd European Workshop on Visual Information Processing*" (EUVIP), 2011, Paris, pp. 216 - 221.

[GAR 11b] Garboan, A., Mitrea, M., Prêteux, F., "Video retrieval by means of robust fingerprinting", Proceedings for the "*15th IEEE Symposium on Consumer Electronics*" (ISCE), 2011, Singapore, pp. 299 - 303.

[GAU 04] Gauch, J. M, "Real-time feature-based video stream validation and distortion analysis system using color moments", United States Patent 6246803.

[GIO 99] Gionis, A., Indyk. P, Motwani., R., "Similarity Search in High Dimensions via Hashing", Proc. 25th VLBD Conference Edinburgh, Scotland 1999

[GRA 12] http://www.gracenote.com/

[GRA 05] K. Grauman, and T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Fea-tures", IEEE ICCV'05, pp. 1458-1465, October 17-21, 2005.

[GOO 12] http://images.google.com/

[HAM 01] Hampapur, A., Bolle, R.M, "Comparison of distance measures for video copy detection", IBM TJ Watson Research Center, *IEEE International Conference on Multimedia and Expo*, 2001, pp. 737 - 740.

[HAM 02] Hampapur, A., Hyun, K-H., Bolle, R., "Comparison of Sequence Matching Techniques for video copy detection", in Proceedings of *Storage and Retrieval for Media Databases* (San Jose, USA, Jan. 20-25, 2002), pp: 194-201.

[HIL 10] Hill, M., Hua, G., Natsev, A., Smith, J.R., Xie, L., Huang, B., Merler M., Ouyang, H., Zhou M., "IBM Research TRECVID-2010 Video Copy Detection and Multimedia Event Detection System", Proceedings of *TRECVID 2010*.

[HUA 04] Hua, X.-S., Chen, X., Zhang, H.-J., 2004. "Robust video signature based on ordinal measure", in: Proceedings of the. *IEEE International. Conference on Image Processing* (ICIP), 2004, Vol. 1, 24–27, 2004, pp. 685–688.

[HU 62] Hu, M.K, "Visual pattern recognition by moment invariants", *Transactions on Information Theory*, Vol. IT-8, pp: 179–187, 1962.

[IPH 12] http://www.ipharro.com/news.html

[INA 12] http://www.institut-national-audiovisuel.fr/

[IND 99] Indyk, P., Iyengar, G., Shivakumar, N., "Finding Pirated Video Sequences on the Internet", Stanford Infolab, 1999.

[JEG 10] Jégou, H., Gros, P., Douze, M., Schmid, C., Gravier, G., "INRIA LEAR-TEXMEX: Video Copy Detection Task", Proceedings of *TRECVID 2010*

[JIA 11] Jiang, M., Shu, F., Tian., Y, Huang., T., "Cascade of Multimodal Features and Temporal Pyramid Matching", Proceedings of *TRECVID 2011*.

[JOL 05] Joly, A., Frélicot, C., Buisson, O., "Content-based video copy detection in large databases: A local fingerprints statistical similarity search approach", in Proceedings of the *International Conference on Image Processing*, 2005.

[KIM 05] Kim, C., Vasudev, B., "Spatio-temporal sequence matching for efficient video copy detection", in Proceedings of the *IEEE Transactions on Circuit Systems Video Technology*, 15 (1), 2005, pp.127–132.

[KIM 09] Kim, J., Nam J., "Content-based video copy detection using spatio-temporal compact feature", *Proceedings of the 11th international conference on Advanced Communication Technology (ICACT)*, Vol. 3, 2009.

[KOCH 85] Koch, c., Ullman, S., "Shifts in selective visual attention: towards the underlying neural circuitry", *Human neurobiology*, Vol. 4, No. 4. (1985), pp: 219-227, 1985.

[LAW 06] Law-To J., Buisson O., Gouet-Brunet, Boujemaa N., "Robust voting algorithm based on labels of behavior for video copy detection", *14th ACM International Conference on Multimedia*, pp.835 – 844, Santa Barbara, USA, 2006.

[LAW 07] Law-To, J., Buisson, O., Gouet-Brunet, Boujemaa, N., "Video copy detection on the Internet": The challenges of copyright and multiplicity", *IEEE International Conference on Multimedia & Expo*, pp. 2082 - 2085, Beijing, 2007.

[LAZ 06] Lazebnik, S., Schmid, C., Ponce, J., "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", *CVPR'06*, Vol. 2, pp. 2169-2178, June 17-22, 2006.

[LEE 08] Lee, S., Yoo C.D., "Robust video fingerprinting for content-based video identification", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 7, 2008.

[LOW 04] Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints", *IJCV*, Vol. 60, No. 2, pp. 91-110, 2004.

[LIU 10] Liu, Y., Zhao, W., Ngo, C., Xu, C., Lu, H., "Coherent Bag-of Audio Words Model For Efficient Large-Scale Vid-eo Copy Detection", *ACM CIVR*, pp. 89-96, July 5-7, 2010.

[MAS 06] Massoudi, A., Lefebvre, F., Demarty, C.H., Oisel L., Chupeau, B., "A Video Fingerprint Based on Visual Digest and Local Fingerprints", *IEEE International Conference on Image Processing*, Issue 8-11, pp. 2297 – 2300, 2006.

[MIT 07] Mitrea ,M., Dumitru, O., Prêteux, F., Vlad A., "Zero-memory information sources approximating to video watermarking attacks", *Proceedings of the International Conference on Computational Science and Its Applications*, Lecture Notes in Computer Science 4707, Vol. 3, pp. 445 – 459, Kuala Lumpur, Malaysia, 2007.

[MUK 11] Mukai, R., Kurozumi, T., Kawanishi, T., Nagano, H., Kashino, H. "NTT Communication Science Laboratories at TRECVID 2011 Content Based Copy Detection", *Proceedings* of *TRECVID 2011*.

[NAP 00] Naphade, M. R., Yeung, M.M., Yeo, B.L., "Novel scheme for fast and efficent video sequence matching using compact signatures", In Proc. *SPIE, Storage and Retrieval for Media Databases* , Vol. 3972, pp. 564-572, 2000.

[OOS 01] Oostveen, J., Kalker, T., Haitsma, J., "Visual hashing of digital video: applications and techniques", *Proceedings of SPIE,* vol. 4472, pp. 121-131, San Diego CA, 2001

[OOS 02] Oostveen, J., Kalker, T., Haitsma, J., "Feature Extraction and a Database Strategy for Video Fingerprinting", Lecture Notes In Computer Science, Vol. 2314 archive, Proceedings of the *5th International Conference on Recent Advances in Visual Information Systems*, pp. 117 – 128, 2002.

[RAD 08] Radhakrishnan, R., Bauer, C., "Robust Video Fingerprints Based on Subspace Embedding", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, Las Vegas, pp. 2245 – 2248.

[ROO 05] Roover, C. De, Vleeschouwer, C. De, Lefebvre, F., Macq B., "Robust video hashing based on radial projections of key frames", *IEEE Transactions on Signal Processing,* Vol 53, Issue: 10, pp. 4020 - 4030, 2005.

[ROY 12] http://royal.pingdom.com/2012/01/17/Internet-2011-in-numbers/

[SAR 08] Sarkar, A., Ghosh, P., Moxley, E., Manjunath, B. S., "Video Fingerprinting: Features for Duplicate and Similar Video Detection and Query-based Video Retrieval", *Proceedings of SPIE - Multimedia Content Access: Algorithms and Systems II*, 2008.

[SEO 03] "A robust image fingerprinting system using the Radon transform", Jin S. Seo, Jaap Haitsma, Ton Kalker, Chang D. Yoo

[SIV 03] Sivic, J., Zisserman, A., "Video Google: A Text Retrieval Approach to Object Matching in Videos", IEEE ICCV'03, pp. 1470-1477, October 13-16, 2003.

[SU 09] Su, X., Huang, T., Gao, W., "Robust video fingerprinting based on visual attention regions", in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2009.

[SAN 99] Sánchez, J. M., Binefa, X., Vitrià, J., Radeva, P., "Local Color Analysis for Scene Break Detection Applied to TV Commercials Recognition", *Proceedings* of the *Third International Conference on Visual Information and Information*, pp: 237 – 244,1999.

[TEC 12] http://www.technicolor.com/uploads/associated_materials/ds_technicolor_fingerprinting_4fed9ee8947e9869656659.pdf

[TRE 12] http://trecvid.nist.gov/

[VOB 12] http://www.vobileinc.com/index.html

[VER 12] http://www.vercury.com/about.htm

[WAL 02] Walpole, R.E., Myers, R.H., Myers, S-L., Ye, K., "Probability & Statistics for Engineers and Scientists", *Pearson Educational International*, 2002.

[WIR 12] http://www.wired.com/threatlevel/2012/04/viacom-youtube-appeals/

[WON 09] Wong, W. K. Yuen, C.W.M., Fan, D.D., Chan L. K., Fung E.H.K., "Stitching defect detection and classification using wavelet transform and BP neural network", *Journal Expert Systems with Applications: An International Journal*, Vol. 36, Issue 2, pp 3845 – 3856, March 2009.

[XER 12] http://www.xerfi.fr/

[YAN 08] Yan, Y., Ooi, C. B., Zhou, A., "Continuous Content-Based Copy Detection over Streaming Videos", *Proceedings* of the *IEEE 24th International Conference on Data Engineering*, pp: 853-862, 2008

[YEH 10] Yeh, M-C., Hsu, C-Y., Lu, C-S.; "NTNU-Academia Sinica at TRECVID 2010 Content Based Copy Detection", i*n Proceedings of TRECVID 2010*.

[YOU 12] http://www.youtube.com/t/press_statistics/

[YUA 04] Yuan, J., Duan, L. Y., Tian, Q., Ranganath S., Xu, C., "Fast and robust short video clip search for copy detection," *in Springer: Lecture Notes in Computer Science* - 3332, pp. 479–488, 2004

[ZIU 12] http://ziuz.com/en/twinmatch.ashx

PART II: VIDEO FINGERPRINTING AT WORK: TRACKART

Abstract

A DWT (discrete wavelet transform)-based video fingerprinting method involving a mathematical decision rule for the detection of replicas is presented.

Summarizing, the contributions of the thesis are threefold:

- a novel fingerprinting feature with a new mathematical matching procedure;
- a dynamic synchronization block addressing for the first time the live camcorder recording;
- a bag of visual words algorithm employed for granting the fingerprinting system scalability to large scale databases;

The fingerprint *per-se* is represented by a set of 2D-DWT coefficients of frames sampled from the video sequence. An in-depth statistical investigation on the 2D-DWT coefficients demonstrated not only the stationarity of such coefficients but also the stationarity of their modifications under the computer-simulated camcorder attacks. These mathematical properties grant the fingerprints the uniqueness property and limits the occurrences of false alarms (*i.e.* fingerprints extracted from different video content have to be different).

The fingerprint matching is done based on a repeated Rho test on correlation which allows the detection of replicas, hence ensuring the robustness property (*i.e.* fingerprints extracted from an original video sequence and its replicas should be similar in the sense of the considered similarity metric).

In order to make the method efficient in the case of large scale databases, a localization algorithm is employed. Consequently, the replica sequence is not matched to the entire reference video collection but only with a few candidates determined based on a bag of visual words representation (concept introduced by Sivic and Zisserman in 2003) of the video keyframes. An additional synchronization mechanism able to address the strong distortions from difficult use-cases such as camcorder recording in cinema was also designed.

The method scalability is granted by the localization and synchronization procedures and by its low complexity which is kept under the $O(n\log n)$ limit.

Keywords

2D-DWT coefficients, normalized cross-correlation, Rho test on correlation, localization, bag of visual words, synchronization.

Resumé

Ce chapitre présent une nouvelle méthode de traçage du contenu vidéo, basée sur la transformée en ondelettes discrète (pour définir les empreintes numériques) et sur une règle de décision mathématique définie a partir du teste statistique Rho sur la corrélation appliquée selon une procédure répétitive (pour l'appariement des empreintes numériques).

Les contributions de cette méthode se situent à trois niveaux:

- une nouvelle empreinte numérique et une nouvelle procédure mathématique pour la détection des copies;
- un bloc de synchronisation dynamique pour adresser pour la première fois l'enregistrement en salle de cinéma;
- un algorithme sac de mots visuels (*i.e.* bag of visual words) utilisé pour assurer la scalabilité du système pour des bases de données à grande échelle.

L'empreinte numérique *per-se* est représentée par un ensemble de coefficients 2D-DWT obtenu à partir de trames échantillonnées de la séquence vidéo. Une analyse statistique approfondie sur les coefficients 2D-DWT a démontrée non seulement la stationnarité de ces coefficients, mais aussi, la stationnarité de leurs modifications sous des distorsions simulées par l'ordinateur. Ces comportements mathématiques assurent la propriété d'unicité et limite les occurrences de fausses alarmes (c'est-à- dire les empreintes extraites de contenu vidéo différente doit être différent).

L'appariement des empreintes numériques est réalisé avec un test Rho sur la corrélation, qui permet la détection des copies, assure la propriété de robustesse et limite les occurrences de pertes (c'est-a-dire les empreintes numériques extraites de la séquence vidéo originale et des ses répliques doivent être similaires dans le sens de la métrique de similarité considéré).

Afin de rendre la méthode efficace dans le cas de bases de données à grande échelle, un algorithme de localisation a été proposé. Par conséquent, la séquence copie n'est pas apparie avec toutes les séquences video de la basse de donne, mais seulement avec quelques candidates déterminées sur la représentation sac des mots visuels (concept introduit par Sivic et Zisserman en 2003) des images clés vidéo. Un mécanisme de synchronisation supplémentaire, capable de répondre aux fortes distorsions qui apparait dans les cas d'usage difficiles comme celui d'enregistrement du caméscope dans le cinéma a également été conçu.

La scalabilité de la méthode est assurée par les procédures de localisation et synchronisation et par leur basse complexité qui est maintenue sous la limite $O(n \log n)$.

Mots clés

Coefficients 2D-DWT, corrélation croisée normalisée, Rho test sur la corrélation, localisation, sac à mots visuels, synchronisation.

II.1 TrackART: Synopsis

The present thesis advances a novel video fingerprinting methodology called TrackART able to identify visual content subjected to different types of user induced, mundane or malicious distortions.

Concisely, the challenges the TrackART video fingerprinting method takes are threefold:

- Uniqueness: the TrackART method aims at proposing a video fingerprint which represents the video content with mathematical accuracy and rigor.
- Robustness: the TrackART method aims at providing a general mathematical decision rule for the robustness to distortions and at addressing the challenging use case of live camcorder recording (which has not been yet addressed in the state of the art).
- Scalability: the TrackART method aims at being operative even in large scale databases.

The functioning principle of the TrackART method consists in two phases: the offline phase and the online phase, as illustrated in Fig.II.1.

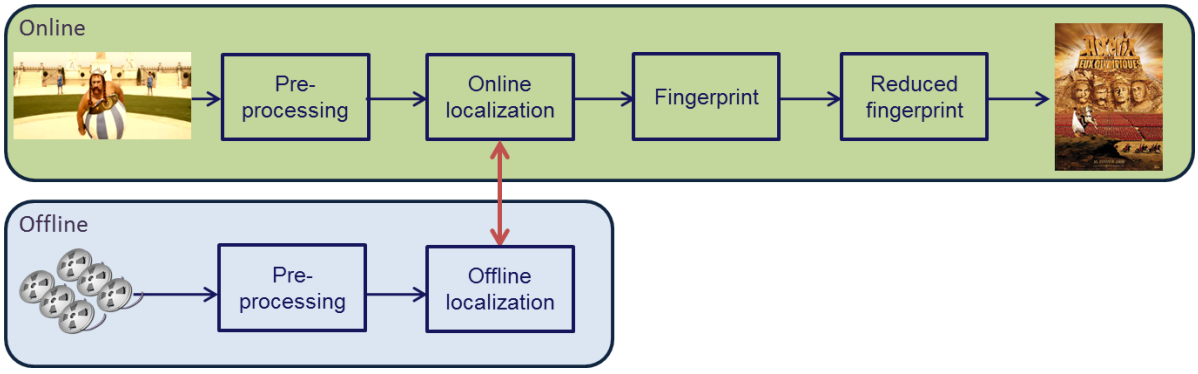


Fig.II.1 TrackART system functional schema

As the word “offline” suggests, the offline phase holds the computation executed before the run-time phase. Its purpose is to process the reference video collection in order to enable the retrieval (if existing) of the original version of the query from the reference database, *i.e.* to enable the localization and the fingerprint modules. The offline phase consists of two modules: pre-processing and offline localization.

The pre-processing stage prepares the reference video sequences for the further processing by performing some parameter setting and common image processing operations as detailed in Section II.2.1).

The offline localization stage (detailed in Section II.2.2) consists in mapping the reference video content to a representation space which allows the matching of video content and enables the localization module.

In the online, a query video sequence is proposed for identification by a user or by another system. By passing the query through the modules of the run-time phase, its original version (if existing in the reference data set) has to be identified. The online phase consists in four modules whose role is briefly given here and further detailed in the rest of this chapter: pre-processing, localization, fingerprint and reduced fingerprint.

The pre-processing module (detailed in Section II.3.1) sets the parameters of the query video sequence to predefined values in order to avoid the variations induced by distortions.

The online localization module (detailed in Section II.3.2) aims at significantly reducing the multitude of reference sequences which are candidates for matching the query (*i.e.* all the video sequences) and to identify just a few nominees for further testing. Moreover, in the localization module, for each nominated video sequence, a potential starting position (*i.e.* the frame number) of the query sequence is obtained.

The Fingerprint module computes and matches the fingerprints of the query and reference video sequences. It consists in three blocks: fingerprint computation (detailed in Section II.3.3.2), fingerprint matching (detailed in Section II.3.3.3) and synchronization (detailed in Section II.3.3.4). The synchronization module is designed to ensure the correct content correspondence between the query and reference video sequences which can be altered by video format distortions.

The reduced fingerprint module (detailed in Section II.3.4) consists in two blocks: reduced fingerprint computation (detailed in Section II.3.4.1) and reduced fingerprint matching (as detailed in Section II.3.4.2) and aims at reducing the amount of information needed for identifying a query video sequence.

II.2 Offline phase

The offline phase enables the localization of a query sequence within a reference sequence. Its purpose is to process the reference video collection and to map the visual content to a representation space. The representation of the content in a new space enables the comparison of the reference and query sequences with respect to certain similarity measures and under different types of distortions. The offline phase consists of two modules: pre-processing (Section II.2.1) and offline localization (Section II.2.2).

II.2.1 Pre-processing

The pre-processing stage aims at achieving a common formatting for the reference video sequences in order to reduce the influence of video format distortions (detailed in Section I.3.4.1) as follows.

Due to the multitude of different existing video formats and to the manipulations that video sequences subsist through their consumption chain (*i.e.* encodings, transcodings), the video frame-rate can have a large variation. In order to enable the TrackART method to cope with this fact, the reference video sequences are all brought to the same frame-rate value. In the current implementation, the frame rate was chosen to be 25 fps due to its frequent use in video formats, but another value can be equally chosen. The changes of the video parameters are done with the ffmpeg library which contains dedicated functions for controlling the video parameters.

The black keyframes were discarded and the letterboxing (if existing) was removed in order to consider only the valid visual information.

The length of a typical film is between 150-250 000 frames (*i.e.* between 1 hour and a half and two hours and a half at a 25 frames per seconds); in order to reduce the complexity, keyframes are extracted uniformly, one frame per second.

Note: Shot boundaries keyframes were also considered in the current study, but as they are not repeatable under video distortions they yield poor results (*i.e.* due to various distortions, the shot boundaries of an original video and its distorted version will not be the same).

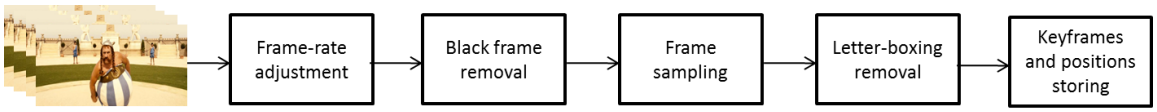


Fig.II.2. Offline pre-processing module

In order to enable the localization step, the sampled keyframes are stored with their position within the reference video sequence (*e.g.* seq1_000006_000131.jpg).

II.2.2 Offline localization

The role of the offline localization module is to provide to the online localization module in the online phase potential positions within the reference sequences which might be the start of the query sequence. Calculating these potential positions relies in identifying matching keyframes between the reference and query video sequences. Therefore the offline localization module reduces to identifying the original version of a query keyframe within the reference keyframe collection.

Establishing keyframe similarity under different frame content and frame aspect modifications (described in detail in Section I.3.4.2 and Section I.3.4.3) induced by user processing is a challenging task. Efficient approaches to this task are provided by methods which exploit the local image features and "Bag of visual Words" (BoW) model of image representation as proved in [CSU 04], [DOU 08] [SIV 06].

Aiming at providing an efficient and accurate keyframe matching procedure, the offline localization module of the TrackART method uses the approach based on the local features and the bag of visual

words model, advanced by Sivic and Zisserman in [SIV 03]. Similar to terms in a text document, an image has local interest points or keypoints defined as patches (small regions) that contain rich local information of the image. Inspired by text retrieval techniques, Sivic and Zisserman developed an algorithm for image search based on representing the images as bags (*i.e.* collections) of visual words (*i.e.* visual descriptors).

The matching of images is assured by comparing the associated bag of words and by testing their spatial consistency.

Being a search algorithm, the Bag of Words framework has two phases: the offline phase and the run-time phase as illustrated in Fig.II.3.

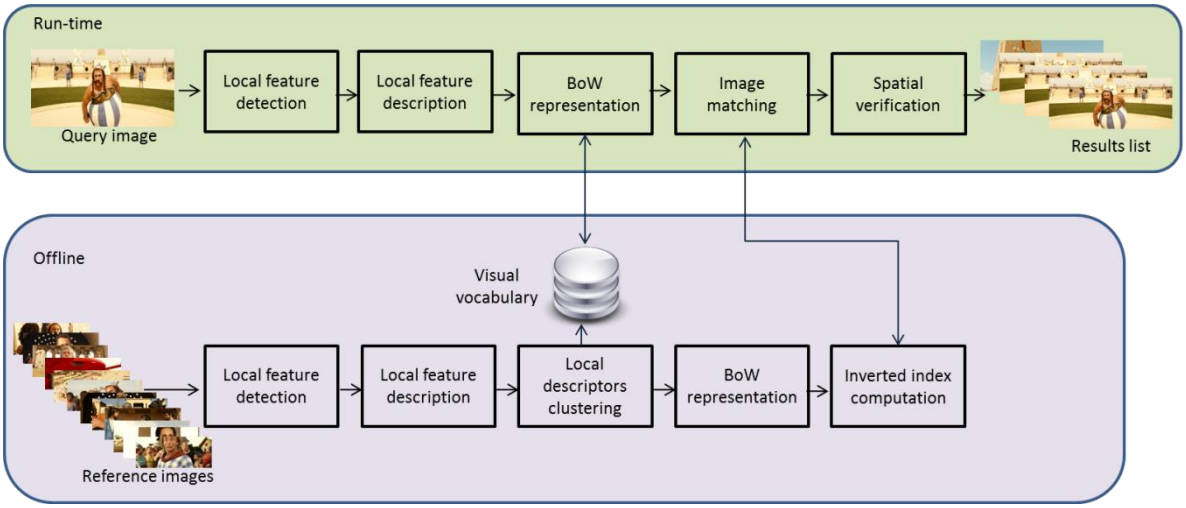


Fig.II.3. The bag of words framework

The scope of the offline phase is to build a visual word representation space based on the reference image collection and to represent each reference image as a collection of visual words from the representation space.

The representation space is called a visual vocabulary. It is built by detecting the local features (detailed in Section II.2.2.1) in all the reference images, by describing these local features with a formalized descriptor (detailed Section II.2.2.2) and by clustering the descriptors into visual words (detailed Section II.2.2.3). Each image in the reference is then expressed as a collection of weighted visual words from the vocabulary (detailed Section II.2.2.4). In order to ensure the retrieval of images in the run-time phase, an inverted file structure stores for every visual word its occurrences in the reference images (detailed Section II.2.2.5).

Within the framework of the TrackART method, the run-time phase of the bag of words framework takes place in the localization module of the TrackART fingerprinting system and is further detailed Section II.3.2.

II.2.2.1 Local feature detection



A local feature is an image pattern which differs from its immediate neighborhood, [TUY 08]. It is usually associated with a change of an image property (*e.g*. intensity, color and texture) or several properties simultaneously; a few examples are illustrated in Fig II.4. To identify local features in images, the underlying intensity patterns in a local neighborhood of pixels needs to be analyzed by using a local feature detector**.** Local features can be interest points, regions (blobs) or edge segments**.**

A set of local features can be used as a robust image representation that allows recognizing objects or scenes without the need for segmentation, [TUY 08].Consequently, local features have gained a lot of momentum in computer vision in the last fifteen years because they are powerful tools in applications like image retrieval from large databases [SCH 97], object retrieval in video [SIV 03], [SIV 04a], visual data mining [SIV 04b], texture recognition [LAZ 03a], [LAZ 03b], shot location [SCH 03], robot localization [SE 02], recognition of object categories [DOR 03]. The relevance of local features has also been demonstrated in the context of object recognition by the human visual system [BIE 98]. Their experiments shown that removing the corners from images impedes human recognition, while removing most of the straight edge does not.

Good local features prove a few properties which make them useful in the above applications, [TUY 08]:

(1) – repeatability: the propriety of local region of being re-detected in other image under different camera viewpoints, illumination conditions and noise)

(2) distinctiveness*:* the intensity patterns underlying the detected features should show a lot of variation, such that features can be distinguished and matched)

(3) – locality: the features should be local, so as to reduce the probability of occlusion and to allow simple model approximations of the geometric and photometric deformations between two images taken under different viewing conditions

(4) quantity: the number of detected features should be sufficiently large, such that a reasonable number of features are detected even on small objects, ideally, the number of detected features should be adaptable over a large range by a simple and intuitive threshold

(5) accuracy: the detected features should be accurately localized, both in image location, as with respect to scale and possibly shape

(6) efficiency: the detection of features in a new image should allow for time-critical applications.

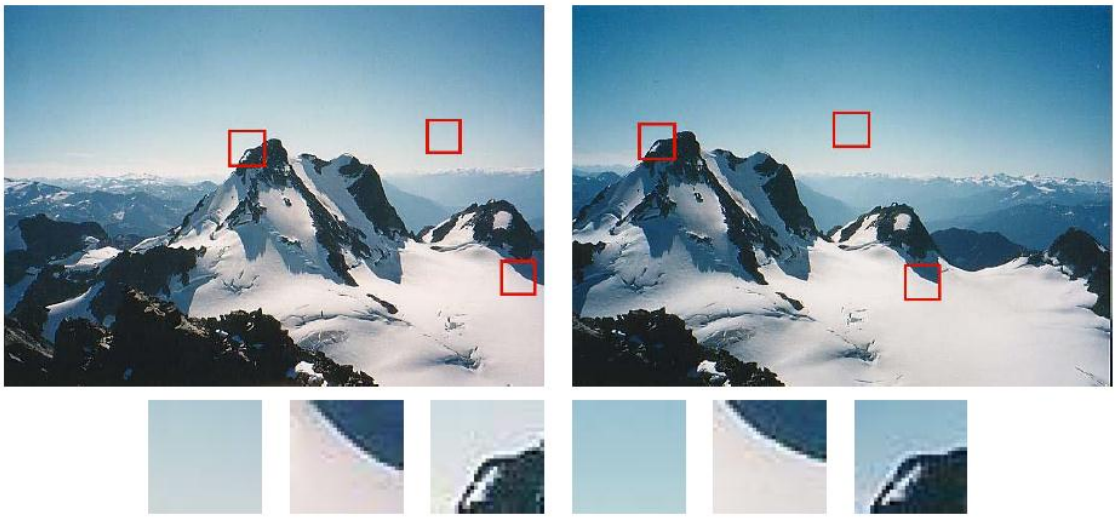


Fig. II.4: Examples of local features, [SZE 10]

Under the framework of a fingerprinting system, the role of local features is to allow the identification of the original version of query keyframes, within the reference keyframes. This requirement can be reformulated: the local features employed in a fingerprinting system need to be robust (*i.e.* invariant) to the distortions arising in different video fingerprinting use cases. Thus the local features employed in the TrackART video fingerprinting method need to cope with frame content and frame aspect modifications (as detailed in Section I.3.4.2 and Section I.3.4.3). The frame aspect distortions are preponderantly photometric distortions, whereas the frame content distortions are affine distortions or content insertion/cropping distortions.

Intuitively, in order to obtain a set of local features robust to a set of distortions, an approach can be to estimate the distortions through a mathematical transformation and then to look for the local features which are invariant to such transformations.

Following this approach, the mathematical transform which can model photometric distortion is a linear transform of pixel intensities whereas the transform which can model the image distortions like affine manipulations, (*i.e.* viewpoint changes, scale changes, rotations), partial occlusion or cropping is the affinity, [MIK 05a]. Consequently, the local features necessary for the TrackART video fingerprinting method have to be robust to linear and affine transformations. In order to obtain such features, local feature detectors which detect features with such properties have to be investigated.

The state of the art exhibits many approaches for local features and feature detectors, such as:

- the SIFT detector [LOW 99], the Harris detector, [HAR 84] and the SUSAN detector [SMI 97] detect corners as local features, robust to translation, rotation and stable under varying lighting conditions;
- the Harris-Laplace detector yields scale and rotation invariant regions, and the Harris-Affine detector yields affine invariant regions as showed in [MIK 04];

- the Hessian-Laplace and Hessian-Affine detectors identify scale invariant and respectively affine invariant blobs (ellipsoid regions) as local features in [MIK 04]; the intensity based detector in[TUY 00], [TUY 04] leads to affine covariant regions as local features;
- the MSER (Maximally Stable Extremal Regions) detector proposed in [MAT 02] identifies affine covariant (photometrically and geometrically) regions.

Note that the local features identified with the help of detectors have been referred in the state of the art, both as invariant or covariant. On the one hand, the local features are detected invariant to the image distortions and on the other hand, the local features covariantly change with the image distortions (*i.e.* with the 2D affine image transform which models the distortions).

Taking into account the requirements set for the TrackART video fingerprinting method, the local features that qualify the best are the affine covariant regions. They can cope with the geometric and photometric deformation of images.

Typically, such regions have an elliptical shape while other approaches such as the DoG (Difference of Gaussians) [LOW 99] use fixed shape circular support regions. The advantage of elliptical shaped regions over circular ones is illustrated in Fig.II.5.

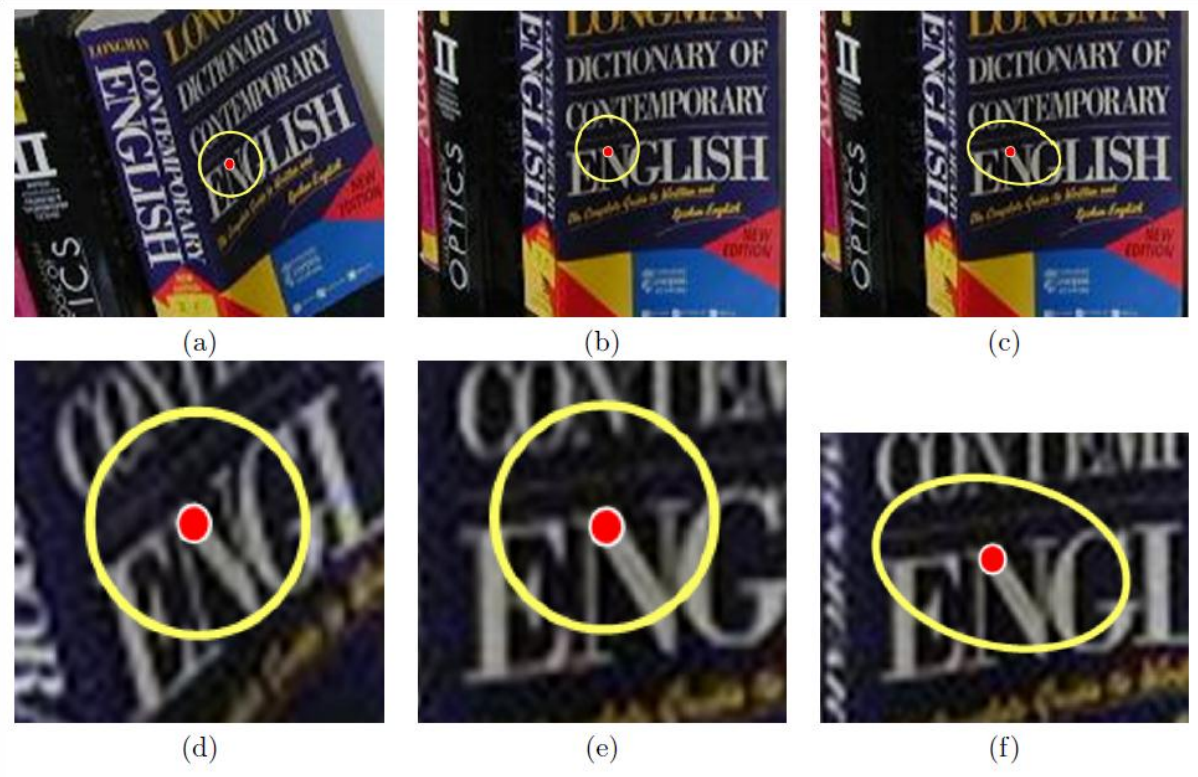


Fig.II.5 Limitation of circular support regions under large viewpoint changes [MIK 05a]: (a) First viewpoint. (b)-(c) second viewpoint. The circular region in (b) does not cover the same object surface patch as the circular region in (a). What is needed is a deformation of the circular region in (b) by an anisotropic scaling to the ellipse shown in (c). Note that regions in (a) and (c) cover approximately the same surface patch on the book. (d)-(f) close-ups of (a)-(c).

The key idea of the affine covariant region detectors is that the shape of the region is automatically adapted to underlying image intensities in a single image in such a way that regions detected independently in each image correspond to the same 3-dimensional surface patch. The size and shape of such regions are covariantly transformed under a particular 2-D image distortion. In most of the cases an affine transformation is a reasonably good local approximation to transformations arising from viewpoint changes for locally planar surfaces. In the case of video content, most of the objects and characters are in motion and suffer different transformations. Consequently, affine covariant regions are *a priori* a more suitable solution to this problem comparing to its DoG correspondent.

Concerning the detector, the study and comparison of affine covariant region detectors in [MIK 05a] concludes that no detector shows superiority in all experiments. However, the MSER and the Hessian-Affine detectors had consistently higher scores. The MSER detector performs well on images containing homogenous regions with distinctive boundaries but the number of detected regions is rather reduced comparing to Hessian-Affine. The Hessian-Affine detector provides more regions than other detectors, making them more suitable for identifying cluttered or occluded objects.

Taking into account all the considerations above, for the TrackART method the Hessian-Affine region detector is used for detecting the affine covariant regions which will be used as local features.
The Hessian-Affine *detector* algorithm consists in three major operations, [MIK 02]: the detection of the interest points, the detection and selection of scale for the interest points and the estimation of the region shape.
The Hessian-Affine detector is applied on gray scale images, hence they consider the intensity information within an image.
The interest points and their scales are computed and selected with the Hessian matrix. The algorithm searches in the Gaussian scale space over a fixed number of predefined scales $\sigma_1 \dots \sigma_n$, with $\sigma_i = k^i \sigma_0$ and k = 1.4 [MIK 04]. For each integration scale σ_I, chosen from this set, an appropriate local differentiation scale is chosen to be a constant factor of the integration scale $\sigma_D = s\sigma_I$, where $s = 0.7$.
Considering an image I, the interest points a are detected in the Gaussian scale space with the Hessian matrix H. For each integration scale σ_I, the Hessian matrix is issued from the Taylor expansion of the intensity function $I(a)$.

$$H = H(a,\sigma_D,\sigma_I) = \begin{bmatrix} I_{xx}(a,\sigma_D,\sigma_I) & I_x I_y(a,\sigma_D,\sigma_I) \\ I_x I_y(a,\sigma_D,\sigma_I) & I_{yy}(a,\sigma_D,\sigma_I) \end{bmatrix} \qquad \textit{(II.1)}$$

where the local image derivates are computed with Gaussian kernels of local derivation scale σ_D. $I_{xx} = \frac{\partial^2 I}{\partial x^2}$ is second order partial derivate in the x direction and $I_{xy} = \frac{\partial^2 I}{\partial x \partial y}$ is the mixed partial second order derivative in the x and y directions. The derivatives are computed at the current integration scale and are thus the derivatives of an image smoothed by a Gaussian kernel $g(\sigma_I)$.

$$I(a,\sigma_D) = g(\sigma_D) * I(a) \qquad \text{(II.2)}$$

The components of the Hessian matrix are illustrated in Fig.II.6.

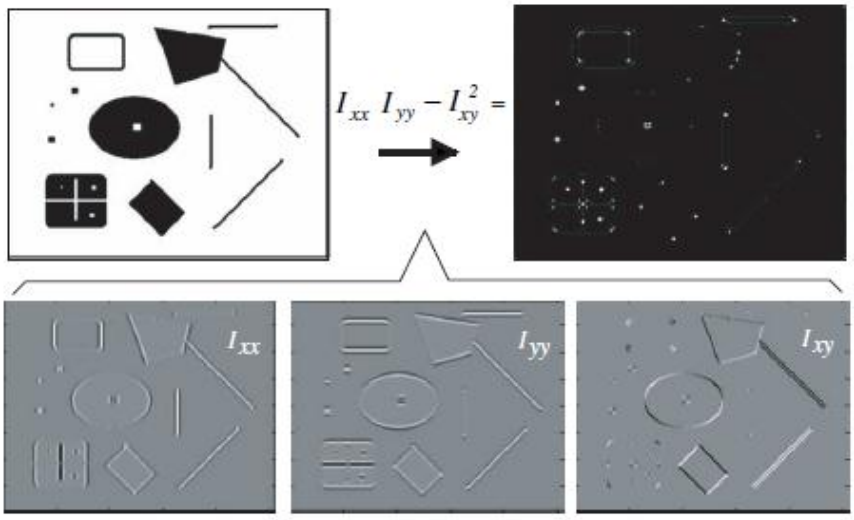


Fig.II.6 Illustration of components of the Hessian matrix and Hessian determinant [TUY 08]

The algorithm starts from the determinant and the trace of the Hessian similarly with the Laplacian of Gaussians [LIN 98], (the trace of this matrix is also referred to as the Laplacian):

$$\det(H(a,\sigma_D)) = \sigma_D(I_{xx}(a,\sigma_D)I_{yy}(a,\sigma_D) - I_{xy}^2(a,\sigma_D)) \quad \text{(II.3)}$$

$$trace(H(a,\sigma_D)) = \sigma_D(I_{xx}(a,\sigma_D) + I_{yy}(a,\sigma_D)) \quad \text{(II.4)}$$

At each scale, the interest points are detected as those points that are simultaneously local extrema of both the determinant and trace of the Hessian matrix (*i.e.* a local maximum of the determinant decides if it is an interest point, a local maximum in the trace decides its characteristic scale).

By thresholding the Hessian determinant and the Laplacian response, the number of regions detected can be controlled.

Having the interest points extracted at their characteristic scale, the shape of the affine elliptical region of the point neighborhood is determined based on the eigenvalues of the second moment matrix (*i.e.* called the autocorrelation matrix) with an iterative region estimation algorithm as described in [LIN 97] and illustrated in Fig.II.7. The autocorrelation matrix describes the gradient distribution in a local neighborhood of a point a. The eigenvalues of the autocorrelation matrix represent the principal signal changes in two orthogonal directions in a neighborhood of the point a at the scale σ_I. The matrix must be adapted to scale changes to make it independent of the image's resolution.

	Iteration 1
	Iteration 2
	Iteration 3
	Iteration 4
	Iteration 5 – final result

Fig.II.7 Obtaining the affine shape of a region through the iterative algorithm in [LIN 97]

The scale-adapted second moment matrix for a point a is defined by:

$$M = \mu(x, \sigma_D, \sigma_I) = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} = \sigma_D^2 g(\sigma_I) * \begin{bmatrix} I_{xx}(a, \sigma_D) & I_x I_y(a, \sigma_D) \\ I_x I_y(a, \sigma_D) & I_{yy}(a, \sigma_D) \end{bmatrix} \quad \text{(II.5)}$$

where I_x is the derivative computed in the x direction $I_x = \frac{\partial I}{\partial x}$. The eigenvalues of the second moment matrix are used to measure the affine shape of the point neighborhood, by computing the transformation that projects the intensity pattern of this neighborhood to one with equal eigenvalues. Practically, the affine region is skewed or stretched to a normalized circular region where the second moment matrix is isotropic. A new location and scale are detected in the normalized region. If the eigenvalues of the second moment matrix for the new point are equal, the estimation is correct. Otherwise, a new affine shape is estimated with the second moment matrix and tested. When estimating the shape and size of the affine region, with the algorithm proposed in [LIN 97] there can be a maximum of 16 iterations. If the shape of the region is not detected after 15 iterations, the algorithm discards the current interest point and takes another one.

The resulting shapes of the regions will be adapted to the underlying intensity patterns and ensure in this manner that the same parts of different instances of the same region are covered in spite of deformations caused by viewpoint change or rotations.

A few examples of Hessian-Affine regions are illustrated in Fig.II.8. Note that only 10% of the regions detected in the images are shown so that the illustration is not overwhelmed by the amount of regions.

The software implementation of the Hessian-Affine detector was used from [PER 09].



Fig.II.8.Examples of Hessian-Affine regions (only 10% of the regions are illustrated)

II.2.2.2 Local feature description



Once the local features have been identified, they have to be formalized into a description which can allow their matching or further use. Choosing or designing a descriptor preserving and enhancing the affine invariance of the Hessian-Affine regions is not an easy task.

The study in [MIK 05b] investigates the performances of certain descriptors in the context of descriptor matching (for recognition of the same object or scene purposes) between images under the following distortions: affine transformations, scale changes, rotation, blur, jpeg compression and illumination changes. The investigated descriptors are: SIFT [LOW 04], PCA-SIFT [KE 04], gradient location and orientation histogram (GLOH) [MIK 05b], shape context [BEL 02], spin images [LAZ 03a], steerable filters [FRE 91], moment invariant [GOO 96], and cross-correlation of sampled pixel values which are all computed on Hessian-Affine regions.

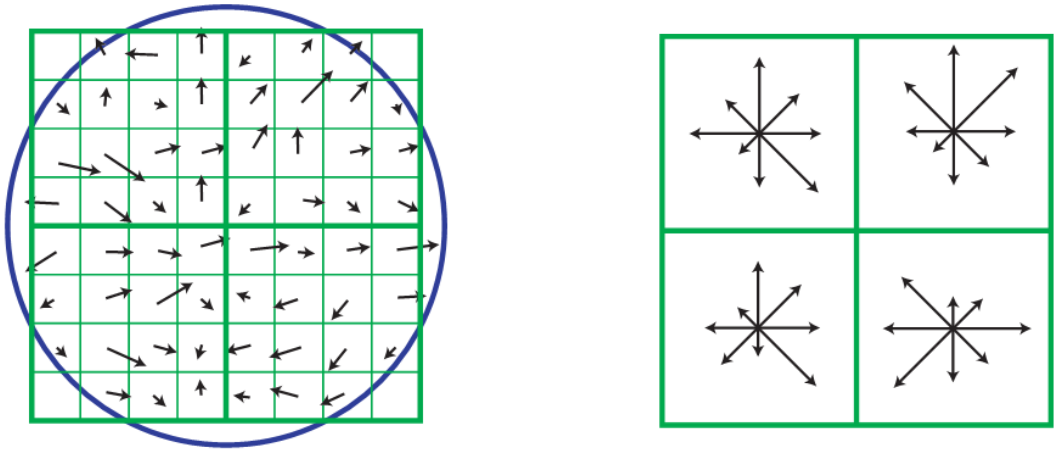
The results of the study brought to light the superiority of the SIFT and GLOH descriptors and proved the robust and distinctive character of the region based SIFT descriptor (*e.g.* SIFT and GLOH had the highest matching accuracies for affine transformation of 50°; for scale changes in the range 2-2.5 and image rotations with 30-45° the SIFT descriptor outperformed the others; the SIFT descriptor proved better on both textured and structured images; the introduction of blur also pointed to the superiority of the SIFT descriptor; in terms of distinctiveness, the SIFT was ranked top three).

While [MIK 05b] investigates the SIFT descriptor used as an entity for matching (*i.e.* the descriptors of the query and reference images are matched one to one based on a nearest neighbor similarity metric), its good properties have been confirmed in bag of visual words approaches as well.

The SIFT descriptors have been employed successfully in other fingerprinting techniques detailed in [DOU 08], [JIA 11], [BER 11, [LIU 11], [ZHA 11].

Therefore due to its proven accuracy of describing local regions confirmed by the wide spread use, Lowe's SIFT (Scale Invariant Feature Transform) descriptor was chosen for the TrackART video fingerprinting method.

The SIFT descriptor is illustrated in Fig.II.9 and is obtained from DoG (Difference of Gaussian) points. It is a 128-histogram storing in each bin the magnitude of a local gradient in a certain direction, every bin representing a direction. SIFT is constructed from a 4 × 4 grid centered on the interest points. Each cell of the grid quantizes gradient direction into 8 bins.



(a) Image gradients (b) Interest point descriptor

Fig.II.9 SIFT descriptor illustration [LOW 04]. Image gradients within a patch (left) are accumulated into a coarse 4 × 4 spatial grid (right). In this example we show only a 2 × 2 grid. A histogram of gradient orientations is formed in each grid cell. 8 orientation bins are used in each grid cell giving a descriptor with the dimension 128 = 4 × 4 × 8)

A smoothing Gaussian function with σ equal to one half the width of the descriptor window is added in order to emphasize the information in the close neighborhood of the interest points. This is illustrated with the circular window in Fig.II.9.a. The purpose of the Gaussian window is to avoid sudden changes in the descriptor with small changes in the position of the window and to give less weight to the gradients which are far from the center of the descriptor.

In order to achieve the rotations invariance, all gradients within the patch are computed relative to a dominant gradient orientation, which is obtained as the highest peak in a histogram of all gradient orientations within the patch. The gradients are illustrated with small arrows at each sample location in Fig.9.II.a.

The interest point descriptor is shown in Fig.II.9.b. It allows significant shift in gradient positions by creating orientation histograms over 4 × 4 sample regions. The eight directions for each orientation histogram with the length of each arrow corresponding to the magnitude of the histogram entry.

The descriptor is formed from a vector containing the values of all orientation histogram entries, corresponding to the lengths of the arrows in the Fig.II.9b. Whereas Fig.II.9 shows a 2 × 2 array of orientation histograms, the best results are achieved with a 4 × 4 array of histograms with 8 orientation bins in each therefore the descriptor used is a 4 × 4 × 8 = 128 feature vector for each interest point.

The SIFT descriptor is also robust to affine illumination change effects because the feature vector is normalized to unit length (*i.e.* a change in image contrast in which each pixel value is multiplied by a constant multiplies gradients by the same constant, so this contrast change is canceled by the vector normalization; a change in brightness in which a constant is added to each image pixel does not affect the gradient values because they are computed from pixel differences).

The SIFT descriptors are computed for the Hessian-Affine regions in a keyframe by warping the elliptical patches into a circular patch of 41 × 41 pixels and rotated based on the dominant gradient orientation to compensate for the affine geometric deformations.

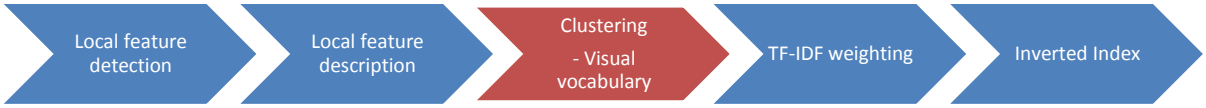
In Fig.II.10 an affine covariant region is warped to a circular shape and then brought to a 41 × 41 pixels patch from which the SIFT descriptor is computed.



(a)Affine covariant regions	(b) Warped affine covariant region	(c) The patch of the region

Fig.II.10.The image patch on from which the SIFT descriptor is computed

II.2.2.3 Clustering



II.2.2.3.1 Visual vocabulary

The local features detected in the reference keyframes and their formalized descriptors pile up to a significant amount of data which is not easily comparable without an *a priori* structuring.

Consequently, in order to reduce the number of considered features and to optimize the search times, large scale image or video retrieval systems cluster the high-dimensional descriptors into a limited set of descriptors, a so called vocabulary of visual words. The visual words vocabulary has been also referred in the state of the art as visual codebook or visual dictionary.

Based on the visual vocabulary, images are mapped into the bag of words representation by assigning to their local features the corresponding visual word in the vocabulary. The advantage of this approach is the increased efficiency: the assignment of visual words labels to image local features, leads to matching becomes labels (*i.e.* the visual words) instead of matching the high dimensional descriptors of the local regions.

The visual vocabulary is a key component for a large scale image retrieval system because it enhances the efficiency in the online localization stage. However, the creation of the vocabulary, *i.e.* the clustering, is in general the most computational expensive stage of the offline phase.

Considering the case of the proposed video fingerprinting technique, the amount of descriptors to be clustered depends on the number of reference keyframes in the reference video database and can easily vary between 50 000 and millions of keyframes. Considering the amount of affine covariant regions within a keyframe, it can vary between a few hundreds and up to thousands of regions, hence yielding millions of SIFT descriptors.

Generating clusters from large collections of high dimensional descriptors presents computational costs which cannot be surmounted by typical clustering algorithms such as k-means, mean-shift, spectral and agglomerative. The proof comes from the study reported in [SIV 03] which uses flat k-means clustering effectively but concludes that it is impossible to scale it to large vocabularies. Therefore k-means algorithms, scalable to high dimensional spaces need to be investigated.

A big step towards this direction is made by Nister and Stewenius [NIS 06] who introduce the vocabulary tree obtained from hierarchical k-means (HKM) clustering and brought significant improvements in the retrieval accuracy. Due to its reduced complexity, the method can scale to very large numbers of clusters and feature points (*i.e.* more than 1 million visual words).

An advantage of the HKM and vocabulary tree is the hierarchical scoring, which considers nodes from several levels in the similarity score, weighting the contribution of each level to the score with an entropy weight relative to the root of the tree and ignoring dependencies within the path. In this way, possible quantization errors can be overcome.

The study in [PHI 07] introduced the Approximate k-Means (AKM) which is a scalable version of the k-means algorithm.

A typical k-means algorithm is a method which aims at partitioning a set of *n* observations, $x_1, x_2, ..., x_n$, into *k* clusters, $S = \{S_1, S_2, ..., S_k\}$ in which each observation belongs to the cluster with the nearest mean, *i.e.* $\arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$, where μ_i is the mean of points in S_i.

In the typical k-means algorithm, the majority of computational cost is due to the computation of the nearest neighbors between the points and the cluster centers. In the approximate k-means algorithm, the nearest neighbor operation is replaced with an approximate nearest neighbor method and 8 randomized k-d trees are built over the clusters centers at the beginning of each iteration to increase speed, as proposed in [LEP 05] and [MUJ 09].

In a typical k-d tree, [FRI 77] each node splits the dataset using the dimension with the highest variance for all the data points falling into that node and the value to split on is found by taking the median value along that dimension. Concerning the randomized k-d trees, the splitting dimension is chosen at random from a set of the dimensions with highest variance and the split value is randomly chosen using a point close to the median. The union of these trees creates an overlapping partition of the feature space and helps to diminish the quantization effects, where features which fall close to a partition boundary are assigned to an incorrect nearest neighbor.

A new data (*i.e.* descriptor) is assigned to the approximately closest cluster center as follow. Initially, each tree is descended to a leaf and the distances to the discriminating boundaries are recorded in a single priority queue for all trees. Then, the most promising branch from all trees is iteratively chosen and unseen nodes are added into the priority queue. The stop condition is when a fixed number of tree paths have been explored.

The algorithmic complexity of a single k-means iteration is reduced from $O(Nk)$ to $O(N \log k)$, where N is the number of SIFT descriptors that is being clustered. It was proved by [PHI 10] that for moderate values of k, the percentage of points assigned to different cluster centers differs from the exact version of the algorithm by less than 1%.

The choice between the two state of the art scalable k-means algorithms is for the AKM due to its close to optimal results. The main drawback of the HKM method is that at each level in the tree, a decision is made on cluster ownership. At each such point, a wrong decision can be taken and thus the HKM features multiple occasions of causing quantization errors. Since it relies on the flat k-means, the AKM involves only a single decision, so the probability of a quantization error is reduced. Moreover, the studies in [PHI 07], [PHI 10] show that the AKM algorithm achieves better results than the HKM.

Another parameter to be set is the size of the visual vocabulary. No guidelines have been yet derived in order to establish a clear relationship between the number of descriptors to be clustered, the clustering techniques and the results they provide.

Unlike the vocabulary of a text corpus whose size is relatively fixed, the size of a visual word vocabulary is controlled by the number of clusters in the clustering process, [YAN 07]. A good

vocabulary size involves the trade-off between discriminability and generalization. In the case of small vocabularies, the visual words are not very discriminative because dissimilar points can be mapped to the same visual word. With the increase in vocabulary size, the visual words become more discriminative, but in the same time less generalizable. Using a large vocabulary also increases the cost of clustering the descriptors and the computation of the bag of words representation. Usually smaller vocabularies are used for image classification whereas for object or image retrieval, larger and more discriminative vocabularies are considered. However, in general the increase in vocabulary size yields better results. Philbin [PHI 10] tested vocabulary sized to 2 million visual words, while Nister and Stewenius [NIS 06] reached 16 million points for a vocabulary size. The conclusion of the studies was that in general the large vocabularies yield better results and that an increase in the number of clusters achieves a better improvement than taking more data in the clustering process, *i.e.* more descriptors.

Table II.1 illustrates the choice made in [PHI 07] for the size of the vocabulary according to the number of descriptors. The obtained mean Average Precision (mAP) in the case of the typical k-means algorithm and in the case of the approximate k-means (AKM) is also mentioned for the different size vocabularies.

Clustering parameters		mAP	
# of desc	Voc. size	k-means	AKM
800K	10K	0.355	0.358
1M	20K	0.385	0.385
5M	50K	0.464	0.453
16.7M	1M		0.618

Table II.1. Clustering parameters and their performances as shown in [PHI 07]

For the TrackART video fingerprinting method, the uniform sampling of the video sequences in the reference database yielded 47 163 keyframes. For the sampled keyframes, the local feature detection and description lead to 38 466 280 descriptors.

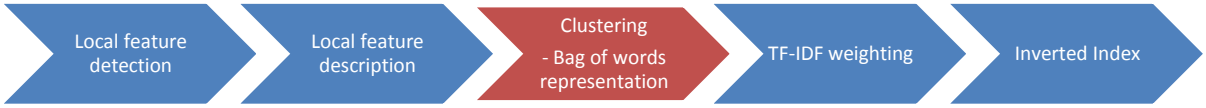
Based on the examples in other studies [JEG 08], [PHI 07] and in order to achieve an accurate retrieval of keyframes even in the context of distortions while keeping a reasonable computational cost, the size of the vocabulary was chosen to be 250 000 clusters.

Following the approach in [PHI 07] illustrated in Table II.1, the ratio between the number of descriptors and the vocabulary size, K:N was chosen to be 1:15, therefore the a subset of the descriptors from the dataset was sampled. The sampling yielded a subset of 10% descriptors, totalizing 3 846 628 descriptors.

Note that the sub-sampling of the descriptors is done randomly: the order of the images is randomized and the order of the descriptors for each image is randomized as well. Thus, all bias is avoided and the sub-sampling is done in a completely automatic manner.

Concerning the practical implementation of the AKM technique, the specifications of the authors [PHI 10] were used: 30 iterations, 8 trees and 1024 distance computations, for the clustering 784 checks per point and 1500 for the final assignment. The AKM is easily parallelizable and its distributed memory computed can be achieved with the open source MPI library [MPI 12]. A publicly available implementation of AKM is available from [PHI 12].
For the reference dataset of *N* = 3 846 628 descriptors the clustering into *K* = 250 000 clusters on a 4 core machine took 3 hours.

II.2.2.3.2 Bag of words representation



Having the visual vocabulary computed, the next step consists in expressing the reference keyframes as a collection of visual words from the vocabulary, the bag of words representation (BoW). This is achieved by assigning each SIFT descriptors from every reference keyframe to the most similar visual words in the vocabulary. The descriptor assignment is done with a fast approximate nearest neighbor strategy proposed in [MUJ 09]

The nearest neighbor search problem can be formulated as follows: given a set of points $P = \{p_1, p_2, ..., p_n\}$ in a vector space $q \in X$, these points must be preprocessed in such a way that given a new query point $q \in X$, finding the points in P that are nearest to q can be performed efficiently. For high dimensional spaces (as it is the case of the SIFT descriptor which has 128 dimensions) there are no known algorithm for nearest neighbor search that are more efficient than linear search. As linear search is too costly for many applications, the algorithms which compute an approximate nearest neighbor search have been considered. Such approximate algorithms can be orders of magnitude faster than exact search, while still providing near-optimal results.

The two approaches for computing the fast approximate nearest neighbor proposed in [MUJ 09] are the randomized kd-tree algorithm and the hierarchical k-means tree algorithm which were implemented in the C++ FLANN library (Fast Library for Approximate Nearest Neighbors).

The classical kd-tree algorithm [FRE 77] is efficient in low dimensions but its performances degrade rapidly in high dimensions. In its original form, the kd-tree algorithm splits the data in half at each level of the tree on the dimension for which the data exhibits the greatest variation. [SIL 08] improved the algorithm by using multiple randomized kd-trees, which are built by choosing the split dimension randomly from the first D dimensions on which data has the greatest variance. The fixed value D = 5 was used as it was proven to perform well in the [MUJ 09] and does not benefit greatly from further tuning.

When searching the trees, a single priority queue is maintained across all the randomized trees so that search can be ordered by increasing distance to each bin boundary. The degree of

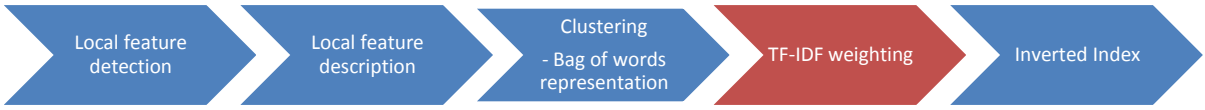
approximation is determined by examining a fixed number of leaf nodes, at which point the search is terminated and the best candidates returned.

The hierarchical k-means tree is constructed by splitting the data points at each level into K distinct regions using a k-means clustering, and then applying the same method recursively to the points in each region. The recursion stops when the number of points in a region is smaller than K. The algorithm explores the hierarchical k-means tree in a best-bin-first manner by analogy to what has been found to improve the exploration of the kd-tree. The algorithm initially performs a single traversal through the tree and adds to a priority queue all unexplored branches in each node along the path. Next, it extracts from the priority queue the branch that has the closest center to the query point and it restarts the tree traversal from that branch. In each traversal, the algorithm keeps adding to the priority queue the unexplored branches along the path. The degree of approximation is specified in the same way as for the randomized kd-trees, by stopping the search early after a predetermined number of leaf nodes (dataset points) have been examined.

Within the FLANN library, the selection of the algorithm for approximate nearest neighbor can be done automatically, based on a precision wanted by the user, *e.g.* considering a precision of 60% it is assume that 40% of the nearest neighbors returned are not the exact nearest neighbors, but just approximations under the advantage of greatly reducing the computational cost. Within the algorithm the precision parameter given by the user is represented as a cost function which allows the algorithm to choose between the two possibilities of computing the approximate nearest neighbors, either the kd-tree method, either the hierarchical k-means tree. The cost function is based on the method's specificities search time, tree build time and tree memory overhead.

For the TrackART video fingerprinting method, the kd-tree algorithm was chosen and not the precision auto-tuned version. The kd-tree algorithm was chosen because it is the fastest to build and most memory efficient. The number of randomized trees was set to D = 4 trees and the number of examined leaf nodes to 32.

II.2.2.4 TF-IDF weighting



The bag of words representations obtained in the previous stage for the reference keyframes are further used to build a vector of visual word frequencies for each keyframe. This is achieved by employing text retrieval and statistical text analysis techniques as proposed in [SIV 06].
The components of this vector are weighted in order to overcome biases related to the uneven number of visual words per image or to the reachability of often encountered visual words or less encountered ones.

These vectors are weighted with the standard weighting, [BAE 99], known as term frequency–inverse document frequency (tf–idf) and is computed as follows.

Assuming the computed visual vocabulary has K words, then each image (*i.e.* each keyframe) is represented by a vector:

$$v_d = (t_1,...,t_i,...,t_K)^T \quad \text{(II.6)}$$

of weighted word frequencies with components:

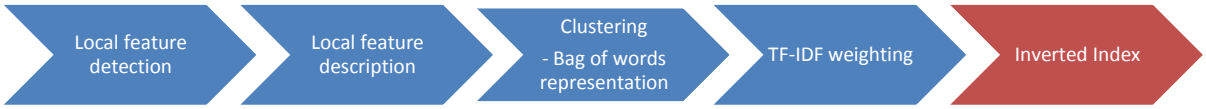
$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{N_i} \quad \text{(II.7)}$$

where n_{id} is the number of occurrences of word i in document d, n_d is the total number of words in the document d, N_i is the number of documents containing term i, and N is the number of documents in the whole database. The weighting is a product of two terms: the word frequency, $tf = \frac{n_{id}}{n_d}$ and the inverse document frequency, $tidf = \log \frac{N}{N_i}$.

Intuitively, it can be observed that on the one hand, the word frequency (tf) gives a higher weight to the words occurring more often in a particular document in comparison with words that do not appear at all, hence offering a relevant representation.

On the other hand, the inverse document frequency downweights words that appear often in the database, which do not help to discriminate between different documents, hence contributing to the relevant representation of the image through visual words.

II.2.2.5 Inverted index



Having the reference keyframes represented as bags of visual words, *i.e.* as tf-idf vectors, efficient matching can be achieved with the help of an inverted file structure.

An inverted file [WIT 99] is a commonly used indexing structure in text retrieval and its structure is analogue to a complete book index.

In the current implementation the format of the inverted index was chosen as illustrated in Fig.II.11. For each visual word in the vocabulary, an entry is considered in the inverted index table. For each entry, a list of all the occurrences of the considered visual word in the reference keyframes is attached.

Word ID	Keyframe list
1	5, 8, 50, …
2	425, 789, 20
…	….
K	44879, 28900

Fig.II.11 Inverted index

Depending on the application and on the computation needed in the online phase, the structure of the inverted index can be more complex.

For example, [SIV 09] stores for each occurrence of the visual word, its position in the corresponding frame and the distance to the 15th nearest neighbor in the image, in order to achieve a fast geometrical consistency verification. In [JEG 07], a tree-based inverted index for fast search is designed and in [PHI 10] the position and the coordinates of the elliptical shape of each point is stored.

The advantage of using an inverted index is shown in the online localization phase (Section II.3.2.), when a query keyframe is given in its BoW representation and the reference keyframes which contain the same visual words are inquired. The visual words of the query keyframe are searched in the inverted index and are used to generate a list of plausible reference keyframe candidates by selecting only images that contain at least an occurrence of one query visual word. Consequently, this strategy reduces greatly the number of keyframes to be compared, while ensuring that no plausible candidates have been omitted.

II.3 Online phase

In the online phase of the TrackART video fingerprinting system, a query video sequence is considered to be given by a user or another system. Depending on the use case scenario, the identity of the query or its existence in a database is inquired. The run-time phase consists of the following modules: pre-processing, localization, fingerprint and reduced fingerprint.

II.3.1 Pre-processing

The pre-processing step in the run-time phase is identical to the pre-processing step in the offline phase (detailed in Section II.2.1) with the difference that it is applied to the query video sequence.

It consists in the following operations: frame rate adjustment to 25 fps, removal of black keyframes, uniform frame sampling at 1 fps, letter-boxing removal and storing of the sampled keyframes together with their positions (*i.e.* frame number).

II.3.2 Online localization

In general, the query video sequence is just a part of a reference video sequence and therefore needs to be localized within that particular reference video sequence. The online localization procedure refers to identifying the position (*i.e.* the frame number) at which the query video sequence is located within the reference video sequence. Due to the distortions (described in Section I.3.5) which are induced in the video queries, and which transform the video sequence, localizing a query is a challenging task.

The online localization step consists in the run-time phase of the bag of words algorithm introduced in Section.II.2.2 and illustrated in Fig.II.11.

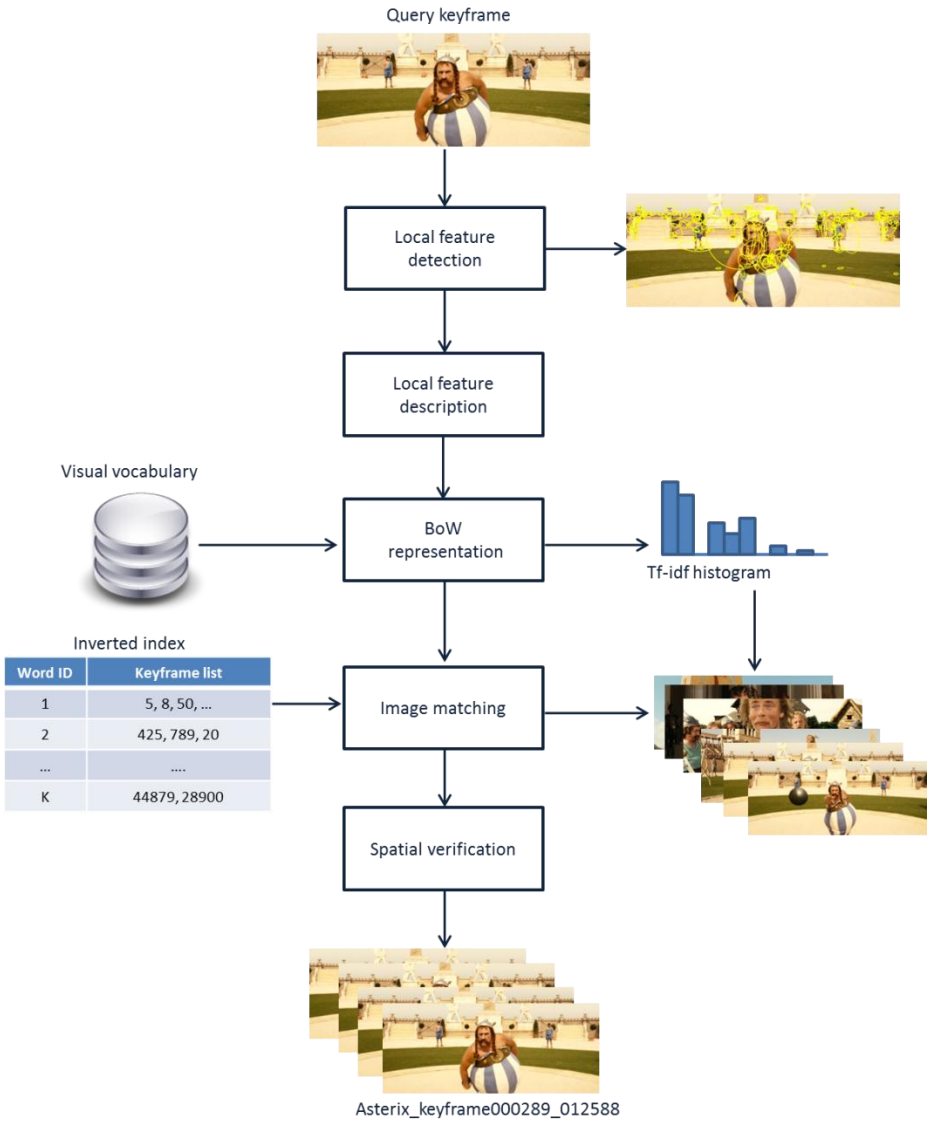


Fig.II.12. Run-time phase of the Bag of Words framework

In the run-time phase of the bag of words algorithm, a query image is considered and its original version it's searched for in the reference image database (*i.e.* in the case of the TrackART method, the original version of a query keyframe is searched for among the reference keyframes). This is achieved by detecting the local features in the query image, by computing their descriptors and by quantizing the descriptors into visual words. Having the bag of words representation, the matching between the query and the reference images can take place. Based on the inverted index, reference image candidates are obtained and matched based on the normalized scalar product similarity metric (detailed in Section II.3.2.1). The matching reference images are further filtered through a geometrical verification as detailed in Section II.3.2.2.

The localization step for the TrackART method takes as input the sampled keyframes of the query video sequences and returns their matches in the reference keyframes database and implicitly their location within the reference video sequences.

II.3.2.1 Keyframe matching

In the keyframe matching stage, the similarity between query keyframes and reference keyframes is inquired. The similarity between keyframes is formulated under the bag of words framework as a histogram similarity measure, where the tf-idf BoW representations of keyframes are considered as histograms with equal number of bins (*i.e.* the size of the vocabulary).
The keframe matching stage consists in the following steps: (1) – considering a given query keyframe, its visual words are searched within the inverted index and a list of reference keyframes which contain the same visual words is retrieved; (2) – a histogram similarity distance is employed between the tf-idf vectors of the query and reference keyframe.

In order to achieve the matching of these keyframes, the normalized scalar product is considered between their tf-idf vectors, as proposed in [SIV 06].

The normalized scalar product between the query vector v_q and all reference vectors v_d in the database is:

$$f_d = \frac{v_q^T v_d}{\|v_q\|_2 \|v_d\|_2} \tag{II.8}$$

where $\|v\|_2 = \sqrt{v^T v}$ is the L_2 norm of v.

Note that when the vectors are normalized using the L_2, then $\|v_q\|_2 = \|v_d\|_2 = 1$ and (eq f_d) becomes:

$$f_d = v_q^T v_d = 1 - \frac{1}{2}\|v_q - v_d\|_2^2 \tag{II.9}$$

Therefore, ranking keyframes in ascending order of their L_2 distance (*i.e.* $\|v_q - v_d\|$), is equivalent to ranking them in the ascending order of $v_q^T v_d$. When v_q and v_d are very sparse then the dot product can be computed very quickly by only considering terms which are non-zero in both v_q and

v_d. Computing the ranking for all images in the dataset where, $D = [d_1, d_2, ..., d_N]$, can just be considered as a sparse matrix multiplication, $s = D^T v_q$, where s_i is then the similarity of the i^{th} document to the query.

The inverted index is essentially a compressed sparse row representation of the matrix D. The sparse matrix-vector product only needs to explicitly consider the non-zero contributions to this product.

In our case the query vector is given by the frequencies of visual words contained the query keyframe, weighted by the inverse document frequencies computed on the visual word.

Retrieved keyframes are ranked according to the similarity of their weighted vectors to this query vector.

Other weighting schemes and distances between the tf-idf vectors were tested in the state of the art. The study in [TIR 10] proposed probabilistic models for weighting such as BM25 [ROB 77] which weights the tf (by considering that word occurrences are distributed following two Poisson distributions) and the idf terms (according to a probability ranking principle which ranks the results according to their relevance with the query); different variants of the *Ln* histogram distance have been tested by varying the value of *n.* No method showed superior performance and the authors concluded that the choice of the technique depends on the dataset and its size, on the size of the vocabulary and on the use case.

Sivic and Zisserman [SIV 09] investigated several weighting schemes (including different normalizations of the tf and tf-idf vectors) with corresponding similarity measures (*e.g.* L1, L2, $\chi 2$ [LEU 01], Kullback–Leibler (KL) divergence [VAR 05], Bhattacharyya [AHE 98]). Their experiments found that the standard tf-idf and Bhattacharyya ranking had the best scores, followed closely by the Kullback-Liebler divergence methods and the standard tf-idf method with L2 ranking.

Considering that no tf-idf weighting technique and histogram distance proved better results and due to the fact that it is difficult to estimate automatically the most suitable choice given a dataset and a query, for the TrackART video fingerprinting method, the standard tf-idf weighting and the normalized scalar product.

II.3.2.2 Geometric consistency verification

In the keyframe matching step, the reference keyframes which are the most similar to a given query keyframe are returned based on the normalized scalar product and the bag of words representation. The bag of words representation does not take into account the spatial configuration of the visual words within the keyframes. However, it was proved in [PHI 07] that a spatial consistency between the visual words of the query and reference images can improve the results.

Therefore, the aim of the geometric consistency verification block is to establish a spatial coherence between the visual words of the query keyframe and the visual words of the reference keyframes returned by the keyframe matching block.

The geometric consistency verification has been done with several algorithms in the literature. The standard solution is to use the RANSAC (RANdom SAmple Consensus, proposed by [FIS 81])

algorithm which is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers. A basic assumption for RANSAC is that the data consists of "inliers", *i.e.*, data whose distribution can be explained by some set of model parameters, and "outliers" which are data that do not fit the model. The RANSAC algorithm consists in generating hypothesis concerning the targeted geometrical transformation using a minimal number of correspondences between two images. Then, each hypothesis is evaluated based on the number of inliers among all features under the hypothesis. The transformation hypotheses are scored by maximum number of inliers. The algorithm checks the top results yield by the bag of words matching and re-ranks them according to their spatial consistency with the query.

Other versions of the RANSAC algorithm were proposed in the state of the art for geometric verification: the PROSAC (Progressive Sample Consensus), [CHU 05] method weights correspondences by employing an external measure of confidence, which is used as a priority for guiding the search towards good solutions; the GroupSAC [NI 09] partitions points into groups based on similarity information; for their large scale object retrieval, Philbin *et al.* [PHI 07] don't use RANSAC because the estimation of a full 3-D fundamental matrix or 2-D projective homography between two images is too general and runs very slowly. They use LO-RANSAC (Locally Optimized-RANSAC) [CHU 04] a variant of RANSAC which consists in (1) - generating hypotheses of an approximate model and then (2) - iteratively re-evaluating promising hypotheses using the full transformation. The approximate model is built iteratively from single pairs of correspondences verified through a class of transformations of the affine-invariant regions corresponding to the matched points.

For the TrackART video fingerprinting method, the approach proposed in [PHI 07] was chosen because it is fast, effective and can generate transformation hypotheses even with a single pair of corresponding features, which is very useful the case of distortions which induce content addition or content cropping.

The geometrical consistency verification proposed in [PHI 07] starts from one pair of matched points (correspondences) and uses a 5 degrees of freedom affine transform combined with a decision threshold (for deciding the points that feature geometric consistency). The threshold is chosen in order to allow the matching of images with significant perspective distortions which can appear in camcorder recording use-cases. The transformation between the points of the correspondence is used for generating a hypothesis, which is further applied to the rest of the correspondences. The correspondences with the distances lower than the threshold are considered to be verified for the assumed hypothesis and added to a list of verified inliers.

The inliers from the list are then used to re-estimate a full affine homography with the least-squares method. In practice, this step can be discarded as accurate results can be obtained also by simply counting the verified inliers for each hypothesis and selecting the one with the highest number of inliers.

The 5 degrees of freedom transform allows translation, anisotropic scaling and vertical-preserving shear. The elliptical regions are constrained to be oriented "up" as it is a good assumption for videos to be filmed in without significant rotations in the viewpoint. This transform is computed from a single correspondence of two elliptical regions C_p and C_q from the reference image and the query

image, respectively. The region centroids are used to compute the translation, while the affine *2* x *2* sub-matrix H_{qp} is computed as $H_{qp} = H_p^{-1} H_q$ (Fig.II.13), where H_p and H_q project the ellipses to a unit circle such that the orientation of the unit vector in the y direction (*i.e.* "up") is maintained. The matrices H_p and H_q can be computed in closed form using a transposed Cholesky decomposition, $C = H^T H$. The transformation considered in this case is modeled with the following functional matrix:

$$A = \begin{bmatrix} a & 0 & t_x \\ b & c & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad \text{(II.10)}$$

The geometrical consistency verification is applied as following for the TrackART method. In the keyframe matching block the distances between the BoW representations of the query keyframe and the reference keyframes are computed and a ranked list of reference keyframes is generated for the query keyframe.

The geometrical consistency verification is performed on the first *N* reference keyframes which are further re-ranked. Firstly, a matching between the affine covariant regions of the query keyframe and the reference keyframe is performed with an approximate nearest neighbors algorithm (FLANN) proposed in [MUJ 09].

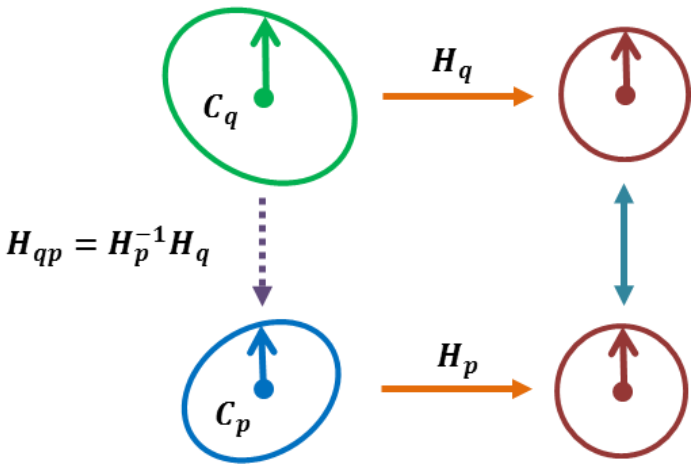


Fig.II.13 Computing $H_{qp} = H_p^{-1} H_q$ by projecting ellipses C_p and C_q to a unit circle and preserving the "up" orientation

Secondly, once this list of matches between the covariant regions is computed, they are checked iteratively by employing the 5 degrees of freedom transform. This 5 degrees of freedom transformation is considered as a hypothesis and applied to all matched points from the query keyframe, and then projected into the reference keyframe. The projected points that are localized close to their corresponding points are considered as inliers and added to the list of verified matches for this transform. The configuration which yields the highest number of inliers is returned and the *N* considered frames are re-ranked in decreasing order of the number of verified inliers, leaving the ranking of the rest of the results unchanged.

II.3.3 Fingerprint

The input for the Fingerprint block is a pair of matching query and reference keyframes provided by the localization module. Implicitly, the associated query frame f_q, reference frame f_r and their position in the video sequences are known.

In order to grant a mathematical based video feature as fingerprint, the properties and statistics of the wavelet coefficients have been reconsidered and investigated in Section II.3.3.1.

Having identified potential locations where the query sequence can start in the reference video sequence, the fingerprints of the query and reference videos can to be computed (detailed in Section II.3.3.2) and matched (detailed in Section II.3.3.3). Moreover in order to enable the proposed video fingerprint to resist inner time-variant desynchronization, a synchronization block has been designed (as detailed in Section II.3.3.4).

II.3.3.1 Discrete Wavelet Transform coefficients statistics

Having in view the fingerprinting and watermarking applications [MIT 04b], [DUM 07] investigated the probability density law modelling the 2D-DWT coefficients and established at what extent the ergodicity hypothesis of these law holds.

A concise presentation of the proposed procedure [MIT 04a] follows:

- Be there a video sequence sampled from the 2D random process representing the video.
- Consider the video sequences as a set of *L* successive frames.
- Compute the DWT to the V component of each frame; these coefficients can be either considered according to their spatial position; sort coefficients in a decreasing order; record the largest *R* coefficients.
- Partition the *L* values corresponding to an $r = 1,...,R$ location (spatial or rank), into *D* classes by using a fixed period sampling of period $D = 250$ and by shifting the sampling origin.
- Apply for each class the Chi-square test on concordance, the Ro test on correlation, the Fisher test on equality between two variances, and the Student test on equality between two means; all these tests are applied at an $\alpha = 0.05$ significance level.

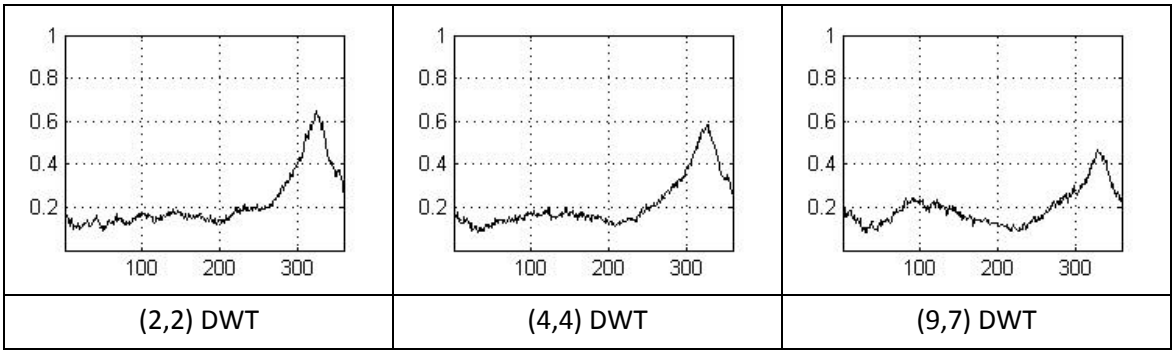
Note that if *D* is large enough, the elements in each class are independent.

The results are illustrated in Fig.II.14 and Fig.II.15, for the particular case of coefficient sorted in decreasing order. In Fig.II.14 the statistical investigation was applied for three bi-orthogonal DWTs, namely (2,2), (4,4), and (9,7) DWT [DAU 92], [MIT 04c]. The abscissa corresponds to the investigated rank and the ordinate corresponds to the relative number of the Chi-square tests which are not passed.

It can be seen that for the considered DWTs, more than 75% of tests are passed only when:

- $r \in [10; 50] \cup [150; 250]$ in the (9,7) DWT case;

- $r \in [0;100] \cup [170;230]$ in the (4,4) DWT case;
- $r \in [0;210]$ in the (2,2) DWT case.



(2,2) DWT	(4,4) DWT	(9,7) DWT

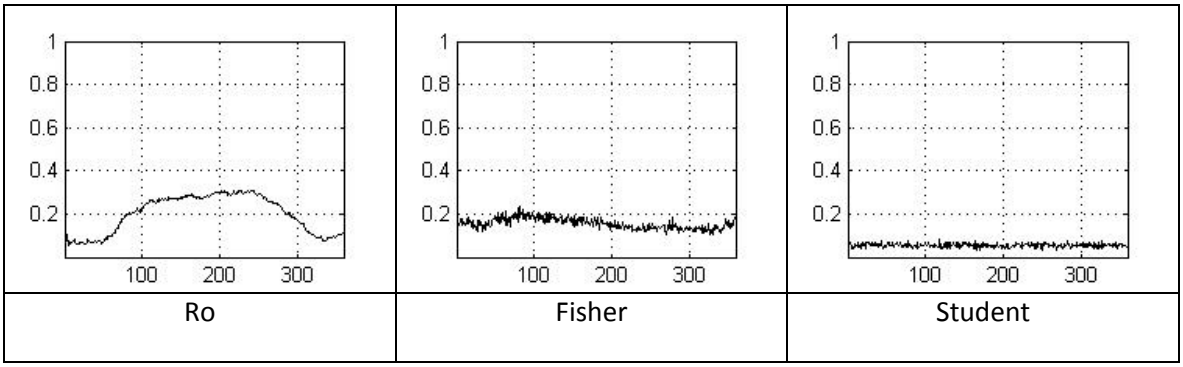
Fig.II.14: The relative number of Chi-square tests on concordance with the Gaussian law which are not passed ($R = 360$ coefficients, $L = 35000$ frames, and $D = 250$ frames)

For others ranks the Gaussian behaviour has been refuted.

Concerning the DWT coefficients selected according to their spatial frequency, the Gaussian behaviour can be always accepted at least as a first hand approximation [MAL 99], [DAU 92].

When the Chi-square tests were not passed, the Ro, Fisher and Student tests cannot be properly run (such tests are mathematically proved only for Gaussian data). However the very high ratio of the Ro tests which are passed are considered as an encouraging hint in data independency and stationarity: the mean value and the variance are independent with respect to a translation on the time axis.

The Ro, the Fisher and Student tests were applied and the results are illustrated in Fig.II.15 for the (9,7) DWT. The figure axes correspond to the investigated ranks *vs.* the relative number of tests (Ro, Fisher and Student) which are not passed.



Ro	Fisher	Student

Fig.II.15: The results of the Ro, Fisher, and Student tests in (9,7) DWT domain

($L = 35000$ frames, $R = 360$ coefficients, and $D = 250$ frames)

Although the estimation of the probability density law modelling the 2D-DWT coefficients and its ergodicity has been determined for DWT coefficients disposed on ranks, a similar behaviour can be presumed for all the DWT coefficients. In the sequel, a collection of DWT coefficients are employed as video features for the TrackART video fingerprinting system.

II.3.3.2 Fingerprint computation

The fingerprint for the TrackART method was computed in the DWT domain due to its capacity of identifying the overall salient content of images and representing it through edges in the high frequency sub bands and due to the fine statistical properties featured by the wavelet coefficients. Moreover the Daubechies (9, 7) wavelets were used due to their very fine capacity of approximating the visual content.

The 2D wavelet coefficients are computed as following.

Assuming a pair of sampled query and reference frames, the fingerprint computation can take place. The fingerprint computation module consists of four main steps: spatial subsampling, color space conversion, wavelet transform and coefficients selection as illustrated in Fig.II.16.

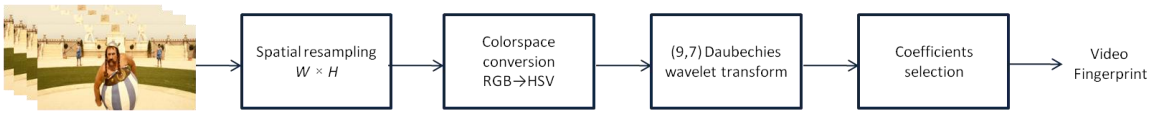


Fig.II.16. Fingerprint computation principle

In the first step, the spatial re-sampling to $W \times H$ pixels is performed on the reference and query frames. The CIF format (*i.e.* $W = 352$ and $H = 288$) was chosen in the current implementation in order to decrease the computational time, *i.e.* the computation of the fingerprint is directly proportional to the size of the video frames, therefore the smaller the size of the video frame, the shorter the processing time. Note that the CIF resolution is not mandatory and other formats can be equally chosen as they do not influence the stability of the video fingerprint.

In the second step, the color space is changed from the native RGB to HSV (Hue - Saturation - Value) and only the V component is considered further in the fingerprint computation. The HSV color space separates the luma (*i.e.* the image intensity, in the H and S components) from chroma (*i.e.* the color information in the V component), thus making the fingerprint invariant to color changes or distortions and increasing the robustness of the proposed method.

In the third step, a (9, 7) Daubechies wavelet transform at the resolution level of $Nr = 3$ is applied on the V component of every sampled frame.

Fourthly, the fingerprint is computed by selecting DWT coefficients from the query and the reference frames. The coefficient selection aims at conveying as much information as possible about the frames in order to achieve robustness to transformations like the frame aspect and frame content modifications (detailed in Section I.3.4.2-3). Consequently, all the DWT coefficients in each frequency sub-band (LL, LH, HL, HH) yielded by wavelet transform are selected.

Note that The fingerprint was computed in the DWT domain due to its capacity of identifying the overall salient content of images and representing it through edges in the high frequency sub bands. Moreover the Daubechies (9, 7) wavelets were used due to their very fine capacity of approximating the visual content.

However, other types of DWT like (2,2) or (4,4) can be used with a very low impact on robustness, while keeping the same uniqueness and reducing the computational time.

II.3.3.3 Fingerprint matching

Once the fingerprints of the query and the reference frames are computed, they have to be matched.

The proposed similarity measure between the fingerprints is the normalized correlation as given by the formula in (II.12).

$$\rho = corr_k(f,t) = \frac{1}{N-1}\sum_{x,y}\frac{(f_k(x,y)-\bar{f}_k)(t_k(x,y)-\bar{t}_k)}{\sigma_{f_k}\sigma_{t_k}} \qquad \text{(II.12)}$$

In (II.12), f_k and t_k designate the 2D-DWT coefficients of the query and the reference frames respectively, in a frequency sub-band *k*, $\bar{f}_k, \bar{t}_k$ are the mean values of the 2D-DWT coefficients in the considered frequency sub-band, while σ_{fk}, σ_{tk} are the related standard deviations, respectively. *N* designates the number of 2D-DWT coefficients in every frequency sub-band *k*. is each of the frequency sub-bands: LL, LH, HL, HH yield by the wavelet transform.

A perfect match (identity) between the query and the reference fingerprints is obtained when $|\rho| = 1$; a value $\rho = 0$ indicates no correlation between f_k and t_k .

In practice, in order to be able to also retrieve content preserving replicas, the absolute value of the normalized correlation should be compared to some threshold *T*; should $\rho \geq T$, then the query and the reference ranks are considered as similar.

The value of the *T* threshold is statistically determined according to the Rho test on correlation [WAL 02]. This test is individually applied to each frequency band under investigation; the null/alternative hypotheses are:

$$\begin{cases} H_0: \text{the coefficients in a sub-band are not correlated} \\ H_1: \text{the coefficients in a sub-band are correlated} \end{cases}$$

A match between the query and the reference frames is obtained when the coefficients in all four frequency sub-bands are correlated. Should the coefficients in one of the frequency sub-band be uncorrelated, the query and reference frames are considered as distinct.

Assuming the *N* 2D-DWT coefficients from a frequency sub-band are *i.i.d.* (identically and independently distributed) and that they follow a Gaussian distribution, and assuming the H_0 is true, the t_{test} value of the test statistics, see (II.13), follows a Student probability density function of *N-2* degrees of freedom:

$$t_{test} = \rho \cdot \sqrt{\frac{(N-2)}{1-\rho^2}}, \tag{II.13}$$

where *N* and ρ are the same as above.

As illustrated in Fig.II.17, if $t_{test} \le z_{\alpha/2}$ (where $z_{\alpha/2}$ is the α-point value of the above-mentioned Student law), then the H_0 hypothesis is accepted, *i.e.* the 2D-DWT coefficients in a considered frequency sub-band are not correlated. If $t_{test} > z_{\alpha/2}$ the H_1 , hypothesis is accepted, *i.e.* the 2D-DWT coefficients in a particular frequency sub-band are correlated.

In the experiments presented in this thesis, a significance level of $\alpha = 0.05$ was considered.

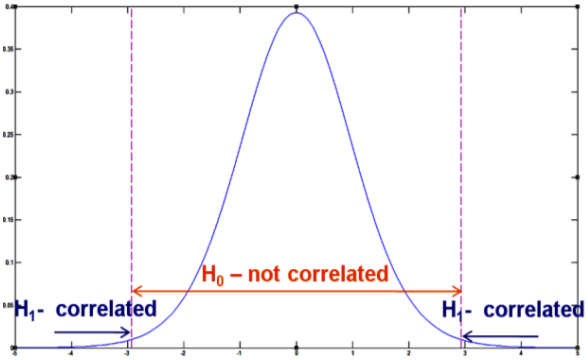


Fig.II.17 Student probability density function (the illustration corresponds to 25 degrees of freedom)

Considering that frames throughout a video sequence can be very similar, (for example in the case of scene with no motion) a query keyframe can turn out to be correlated with more than one reference frames. In order to establish which of the reference frames is the original version of the query frame, the match between the query and the reference frames is established with a correlation score as defined in (II.14) used. The correlation score is computed as the average of the correlation coefficients on all four sub-bands

$$\rho_{score} = (\rho_{LL} + \rho_{LH} + \rho_{HL} + \rho_{HH})/4 \tag{II.14}$$

The higher ρ_{score}, the higher the correlation and hence the similarity between two frames. Consequently the pair of query-reference with the highest correlation score will be considered as matched.

The match between the query and video sequences is achieved by using a threshold *E* of correlated sampled frames. Should the number of sampled correlated frames be equal or higher than E the video sequences are considered as matched. Should the number of sampled correlated frames be lower than *E* the video sequences are distinct.

To conclude with, the fingerprints of a video sequence are the wavelet coefficients in every frequency sub-band for a selection of sampled frames. The matching between the wavelet coefficients is assured by the Rho test on correlation, whereas the matching between query and reference frames in hence based on the correlation score. The statistical error control is given by the type I statistical error in the test.

II.3.3.4 Synchronization

Some video processing operations (*e.g.* change of frame rate) or user manipulations like camcorder recording can induce in a video sequence, an inner time-variant desynchronization. Examples of these desynchronizations are the combined frames at shot transitions or slightly different shot durations between the original and the attacked video sequences.

Consequently, between the distorted and the original version video, the "video content – frame number" correspondence is not identical anymore. Therefore an operation which ensures that the features that will be used as fingerprints are computed from the same visual content in the original and distorted video sequences is needed.

Moreover, as the video content exhibits a high redundancy between adjacent frames, a frame sub-sampling is necessary in order to reduce the computational cost.

Actually, prior to the fingerprint computation, this synchronization should be achieved. While conceptually different tasks, in order to increase the efficiency of the TrackART method, the fingerprint computation/matching and the synchronization are nested. Actually, in the synchronization procedure described below, the fingerprint computed from candidate frames represents the feature on which the synchronization relies. Similarly, the fingerprint matching gives the similarity measure considered in the synchronization procedure.

To serve these aims, a procedure trying to synchronize the video content and to reduce the computational cost was designed to complement the fingerprint computation step.

The synchronization block consists in two stages: the first solves the light time-variant desynchronization whereas the second aims at solving distortions which induce a higher degree of desynchronization.

In the first stage of the algorithm, when attempting to synchronize the query with the reference sequence, the query frame is not matched to a single reference frame but to several frames from its neighborhood (*e.g.* $L=10$ frames), as illustrated in Fig.II.18.a.

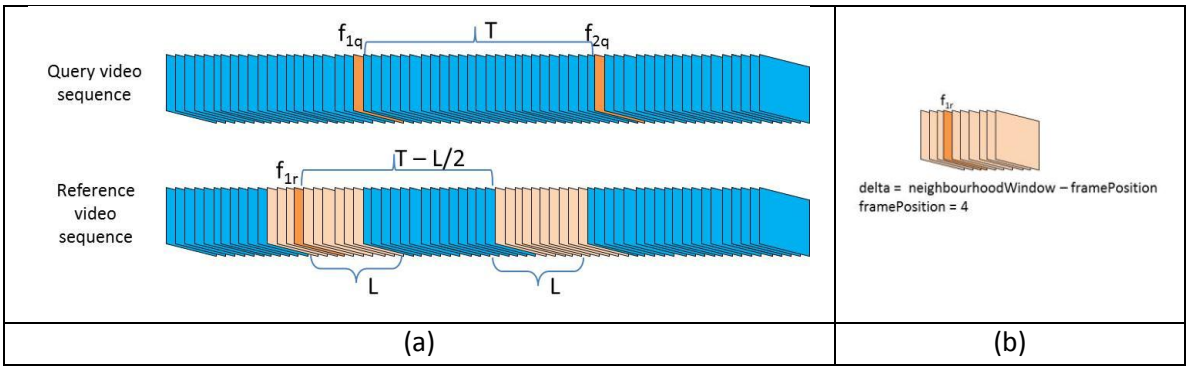


Fig.II.18 Synchronization algorithm first stage

The position of the selected reference frame in the neighborhood window, $framePos$ is retained and used further for synchronization purposes, Fig.II.18.b.

In case a match between the query and the reference frames is obtained, the matching reference-query pair is stored and the process continues as the scope is to synchronize the entire query sequence with the frames of the reference sequence.

In order to reduce the computational cost, just a selection of query frames are matched to their reference frames. This selection is obtained by using a sampling strategy: considering a general sampling rate $T=25$, the query is sampled with T and the reference with $T-L-\Delta$, where $\Delta=L-framePos$ The Δ factor is used to compensate the desynchronisation that might exist between the query and reference frames.

In case no match is obtained after the reference neighborhood window $L=10$ is browsed, the second stage of the algorithm starts: the window is enlarged, by doubling its size $2L$, and a new browsing is done as illustrated in Fig.II.19. If a match occurs, the reference and the query sampling rates are set as before, whereas if no match occurs after iterating the procedure $X=10$ times, both sampling rates are lowered in order to try to resynchronize in the neighborhood; the query sampling rate is set to $T'=10$ and the reference sampling rate to in $T'+\Delta$.

The synchronization stage finishes when all the query frames corresponding to the query keyframes were processed as explained above, or when no matching pair query-reference frames are encountered for $Z=5$ processed query keyframes.

This condition verifies the temporal consistency of the query sequence with respect to the reference sequence (*i.e.* if the succession of frames in the query sequence is the same as in the reference). Moreover, this condition automatically discards the similar, but not original versions reference frames retrieved by the localization algorithm as potential matched for the query keyframes.

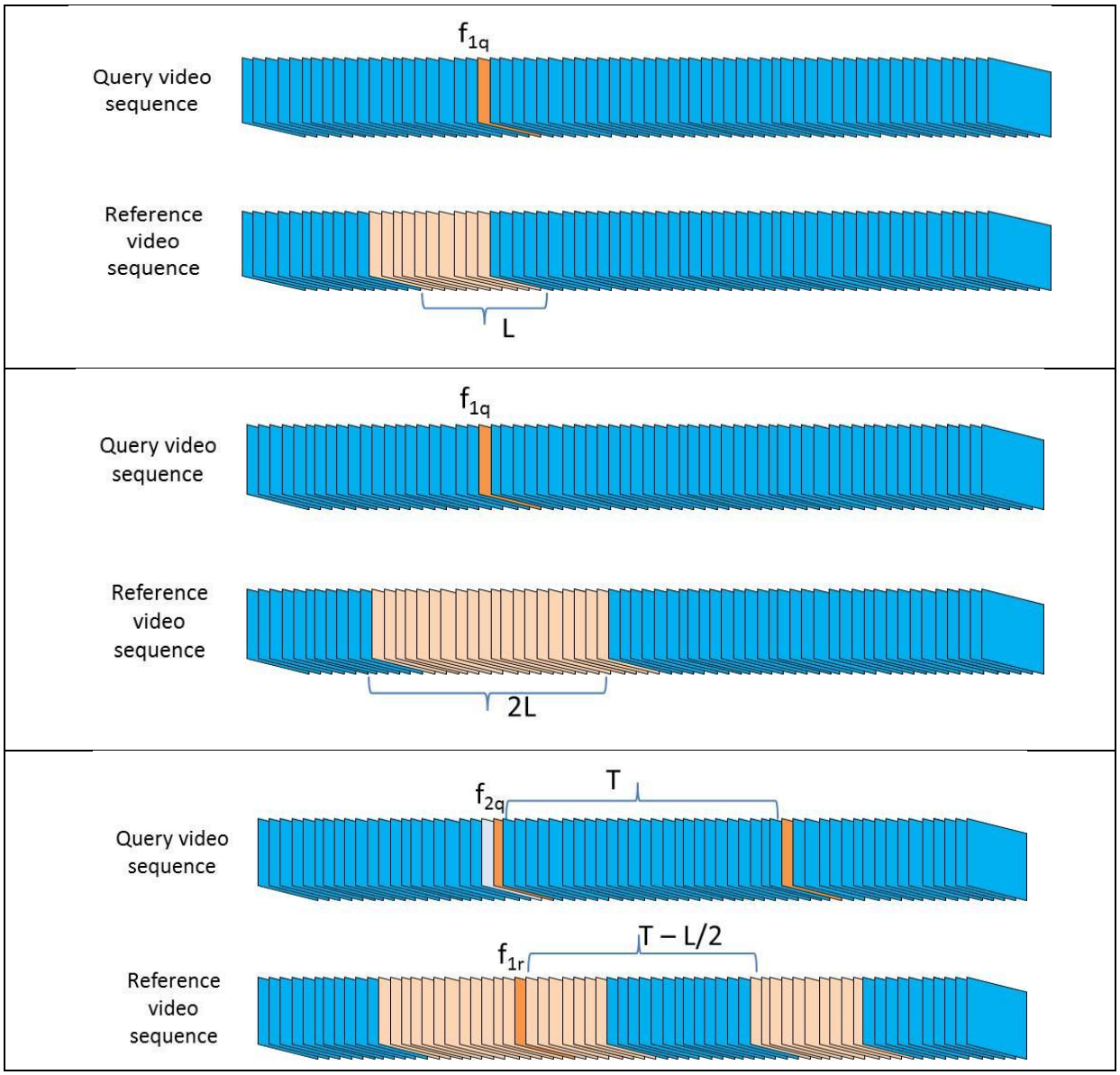


Fig.II.19 Synchronization algorithm second stage

II.3.4 Reduced fingerprint

II.3.4.1 Reduced fingerprint computation

The synchronization block can retrieve a query video sequence based on the correlation of all DWT coefficients in each frequency band of a frame. The Reduced fingerprint block aims at reducing the number of DWT coefficients required for matching a video sequence.

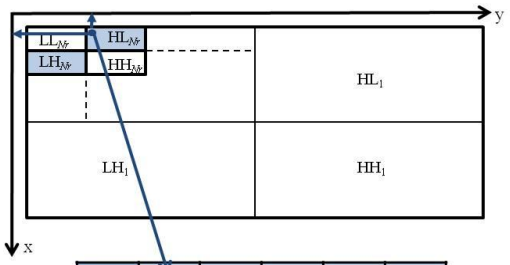
The reduced fingerprint is computed similarly to the fingerprint proposed in Section II.3.3.2 following the steps 1-3 (formatting, color space conversion, wavelet coefficients). The difference appears in the fourth step, the coefficients selection. In the case of the reduced fingerprint, the coefficients selection (*i.e.* the fingerprint) aims at conveying information about the spatial

distribution of salient features within the frames. Consequently, the 2D-DWT coefficients are selected depending on the role of the video sequence (reference or query).

For the reference video sequences, the *R* = 360 highest absolute value coefficients from the HL_{Nr} and LH_{Nr} frequency sub-bands of the transform *V* component, together with their locations are selected and stored in the coefficients matrix (as illustrated in Fig.II.20). The coefficient matrix in Fig.II.20.a illustrates the fingerprint of a sampled frame, while the fingerprint of an entire reference video sequence is presented in Fig.II.20.b and it is called the rank matrix.

The rank matrix is filled-in with all the fingerprints computed on then *N* sampled frames. The fingerprints of the frames consist of *R* = 360*,* 2D-DWT coefficients sorted in a decreasing order of their absolute values. It can be considered that the coefficients are disposed on 360 ranks (where "1" corresponds to the highest absolute value coefficient). This approach will turn to be particularly useful for fingerprint matching.

In the computation of the fingerprint for a query video sequence, the absolute value 2D-DWT coefficients are selected from the HL_{Nr} and LH_{Nr} frequency sub-bands of the V transform component from the locations indicated as salient by the reference coefficients matrices. After selecting the salient coefficients from every selected frame of the reference video, the rank matrix will be obtained.



DWT coefs	0.82	0.79	0.65	...	0.02
x	15	39	43	....	3
y	62	22	29	....	47

(a)

Frames \ Ranks	1	2	3	...	359	360
1	0.822	0.540	0.272	....	0.056	0.056
2	0.465	0.458	0.405		0.031	0.030
...	....	....	....	....	....	....
N-1	0.794	0.687	0.545	....	0.165	0.165
N	0.530	0.491	0.436	....	0.072	0.072

(b)

Fig.II.20: (a) Coefficients matrix for a frame, (b) Rank matrix of DWT coefficients

II.3.4.2 Reduced fingerprint matching

Having the reduced fingerprints of the query and the reference they are matched analogously to the fingerprint matching done in Section II.3.3.3 with the normalized correlation coefficient.

The difference is the fact that the wavelet coefficients are disposed in ranks for the reduced fingerprint, hence the measures in the normalized correlation given by the formula in (II.13) are the following: f_k and t_k designate the 2D-DWT coefficients of the query and the reference videos respectively, on a rank *k*, $\bar{f}_k, \bar{t}_k$ are the mean values of the 2D-DWT coefficients on the considered rank, while $\sigma_{f_k}, \sigma_{t_k}$ are the related standard deviations, respectively. *N* designates the number of 2D-DWT coefficients in every rank *k*, *i.e.* the number of selected frames in each video sequence.

A perfect match (identity) between the query and the reference rank is obtained when $|\rho| = 1$; a value $\rho = 0$ indicates no correlation between f_k and t_k.

$$\rho = corr_k(f,t) = \frac{1}{N-1}\sum_{x,y}\frac{(f_k(x,y) - \bar{f}_k)(t_k(x,y) - \bar{t}_k)}{\sigma_{f_k}\sigma_{t_k}} \qquad \text{(II.13)}$$

The normalized correlation is computed between the absolute values of the 2D-DWT coefficients disposed on ranks, *i.e*. the columns of the rank matrix. Such a strategy is justified by the statistical investigation of the 2D-DWT coefficient behavior in [MIT 07], [DUM 08]: it was proved that the values taken by a rank in the 2D-DWT coefficient hierarchy feature stationarity and the corresponding probability density function was estimated using a mixture of Gaussian laws. Hence, the stationarity property of these coefficients ensures a certain degree of independence of the results with respect to the experimental corpus.

In practice, in order to be able to also retrieve content preserving replicas, the absolute value of the normalized correlation should be compared to some threshold *T*; should $\rho \geq T$, then the query and the reference ranks are considered as similar.

The value of the *T* threshold is statistically determined according to the Rho test on correlation [WAL 02]. This test is individually applied to each of the *R* = 360 ranks under investigation; the null/alternative hypotheses are:

$$\begin{cases} H_0: \textit{the ranks are not correlated} \\ H_1: \textit{the ranks are correlated} \end{cases}$$

A match between the query and the reference video sequences is obtained when the majority of ranks (*i.e.* more than *R/2* = 180) are correlated and when the number or selected frames N is larger than or equal to a threshold *E* = 10 frames. Should the majority of ranks be uncorrelated, or the threshold *E*< 10, the query and the reference video sequences are considered as distinct.

Assuming the *k* ranked absolute value 2D-DWT coefficients from the query and from the reference video sequence are *i.i.d.* (identically and independently distributed) and that they follow a Gaussian distribution, and assuming the H_0 is true, the t_{test} value of the test statistics, see (II.13), follows a Student probability density function of *N-2* degrees of freedom:

$$t_{test} = \rho \cdot \sqrt{\frac{(N-2)}{1-\rho^2}}, \qquad \text{(II.13)}$$

where *N* and ρ are the same as above.

If $t_{test} \leq z_{\alpha/2}$ (where $z_{\alpha/2}$ is the α-point value of the above-mentioned Student law), then the H_0 hypothesis is accepted, *i.e.* the 2D-DWT coefficients on the *k* rank are not correlated. If $t_{test} > z_{\alpha/2}$ the H_1, hypothesis is accepted, *i.e.* the 2D-DWT coefficients on the *k* rank are correlated.

In the experiments presented in this thesis, a significance level of $\alpha = 0.05$ was considered.
Note that in our application, the Rho test is run properly. First, the stationarity behavior of the 2D-DWT coefficients [MIT 07], [DUM 08] and the original video pre-processing ensures the *i.i.d.* behavior for the tested coefficients. Secondly, the robustness of the Rho test for non-Gaussian data may be invoked [WAL 02] in this case.

II.4 TrackART possible configurations

Due to the mathematical principles on which the TrackART method is built upon, the method can be used in two configurations.

The first configuration, denoted as TrackART Full Fingerprint consists the system illustrated in Fig.II.21: offline phase: pre-processing and offline localization and online phase: pre-processing, online localization, and fingerprint. Consequently, the TrackART Full Fingerprint video fingerprinting method outputs results based on the fingerprint block. The decision in this block is based on the Rho test on correlation between full fingerprints of the query and reference video sequences.

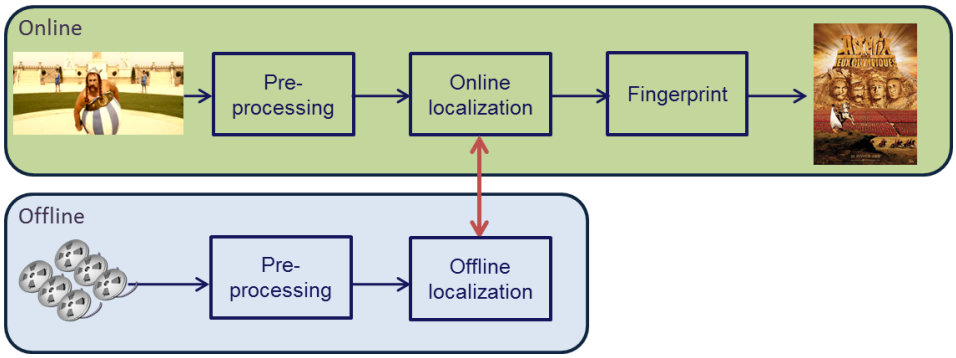


Fig.II.21 TrackART Full Fingerprint functional schema

The second configuration, denoted as TrackART Reduced Fingerprint consists in the system illustrated in Fig.II.22: offline phase - pre-processing and offline localization and online phase: pre-processing, online localization, fingerprint, reduced fingerprint. Consequently, the TrackART Reduced Fingerprint video fingerprinting system outputs the results based on the fingerprint matching block. The decision in this block is based on the Rho test on correlation between the reduced fingerprints of the query and reference video sequences.

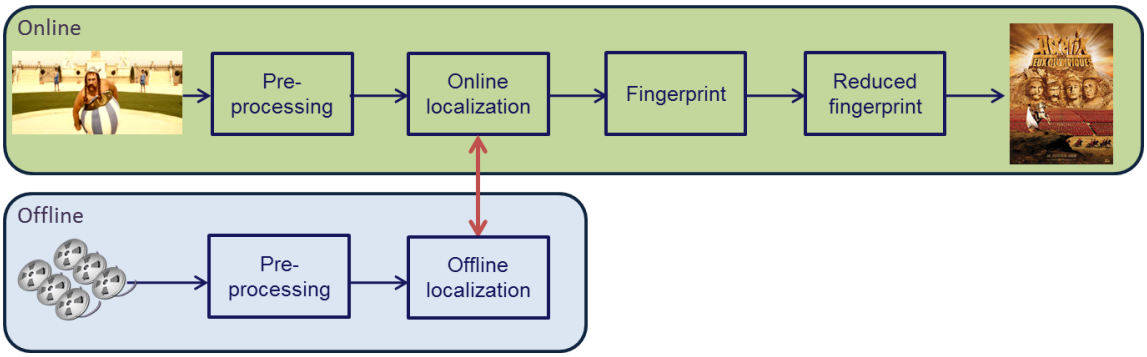


Fig.II.22 TrackART Reduced Fingerprint functional schema

II.5 Conclusion

Part II of the present thesis is devoted to the specification of a new fingerprinting system referred to as TrackART, *cf*. Fig.II.1.

From the structural point of view, TrackART is characterized by the following building blocks:

- The offline phase: enables the localization of a query sequence within a reference video sequence. Its purpose is to process the reference video collection and to map the visual content to a new representation space. It consists of two stages:
    - Pre-processing stage: achieves a common formatting for the reference video sequences by means of sequence of basic operations like spatial and temporal sampling, letterbox removal.
    - Offline localization: provides a framework which can ensure the localization of the query sequence among the reference sequences. This framework is the bag of words approach which consists in: (1) - identifying local features in all the reference keyframes, (2) -describing the local features with a formal descriptor, (3) clustering all the local descriptors into a visual word vocabulary, (4) describing each reference keyframe as a collection of visual words (bag of words), (5) weighting the visual words in each keyframe according to their relative frequencies in both the keyframe and the reference vocabulary, (6) organizing an inverted index file which keeps for each visual words its occurrences in the reference keyframes.
- The online phase: a query video sequence is given to the system and its identity is inquired. It consists in four stages:
    - Pre-processing stage: formats the query video sequence with the common formatting done for the reference video sequences.
    - Online localization: provides possible starting locations of the query within the reference video sequence. It consists in two steps: (1) –identifying matching reference keyframes for the query keyframes, (2) re-ranking the matched reference keyframes according to a geometric consistency verification.

- Fingerprint: computed the fingerprints of the query and reference video fingeprints
- Reduced fingerprint: reduces the amount of information needed for identifying a query video sequence. It consists of two steps: the reduced fingerprint computation and the reduced fingerprint matching.

The novelty of the TrackART video fingerprinting system can be identified at two levels.

First, at the offline phase was obtained by reconsidering, adapting and integrating state of the art image processing algorithms for fingerprinting purposes.

Second, at the online phase, the fingerprint and reduced fingerprint blocks are proposed in the present thesis. They are specified and designed so as to empower the fingerprint with the mathematical properties of the DWT coefficients and to grant statistical error control in the fingerprint matching.

The method thus obtained is *a priori* able to cope with two real life applicative characteristics:

- No constraint is imposed on the query and reference sequences length; the query can have an arbitrarily length and the localization and fingerprint modules are able to position it at the corresponding starting point in the reference sequence.
- No constraint is imposed on the distortions; the localization procedure was designed so as to take into account the effects of not only computer generated distortions but also live camcorder recording as well.

From the functional point of view TrackART is expected to answer the main challenges of a video fingerprinting system:

- The uniqueness property of fingerprints should be ensured by the fact that the video features are selected according to a mathematical model representing the visual content (the wavelet coefficients).
- The robustness property of fingerprints should be achieved by the fact that the mathematical models governing the selected features are robust to frame content and aspect distortions as well as video format distortions.
- The scalability to large scale databases should be ensured by the fact that a query localization procedure is employed and by the fact that the all the algorithms have fast implementations.

The relation between the TrackART method and the state of the art limitations are presented in Table II.2.

These *a priori* properties are experimentally validated in Part III of the present thesis.

Constraints	Challenge	Current limitation	Thesis contributions
Uniqueness	Accurate representation of the video content	Heuristic procedures	Fingerprint computation independent of random, time-variant conditions: ❑ stationary/ergodic fingerprints ❑ 2D-wavelet coefficients
Robustness	Mathematical ground In-theater live camcorder recording	Heuristic procedures No related method reported in the state-of-the-art	Mathematical decision rule in fingerprint matching: ❑ method based on a repeated statistical test ❑ statistical error control
Search efficiency	Scalability	Very few full scalable mono-modal methods reported in the state-of-the-art	Scalable method ❑ automatic retrieval procedure ❑ O(n) complexity for fingerprint computation ❑ O(nlog(n)) complexity for fingerprint matching with respect to the fingerprint size

Table II.2. Camcorder recording robust video fingerprinting: constraints, challenges, state of the art limitations and thesis contributions.

References

[AHE 98] Aherne, F.,Thacker, N., Rockett, P., "The Bhattacharyya metric as an absolute similarity measure for frequency coded data", *Kybernetika*, Vol. 34, pp. 363–368, 1998.

[LEU 01] Leung, T., Malik, J., "Representing and recognizing the visual appearance of materials using three-dimensional textons", *International Journal of Computer Vision*, vol. 43, pp.29–44, June 2001

[LIN 98] Lindeberg, T., "Feature detection with automatic scale selection", *International Journal*

of Computer Vision, Vol. 30, pp. 77–116, 1998.

[BAE 99] Baeza-Yates, R., Ribeiro-Neto, B. "Modern Information Retrieval", *ACM Press*, ISBN: 020139829, 1999.

[BAY 08] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., "Speeded-up robust features (SURF)", *Computer Vision and Image Understanding*, 2008.

[BEL 02] Belongie, S., Malik, J.,Puzicha, J., "Shape matching and object recognition using shape contexts", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 509–522, 2002.

[BER 11] Beran, V., Řezníček, I., ''Brno University of Technology at TRECVid 2011 - Content-based Copy Detection", 2011

[BIE 98] Biederman,I. "Recognition-by-components: A theory of human image understanding", *Psychological Review*, vol. 2, no. 94, pp. 115–147, 1987.

[BUC 99] Buccigrossi, R., Simoncelli, E., "Image Compression via Joint Statistical Characterization in the Wavelet Domain", *IEEE Trans. on Image Processing*, Vol.8, pp. 1688 – 1700, 1999.

[CHU 04] Chum, O., Matas, J., Obdrzalek, S., "Enhancing RANSAC by generalized model optimization", *In Proc. ACCV*, 2004

[CSU 04] Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C., "Visual Categorization with Bags of Keypoints", *Workshop on statistical learning in computer vision*, ECCV 1, 22

[CHU 05] Chum, O., Matas, J., "Matching with PROSAC - progressive sampling consensus", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005

[DAU 92] Daubechies, I., Ten lectures on wavelets, SIAM 1992.

[DOU 08] Douze, M., Gaidon, A., Jegou, H., Marszałek, M., Schmid, C., "INRIA-LEAR'S VIDEO COPY DETECTION SYSTEM", *TREC Video Retrieval Evaluation*, 2008

[DOR 03] Dorko, G. ,Schmid, C., "Selection of scale-invariant parts for object class recognition", *Proceedings. Ninth IEEE International Conference on Computer Vision*, Vol. 1,pp. 634- 639, 2003.

[DUM 07] Dumitru, Duta.S, O., Mitrea, M., Prêteux, F.,"Gaussian hypothesis for video watermarking attacks:Drawbacks and Limitations", EUROCON, 2007, Warsaw, Sept 2007, pp 849-855.

[DUM 08] Dumitru, O., Mitrea, M., Prêteux, F., "Video Modelling in the DWT domain", *Proceedings. SPIE*, Vol. 7000, pp. 7000 OP: 1-12. Strasbourg, 2008.

[DUM 10] Dumitru, O., Mitrea, M., Prêteux, F "Noise sources in robust uncompressed video watermarking", PhD thesis, Université Pierre et Marie Curie, Paris, 2010

[FIS 81] Fischler, A. M., Bolles, C.R., "Random sample consensus", Comm. ACM, Vol. 24, pp. 381–395, 1981.

[FRE 77] Freidman, J. H., Bentley, J. L., Finkel, R. A., "An algorithm for finding best matches in logarithmic expected time", *ACM Trans. Math. Softw*, Vol. 3, pp. 209–226, 1977

[FRE 91] Freeman, W., Adelson, E., "The design and use of steerable filters", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.13, pp. 891–906, 1991.

[GAR 96] Garding, J., Lindeberg, T., "Direct computation of shape cues using scale-adapted spatial derivative operators", *International Journal of Computer Vision*, Vol. 17, pp. 163–191, 1996.

[GOO 96] Van Gool, L., Moons, T., Ungureanu, D., "Affine / photometric invariants for planar intensity patterns", *In Proceedings of the 4th European Conference on Computer Vision*, pp. 642–651, Cambridge, UK,, 1996.

[HAR 84] Harris, C., Stephens, M. "A combined corner and edge detector", *in Alvey Vision Conference*, pp. 147–151, 1988.

[HAR 88] Harris, C., Stephens, M., "A combined corner and edge detector", *In M. M. Matthews, Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, 1988.

[HAR 04] Hartley, I. R., Zisserman, A., "Multiple View Geometryin Computer Vision", *Cambridge University Press*, ISBN: 0521540518, second edition, 2004.

[JEG 08] Jégou, H., Douze, M., Schmid, C., "Hamming embedding and weak geometric consistency for large scale image search", *European Conference on Computer Vision–ECCV,* pp. 304-317, 2008.

[JEG 07] Jégou, H., Harzallah, H., Schmid, C., "A contextual dissimilarity measure for accurate and efficient image search", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007

[JIA 11] Jiang, M., Shu, F., Tian., Y, Huang., T., "Cascade of Multimodal Features and Temporal Pyramid Matching", *Proceedings* of *TRECVID 2011*.

[KE 04] Ke, Y., Sukthankar, R., " PCA-SIFT: A more distinctive representation for local image descriptors", *In Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 511-517, Washington, USA, 2004.

[LAZ 03a] Lazebnik, S., Schmid, C., and Ponce, J. "A sparse texture representation using affine-invariant regions". *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 319–324, Madison, Wisconsin, USA, 2003.

[LAZ 03b] Lazebnik, S., Schmid, C., and Ponce, J. "Affine-invariant local descriptors and neighborhood statistics for texture recognition", *In Proceedings of the International Conference on Computer Vision*, pp. 649–655, Nice, France, 2003.

[LEP 05] Lepetit, V., Lagger, P., Fua, P., "Randomized trees for real-time keypoint recognition." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2005.

[LIN 97] Lindeberg, T, Garding, J., ”Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure”, *Image and Vision Computing*, Vol. 15, pp. 415–434, 1997.

[LIN 98] Lindeberg, T. “Feature detection with automatic scale selection,” *IJCV*, vol. 30, no. 2, pp. 77–116, 1998.

[LIU 11] Liu, Z., Zavesky, E., Zhou, N., Shahraray, B., “AT&T Research at TRECVID 2011 – Content Copy Detection Task”, 2011

[LOW 99] Lowe, D. G., “Object recognition from local scale-invariant features”, *Proceedings of the International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[LOW 04] Lowe, D. G., “Distinctive Image Features from Scale-Invariant Keypoints”, *international journal of computer vision*, Vol. 60, pp. 91-110.

[MAL 99] Mallat, S., A wavelet tour of signal processing, ProcessingAcademic Press, San Diego 1999

[MAT 02] Matas, J., Chum, O., Urban, M., Pajdla, T., “Robust wide-baseline stereo from maximally stable extremal regions,” in Proceedings of the British Machine Vision Conference, pp. 384–393, 2002.

[MIK 02] Mikolajczy, K., Schmid, C., “An affine invariant interest point detector”, *ECCV* 2002

[MIK 04] Mikolajczyk, K, Schmid, C., “Scale and affine invariant interest point detectors”, *International Journal of Computer Vision*, vol. 1, no. 60, pp. 63–86, 2004.

[MIK 05a] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L., “A Comparison of Affine Region Detectors”, Journal International Journal of Computer Vision archive, Vol. 65, pp. 43 – 72, Nov 2005

[MIK 05b] Mikolajczy, K., Schmid, C., “A performance evaluation of local descriptors”, *IEEE Transactions on Pattern analysis and Machine Inteligence*, Vol.27, No 10, Oct 2005

[MIT 04a] Mitrea, M., Prêteux, F., Vlad, A., Fetita, C. “The 2D-DCT coefficients statistical behaviour: A comparative analysis on different types of image sequences”, Journal of Optoelectronics and advanced materials, Vol. 6, No. 1, 2004, pp.95-102.

[MIT 04b] Mitrea, M., Prêteux, F., Vlad, A., “Watermarking oriented video modelling in the wavelet domain”, WSEAS Transactions on Mathematics, Vol. 3, No. 1, January 2004, pp.282-287.

[MIT 04c] Mitrea, M., Zaharia, T., Prêteux, F., Vlad, A., “Accurate data modelling for watermarking applications”, Proc. IMA International Conference on Mathematics in Signal Processing VI, Cirenster, UK, 2004, pp.167-170.

[MIT 07] Mitrea, M., Dumitru, O., Prêteux, F., Vlad, A., “Zero-memory information sources approximating to video watermarking attacks”, *Proceedings of the International Conference on Computational Science and Its Applications*, Vol. 3, pp. 445 – 459, Kuala Lumpur, Malaysia, 2007.

[MPI 12] http://www.open-mpi.org/

[MUJ 09] Muja, M., Lowe, G. D., "FAST APPROXIMATE NEAREST NEIGHBORS WITH AUTOMATIC ALGORITHM CONFIGURATION", *in VISAPP International Conference on Computer Vision Theory and Applications* 2009

[NI 09] Ni, K., Jin, H., Dellaert, F., "GroupSAC: Efficient Consensus in the Presence of Groupings.", *In International Conference on Computer Vision*, 2009.

[NIP 06] Nister, D., Stewenius, H., "Scalable recognition with a vocabulary tree", *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 2161-2168, 2006.

[LAF 01] Lafferty, J., Zhai, C., "Document language models, query models, and risk minimization for information retrieval", *In SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 111–119, New York, NY, USA, 2001.

[LIN 97] Lindeberg, T., Garding, J., "shape-adapted smoothing in estimation of 3-D shape cues from a_ne deformations of local 2-D brightness structure", *Image and Vision Computing*, vol. 15, pp. 415-434, 1997.

[OGI 02] Ogilvie, P., Callan, J., "Language models and structured document retrieval. In Proceedings of the Initiative for the Evaluation of XML Retrieval", *Workshop (INEX 2002)*, 2002.

[PER 09] Perdoch, M. and Chum, O. and Matas, J. "Efficient Representation of Local Geometry for Large Scale Object Retrieval", *In proceedings of CVPR09*. Jun 2009.

[PHI 07] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., "Object retrieval with large vocabularies and fast spatial matching", *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[PHI 10] Philbin, J., "Scalable Object Retrieval In Very Large Image Collections", PhD thesis 2006

[PHI 12] Philbin, J., "FASTCLUSTER: A library for fast, distributed clustering". http://github.com/philbinj/fastcluster, 2012.

[ROB 77] Robertson, S., "The probability ranking principle in information retrieval", *Journal of documentation*, pp. 294 – 304, 1977.

[SCH 97] Schmid, C. and Mohr, R., "Local grayvalue invariants for image retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19 (5), pp. 530–535, 1997

[SE 02] Se, S., Lowe, D., and Little, J., " Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks", *International Journal of Robotics Research* 21(8), pp. 735–758

[SIV 03] Sivic, J., Zisserman, A., "Video Google: A Text Retrieval Approach to Object Matching in Videos", *Proc. of the Ninth IEEE International Conference on Computer Vision*, Vol. 2, pp. 470-1477, 2003

[SIV 04a] Sivic, J., Schaffalitzky, F., Zisserman, A. "Object level grouping for video shots", *In Proceedings of the 8th European Conference on Computer Vision*, pp. 724–734, Prague, 2004.

[SIV 04b] Sivic, J., Zisserman, A., "Video data mining using configurations of viewpoint invariant regions", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 488–495, Washington, 2004.

[SIV 06] Sivic, J., “Efficient visual search of images and videos”, PhD thesis 2006

[SCH 03]Schaffalitzky, F. and Zisserman, A. “ Automated Location matching in movies”, *Computer Vision and Image Understanding*, 92(2), pp. 236–264, 2003.

[SMI 97] Smith, M. S., Brady, M. J, “SUSAN — A new approach to low level image processing”, *International Journal of Computer Vision*, vol. 23, no. 34, pp. 45–78, 1997.

[SIV 09] Sivic, J., Zisserman, A., “Efficient visual search of videos cast as text retrieval”, *in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*

[SIL 08] Silpa-Anan, C. and Hartley, R., “Optimised KD-trees for fast image descriptor matching”, *In CVPR,* 2008.

[SZE 10] Szeliski, R., “Computer Vision: Algorithms and Applications”, ISBN 978-1-84882-935-0, 2010

[TUY 00] Tuytelaars, T., Van Gool, L., “Wide baseline stereo matching based on local, affinely invariant regions,” *in Proceedings of the British Machine Vision Conference*, pp. 412–425, 2000.

[TUY 04] Tuytelaars, T., Van Gool, L., “Matching widely separated views based on affine invariant regions”, *International Journal of Computer Vision*, vol. 1, no. 59, pp. 61–85, 2004.

[TUY 08] Tuytelaars, T., Mikolajczyk, K.,“Local Invariant Feature Detectors: A Survey”, *Journal Foundations and Trends in Computer Graphics and Vision archive*, Vol. 3, pp. 177-280, Jan 2008.

[TIR 10] Tirilly, P., Claveau, V., Gros, P., “Distances and weighting schemes for bag of visual words image retrieval”, *In Proceedings of the international conference on Multimedia information retrieval*, pp. 323-332, 2010.

[VAR 05] Varma, M., Zisserman, A., “Unifying statistical texture classification frameworks. Image and Vision Computing”, Vol. 22, pp. 1175–1183, 2005.

[YAN 07] Yang, J., Jiang, G. Y., Hauptmann, G. A., Ngo, W. C., “Evaluating Bag-of-Visual-Words Representations in Scene Classification”

[WIT 99] Witten, H. I., Moffat, A., Bell, T., “Managing Gigabytes: Compressing and Indexing Documents and Images”, Morgan Kaufmann Publishers, ISBN:1558605703, 1999

[WAL 02] Walpole, E. R, Myers, R.H, Myers, L. S., Ye, K., “Probability & Statistics for Engineers and Scientists”, Pearson Educational International, 2002

[ZHA 11] Zhao, L. W., Borth, D., Breuel, M. T., “University of Kaiserslautern at TRECVID 2011 - Content-based Copy Detection Task”, 2011

PART III: TRACKART – EXPERIMENTAL RESULTS

Abstract

The present section relates to the experiments. The TrackART video fingerprinting system advanced by the present thesis is evaluated in industrial partnership with professional players in cinematography special effects (Mikros Image) and with the French Cinematography Authority (CST - Commision Supérieure Technique de l'Image et du Son).

Two use cases have been incrementally considered: (1) computer generated replica video retrieval and (2) live camcorder recorded video retrieval. The reference dataset was composed of 14 hours of video content from different movies produced in Ile de France (*e.g.* Asterix), under the framework of the HD3D-IIO and HD3D2 CapDigital Competitiveness Cluster Projects. The query dataset was organized differently for each use case. For computer generated replica video retrieval, the query dataset consists of 24 hours of replica video content generated obtained by applying eight types of distortions (*i.e.* brightness increase/decrease, contrast decrease, conversion to grayscale, Gaussian filtering, sharpening, rotations with 2°, stirMark) on 3 hours of original video content from the reference dataset. For the live camcorder recording, the query corpus consisted of 1 hour of live camcorder recorded video content from the reference dataset.

The properties of the TrackART video fingerprinting system were evaluated as following: the robustness property is assessed by two objective evaluation criteria, namely the probability of missed detection (P_{md}) and the recall rate (Rec); the uniqueness property is assessed by two objective evaluation criteria: the probability of false alarm (P_{fa}) and the precision rate (Prec); the scalability property is assessed by an in depth complexity evaluation.

The inner 2D-DWT properties with respect to content preserving attacks (such as linear filtering, sharpening, geometric, conversion to grayscale, small rotations, contrast changes, brightness changes, live camcorder recording), ensure the following results: in the first use case the probability of false alarm reached its null ideal value whereas the missed detection was lower than 0.025, precision and recall were higher than 0.97; in the second use case, the probability of false alarm was 0.000016, the probability of missed detection was lower than 0.041, precision and recall were equal to 0.93

Keywords

Computer generated replica video retrieval, brightness increase/decrease, contrast decrease, conversion to grayscale, Gaussian filtering, sharpening, rotations with 2, StirMark, camcorder recorded replica video retrieval, probability of false alarm, probability of missed detection, precision, recall, complexity evaluation.

Resumé

Ce chapitre porte sur la validation expérimentale. TrackART, le système de traçage du contenu vidéo avancé dans cette thèse, est évalué en partenariat avec des professionnels de l'industrie des effets spéciaux (Mikros Image) et avec l'autorité cinématographie française (CST - Commission Technique Supérieure de l'Image et du Son).

Deux cas d'usages ont été examinés: (1) - recherche des séquences vidéo qui comporte des distorsions générée par l'ordinateur et (2) - recherche des séquences vidéo qui comporte des

distorsions générée par l'enregistrement en salle de cinéma. La basse de donnes de référence est composée par 14 heures de contenu vidéo obtenu à partir de différents films produits en Ile de France (par exemple Astérix), dans le cadre de projets pôle de compétitivité CapDigital, HD3D-IIO et HD3D2. La basse de donne de requête a été organisée différemment pour chaque cas d'usage. Pour le cas de recherche des séquences vidéo qui comporte des distorsions générée par l'ordinateur, la basse de données de requête est constituée par 24 heures de contenu vidéo obtenu en appliquant huit types de distorsions (augmentation/diminution de la luminosité, diminution du contraste, conversion en niveaux de gris, filtrage Gaussien, le rehaussement, rotation 2°, StirMark) sur 3 heures de contenu vidéo original. Pour le cas d'enregistrement en salle de cinéma, le corpus requête consiste en 1 heure de contenu vidéo, *i.e.* 1 heure de originale a été enregistré avec un caméscope.

Les propriétés du système TrackART ont été ensuite évaluées: la robustesse est évaluée selon deux critères d'évaluation objectifs, *i.e.* la probabilité de pertes et le taux de rappel ; l'unicité est également évaluée par deux critères objectives, à savoir la probabilité de fausse alarme et le taux de précision ; la scalabilité est évaluée par la complexité du calcul du chaque block fonctionnelle du système.

Les propriétés intrinsèque des coefficients 2D-DWT en ce qui concerne le distorsions préservant le contenu (tels que le filtrage linéaire, rehaussement, conversion en niveaux de gris, les petits rotations, les changements de contraste et luminosité, l'enregistrement en salle de cinéma), assurent les résultats suivants: dans le premier cas d'usage la probabilité de fausse alarme atteint sa valeur idéale (nulle), la probabilité de détection manquée est inférieure à 0.025, la précision et le rappel sont plus élevé que 0,97 ; dans le deuxième cas d'usage d'autre, la probabilité de fausse alarme est 0.000016, la probabilité de détection manquée était inférieur à 0.041 , la précision et le rappel sont 0.93.

Mots clés

Recherche des séquences vidéo qui comporte des distorsions générée par l'ordinateur, augmentation/diminution de la luminosité, diminution du contraste, conversion en niveaux de gris, filtrage Gaussien, le rehaussement, rotation 2°, StirMark, recherche des séquences vidéo qui comporte des distorsions générée par l'enregistrement en salle de cinéma, la probabilité de fausse alarme,), la probabilité de détection manquée, la précision, le rappel.

III.1 Context

Part III of the current thesis experimentally validates the TrackART method proposed in Part II.

The evaluation of the TrackART method was accomplished under the framework of the HD3D2 project [HD3 11] and in direct partnership with Mikros Image [MIK 12] and the CST-Commission Supérieure Technique de l'Image et du Son [CST 12].

The aim of the HD3D2 project was to develop a platform able to provide film and animation producers with all the tools needed for film making, starting from content production, management and finishing with copyright protection and legal matters.

Mikros Image is a major player in the post production industry (the Oscar in 2010 for short animation movies, La Palme d'Or Cannes 2012), dedicated to high-end visual effects. The CST is an association of professionals from the audiovisual field, in charge of supervising the quality of the production and broadcast of sound and images, whether they are intended for cinema, television or any other medium.

Under this framework, the role of the ARTEMIS department of Institut Telecom; Telecom SudParis was in charge of investigating the state of the art fingerprinting methods and of providing a novel fingerprinting method able to cope with the particularities of the use cases stated by the industrial partners.

Both partners, Mikros Image and the CST proposed a use-case which they found relevant in their business activities. On the one hand, Mikros Image as a post-production company, was interested in having a method able to cope with the distortions induced in video content with the help of the computer. On the other hand, the CST as quality supervisors were interested in researching a fingerprinting method able to address the challenging case of live camcorder recording. To our best knowledge (Section I.5.2), camcording has been addresses by the state of the art video fingerprinting methods only in its computer simulated form and not in its live version.

These two use case have been considered in the present thesis.

III.2 Testing corpus

The reference database for the TrackART method is the HD3D-IIO video corpus which was compiled by the HD3D2 project partners. It consists of 8 video videos Asterix, Chromophobia, Fauteuil d'Orchestre 3, Femme Fatale, Hannibal, Hitman, La Mome and The Last Legion which totalize 14 hours of video content.

The films are divided into chapters as follows: Hitman – 7 chapters (denoted as seqRef1, …, seqRef7), Chromophobia – 7 chapters (denoted as seqRef8, …, seqRef14), Femme Fatale – 6 chapters (denoted as seqRef15, …, seqRef20), The Last Legion – 1 chapter (denoted as seqRef21), Fauteuil d'Orchestre – 5 chapters (denoted as seqRef22, …, seqRef26), Hannibal – 3 chapters (denoted as seqRef27, …, seqRef29), Asterix – 1 chapter (denoted as seqRef30), La Mome – 1 chapter (denoted as seqRef31).

The films in the reference corpus were provided as a collection of images encoded with the tiff format and with HD definition. The resolutions are presented in Table III.1.

Film	Resolution
Asterix	1920 × 1080
Chromophobia	1920 × 1080
Fauteuil d'Orchestre 3	720 × 506
Femme Fatale	720 × 506
Hannibal	1920 × 1080
Hitman	1920 × 1080
The Last Legion	1920 × 1080
La Mome	1920 × 1080

Table III.1 HD3D-IIO video corpus resolution

The content of the HD3D-IIO corpus encompasses scenes with high and still motion, indoor and outdoor scenes, stable and unstable lighting conditions, as illustrated in Fig.III.1.

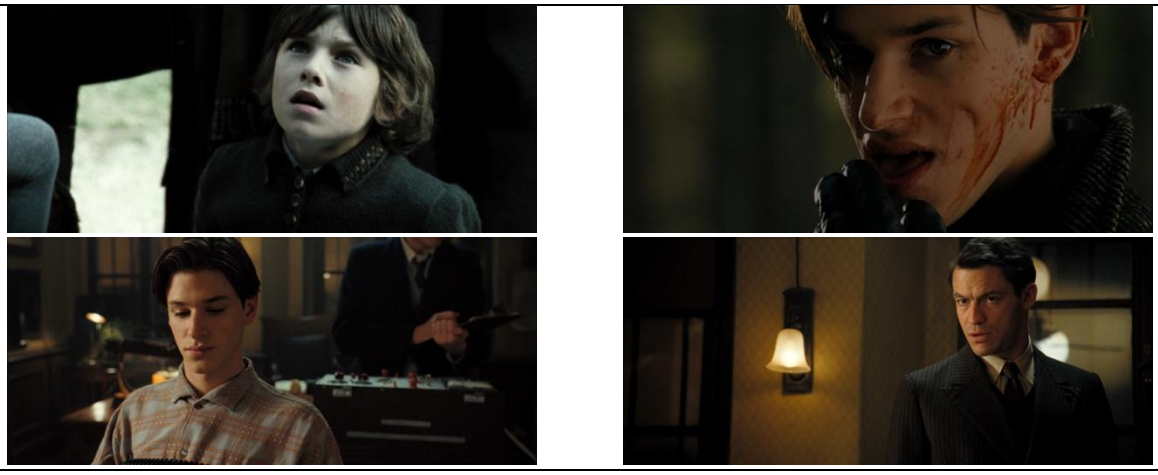
In order to assess the performances of the TrackART video fingerprinting system, the processing of the HD3D-IIO corpus is detailed for every use case in the sequel.

<table>
<tr><td colspan="2">Asterix</td></tr>
<tr><td></td><td></td></tr>
<tr><td></td><td></td></tr>
</table>

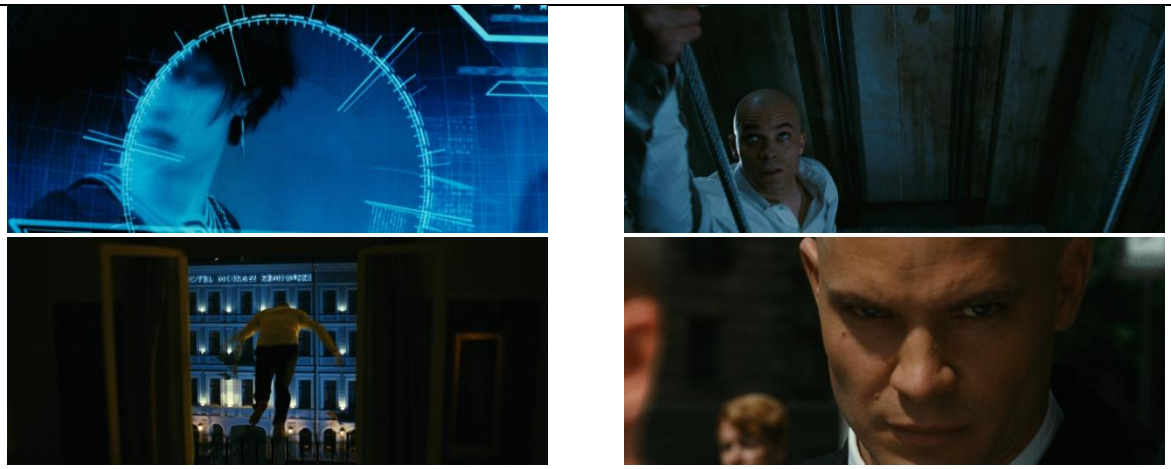
Chromophobia	
	
	

Fauteuil d'Orchestre 3	
	
	

Femme Fatale	
	
	
Hannibal	



Hitman



La Mome



he Last Legion	
	
	

Fig.III.1 Frames from the HD3D-IIO video corpus

III.3 Video retrieval use-case

A video identification and retrieval use case consists in identifying a query video sequence in a reference database of video sequences. When consulting the reference database with a query video sequence, all its replicas should be retrieved. Irrelevant video sequences (not connected to the query) should be ignored.

The reference video database consists of the entire HD3D-IIO corpus totalizing 14 hours of video content. The reference content is structured in its original chapter format, as detailed in Section III.2.

			
Original version			
			
Brightness decrease	Brightness increase	Contrast decrease	Gaussian filtering
			
Grayscale conversion	Rotation 2°	Sharpening	StirMark

Fig.III.2: The replica video sequences

The query video corpus consisted of 1440 video sequences chosen from the reference database, *i.e.* 180 replica video sequences with length of 1 minute for each of the distortions. The replica video sequences were obtained by applying distortions in the video sequences selected from the reference database. The distortions considered are the following: brightness decrease (25%), brightness increase (20%), contrast decrease (25%), linear filtering (Gaussian filter), conversion to grayscale, rotations by 2°, sharpening, and and StirMark attack. The effect of the attack on the video frames is illustrated in Fig.III.2.

Note that the distortions applied on the query video sequences include only frame aspect and frame content distortions and no video format distortions were induced at this stage of experimental work.

Having this experimental set-up, the TrackART video fingerprinting method will be tested under the video retrieval use case in two configurations denoted as TrackArt Full Fingerprint (*cf.*Fig.II.21) and TrackART Reduced Fingerprint (*cf.* Fig.II.22). The difference between the two configurations of the TrackART method is the functional block of the method which provides the results of the system.

In the TrackART Full Fingerprint configuration, the result is given by the fingerprint block which computes the number of matching frames (according to the Rho test on correlation) between the query and the reference video sequences. Consequently the amount of matching frames is considered as the criterion for the video sequences matching. A query is retrieved if *E* of its frames are correlated with the frames of a reference sequence. The *E* threshold needs to be set by taking into account the sampling rate performed in the synchronization step of the Fingerprint block (Section II. 3.3.4) of 1 frame per second (a frame every 25 frames) and the length of the query sequence. Considering the size of the query sequences was of 1 minute (25 x 60 = 1500 frames) each, and that general sampling rate is 1 frame per second, the minimum amount of matching frames was set at *E= 20 i.e.* a third of the total sampled frames. Consequently, if equal or more than *E= 20* matching frames are encountered, the query sequence has been identified in the reference database.

In the TrackART Reduced Fingerprint configuration, the decision criterion is based on the Rho statistical test between the reduced fingerprints of the query and reference video sequences.

III.3.1 TrackART Full Fingerprint evaluation

Analogous to the the evaluation of TrackART Full Fingerprint for the evaluation of the TrackART Reduced Fingerprint configuration, the average precision (Prec) and recall (Rec) rates and as well as the probabilities of false alarm (P_{fa}) and missed detection (P_{md}) are investigated.

In Table III.2 and in Fig.III.3-4 the experimental results are reported. In the Fig.III.3, the precision and recall rated are equal and are illustrated by the red squares. In Fig.III.4, the probability of false alarm is illustrated with blue diamonds and probability of miss detection with red squares.

As it can be observed, the results obtained are excellent, with precision, recall rates and probabilities of false alarm and missed detection close to their ideal values.

Distortion	Precision	Recall	P_{fa}	P_{md}
Brightness decrease	1	1	0	0
Brightness increase	0.983	0.983	0	0.016
Contrast decrease	0.988	0.988	0	0.011
Gaussian filtering	0.994	0.994	0	0.005
Grayscale conversion	0.988	0.988	0	0.011
Rotation 2°	0.866	0.866	0	0.133
Sharpening	0.988	0.988	0	0.011
StirMark	0.988	0.9888	0	0.011
Average	0.975	0.975	0	0.025

Table III.2 Average results for precision and recall rates and for the probabilities of false alarm and miss detection for TrackART Full Fingerprint under the video retrieval use case for different distortions

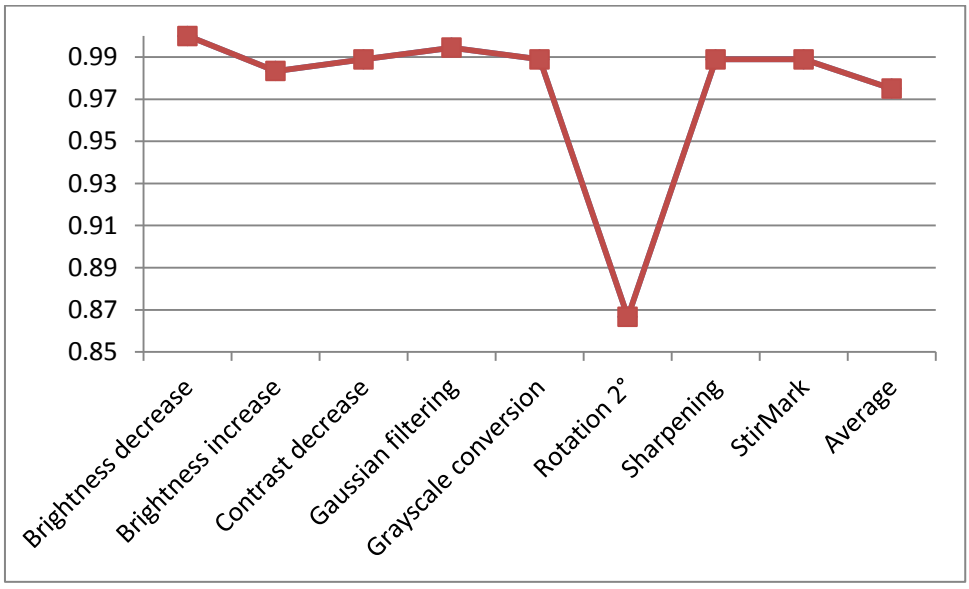


Fig.III.3 Precision and recall rates for TrackART Full Fingerprint in the video retrieval use case depending on the distortions

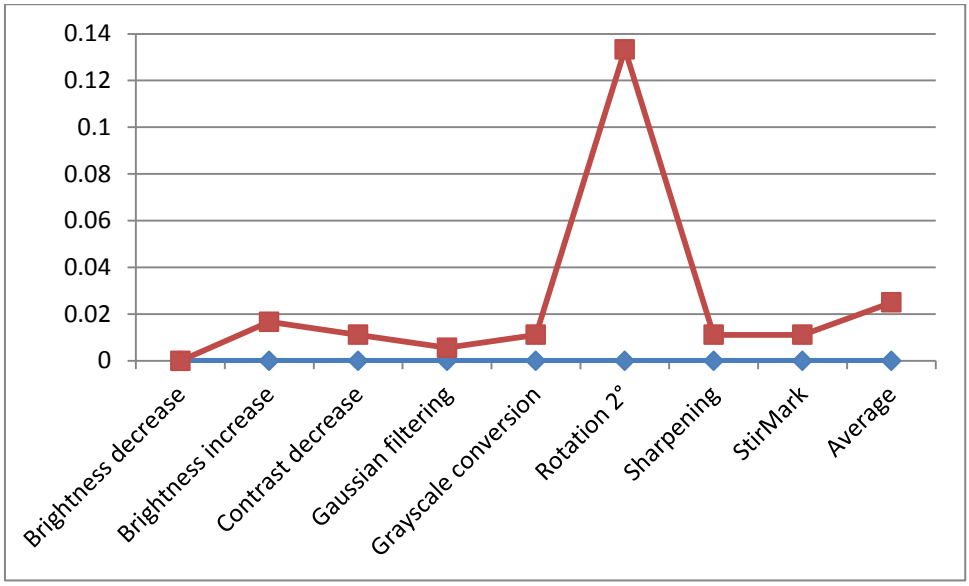


Fig.III.4 The probabilities off false alarm and miss detection for TrackART Full Fingerprint in the video retrieval use depending on the distortions

III.3.2 TrackART Reduced Fingerprint evaluation

This section investigates whether the fingerprint size can be reduced.

In the evaluation of the TrackART Reduced Fingerprint method, the precision and recall rates and the probabilities of false alarm and missed detection are investigated.

Table III.3 presents the average values for the precision (Prec) and recall (Rec) rates, as well as for the probabilities of false alarm (P_{fa}) and miss detection (P_{md}) for the considered distortions. The average values are obtained by averaging the precision/recall/Probability of false alarm/Probability of miss detection obtained individually for each query.

Distortion	Prec	Rec	P_{fa}	P_{md}
Brightness decrease	0.983	0.983	0	0.016
Brightness increase	0.966	0.966	0	0.033
Contrast decrease	0.972	0.972	0	0.027
Gaussian filtering	0.972	0.9722	0	0.022
Grayscale conversion	0.972	0.972	0	0.027
Rotation 2°	0.461	0.461	0	0.466
Sharpening	0.966	0.966	0	0.033
StirMark	0.883	0.883	0	0.094
Average	0.897	0.897	0	0.090

Table III.3: Average results for precision and recall rates and for the probabilities of false alarm and miss detection for TrackART Reduced Fingerprint under the video retrieval use case for different distortions

In Fig.III.5 the average values for the precision and recall rates are illustrated for each distortion. As the values of precision and recall are identical, they are represented by the squares in red.

Fig.III.6 graphically illustrates the average values of the probability of false alarm (the diamonds in blue) and of the probability of miss detection (the squares in red) depending on the particular attack.

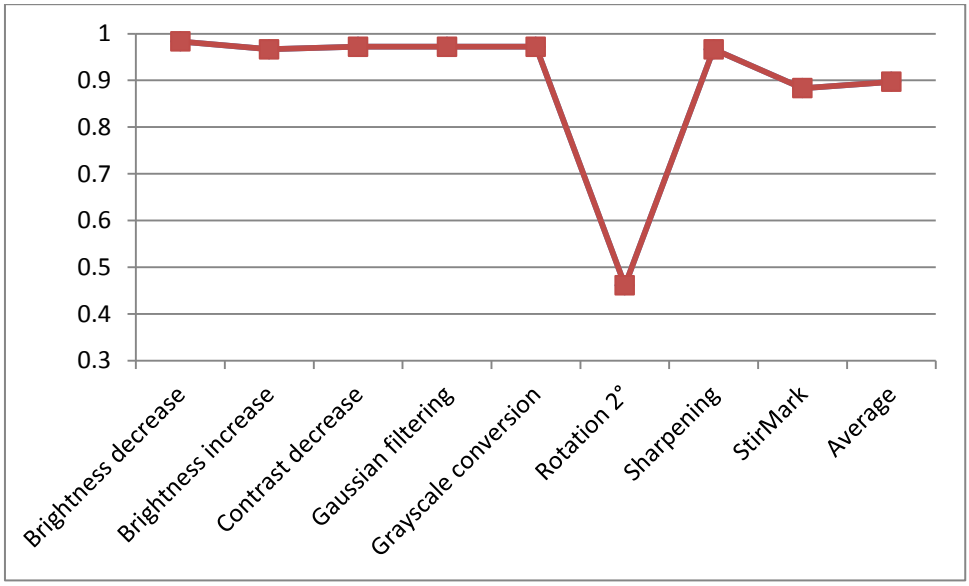


Fig.III.5 Precision and recall rates for TrackART Reduced Fingerprint in the video retrieval use case depending on the distortions

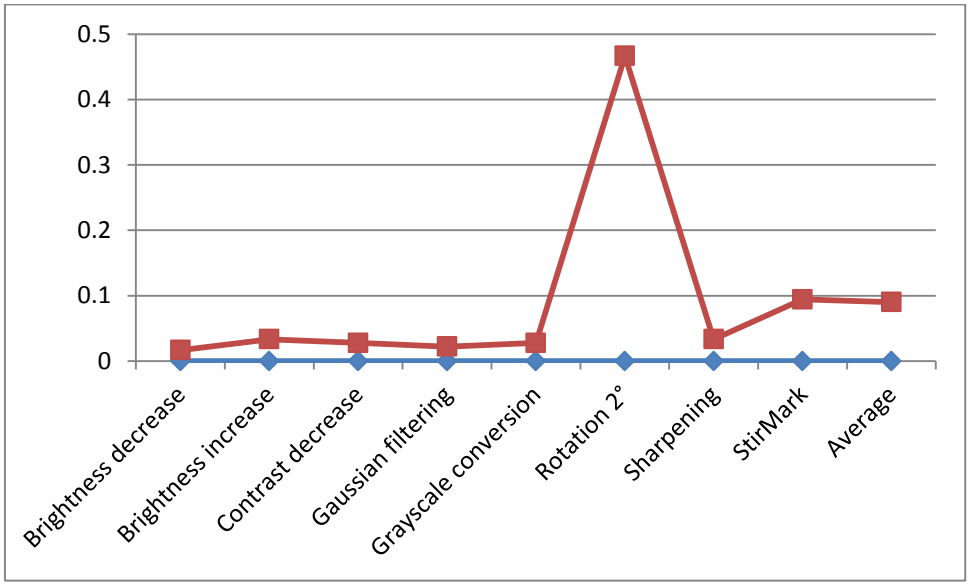


Fig.III.6 The probabilities of false alarm and miss detection for TrackART Reduced Fingerprint in the video retrieval use depending on the distortions

As it can be observed from Table III.3 and from Fig.III.5-6, the retrieval accuracy in terms of average results can be considered as satisfactory, as Prec =0.89 and Rec = 0.89. However, while for some

distortions like brightness decrease/increase, contrast decrease, Gaussian filtering, conversion to grayscale, and sharpening the results are very good, the robustness to rotations by 2° and to StirMark yields quite poor results. This results can be explained as following.

A change in image contrast consists in multiplying each pixel value by a constant; a change in brightness consists in adding a constant to each image pixel. The very good results obtained in the case of brightness increase/decrease, and contrast decrease are achieved due to the normalized correlation coefficient which normalizes the DWT coefficients before computing their correlations, and hence, the changes induced in the images by these distortions are discarded.

The Gaussian filtering and the sharpening multiply each pixel value with a filtering kernel which takes into account a 3×3 pixel neighborhood. The effect of the Gaussian filtering is a smoothing, a blurring on the image, *i.e.* the high frequencies are attenuated. The effect of the sharpening filtering is contrary to the Gaussian filtering, the contours and the edges in the image are enhanced, *i.e.* the low frequencies are attenuated. The wavelet transform is computed as weighted averages and differences of the pixel values, and separate the image content into high and low frequencies, hence discarding the changes induced by such filters.

In the case of the small rotations, the unsatisfactory results can be explained by the fact that the content of the frames is changed (*i.e.* the frames are rotated by 2°, cropped and brought to the resolution of the original frame) and consequently the fingerprint of the rotated video sequence is computed from different content compared to the original sequence. Moreover, the fingerprint is dependent on the positions of the DWT coefficients, while the cropping and resizing change those positions in the rotated sequence.

Considering the StirMark attack, the results can be explained by the nature of the attack which performs local de-synchronization in the frame. The StirMark attack performs a global bending and random displacement in the image, followed by a slight deviation of each pixel (greatest at the center of the picture and almost null at the borders) and a higher frequency displacement. A transfer function that introduces a small and smoothly distributed error into all sample values is applied and a medium jpg compression is performed.

Analogous, to the case of rotations, the StirMark attack modifies the content of the StirMarked frames, and hence the fingerprints are computed from different content.
The conclusion which can be drawn from the results of the first experiment is that the proposed fingerprint is robust to distortions which preserve the content of the video frames and which do not induce local de-synchronizations inside the video frames.

By comparing the results obtained with the TrackART Full Fingerprint configuration and TrackART Reduced Fingerprint, it can be observed that for all the distortions, a significant gain is achieved by the TrackART Full Fingerprint, as presented in Table III.4. The gain is computed as:

$$\text{Gain} = \text{abs}(\text{EvaluationMetric}_{\text{TrackART Full Fingerprint}} - \text{EvaluationMetric}_{\text{TrackART Reduced Fingerprint}}).$$

Specifically, in the case of rotations with 2°, precision and recall are improved with 40%, probability of miss detection decreased with 33%. In the case of the StirMark attack, precision and recall are increased with 10%, while the probability of miss detection is reduced with 9%.

	Gain(%)			
Distortion	Precision	Recall	P_{fa}	P_{md}
Brightness decrease	1.66	1.66	0	1.66
Brightness increase	1.66	1.66	0	1.66
Contrast decrease	1.66	1.66	0	1.66
Gaussian filtering	2.22	2.22	0	1.66
Grayscale conversion	1.66	1.66	0	1.66
Rotation 2°	40.55	40.55	0	33.33
Sharpening	2.22	2.22	0	2.22
StirMark	10.55	10.55	0	8.33
Average	7.77	7.77	0	6.52

Table III.4 Gain obtained in the performances of TrackART Full Fingerprint configuration over the TrackART Reduced Fingerprint configuration in the video retrieval use case

The gain obtained in results for the TrackART Full Fingerprint can be explained by two facts. Firstly, the correlation between the frames of the query and reference video sequences performed in the Fingerprint block is based on a statistical ground and can provide reliable results. Secondly, the computation of the video fingerprint employed information (*i.e.* location of the DWT coefficients) from the video frames which is subject to change in distortions like the rotations with 2° or the StirMark attack and therefore is sensitive to this type of distortions.

These functional gains are obtained at the expense of the fingerprint length, which is in the TrackART Full Fingerprint 17 times larger than the TrackART Reduced Fingerprint. Hence, assuming the case no malicious distortion occurs (*i.e.* content in large archives such as INA [INA 12]) the TrackART Reduced Fingerprint is the best solution as it withstands all the mundane signal processing operations like: Gaussian filtering, sharpening, brightness and contrast changes, or conversion to grayscale.

III.4 Live camcorder recording use-case

The camcording use case consists of tracking an in-theatre camcorder recorded video sequence in a database of original video sequences. In such a case, the attacked video sequences are obtained by capturing with a non-professional camcorder the original sequence which is displayed on a screen.

The reference video database for the camcording use case consists of the entire HD3D-IIO corpus totalizing 14 hours of video content structured in chapters, as detailed in Section III.2.

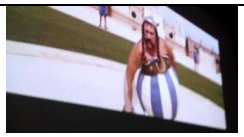
From the HD3D-IIO corpus, a random selection of 60 video sequences (*i.e*. 1 hour) was performed and afterwards camcorder recorded, yielding a query corpus of 60 video replicas. A few frames from the camcorder recorded replicas are exhibited in Fig.III.7.

The camcorder recording was performed in two experimental set-ups, each of them contributing with 30 minutes of recorded video.

The first set-up consisted of video projection in the cinema theatre located at the Commision Supérieure Technique de l'Image et du Son (CST, [CST 12]); the capturing devices were the video cameras of two cell phones, namely an iPhone4 and a Nokia 5800.



Original version



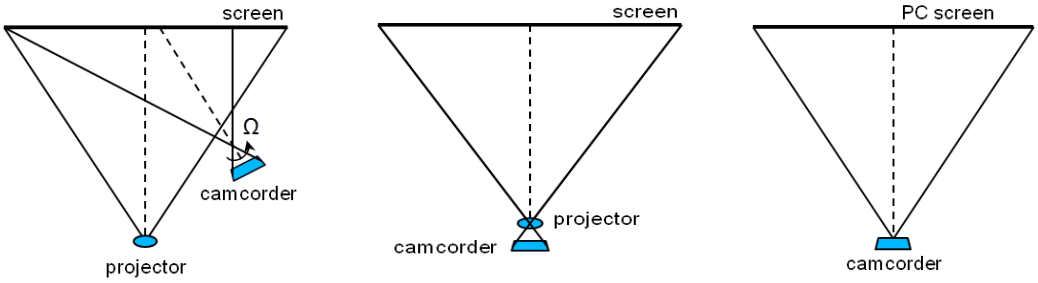



CST captures | ARTEMIS capture

Fig.III.7 Frames from camcorder recorded video sequences

The second set-up consisted of video playing on a PC monitor (DELL 1680 x 1050 pixel resolution, 22" LCD display screen) at the ARTEMIS department [ART 12]; the capturing devices were three cameras: a Canon Legria HF20, a Sanyo Xacti HD1010 and a Canon EOS 7D with a Tokina AT-X PRO objective.

A simplified geometrical representation of any recording process performed in a cinema theatre is given in Fig.III.8.a, the theatre being viewed from the top side view [CHU 08]. The optical axes of the camcorder and of the projector do not usually intersect with the screen at the same point and are not parallel with each other. The angle Ω measures the rotation of the camcorder around its optical axe.



(a) (b) (c)

Fig.III.8 Projection and capture-set; top view

Our experimental set-up for the CST captures is illustrated in Fig.III.8.b: the camcorder was positioned parallel with the axe of the projector; in the ideal case, Ω = 0. However, by its very nature, live camcording introduces random, time variant capturing angles induced by the pirate's involuntarily movements; in our experiments -2° ≤ Ω ≤ 2°.
The experimental set-up for the ARTEMIS video captures is depicted in Fig.III.10.c: the PC screen has two functions, *i.e.* screen and projector, while the camcorder was positioned with its optical axe perpendicular on the PC screen, but the same random capturing angles, -2° ≤ Ω ≤ 2°. were encountered.
In the proposed experimental set-up, the angle Ω was not considered larger than ±2° and the position of the camcorder was approximately maintained in a central position of the screen in order to capture the entire video content displayed on the screen.

III.4.1 TrackART Full Fingerprint evaluation

In the described set-up for the live camcorder recording use case, the evaluation of the TrackART Full Fingerprint video fingerprinting system feature excellent results, in terms of average values of precision, recall, probability of false alarm and miss detection as presented in Table III.5.

Distortion	Precision	Recall	P_{fa}	P_{md}
Live camcorder recording	0.930	0.930	0.000016	0.041

Table III.5 Average results for the live camcorder recording use-case

III.4.2 TrackART Reduced Fingerprint evaluation

The evaluation of the TrackART Reduced Fingerprint video fingerprinting system under the live camcorder use case has been also performed and the results obtained in terms of average values of precision, recall, probability of false alarm and miss detection are presented in Table III.6.

Distortion	Precision	Recall	P_{fa}	P_{md}
Live camcorder recording	0.611	0.611	0	0.333

Table III.6 Average results for the live camcorder recording use-case

The values in Table III.6 point to results far below the minimal requirements for a practical application such as live camcorder recording.

It can be observed that the values in Table III.6 indicate an intuitive discrepancy between the precision and false alarm. This can be explained by the computation formulas of the two metrics.

The probability of false alarm is computed as a rate by taking into the account the length of the database and the length of the query sequences.

Precision denotes the probability of retrieving replica video sequences for a given query out of all the retrieved video sequences. Hence, the precision does not take into account the length of the query nor of the database.

Considering the case there is no false positive and no true positive detected by the system for a certain query video, the value of precision will be put to zero, although the result is a division by zero. This situation yields the worst case for the precision metric, while for the probability of false alarm metric it would yield zero which constitutes the best case.

In Table III.7 the gain the TrackART Full Fingerprint configuration achieves over the TrackART Reduced Fingerprint configuration is presented for each evaluation metric.

It can be observed that the TrackART Full Fingerprint configuration improves the precision and recall rates with 31%, whereas the probability the false alarm is decreased by 29% and the probability of false alarm is increased with 0.001.

The gain obtained by the TrackART Reduced Fingerprint method over the TrackART Full Fingerprint can be explained as in the video retrieval use case by two facts. Firstly, the correlation between the frames of the query and reference video sequences performed in the Fingerprint block is based on a statistical ground and provides reliable results. Secondly, the computation of the video fingerprint employs information (*i.e.* location of the DWT coefficients) from the video frames which is subject to change in distortions like the camcorder recording, *i.e.* random and abrupt geometric transformations.

	Gain(%)			
Distortion	Precision	Recall	P_{fa}	P_{md}
Live camcorder recording	31.94	31.94	-0.001	29.16

Table III.7 Gain obtained in the performances of Partial TrackART over the Full TrackART methods in the camcorder recording use case

Of course, the gain in performances is obtained at the expense of the fingerprint size which is 17 times larger in the case of the TrackART Full Fingerprint configuration. However, TrackART Full Fingerprint is the only solution currently available able to identify live camcorder recorded video content.

III.5 Computational cost

The computational cost for the TrackART video fingerprinting method can be computed by assessing the computational costs of every functional block of the method. Considering the method has two phases (the offline and the online), the computational cost will be computed individually for each of the two phases.

The parameters employed in the TrackART video fingerprinting system are presented in Table III.8 and the computational complexity of the algorithms is presented in Table III.9.

It can be noticed that:

- each operation has a maximum complexity of $O(n \log n)$, where n is the underlying data size; there is only one exception, namely the computation of the SIFT descriptor which has an $O(J \times G^2)$ complexity. However, *J = 8* and *G = 4*, irrespective to the frame size.
- the heavier computational cost is performed in the offline phase, whereas the online phase performs light computational operations.
- the offline phase is computed only once, and then used each time there is a query in the online phase.

The properties featured by the computational cost complements the automatic localization procedure thus granting the scalability for the TrackART method.

Parameters	Meaning
N = 47 163	• total number of reference keyframes
W_{orig}	• the original width for a frame (content dependent)
H_{orig}	• the original height for a frame (content dependent)
L = 16	• the number of iterations employed by the shape adaptation [LIN 97] algorithm to estimate the hessian-affine regions
S = 3	• the size of the scale search (the number of scales investigated) in the shape adaptation algorithm in [LIN 97]
M	• the number of potential interest point detected by the Hessian-Affine detector
J = 8	• number of orientations for the SIFT descriptor
G = 4	• the size of orientation histogram of the SIFT descriptor
*T =*38 466 280	• 10% of the total numbers of SIFT descriptors computed from all the reference keyrames
K=250 000	• the total number of visual words of the vocabulary
P_1	• the number of interest point detected in a keyframe (content dependent); typical values can be between 0 and a few thousands, depending on the size of the image and the content)
P_2	• the number of interest point detected in a keyframe (content dependent); typical values can be between 0 and a few thousands, depending on the size of the image and the content)
W = 352	• the predefined width for the frames before the wavelet computation
H = 288	• the predefined height for the frames before the wavelet computation
B	• the number of matching interest points between two reference keyframes (content dependent)

Table III.8: Parameters employed in the TrackART video fingerprinting method

	Operation		Complexity
Offline phase	Local feature detection	Interest point detection algorithm	$O(W_{orig} \times H_{orig})$
		Scale and affine shape estimation	$O(M \times (S+L))$
	Local descriptor computation		$O(J \times G^2)$
	Vocabulary computation		$O(T \log K)$
	BoW keyframe representation		$O(K)$
	TF-IDF weighting		$O(V)$
	Inverted index		$O(V)$
Online phase	Local feature detection	Interest point detection algorithm	$O(W_{orig} \times H_{orig})$
		Scale and affine shape estimation	$O(M \times (S+L))$
	Local descriptor computation		$O(J \times G^2)$
	Keyframe matching		$O(K)$
	Geometrical verification	Matching between the interest points in two keyframes	$O(P_1 \times P_2 \log P_2)$
		Geometrical transformation estimation and verification	$O(B)$
	Synchronization		$O(W \times H)$
	Computation of the DWT		$O(W \times H)$
	Sorting of the coefficients		$O(R \log R)$
	Fingerprint matching		$O(R \log R)$

Table III.9 The computational complexity of the algorithms employed in the TrackART video fingerprinting method

III.6 Video fingerprint demonstrator

Under the HD3D2 project, one of the deliverables consisted in the implementation and comparison of the TrackART method with two state of the art competitors, namely the 3D-DCT based fingerprint advanced in the study in [COS 06] the visual attention regions based method advanced in the study in [SU 09].

The functionality of the fingerprinting demonstration software is assured by four steps corresponding to the four blocks in Fig.III.9 In the first step the system takes as input an unknown/not identified sequence of video. The second step introduces the fingerprinting process by computing the fingerprint of the input video whereas the third step performs a search in a database of fingerprints. This search step will provide a fingerprint which is the closest match for the fingerprint of the input video sequence. Given the fact that each fingerprint from the database references a known video, the system is able to retrieve the identity of the unknown input video sequence.

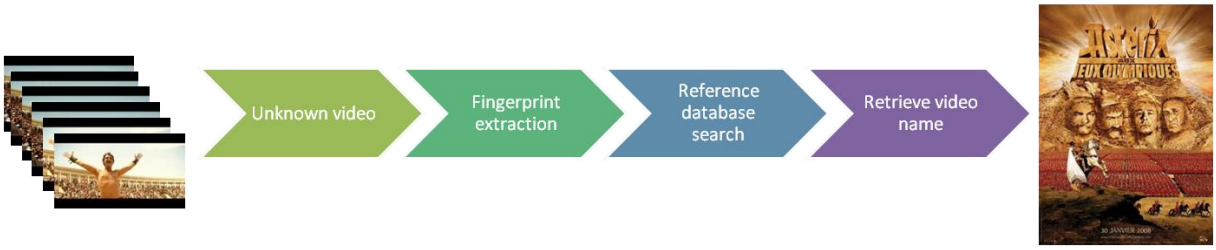


Fig.III.9 Video fingerprinting demonstrator

A demo can have four steps as follows:

- Step 1: Fingerprinting application: a video can be selected from the first drop-down menu “Choose video” and viewed in the Windows Media Player window. From the second drop-down menu “Choose fingerprinting method” a fingerprinting method can be selected. By clicking on the “Generate fingerprint” button, the fingerprint of the selected video will be computed according to the desired method. Step 1 is illustrated in Fig.III.10.
- Step 2: The message box with the text “Fingerprint computed” announces that the fingerprint computation process ended. Following, the extracted fingerprint has to be looked-up in the reference fingerprints database so that the name of the video will be retrieved. The look-up step starts when the “Search the database” button is pressed. Step 2 is illustrated in Fig.III.11.
- Step 3: The message box with the name of the videos indicated as matching by the fingerprinting methods pops up. Step 3 is illustrated in Fig.III.12.



Fig.III.10 Illustration of the step 1 of a fingerprinting demonstrator system

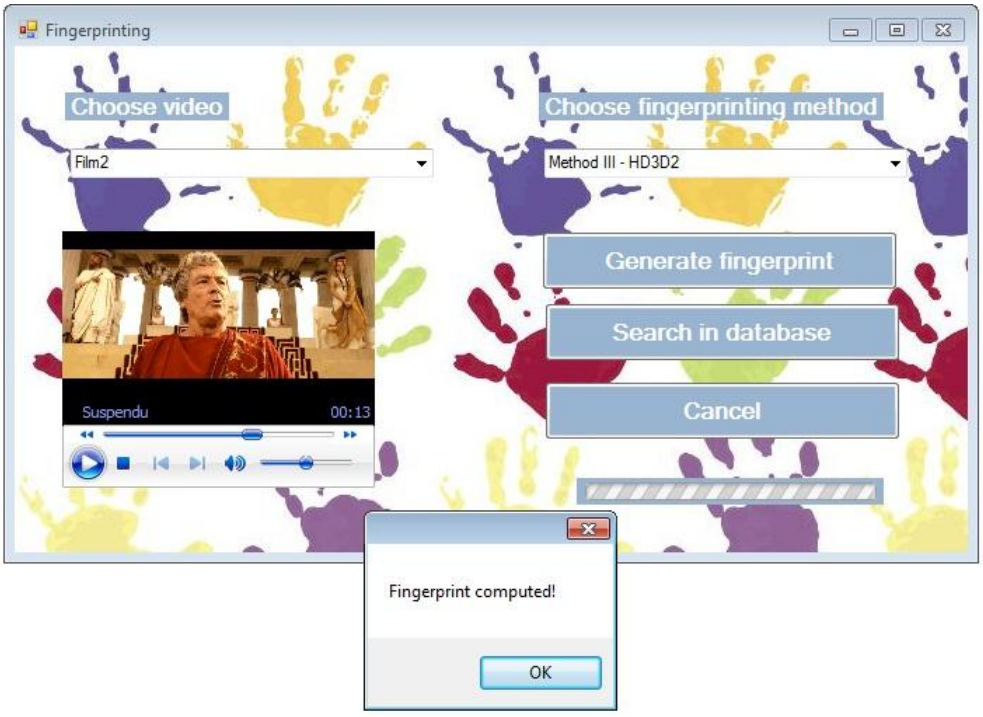


Fig.III.11 Illustration of the step 2 of a fingerprinting demonstrator system



Fig.III.12 Illustration of the step 3 of a fingerprinting demonstrator system

III.7 Conclusion

The TrackART framework introduced in PART II is experimentally validated for two configurations as explained in Section II.4: TrackART Full Fingerprint (*i.e.* which employs all the wavelet coefficients from video frames) and TrackART Reduced Fingerprint (*i.e.* which employs a reduced selection of wavelet coefficients).

The evaluation considers two use cases, namely, the retrieval of video sequences under computer generated dirtsortions and live camcorder recording. The former use case emcompases both mundane (like the Gaussian filtering, sharpening, contrast and brightness changes) and malicious (like small rotations and the StirMark attack) distortions. The latter use case deals with the very sophisticated malicious distortion of live camcorder recording which includes and combines abrupt geometric transformations with brightness, color and contrast variations.

The quantitative results are obtained on a corpus totalizing 14 hours of reference video content and 24 hours of distorted video content for the video retreival use case and 1 hour of live camcorded content for the live camcorder recording use case.

The results show that when assuming mundane computer generated distortions, the TrackART Reduced Fingerprint reaches very good values: the probability of false alarm reaches its ideal null value, the probability of missed detection is 0.026, precision and recall equal to 0.96. This configuration is particularly useful when no malicious distortion occur which is the case for example in large scale archives of INA.

When malicious distortions are encountered, the TrackART Reduced Fingerprint has to be replaced by the TrackART Full Fingerprint. With this configuration, the results are kept at a very good level: the probability of false alarm maintains its ideal null value, the probability of missed detection is 0.025 whereas precision and recall are equal to 0.97. It can be observed that the TrackART Full Fingerprint configuration ensures very good results even under malicious computer generated results, but at the expense of increasing the length of the fingerprint (*i.e.* the size of the full fingerprint is 17 times larger than the size of the reduced fingerprint).

When considering the complex case of live camcorder recording use case, the TrackART Full fingerprint configuration proved itself to be strong enough and featured very good results: the probability of false alarm equal to 0.000016, the probability of missed detection equal to 0.041, precision and recall equal to 0.93. When compared to the CST limits (*i.e.* probabilities of false alarm and missed detection lower that 0.05 and precisions and recall higher than 95%), the results can be considered satisfactory. While complying with the limits set for the probabilities of false alarm and missed detection, the precision and recall are just 2% lower.

The in depth analysis of the computational cost for the TrackART video fingerprinting method proved its feasibility and scalability. Firstly, each operation has a maximum complexity of $O(n \log n)$, where n is the underlying data size; there is only one exception, namely the computation of the SIFT descriptor which has an $O(J \times G^2)$ complexity. However, *J = 8* and *G = 4*, irrespective to the frame size. Secondly, the heavier computational cost is performed in the offline phase, whereas the online phase performs light computational operations. Thirdly, the offline phase is computed only once, and then used each time there is a query in the online phase.

Reference

[ART 12] http://www-artemis.it-sudparis.eu/

[CHU 08] Chupeau, B., Massoudi, A., Lefèbvre F., “In-theater piracy: Finding where the pirate was“ SPIE'08, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, 2008.

[CST 12] http://www.cst.fr/

[HD3 11] http://www.hd3d.fr/

[LIN 97] Lindeberg, T, Garding, J., ”Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure”, *Image and Vision Computing*, Vol. 15, pp. 415–434, 1997.

[MIK 12] http://mikrosimage.eu

PART IV: FINAL CONCLUSIONS AND PERSPECTIVES

Abstract

From the functional point of view, TrackART answers the main challenges for a video fingerprinting system: the uniqueness property of fingerprints is ensured by the fact that the video features are selected according to a mathematical model representing the visual content (the wavelet coefficients); the robustness property of fingerprints is achieved by the fact that the mathematical models governing the selected features are robust to frame content and aspect distortions as well as video format distortions; the scalability to large scale databases is ensured by the fact that a query localization procedure is employed and by the fact that the all the algorithms have fast implementations.

The advanced video fingerprinting method has been tested in two practical use cases proposed by the industrial partners, namely the retrieval of video sequences from database under computer generated distortions and the live camcorder recording.

While the present thesis offers a solution to the nowadays limitations in deploying video fingerprinting in two real-life applications, the perspectives are connected to defining a theoretical model for video fingerprinting. As no such model is currently available in the literature, the video fingerprinting theoretical limits can neither be computed nor explored. To offer a solution to this problem, an information theory based model is advanced. Such a model allows the investigation of the minimal fingerprint size able to identify a video sequence of a given length under prescribed robustness/uniqueness constraints, can be established.

Keywords

Video fingerprinting theoretical model, theoretical limits for video fingerprinting systems, information theory model.

Resumé

D’un point de vue fonctionnel, le système TrackART répond aux enjeux d’aujourd’hui: la propriété d'unicité est assurée par le fait que les empreintes numérique ont été sélectionnés selon un modèle mathématique représentant le contenu visuel (les coefficients d'ondelettes) ; la propriété de robustesse est atteinte par le fait que les modèles mathématiques régissant les empreintes numériques sélectionnées sont robuste à des distorsions du contenu, d’aspect ainsi que de format vidéo ; la propriété de scalabilité pour des bases de données à grande échelle est assurée par le la procédure de localisation de requête et par le fait que tous les algorithmes ont des implémentations rapides.

La méthode de traçage de contenu vidéo avance a été testée dans deux cas d'usage proposés par nos partenaires industriels, notamment (1) - la recherche des séquences vidéo qui comportent des distorsions générées par ordinateur et (2) - la recherche des séquences vidéo qui comportent des distorsions générées par l’enregistrement en salle de cinéma.

Les perspectives sont liées à la définition d’un modèle théorique pour les systèmes de traçage du contenu. En l'absence d'un tel modèle, les limites théoriques d’un système de traçage de contenu vidéo ne peuvent être ni calculées ni explorées. Pour apporter une solution à ce problème, un modèle basse sur la théorie de l'information est avancé. Un tel modèle permet notamment une

étude sur la taille minimale de l’empreinte capable d'identifier une séquence vidéo d'une longueur donnée, sous contraintes d’unicité et robustesse pré-imposées.

Mots clés

Model théorique pour le traçage du contenu vidéo, les limites théoriques d’un système de traçage de contenu vidéo, modèle basse sur la théorie de l'information.

Conclusions

The worldwide mass production context brings technology closer to people. Affordable capturing, processing and storage devices along with wide spread broadband Internet access, empowers people to easily produce, manipulate and distribute large amounts of visual content. Hence, efficient tools for searching, retrieving and tracking distorted video content in very large video databases have to be deployed in order to serve the purposes of applications like copyright protection, parental control. Video fingerprinting is an appealing solution to these issues

Despite the wide range of methods that have been investigated in the state of the art for video fingerprinting methods, limitations have been identified at three levels. Firstly, the uniqueness property of fingerprints is not granted by a mathematical comprehensive approach. Secondly, the robustness property of fingerprints is based on partial mathematical models without a general framework able to address the wide variety of existing distortions. Moreover, the academic state of the art methods have not addressed yet, at our best knowledge, the challenging case of live camcorder recording. Thirdly, in general the state of the art video fingerprinting methods do not have query localization support able to result in scalable solutions for large scale databases.

The video fingerprinting method advanced in the present thesis (TrackART) is characterized from the structural point of view, by two phases: the offline phase and the online phase.

The offline phase enables the localization of a query sequence within a reference video sequence. Its purpose is to process the reference video collection and to map the visual content to a new representation space. This phase comports a usage innovation: it reconsiders, adapts and integrates state of the art image processing algorithms for fingerprinting purposes.

The online phase: a query video sequence is given to the system and its identity is inquired. This phase comports a design innovation: the fingerprint and reduced fingerprint blocks are advanced in the present thesis so as to empower the fingerprint with the mathematical properties of the DWT coefficients and to grant statistical error control in the fingerprint matching.

From the functional point of view, TrackART answers the main challenges for a video fingerprinting system (described in Table IV.1):

- The uniqueness property of fingerprints is ensured by the fact that the video features are selected according to a mathematical model representing the visual content (the wavelet coefficients).
- The robustness property of fingerprints is achieved by the fact that the mathematical models governing the selected features are robust to frame content and aspect distortions as well as video format distortions.
- The scalability to large scale databases is ensured by the fact that a query localization procedure is employed and by the fact that the all the algorithms have fast implementations.

The advanced video fingerprinting method has been tested in two practical use cases proposed by the industrial partners, namely the retrieval of video sequences from database under computer generated distortions and the live camcorder recording.

The *a priori* properties described above are validated by the experimental results: in the video retrieval use case, the probability of false alarm reached is null i whereas the missed detection was lower than 0.025, precision and recall were higher than 0.97; in the live camcorder recording use case, the probability of false alarm was 0.000016, the probability of missed detection was lower than 0.041, precision and recall were equal to 0.93.

To conclude with, the present thesis offers a solution to the nowadays limitations in deploying fingerprinting in two real-life applications.

Perspectives

The perspectives of the present thesis are connected to the fingerprinting theoretical model. Actually, no such model is currently available in the literature and hence the fingerprinting theoretical limits cannot be explored. In order to establish the minimal fingerprint size able to identify a video sequence of a given length under prescribed robustness/uniqueness constraints, the model presented in the Fig. IV.1 is advanced:

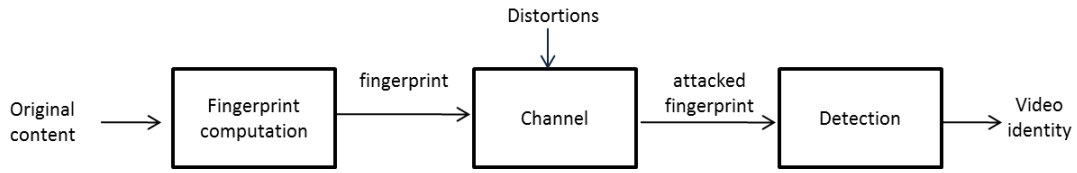


Fig.IV.1 Theoretical model for video fingerprinting

Constraints	Challenge	Current limitation	Thesis contributions
Uniqueness	Accurate representation of the video content	Heuristic procedures	Fingerprint computation independent of random, time-variant conditions: ❑ stationary/ergodic fingerprints ❑ 2D-wavelet coefficients
Robustness	Mathematical ground In-theater live camcorder recording	Heuristic procedures No related method reported in the state-of-the-art	Mathematical decision rule in fingerprint matching: ❑ method based on a repeated statistical test ❑ statistical error control
Search efficiency	Scalability	Very few full scalable mono-modal methods reported in the state-of-the-art	Scalable method ❑ automatic retrieval procedure ❑ O(n) complexity for fingerprint computation ❑ O(nlog(n)) complexity for fingerprint matching with respect to the fingerprint size

Table IV.1 Camcorder recording robust video fingerprinting: constraints, challenges, state of the art limitations and thesis contributions.

Appendix

A.1. Online localization illustrations

In order to have a visual explanation for the Online localization block of the TrackART method (detailed in Section II.3), some situations are illustrated below.

To recap, the function online localization block is to position the query starting point in the reference database. As explained in Section II.3 the query is represented as a set of successive keyframes.

Fig.A.1.a-b presents the case of two video sequences, namely seqQuery7 and seqQuery27 respectively.

Note that both the query and reference keyframes have attached their corresponding position (frame number) from the sequences from where they originate, *i.e.* the parent video sequence. For example, for the keyframe named *seqQuery7_000001_000001.jpg*, the name of the parent video sequence is the query sequence seqQuery7, the position of the keyframe is 1 and the position of the frame is 1.

The localization system starts by processing the first keyframe in the query and returns a list of 10 candidates for the starting point of the query sequence in the reference sequence, as illustrated in Fig.A.1. By running the fingerprinting algorithm (detailed in Section II.3.3), the true starting point is identified, see the green highlighted reference keyframe in Fig.A.1.

However, there is no *a priori* evidence about the fact that one of the 10 reference queries is the true starting point. Such a case is illustrated in Fig.A.2 Consider the case of query seqQuery3. All the 10 candidates returned by the localization system for its first keyframe are refuted by the fingerprinting algorithm. Consequently, the online localization process is resumed on the second keyframe of seqQuery3. This time, the fingerprinting algorithm confirms seqRef2_000002-27.jpg as the true starting point in the reference (see the green highlighted reference keyframe in Fig.A.2.b)

In our experiments, the maximum number of tested query keyframes in order to obtain a positive answer from the fingerprinting block is 5.

Note that the illustrations are made only for the use case of live camcorder recording because the use case of computer generated distortions (*i.e.* Gaussian filtering, sharpening, contrast and brightness changes, conversion to grayscale) are practically included within the distortions induced by live camcorder recording (*i.e.* abrupt geometric transformations with brightness, color and contrast variations).

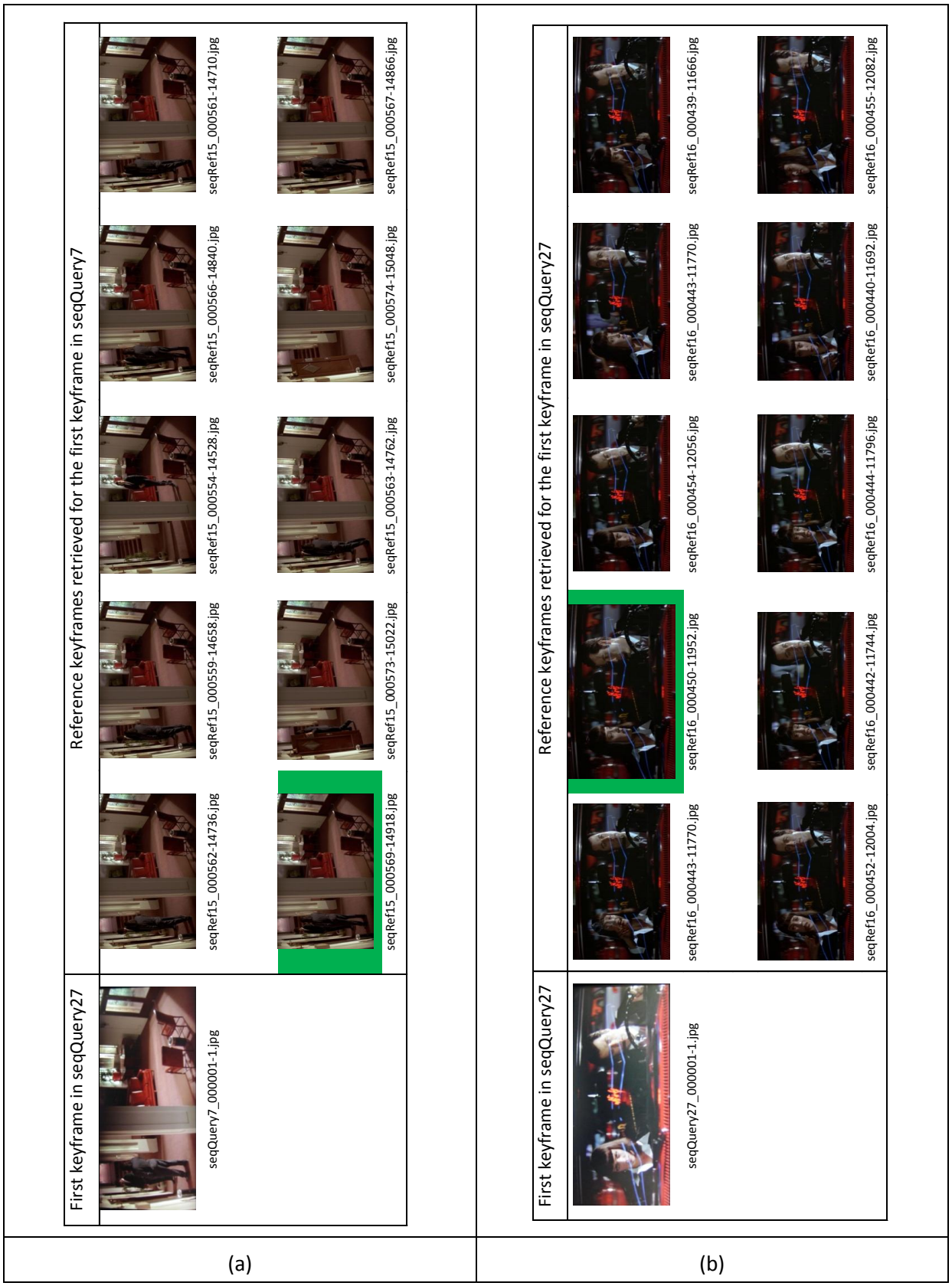


Fig.A.1 Illustration of the first 10 best matching reference keyframes for the first query keyframe of: (a) query sequence seqQuery7, (b) query sequence seqQuery27



(a) seqQuery3, keyframe 1

(b) seqQuery3, keyframe 2

Fig.A.2 Illustration of the first 10 best matching reference keyframes for the first 2 query keyframes of query sequence seqQuery3

A.2. Publications

The incremental results obtained during the thesis were included in one journal papers and four conference proceddings:

Journal

- Garboan, A., Mitrea, M., Prêteux, F., "Cinematography sequences tracking by means of fingeprinting techniques", Annals of Telecommunications, no. 2013; on-line available DOI: 10.1007/s12243-012-0334-7.

Conference papers

- Garboan, A., Mitrea, M., Prêteux, F., "Camcorder recording robust video fingerprinting", Proceedings for the *IEEE 16th Symposium on Consumer Electronics* (ISCE), 2012, Harrisburg-US, pp. 1 – 4.
- Garboan, A., Mitrea, M., Prêteux, F., "Video retrieval by means of robust fingerprinting", Proceedings for the *IEEE 15th Symposium on Consumer Electronics* (ISCE), 2011, Singapore, pp. 299 - 303.
- Garboan, A., Mitrea, M., Prêteux, F., "DWT-based Robust Video Fingerprinting", Proceedings for the *3rd European Workshop on Visual Information Processing* (EUVIP), 2011, Paris, pp. 216 - 221.
- Garboan, A., Mitrea, M., Prêteux, F., "Statistical counter-attacks in MPEG-4 AVC watermarking", Proc. SPIE, Vol. 7723, 2010.

A.3. Selection of publications

A.3.1. Journal paper

Garboan, A., Mitrea, M., Prêteux, F., "Cinematography sequences tracking by means of fingeprinting techniques", Annals of Telecommunications, 2013 on-line available DOI: 10.1007/s12243-012-0334-7.

Cinematography sequences tracking by means of fingerprinting techniques

A. Garboan[1], *M. Mitrea*[1,3], *F. Prêteux*[2, 3]
[1]Institut Télécom - Télécom SudParis, Department ARTEMIS;
[2]MINES ParisTech; [3]UMR CNRS 8145 MAP5
9, rue Charles Fourier, 91011 Evry France
Phone : +33 1 60 76 44 24, Fax : +33 1 60 76 43 81
adriana.garboan@it-sudparis.eu, mihai.mitrea@it-sudparis.eu, francoise.preteux@mines-paristech.fr

ABSTRACT

By advancing a new robust fingerprinting method, the present paper takes the challenge of designing an enabler for the use of Internet as a distribution tool in cinematography. Video fingerprints are short features extracted from a video sequence in order to uniquely identify that visual content and its replicas. This paper develops a new 2D-DWT-based robust video fingerprinting method able to address two use cases related to the cinematography industry, namely the retrieval of video content from a database and the tracking of in-theater camcorder recorded video content. In this respect, a set of largest absolute value wavelet coefficients is considered as the fingerprint and a repeated statistical test is used as the matching procedure. The video dataset consists of two corpora, one for each use case. The first corpus regroups 3 hours of heterogeneous original content (organized under the framework of the HD3D-IIO French national project) and of its attacked versions (a total of 21 hours of video content). The second corpus consists of 3 hours of heterogeneous content (i.e. HD3D-IIO corpus) and of 1 hour of live camcorder recorded video content (a total of 4 hours of video content). The inner 2D-DWT properties with respect to content preserving attacks (such as linear filtering, sharpening, geometric, conversion to grayscale, small rotations, contrast changes, brightness changes, live camcorder recording), ensure the following results: in the first use case the probability of false alarm and missed detection were lower than 0.0005, precision and recall were higher than 0.97; in the second use case, the probability of false alarm was 0.00009, the probability of missed detection was lower than 0.0036, precision and recall were equal to 0.72.

Keywords — robust video fingerprinting, DWT, robustness, uniqueness, live camcorder recording.

1. INTRODUCTION

The worldwide mass production context brings technology closer to people. Affordable capturing, processing and storage devices along with wide spread broadband Internet access, empowers people to easily produce, manipulate and distribute large amounts of visual content.

Such a situation raises complex challenges in various multimedia domains (copyright protection, illegal distribution and management of massive databases, …). Despite the particular applicative target, issues connected to video identification, authentication, indexation, retrieval, searching, navigation, organization and manipulation have to be always addressed.

The bottleneck in developing practical solutions for such problems makes the cinematography industry very suspicious in using Internet as a main movie distribution support.

Currently, a solution intensively considered in research studies is ***video fingerprinting*** also referred to as ***content-based copy*** detection or ***near-duplicate copy detection***.

Throughout the current study, ***a copy***, ***a replica*** or ***an attacked video*** is obtained from some original video excerpt by means of any transformation/distortion, such as addition, deletion, modifications (of aspect, color, contrast, encoding, …), or camcording [1], see Table 1.

Distortions	Examples
Video format	▪ encoding ▪ transcoding ▪ bitrate changes ▪ D/A and A/D conversions
	▪ frame dropping ▪ frame addition ▪ framerate changes ▪ frames substitution
Frame aspect	➢ geometric modifications: scaling, rotations, shifting
	➢ color modifications: conversion to grayscale, sepia, color filtering
	➢ illumination changes: brightness, contrast, saturation, gamma correction modifications, histogram equalization
	➢ compression
	➢ filtering: linear (Gaussian, sharpening), non-linear (median filter)
	➢ noise addition
	➢ aspect ratio changes
Video content	▪ cropping ▪ text insertion, caption insertion ▪ letter-box insertion ▪ affine transformations
Mixed	▪ combinations of all the above modifications

Table 1: Types of computer or camcording generated video modifications

Due to the practical applicative field of video fingerprinting, although the modifications considered in this classification alter the quality of the video content, they do not destroy its commercial or entertainment value. These modifications can be classified in three major categories depending on the domain they affect, namely the video format, the frame aspect and the video content.
Video fingerprints can be best defined in relation with human fingerprints, [2], as illustrated in Figure 1.

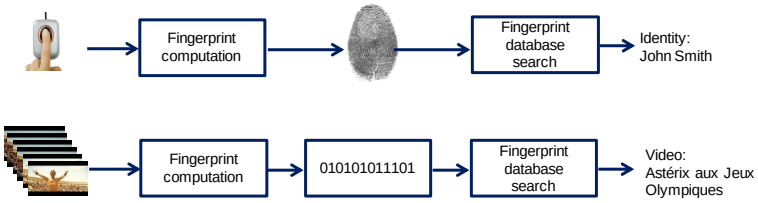


Figure 1: Human *versus* video fingerprinting

While the human fingerprint can be seen as a human summary (a signature) that is unique for every person, the video fingerprint can be seen as some short video feature (*e.g.* a string of bits, color histograms, …) which can uniquely identify that video.In practice, video fingerprints are used just as human fingerprints: they are first computed and then searched for in a database, according to a given similarity measure.

Assume the case in which ***S*** video sequences have their fingerprints computed and are sequentially searched for in the database. A correct answer in such a matching procedure is obtained when the same visual content is detected not only in its original video sequence but also in all its replica videos; be there ***tp*** the number of such correct answers. Practical fingerprinting methods may also come across with two types of matching errors. First, some video content existing in the database might not be retrieved; be ***fn*** the number of such erred decisions. Secondly, the detection procedure can also yield a false positive *i.e.* take some visual content for another one. Be ***fp*** the number of such situations. Note that ***S = tp + fp +fn.***

Video fingerprints have two main properties:

- ***Uniqueness*****:** fingerprints extracted from different video content should be different. This property is assessed by two objective evaluation criteria: the ***probability of false alarm*** (P_{fa}) and the ***precision rate*** (*Prec*), defined by the following formulas:

$$P_{fa} = \frac{fp}{tp + fn + fp} \qquad (1) \qquad\qquad Prec = \frac{tp}{tp + fp} \qquad (2)$$

- ***Robustness to distortions*****:** fingerprints extracted from an original video sequence and its replicas should be similar in the sense of the considered similarity metric. The robustness property is also quantified by two objective evaluation criteria, namely the ***probability of missed detection*** (P_{md}) and the ***recall rate*** (*Rec*), as defined below:

$$P_{md} = \frac{fn}{tp + fn + fp} \quad (3) \qquad Rec = \frac{tp}{tp + fn} \quad (4)$$

On the one hand, an efficient fingerprinting method should ensure a low probability of false alarm (*i.e.* low probability of retrieving video sequences which are neither the query nor its replicas) and low probability of missed detection (*i.e.* a low probability of not retrieving replica video sequences of the query). On the other hand, high values for precision (*i.e.* a high probability of retrieving replica video sequences for a given query out of all the retrieved video sequences) and recall (*i.e.* a high probability in retrieving all the replica video sequences existing in a database for a given query) should also be obtained.
Additional functional properties (database search efficiency, automatic processing, localization of a query in the reference video) can be set, according to the targeted practical application.

The present study is focused on two applicative use cases of relevance for the cinematography industry: database video retrieval and live camcorder recording tracking. The former takes a video sequence (arbitrarily chosen) as a query and searches for its potential replicas in the database. The latter covers the case in which an arbitrarily chosen sequence from the reference database is live camcorder recorded and its original version is searched in the reference database. French cinematography authorities (CST - Commission Supérieure
Technique de l'Image et du Son [3]) have set for the fingerprinting methods serving these two cases the following constraints: probabilities of false alarm/missed detection lower than 5% and precision/recall rates higher than 95%.
In order to reach these performances, the present research study advances a 2D-DWT (Discrete Wavelet Transform)-based video fingerprinting method. The fingerprint itself consists of highest absolute value 2D-DWT coefficients, computed on video key-frames. As already known in the literature [4], such coefficients feature very fine statistical behaviors; hence, repeated statistical tests can be considered in the fingerprint matching, thus granting mathematical relevance to the experimental results.
The present paper is structured as follows. Section 2 presents the state of the art for video fingerprinting. Section 3 advances an original method for video fingerprinting. Section 4 experimentally validates the proposed method, according to the two above mentioned use cases. Conclusions are drawn and perspectives are opened in Section 5.

2. STATE OF THE ART

Despite its young age, video fingerprinting can serve a large variety of applications: detection of copyright infringement, detection of known illegal content, control and management of video content, broadcast and advertisement monitoring, audience measurement, business intelligence, … . Consequently, the research studies cover a large area of methodological tools from pixel difference of consecutive frames or RGB histograms to transform domain based fingerprinting approaches.

In the sequel, the fingerprinting state of the art will be structured according to the type of feature representing the fingerprint (Table 2) and the similarity metric achieving the fingerprints matching (Table 3).

The video features used as fingerprints can be computed only from the visual content (*i.e*. the case of ***mono-modal methods***) or from visual and audio content (*i.e*. the case of ***multi-modal methods***). Independently with respect to its type, the video fingerprint can be computed at different ***granularity*** levels, *e.g*. frames, keyframes, blocks or regions of frames, group of frames, points of interest.

According to the domain in which the fingerprints are computed, the group of mono-modal methods can be of four types: spatial, temporal, transform and color.
The **spatial fingerprints** computed on blocks, regions of frames or whole frames are robust to non-geometric distortions, but they lack in robustness against geometric modifications (*e.g*. cropping, rotations). The interest points based features have a high robustness against the geometric distortions and transcoding transformations but lack in resilience against changes in color, illumination and filtering. Moreover, this type of features poses problems of uniqueness in the case of very similar video sequences, (*e.g*. TV news) therefore needs to be used in combination with other features.
The category of **temporal fingerprints** is generally robust to global changes in the quality of the video like non-geometric modifications of the frame aspect and they can resist several encoding (*e.g*. MPEG compression), but they are generally sensitive to distortions affecting the video format (*e.g*. frame-rate changes frame-dropping, transcoding) and to geometric modifications).
Transform based fingerprints ensure robustness to geometric and non-geometric frame aspect modifications and to video format modification but are sensitive to modifications of video content such as cropping and content addition.
The **color based** category of fingerprints lacks resilience to global variations in color and illumination but can be used along with other features in order to enhance discriminability.
As explained above, the mono-modal methods employ a reduced number of visual features as fingerprints in order to identify the limitations that they pose and their possible applications. The multi-modal types of fingerprints combine the advantages of video and audio features of videos can achieve better results with faster computation time than the mono-modal methods.
The frequent disadvantage of the multi-modal types of fingerprints is their excessive number of computed features, which leads to redundant video information used as fingerprint (*e.g*. [5] using SIFT and SURF features simultaneously). As the computational resources increase steadily due to technological development, extra computation is not considered a prohibitive factor. However, a clear mathematical ground for video fingerprinting should not be ignored.
According to the similarity distance employed for matching, the fingerprinting methods can be divided in two categories, distance based and probability based, as illustrated in Table 2.

Types of fingerprints		Granularity	Fingerprint examples
Mono-modal methods (Video content features)	Spatial	Blocks, regions of frames, frames, keyframes	▪ visual attention regions, [6] ▪ ordinal ranking of average gray level of frame blocks [7], [8] ▪ quantized block motion vectors of frames [7] ▪ invariant moments of frames edge representation, [9] ▪ centroid of gradient orientations, [10] ▪ dominant edge orientation, [11]
		Points of interest	▪ signal description of motion of interest points (corner features, Harris points), across videos [12], [13], [14] ▪ scale-space features (*e.g.* SIFT), [15] ▪ descriptors of interest points [16]
	Temporal	Group of frames	▪ differential block luminance features between consecutive frames, [2] ▪ temporal ordinal measure (ordering of intensity blocks in successive frames depending on their average intensity), [17], [18], [19], [11] ▪ ordinal histogram over the frames of the entire video [15], [20] ▪ pixel differences between consecutive frames [11] ▪ shot duration sequence, [21]]
		Down-sampled frames	
		Keyframes	
		Every frame	
	Transform-D (2D, 3D)	GOP	▪ quantized compact Fourier-Mellin transform coefficients of keyframes, [15] ▪ subspace embedding using the singular value decomposition [22] ▪ 3D DCT coefficients of sub-sampled keyframes, [23] ▪ DCT coefficients of the radial projection vector of the keyframes pixels, [24] ▪ 2D wavelet transform [25], [26], [27]
		Re-sampled video	
		Frame transform	
	Color	Histogram based	▪ YUV histograms of the DC sequence of MPEG videos [28], [7] ▪ YCbCr histogram of a group of frames, [15] ▪ color moment representation [29] ▪ RGB, HSV histogram of frames [11] ▪ the principal component of the color histograms of keyframes [30]
Multi-modal methods (Video + Audio features)	Combined	Combined approaches	▪ SIFT, GIST and color correlogram features for keyframes, [31] ▪ global visual feature (DCT), local visual feature (SIFT, SURF), audio feature (WASF, modified MPEG-7 descriptor ASF), [5] ▪ visual feature: center-symmetric local binary pattern (CS-LBP), hamming embedding; audio feature: filter banks, [32] ▪ coarsely quantized area matching – visual feature, divide and locate – audio feature [33],[34] ▪ cascade of multimodal features (DC SIFT BoW, DCT, WASF) and temporal pyramid matching [35]

Table 2: Types of video fingerprints

Types of similarity measures	Similarity measure	Applicability
Distance based	L1 distance (Manhattan)	▪ non-binary fingerprints, [11]
	L2 (Euclidian) distances	▪ non-binary fingerprints [10]
	Hamming distance	▪ binary fingerprints [23], [6], [2]
	Haussdorff distance	▪ edge points based fingerpints [11]
	Normalized histogram intersection	▪ histogram based fingerprints [37]
	Normalized correlation coefficient	▪ histogram of block motion vectors [7]
	k-nn, voting function	▪ interest point-based fingerprints [12, [13], [14]
Probability based	Based on statistical tests	▪ hypothesis testing, multivariate Wald-Wolforwitz [27] ▪ Rho test on correlation [25]

Table 3: Types of similarity measures

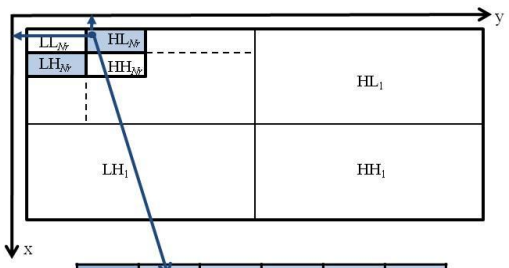
The distance-based group of methods has the advantage of allowing a decision based on an experimentally determined threshold. While they are easier to use, they don't permit in the majority of cases a decision based on a mathematical ground. Therefore the alternative is the probability-based similarity measures which can grant a statistical rule for decision.
As it can be seen, although in the last decade the applicability field of video fingerprinting grew steadily and despite the wide range of methods that have been investigated, at least two challenges are still to be taken.
First, the state of the art methods presented in Tables 2 and 3 are generally tested on TV content data sets and don't take into account the particularities of the cinema content. These particularities are twofold and refer to the types of visual content and to the types of distortions that need to be addressed by the fingerprinting method. For the former particularity, the cinema visual content has HD quality and presents a high dynamics of the visual content, outdoor/indoor scenes and arbitrarily changing lighting conditions. For the latter particularity, the category of distortions introduced by live camcording is one of the most complexes because it includes and combines abrupt geometric transformations with brightness, color and contrast variations.
Secondly, the trade-off among the probability of false alarm, the probability of missed detection, the precision, the recall and the computational time required by such a use case has not yet been investigated. Therefore the objective of this paper is to advance a DWT-based video fingerprinting method using a mathematical decision rule for the detection of replicas, capable of addressing not only the use case of video retrieval but also the complex use case of camcording in movie theatres.

3. DWT-BASED VIDEO FINGERPRINTING

Due to its possibility of representing, in a very compact way, salient characteristics of the video content and also to its low complexity, the 2D-DWT is already intensively considered in practically all image processing applications, from compression and watermarking to defect detection in garments. Our study investigates whether the 2D-DWT can be employed in order to uniquely and robustly identify the visual content.
In this respect, a new video fingerprinting method is advanced. In order to extract a fingerprint from a video sequence (arbitrarily chosen), that sequence is first pre-processed, then a 2D-Wavelet transform is applied to its frames and finally a certain selection of the coefficients is carried on in order to obtain the fingerprint *per-se,* see Section 3.1. The fingerprint matching is achieved by a repeated test on correlation, see Section 3.2.

3.1. Fingerprinting computation principle

Be there a video sequence, represented in a given format (compressed or not).
The pre-processing step aims at increasing the invariance of the envisioned fingerprint to different video processing operations, be they malicious (attacks) or mundane (ordinary video manipulations).
First, the video sequence is decoded into frames in order to diminish the influence of a particular video format or codec.
Second, a temporal sub-sampling to 1 fps is performed in order to eliminate the redundancy between adjacent frames and to speed-up the fingerprint computation.
Third, a spatial re-sampling to $W \times H$ pixels (in our experiments, $W = 352$ and $H = 288$) is performed on the sampled frames, thus increasing the robustness of the method to frame size changes attacks.
Fourthly, in order to extract the salient information contained within the frame while reducing the computational costs, the down-sampled frames are represented in the HSV (Hue - Saturation - Value) color space with the V component normalized to the [0, 1] interval.
In the 2D-DWT Transform step, a (9, 7) Daubechies wavelet transform at the resolution level of $Nr = 3$ is applied on the V component of every sampled frame.
The coefficients selection (*i.e.* the fingerprint) aims at conveying information about the spatial distribution of salient features within the frames.
In the coefficients selection step, the 2D-DWT coefficients are selected depending on the role of the video sequence (reference or query).
For the reference video sequences in the database, the $R = 360$ highest absolute value coefficients from the HL_{Nr} and LH_{Nr} frequency sub-bands of the transform V component, together with their locations are selected and stored in the coefficients matrix (as illustrated in Figure 2.a).



DWT coefs	0.82	0.79	0.65	...	0.02
x	15	39	43	....	3
y	62	22	29	....	47

(a)

Ranks / Frames	1	2	3	...	359	360
1	0.822	0.540	0.272	....	0.056	0.056
2	0.465	0.458	0.405		0.031	0.030
...	....	....	....	....	....	....
N	0.794	0.687	0.545	....	0.165	0.165
N-1	0.530	0.491	0.436	....	0.072	0.072

(b)

Figure 2: (a) Coefficients matrix for a frame, (b) Rank matrix of DWT coefficients

The coefficient matrix in Figure 2.a illustrates the fingerprint of a sampled frame, while the fingerprint of an entire reference video sequence is presented in Figure 2.b and it is called the rank matrix.

The rank matrix is filled-in with all the fingerprints computed on then *N* sampled frames. The fingerprints of the frames consist of *R* = 360, 2D-DWT coefficients sorted in a decreasing order of their absolute values, it can be considered that the coefficients are disposed on 360 ranks (where "1" corresponds to the highest absolute value coefficient). This approach will turn to be particularly useful for fingerprint matching.

In the computation of the fingerprint for a query video sequence, the 2D-DWT coefficients are selected from the HL_{Nr} and LH_{Nr} frequency sub-bands of the V transform component from the locations indicated as salient by the reference coefficients matrices. After selecting the salient coefficients from every sampled frame of the reference video, the rank matrix will be obtained.

3.2 Fingerprint matching

The proposed similarity measure between fingerprints is the normalized correlation as given by the formula in (5).

$$\rho = corr_k(f,t) = \frac{1}{N-1}\sum_{x,y}\frac{(f_k(x,y)-\bar{f}_k)(t_k(x,y)-\bar{t}_k)}{\sigma_{f_k}\sigma_{t_k}} \qquad (5)$$

In (5), f_k and t_k designate the 2D-DWT coefficients of the query and the reference videos on a rank *k*, $\bar{f}_k, \bar{t}_k$ are the mean values of the 2D-DWT coefficients on the considered rank, while $\sigma_{f_k}, \sigma_{t_k}$ are the related standard deviations, respectively. *N* designates the number of 2D-DWT coefficients in every rank *k*, *i.e.* the number of sampled frames in each video sequence.

A perfect match (identity) between the query and the reference rank is obtained when $|\rho| = 1$; a value $\rho = 0$ indicates no correlation between f_k and t_k.

The normalized correlation is computed between the 2D-DWT coefficients disposed on ranks, *i.e*. the columns of the rank matrix. Such a strategy is justified by the statistical

investigation of the 2D-DWT coefficient behavior in [37] and [38]: it was proved that the values taken by a rank in the 2D-DWT coefficient hierarchy feature "stationarity" and the corresponding probability density function was estimated using a mixture of Gaussian laws. Hence, the "stationarity" property of these coefficients ensures a certain degree of independence of the results with respect to the experimental corpus.
In practice, in order to be able to also retrieve content preserving replicas, the absolute value of the normalized correlation should be compared to some threshold *T*; should $\rho \geq T$, then the query and the reference ranks are considered as identical.
The value of the *T* threshold is statistically determined according to the Rho test on correlation [39]. This test is individually applied to each of the $R = 360$ ranks under investigation; the null/alternative hypotheses are:

$$\begin{cases} H_0 : \textit{the ranks are not correlated} \\ H_1 : \textit{the ranks are correlated} \end{cases}$$

A match between the query and the reference video sequences is obtained when the majority of ranks (*i.e.* more than $R/2 = 180$) are correlated. Should the majority of ranks be uncorrelated the query and the reference video sequences are considered as distinct.
Assuming the *k* ranked 2D-DWT coefficients from the query and from the reference video sequence are *i.i.d.* (identically and independently distributed) and that they follow a Gaussian distribution, and assuming the H_0 is true, the t_{test} value of the test statistics, see (6), follows a Student probability density function of $N-2$ degrees of freedom:

$$t_{test} = \rho \cdot \sqrt{\frac{(N-2)}{1-\rho^2}}, \qquad (6)$$

where *N* and ρ are the same as above.
As illustrated in Figure 3, if $t_{test} \leq z_{\alpha/2}$ (where $z_{\alpha/2}$ is the α-point value of the above-mentioned Student law), then the H_0 hypothesis is accepted, *i.e.* the 2D-DWT coefficients on the *k* rank are not correlated. If $t_{test} > z_{\alpha/2}$ the H_1, hypothesis is accepted, *i.e.* the 2D-DWT coefficients on the *k* rank are correlated.
In our experiments, we considered a significance level of $\alpha = 0.05$.
Note that in our application, the Rho test is run properly. First, the "stationarity" behavior of the 2D-DWT coefficients [37], [38] and the original video pre-processing ensures the *i.i.d.* behavior for the tested coefficients. Secondly, the robustness of the Rho test for non-Gaussian data may be invoked [39] in this case.

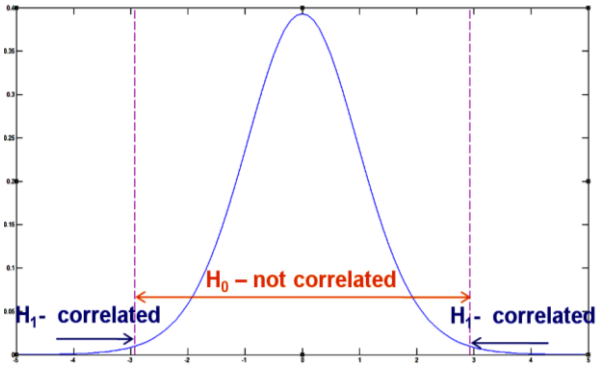


Figure 3: Student probability density function of *N-2* degrees of freedom

4. EXPERIMENTAL RESULTS

The proposed 2D-DWT-based fingerprinting technique was tested in two applicative use cases: (1) – video identification and retrieval in a database and (2) – live camcorder recording.

In the experiments, the probability of false alarm (P_{fa}), the probability of missed detection (P_{md}), the precision (*Prec*) and the recall (*Rec*) rates were computed for every query and average values were obtained by successively considering all the sequences in the database and by averaging the corresponding results.

4.1 Video retrieval use-case

A video identification and retrieval use case consists of identifying a video sequence (further referred as query sequence) in a database of video sequences, called the reference video database, as illustrated in Figure 4. The computation of the fingerprint for the query video is done online, whereas the computation for the reference videos is done offline. The identification process is based on the video fingerprints and on the matching between them.

When consulting the reference database with a query video sequence, all its replicas should be retrieved. Irrelevant video sequences (not connected to the query) should be ignored.

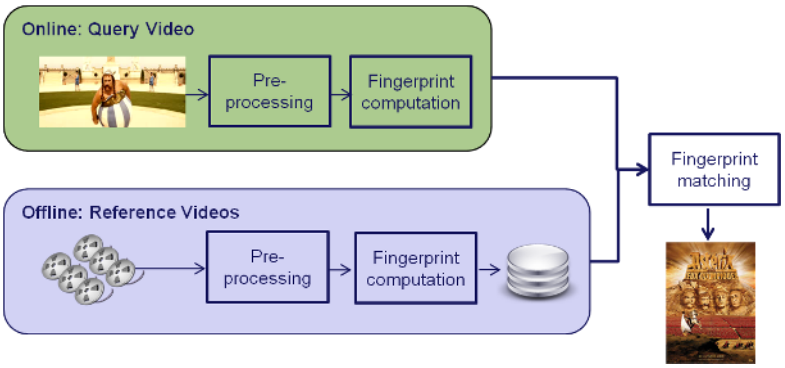


Figure 4: Video identification and retrieval use case

The reference video database contains original video sequences from the HD3D-IIO [40] corpus and computer generated video replicas totalizing 21 hours of video content. It is structured in 1260 video sequences of 1 minute each, 180 original sequences and 1080 replicas.
The HD3D-IIO video corpus consists of 180 original video sequences of 1 minute each and totalizes 3 hours of visual content. The HD3D-IIO video content belongs to 7 different movies and combines indoor and outdoor scenes, unstable and arbitrary lighting conditions, still and high motion scenes as illustrated in Figure 5.

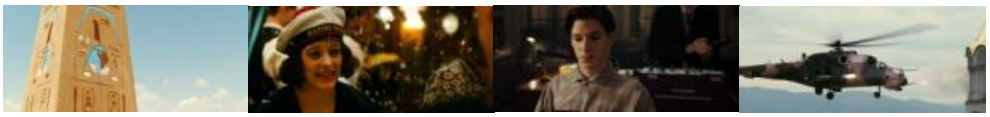


Figure 5: Frames from the HD3D-IIO video corpus

The query video corpus consisted of 140 video sequences chosen from the reference database (20 original video sequences and 20 replica video sequences for each of the attacks).
The replica video sequences were obtained by considering the following modifications: contrast decrease (25%), linear filtering (Gaussian filter), conversion to grayscale, sharpening, brightness increase (20%) and brightness decrease (25%) as illustrated in Figure 6.



Original version



Contrast decrease



Gaussian filtering



Grayscale conversion



Sharpening



Brightness increase



Brightness decrease

Figure 6: The replica video sequences

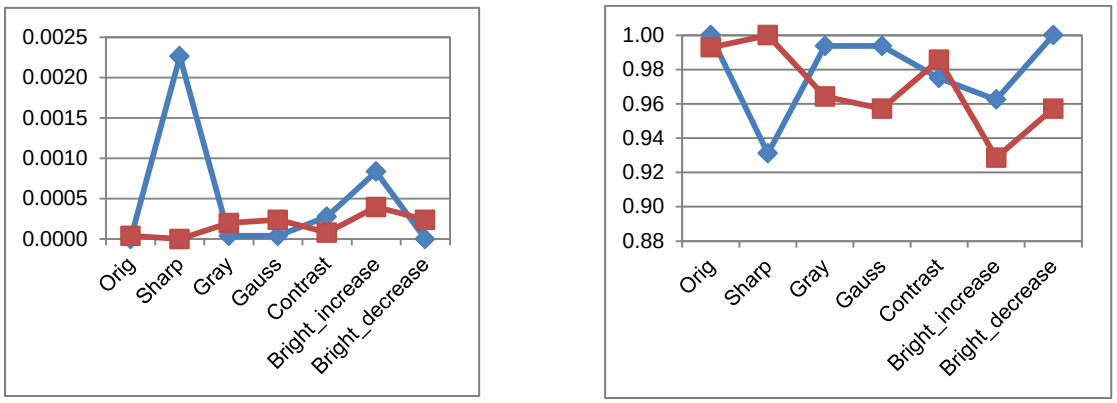
Having this experimental setup, the average results (obtained by averaging the results for all the considered queries) are illustrated in Table 4:

Probability of false alarm	0.0005	Precision	0.98
Probability of missed detection	0.0002	Recall	0.97

Table 4: Average results for the video database use-case

The overall results, point to a very good retrieval accuracy, with average false alarm probability lower than 0.0005 and average missed detection probability lower than 0.0002. The results are reinforced by the average precision higher than 0.98 and average recall higher than 0.97.
Figure 7.a illustrates the average values (obtained by averaging the results for all the considered queries) of the probability of false alarm (the diamonds in blue) and of the probability of missed detection (the squares in red) depending on the particular attack.

Similarly, Figure 7 (b) illustrates the average values of precision (the diamonds in blue) and recall (the squares in red) as functions of attacks



(a) (b)

Figure 7: Probability of false alarm and missed detection (a), precision and recall (b) depending on the attacks

The quantitative results in Figure 7 show that the most disturbing effects are induced by the sharpening and by the brightness increase attacks. This can be explained by the fact that these two types of attacks follow the stationarity investigation with less accuracy than the other four types [4]. Consequently a matching rule devoted to non-stationary information sources is expected to improve the results.

4.2 Camcording use-case

The camcording use case consists of tracking an in-theatre camcorder recorded video sequence in a database of original video sequences, as illustrated in Figure 8. In such a case, the attacked video sequences are obtained by capturing with a non-professional camcorder the original sequence which is displayed on a screen.

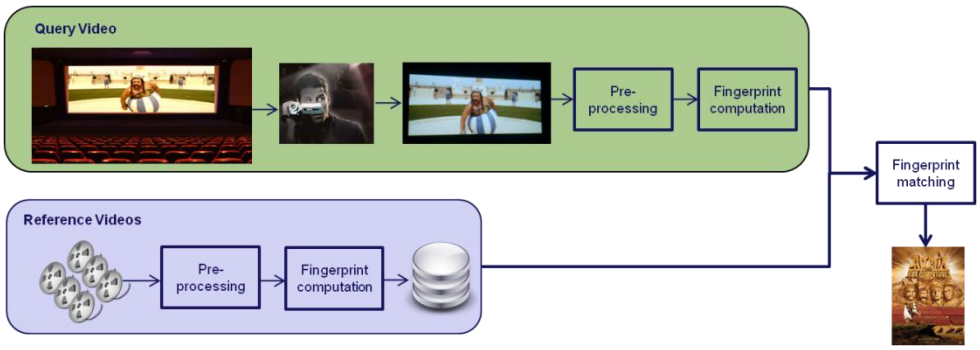
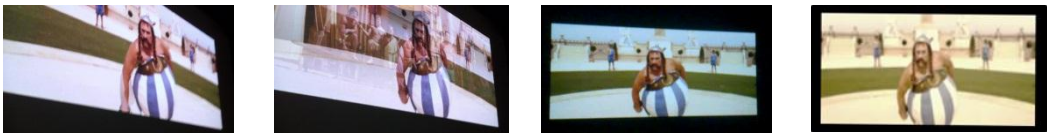


Figure 8: In-theater camcording use case

The reference video database for this use case contains original video sequences from the HD3D-IIO corpus.
From the HD3D-IIO corpus, a random selection of 60 video sequences (*i.e*. 1 hour) was performed and afterwards camcorder recorded, yielding a query corpus of 60 video replicas. A few frames from the camcorder recorded replicas are exhibited in Figure 9.
The camcorder recording was performed in two experimental set-ups, each of them contributing with 30 minutes of recorded video.
The first set-up consisted of video projection in the cinema theatre located at the Commision Supérieure Technique de l'Image et du Son (CST, [3]); the capturing devices were the video cameras of two cell phones, namely an iPhone4 and a Nokia 5800.



Original version



CST captures ARTEMIS capture

Figure 9: Frames from camcorded video sequences

The second set-up consisted of video playing on a PC monitor (DELL 1680 x 1050 pixel resolution, 22" LCD display screen) at the ARTEMIS department [41]; the capturing devices were three cameras: a Canon Legria HF20, a Sanyo Xacti HD1010 and a Canon EOS 7D with a Tokina AT-X PRO objective.
A simplified geometrical representation of the recording process performed in the CST cinema theatre is given in Figure 10.a, the theatre being viewed from the top side view [42]. The optical axes of the camcorder and of the projector do not usually intersect with the screen at the same point and are not parallel with each other. The angle Ω measures the rotation of the camcorder around its optical axe.

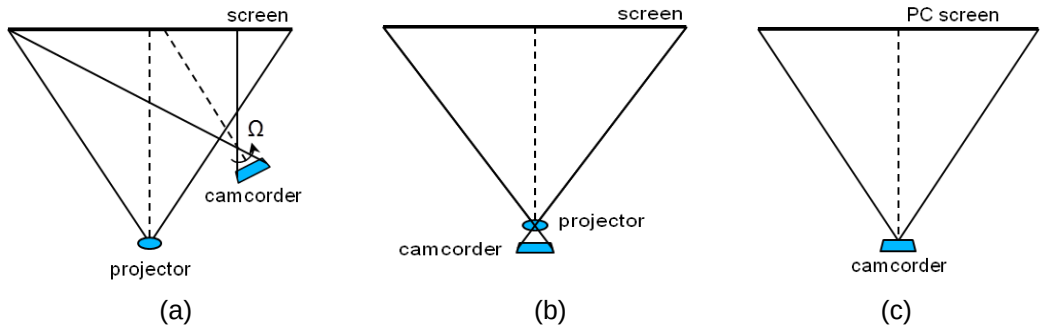
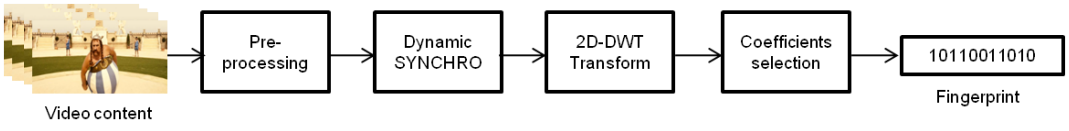


Figure 10:Projection and capture-set; top view

Our experimental set-up for the CST captures is illustrated in Figure 10.b: the camcorder was positioned parallel with the axe of the projector; in the ideal case, Ω = 0. However, by its very nature, live camcording introduces random, time variant capturing angles induced by the pirate's involuntarily movements; in our experiments $-2° \le \Omega \le 2°$.
The experimental set-up for the ARTEMIS video captures is depicted in Figure 10.c: the PC screen has two functions, *i.e.* screen and projector, while the camcorder was positioned with its optical axe perpendicular on the PC screen, but the same random capturing angles, $-2° \le \Omega \le 2°$ were encountered.
In the proposed experimental set-up, the angle Ω was not considered larger than $\pm 2°$ and the position of the camcorder was approximately maintained in a central position of the screen in order to capture the entire video content displayed on the screen.

Due to the severe modifications induced in the camcorded video, a more elaborated approach is needed. Firstly, in order to address the modifications induces in the video by the re-encoding performed by the camcorder (frame-rate changes, bitrate changes, A/D, D/A conversions) the pre-processing step (described in Section 3.1) needs two extra operations: a transcoding to the original parameters (bit rate, frame rate, GOP) by using the ffmpeg libraries [44] and the black letterboxing removal.
Secondly, in order to eliminate as much as possible the inner time-variant de-synchronization induced between the query and the reference frames by the very camcording mechanisms a dynamic synchronization block was designed, as illustrated in Figure 11.



Figure°11: Fingerprint computation principle

To this aim, the fingerprinting matching is no longer performed between frames sampled according to a fixed sampling period but an adaptive mechanism is considered *cf.* Figure 12.
In the experiments of this study, a coarse synchronization of the starting time of the reference and query is already available (*e.g.* obtained through a shot detection procedure).

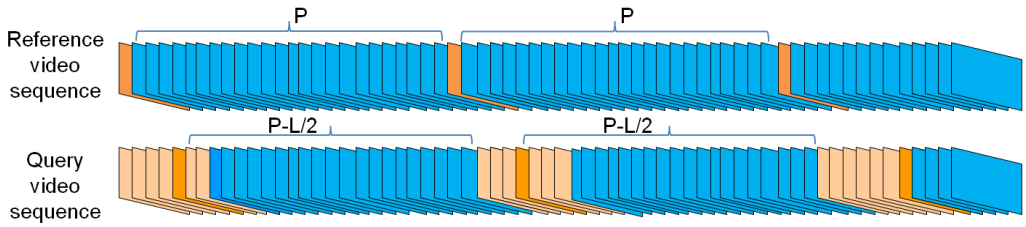


Figure 12: Frame sampling strategy

Be there f_{1r} the first frame sampled from the reference sequence by using a fixed-period sampling of *P* frames. The same *P* period is considered to sample the query and a neighborhood window of *L* frames is selected accordingly. In this window, the frame which is the closest to the f_{1r} frame (in the sense of some similarity measure, *e.g.* relative mean square error, eq. 7) is selected; be this frame f_{1q}. This procedure is recursively applied, by considering f_{1r} and f_{1q} as the starting frames for the rest of the reference and query sequence.

This way, dynamic desynchronization lower than *L/2* frames can be compensated inside each *P* frame interval.

$$MSE(f,t) = \frac{\sum_x (f(x) - t(x))^2}{\sum_x (f(x))^2} \qquad (7)$$

In the formulas above, f and t designate the 2D-DWT coefficients of the query and the reference images respectively.

In our implementation the size of the window was considered *L* = 7 frames and the frame sampling period was *P* = 25.

The results are presented in Table 5 and point to very good false alarm and missed detection probabilities, significantly lower than the limits imposed by the CST. Precision and recall rates indicate reasonable retrieval accuracy but still need to be improved in order to cope with the CST constraints.

The results can be further improved by an adapted shot detector and by finding an efficient matching procedure between the reference and query sampled frames in the re-synchronization step.

Probability of false alarm	0.00009	Precision	0.72
Probability of missed detection	0.0036	Recall	0.72

Table 5: Average results for the camcording use-case

4.3 Computational time

The main steps in our method are the computation of the DWT (hence, an $O(W \times H)$ complexity), the sorting of the corresponding coefficients (hence, an $O(R \log R)$ complexity and the matching of the fingerprints by a normalized correlation coefficient (hence, an $O(R \log R)$ complexity). Figure 13 illustrates the computation times for the proposed

fingerprinting method. The task durations were computed using a system with the following configuration: Intel Xeon CPU processor at 2.8 GHz with 3.5 Go of RAM memory, with an operating system working on 32 bits.

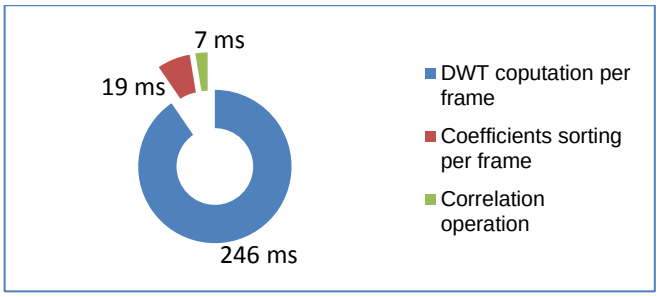


Figure 13: Computation time for the proposed video fingerprinting method

4.4 Parameter choice

Throughout the presentation of the fingerprinting method, several choices for parameters have been made. This section discusses their practical impact.

The fingerprint was computed in the DWT domain due to its capacity of identifying the overall salient content of images and representing it through edges in the high frequency sub bands. Moreover the Daubechies (9, 7) wavelets were used due to their very fine capacity of approximating the visual content.

However, other types of DWT like (2,2) or (4,4) can be used with a very low impact on robustness, while keeping the same uniqueness and reducing the computational time.

For the fingerprint itself, the R = 360 2D-DWT coefficients from every frame were chosen due to their good [4] statistical properties, *i.e.* due to their stationarity. However, should the user be interested in a shorter fingerprint, the R value can be reduced. For instance, when considering R = 250, the P_{fa}, P_{md}, *Prec* and *Rec* values are affected by less than 0.03 (absolute value).

The fingerprint computation on the V component ensures total robustness against color editing.

The pre-processing step (temporal down sampling at 1 frame per second and spatial down sampling at 352 x 288) is meant to ensure robustness against re-encoding. While these values are chosen according to the state of the art hints ([44], [45], [23]) they can be modified according to the practical application. For instance assuming a high motion FX sequence, the temporal down sampling can be neglected.

When considering the live camcording case, the size of the dynamic synchronization window should be set so as to reach the trade-off between time jitter compensation and computational time. While the results reported in the paper corresponds to L = 7, a value L = 11 can slightly improve the results (by 2%).

5. CONCLUSION AND PERSPECTIVES

This paper advances a simple yet very efficient video fingerprinting method. The fingerprint is represented by the hierarchy of largest absolute value 2D-DWT coefficients selected from

two low-frequency sub-bands. The fingerprint matching is carried on by a normalized correlation coefficient and the decision is based on a repeated Rho test on correlation, applied at an $\alpha = 0.05$ significance level. Note that the stationarity of the information sources modeling the hierarchy of the 2D-DWT coefficients allows us to run such a test with statistical rigor. This very fine mathematical model also allows us to consider results obtained on particular databases as being representative for larger corpora. A future direction of our research will be to exploit the probability density function of the 2D-DWT coefficients during the fingerprinting matching procedure.

The advanced method was tested in two applicative use cases related to the cinematography industry (the experiments being jointly conducted with CST experts): video retrieval in databases and live camcorded sequence tracking. In the former case, very good results in terms of probability of false alarm and missed detection lower than 0.0005, precision and recall higher than 0.97 were obtained. Note that the procedure involves only low computational complexity algorithms (the 2D-DWT computation with linear complexity $O(n)$ and R = 360 correlation computations, with a complexity of $O(n \log n)$). While imposing the same computational constraints on the algorithm, the latter case resulted in practical acceptable performances: probability of false alarm equal to 0.00009, probability of missed detection equal to 0.0036, precision and recall equal to 0.72. To our best knowledge, this is the first time when the tracking of the live camcorder recording video sequences can be achieved in practice by completely automatic procedures.

Further work will be devoted to developing an adapted shot detector and to optimizing the matching procedure between the reference and query sampled frames in the re-synchronization step. In this respect, SIFT techniques can be completed with some MPEG-7 descriptors. Obtaining significant gain in matching speed is also part of our future work. This can be obtained by performing an offline clustering of the database, *e.g.* by PCA means.

6. REFERENCES

[1] http://trecvid.nist.gov/

[2] Oostveen, J., Kalker, T., Haitsma, J., ”Feature Extraction and a Database Strategy for Video Fingerprinting”, Lecture Notes In Computer Science, Vol. 2314 archive, Proceedings of the *5th International Conference on Recent Advances in Visual Information Systems*, 2002, pp. 117 - 128.

[3] http://www.cst.fr/

[4] Mitrea ,M., Dumitru, O., Prêteux, F., Vlad A., “Zero-memory information sources approximating to video watermarking attacks”, Proceedings of the *International Conference on Computational Science and Its Applications*, Kuala Lumpur, Malaysia - Lecture Notes in Computer Science 4707, Vol. 3, 2007, pp. 445 - 459.

[5] Gao, W., Huang, T., Tian, Y., Wang Y., Li, Y., Mou, L., Su, C., Jiang, M., Fang, X., Qian, M., “PKU-IDM@TRECVID-CCD 2010: Copy Detection with Visual-Audio Feature Fusion and Sequential Pyramid Matching”, Proceedings of *TRECVID 2010*.

[6] Su, X., Huang, T., Gao, W., “Robust video fingerprinting based on visual attention regions”, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2009.

[7] Hampapur, A., Hyun, K-H., Bolle, R., “Comparison of Sequence Matching Techniques for video copy detection”, in Proceedings of *Storage and Retrieval for Media Databases* (San Jose, USA, Jan. 20-25, 2002), pp: 194-201.

[8] Kim, J., Nam J., “Content-based video copy detection using spatio-temporal compact feature”, Proceedings of the *11th international conference on Advanced Communication Technology* (ICACT), Vol. 3, 2009.
[9] Hu, M.K, “Visual pattern recognition by moment invariants”, *Transactions on Information Theory*, Vol. IT-8, pp: 179–187, 1962.
[10] Lee, S., Yoo C.D., “Robust video fingerprinting for content-based video identification”, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 7, 2008.
[11] Hampapur, A., Bolle, R.M, “Comparison of distance measures for video copy detection”, IBM TJ Watson Research Center, *IEEE International Conference on Multimedia and Expo*, 2001, pp. 737 - 740.
[12] Law-To J., Buisson O., Gouet-Brunet, Boujemaa N., “Robust voting algorithm based on labels of behavior for video copy detection”, *14th ACM International Conference on Multimedia*, 2006, Santa Barbara, USA, pp.835 - 844.
[13] Law-To, J., Buisson, O., Gouet-Brunet, Boujemaa, N., “Video copy detection on the internet”: The challenges of copyright and multiplicity”, *IEEE International Conference on Multimedia & Expo*, 2007, Beijing pp. 2082 - 2085.
[14] Joly, A., Frélicot, C., Buisson, O., “Content-based video copy detection in large databases: A local fingerprints statistical similarity search approach”, in Proceedings of the *International Conference on Image Processing*, 2005.
[15] Sarkar, A., Ghosh, P., Moxley, E., Manjunath, B. S., “Video Fingerprinting: Features for Duplicate and Similar Video Detection and Query-based Video Retrieval”, Proceedings of *SPIE - Multimedia Content Access: Algorithms and Systems II*, 2008.
[16] Massoudi, A., Lefebvre, F., Demarty, C.H., Oisel L., Chupeau, B., “A Video Fingerprint Based on Visual Digest and Local Fingerprints”, 2006 *IEEE International Conference on Image Processing*, Issue 8-11, 2006, pp. 2297 - 2300.
[17] Chen, L., Stentiford,F. W. M., “Video sequence matching based on temporal ordinal measurement”, *Pattern Recognition Letters*, Vol. 29 , Issue 13, 2008, pp. 1824 – 1831.
[18] Hua, X.-S., Chen, X., Zhang, H.-J., 2004. “Robust video signature based on ordinal measure”, in: Proceedings of the. *IEEE International. Conference on Image Processing* (ICIP), 2004, Vol. 1, 24–27, 2004, pp. 685–688.
[19] Kim, C., Vasudev, B., “Spatio-temporal sequence matching for efficient video copy detection”, in Proceedings of the *IEEE Transactions on Circuit Systems Video Technology*, 15 (1), 2005, pp.127–132.
[20] Yuan, J., Duan, L. Y., Tian, Q., Ranganath S., and Xu C., “Fast and robust short video clip search for copy detection,” in *Springer: Lecture Notes in Computer Science* - 3332, pp. 479–488, 2004
[21] Indyk, P., Iyengar, G., Shivakumar, N., “Finding Pirated Video Sequences on the Internet”, Stanford Infolab, 1999.
[22] Radhakrishnan, R., Bauer, C., “Robust Video Fingerprints Based on Subspace Embedding”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, Las Vegas, pp. 2245 – 2248.
[23] Coskun, B., Sankur, B., Memon, N., “Spatio-temporal transform based video hashing”, *IEEE Transactions on Multimedia*, Vol. 8, No. 6, 2006.
[24] Roover, C. De, Vleeschouwer, C. De, Lefebvre, F., Macq B., “Robust video hashing based on radial projections of key frames”, *IEEE Transactions on Signal Processing,* Vol 53, Issue:10, 2005, pp. 4020 - 4030.

[25] Garboan, A., Mitrea, M., Prêteux, F., "DWT-based Robust Video Fingerprinting", Proceedings for the "*3rd European Workshop on Visual Information Processing*" (EUVIP), 2011, Paris, pp. 216 - 221.
[26] Garboan, A., Mitrea, M., Prêteux, F., "Video retrieval by means of robust fingerprinting", Proceedings for the "*15th IEEE Symposium on Consumer Electronics*" (ISCE), 2011, Singapore, pp. 299 - 303.
[27] Dutta, D, Saha, S.K., Chanda, B., "A hypothesis test based robust technique for video sequence matching", *International Journal of Future Generation Communication and Networking*, Vol. 3, No.3, 2010.
[28] Naphade, M. R., Yeung, M.M., Yeo, B.L., "Novel scheme for fast and efficent video sequence matching using compact signatures", In Proc. *SPIE, Storage and Retrieval for Media Databases 2000*, Vol. 3972, 2000, pp. 564-572.
[29] Gauch, J. M, "Real-time feature-based video stream validation and distortion analysis system using color moments", United States Patent 6246803.
[30] Sánchez, J. M., Binefa, X., Vitrià, J., Radeva, P., "Local Color Analysis for Scene Break Detection Applied to TV Commercials Recognition", Proceedings of the *Third International Conference on Visual Information and Information*, 1999, pp: 237 – 244.
[31] Hill, M., Hua, G., Natsev, A., Smith, J.R., Xie, L., Huang, B., Merler M., Ouyang, H., Zhou M., "IBM Research TRECVID-2010 Video Copy Detection and Multimedia Event Detection System", Proceedings of *TRECVID 2010*.
[32] Jégou, H., Gros, P., Douze, M., Schmid, C., Gravier, G., "INRIA LEAR-TEXMEX: Video Copy Detection Task", Proceedings of *TRECVID 2010*
[33] Mukai, R., Kurozumi, T., Kawanishi, T., Nagano, H., Kashino, H. "NTT Communication Science Laboratories at TRECVID 2011 Content Based Copy Detection", Proceedings of *TRECVID 2011*.
[34] Foucher, S., Lalonde, M. Gupta, V., Darvish, P., Gagnon L., Boulianne, G., "CRIM Notebook Paper - TRECVID 2011 Surveillance Event Detection", Proceedings of *TRECVID 2011*.
[35] Fang, X. Su, C. Xu, T., Xia, Z., Peng, P., Wang, Y., Tian, Y., Zhang, H., Wang, F., using a Cascade of Multimodal Features & Temporal Pyramid Matching", Proceedings of *TRECVID 2011*.
[36] Bhat, D. N., Nayar, S. K., "Ordinal Measures for Visual Correspondence", in Proceedings of the Conference on Computer Vision and Pattern Recognition, 1996.
[37] Dumitru, O., Mitrea, M., Prêteux, F., "Video Modelling in the DWT domain", Proceedings. *SPIE*, Vol. 7000, 2008, Strasbourg, pp. 7000 OP: 1-12.
[38] Buccigrossi, R., Simoncelli ,E., "Image Compression via Joint Statistical Characterization in the Wavelet Domain", *IEEE Transactions. on Image Processing*, Vol.8, No.12, 1999, pp. 1688 - 1700.
[39] Walpole, R.E., Myers, R.H., Myers, S-L., Ye, K., "Probability & Statistics for Engineers and Scientists", *Pearson Educational International*, 2002.
[40] http://www.hd3d.fr/
[41] http://www-artemis.it-sudparis.eu/
[42] Chupeau, B., Massoudi, A., Lefèbvre F., "In-theater piracy: Finding where the pirate was" *SPIE'08, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, 2008.
[43] http://ffmpeg.org/

[44] Yeh, M-C., Hsu, C-Y., Lu, C-S.; “NTNU-Academia Sinica at TRECVID 2010 Content Based Copy Detection” in Proceedings of *TRECVID 2010*.

[45] Barrios, J. M., Bustos, B., “Content-Based Video Copy Detection: PRISMA at TR*ECVID 2010*”, in Proceedings of *TRECVID*

A.3.2. Conference paper

Garboan, A., Mitrea, M., Prêteux, F., "Video retrieval by means of robust fingerprinting", Proceedings for the *IEEE 15th Symposium on Consumer Electronics* (ISCE), 2011, Singapore, pp. 299 - 303.

Video retrieval by means of robust fingerprinting

A. Garboan[1,3], *M. Mitrea*[1,3], *F. Prêteux*[2, 3]
[1]Institut Télécom - Télécom SudParis, Department ARTEMIS ; [2]MINES ParisTech;
[3]UMR CNRS 8145 MAP5
9, rue Charles Fourier, 91011 Evry France
adriana.garboan@it-sudparis.eu, mihai.mitrea@it-sudparis.eu, francoise.preteux@mines-paristech.fr

Abstract

Uniquely identifying visual content remains a challenging issue for a large variety of nowadays applications, as video browsing, database search and multimedia security, for instance.

In this respect, our study brought to light a simple yet efficient fingerprinting technique allowing short video sequences to be tracked. Three corpora, all of them containing 3780 video excerpts, with different excerpts lengths (20 seconds, 40 seconds and 60 seconds) were considered in the experiments. The quantitative results established that the average probability errors for both missed detection and false alarm are lower than 0.0007. These good practical results derive from the very fine mathematical properties of stationarity governing the DWT coefficients representing the fingerprint.

Index Terms — robust video fingerprinting, DWT, robustness, uniqueness.

1. INTRODUCTION

The worldwide mass production context brought technology closer to people. Affordable capturing, processing and storage devices along with the wide spread of broadband Internet access, empowered people to easily produce, manipulate and distribute large amounts of visual content.

Such a situation raises complex challenges in various multimedia domains (copyright protection, illegal distribution and management of massive databases, …). Despite the particular applicative challenge, issues connected to identification, authentication, indexation, retrieval, searching, navigation, organization and manipulation have to be addressed.

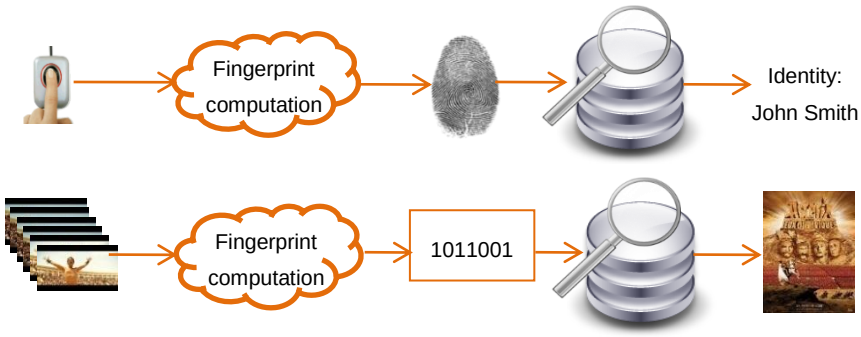


Figure 1: Human and video fingerprinting

A solution that is currently being intensively considered in research studies is video fingerprinting.

Video fingerprints can be best defined in relation with the human fingerprints, [1], as illustrated in Figure 1.

While the human fingerprint can be seen as a human summary (a signature) that is unique for every person, the video fingerprint can be seen as short video features (*e.g.* a string of bits with no particular format constraint) which can uniquely identify every video.

Fingerprinting methods have three main characteristics:

- *Robustness to distortions*: fingerprints extracted from a video subjected to content-preserving distortions (attacked video) should be similar to the fingerprints extracted from the original video. A robust fingerprinting method should ensure low probability of missed detection, *i.e.* a low probability of not retrieving an attacked video registered in the database.
- *Uniqueness*: fingerprints extracted from different video clips should be considerably different. A fingerprinting technique featuring uniqueness should ensure low probability of false alarm, *i.e.* low probability in retrieving a video sequence which is neither the query nor its replicas.
- *Database search efficiency*: for applications with a large scale database, fingerprints should be conducive to efficient database search (fast fingerprint computation and matching, compact form, …).

The present study is focused on a video retrieval applicative scenario. In this respect, a video sequence (arbitrarily chosen) is used as a query, and its would-be replicas are searched for in the database (*i.e.* within the reference video sequences.

In a previous study [2], the authors have already addressed the Discrete Wavelet Transform (DWT) video fingerprinting issue. Although that method proved to be very efficient in practice (resulting in probability of false alarm and missed detection lower then 0.005), it is intrinsically limited by its empirical approach.

The present paper also deals with the DWT-based video fingerprinting, this time focusing on deriving a related method allowing for an objective, mathematical decision rule to be specified.

The present paper is structured as follows. Section 2 introduces an original method for video fingerprinting which is experimentally validated in Section 3. Conclusions are drawn and perspectives are opened in Section 4.

2. DWT-BASED VIDEO FINGERPRINTING

Due to its possibility of representing in a very compact way salient characteristic of the video content and to its low complexity, the 2D-DWT is already intensively considered in practically all image processing applications, from compression and watermarking to default finding in garments. Our study investigated whether the 2D-DWT can be employed in order to uniquely and robustly identify the visual content.

In this respect, a new video fingerprinting method is advanced, Figure 2.

In order to extract a fingerprint from a video sequence (arbitrarily chosen), that sequence is first pre-processed, then a 2D-wavelet transform is applied to its frames and finally a certain selection of the coefficients will be carried on in order to obtain the fingerprint *per-se*.

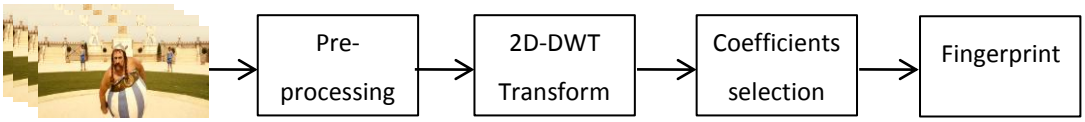


Figure 2: Overall fingerprinting method principle

2.1. Fingerprinting computation principle

Be there a video sequence, represented into a given format (compressed or not).

The pre-processing step is aimed at increasing the invariance of the envisioned fingerprint to different video processing operations, be they malicious (attacks) or mundane (ordinary video manipulation).

First, the video sequence is decoded in frames in order to diminish the influence of a particular video format or codec. Second, a temporal sub-sampling to 1 fps is performed in order to eliminate the redundancy between adjacent frames and to speed-up the fingerprint computation.

Third, a spatial re-sampling to $W \times H$ pixels (in our experiments, $W = 352$ and $H = 288$) is performed on the sampled frames, thus increasing the robustness of the method to frame size changes attacks.

Fourth, in order to extract the salient information within the frame while reducing the computational costs, the down-sampled frames are represented in the HSV (Hue – Saturation – Value) color space with the V (luminace) component normalized to the $[0,1]$ interval.
In the 2D-DWT Transform step, a $(9,7)$ Daubechies wavelet transform at the resolution level of $Nr = 3$ is applied on the V component of every sampled frame.

The coefficients selection step leads to a dichotomy in the study of wavelet coefficients employed as video fingerprints.

First, a previous study reported in [2] employed as fingerprint the 360 highest absolute value coefficients from the HL_{Nr} and LH_{Nr} frequency sub-bands of the transformed V component of every frame in the two considered video sequences (the query and the reference).

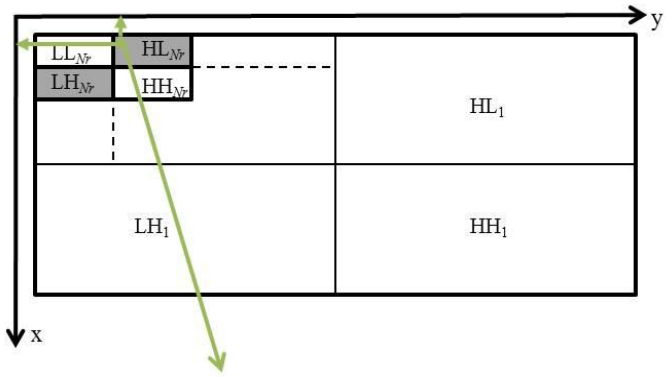
This approach leads to an empiric matching procedure yielding satisfactory results on the testing database. The testing database was the same as in the current study [3]. For a heuristically determined optimal matching threshold, the average probability of false alarm and the average probability of missed detection were both lower than 0.005.

However, as the method did not rely on scientific ground it cannot be proposed for general applications (its performances are *a priori* depending on the investigated corpus).

Second, the new strategy employed in the present paper relies on a coefficients selection dependent on the role of the video sequence (query or reference).

The coefficients selection (*i.e.* the fingerprint) proposed here aims at conveying more information about the spatial distribution of salient features within the frames as compared to the proposal in [2]. In this respect, the fingerprint of the query video sequence is computed and then, the fingerprints of the reference video sequences are obtained by using some spatial information provided by the query fingerprint.

For the query video sequence, the $R = 360$ highest absolute value coefficients from the HL_{Nr} and LH_{Nr} frequency sub-bands of the transform V component, together with their locations are selected and stored in the coefficients matrix (as illustrated in Figure 3).



Coef DWT	0.82	0.79	0.65	…	0.02
x	15	39	43	…	3
y	62	22	29	…	47

Figure 3: Coefficients matrix for a frame

While only the DWT coefficients compose the fingerprint of the query video sequence, the locations of these coefficients will be used to compute the fingerprint of the reference video sequences.

The coefficient matrix in Figure 3 illustrates the fingerprint of a sampled frame, while the fingerprint of the entire query video sequence is presented in Figure 4 and it is called the rank matrix.

The rank matrix is filled-in with all the fingerprints computed on the N sampled frames. Because the fingerprints of the frames consist of $R = 360$ DWT coefficients sorted in a decreasing order, it can be considered that the coefficients are disposed on 360 ranks, with "1" being the highest and "360" being the smallest value coefficient. This approach will turn to be particularly useful for fingerprint matching.

Ranks / Frames	1	2	3	…	359	360
1	0.822	0.540	0.272	….	0.056	0.056
2	0.465	0.458	0.405		0.031	0.030
…	….	….	….	….	….	….
N	0.794	0.687	0.545	….	0.165	0.165
N-1	0.530	0.491	0.436	….	0.072	0.072

Figure 4: Rank matrix of DWT coefficients

In the computation of the fingerprint for a reference video sequence, the DWT coefficients are selected from the HL_{Nr} and LH_{Nr} frequency sub-bands of the V transform component from the locations indicated as salient by the query coefficients matrices. After selecting the salient coefficients from every sampled frame of the reference video, the rank matrix will be obtained.

2.2 Fingerprint matching

The proposed similarity measure between fingerprints is the normalized correlation as given by the formula in (1).

$$\rho = corr_k(f,t) = \frac{1}{N-1}\sum_{x,y}\frac{(f_k(x,y)-\bar{f}_k)(t_k(x,y)-\bar{t}_k)}{\sigma_{f_k}\sigma_{t_k}} \qquad (1)$$

In (1), f_k, t_k designate the DWT coefficients of the query and the reference videos on a rank k; $\bar{f}_k, \bar{t}_k$ are the mean values of the DWT coefficients on the considered rank, while $\sigma_{f_k}, \sigma_{t_k}$ are the related standard deviations, respectively. N designates the number of DWT coefficients in every rank k, *i.e.* the number of sampled frames in each video sequence.

A perfect match (identity) between the query and the reference rank is obtained when $|\rho| = 1$; a value $\rho = 0$ indicates no correlation between f_k and t_k

While for the method in [2] the correlation was computed between the DWT coefficients without any organization or hierarchy of the coefficients, for the current method, the correlation is computed between the coefficients disposed on ranks, *i.e*. the columns of the rank matrix. Such a strategy is justified by the statistical investigation on the DWT coefficient behavior in [4]: it was proved that the values taken by a rank in the DWT coefficient hierarchy feature stationarity and the corresponding probability density function was estimated by a mixture of Gaussian laws. Hence, the stationarity property of these coefficients ensures a certain degree of independence of the results with respect to the experimental corpus.

In practice, in order to be able to also retrieve content preserving replicas, the absolute value of the normalized correlation should be compared to some threshold T; should $\rho \geq T$, then the query and the reference ranks are considered as identical.

The value of the T threshold is statistically determined according to the Rho test on correlation [5]. This test is individually applied to each of the $R = 360$ ranks under investigation; the null/alternative hypotheses are:

$$\begin{cases} H_0 : the\ ranks\ are\ not\ correlated \\ H_1 : the\ ranks\ are\ correlated \end{cases}$$

A match between the query and the reference video sequences is obtained when the majority of ranks (*i.e.* at least 181) is correlated. Should the majority of ranks be uncorrelated the query and the reference video sequences are considered as distinct.

Assuming the k ranked DWT coefficients from the query and from the reference video sequence are *i.i.d.* (identically and independently distributed) and that they follow a Gaussian distribution, and assuming the H_0 is true, the t_{test} value of the test statistics, see (2), follows a Student probability density function of $N-2$ degrees of freedom:

$$t_{test} = \rho \cdot \sqrt{\frac{(N-2)}{1-\rho^2}}, \qquad (2)$$

where N and ρ are the same as above.

If $t_{test} \leq z_{\alpha/2}$ (where $z_{\alpha/2}$ is the α-point value of the above-mentioned Student law), then the H_0 hypothesis is accepted, *i.e.* the DWT coefficients on the k rank are not correlated. If $t_{test} > z_{\alpha/2}$ the H_1 hypothesis is accepted, *i.e.* the DWT coefficients on the k rank are correlated.
In our experiments, we considered a significance level of $\alpha = 0.05$.

3. EXPERIMENTAL RESULTS

3.1 Video corpus

The quantitative results were obtained by processing 3 corpora consisting of 3, 6 and 9 hours of original content, belonging to 7 different movies from the HD3D-IIO corpus [3]. The content combines indoor and outdoor scenes, unstable and arbitrary lighting conditions, still and high motion scenes.

The corpora are composed of 540 video excerpts of either 20, 40, 60 seconds each, and of their 6 attacked versions. The following attacks have been considered: conversion to grayscale, contrast decrease, sharpening, small rotations (2°), linear filtering (Gaussian filter) and geometric (StirMark random bending) attacks simulating the in-theater camcordering.

Consequently, the final corpora are composed of 3780 sequences of 20, 40, 60 seconds each, *i.e.* 21, 42, 63 hours of video for each corpus, respectively.

3.2 Targeted application

The present experimental study is focused on video identification, Figure 5. The database is represented by one of the above corpuses containing 3780 video sequences of 20, 40, 60 seconds respectively.

When inquiring this database with one of its sequences, the 7 versions (one original and six attacked) should be retrieved. A missed detection occurs when at least one of the expected 7 versions is not retrieved. A false alarm is encountered when at least one sequence which is not related to the query is retrieved.

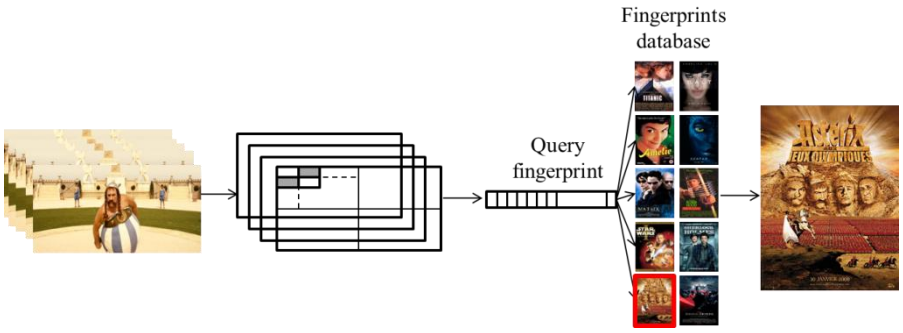


Figure 5: Video retrieval

Consequently, the robustness and the uniqueness of the method can be evaluated by computing the probability of false alarm and the probability of missed detection, according to (3) and (4), respectively:

$$P_{fa} = \frac{\#\ false\ alarm}{\#\ total\ tests} \tag{3}$$

$$P_{md} = \frac{\#\ missed\ detection}{\#\ total\ tests} \tag{4}$$

It can be noticed that the values in (3) and (4) are computed on a particular query. Average values for P_{fa} and P_{md} can be obtained by successively considering all the sequences in the database and by averaging the corresponding results.

3.3 Experimental results

The overall results, Figure 6, point to a very good retrieval accuracy, with false alarm and missed detection probability lower than 0.0007 for all the three considered corpora.

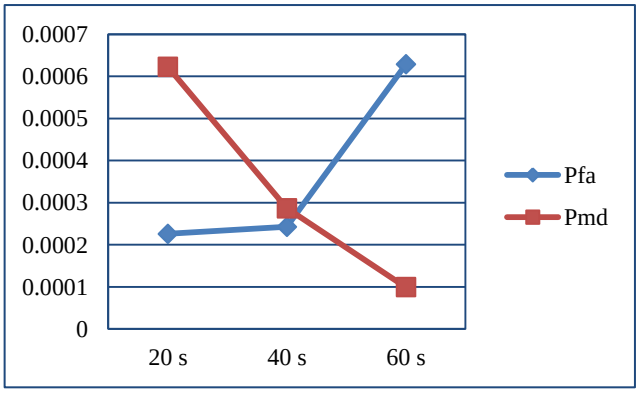


Figure 6: Average false alarm and missed detection probability

On the one hand, as an empiric rule (Figure 6), the missed detection probability decreases with the length of the video sequences, *i.e.* with the temporal information conveyed by the fingerprint.

On the other hand the probability of false alarm increases as the length of the video sequences increases. The balanced is reached for the 40 seconds corpus.

Figure 7 details the impact of each type of attack on the missed detection probability.

The quantitative results show that the most disturbing effects are induced by the geometrical attacks, *i.e.* StirMark and rotations with 2°.

It should be also noticed that for a given attack the longer the video sequence, the better the robustness of the fingerprint.

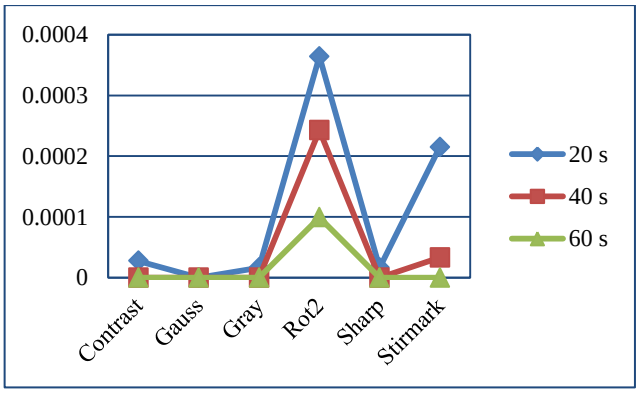


Figure 7: Missed detection probability for different attacks

3.4 Computational complexity

The main steps in our method are the computation of the DWT (hence, an $O(W \times H)$ complexity), the sorting of the corresponding coefficients (hence, an $O(R \log R)$ complexity and the matching of the fingerprints by a normalized correlation coefficient (hence, an $O(R \log R)$ complexity). Figure 8 illustrates the computation times for the proposed fingerprinting method.

The task durations were computed using a system with the following configuration: Intel Xeon CPU processor at 2.8 GHz with 6 Go of RAM memory, with an operating system working on 32 bits.

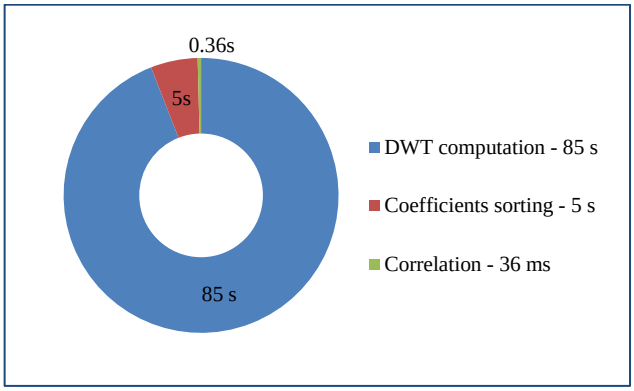


Figure 8: Computation time for the proposed video fingerprinting method

4. CONCLUSION AND PERSPECTIVES

The present paper presents a simple yet very efficient video fingerprinting method. The fingerprint is represented by a hierarchy of largest 2D-DWT coefficients selected from two low-frequency sub-bands. The fingerprint matching is carried on by a normalized correlation coefficient and the decision is based on the Rho test on correlation. Applied to different lengths of the video sequence (*i.e.* 20, 40, 60 seconds) and tested on reference corpora of 21, 42, 63 hours of visual content respectively, the method featured $P_{fa} \leq 10^{-3}$ and $P_{md} \leq 10^{-3}$, while ensuring a low complexity.

These good performances result from the stationarity of the information sources modeling the hierarchy of the DWT coefficients. A future direction of our research will be to exploit the probability density function of the DWT coefficients during the fingerprinting matching procedure.

The experiments performed in the present paper pointed out that the optimal length for a video sequence to be identified would be 40 seconds.

This 40 seconds duration seems to be long enough so to reflect the inner salient content of the sequence (hence, to reduce the probability of missed detection) but short enough so as to ensure low values for the false alarm probability.

Future work will be devoted to the integration of the proposed fingerprinting method for a movie identification application. This would suppose first a shot detection and then a search according to

our method. From the practical point of view, issues connected to the shot detection jitter are expected.

5. REFERENCES

[1] J. Oostveen,T. Kalker, J. Haitsma, ”Feature Extraction and a Database Strategy for Video Fingerprinting”, Lecture Notes In Computer Science, Vol. 2314 archive, Proceedings of the “*5th International Conference on Recent Advances in Visual Information Systems”*, pp: 117 - 128, 2002.

[2] A. Garboan, M. Mitrea, F. Prêteux, “DWT-based Robust Video Fingerprinting”, submitted to the “*European Workshop on Visual Information Processing (EUVIP) 2011”*, July 2011, Paris, France

[3] http://www.hd3d.fr/

[4] O. Dumitru, M. Mitrea, F. Prêteux, “Video Modelling in the DWT domain”, *Proc. SPIE*, Vol. 7000, April 2008, Strasbourg, France, pp. 7000 OP: 1-12

[5] R.E. Walpole, R.H. Myers, S. L. Myers, K. Ye, “*Probability & Statistics for Engineers and Scientists*”, Pearson Educational International, 200

Traçage de contenu vidéo : une méthode robuste à l'enregistrement en salle de cinéma

RESUME

Composantes sine qua non des contenus multimédias distribués et/ou partagés via un réseau, les techniques de *fingerprinting* permettent d'identifier tout contenu numérique à l'aide d'une signature de taille réduite, calculée à partir des données d'origine. Cette signature doit être invariante aux transformations du contenu. Pour des vidéos, cela renvoie aussi bien à du filtrage, de la compression, des opérations géométriques (rotation, sélection de sous-région…) qu'à du sous-échantillonnage spatio-temporel. Dans la pratique, c'est l'enregistrement par caméscope directement dans une salle de projection qui combine de façon non linéaire toutes les transformations pré-citées.

Par rapport à l'état de l'art, sous contrainte de robustesse à l'enregistrement en salle de cinéma, trois verrous scientifiques restent à lever : (1) unicité des signatures, (2) appariement mathématique des signatures, (3) scalabilité de la recherche au regard de la dimension de la base de données.

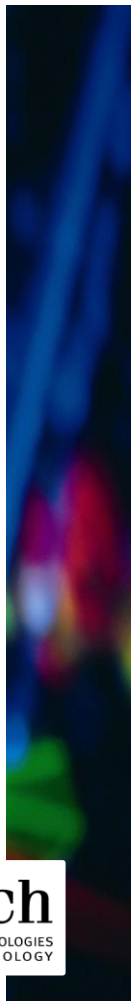
La principale contribution de cette thèse est de spécifier, concevoir, implanter et valider TrackART, une nouvelle méthode de traçage des contenus vidéo relevant ces trois défis.

L'unicité de la signature est obtenue par sélection d'un sous-ensemble de coefficients d'ondelettes, selon un critère statistique de leurs propriétés. La robustesse des signatures aux distorsions lors de l'appariement est garantie par l'introduction d'un test statistique Rho de corrélation. Enfin, la méthode développée est scalable : l'algorithme de localisation met en œuvre une représentation par sac de mots visuels. TrackART comporte également un mécanisme de synchronisation supplémentaire, capable de corriger automatiquement le *jitter* introduit par les attaques de désynchronisation variables en temps.

La méthode TrackART a été validée dans le cadre d'un partenariat industriel, avec les principaux professionnels de l'industrie cinématographique et avec le concours de la Commission Technique Supérieure de l'Image et du Son. La base de données de référence est constituée de 14 heures de contenu vidéo. La base de données requête correspond à 25 heures de contenu vidéo attaqué, obtenues en appliquant neuf types de distorsion sur le tiers des vidéo de la base de référence.

Les performances de la méthode TrackART ont été mesurées objectivement dans un contexte d'enregistrement en salle : la probabilité de fausse alarme est inférieure à 16 10^{-6}, la probabilité de perte inférieure à 0,041, de précision et de rappel égaux à 0,93. Ces valeurs représentent une avancée par rapport à l'état de l'art qui n'exhibe aucune méthode de traçage robuste à l'enregistrement en salle et constituent une première preuve de concept pour les technologies sous-jacentes.

MOT CLES : unicité, robustesse, scalabilité, sac à mots visuels, ondelettes, synchronisation, distorsions, augmentation/diminution de la luminosité, diminution du contraste, conversion en niveaux de gris, filtrage Gaussien, le rehaussement, rotation 2°, StirMark.



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

Towards camcorder recording robust video fingerprinting

ABSTRACT

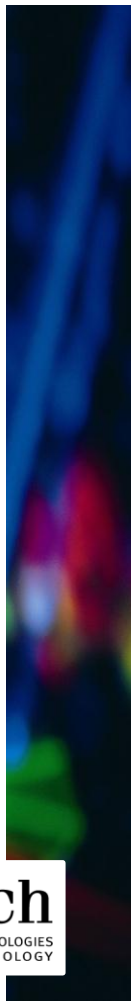
Sine qua non component of multimedia content distribution on the Internet, video fingerprinting techniques allow the identification of content based on digital signatures computed from the content itself. The signatures have to be invariant to content transformations like filtering, compression, geometric modifications, and spatial-temporal sub-sampling/cropping. In practice, all these transformations are non-linearly combined by the live camcorder recording use case.

The state-of-the-art limitations for video fingerprinting can be identified at three levels: (1) the uniqueness of the fingerprint is solely dealt with by heuristic procedures; (2) the fingerprinting matching is not constructed on a mathematical ground, thus resulting in lack of robustness to live camcorder recording distortions; (3) very few, if any, full scalable mono-modal methods exist.

The main contribution of the present thesis is to specify, design, implement and validate a new video fingerprinting method, TrackART, able to overcome these limitations. In order to ensure a unique and mathematical representation of the video content, the fingerprint is represented by a set of wavelet coefficients. In order to grant the fingerprints robustness to the mundane or malicious distortions which appear practical use-cases, the fingerprint matching is based on a repeated Rho test on correlation. In order to make the method efficient in the case of large scale databases, a localization algorithm based on a bag of visual words representation (Sivic and Zisserman, 2003) is employed. An additional synchronization mechanism able to address the time-variants distortions induced by live camcorder recording was also designed.

The TrackART method was validated in industrial partnership with professional players in cinematography special effects (Mikros Image) and with the French Cinematography Authority (CST - Commission Supérieure Technique de l'Image et du Son). The reference video database consists of 14 hours of video content. The query dataset consists in 25 hours of replica content obtained by applying nine types of distortions on a third of the reference video content. The performances of the TrackART method have been objectively assessed in the context of live camcorder recording: the probability of false alarm lower than 0.000016, the probability of missed detection lower than 0.041, precision and recall equal to 0.93. These results represent an advancement compared to the state of the art which does not exhibit any video fingerprinting method robust to live camcorder recording and validate a first proof of concept for the developed statistical methodology.

KEYWORDS: uniqueness, robustness, scalability, bag of visual words, wavelets, synchronization, distortions, computer generated distortions, brightness increase/decrease, contrast decrease, conversion to grayscale, Gaussian filtering, sharpening, rotations with 2, StirMark live camcorder recording generated distortions.



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY