CONTENT-BASED VIDEO COPY DETECTION, TRACKING AND IDENTIFICATION OF MOVIE PIRATES

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

ROOPALAKSHMI

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE – 575025

APRIL, 2014

DECLARATION

By the Ph.D. Research Scholar

I hereby declare that the Research Thesis entitled **CONTENT-BASED VIDEO COPY DETECTION, TRACKING AND IDENTIFICATION OF MOVIE PIRATES** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfillment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Information Technology** is a *bonafide report of the research work carried out by me.* The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

(IT09F04, Roopalakshmi)

Department of Information Technology

Place: NITK, Surathkal.

Date:

CERTIFICATE

This is to *certify* that the Research Thesis entitled **CONTENT-BASED VIDEO COPY DETECTION, TRACKING AND IDENTIFICATION OF MOVIE PIRATES** submitted by **ROOPALAKSHMI**, (Register Number: IT09F04) as the record of the research work carried out by her, is *accepted as the Research Thesis submission* in partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy**.

Prof. G.Ram Mohana Reddy

(Research Guide)

Prof. Ananthanarayana V.S

(Chairman- DRPC)

Acknowledgements

I would like to take this chance to thank those people who have made this thesis possible. First and foremost, I would like to express my deepest gratitude to my research guide, *Prof. G. Ram Mohana Reddy*, of Information Technology Department, for giving me guidance and support throughout my research work. This thesis would not have been possible without his support and suggestions.

I express my heartfelt thanks to *Prof. V. S. Ananthanarayana*, HOD, Information Technology Department, for his timely suggestions and guidance. I would also like to thank members of my RPAC committee, *Prof. S. M. Hegde* and *Prof. K. P. Vittal*, for their valuable suggestions during my research work.

I acknowledge the financial support, which I have received from the Department of Science and Technology, Government of India, for pursuing this research work.

I would like to thank all my co-researchers *Mr.Kiran*, *Mrs.Megha*, *Mr.Melwyn* and *Ms.Pushpalatha* and others, who made the past three years more enjoyable and fun. I extend my sincere thanks to all teaching and administrative staff of the Information Technology department, for their support to complete this research work. I am grateful to all those people, who supported me in any kind during the completion of my Ph.D work.

Last, but not the least, my gratitude as well as sincere appreciations go towards my family including my mother-in-law, who gave unconditional cooperation, love and inspiration throughout my research work. Without them, surely, this research work would not have been possible.

Place: NITK, Surathkal ROOPALAKSHMI

Date: April, 2014.

ABSTRACT

Due to the exponential growth of multimedia technologies, numerous pirated contents are proliferating on the Internet and causing huge piracy as well as copyright issues. Therefore, this thesis investigates four different methodologies for combating piracy namely, Content-Based video Copy Detection (CBCD), duplicate video registration, geometric distortions computation and pirate position estimation in a movie theater.

In the first methodology, this thesis targets CBCD problem, by introducing different copy detection techniques, which employ efficient video fingerprints for detecting duplicate video clips. Precisely, this research work attempts to solve some of the issues of the CBCD domain, by proposing novel video copy detection techniques, that employ color, motion activity, audio and multimodal signatures.

In the second methodology, this thesis addresses the problem of video copy localization, by proposing robust registration schemes, which guarantee the accurate frame alignments of the pirate video with the master content. Specifically, this research study contributes robust temporal as well as spatio-temporal registration frameworks, which exploit visual-audio fingerprints for obtaining frame-to-frame alignments of the two video sequences.

In the third methodology, this thesis aims at geometric distortions estimation, by presenting a framework using visual-audio features, which computes the geometric distortions present in the duplicate video. In the fourth methodology, this thesis attempts to emphasize the capability of video fingerprints towards the pirate position estimation problem, by performing a Case study for investigating the illegal capture location in a theater.

Video copy detection, tracking, distortion estimation and pirate position approximation frameworks presented in this thesis, are evaluated with extensive experiments on different datasets. More specifically, the experimental results demonstrate the efficiency of the proposed methods over standard datasets such as TRECVID dataset, Open Video Project dataset, CC_WEB_VIDEO collection and real datasets comprising camcorded versions of popular movies. Further, In-Theater experiments and evaluations demonstrate the satisfactory performance of the proposed forensic framework in terms of statistical results, 2D and 3D views of position estimations.

Keywords: *Content-Based video Copy Detection (CBCD), Video copy registration, Geometric distortions estimation, Camcorder piracy, Video fingerprinting.*

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction of Piracy

Due to the exponential growth of multimedia and Internet technologies, numerous pirated (unauthorized or illegal copies) movies as well as videos are proliferating on the Internet, which cause huge piracy issues. In addition, due to the existence of these pirated clips, web search engines are facing severe problems for monitoring and managing their digital contents. For example, for a sample set of 24 popular queries from YouTube [1], Google Video [2] and Yahoo! Video [3], the search results are comprising 27% duplicates or near-duplicates (Wu et al. 2007). On the other hand, due to the easily available multimedia applications, simple steps are sufficient to duplicate, manipulate and distribute a video content. As a result, the downloading as well as distribution of illegal video contents on the Internet is unprecedented, which in turn triggers Internet piracy and also copyright issues. In this way, piracy is creating a devastating impact on the entertainment industry.

Strictly speaking, due to movie piracy, worldwide motion picture industry is losing billions of dollars from the past few years. For instance, the latest Canadian Motion Picture Distributors Association (CMPDA)-2011[4] report alarms that, 133 million pirated movies are viewed in Canada in 2010. This report also estimates the total loss to Canadian economy as C$895 million in 2010 due to movie piracy. Moreover, according to Motion Picture Association (MPA)[5], over 90% of the pirated

[1] YouTube. Available: `http://www.youtube.com`.

[2] Google Video. Available: `http://www.video.google.com`.

[3] Yahoo! Video. Available: `http://www.video.yahoo.com`.

[4] *Economic consequences of movie piracy-* CMPDA Feb 2011 report. `http://www.mpa-canada.org/press/IPSOS-OXFORD-ECONOMICS-Report_February-17-2011.pdf`

[5] *2005 US Piracy Fact Sheet.* Motion Picture Association of America. `http://www.mpaa.org/USPiracyFactSheet.pdf`

versions of newly released movies are created by camcorder piracy, which denotes illegal camcorder captures in theaters. Precisely, a typical camcorder piracy scenario is illustrated as follows:

- First, a pirate illegally captures a film in a movie theater using the camcorder and sells the recorded movie to replicators or illicit source labs.
- Then, the replicators illegally duplicate the recorded movies and rapidly produce thousands of pirated DVDs for sale.
- Consequently, the unauthorized copies of movies are distributed and downloaded through illegal file sharing networks, which cause Internet piracy.

In this way, the pirated versions of movies appear on the Internet or on the street market, within hours of official release of a movie.

Thanks to technical advances in camcorders, camcorder piracy is raised as a major issue for the movie industry over the past few years. Therefore, video copy detection as well as tracking methods are very much essential for restricting piracy and also for effective video retrieval. Due to these reasons, duplicate video detection and tracking techniques are evolving as active research fields from the past decades. More precisely, video copy detection techniques are required to prevent downloading as well as distribution of illegal contents on the Internet and as a result *Internet piracy can be controlled*. Further, illegal video investigation frameworks are needed to track the movie pirates, which eventually *leads to reduction of camcorder piracy*.

1.2 Motivation

Combating camcorder piracy requires duplicate video detection followed by the precise frame alignments of the duplicate video with the master content, in order to estimate the distortion model and camcorder capture location in a theater. In addition, detection and tracking of video copies are also useful in a large number of applications. Possible categories of applications include:

- **Digital Copyrights Protection and Content Management**

  Copy detection techniques are used to identify the illegal videos and restrict their distribution on the Internet; hence, they deter copyright violations. Further, duplicate video detection methods check the integrity of user uploads in User Generated Content (UGC) websites such as YouTube and thus, assist in digital content management.

- **Content-Based Video Retrieval (CBVR)**

  Massive growth of illegal videos is becoming a serious issue for online video repositories. For example, Cheung and Zakhor (2000) showed that, each web video in the database of size 45,000 clips, includes an average of five similar copies. Further, Yahoo![6] search engine returned 2 to 3 duplicates among the top ten retrievals for some popular queries (Liu et al. 2007). Therefore, identification and tracking of illegal video clips, enhance the retrieval capabilities of the CBVR system by eliminating duplicate contents.

- **Video Indexing and Mining**

  Duplicate video detection techniques are useful to efficiently index the video contents, since they provide compact fingerprints and support faster similarity searches (Sarkar et al. 2008). In addition, copy detection methods enable scalable mining of huge video databases by identifying the content links between the video sequences and their modified versions (Poullot et al. 2008).

- **Identification of TV Commercials**

  Copy identification methods are employed to search the repeated instances of a video clip inside a long video or video collection (Schoeffmann and Boeszoermenyi 2011). For example, if the companies want to monitor whether their commercials are broadcasted or not, for planning their marketing strategies, then it requires the identification of repeated instances (here, commercials) in a specific broadcast timeslot.

- **Tracking News Stories**

  Information association is very essential in this digital era. Duplicate content identification and tracking techniques can be used for semantically linking the news stories on the same topic across different sources, in order to provide a comprehensive view to the audience (Zhai and Shah 2005).

Though duplicate video detection and tracking methods are investigated from the past several years in the multimedia research; yet, efficient illegal video detection and investigation are really challenging issues, as they require compact fingerprints and involve considerable computational costs.

1.3 Issues and Challenges

Whether the problem is copy detection or pirate video alignment or distortion estimation, the major challenge lies often in compact representation of the video sequence

[6]Yahoo!. Available: `http://www.yahoo.com`.

using efficient fingerprints/feature descriptors. Hence, some of the prominent challenges of these problem domains are listed below:

- *Fingerprints extraction*
  Video fingerprints play a vital role in determining the performance of the video copy detection as well as tracking system, since huge databases need to be checked. Henceforth, while extracting fingerprints, the following properties must be considered:
  - *Robustness*
    Must be invariant to different patterns generated by the same source video.
  - *Discrimination ability*
    Two perceptually different videos must have distinguishable signatures.
  - *Compactness*
    Support low dimensional representation.
  - *Computationally efficient*
    Examine thousands of videos using minimum amount of computations.
  - *Fast search*
    Find a match in a very large database within a finite amount of time.
- *Computational cost*
  A major demanding issue in most of the copy detection systems is, the computational cost of feature extraction and similarity matching tasks, since multi-dimensional data is involved.
- *Size of the problem space*
  In duplicate detection systems the problem space is extremely large, which involves millions of video sequences. For instance, YouTube has to process 20 hours of uploaded videos in every minute (Junee 2009), which in turn demands computationally efficient real-time solutions.

1.4 The Complete Research Framework

The research work described in this Doctoral thesis comprises the following methodologies/fields, which can be used to track as well as control video piracy as follows:

1) Video copy detection
2) Pirate video tracking/registration
3) Geometric distortions estimation
4) Use case: Pirate position estimation and identification

Figure 1.1 shows the complete research framework carried out in this study, which comprises four different methodologies as illustrated below.

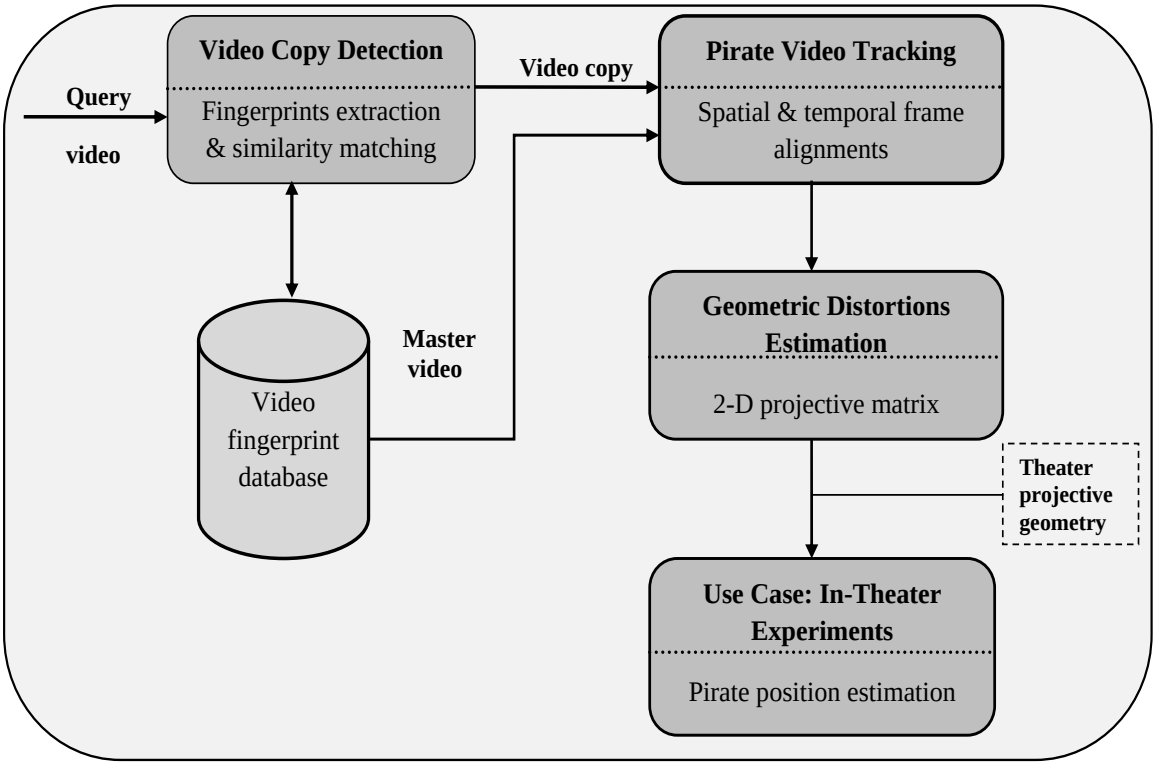


Figure 1.1: The complete research framework

1) Video copy detection: Video copy detection is commonly known as fingerprint-based video identification or more generally *Content-Based video Copy Detection (CBCD)*. A *video copy/pirate video* is a distorted video sequence derived from the *master/database video* by applying several video editing and transformations such as noise, cropping, zooming and caption insertions. In CBCD, a video copy is detected by comparing the fingerprints of the query video with the fingerprints of the database videos. Here, the media (image, audio etc.,) itself acts like a fingerprint to assess a duplicate video, which is similar to human fingerprints; hence called as video fingerprints. This research work targets CBCD problem, by introducing different copy detection schemes to detect illegal videos, which employ fingerprints derived from visual, audio and multimodal signatures.

2) Pirate video tracking/registration: After copy detection, it is essential to compute the accurate frame-to-frame alignments of the pirate content with the master sequence. For this purpose, video copy tracking techniques are presented in this thesis, which locate the given query clip within the master video content and

compute frame matches of two video sequences. More precisely, this research study targets pirate video tracking problem, by presenting robust spatial and temporal registration schemes to guarantee the accurate frame alignments of the copied video with the master content. In this research work, the terms *tracking* or *registration* define a way of mapping the pirate and master sequences with an objective to calculate frame-to-frame alignments.

3) Geometric distortions estimation: It is known that, when the pirate is recording a movie in a theater using a camcorder, the captured image is geometrically distorted, because capturing axis is not perpendicular to the screen. Therefore, the resultant geometric distortions in the illegal video need to be estimated, so that the subsequent forensic activities such as detecting embedded watermarks and estimating illegal capture locations can be carried out. For this reason, distortion estimation schemes are introduced in this thesis, which estimate the geometric distortions present in the duplicate video in terms of projective matrix.

4) Pirate position estimation and identification: Besides geometric distortion estimation, the application of video fingerprints could be further extended to estimate the illegal capture location in a theater. Specifically, it is possible to estimate the pirate position in a theater, by performing in-depth analysis of geometric distortions and the theater projective geometry. On the other hand, preventing camcorder piracy, by introducing forensic tracking frameworks to identify the movie pirate, is *certainly not* the aim of this thesis. Instead of that, this research study attempts to highlight the capability of video fingerprints towards the pirate position estimation problem. In other words, current research work tries to prove that, the illegal capture location in a theater could be approximated, by performing the exhaustive investigation of geometric distortions and the theater projective geometry. *To validate this viewpoint, In-Theater experiments are conducted and evaluated as a case study in this research work.* More specifically, the *purpose of the use case is to emphasize the ability of the video fingerprints towards the pirate position estimation problem.* It is, therefore *certainly not* the main goal of this thesis, to propose pirate position estimation frameworks, which could be used by theater owners for pin pointing exact seats and identifying the actual pirates.

1.5 Summary of Contributions

This thesis introduces different CBCD techniques for detecting the illegal video sequences on the Internet and thereby limits Internet piracy. Further, this research study also presents several pirate video tracking techniques, which compute spatio-

temporal frame alignments and estimate geometric distortion model of two video sequences, by employing multimodal fingerprints. Furthermore, this thesis investigates a forensic tracking framework, that attempts to estimate the illegal capture location in a theater with the help of audio-visual fingerprints. Brief explanation of each of the contributions is given below:

- **Content-Based Video Copy Detection (CBCD) Schemes**

  ⋆ ***Using Color-Based Visual Features***

  Color is one of the dominant and distinguishing visual feature of an image; hence, this thesis first introduces two CBCD techniques, that employ compact and computationally efficient visual fingerprints derived from Dominant Color Descriptors (DCDs) of MPEG-7 standard (Manjunath et al. 2002) to detect duplicate video contents. More precisely, DCD indicates the complete color information of an image with a small number of representative/dominant colors. The Generalized Lloyd algorithm (GLA) is the most extensively used algorithm to extract the dominant colors of an image (Lloyd 1982); yet, GLA suffers due to its expensive computational cost. From another perspective, a main challenging issue in CBCD systems is, the computational cost of fingerprints extraction and similarity matching, because huge video databases need to be checked. To solve this issue, compact feature descriptors with low computational cost are needed for effective video copy detection systems. Therefore, the current research study proposes a novel CBCD framework, which exploits a simple and computationally inexpensive DCD extraction scheme for detecting illegal videos, compared to the existing methods. In addition, this framework also incorporates an adaptive signature pruning method, which noticeably reduces the total DCD's extracted from a video sequence.

  Although two images may have similar dominant colors, but the spatial distribution of same color pixels in the two images, may not be always same. Based on this aspect, this thesis introduces an another video copy detection technique, by enhancing the previously proposed CBCD framework, which integrates spatial coherency factor along with the dominant color values. Specifically, spatial coherency describes the spatial distribution of pixels associated with each dominant color, i.e. pixels of same color are how much co-located. Inclusion of this spatial coherency factor in the video fingerprints significantly improves the detection performance. Experiments conducted on a database of 101 video sequences, demonstrate

the efficiency of the proposed CBCD schemes in terms of standard Precision and Recall metrics against various video transformations such as zooming, noise, resolution change and rotation.

⋆ ***Using Motion Activity Features***

Motion features contribute significant information about a video content. However, in the CBCD literature motion vectors are considered as poor descriptors (Hampapur et al. 2001), due to the following reasons:
a) They are close to zero values when captured at normal frame rates;
b) Raw motion vectors are noisy in nature and
c) Huge amount of information is needed to describe the motion content.
In addition, frame-based motion features may describe the temporal content of a video clip, yet they may not effectively characterize the overall activity of a video sequence. To tackle these discrepancies, this thesis introduces a new CBCD method, that integrates the temporal behavior and spatial distribution of motion activity for describing the overall activity of a video sequence. More precisely, the proposed CBCD framework employs the robust video fingerprints derived from the attributes of Motion Activity Descriptor of MPEG-7 standard (Manjunath et al. 2002) such as motion intensity, dominant direction and spatial distribution of activity for the copy detection task. Further, clustering based pruned search is performed, to speed up the similarity matching process. Experiments conducted on 75 hours of TRECVID 2007 Sound and Vision dataset[7], prove the improved detection efficiency of the proposed CBCD technique compared to reference methods.

⋆ ***Using Audio Fingerprints***

Acoustic features are significant and powerful in describing a video content; hence their exploitation may considerably enhance the performance of the copy detection task. On the other hand, past acoustic investigations prove that, the most important perceptual audio features exist in the frequency domain (Tang et al. 2009). By considering these factors, this thesis contributes a novel video copy detection system using robust audio spectral features. More precisely, spectral descriptors including Centroid, Energy, Roll-off and Flux are extracted from power spectrum of the audio signal. However, direct processing of resultant spectral features is computation-

[7]TRECVID 2010 Guidelines [Online]. Available: `http://www.nlpir.nist.gov/projects/tv2010/tv2010.html`.

ally expensive; hence they are combined into *SPectral Descriptive (SPD)* words and utilized as video fingerprints for the CBCD task. The experiments evaluated on 75 hours of TRECVID 2008 dataset, demonstrate the better accuracy of the proposed CBCD framework, when compared to the reference methods.

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used by the audio processing community to obtain discriminative performance with reasonable noise robustness (Rabiner and Juang 1993). By keeping this factor in mind, this thesis presents an another copy detection technique, by enhancing the previously presented CBCD method, which incorporates MFCCs and spectral descriptors for detecting duplicate video sequences. Specifically, MFCCs and four spectral descriptors are extracted from the spectral decomposition of the downsampled audio signal. Then, the resultant intra-frame features are concatenated into *Multi-Feature (MF)* vectors, which effectively represent frame-level and clip-level audio content of a video. After this step, Principal Component Analysis (PCA) is applied on the MF vectors, in order to get compact video fingerprints. Experiments conducted on 25 hours of TRECVID 2007 dataset prove the efficiency of the proposed scheme against 16 different types of video modifications such as color change, pattern insertion, moving caption and combined attacks.

⋆ ***Using Multimodal Features***

In general, if audio content is available, then the joint utilization of visual-audio fingerprints for detecting illegal videos may significantly improve the copy detection performance. Based on this aspect, this thesis contributes a robust CBCD framework, which employs visual fingerprints derived from DCDs and audio signatures extracted from MFCCs for detecting duplicate videos. More precisely, a novel visual signature called as *Spatio-Temporal DCDs* are generated, which effectively characterize the region-based dominant color features as well as temporal color information present in the given video. Then, the resultant visual signatures are jointly employed along with the robust audio fingerprints for identifying the duplicate video clips. The results tested on 100 hours of TRECVID 2008 and 2009 datasets indicate that, the proposed framework consistently outperforms the reference methods against 17 different types of video transformations.

Usually, exploiting several complementary features of a video sequence for the copy detection task, not only improves the detection performance,

but also widens the coverage to more number of video transformations. Based on this viewpoint, this thesis proposes a new copy detection system, which integrates motion activity features and audio spectral descriptors to detect the pirate video sequences. More specifically, *Motion Activity (MA)* words describing the spatio-temporal activity of a video as well as *Spectral Descriptive (SD)* signatures illustrating audio spectrum of a clip are jointly utilized for detecting illegal videos. Experiments on 200 hours of TRECVID 2009 dataset, prove the better detection rates of the proposed in terms of standard Precision, Recall and F-Measure metrics.

- **Pirate Video Tracking Techniques**

  - ⋆ ***Temporal Registration***

    Tracking piracy needs illegal video detection followed by the accurate frame alignments of the duplicate video with the master content, in order to estimate the geometric distortion model of the two video sequences. In this perspective, this thesis contributes a new temporal registration technique, that exploits MFCCs and motion activity features for obtaining frame-to-frame alignments of the two video contents. More precisely, the proposed registration scheme first extracts the robust motion profile, which describes the temporal and spatial motion activity of video contents. Then, it selects the most similar segment of the master video by employing sliding window-based dynamic programming technique. After this step, multi-features based frame matching scheme is employed in order to obtain the accurate frame alignments of the pirate video with the master content. Experiments are conducted on a master database comprising 150 hours of TRECVID-2008, 2009 datasets and a query dataset including 850 video clips. The registration results demonstrate the robustness of the proposed framework against a wide range of video transformation types such as scaling, temporal, filtering, geometric and combined distortions.

  - ⋆ ***Spatio-Temporal Registration Using Visual Features***

    Followed by the temporal registration scheme, this thesis presents a spatio-temporal alignment technique for mapping the pirate and master video contents, which employs visual signatures derived from Speeded Up Robust Features (SURF) descriptors (Bay et al. 2008). Precisely, first the proposed registration scheme computes a new visual signature called as *1-D SURF* of video contents, which are subsequently mapped to achieve temporal frame-

to-frame matches. Then, a small set of representative frame pairs from the two video sequences are extracted and consecutively mapped by means of their SURF key points in order to get accurate geometric frame alignments. Experiments are carried out on a master database consisting 100 hours of TRECVID 2009 dataset plus another 30 hours of real data comprising camcorded copies of master sequences. The experimental results indicated in terms of *percentage of perfectly Matched Frames* (*MF*) and *Average Distance (AD) between frame indexes*, demonstrate the efficiency of the proposed scheme against a broad range of video transformations.

⋆ ***Spatio-Temporal Registration Using Visual-Audio Features***

Generally, if audio content is available, then the joint exploitation of visual-audio fingerprints for alignment task, could enhance the registration accuracy to a greater extent. Based on this aspect, this thesis contributes a new spatio-temporal registration framework, which utilizes content-based multimodal signatures for obtaining the accurate frame alignments of pirate and master video sequences. More precisely, first the proposed registration framework introduces a new visual fingerprint denoted as *1-D visual profile*, which is extracted from SURF key points of frames. Then, the most similar segment of the master video known as *Candidate Segment* is selected using dynamic time warping (DTW) algorithm, which substantially reduces the frame matching cost. Further, the proposed framework employs a multimodal frame matching scheme, which aligns visual-acoustic fingerprints and noticeably reduces false frame matches. Furthermore, *principal frames* extraction algorithm is introduced, which extracts the most similar frames from the temporally aligned master and pirate video sequences. Finally, the resultant *principal frames* are mapped using their SURF descriptors in terms of control points, to get accurate spatial frame alignments.

The proposed spatio-temporal registration framework is evaluated on three different datasets, namely TRECVID Sound & Vision dataset, CC_WEB_VIDEO dataset [8] and a set of real data comprising camcorded versions of master videos. The TRECVID master database includes approximately 190 hours of data collected from TRECVID 2008 as well as 2009 datasets and 750 query video clips. The CC_WEB_VIDEO master database consists of 24 most viewed videos provided by CC_WEB_VIDEO collection. The CC_WEB_VIDEO query dataset includes approximately 600 video

[8] CC_WEB_VIDEO: Near-Duplicate Web Video Dataset. `http://vireo.cs.cityu.edu.hk/webvideo/`

files with two different classes of distortions namely formatting and content modifications. To evaluate the performance of proposed framework against the camcorder captured video clips, a database of 30 master video sequences and 75 camcorded copies ranging from 1.55min to 15min are considered. The proposed spatio-temporal registration framework achieves promising results in terms of better *MF* and *AD* rates compared to the reference methods. Frame alignment using the sliding window technique, is an another beneficial characteristic of the proposed framework, which demonstrates that, the effective performance can be obtained with the lowest computational cost, although the fingerprint extraction cost is higher.

- **Geometric Distortions Estimation Method**

  Estimating geometric distortions in the pirate video is prerequisite, in order to proceed with the forensic activities such as approximating the pirate position in theater and recovering embedded watermarks. Therefore, this thesis contributes a new distortion model estimation framework, that employs multimodal fingerprints compared to the existing visual-features based methods. More specifically, the proposed framework first introduces a novel visual fingerprint, denoted as *Compact Spatio-Temporal (CST) SURF* signature, which represents spatial and temporal content of videos. Then, robust audio signatures derived from MFCCs and *CST-SURF* signatures are utilized to obtain the accurate frame alignments of pirate and master video contents. Further, it presents the *Most Similar* Segment selection algorithm to obtain effective temporal alignments, by using Minimum Weight Perfect Bipartite Matching technique. Furthermore, the proposed framework introduces *stable frame pairs selection* algorithm, for extracting the most similar frame pairs and also two filtering policies for obtaining the robust key point pairs. After this step, the proposed scheme estimates the geometric distortion model in terms of 8-parameter homographic matrix using Normalized Direct Linear Transformation (DLT) algorithm (Hartley and Ziserman 2004). The results of experiments on 7 popular movies and 84 camcorded copies of movies prove the promising results of the proposed framework compared to the reference methods.

- **Pirate Position Estimation Framework**

  Followed by the estimation of geometric distortion model, the resultant visual-audio fingerprints could also be successfully exploited for approximating the position of pirate in a theater. In order to prove this view, this thesis contributes

a forensic tracking framework using acoustic-visual features for estimating the position of the pirate in a movie theater. More specifically, first the proposed position estimation framework computes spatio-temporal frame alignments of the source movie and pirate video contents by making use of both the acoustic as well as visual features. Then, the geometric distortions in the pirate video are estimated in terms of 3×3 projective matrix. Consequently, the camcorder optical axis to the screen perpendicular is determined by redefining the theater projective geometry and eventually the position of the pirate in the movie theater is estimated.

To analyze the performance of the proposed position estimation framework, the experiments are conducted in a large-scale test environment with 176 seats and ten arbitrary locations are employed for camcorder captures. The statistical analysis of position estimation results demonstrate the satisfactory performance of the proposed forensic framework. More precisely, the mean absolute error of estimation results is (38.25, 22.45, 11.11)cm and the standard deviation of the estimation errors is (22.26, 12.97, 7.29)cm respectively. Further, mean width error ranges from 2.5-63.5cm, while mean depth error ranges from 0.6-30.1cm. This estimation accuracy is quite reasonable, as the distance between two seats in a row is about 35cm and the distance between two rows is about 100cm respectively.

1.6 Outline of the Thesis

This thesis is organized as follows. Chapter 2 first introduces the concepts of content-based video copy detection, tracking, distortion model computation and pirate position estimation methodologies followed by the discussion of respective state-of-the-art techniques. Chapter 3 illustrates the proposed CBCD techniques, that employ features such as color, motion activity, audio and multimodal signatures for detecting illegal videos.

Chapter 4 describes video copy registration frameworks, which focus on spatial and temporal alignments of the pirate video with the master sequence. The key contribution of this chapter is, a novel spatio-temporal registration technique using visual-audio fingerprints for the localization of video copies. Further, this chapter details the experimental evaluations on 3 different datasets, which prove the efficiency and effectiveness of the proposed framework against a wide range of video transformations. Chapter 5 addresses geometric distortion estimation problem by proposing

a new framework to estimate the geometric distortions in video copies, by employing acoustic and visual features. The experimental evaluations on popular movies and their camcorded versions, demonstrate the consistent performance of the proposed scheme compared to the reference methods.

Chapter 6 describes the case study involving In-Theater experiments, that deals with pirate position estimation problem. Precisely, this chapter details the proposed forensic tracking framework, that exploits visual-audio fingerprints for estimating the position of the pirate in a movie theater. The experimental results in terms of top, isometric and 3-D views of actual as well as estimated camcorder locations, prove the satisfactory performance of the proposed forensic tracking framework. Chapter 7 summarizes the contributions and highlights the possible directions for future work.

1.7 Summary

This chapter first introduces the piracy problem in terms of Internet as well as Camcorder piracy issues, followed by their devastating impact on the entertainment industry. To deal with these issues, a research framework comprising four different methodologies namely, CBCD, video copy registration, geometric distortions estimation and pirate position approximation is introduced in this chapter. Further, this chapter also describes some of the prominent challenges of these problem domains, followed by the summary of contributions of the current research study.

Chapter 2

Literature Survey

2.1 Content-Based Video Copy Detection (CBCD)

2.1.1 Why CBCD?

digital watermarking and Content-Based video Copy Detection (CBCD) techniques are widely used in the literature to detect illegal video contents (Sarkar et al. 2010). The watermarking approach embeds an identifier (watermark) in to the master video before distribution and checks the identifier in the query video to detect whether it is copyrighted or not. The alternative CBCD techniques employ the media (like image, audio, video etc.,), that contain unique information for detecting illegal video clips (Chiu et al. 2010). In other words, CBCD methods utilize video fingerprints extracted from the content-based features to assess a video copy (Esmaeili et al. 2011). Watermarking-based systems are well studied and explored, whereas CBCD approaches are still in the early stages (Li et al. 2010). However, the CBCD techniques are more successful compared to digital watermarking methods due to the following key features:

- Fingerprint generation neither destroys nor damages the video content.
- More robust than fragile watermarking techniques, since the fingerprints remain mostly unchanged even after various video modifications.
- Capable of detecting video copies, even if the master video is not watermarked.
- Fingerprint extraction is possible even after the distribution of digital media, whereas embedding watermarks is difficult after distribution, since the media gets distorted.

Hence, the complementary CBCD methods are emerging as primary tools for dealing with digital video piracy and also generated a great deal of research interest recently.

2.1.2 Related work

In the copy detection literature, considerable efforts are made to propose efficient video fingerprints and effective similarity matching techniques. Early research on CBCD can be broadly classified into two groups namely, global features and local descriptors techniques.

Global descriptors/features based CBCD methods

Global descriptors summarize the global statistics of low level features in the entire frame; hence they are also called as frame-level descriptors (Chiu and Wang 2010). Initial work in this group comes from Shivakumar (1999) and Indyk (2000), in which fingerprinting technique is introduced to identify the pirated video sequences on the Internet. Hampapur et al. (2001) performed a comparative study of different descriptors such as motion, intensity and color-based signatures, that are employed in the CBCD domain. At present, global features such as the Ordinal measure (Hua et al. 2004; Chaisorn et al. 2010) and color histograms (Naphade and Yeo 2000; Liu et al. 2007) are very popular in the CBCD domain. However, Ordinal measure is less sensitive to block-based modifications such as logo insertions and subtitles. On the other hand, color histograms are highly susceptible to global and local color variations. Hoad et al. (2006) introduced color-shift and centroid based descriptors to identify the duplicate video sequences. Color-shift descriptor is less effective to black and white contents, whereas centroid-based descriptor gives poor performance for pixel luminance degradations.

Kim et al. (2008) introduced a group based copy detection scheme using video linkage, which transforms a video into a group of frames. Xu et al. (2009) and Uchida et al. (2012) presented CBCD approaches based on DCT coefficients for detecting duplicate video contents, which are less effective against image cropping attacks. Cui et al. (2010) presented a copy detection method based on the Slice Entropy Scattergraph (SES), that utilizes spatio-temporal video slices to identify the video copies. Zhang and Zou (2010) proposed a CBCD technique, which operates directly in the DCT domain and obtains edge information in video frames in order to detect video copies.

Xu et al. (2012) introduced Eudemon system, for online video frame copy detection, which uses Earth Movers Distance (EMD) similarity measure; yet it suffers due to complex EMD computations. Gupta et al. (2012) introduced a copy detection system using nearest neighbor mapping based on sliding window mechanism. However, this method scores poor results for the sequence of temporally invariant frames.

Recently, Jiang et al. (2013) proposed a copy detection framework, which employs a frame-level descriptor computed from relative mean intensity of frames. This method is robust against video modifications such as rotation and flipping; yet, it is less effective towards region-oriented transformations. Though global descriptors are compact and effective for frame-level transformations; yet they are less robust against region-based attacks such as letter-box insertions, picture-in-picture and cropping.

Local features based CBCD schemes

Due to the limited capability of global descriptors, many researchers introduced local descriptors, which compute a set of interest points to facilitate local matching. For instance, SIFT (Lowe 2004) and SURF (Bay et al. 2008) descriptors are widely popular in the CBCD domain, since they are more robust than global descriptors in handling several video manipulations (Law-To et al. 2007). Visual words or Bag-of-words model is introduced by Poullot et al. (2008) and Ren et al. (2009) in order to detect illegal video contents. Selecting perfect vocabulary size is a critical issue in bag-of-words model, which may result in poor retrieval rates. Natsev et al. (2010) designed a video copy detection system, that uses SIFT and color correlogram descriptors for detecting duplicate videos.

Roth et al. (2010) and Zhang et al. (2010) introduced CBCD systems by employing SURF descriptors, which require multidimensional index structures for indexing the video signatures. Recently, Ren et al. (2012) introduced a pirate video detection technique, which employs global as well as local detectors as signatures; however, it performs poor for complicated attacks such as camcording. Although, local descriptors are more robust than global descriptors; still, they are computationally expensive and require multidimensional indexing structures.

Notable CBCD methods using visual features

There are notable works in the literature on CBCD problem, which utilize only visual signatures for detecting illegal videos. For example, Kim and Vasudev (2005) proposed a spatio-temporal sequence matching scheme for detecting video copies, which performs spatial matching using ordinal signatures and temporal matching using temporal signatures extracted from the image frames. Joly et al. (2007) introduced a copy detection technique by employing Harris detector (Schmid and Mohr 1997) and proposed a new probabilistic similarity search technique to retrieve video copies. Though this method provides very good performance in terms of robustness and discrimination; yet it suffers due to its computational complexity. Further from database point

of view, the redundancy of local features proves problematic for searching speed. On the other hand, the authors addressed only a few kinds of distortions such as resize, shift, contrast, gamma correction and noise addition; while this technique fails to handle modifications such as pattern insertion, moving captions and flipping.

Chiu et al. (2008) proposed a probabilistic framework for identifying duplicate videos, which transforms the copy detection task into a shortest-path problem and computes matching pairs of frames between the video sequences. This method uses compact visual signatures; that is, only 9D ordinal signature is generated for a block of size 3×3. However, the ordinal signature they use, limits the resistance to region-based attacks such as zooming, cropping and pattern insertions. Furthermore, the authors experimented on a limited set of distortions such as brightness enhancement, speed change and resize, while transformations such as noise, pattern insertion, zoom in and color change are not tackled in their study. Douza et al. (2010) introduced an image-based approach, which employs local feature indexing method and a spatio-temporal post filtering step to identify the video copies.

Küçüktunç et al. (2010) presented a copy detection framework by employing MPEG-7 descriptors, facial shot mapping and activity subsequence matching techniques, which is less effective against complex transformations such as camcording and picture-in-picture. Sarkar et al. (2010) designed a duplicate video detection framework using color layout descriptors and proposed a non-metric distance measure to search efficiently in the high-dimensional space. however, this scheme may score inaccurate results, if the query video contains portions of multiple master video sequences. Chiu and Wang (2010) introduced a time-series linear search (TLS) method for detecting illegal videos, which combines compact video signatures derived from min hash theory and efficient fingerprint generation process based on heap operations.

As mentioned in Section 1.5, motion features are considered as poor descriptors in the CBCD literature (Hampapur et al. 2001), since they are close to zero values, when they are captured at normal frame rates (25-30 fps). However, Tasdemir and Cetin (2010) attempted to employ motion features for their CBCD task, by capturing frames at a lower rate (i.e. 5fps) and using motion vector magnitudes. Though, the proposed motion vector-based signatures are resistant to illumination and color changes; still, they fail to describe the spatio-temporal motion activity of a video sequence.

Wei et al. (2011) presented a frame fusion based copy detection scheme, which exploits Viterbi-like dynamic programming algorithm with on-line backtracking strategy for detecting copied videos. Esmaeili et al. (2011) proposed a CBCD system, which utilizes fingerprints extracted from special images of videos and introduced a fast approximate search algorithm to identify the duplicate video contents. Recently,

Lei et al. (2012) introduced a video sequence matching method based on the invariance of color correlation of RGB components. This method achieves satisfactory performance in terms of both time and space complexity. However, it fails to provide better detection rates for transformations such as camcording, color phase modification and color adjustment; since, they considerably alter the color components of the video sequences.

CBCD frameworks using audio and multimodal signatures

Though audio content is an essential information source of a video sequence; yet, only very few attempts are made to detect video copies using acoustic signatures. For instance, Itoh et al. (2010) utilized acoustic power features for their copy detection task. Although this method is efficient; still, its high computational cost may degrade the performance of the system. Anguera et al. (2009) designed a multimodal video copy detection scheme by fusing visual fingerprints extracted from luminance variations and audio signatures derived from spectral coefficients. However, this method fails to provide better accuracy for region-based attacks such as logo insertions and subtitles.

Saracoğlu et al. (2009) presented a framework by employing coarse visual-audio fingerprints for identifying video copies. Although the performance of this method is reasonably good, still it is less effective against video attacks such as picture-in picture, due to the limitations of the proposed global descriptors. Moreover, the authors combined the audio-visual fingerprints in a simplified manner based on decision scores, which may not guarantee a stable performance for all the cases. Tian et al. (2011) presented a video copy detection framework, which exploits visual-audio features and sequential pyramid matching technique for detecting duplicate videos. However, this method uses simple fusion strategy, which may not provide optimum performance for complicated video transformations such as camcording and combined attacks. Recently, Wu and Zhao (2012) presented a multimodal copy detection approach, which integrates DCT, SIFT and audio features for detecting video copies. However, the computational cost of this scheme is slightly high, since it involves complex computations.

2.1.3 Research challenges of the CBCD Domain

Numerous state-of-the-art CBCD schemes are exploiting only visual signatures of videos for detecting video copies. Instead, if the video content is utilized in an intuitive and natural way, then the performance of a copy detection system can be significantly improved (Roopalakshmi and Reddy 2010). By keeping the above factors in mind,

this thesis highlights some of the research challenges in the context of CBCD domain:

- ⋆ **Utilization of Audio Fingerprints**
  Audio content is an indispensable information source of a video sequence. However, most studies on CBCD concentrate only on visual signatures, while very few efforts are made to exploit audio features. From piracy perspective, in most of the copy detection cases, the audio data is less manipulated compared to its counterpart (Saracoğlu et al. 2009). Therefore, the combined utilization of visual-audio fingerprints for the copy detection task, not only enhances the detection performance, but also extends the coverage to more number of video editing and transformations.
- ⋆ **Fingerprints using MPEG-7 Descriptors**
  A major challenge in most of the copy detection systems is, the computational cost of fingerprint extraction and similarity matching tasks, since huge amount of databases need to be verified. Though, feature extraction can be done effectively and efficiently using compressed domain features such as MPEG-7 descriptors; still, it is not much explored in the CBCD domain.
- ⋆ **Complicated Image Transformations**
  In CBCD paradigm, for video transformations such as rotation, scaling, zooming and brightness change, numerous solutions are proposed. However, complicated video manipulations such as camcorder captures, picture-inside-picture, gamma-corrections and combined visual-audio attacks pose specific challenges in the context of CBCD domain.
- ⋆ **Incorporating Visual Cues**
  Most studies in CBCD, concentrate on low level features which describe the image content; yet, there is a semantic gap between the low level image features and user's high level representation of images. In order to handle this disparity, if visual cues are incorporated in the copy detection task, then the detection performance can be significantly enhanced.
- ⋆ **Combining Content and Context**
  In video copy detection, when a query video is presented, huge database of videos need to be compared, to confirm whether it is copied or not. Hence, if the contextual information is also incorporated along with the video content for detecting the video copies, then the detection accuracy can be considerably improved.
- ⋆ **Using Compact Video Fingerprints**
  In CBCD systems, multi-dimensional master videos need to be checked to detect

duplicate videos; hence, compact representation of a video sequence using efficient fingerprints is a major challenge, which considerably affects the detection performance of the CBCD system.

- ⋆ **Employing Effective Indexing Techniques**
  A main challenging issue in most of the CBCD systems is, the computational cost of indexing task, since huge database of videos need to processed. Therefore, effective indexing techniques could be employed for enhancing the detection performance of the given CBCD system.
- ⋆ **Exploiting Efficient Similarity Matching Methods**
  If efficient similarity matching techniques are exploited, then the computational cost of the CBCD framework could be reduced to a greater extent, which in turn may considerably improve the effectiveness of the CBCD framework. From another perspective, if the pirate video is derived from multiple master video sequences, then effective similarity matching algorithms are essential to ensure the best matching master video.
- ⋆ **Using Appropriate Fusion Schemes**
  It is known that, exploiting both the visual-audio features of a video sequence for the CBCD problem, not only enhances the detection accuracy, but also widens the coverage to more number of video modifications. For this reason, suitable fusion schemes are required for combining the audio-visual fingerprints in an effective manner.

This research study attempts to solve some of these issues, by proposing various copy detection schemes, which employ video signatures such as color, audio, motion activity and multimodal features. The contributions of this thesis towards the CBCD problem are illustrated in Chapter 3.

2.2 Video Copy Tracking/Registration

2.2.1 Basics of pirate video registration

Fighting against video piracy needs duplicate detection as the first step, which aims to find out the best matching master video for the given pirate/query clip. As mentioned in Section 2.1.1., content-based video copy detection (CBCD) techniques use content-based features of the media to detect duplicate video sequences; thus, they are widely popular. However, existing CBCD schemes do not deal with the accurate frame-to-frame alignments of a pirate video with the master sequence, since their ultimate goal

is to detect the duplicate video clips by comparing the perceptual similarity between the two video sequences.

On the other hand, in case of illegal camcorder captures, the captured images are mostly distorted, since the capturing axis is not exactly perpendicular to the theater screen. Due to this reason, significant frame misalignments exist between the pirate video and the master sequence, where the misalignments could be temporal, geometric or combination of both. Therefore, followed by copy detection, accurate frame alignments of the pirate video with the master content is very much essential, for a number of applications such as estimation of geometric distortions, detection of forensic watermarks and approximation of the pirate location in a movie theater.

2.2.2 State-of-the-art schemes and their shortcomings

The research of pirate video registration is a brand new and the early research concentrates on visual features for perfectly registering the pirate video frames with the master sequence. For instance, Delannay et al. (2003) presented a temporal registration technique using key frames to compute frame alignments of watermarked documents, in which frame rates are assumed constant. However, in case of high motion activity, this scheme extracts different sets of key frames from the pirate and master video sequences.

Cheng (2003) proposed an algorithm for temporally matching two video contents using dynamic programming. Though, this method scores good registration results; yet it is severely altered by distortions such as noise addition. Cheng and Isnardi (2003) developed a spatial, temporal and histogram registration scheme by including contextual costs, which can be applied to detect forensic watermark information. In 2004, Cheng reviewed and compared three different video registration algorithms, which detect forensic watermarks in digital cinema applications.

Chupeau et al. (2006) exploited color histograms to map two video sequences using dynamic programming. This scheme achieves poor results for region-based transformations, because of the global descriptive nature of color histograms. In 2007, Chupeau et al. presented a registration scheme for estimating the distortion model and compensating geometric distortions in video copies. This method attempts to match the pirate video frames with the master sequence as a prerequisite for recovering the embedded watermarks. Chen et al. (2008) utilized temporal ordinal measurements for matching two video contents. Although this method achieves precise temporal localization; still, it is much affected by picture-in-picture and cropping transformations. Baudry et al. (2009) employed both the local and global descriptors

for aligning two video sequences. However, this method provides poor registration results for low motion frames and complex transformations such as letter-box insertions and subtitles.

Lee et al. (2009) presented a video frame matching scheme using dynamic programming, which considerably decreases the probability of matching errors by incorporating an effective matching cost function. However, this frame matching scheme addresses only a few types of video attacks such as frame insertions, shuffle, removal and compression attacks. Recently, Baudry et al. (2010) designed a temporal registration technique for video copies, in which wavelet coefficients are hierarchically encoded in order to get video fingerprints. Though, this registration technique guarantees accurate alignments, the encoding of wavelet coefficients is expensive in terms of CPU and memory.

To summarize, existing pirate video registration techniques employ only visual features of videos for achieving accurate frame-to-frame alignments of the pirate and master video sequences (Delannay et al. 2003; Cheng 2003; Cheng and Isnardi 2003; Chupeau et al. 2006; Chupeau et al. 2007; Chen and Stentiford 2008; Baudry et al. 2009; Lee et al. 2009; Baudry et al. 2010). Although, audio content constitutes an important information source of a video; yet, no attempts are made to exploit acoustic features for obtaining the accurate frame alignments of the master and duplicate video contents. Further, from the video-piracy point of view, in most of the illegal camcorder captures, the audio content is less affected compared to its counterpart (i.e. visual data)(Saracoğlu et al. 2009). Due to the above reasons, if the visual and audio fingerprints are jointly exploited for the registration task, then the accuracy could be considerably enhanced. Therefore, *promising frameworks for pirate video registration, which exploit visual and acoustic features in a unified framework are needed.*

From another perspective, most of the current video copy registration schemes are concentrating on the alignment of watermarked documents (Delannay et al. 2003; Cheng 2003; Cheng and Isnardi 2003; Chupeau et al. 2006); while very few efforts are made to address the alignment of non-watermarked video sequences. However, it is to be noted that, all master/copyrighted contents are not watermarked (Lefèbvre et al. 2009). Therefore, *novel video copy registration frameworks, which could compute accurate frame-to-frame alignments of the pirate video with the master sequence, irrespective of presence/absence of forensic watermarks are required.*

This thesis attempts to address the shortcomings of the existing registration schemes, by introducing different spatial as well as temporal alignment frameworks, which exploit content-based multimodal features. More precisely, the proposed frameworks, locate a given pirate clip within the master content and compute the accurate

frame alignments of two video sequences, even in the absence of forensic watermarks. Thus, the scholarly contributions of this thesis towards the pirate video registration problem are described in Chapter 4.

2.3 Geometric Distortions Estimation

2.3.1 Estimating geometric distortions in video copies

Due to the exponential growth of multimedia technologies and online streaming activities, numerous illegal videos are proliferating on the Internet and causing digital video piracy. Therefore, rigorous countermeasures and forensic activities are needed to control Internet piracy as well as camcorder piracy. Further, it is observed that, in most of the illegal camcorder captures, the capturing axis is not perpendicular to the screen, which results in distorted images. Because of this reason, significant frame misalignments exist between the pirate and master video sequences, where the misalignments might be temporal, geometric or combination of both. Therefore, estimation of geometric distortion model between the pirate and master video contents is a prerequisite step, for approximating the illegal capture location in a movie theater.

2.3.2 Geometric distortions and 2-D homography

As mentioned above, in illegal camcorder captures, the captured images are coupled with severe geometric distortions. The resultant geometric distortions can be well described by perspective projection, which models the imaging process of a pinhole camera. Strictly speaking, a projective transformation or a homography is an invertible mapping of points and lines on the projective plane $\mathbb{P}^2$, which is defined as follows (Hartley and Ziserman 2004):
A mapping $h : \mathbb{P}^2 \longrightarrow \mathbb{P}^2$ *is a projectivity if and only if there exists a non-singular* 3×3 *matrix* $\mathbf{H}$ *such that for any point in* $\mathbb{P}^2$ *represented by a vector* x*, it is true that* $h(x) = Hx$.

Precisely, the projective transformation deals with 2D to 2D transformations. Perspective projection indicates the projection of 3D points in space to 2D points in the image plane. More specifically, perspective projection occurs, when a camera takes the image of world and displays the result on the its image plane. A planar projective transformation is a linear transformation on homogeneous 3-vectors represented by a

non-singular 3×3 matrix as (Hartley and Ziserman 2004),

$$x'_1 = \mathbf{H}x_1, \; where \; \mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{2.1}$$

Here **H** is a homogeneous matrix; hence, only the ratio of matrix elements is significant in the homogeneous form of representation. There are eight independent ratios amongst the nine elements of **H**; thus, a projective transformation has eight degrees of freedom (DoF). In this way, four 2D to 2D point correspondences derived from the source and pirate video frames are required in order to estimate the 8-parameter homographic matrix **H**, which represents the distortion model between the two videos.

2.3.3 Related work and research challenges

In the literature, there exist only a very few papers, which focus on estimation of geometric distortions in illegal video contents. For example, Delannay et al. (2001) presented a framework for estimating and compensating the geometric distortions in pirated contents, which occur due to handy cam attacks. The authors employed displacement vectors to estimate the distortions and utilized 12-parameters bilinear transformation model for compensating the distortions. Though, this method is useful in digital cinema applications; yet, it is specifically designed to estimate the distortions in watermarked documents. Furthermore, the performance of this method is more sensitive to the underlying watermark embedding algorithm.

Chupeau et al. (2007) introduced a technique for estimating and compensating the geometric distortions in video copies. In this scheme, first the master and duplicate video contents are temporally mapped by utilizing a visual descriptor based on luminance values. After this step, from the temporally registered frames, the projective matrix is estimated and distortion compensations are performed in order to recover the embedded watermark information.

To summarize, existing works on estimation of geometric distortions in video copies are employing only visual signatures of the video sequences (Delannay et al. 2001; Chupeau et al. 2007). However, as described earlier in Section 2.2.2, audio content is a powerful source of any video, which remains less affected compared to visual data in illegal captures. In addition, if audio data is available, then the combined utilization of visual-acoustic fingerprints for estimating distortions, would possibly enhance the estimation accuracy. Therefore, *promising approaches utilizing multimodal fingerprints are required for estimating geometric distortions in video copies.*

From another perspective, current distortion estimation approaches are concentrating towards estimating the geometric distortions in watermarked video contents (Delannay et al. 2001; Chupeau et al. 2007). However, it is well known that, all copyrighted/master contents are not watermarked (Lefèbvre et al. 2009). Therefore, *robust frameworks employing visual-audio features for the estimation of geometric distortions in illegal videos are needed, which are useful for both the watermarked as well as non-watermarked video contents.* This thesis attempts to solve these discrepancies, by contributing a novel distortion estimation framework, which utilizes content-based multimodal features. Precisely, the scholarly contribution of this thesis towards the geometric distortion estimation problem is illustrated in Chapter 5.

2.4 Pirate Position Estimation

2.4.1 Estimating the position of the pirate in a theater

As described in Section 1.1, the exponential growth of on-line publishing activities are increasing the proliferation of illegal video contents on the Internet at an impressive rate, which leads to video piracy. Further, 90% of the pirated versions of movies are created by illegal camcorder captures in theaters. Hence, camcorder piracy is emerging as a serious issue for the entertainment industry from the past few years, which needs to be solved.

Recently, digital cinema system is introduced to uniformly project and distribute motion pictures and also to protect digital cinema. Strictly speaking, Digital Cinema Initiatives (DCI)[1] is the entity, that is created to establish the technical specifications and requirements for mastering, distributing as well as theatrical playback of digital cinema content. DCI defines a forensic watermarking system for copyright protection and also specifies that, the payload of the forensic watermark should contain time stamp and the theater information of movie playback. Thus, the forensic watermark helps to detect and warn the designated theater against camcorder piracy. However, as per the requirements for protecting digital cinema, detecting the theater and time stamp information is not sufficient, it is necessary to identify the pirate so that the number of piracy suspects is restricted.

On the other hand, as explained in Section 1.4, followed by video copy detection, tracking and distortions estimation, the resultant (i.e. precomputed) video fingerprints could also be exploited for estimating the position of the movie pirate in a

[1] *Digital Cinema System Specification Version 1.2. Digital Cinema Initiatives, LLC, 2008.* [Online]. Available: `http://www.dcimovies.com/DCIDigitalCinemaSystemSpecv1_2.pdf`.

theater. More precisely, in order to emphasize the capability of video fingerprints towards pirate location estimation problem, In-theater experiments are evaluated as a Case study in this research work. Therefore, *addressing movie piracy by means of identifying the exact pirate and bringing him/her to justice, surely falls beyond the scope of this thesis.* Further, it is certainly possible to estimate the pirate location in a movie theater, by performing exhaustive investigation of the theater projective geometry and the geometric distortion model between the two video sequences. *In order to prove this viewpoint, a Case-study is conducted in this research work towards the pirate position estimation problem*, with the following assumptions:

- Since the ability of precomputed video fingerprints towards the pirate location estimation problem is to be validated, the position approximation is implemented by utilizing content-based feature descriptors/key points; though, it might be slightly complicated.
- Although position estimation problem is extremely well studied in computer vision literature; yet, this thesis focuses specifically on In-theater location approximations with the help of Perspective projection.

Based on these assumptions, only relevant existing literature is discussed in the subsequent section as given below.

2.4.2 Existing frameworks and their limitations

The research of pirate position estimation is quite challenging, hence only very few attempts are made to approximate the illegal capture location in a movie theater. Chupeau et al. (2008) presented a forensic tracking framework using visual signatures to determine the camcorder viewing axis and derived the approximate position of the pirate in a theater. This framework estimates the capture location using the control points derived from the temporally aligned source and pirate video frames. However, this scheme employs only visual signatures for aligning the two video contents and estimating the homographic distortion model.

Nakashima et al. (2009) proposed a position estimation system, which embeds audio watermarking signal into the movie soundtrack and utilizes detection strength for deriving the position of the pirate in a theater. Still, the estimation accuracy of their system mainly depends on the interior construction of the theater such as location or number of loudspeakers and microphones. Further, the performance of their approach may severely affected by the environmental factors such as background noise and frequency response of the auditorium. Furthermore, their position estima-

tion framework is less robust against attacks such as pitch shifting, lossy compression and collusion, which may decrease the detection strength of the system.

Recently, Lee et al. (2010) presented a framework for estimating the position of the pirate using a video watermarking scheme based on local-auto correlation function (LACF). The authors employed corner points of the video frames and estimated the projective transformation using LACF, from which the position of the pirate is estimated. However, the watermark embedding process of this framework needs to be done in real-time and the embedded watermark must survive to camcorder piracy.

To summarize, existing pirate position estimation frameworks are employing only watermarking techniques for approximating the location of pirate in a movie theater (Nakashima et al. 2009; Lee et al. 2010). However, watermarking techniques suffer due to these drawbacks:

a) Fragile in nature.
b) Insertion/extraction of watermarks involve complicated procedures.
c) Embedding and decoding of watermarks may damage the video content.

In addition, current position estimation schemes are using only visual features of videos for approximating the illegal capture location in a theater (Chupeau et al. 2008). However, if both the visual-acoustic fingerprints are jointly employed, then the position estimation accuracy might be enhanced. Therefore, *promising forensic tracking schemes utilizing content-based multimodal features are needed, which can estimate the pirate position in a movie theater.*

This thesis attempts to solve the issues of existing position estimation techniques, by presenting a forensic tracking scheme, which employs content-based multimodal features. The scholarly contribution of this thesis towards the pirate position estimation problem is discussed in Chapter 6.

2.5 Outcome of Literature Survey

Generally, a video comprises a collection of multimodal features such as visual, audio, motion activity and textual information. In CBCD literature, numerous state-of-the-art techniques are primarily focusing on the visual features of the video contents. However, exploiting audio fingerprints for detecting video copies is essential due to the following reasons:

(a) The audio content constitutes an indispensable information source of a video.
(b) In case of illegal camcorder captures, the audio content is less altered compared to visual data.

Therefore, promising methods employing visual and acoustic features for detecting video copies are required, which can significantly enhance the detection accuracy of a CBCD system.

Most studies on pirate video registration and distortion model estimation utilize only visual features of the video sequences, while not much efforts are made to employ acoustic features. Further, existing schemes are focusing on the alignment of watermarked documents, while limited efforts are made towards non-watermarked video contents, even though all copyrighted contents are not watermarked. Therefore, robust frameworks employing visual-acoustic fingerprints are needed for aligning the pirate video with the master content and estimating geometric distortions in video copies, which can be used irrespective of presence/absence of forensic watermarks.

State-of-the-art forensic tracking frameworks employ watermarking techniques to estimate the position of the pirate in a theater. However, watermarks are fragile in nature and may destroy the video content. Therefore, promising forensic tracking frameworks utilizing content-based multimodal features are required, which can estimate the location of the pirate in a movie theater.

2.6 Problem Statement

Based on the outcome of the literature survey, this research study attempts to solve the mentioned issues, by contributing novel and efficient techniques. Precisely, the problem statement of this research study is defined as follows, *To develop effective and efficient frameworks for detecting, tracking video copies and also for estimating the geometric distortions as well as the illegal capture location in a theater, and thereby restrict video piracy.* More specifically, the objectives of the current research study are illustrated as follows.

2.7 Research Objectives

(a) To develop efficient CBCD methods, by employing compact video fingerprints or feature descriptors for detecting the duplicate video sequences.

(b) To design new spatio-temporal video copy registration frameworks in order to guarantee the accurate frame alignments of the copied clip with the master video sequence.

(c) To present robust distortion estimation techniques, in order to compute the geometric distortions present in the illegal video.

(d) To propose novel pirate position estimation frameworks, so as to approximate the illegal capture location in a theater, by performing in-depth analysis of theater projective geometry and distortion model of the duplicate video.

In this way, designing efficient methods based on (a)-(d) objectives, to restrict digital video piracy as well as camcorder piracy is the primary goal of this thesis. To accomplish these objectives, the research framework comprising four different methodologies is implemented, as specified in Section 1.4. More precisely, various modules of the current research framework, followed by the contributions with respect to each of the modules are clearly illustrated in the subsequent chapters.

2.8 Experimental Datasets

Initially, a set of video sequences collected from Open Video Project[2] are employed in this research study, to evaluate the proposed copy detection methods. Since, Open Video Project datasets are readily available and also popularly utilized in the copy detection problem domain (Chiu et al. 2010; Chiu et al. 2008), they are exploited for implementing the CBCD task.

Followed by the satisfactory performance of the proposed CBCD techniques on Open Video Project datasets, TRECVID[3] datasets are also utilized in this research study, because of the widespread usage of the latter datasets. Specifically, TRECVID 2007, 2008, 2009 Sound and Vision datasets are obtained and utilized for the copy detection as well as registration experiments. Though TRECVID Sound and Vision datasets are extensively popular in the CBCD literature (Chiu and Wang 2010; Küçüktunç et al. 2010; Wei et al. 2011), only subsets of TRECVID database videos are employed for duplicate video detection and registration experiments, due to the following reasons:

- TRECVID database size is huge; hence, exploiting the complete database (all 2007, 2008 and 2009 datasets) is quite challenging in terms of size constraints.
- Further, using a subgroup of TRECVID database videos, that is, either 2007 or 2008 or 2009 datasets, is common in the CBCD literature. For instance, Wei et al. (2011) employed 200 hours of TRECVID 2008 dataset for detecting illegal video clips.

[2] `http://www.open-video.org`

[3] TRECVID 2010 Guidelines [Online]. Available: `http://www.nlpir.nist.gov/projects/tv2010/tv2010.html`.

- Furthermore, combining two or more category of TRECVID database videos for the copy detection task, is also popular in the CBCD domain. For example, Küçüktunç et al. (2010) utilized 100 hours of TRECVID 2007 sound & vision data, plus another 100 hours of TRECVID 2008 sound & vision data for implementing their video copy detection task.

Based on these factors, this thesis utilizes different groups of TRECVID database videos for implementing the proposed video copy detection and tracking frameworks.

Further, this research work also employs popular CC_WEB_VIDEO datasets for video copy registration experiments. Specifically, CC_WEB_VIDEO collection includes 24 top favorite videos collected from YouTube, Google Video and Yahoo! Video followed by the duplicate and near-duplicate video clips of the corresponding master video sequences.

Furthermore, in order to assess the performance of the proposed frameworks against camcorder captured videos, the present research work also utilizes a set of real data comprising camcorded copies of master video sequences. The experimental setup of each of the proposed frameworks in terms of dataset generation is clearly illustrated in the appropriate sections of the subsequent chapters.

2.9 Summary

This chapter describes the problem domains such as content-based video copy detection (CBCD), video copy registration, distortion computation and pirate position estimation. Further, this chapter also illustrates the existing literature and their corresponding limitations with respect to these four problem areas. Specifically, this chapter also summarizes the some of the prominent state-of-the art techniques of CBCD and pirate video registration domains along with their drawbacks in Tables 2.1-2.3 as follows:

Table 2.1: State-of-the-art CBCD techniques

Author/Authors	**Method description**	**Limitations**
Hua et al. 2004; Chiu et al. 2008	Uses Ordinal Measure using average intensities of blocks	Less sensitive to block-based modifications
Hoad and Zobel 2006	Color-shift and Centroid-based descriptors are used	Less effective to B/W contents and pixel luminance degradations
Naphade and Yeo 2000; Liu et al. 2007; Chiu et al. 2010	YUV/RGB/HSV color space and histogram distributions are exploited	Highly susceptible to global and local color variations
Xu et al. 2009 Uchida et al. 2012	Uses DCT coefficients in the compressed domain	Less effective against image cropping attacks
Gupta et al. 2012	Uses nearest neighbor mapping based on sliding window	Scores poor for temporally invariant frames
Roth et al. 2010; Zhang et al. 2010	Employ SURF (Speeded UP Robust Features)-based fingerprints of frames	Require multi-dimensional index structures for indexing
Ren et al. 2012	Global descriptors and local features of frames is used	Scores poor results for complicated attacks such as Camcording
Kim and Vasudev 2005	Spatial matching using ordinal signatures and temporal matching using temporal signatures of frames are employed	Achieves low PR rates for combined and region-based video transformations

Table 2.2: State-of-the-art CBCD techniques (contd..)

Joly et al. 2007	Uses Harris detector and new probabilistic similarity search technique	High computational complexity
Chiu et al. 2008	Probabilistic framework for transforming the CBCD task into a shortest-path problem	Low resistance to region-based attacks such as cropping
Küçüktunç et al. 2010	MPEG-7 descriptors, facial shot mapping and activity subsequence matching technique are utilized	Less effective against complex video transformations
Sarkar et al. 2010	Color layout descriptors and non-metric distance measure are employed	Performs poor for the query clip, containing parts of multiple master sequences
Tasdemir and Cetin 2010	Motion vector magnitudes of frames are exploited	Fails to describe spatio-temporal motion activity of video sequences
Saracoğlu et al. 2009	Utilizes coarse visual-audio fingerprints	Fails for transformations such as camcording, color phase modification and color adjustment
Lei et al. 2012	Invariance of color correlation of RGB components of frames are employed	Less effective against region-based attacks such as cropping
Jiang et al. 2013	Frame level descriptor using relative intensity of frames is used	Scores poor results for region-oriented transformations

Table 2.3: State-of-the-art video copy registration methods

Author/Authors	**Method description**	**Limitations**
Delannay et al. 2003	Temporal registration technique using key frames for watermarked documents	Extracts different sets of key frames in case high motion activity
Cheng and Isnardi 2003	Spatial, temporal and histogram registration scheme including contextual costs	Severely altered by distortions such as noise addition
Chupeau et al. 2006	Uses color histograms and mapped sequences using dynamic programming	Achieves poor results for region-based transformations
Chen and Stentiford 2008	Uses temporal ordinal measurements for mapping sequences	Much affected by picture-in-picture & cropping transformations
Baudry et al. 2009	Employs both the local and global descriptors for matching	Provides poor results for low motion frames & complex transformations
Lee et al. 2009	Video frame matching scheme using dynamic programming and effective matching cost function	Deals only a few types of video attacks & such as frame insertions shuffle and removal
Baudry et al. 2010	Wavelet coefficients are hierarchically encoded in to get signatures matching cost function	Encoding of wavelet coefficients is expensive & in terms of CPU and memory

Chapter 3

Content-Based Video Copy Detection (CBCD) Schemes

This thesis elaborates the scholarly contributions towards the Content-Based video Copy Detection problem in this chapter. More precisely, this chapter attempts to solve some of the challenges of CBCD domain as mentioned in Section 2.1.3, by contributing different pirate video detection methods. The proposed CBCD techniques employ video fingerprints derived from features such as audio, visual and motion activity, in order to detect illegal video sequences, which are illustrated in the subsequent sections of this chapter.

3.1 Copy Detection Using Color Features

Color is one of the principal visual features of an image; hence, this thesis presents two CBCD techniques, that employ Dominant Color Descriptor (DCD) of MPEG-7 standard (Manjunath et al. 2002), to facilitate the detection of duplicate videos, as detailed below.

3.1.1 CBCD scheme based on dominant color features

Dominant Color Descriptor (DCD)[1]of MPEG-7 standard effectively describes the color information in an image, by capturing the dominant or representative colors of that image. Specifically, DCD compactly represents the color distribution present in an image, with a small number of dominant colors and their relative distribution.

[1]ISO/IEC/JTC1/SC29/WG11 MPEG 2001/N4358, Text of ISO/IEC 15938-3/FDIS Information technology - *Multimedia content description interface - Part 3 Visual*, Sydney, Australia, July 2001

More precisely, MPEG-7 standard defines DCD as,

$$DCD = \{\{c_i, p_i\}, v_i, s\},\ i \in [1 : m] \tag{3.1}$$

Here, c_i represents 3D color vector, while p_i indicates the percentage of distribution of each dominant color so that $\sum p_i = 1$ and m represents the total dominant colors of an image (Deng et al. 2001). Two optional fields, spatial coherency s and color variance v_i, precisely characterize the color distribution in spatial and color space domains respectively (Manjunath et al. 2002). Specifically, spatial coherency describes the spatial distribution of pixels associated with each representative color. Color variance explains the variation of color values of the pixels in the surroundings of a corresponding representative color.

In the past literature, most of the works utilize color clustering algorithms such as Generalized Lloyd Algorithm (GLA)(Lloyd 1982) to extract DCDs from an image (Yang et al. 2008; Deng et al. 2001). However, GLA suffers due to the following limitations: (a) It is computationally intensive; (b) Its efficiency mainly depends upon the initial specifications such as number of clusters, distance and centroid. From another perspective, a major demanding issue in CBCD systems is, the computational cost of fingerprint extraction and similarity matching tasks, because huge video databases need to be processed. In order to address these issues, *new copy detection frameworks using compact video fingerprints need to be explored, which result in less computational cost*. Based on this aspect, this thesis first proposes a novel CBCD technique, which utilizes compact fingerprints derived from DCDs for the copy detection task with less computational cost. More precisely, the main contributions of the proposed video copy detection method are given by,

- A novel DCDs extraction technique, which is simple and compact, compared to the existing techniques.
- An adaptive video signature pruning method, which considerably reduces the total number of visual signatures of the given video.

The framework of the proposed copy detection scheme including video fingerprints extraction and similarity matching is described as follows.

Proposed CBCD framework based on dominant color features

Figure 3.1 describes the framework of the proposed scheme, in which dominant color descriptors derived from the master and query video frames are employed for executing the copy detection task. Precisely, in the proposed CBCD scheme, the frequency

imaging method introduced by Kashiwagi and Oe (2007) is extended, in order to extract DCDs of images in an easy and compact manner.

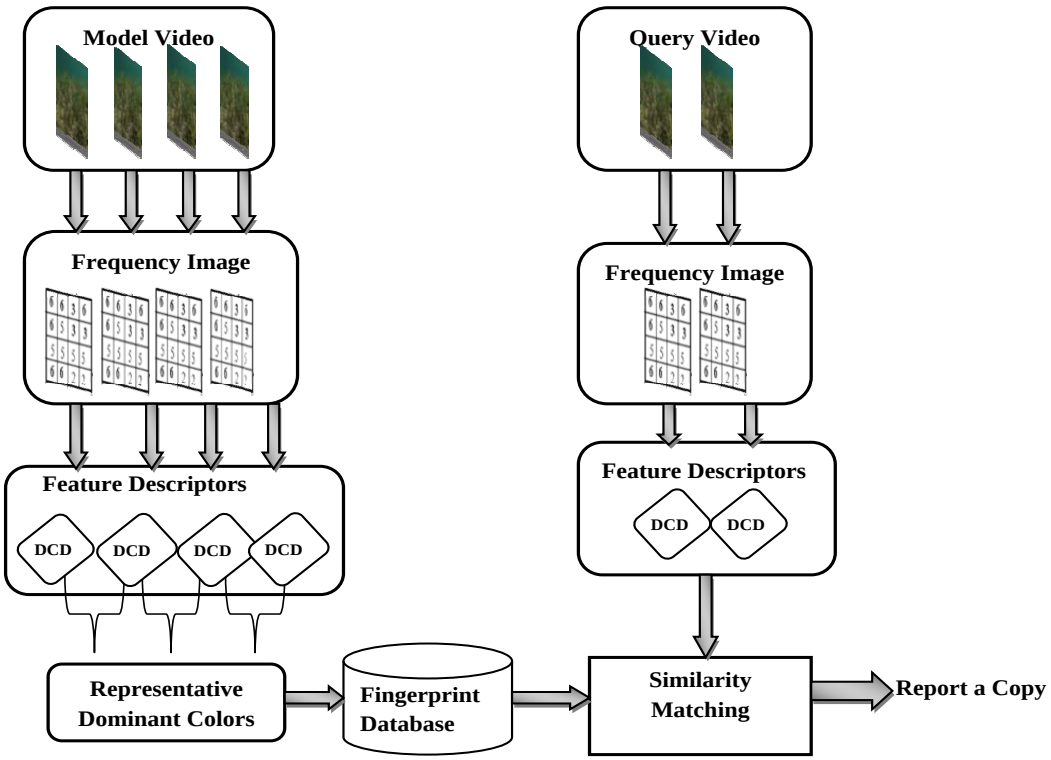


Figure 3.1: Proposed CBCD framework based on dominant color features

Precisely, Frequency Image is a feature image,in which each pixel indicates the frequency of the same color pixels. More precisely, in the proposed CBCD framework, first key frames are obtained from the model video using uniform sampling method. Then, for each key frame, frequency images representing the distribution of same feature pixels is calculated. After this step, an adaptive pruning strategy is applied in order to get the final set of dominant color descriptors of a video sequence and the resultant DCDs of master videos are stored in the fingerprint database. Whenever user presents a query video, frequency images are generated and DCDs are extracted in the similar manner from the query video frames. Then, the fingerprints of the query video are compared with the respective master video feature sequences and consequently, the copy detection results are reported.

Video fingerprints extraction

In the proposed copy detection framework, using uniform sampling at the rate of 10frames/sec (fps), key frames are obtained from the master and the copied video

contents. Without the loss of generality, the proposed scheme employs RGB color space and Look-up tables (LUTs) for computing DCDs. Precisely, R,G and B colors are considered as three features of an image and every pixel is replaced by the frequency value of same feature pixels. The resultant pixel frequencies of an image is collectively called as *Frequency/Feature Image*, which is further processed for calculating the dominant colors and their distribution percentages in the given image.

Though, color histograms are widely popular in the literature, the Frequency Images differ from color histograms, in the following aspects:

- ⋆ Frequency Images are more informative than the color histograms, since they explicitly indicate frequency of every pixel in an image.
- ⋆ For a given color image, multi-dimensional histograms are needed to specify the complete color distribution; whereas 1-dimensional Frequency Images are sufficient to describe the color distribution of that image.

Due to these reasons, Frequency Images are employed in the proposed CBCD framework to extract DCDs of an image.

On the other hand, it is observed that, *Consecutive images in a video sequence may have very similar color statistics* (Roytman and Gotsman 1995). Based on this aspect, a new video signature pruning method is introduced, by exploiting the color similarity existing in the temporal domain, in order to efficiently describe the color content of the given video. Specifically, *the proposed signature pruning method considerably reduces the total number of DCDs of the given video sequence, by exploiting the temporal color statistics.* To validate this statement, two sets of experiments are conducted for extracting DCDs-based fingerprints. In the first method, (called as Baseline method), the DCDs extracted from the Frequency Images are considered as fingerprints of the corresponding video sequences. In the second method (named as Pruning based adaptive method), the DCD of each frame is compared with that of the previous frame, and if the similarity between the DCDs exceeds the threshold, then the latter DCD is considered as a new representative color of the given video sequence. Experiments are conducted for different threshold values ranging from 25-46 and based on the results the threshold value is set as 35.

Table 3.1 shows the details of the extracted feature descriptors, using both the Baseline as well as Pruning based adaptive methods for 1, 3 and 5 minutes videos respectively. Results from Table 3.1 shows that, the Pruning based adaptive extraction method reduces the total number of feature descriptors by 58%, 31% and 29% for 1, 3 and 5 minutes videos respectively. In this way, Pruning based adaptive method substantially decreases the total fingerprints of the video sequence, when compared to

Table 3.1: Comparison of total number of extracted feature descriptors

S.No	Duration (in minutes)	Total no.of feature descriptors		Reduction (in %)
		Baseline method	Pruning based adaptive method	
1	1	247	144	58
2	3	1150	352	31
3	5	1445	407	29

the Baseline technique. Further, as mentioned Section 1.3, a main challenging problem in most of the CBCD systems, is the computational cost of signatures extraction and similarity matching activities. Since, *Pruning based adaptive scheme noticeably reduces the total fingerprints of a given video and thereby also decreases the similarity matching cost.* Due to these reasons, the Pruning based adaptive DCDs extraction method is employed in the proposed framework to implement the copy detection task.

Fingerprint matching

While extracting DCDs of images, it is observed that, single dominant color is necessary for an image. However, each image can be effectively represented using 3 to 5 dominant colors. Since the number of representative colors is less, the feature descriptors are indexed based upon their dominant color values. Therefore, fingerprint matching of the proposed framework includes searching the database for similar color distributions same as the input query, which involves searching for each of the dominant colors separately. If F1 and F2 are two dominant color descriptors such that,

$$F1 = \{\{c_i, p_i\}, i = 1, 2, .., N_1\}, \tag{3.2}$$

$$F2 = \{\{b_j, q_j\}, j = 1, 2, .., N_2\}, \tag{3.3}$$

where c_i, b_j are the dominant color vectors and p_i, q_j are their distribution percentages respectively. Here, N_1 and N_2 represent the total dominant colors. Then the distance $Dist$ between two DCDs (F1 and F2) can be computed as follows (Manjunath et al. 2002),

$$Dist\ (F1, F2) = \sum_{i=1}^{N_1} p_i^2 + \sum_{j=1}^{N_2} q_j^2 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 2a_{i,j}\, p_i\, q_j \tag{3.4}$$

where $a_{i,j}$ is the similarity coefficient between the color vectors c_i and b_j, which is calculated as,

$$a_{i,j} = \begin{cases} 1 - \frac{d_{i,j}}{d_m} & if\ d_{i,j} \leq T_d \\ 0 & if\ d_{i,j} > T_d \end{cases} \tag{3.5}$$

where $d_{i,j}$ is the Euclidean distance between two colors c_i and b_j. The threshold T_d is the maximum distance used to judge whether two color features are similar or not. The distance $d_m = \alpha \times T_d$,where α is set as 1.2 as specified in (Deng et al. 2001). Different T_d values ranging from 20 to 45 are experimented and based on the results, T_d is set as 25 in the proposed copy detection framework.

3.1.2 CBCD using integrated dominant color features

In general, two images may have similar dominant colors, but the spatial distribution of same color pixels in the two images may not be same always. Therefore, if the dominant color features are exploited along with their spatial correlation information, then the performance of the copy detection systems can be improved. Based on this aspect, this thesis contributes an another copy detection technique, by enhancing the previous CBCD scheme described in Section 3.1.1, which integrates the spatial coherency factor along with the dominant color features for the copy detection task. Since spatial coherency uniquely characterizes the color distribution in the spatial domain, the integration of spatial coherency value with the dominant color features improves the CBCD system performance. The proposed CBCD framework including video signatures extraction and matching is illustrated as follows.

Proposed CBCD framework using integrated color features

Figure 3.2 shows the schematic diagram of the proposed duplicate video detection framework, in which first uniform frame sampling technique is used to extract the key frames from the master video contents. Then, Frequency Images are generated as described in Section 3.1.1 and consequently dominant color descriptors of frames are calculated. In addition to dominant colors and their percentage of distribution, spatial coherency values are also exploited to generate the video fingerprints. Then the resultant fingerprints are stored in the fingerprint database of the video sequences. Whenever user uploads a query video, the dominant color descriptors are extracted from the query video frames and the respective video signatures are generated. Finally, the video fingerprints of the query and master video sequences are compared, in order to obtain the copy detection results.

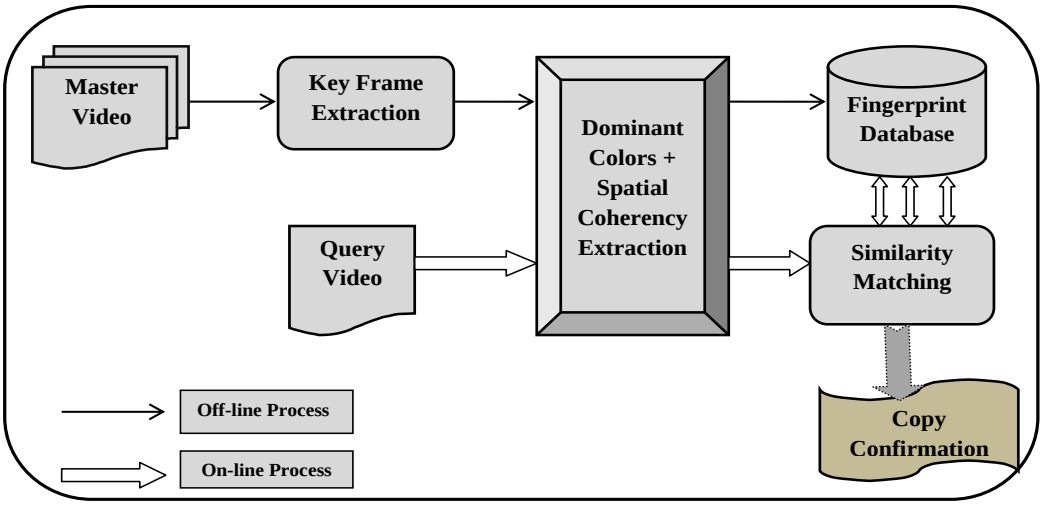


Figure 3.2: Proposed CBCD framework using integrated color features

Video signatures generation

In this method, key frames are extracted using uniform sampling technique with the frame rate of 10fps. Then RGB color space is employed to compute the Frequency Images and consequently dominant colors and their respective distribution values are calculated, as described in Section 3.1.1. The spatial correlation between the dominant colors of an image is computed as follows: First, a video frame is divided into 2×2 blocks; Second, for each block, the maximum distance between two dominant color pixels is computed; Third, the resultant distance is normalized into 1-5 values and considered as spatial coherency factor of the given frame.

Similarity matching

In the proposed copy detection scheme, the video signatures of the master and query video sequences are compared by exploiting the dominant color descriptors as follows. Let $R1$ and $R2$ be two dominant color descriptors such that,

$$R1 = \{\{c_i, p_i\}, i = 1, 2, .., N_1\}, \tag{3.6}$$

$$R2 = \{\{b_j, q_j\}, j = 1, 2, .., N_2\}, \tag{3.7}$$

Then the distance $Dist$ between the dominant color features $R1$ and $R2$ is computed, as specified in Equations (3.4), (3.5). Here, spatial coherency values of dominant colors are also employed for the copy detection task. Therefore, the dissimilarity

Dis_{DC} between the two descriptors $R1$ and $R2$ is computed as follows,

$$Dis_{DC} = w_a \ |s_m - s_q| \ Dist \ (R1, R2) + w_b \ Dist \ (R1, R2) \tag{3.8}$$

where s_m and s_q represent the spatial coherency values of master and query video frames respectively. Here, w_a and w_b are fixed weights, which are set to 0.3 and 0.7 respectively, as specified in (Cieplinski 2001).

3.1.3 Experimental setup

To evaluate the performance of the proposed CBCD techniques (described in Sections 3.1.1 and 3.1.2), two sets of experiments are conducted and the results are indicated in terms of detection accuracy and efficiency. In order to facilitate the discussion of experimental results, hereafter the two proposed CBCD methods are denoted as,
CBCD scheme1: The proposed CBCD method employing only dominant colors and their percentage of distribution values.
CBCD scheme2: The proposed CBCD technique exploiting dominant colors and their distribution percentages along with the spatial coherency values.

As mentioned in Section 2.8, in order to evaluate the proposed CBCD schemes, different video sequences collected from Open Video Project dataset are employed in this research study. More precisely, a video database containing 101 video sequences of Open Video Project dataset is utilized for the copy detection task. The video database contains approximately 305297 frames and the video content includes news, documentaries, education, movies, natural scenes and landscapes. From the video database, 15 videos ranging from 5 to 8 seconds are randomly selected and eight different transformations given by, 1) Blurring, 2) Zooming-in, 3) Zooming-out, 4) Contrast Change, 5) Rotation, 6) Random Noise Addition, 7) Image Ratio and 8) Resolution Change are applied to generate query clips. The resultant 120 (15×8) video copies constitute the query dataset of the proposed copy detection frameworks.

Evaluation metrics

To measure the detection accuracy of the proposed CBCD schemes, standard Precision (P) and Recall (R) metrics are employed, which are given by,

$$Precision = TP/(TP + FP), \tag{3.9}$$

$$Recall = TP/(TP + FN), \tag{3.10}$$

where, True Positives (TP) are positive examples, which are correctly labeled as positives and False Positives (FP) indicate to negative examples incorrectly labeled as positives. False Negatives (FN) represent to positive examples incorrectly labeled as negatives. A detection result is considered as correct, if there is any overlap with the region from which the query is extracted.

Detection accuracy of CBCD scheme1

Table 3.2 shows the detection results of the reference and proposed methods for different video transformations such as blurring, zooming in, zooming out and contrast change. Specifically, in Table 3.2, Recall and Precision rates of the proposed CBCD scheme1 is compared with the corresponding PR rates of Algorithm (1), in which the PR rates fall within the correct intervals. Algorithm (1) is proposed by Cho et al. (2009), in which Ordinal measure is utilized for identifying the duplicate video contents. Table 3.2 results prove that, the proposed CBCD scheme1 achieves better

Table 3.2: PR rates of CBCD scheme1 (at correct intervals)

Transforms	Blurring		Zoom-in		Zoom-out		Contrast	
	R	P	R	P	R	P	R	P
Algorithm (1)	0.2	0.3	0.2	0.45	0.2	0.42	0.2	0.51
	0.4	0.65	0.4	0.5	0.4	0.71	0.4	0.64
	0.6	0.74	0.6	0.62	0.6	0.78	0.6	0.71
	0.8	0.78	0.8	0.65	0.8	0.86	0.8	0.65
	1.0	0.85	1.0	0.70	1.0	0.68	1.0	0.67
Proposed method	0.2	0.48	0.2	0.81	0.2	0.73	0.2	0.74
	0.4	0.75	0.4	0.73	0.4	0.78	0.4	0.69
	0.6	0.85	0.6	0.76	0.6	0.85	0.6	0.79
	0.8	0.79	0.8	0.69	0.8	0.89	0.8	0.74
	1.0	0.88	1.0	0.85	1.0	0.71	1.0	0.70

precision rates compared to Algorithm (1) for the given recall values. Precisely, for recall values 0.4 and above, the proposed CBCD scheme1 scores good precision rates, which vary between 0.69-0.89; whereas the precision rates of the reference method fall between 0.4-0.86. In this way, Table 3.2 results prove the better detection accuracy of the proposed CBCD scheme1 compared to the reference method.

Further, the PR rates of the proposed and the reference methods, which fall between error intervals are indicated in Table 3.3. The results given in Table 3.3 demonstrate that, the proposed CBCD scheme1 provides better detection accuracy

Table 3.3: PR rates of CBCD scheme1 (at error intervals)

Transforms	Blurring		Zoom-in		Zoom-out		Contrast	
	R	P	R	P	R	P	R	P
Algorithm (1)	0.12	0.09	0.14	0.07	0.21	0.10	0.17	0.05
	0.35	0.23	0.30	0.21	0.36	0.20	0.29	0.19
	0.57	0.38	0.62	0.47	0.54	0.38	0.45	0.32
	0.74	0.59	0.84	0.59	0.79	0.52	0.72	0.59
Proposed method	0.18	0.10	0.12	0.10	0.19	0.13	0.19	0.11
	0.32	0.27	0.38	0.23	0.28	0.20	0.32	0.20
	0.64	0.54	0.67	0.52	0.46	0.35	0.56	0.39
	0.83	0.72	0.89	0.76	0.83	0.69	0.84	0.68

compared to the detection rates of the reference method, although the PR results are falling between irregular intervals.

Table 3.4 indicates the detection results of the reference and proposed methods for various video attacks such as rotation, image ratio, noise addition and resolution change. Specifically, in Table 3.4, Recall and Precision rates of the proposed CBCD scheme1 is compared with the respective PR rates of Algorithm (1), in which the PR results fall within the correct intervals. Table 3.4 results demonstrates that, the

Table 3.4: PR rates CBCD scheme1 (at correct intervals)

Transforms	Rotation		Image ratio		Noise		Resolution	
	R	P	R	P	R	P	R	P
Algorithm (1)	0.2	0.45	0.2	0.55	0.2	0.33	0.2	0.55
	0.4	0.6	0.4	0.59	0.4	0.49	0.4	0.59
	0.6	0.53	0.6	0.63	0.6	0.51	0.6	0.63
	0.8	0.64	0.8	0.68	0.8	0.62	0.8	0.68
	1.0	0.72	1.0	0.79	1.0	0.79	1.0	0.79
Proposed method	0.2	0.64	0.2	0.67	0.2	0.66	0.2	0.67
	0.4	0.68	0.4	0.68	0.4	0.72	0.4	0.68
	0.6	0.75	0.6	0.72	0.6	0.78	0.6	0.72
	0.8	0.89	0.8	0.89	0.8	0.89	0.8	0.79
	1.0	0.88	1.0	0.89	1.0	0.78	1.0	0.85

proposed CBCD scheme1 achieves better precision rates compared to Algorithm(1) for the given recall values. In addition, Table 3.5 shows the detection results of the proposed and the reference methods falling between error intervals. The results given in Table 3.5 proves that, the proposed CBCD scheme1 provides better PR

Table 3.5: PR rates of CBCD scheme1 (at error intervals)

Transforms	Rotation		Image-ratio		Noise		Resolution	
	R	P	R	P	R	P	R	P
Algorithm (1)	0.14	0.08	0.11	0.9	0.12	0.04	0.17	0.12
	0.39	0.20	0.29	0.14	0.28	0.12	0.32	0.19
	0.53	0.39	0.48	0.29	0.45	0.24	0.53	0.24
	0.70	0.55	0.69	0.46	0.68	0.38	0.75	0.48
Proposed method	0.17	0.11	0.19	0.12	0.17	0.10	0.19	0.13
	0.35	0.28	0.33	0.27	0.32	0.23	0.38	0.23
	0.68	0.43	0.72	0.56	0.54	0.37	0.62	0.49
	0.89	0.66	0.85	0.69	0.84	0.62	0.86	0.63

rates compared to that of the reference method, though the precision as well as recall measurements are not coming under correct intervals.

Detection efficiency of CBCD scheme1

To evaluate the efficiency, computational cost involving the detection of a single duplicate video is considered. Precisely, the detection efficiency of the proposed CBCD scheme1 is evaluated by comparing its computational cost with that of Kim and Nam's method (2009). Kim and Nam employed luminance of frames as feature descriptors for their CBCD task. The experiments are conducted on a standard PC with 3.2 GHz CPU and 2 GB RAM. Table 3.6 gives the computational cost details of both the proposed and reference methods. The results from Table 3.6 show that, the proposed CBCD scheme1 is more efficient compared to Kim and Nam's method, since it reduces the total computational cost up to 65%.

Table 3.6: Computational cost comparison of CBCD scheme1

Task	Kim and Nam's method (in secs)			Proposed scheme1 (in secs)		
	1 min	3 min	5 min	1 min	3 min	5 min
Fingerprint extraction	16.00	51.00	97.00	13.98	34.84	52.56
Fingerprint matching	6.50	18.70	27.80	0.64	1.14	2.68
Total cost	22.500	69.700	124.800	14.634	35.989	55.250

Detection accuracy of CBCD scheme2

Table 3.7 compares the precision and recall rates of baseline DCD, proposed DCD and Cho et al.'s (2009) methods. Precisely, the baseline DCD method denotes the proposed CBCD scheme1, in which fingerprints are generated using dominant colors and their distribution values. In Table 3.7, the proposed DCD method denotes CBCD scheme2, which employs dominant color features and their spatial correlation values. Cho et al. (2009) employed ordinal measure signatures for detecting video copies.

Table 3.7: Copy detection results of CBCD scheme2

Transforms	Baseline DCD (%)		Cho's method (%)		Proposed DCD (%)	
	Precision	Recall	Precision	Recall	Precision	Recall
Blurring	98.1	83.2	90.1	78.8	100	95.6
Brightness	96.4	81.4	92.3	77.1	100	98.7
Noise addition	90.5	73.3	83.5	65.3	100	92.5
Zooming out	93.7	69.4	84.1	59.2	99.2	84.4
Image ratio	88.6	61.9	66.6	60.8	98.1	79.6
Zooming in	91.3	57.8	59.7	52.4	100	63.7
Image resize	92.4	58.1	73.6	51.9	100	69.3
Rotation	94.4	64.4	69.5	61.5	98.6	78.4

The results from Table 3.7 prove that, the proposed DCD method provides better precision and recall values. Therefore, the proposed CBCD scheme2 improves the detection accuracy (up to 24%) compared to the baseline DCD method. The results also show that, the proposed DCD method provides better detection results and significantly increases detection accuracy up to 38.1% compared to Cho et al.'s method.

Detection efficiency of CBCD scheme2

Figure 3.3 compares the computational cost of the proposed CBCD scheme2 with the reference methods. Specifically, Figure 3.3 shows the computational costs of Kim's method (Kim and Nam 2009), baseline DCD and the proposed DCD methods. Here, the baseline DCD method denotes CBCD scheme1, whereas proposed DCD method represents CBCD scheme2.

The experiments are conducted on a standard PC with 3.2 GHz CPU and 2 GB RAM. Though, the computational cost of the proposed DCD method is slightly higher compared to baseline DCD method (up to 12%); yet it provides improved detection results. However, the proposed DCD method is more efficient compared to Kim and

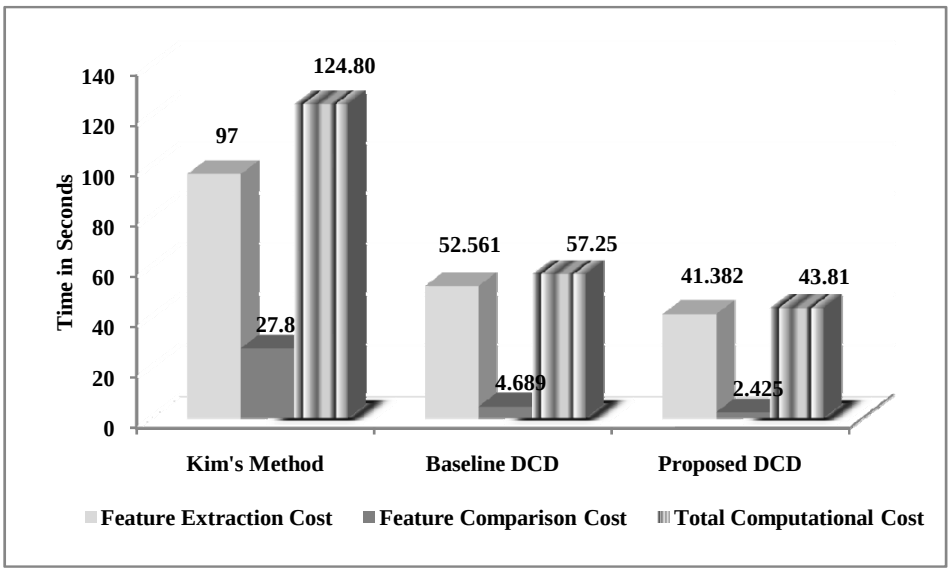


Figure 3.3: Comparison of computational cost

Nam's method. Precisely, Figure 3.3 results prove that, the proposed DCD method is 3 times faster than Kim and Nam's method, since it reduces total computational cost upto 91%.

3.2 CBCD Scheme Using Motion Activity Features

Motion features contribute important information about a video content. Therefore, motion-features based video analysis is used in several applications such as video retrieval, summarization (Divakaran et al. 2001) and characterization (Koprinska and Carrato 2001). However, as mentioned in Section 1.5., motion features are considered as poor descriptors in the CBCD literature(Hampapur et al. 2001), since the raw motion vectors are noisy in nature. Further, the conventional motion features describe the temporal content of a video clip; yet they fail to describe the overall activity of a video sequence. To solve these issues, this thesis contributes a novel CBCD framework, by fusing temporal behavior and spatial distribution of motion activity. Precisely, the main contributions of the proposed CBCD framework are given by,

- Describing the spatial and temporal activity of a video sequence, when compared to the conventional temporal motion vector approaches.
- Combining robust motion activity features such as dominant direction, motion intensity, and spatial distribution of activity to achieve the copy detection task.
- Performing matching using clustering to speed up the similarity mapping task.

The framework of the proposed CBCD method in terms of features extraction and signature matching techniques is detailed below.

3.2.1 CBCD framework using motion activity features

The block diagram of the proposed copy detection framework is shown in Figure 3.4, and the relevant symbols are described in Table 3.8.

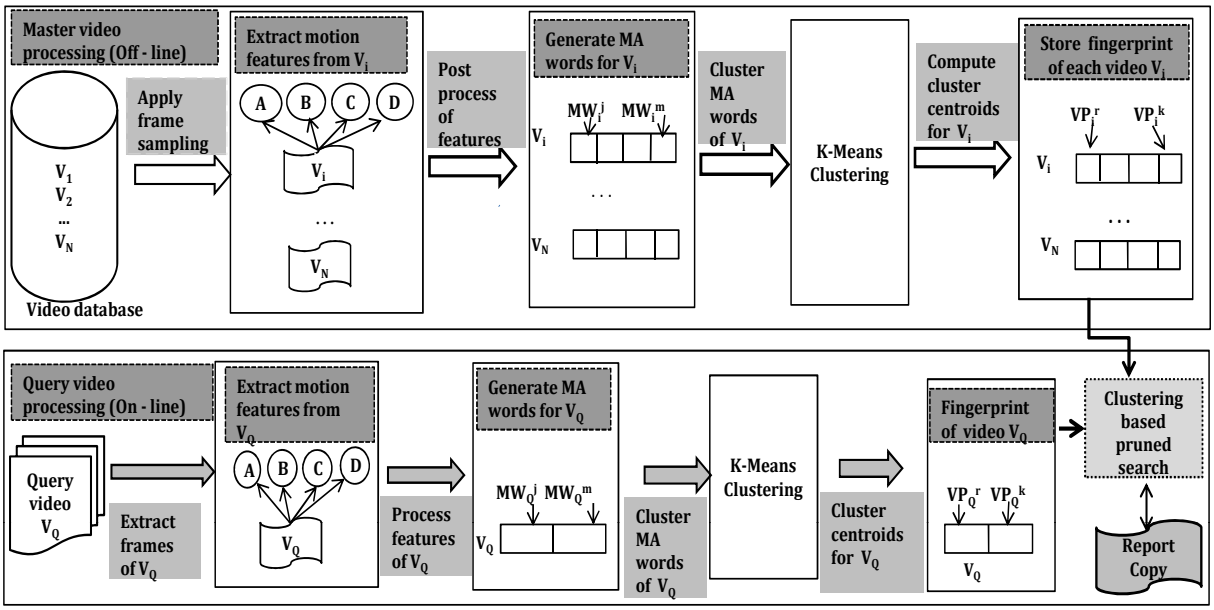


Figure 3.4: Proposed CBCD framework using motion features

Table 3.8: Description of notations used in Figure 3.4

Notation	Description	Notation	Description
N	Total videos in the DB	V_Q	Query video file
V_i	i-th master video in DB	VP_i^r	r-th fingerprint of i-th video V_i, such that r={1, 2, 3, ..., k}
A	Intensity of activity	B	Spatial distribution of activity (No. of active regions)
C	Dominant direction of activity	D	Average MV magnitude
MW_Q^j	j-th MA word of V_Q	VP_Q^r	r-th video fingerprint of V_Q
MW_i^j	j-th MA word of V_i, such that j ={1, 2,.. m}	V_N	N-th master video in DB
m	Number of MA words of V_i	k	Number of signatures of V_i

The proposed framework consists of two main stages: Master video processing

stage (off-line) and Query video processing stage (on-line). In the off-line stage, motion activity based features including intensity of action, spatial distribution, dominant direction of activity and average motion vector magnitudes are extracted from the master video frames. Then the resultant features are further processed and *Motion Activity (MA)* words are computed. MA words integrate raw motion activity features; hence, they comprehensively indicate the overall activity of the video sequences. In order to obtain the compact representation of MA words, the proposed framework employs widely popular K-means clustering algorithm and consequently the resultant cluster centroids are stored as the video fingerprints of the corresponding video sequences.

In the on-line stage, motion activity signatures are derived from the query video frames and MA words are calculated. The resultant MA words are clustered and the respective centroids are stored as video fingerprints. After this step, similarity matching task is performed by employing the clustering technique for detecting the video copies.

Fingerprint Extraction

In the proposed framework, different attributes of the MPEG-7 motion activity descriptor and average motion vector magnitude of frames are jointly exploited for implementing the CBCD task. The reasons for this joint exploitation are: First, average motion vector magnitudes provide better frame-level content of video clips; Second, entire activity of the video sequence can be effectively characterized, by utilizing different attributes of the motion activity descriptor. The MPEG-7 motion activity descriptor and motion activity features extraction are illustrated as follows.

MPEG-7 Motion Activity Descriptor: This descriptor captures the intensity of activity or pace of action in a video segment (Jeannin and Divakaran 2001). For instance, a *high speed car chase* denotes a high activity sequence, while *an interview scene* represents a low activity sequence. The motion activity descriptor is defined in terms of the following four attributes as given by,

$$Motion\ activity = \{I, Dir, Spatial, Temporal\} \tag{3.11}$$

where I indicates intensity of motion activity using an integer value and Dir represents the dominant direction of activity. Spatial distribution of activity ($Spatial$) indicates the number of active regions in a frame and temporal distribution attribute ($Temporal$) represents the variation of activity over the duration of a video sequence.

Motion Intensity (*I*): This attribute provides an effective temporal description of a video shot in terms of different intensity levels (Sun et al. 2001). The statistical properties of motion vectors such as average and standard deviation, can be used to calculate the intensity of motion activity. The Average Motion Vector magnitude (AMV) and Standard deviation of Motion Vector magnitude (SMV) of a frame are given by,

$$AMV = \frac{1}{MN} \times \sum_{i=1}^{M} \sum_{j=1}^{N} mv(i,j) \tag{3.12}$$

$$SMV = \sqrt{\frac{1}{MN} \times \sum_{i=1}^{M} \sum_{j=1}^{N} |mv(i,j) - AMV|^2} \tag{3.13}$$

where $mv(i,j)$ indicates the motion vector of (i,j)-th block and M×N is the frame size in terms of macro blocks. In the proposed copy detection approach, SMV of macro blocks is employed to compute the motion intensity. SMV values are quantized into the range of 1-5 as per MPEG-7 standard (Jeannin and Divakaran 2001), which are given in Table 3.9.

Table 3.9: Quantization thresholds for MPEG-1 video

Activity value	Range of SMV
1	$0 \leq$ SMV <3.9
2	$3.9 \leq$ SMV < 10.7
3	$10.7 \leq$ SMV <17.1
4	$17.1 \leq$ SMV <32
5	$32 \leq$ SMV

Spatial Distribution of Activity (*Spatial*): This attribute represents, whether the activity is spread across many regions or confined to one region (Savakis et al. 2003). The segmentation of frame into $n \times n$ regions has an active role in determining the exact number of active regions in a given frame. Smaller values of n may leave important semantic content, while larger values of n increases the computational complexity.

In order to solve this issue, experiments are conducted for different n values ranging from 2 to 5 and the maximum accuracy is achieved when n=3. Therefore, in the proposed framework, spatial distribution of motion activity of frames is computed, by segmenting the frame into 3×3 regions. Algorithm 3.1, calculating the spatial distribution of activity in a frame is described in Figure 3.5.

Algorithm 3.1: Computing Spatial Distribution of Activity

1: Calculate Spatial Activity Matrix (SAM) of each frame as given by,

$$SAM = \begin{cases} magmv(i,j) & if\ magmv(i,j) \geq AMV \\ 0 & otherwise \end{cases} \tag{3.14}$$

where $magmv(i,j)$ is the magnitude of motion vector of block (i,j).

2: Segment SAM of each frame into non overlapping blocks of size 3×3.

3: Compute the mean motion distribution (MMD) of r-th block of k-th frame as given by,

$$MMD(r) = \frac{Sum\ of\ SAM\ values}{Size\ of\ r} \tag{3.15}$$

4: Sort the regions of a frame in the ascending order of MMD values.

5: Regions with higher MMD values are considered as active regions of a given frame.

Figure 3.5: Algorithm to compute the spatial distribution of activity

Dominant Direction of Activity (*Dir*): Dominant motion directions of a video clip provide important information about its overall activity. Here, the objective is not to calculate the accurate direction of motion of all objects, but to compute the approximate dominant directions for enhancing the robustness of the proposed CBCD system. Therefore, in the proposed framework, the direction vector (Dir) represents the dominant direction of activity by computing the total motion in four major directions, which is formulated as (Benini et al. 2005),

$$Dir = \{Up, Down, Left, Right\} \tag{3.16}$$

Let $mv_x(k)$ and $mv_y(k)$, be the two components of motion vector of k-th block and N indicates total number of blocks, then the motion activity in four directions are calculated as,

$$Up = \sum_{k=1}^{N} (mv_y(k)), \quad if\ mv_y \leq 0 \tag{3.17}$$

$$Down = \sum_{k=1}^{N} (mv_y(k)), \quad if\ mv_y > 0 \tag{3.18}$$

$$Left = \sum_{k=1}^{N}(mv_x(k)), \quad if\ mv_x > 0 \tag{3.19}$$

$$Right = \sum_{k=1}^{N}(mv_x(k)), \quad if\ mv_x \leq 0 \tag{3.20}$$

The largest value of Dir provides dominant direction of motion in a given frame. Direct processing of resultant raw motion activity features is computationally expensive. Therefore, motion activity signatures are first normalized and consequently concatenated into informative MA words. However, the dimension of MA words is large. Hence, K-means clustering algorithm is used to obtain low dimensional representation of MA words.

Fingerprint matching

In the proposed CBCD system, the video signatures of master and duplicate video sequences are grouped into clusters. In experiments, the number of clusters for a video content ranges from 55-213. The cluster centroids of the master and query video sequences are compared and the similarity scores are evaluated against a confidence measure. The reference dataset is experimented with different confidence values varying between 0.50 and 0.75, to reduce the number of false positives. In experiments, good detection accuracy is obtained, when the confidence value is 0.65 and thus it is employed in the proposed copy detection task.

If R_1 and Q_1 are reference and query video clips, fp_r and fp_q are their corresponding video fingerprints, then the similarity score (S) between R_1 and Q_1 is computed using Manhattan distance metric as given by,

$$S(R_1, Q_1) = \sum_{i=1}^{m}\sum_{j=1}^{n}|fp_r(i) - fp_q(j)| \tag{3.21}$$

where m and n indicate total video signatures of R_1 and Q_1 respectively. When the similarity score exceeds the confidence measure, then the query video is indicated as a duplicate video.

3.2.2 Reference database and query dataset construction

As described in Section 2.8, followed by the better performance of the proposed CBCD techniques on Open Video Project datasets, this research study also exploits TRECVID datasets for evaluating the copy detection task. Specifically, the master database includes 75 hours of TRECVID 2007 database videos, covering a wide variety

of contents. If the sampling rates of video sequences are different, then the resultant motion vectors will also be different. To overcome this problem, the entire video data is transformed into 10 frames/sec using resampling technique and the resultant dataset is utilized as the reference database for the proposed CBCD task.

For experimentation purpose, eleven video clips are selected from the reference database and one video clip is collected from the non-reference database, in order to generate the query dataset. Precisely, the query data set totally includes twelve video clips (11+1), while the duration of these clips vary from 15-25 seconds. By applying ten different Transformations given by, T1) Brightness change, T2) Noise addition, T3) Blurring, T4) Color change, T5) Pattern insertion, T6) Moving caption insertion, T7) Slow motion, T8) Fast forward, T9) Cropping and T10) Picture-inside-picture, to the query dataset totally 120 (12×10) video copies are created, which serve as the query clips for the proposed CBCD task. Each video copy is used to detect the corresponding video sequence in the reference database. Figure 3.6 illustrates the sample frames from the transformed query videos. To measure the detection accuracy,

Figure 3.6: Example frames from the transformed query videos

the proposed framework also employs *F-measure* metric along with Precision and Recall, which is given by,

$$F\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.22}$$

where F-measure represents the robustness and discrimination ability of a system.

3.2.3 Copy detection results and discussion

Table 3.10 gives the detection results of Ordinal measure (Hua et al. 2004), Tasdemir's method (Tasdemir and Cetin 2010) and proposed method for T1-T5 transformations. The Ordinal measure (Hua et al. 2004) is a popular global descriptor, which is ex-

Table 3.10: Copy detection results (in %) for T1-T5 transformations

Transformations		**Ordinal Measure** (%)	**Tasdemir's Method** (%)	**Proposed Method** (%)
Type	**Metric**			
T1	P	56.93	70.15	82.85
	R	50.19	68.09	75.00
	F-M	53.34	69.10	78.72
T2	P	41.69	42.17	50.00
	R	40.48	41.18	46.15
	F-M	41.07	41.66	47.99
T3	P	56.86	55.81	57.14
	R	79.14	72.56	92.30
	F-M	66.15	63.09	70.58
T4	P	59.26	60.01	63.63
	R	67.79	69.27	84.83
	F-M	63.23	64.30	72.71
T5	P	79.68	79.82	82.85
	R	80.24	79.47	85.29
	F-M	79.95	79.64	84.05

tracted as follows: Segments the image into N blocks; Then, sorts the blocks according to their average intensity level and the ranking order of blocks are treated as Ordinal signatures. Tasdemir's method utilizes average motion vector magnitudes of frames as fingerprints for the copy detection task.

For T3 (Blurring) transformation, Ordinal measure performs well (66.15%) compared to Tasdemir's method (63.09%), due to its global descriptive properties. However, the proposed method (70.58%) outperforms Ordinal measure, as it uses spatial and temporal motion activity features. For T5 (Pattern insertion) transformation, both Tasdemir's method and Ordinal measure give very similar results (79.95% & 79.64%), which are less than that of the proposed method (84.05%).

Results from Table 3.10 demonstrate the better detection accuracy of the proposed method compared to the reference methods. Integration of spatial as well as temporal motion activity features for the copy detection task is the exact reason for the improved performance of the proposed method (84.05%). Table 3.11 shows the copy detection results of the proposed and reference methods for T6-T10 transformations and the results demonstrate that, the proposed method scores better performance

compared to the reference methods.

Table 3.11: Copy detection results (in %) for T6-T10 transformations

Transformations		**Ordinal**	**Tasdemir's**	**Proposed**
Type	**Metric**	**Measure** (%)	**Method** (%)	**Method** (%)
T6	P	70.64	74.58	82.50
	R	71.15	72.94	84.61
	F-M	70.89	73.75	83.54
T7	P	71.35	71.87	75.00
	R	59.18	70.35	87.80
	F-M	64.69	71.10	80.89
T8	P	60.10	62.71	65.85
	R	61.54	69.64	84.37
	F-M	66.15	63.09	70.58
T9	P	68.29	65.83	80.00
	R	73.58	69.98	82.75
	F-M	70.83	67.84	81.35
T10	P	61.54	60.17	**100.00**
	R	43.67	66.73	74.54
	F-M	51.09	63.28	85.41

For T10 (Picture-inside-picture) transformation, Ordinal measure gives poor Recall rate (43.67%) when compared to proposed and Tasdemir's methods. But the proposed approach provides better Recall (74.54%), Precision (100%) rates when compared to that of Ordinal measure (43.67% & 61.54%) and Tasdemir's methods (66.73% & 60.17%). The reason for the improved performance of proposed method is, the inclusion of dominant direction of activity as one of the features for the CBCD task. In this way, the proposed CBCD scheme improves the detection accuracy up to 13.9%, compared to the reference methods.

3.3 CBCD Systems Using Acoustic Fingerprints

Acoustic features are robust and powerful in describing a video content; yet they are not completely utilized for the emerging Content-Based video Copy Detection (CBCD) problem. On the other hand, as mentioned in Section 1.5., in most of the CBCD cases, the audio content is less affected compared to the visual part. Therefore, it is possible to detect illegal videos using their audio features, even the visual content is badly distorted. By considering these factors, this thesis contributes two copy detection techniques, which employ audio spectral features for detecting video copies, as illustrated below.

3.3.1 Video copy detection using audio spectral features

Audio content is an important information source of video sequence; hence they are widely used in video parsing, indexing and scene categorization approaches (Tsekeridou and Pitas 2001; Zhu et al. 2009). Further, past acoustic investigations prove that, the most important perceptual audio features are existing in the frequency domain (Li et al. 2003; Jie et al. 2009). Therefore, *the main objective of the proposed CBCD approach is to show that, the robust audio spectral features can be efficiently utilized for the copy detection task.* Specifically, the main contributions of the proposed copy detection scheme are given by,

- Novel copy detection framework, which uses audio features, compared to the state-of-the art visual content based CBCD methods.
- Calculating compact spectral descriptive words, which combine the robust spectral features such as signal energy, roll-off, centroid and flux.
- Clustering based pruned similarity matching to speed up the fingerprint mapping process.

The proposed copy detection system including framework, spectral features extraction and fingerprints matching is detailed as follows.

Proposed CBCD framework using audio spectral features

Figure 3.7 shows the block diagram of the proposed copy detection framework, which comprises two stages: Off-line or Master video processing stage and On-line or Query video processing stage.

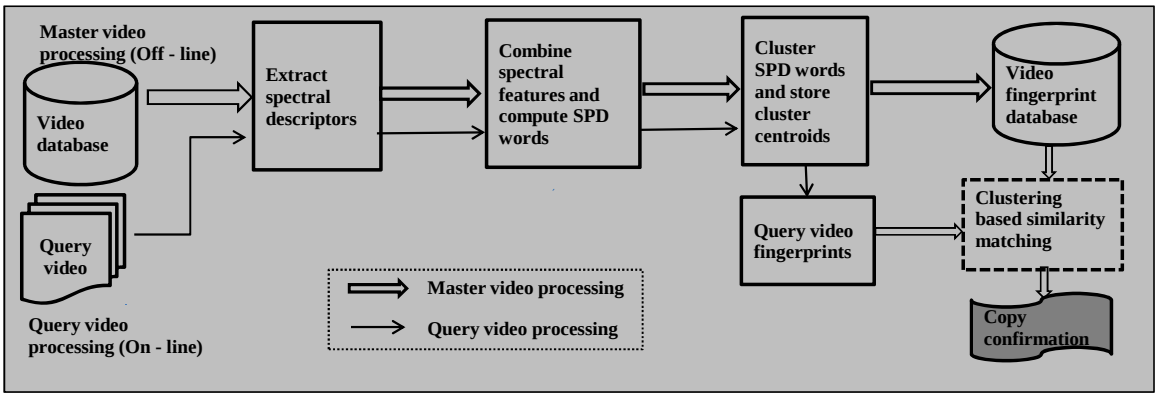


Figure 3.7: Proposed CBCD framework using audio spectral features

In the off-line stage, spectral descriptors including centroid, energy, roll-off and flux are extracted from the master video sequences. These features are further processed and *Spectral Descriptive (SPD)* words are computed. SPD words integrate raw spectral features; hence they summarize the overall audio content of a given video sequence. K-means clustering is employed to obtain low-dimensional representation of SPD words and consequently, the centroids are used as video fingerprints of master video sequences. In the on-line stage, spectral descriptors are extracted from the query video frames and the respective SPD words are calculated. Then the resulting SPD words are clustered and then the cluster centroids are stored as video fingerprints. After this step, clustering based similarity matching is performed for identifying the duplicate videos.

Spectral descriptors extraction

First the audio signal is down sampled to 22050 Hz, in order to decrease the amount of data to be processed. The magnitude spectrum of the audio signal is almost stationary for 10-30ms of window length (Rabiner and Juang 1993). Therefore, the down sampled audio signal is segmented into 11.60ms windows using Hamming window function with an overlapping factor of 86%. Then, audio spectral descriptors including centroid, energy, roll-off and flux are extracted from the short term power spectrum of audio signals, as described below.

Spectral Centroid (SC): Spectral centroid is a timbral feature, which illustrates the brightness of a sound signal (Park 2010). Generally, sound with brighter quality consists of more amount of high frequency components, compared to sound with dark quality. In sound synthesis techniques, spectral centroid is proved to be an important descriptor (Li et al. 2003), which indicates the center of gravity of the signal spectrum. The spectral centroid (SC) is computed as,

$$SC = \frac{\sum_{k=1}^{N} k \times x^d[k]}{\sum_{k=1}^{N} x^d[k]} \tag{3.23}$$

Where $x^d[k]$ represents the magnitude of k^{th} frequency bin of d^{th} frame and N is the frame length. The statistical properties of spectral centroid such as mean, standard deviation and log amplitude are used in various speech analysis and recognition algorithms (Park 2010; West 2008). The average frequency distribution values are utilized as spectral centroids in the proposed CBCD task.

Figure 3.8 shows the example spectral centroid plot of master and copied video sequences. Here, the video copy is created by applying 3 combined transformations including mp3 compression, cropping and pattern insertion. The centroid plots indicate a very high similarity (up to 98.7%) between the master and copied feature sequences and thus prove the robust nature of the spectral descriptor used in the proposed CBCD task.

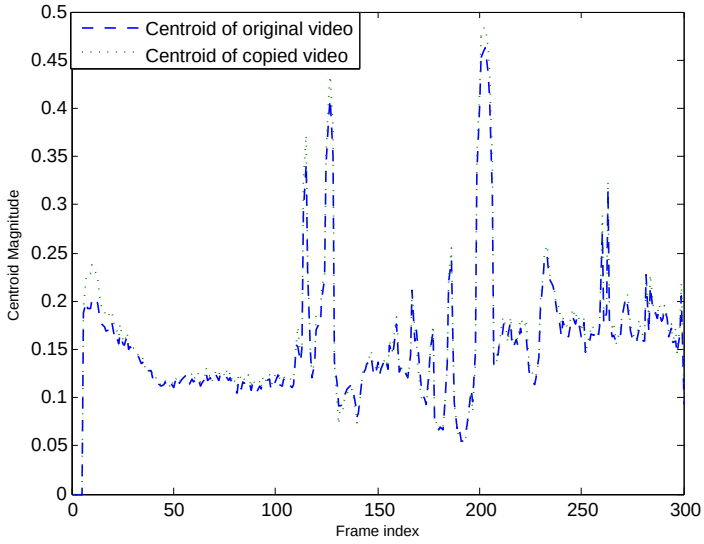


Figure 3.8: Similarity of spectral centroid plots of master and copied videos

Spectral Energy (SE): This descriptor calculates the average short term power of the input signal (West 2008). In this proposed work, the sum of squared magnitude of samples is utilized to calculate the spectral energy, as given by,

$$SE = \frac{1}{N}\sum_{k=1}^{N}\left|x^{d}(k)\right|^{2} \tag{3.24}$$

where N is the length of the frame. Figure 3.9 shows the example spectral energy plot of master and copied video files. Figure 3.9 curves indicate a very high similarity (up to 95.8%) between the two feature sequences and thus proves the robustness of the spectral energy features used in the proposed CBCD framework.

Spectral Roll-off (SR): This feature is commonly referred to as *skew* present in the shape of the power spectrum. Precisely, the roll-off point defines the frequency boundary, in which 85% of the power spectrum energy resides. Therefore, this descriptor is widely used to differentiate between constant as well as highly transient

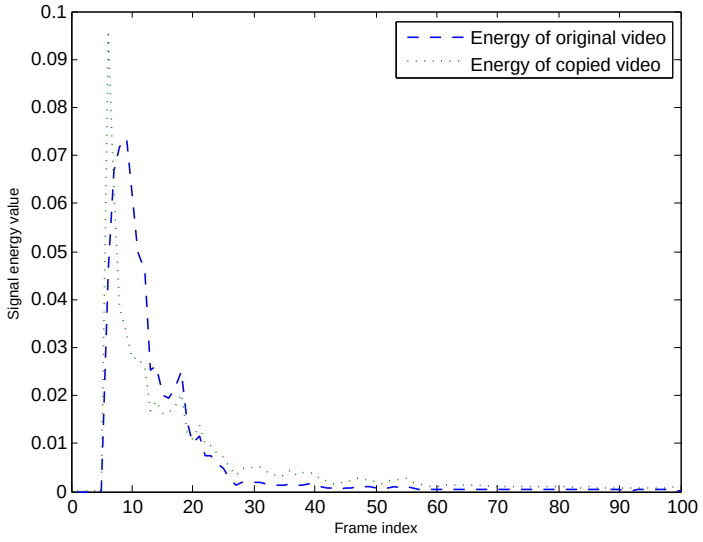


Figure 3.9: Similarity of signal energy plots of master and copied videos

sounds (Burka 2010). The spectral roll-off can be calculated as (Park 2010),

$$SR = \sum_{k=0}^{R} x^d[k] = 0.85 \sum_{k=0}^{N-1} x^d[k] \tag{3.25}$$

where N represents the frame length and $x^d[k]$ indicates the magnitude components of k^{th} frequency bin and R indictes the frequency roll-off point with 85% of energy.

Spectral Flux (SF): Generally, speech signals change at a faster rate, compared to music signals (Park 2010). Spectral flux defines the amount of energy difference between consecutive analysis frames (Burka 2010), which is computed as follows,

$$SF = \mid x_f^d[k] - x_{f\text{-}1}^d[k] \mid \tag{3.26}$$

where $x^d[k]$ represents magnitude of k^{th} frequency component and f,f-1 indicate current and previous frames respectively. Spectral flux is mainly utilized to compare musical and speech signals.

The dimension of resultant spectral descriptors is large (10240/sec), as a result, direct processing of raw features is computationally expensive. Therefore, the resultant feature descriptors are combined into highly informative SPD words. K-means clustering is used to get the compact representation of SPD words. In experiments, the number of clusters of the video sequences vary between 47-413, based upon the individual video contents.

Fingerprint matching

In the proposed duplicate video detection framework, L1-norm Manhattan distance metric is used to compute the similarity between the two video clips. If M_k is k^{th} master video and Q is query video clip, then f_m and f_q are their corresponding video fingerprints. The similarity score (Sim) between M_k and Q is computed as,

$$Sim(M_k, Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} |(f_m(i) - f_q(j)| \tag{3.27}$$

where m, n represent the size of master and query video signatures respectively. The Sim scores are compared against the predefined confidence measure in order to detect the video copies. In the proposed framework, different confidence measures ranging from 0.55-0.73 are experimented and 0.70 confidence measure provides better accuracy, hence it is employed in the proposed copy detection task.

Experimental setup

As mentioned in Section 2.8, TRECVID 2008 Sound and Vision dataset is utilized for evaluating the proposed method. Precisely, the video database contains 75 hours of video covering a wide variety of content including documentaries and science. Table 3.12 lists the audio and visual transformations considered in the proposed CBCD task. In the experiments, seven video clips from the reference database and two video clips

Table 3.12: List of visual and audio attacks considered in the proposed CBCD system

Type	**Transformations**
T1	Blurring
T2	Color change
T3	Slow motion
T4	Fast forward
T5	Pattern insertion
T6	Moving caption insertion
T7	Cropping
T8	Picture-inside-picture
T9	Mp3 compression
T10	singleband companding
T11	Multiband companding
T12	Combination of 3 transformations (Cropping, pattern insertion and mp3 compression)

from Open Video Project are used as query video sequences. The transformations

listed in Table 3.12 are applied to the nine query video sequences, while the duration of these clips vary from 20 to 25 seconds. The resulting 108 (12×9) video clips are employed as query video clips for the proposed copy detection task. To measure the detection accuracy of the proposed scheme, standard performance metrics such as Precision, Recall and F-Measure metrics as specified in Equations (3.9), (3.10) and (3.22) are utilized.

Copy detection results and discussion

The accuracy of the proposed copy detection method is compared with Ordinal measure (Hua et al. 2004) and Itoh et al.'s (Itoh et al. 2010) methods. The Ordinal measure is extracted as follows: First, the image is partitioned into N blocks; Then, the blocks are sorted according to their average intensity values and consequently, the ranking order of blocks are considered as ordinal signatures. Itoh's method employs significant points in acoustic data for identifying duplicate videos, which is executed as follows: First acoustical power envelopes of the input signal are computed; Then the significant points denoting local minimum/maximum values are extracted from the power envelopes, and used as fingerprints for the copy detection task.

Table 3.13 lists the detection results of the proposed and reference methods for T1-T6 transformations. The results from Table 3.13 prove that, the proposed method enhances detection accuracy by 29.78%, when compared to the reference methods.

More precisely, for T3 (slow motion) transformation, Ordinal measure provides poor recall rate (58.45%), when compared to that of Itoh's method (61.97%). The global descriptive nature of Ordinal measure is the reason for this poor performance. Although Itoh's method performs better than the Ordinal measure for T1-T6 transformations; yet, the proposed method outperforms Itoh's method for all six transformations. Specifically, the proposed method scores better recall rate (100%), when compared to that of Ordinal measure (70.11%) and Itoh's method (70.16%) for T4 (fast forward) transformation. The robust nature of the audio spectral features is the exact reason for the better performance of the proposed CBCD method.

Table 3.14 lists the detection results of the proposed and reference methods for T7-T12 transformations and the results demonstrate that, the proposed method improves the detection accuracy by 25.91%, when compared with the reference methods.

For T12 (3 combined transformations), Ordinal measure results in very poor precision, recall rates (58.33% & 51.29%), compared to that of Itoh's (68.39% & 65.11%) and proposed methods (95.96% & 93.21%). Ordinal measure is much affected by region-based transformations; hence, it yields poor results for T12. The detection

Table 3.13: Detection results (in %) for T1-T6 Transformations

Transformations		**Ordinal**	**Itoh's**	**Proposed**
Type	**Metric**	**Measure (%)**	**Method(%)**	**Method (%)**
T1	P	76.24	79.09	**100.00**
	R	70.58	69.08	79.48
	F-M	73.30	73.74	88.56
T2	P	65.35	81.57	97.89
	R	79.14	78.34	96.37
	F-M	71.58	79.92	97.12
T3	P	61.22	71.58	99.39
	R	58.45	61.97	99.40
	F-M	59.80	66.42	99.39
T4	P	74.29	79.59	99.69
	R	70.11	70.16	**100.00**
	F-M	72.13	74.57	99.84
T5	P	68.36	81.62	99.09
	R	69.16	79.31	98.86
	F-M	64.07	80.44	98.13
T6	P	59.69	80.66	97.42
	R	69.16	74.31	98.86
	F-M	64.07	77.35	98.13

Table 3.14: Detection results (in %) for T7-T12 Transformations

Transformations		**Ordinal**	**Itoh's**	**Proposed**
Type	**Metric**	**Measure** (%)	**Method** (%)	**Method** (%)
T7	P	74.24	85.61	99.00
	R	68.86	80.25	92.66
	F-M	71.44	82.84	95.72
T8	P	72.69	88.19	94.44
	R	71.10	80.27	90.26
	F-M	71.88	84.04	92.30
T9	P	73.65	60.24	97.22
	R	72.58	62.33	97.29
	F-M	73.11	61.27	97.25
T10	P	80.06	74.31	93.44
	R	73.64	72.59	90.28
	F-M	76.71	73.44	91.83
T11	P	66.28	69.34	90.36
	R	61.15	60.22	92.22
	F-M	63.61	64.46	91.28
T12	P	58.33	68.39	95.96
	R	51.29	65.11	93.21
	F-M	54.58	66.71	94.57

scores of the proposed method is slightly less for T10-T12 transformations, since spectral descriptors are much altered by these three transformations. However, the proposed method provides better detection results compared to the reference methods by integrating the four robust spectral features for the copy detection task.

3.3.2 CBCD system using audio fingerprints and PCA

MFCCs (Mel-Frequency Cepstral Coefficients) are widely used by the audio processing community to get discriminative performance with reasonable noise robustness (Park 2010). Therefore, this thesis enhances the previous CBCD framework described in Section 3.3.1, by contributing an another copy detection method, which integrates MFCCs and spectral descriptors along with PCA (Principal Component Analysis) for detecting video copies. Specifically, the contributions of the proposed copy detection method are given by,

- Presenting a novel copy detection method by exploiting spectral audio features and MFCCs, compared to conventional visual content based CBCD techniques.
- Construction of multi-feature vectors, by concatenating various spectral feature sequences such as MFCCs and spectral descriptors.
- Dimensionality reduction of multi-feature vectors using PCA.

The framework of the proposed copy detection method along with the fingerprints extraction and similarity matching techniques is described as follows.

Proposed framework using audio fingerprints and PCA

The block diagram of the proposed copy detection framework is shown in Figure 3.10 and the relevant notations are described in Table 3.15.

The proposed framework comprises two main components: Off-line (Master video processing) and Online (query video processing). In the off-line stage, audio spectral features including MFCCs and spectral descriptors are extracted from the master video sequences. These intra-frame features are concatenated into high-dimensional Multi-Feature (MF) vectors of predefined window size. Since MF vectors combine raw features (both intra and inter-frame features), they effectively represent frame-level as well as clip-level information of video contents. Then PCA is applied on high-dimensional MF vectors, in order to get low dimensional representation. The sequence of principal components are subsequently combined and stored as fingerprints of master video sequences. In the online stage, MF vectors are calculated, after extracting audio spectral features from the query video frames. Then, principal

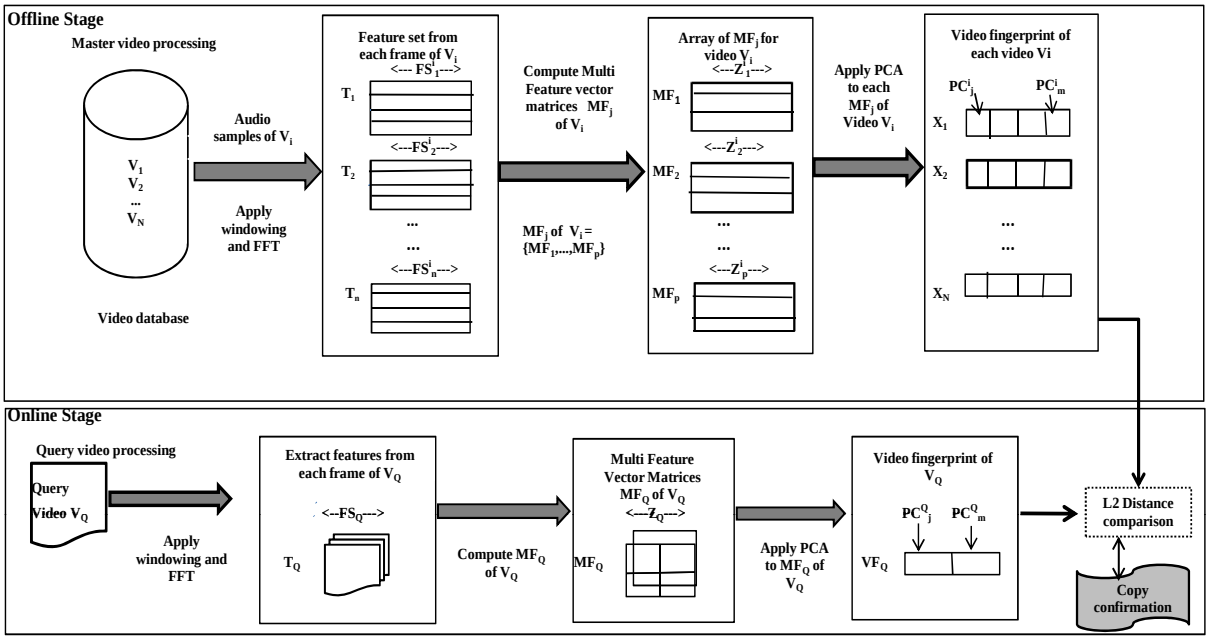


Figure 3.10: Proposed CBCD framework using audio features and PCA

Table 3.15: Description of notations used in Figure 3.10

Notation	Description	Notation	Description
N	Number of master videos	T_Q	Total frames of query video V_Q
V_i	i^{th} master video in the DB	FS_Q	Feature set extracted from V_Q
n	Total frames of video V_i	PC_j^i	j^{th} Principal component of V_i
T_i	i^{th} frame of video V_i, where $V_i = \{T_1, T_2, .., T_n\}$	X_i	Audio fingerprint of i^{th} video V_i where $X_i = \{PC_j^i, ..., PC_m^i\}$
FS_n^i	Feature Set extracted from n^{th} frame of V_i	MF_j	j^{th} MF vector matrix of V_i, where $j = \{1, 2, .., p\}$
Z_Q	Dimension of MF_Q of V_Q	PC_j^Q	j^{th} principal component of V_Q
Z_j^i	Dimension of j^{th} MF of V_i		where $j = \{1, 2, .., m\}$

components of the query video are calculated from MF vectors, and compared against the fingerprints of the master videos. Finally, L2-norm distance based comparison gives the output of proposed CBCD task.

Fingerprint extraction

As described in Section 3.3.1, first the audio signal is down sampled and then segmented into 11.60ms windows with an overlap factor of 86% using Hamming window function. After this step, the spectral representation of each analysis window is com-

puted by applying FFT (Fast-Fourier transform). From the spectral decomposition, two sets of features are extracted: Mel-Frequency Cepstral Coefficients (MFCCs) and spectral distribution descriptors.

MFCCs extraction: The MFCCs are based on the discrete cosine transform of log amplitude Mel-frequency spectrum (Park 2010). In the proposed scheme, FFT spectrum is divided into 24 bands and 40 triangular band pass filters that are placed using Mel-scale. Generally, first 15 cepstrum coefficients are employed in speech signal processing (Park 2010); hence, first 15 MFCC are calculated to capture short term spectral features of video frames and employed in the proposed copy detection framework.

Spectral distribution descriptors: In the proposed CBCD framework, the spectral descriptors such as Spectral Centroid, Energy, Roll-off and Flux are calculated as specified in Equations (3.23)-(3.26), for describing the power spectrum of the input audio signal. The output of the audio features extraction process results in the conversion of 11.60ms frames into a stream of feature vectors with 6 feature values. The resulting feature sequences are concatenated into MF vectors of length 580ms. Since the dimension of MF vector is very high (in the order of 15000), it is not feasible to perform any computations. In order to obtain the compact representation of the MF vectors, two different techniques are employed: (a) Instead of using all 15 MFCCs of frames, only MFCC means and variances are considered in the feature sets of frames; (b) PCA is applied to obtain the low dimensional representation of MF vectors.

Principal component analysis Given d-dimensional MF vectors MF_i ,such that i={1, 2, 3, ..., N}, the mean vector M (Burka 2010) is given by,

$$M = \frac{1}{N}\sum_{i=1}^{N} MF_i \tag{3.28}$$

The mean subtracted data set is given by $B = MF_i$ - M. The covariance matrix (Cov) is given by,

$$Cov = \frac{1}{N-1} BB^T \tag{3.29}$$

where B^T represents transpose of B. Finally, the eigenvectors V and eigen values λ are calculated directly from the covariance matrix by solving the generalized eigenvector problem (Burka 2010) for,

$$Cov.V = \lambda.V \tag{3.30}$$

In the experiments, only K eigenvectors with largest eigen values are considered as fingerprints, where K varies between 2 to 8.

Fingerprint matching

In the proposed CBCD framework, similarity matching is performed using weighted L2 Euclidean distance calculations. If P_1 and Q_1 are master and query video sequences, then f_p and f_q are their corresponding video fingerprints. The master video fingerprint f_p includes p_i eigenvectors and the corresponding λ_i eigen values, while the query video fingerprint f_q contains q_j eigen vectors and the corresponding σ_j eigen values. The distance ($Dist$) between p_i and q_j (Gu et al. 2004) is given by,

$$Dist\ (i,j) = |p_i - q_j|^2 \tag{3.31}$$

In general, eigen vectors with large eigen values specify the most significant relationships between data dimensions and hence the inclusion of eigen values in similarity calculations improves the performance of the CBCD system. Therefore, a weighting factor is considered in the experiments, which is given by,

$$W\ (i,j) = \frac{1}{\sqrt{\lambda_i^2}} . \frac{1}{\sqrt{\sigma_i^2}} \tag{3.32}$$

The similarity (SM) between two video sequences P_1 and Q_1 is defined as the weighted sum of similarity between their fingerprints, given by

$$SM\ (P_1, Q_1) = \sum_{i=1}^{f_p} \sum_{j=1}^{f_q} W(i,j)\ Dist(i,j) \tag{3.33}$$

Experimental setup

The proposed CBCD system is evaluated on TRECVID-2007 Sound & Vision data set. Table 3.16 represents the list of video transformations considered in the proposed CBCD framework, while Figure 3.11 illustrates all the transformations with example frames, extracted from the transformed query videos.

Precisely, the video database includes 25 hours of video covering a wide variety of content. The format of the reference video clips is 352×288 pixels and 30 frames/sec. In the experiments, seven video clips are selected from the reference dataset and one video clip collected from Open Video Project serves as the non-reference video stream. The sixteen types of transformations listed in Table 3.16 are applied to the resultant eight query video clips, while the duration of these clips varies from 30 to 45 seconds. The resulting 128 (16×8) video sequences are used as query video clips for the proposed CBCD task.

Table 3.16: List of transformations considered in the proposed CBCD framework

Category	Type	Description
Transformations -Level 1 (TL1)	T1: Brightness change	Increase brightness by 15% -25%
	T2: Noise Addition	Adding 15% random noise
	T3: Rotation	Rotating up to 90°
	T4: Blurring	Blurring by 20%
	T5: Horizontal flip	Horizontal mirroring up to 90°
	T6: Vertical flip	Vertical mirroring up to 100°
	T7: Color change	Changing color spectrum
	T8: Pattern insertion	Pattern is inserted into selective frames
	T9: Moving caption insertion	Entire video includes moving caption
	T10: Slow motion	Halve the video speed
	T11: Fast forward	Double the video speed
	T12: Zooming in	Zoom in by 15%
Transformations -Level 2 (TL2)	T13: Combination of 3 transformations of TL1	Applying 3 transformations amongst T1-T5
	T14: Combination of 5 transformations of TL1	Applying 5 transformations amongst T1-T4, T6-T8
	T15: Combination of 8 transformations of TL1	Applying 8 transformations amongst T1-T5, T7, T8, T10 and T12
	T16: Combination of 10 transformations of TL1	Applying 10 transformations amongst T1-T12

Figure 3.11: Example frames from the transformed query videos

Copy detection results

The copy detection results of the proposed CBCD method is compared with Cao's method (Cao and Zhu 2009) and baseline methods. Cao and Zhu (2009) employed the mean values of YCbCr components as the feature descriptors for their copy detection task. Baseline method uses only MFCC means and variances as feature descriptors. Table 3.17 compares the PR rates of baseline method, Cao's method and proposed methods for the first eight transformations of type TL1.

Table 3.17: PR rates for T1-T8 of TL1 transformations

Transformations		Cao's Method	Baseline Method	Proposed Method
T1	P	0.70943	0.72775	0.85714
	R	0.69604	0.81861	0.96428
T2	P	0.71621	0.82901	**1.00000**
	R	0.71542	0.80142	0.96825
T3	P	0.69654	0.83675	**1.00000**
	R	0.67554	0.78864	0.92461
T4	P	0.62910	0.71076	0.91541
	R	0.64839	0.67785	0.96923
T5	P	0.74472	0.88675	**1.00000**
	R	0.69843	0.73652	0.88405
T6	P	0.76871	0.81843	**1.00000**
	R	0.57983	0.54908	0.79602
T7	P	0.64911	0.71453	**1.00000**
	R	0.60152	0.62303	0.87341
T8	P	0.63301	0.78994	**1.00000**
	R	0.58973	0.61952	0.83554

For T8 transformation (Pattern insertion), Cao's Method gives poor recall rate (0.58973), when compared to that of the proposed method (0.83544). The reason for the poor performance of Cao's method is, the limited capability of global descriptors. Further, the proposed method yields good precision rates compared to baseline method, especially for T8 and T2 transformations. For Flipping transformation (T6) baseline method gives poor recall rate (0.54908) compared to that of proposed method (0.79602). Therefore, results from Table 3.17 prove that, the proposed method yields better detection rates compared to Cao's method and baseline methods.

Table 3.18 shows the precision and recall rates of Cao's method, baseline and proposed methods for T9-T16 transformations. Since TL2 transformations include multiple video editing tasks, the overall detection rates are slightly less compared to

Table 3.18: PR rates for T9-T16 of TL1 and TL2 transformations

Transformations		Cao's Method	Baseline Method	Proposed Method
T9	P	0.60812	0.73564	**1.00000**
	R	0.43867	0.61762	0.87342
T10	P	0.49972	0.54921	0.74332
	R	0.49889	0.63544	0.83747
T11	P	0.61367	0.65990	0.83875
	R	0.40175	0.59211	0.79002
T12	P	0.61832	0.72178	0.99642
	R	0.53761	0.65156	0.85714
T13	P	0.68120	0.78805	0.99218
	R	0.40961	0.52865	0.69543
T14	P	0.63592	0.79664	0.97564
	R	0.50183	0.65271	0.85285
T15	P	0.64883	0.78853	0.96605
	R	0.54241	0.64400	0.81824
T16	P	0.59971	0.69904	0.90679
	R	0.48762	0.52743	0.79775

that of TL1 type transformations. Although, T16 transformation includes ten types of complicated video editing activities, still the proposed method scores better precision rates (0.90679), compared to that of Cao's method and baseline methods. For T15 and T16 transformations, the recall rates of Cao's method is poor (0.54241 & 0.48762), because YCbCr values are significantly affected by combined visual distortions. In this way, Table 3.18 results demonstrate the improved detection accuracy of the proposed CBCD scheme compared to the reference methods.

3.4 Copy Detection System Using DCDs and Audio Features

As discussed in Section 2.1.3., most studies on Content Based video Copy Detection (CBCD) concentrate only on the visual signatures, while very few efforts are made to exploit audio features. However, audio data if present, is an essential source of a video; hence, the integration of visual-acoustic fingerprints for the CBCD task significantly improves the copy detection performance. Further, the *combined utilization of visual and audio fingerprints not only improves the copy detection performance, but also useful in many applications such as video retrieval, management and multimedia*

fingerprinting.

On the other hand, existing works on DCDs exploit only global description of dominant color features in an image. Therefore, *promising algorithms extracting region-based dominant color features and exploiting temporal color statistics are required, in order to provide compact visual fingerprints using color features.* Based on these factors, this thesis contributes a new CBCD framework, which integrates novel visual signatures extracted from Dominant Color Descriptors (DCDs) and robust acoustic fingerprints derived from Mel-Frequency Cepstral Coefficients (MFCCs) to detect duplicate videos. Precisely, the contributions of the proposed CBCD framework are given by,

- A novel DCD extraction algorithm denoted as *RGB-Feature Image* is introduced, which efficiently extracts the dominant color features from an image.
- A new visual signature called as *Spatio-Temporal DCDs* is presented, which effectively characterizes the region-based dominant color features and temporal color information present in a video sequence.
- A new approach for fusing visual-audio fingerprints is proposed, which employs combination rule and weight factor strategies.

The proposed copy detection system including framework, visual-audio fingerprints extraction and fusion is illustrated below.

3.4.1 Proposed CBCD system using DCDs & audio features

The overview of the proposed CBCD framework is shown in Figure 3.12 and the relevant symbols are explained in Table 3.19. The proposed framework consists of two main stages: off-line (Master video processing) and online (query video processing). In the off-line stage, visual fingerprints based on dominant color features and audio fingerprints based on MFCCs are extracted from master video contents. The resultant visual-audio fingerprints are stored in the fingerprint database.

In the online stage, when a query clip is given, resampling technique is employed in order to synchronize the query and master video frame rates. Then visual and acoustic fingerprints are extracted from the query video. After this step, audio-visual fingerprints are matched separately and the individual matching results are combined, in order to compute the detection results. The visual-audio fingerprints extraction and fusion techniques are detailed below.

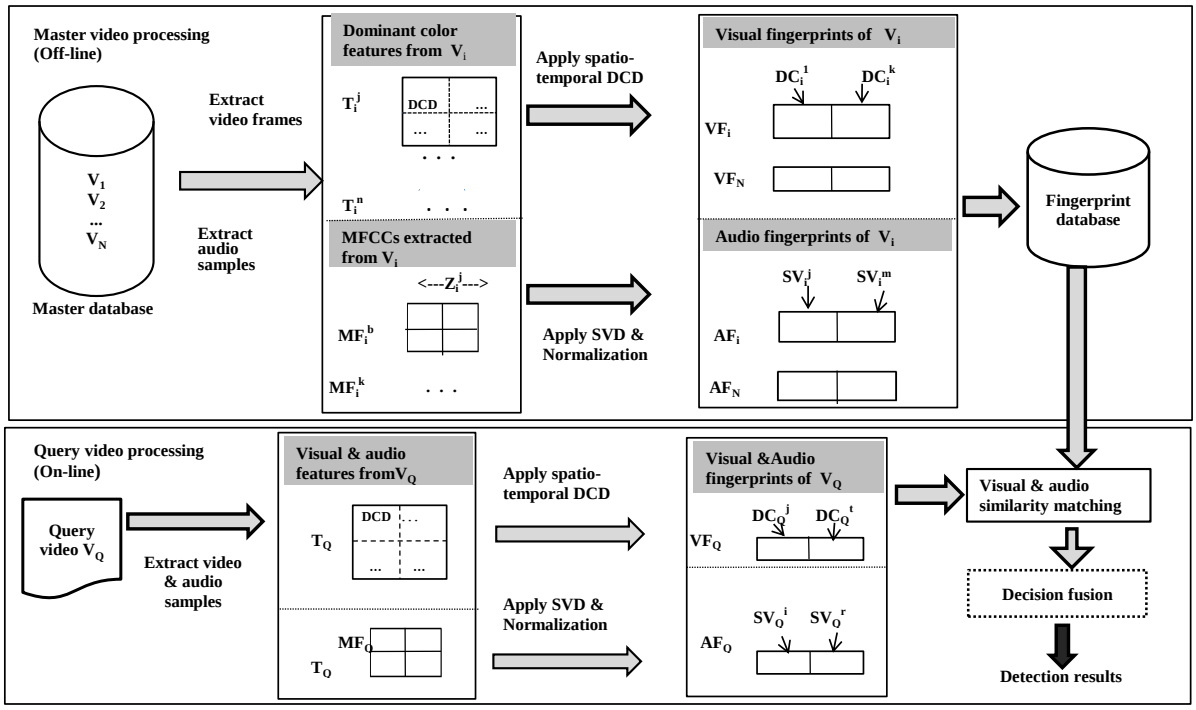


Figure 3.12: Proposed CBCD system using DCDs and audio features

Table 3.19: Description of notations used in Figure 3.12

Notation	Definition	Notation	Description
V_i	i^{th} master video	N	Number of master videos
T_i^j	j^{th} frame of V_i, where $j = \{1, 2, .., n\}$	MF_i^b	b^{th} MFCC feature of V_i, where $b = \{1, 2, ..k\}$
Z_i^j	Dimension of j^{th} MFCC features of V_i	VF_j	Visual fingerprint of V_j, and $VF_j \in [DC_i^l, ..., DC_i^k]$
DC_i^l	l^{th} DCD of V_i, $l = \{1, 2, .., k\}$	T_Q	Number of frames of V_Q
SV_i^j	j^{th} singular value of V_i, where $j = \{1, 2, .., m\}$	AF_j	Audio fingerprint of V_j, where $AF_j \in [SV_j^l, .., SV_j^m]$
MF_Q	MFCC features of V_Q	DC_Q^j	j^{th} DCD of V_Q, $j = \{1, 2, .., t\}$
AF_Q	Audio fingerprint of V_Q, where $AF_Q \in [SV_Q^i, ..., SV_Q^r]$	VF_Q	Visual fingerprint of V_Q $VF_Q \in [DC_Q^i, .., DC_Q^t]$
SV_Q^i	i^{th} singular value of V_Q, where $i = \{1, 2, .., r\}$	t and r	Total visual and audio fingerprints of V_Q

3.4.2 Visual-audio fingerprints generation

As described in Section 3.1.1, Dominant Color Descriptor (DCD) of MPEG-7 standard effectively describes the color information in an image, by capturing the dominant or

representative colors from that image (Manjunath et al. 2002). In the past studies, color clustering algorithm such as Generalized Lloyd algorithm (GLA) (Lloyd 1982) is widely utilized to extract DCDs from an image (Yang et al. 2008; Deng et al. 2001); but GLA suffers due to its high computational cost.

In order to solve this problem, a novel DCD extraction scheme denoted as *RGB-Feature Image* is proposed to efficiently extract the dominant colors from an image. More precisely, the proposed CBCD system expands the frequency imaging method introduced by Kashiwagu and Oe (2007) and derives a novel DCD extraction scheme called as *RGB-Feature Image*. Without the loss of generality, RGB color space is employed in the proposed framework to compute *RGB-Feature Images*, which extracts dominant colors and their relative distribution present in an image. Algorithm 3.2, given in Figure 3.13 details the steps used to compute the *RGB-Feature Image* from a given image/region.

Algorithm 3.2: *RGB-Feature Image* Computation

1: Let $C(x, y)$ be a color image having $m \times n$ pixels.

2: Translate each pixel p_i^c of $C(x, y)$ to a color histogram space HS, where $1 \leq i \leq (m \times n)$.

3: Calculate the frequency of each color in HS.

4: The frequency value in HS corresponding to each p_i^c of $C(x, y)$ is denoted as μ_i^{hs}.

5: Construct RGB feature image $R(x, y)$, by replacing each pixel value with the frequency value, which is formulated as,

$$R(x, y) = \forall\ p_i^c\ \exists\ C(x, y) : \quad \mu_i^{hs} \ \mapsto\ p_i^c \tag{3.34}$$

6: Compute dominant colors from the dominant frequencies of $R(x, y)$.

7: Calculate the distribution of each dominant color in $C(x, y)$ in terms of percentages.

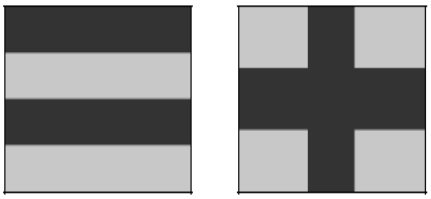
8: To improve robustness, the percentage of distribution of each dominant color is normalized using the formula given by,

$$\sigma_i = \frac{pd_i}{\Sigma pd_i}, \quad i \in [1 : r]. \tag{3.35}$$

where σ_i is the normalized distribution value and pd_i is the percentage of distribution of i^{th} dominant color and r indicates total dominant colors.

Figure 3.13: *RGB-Feature Image* computation algorithm

Most studies in DCD extraction exploit spatial information and provide only the global description of dominant colors in an image. For example, Yang et al. (2008) and Deng et al. (2001) extract dominant colors from the whole image/frame and utilize global description of colors to characterize an image. However, such global color descriptions may be inadequate to obtain precise localization of dominant colors in an image. In practice, many video or movie scenes exist with same dominant color patterns; though they have entirely different contents.



(a) (b)

Figure 3.14: Sample images having same global description of dominant colors irrespective of their different contents. (a) Image-1 having two horizontal lines with gray level 200 and background with gray level 50. (b) Image-2 having two cross lines with gray level 200 and background with gray level 50.

For instance, the images given in Figure 3.14(a)-(b) show the scenario in which global description of dominant colors is same irrespective of their contents; hence, in this case, global color description fails to differentiate these two images. Therefore, *region-wise distribution information of dominant colors is essential* in many applications such as content-based image retrieval and indexing.

On the other hand, it is to be noted that, *Consecutive images in a video sequence have very similar color statistics* (Roytman and Gotsman 1995). Thus, the proposed copy detection framework exploits the color similarity existing in the temporal domain for efficiently describing the color information of a video. Further, on average 3-8 dominant colors are needed to represent an image (Deng et al. 2001). Hence minimum number of DCDs extracted from a video V_i of size N frames is $N \times 3$, and size of the visual signature becomes approximately $N \times 3 \times 4$ (4digits/DCD). If size of N increases, then the computational cost also increases. However, *compact visual signatures are needed* to enhance the detection performance of a CBCD system.

In order to tackle the above mentioned issues, a new visual signature called as *Spatio-Temporal DCDs* is presented, which efficiently characterizes the color information of a given image. Precisely, *Spatio-Temporal DCDs* extract region-based dominant color features and exploit color statistics present in the consecutive frames, in order to provide efficient visual fingerprints of video sequences. Algorithm 3.3,given

in Figure 3.15 illustrates the steps used to compute *Spatio-Temporal DCD* signatures of a video file. Figure 3.16 illustrates *Spatio-Temporal DCDs* extraction from the sample frames on a 2×2 blocks. Experiments are conducted with different block sizes ranging from 2-6 for Spatio-Temporal DCD extraction. Since, 2×2 blocks provide better accuracy, hence it is employed in the proposed CBCD task.

Algorithm 3.3: *Spatio-Temporal DCDs* Computation

a: Let the video $V_i \in \{f_i | 1 \leq i \leq n\}$, where f_k is the k^{th} frame and n is total frames of V_i.

b: Segment each f_k into non-overlapping blocks of size 2×2.

c: Compute RGB feature image for each block using the algorithm given in Figure 3.15.

d: Extract dominant colors and their normalized distribution values for each block of f_k.

e: Repeat steps (3)-(5) for frame f_{k+1}.

f: Let $dc_k^{i,j}$ be j^{th} dominant color of i^{th} region of f_k.

g: Compute the dominant color features DF from the frame f_{k+1} using the formula,

$$\mathrm{DF}_{k+1}^{i,j} = \begin{cases} dc_k^{i,j} & dist \leq T \\ dc_{k+1}^{i,j} & dist > T \end{cases} \tag{3.36}$$

where $1 \leq i \leq 4, 1 \leq j \leq 12$ and $1 \leq k \leq n$. Here, the distance $dist$ indicates the distance between two colors $dc_k^{i,j}$ and $dc_{k+1}^{i,j}$ of frames f_k and f_{k+1} respectively. The threshold T is the minimum distance used to judge the similarity between two colors of consecutive frames and T is set to 15 in this work.

Figure 3.15: Algorithm for computing *Spatio-Temporal DCDs*

In order to emphasize the benefits of the proposed *Spatio-Temporal DCD* scheme, two sets of experiments are performed namely, spatial DCDs and Spatio-Temporal DCDs extraction. In (Roopalakshmi and Reddy July-2011), spatial DCD extraction scheme is employed, in which DCDs extracted from *RGB-Feature Images* are considered as visual signatures of video files. Table 3.20 shows total color features extracted by the spatial and Spatio-Temporal DCD methods for different video sequences. Table 3.20 results indicates that, the proposed *Spatio-Temporal DCD* scheme, reduces the amount of dominant color features up to 59%; hence, the *Spatio-Temporal DCD* scheme is more compact and profitable than the conventional spatial DCD extraction methods.

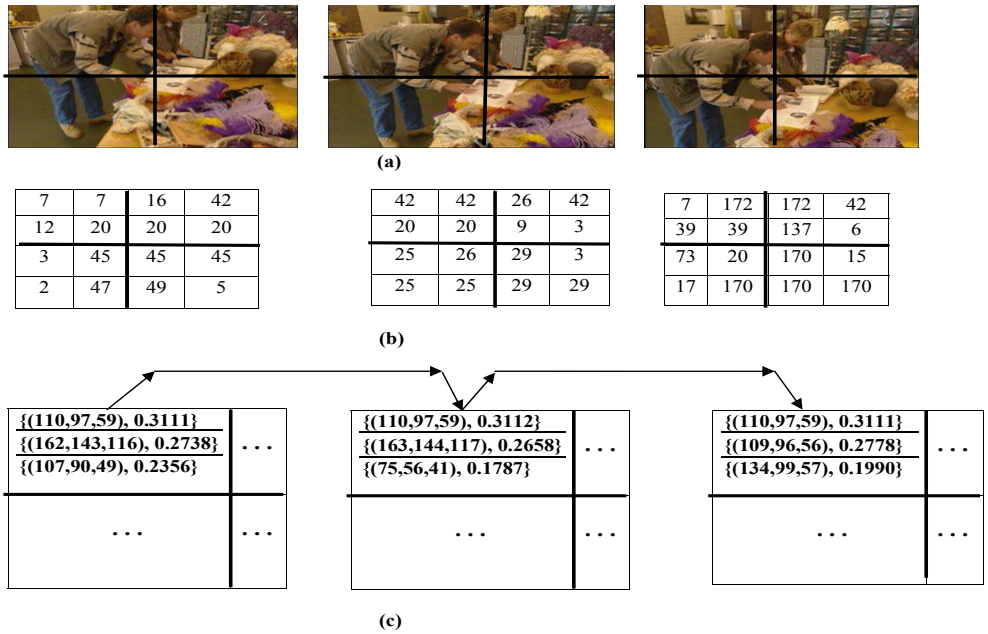


Figure 3.16: Spatio-Temporal DCDs extraction. (a) A video clip with three frames partitioned into 2×2 regions. (b) RGB-Feature Image of each region. (c) Comparing dominant color features of each region with time series

Audio fingerprints generation

In this proposed framework, Mel-Frequency Cepstral Coefficients (MFCCs) are employed, to obtain the acoustic fingerprints from the video contents. MFCCs are highly robust and discriminative spectral features; thus, they are widely used in video indexing and segmentation methods (Boreczky and Wilcox 1998). In the mel-frequency cepstrum, the frequency bands are equally spaced on mel-scale, which closely approximates response of the human hearing systems with respect to different frequencies (Park 2010; Wang et al. 2000). This frequency warping allows better representation of sound signals in automatic speech recognition applications. Further, MFCCs represent perceptual feature of the audio signals very well. Therefore, it is quite difficult to alter MFCCs, even for manipulations such as mp3 compression. Due to these reasons, perceptually robust MFCCs are utilized to extract the audio signatures of video contents.

Figure 3.17 shows the block diagram of audio fingerprints generation technique based on MFCC features. Specifically, first the audio signal is down sampled and consequently segmented into 11.60ms windows using Hamming windowing with an overlap factor of 76% (Roopalakshmi and Reddy Sep-2011). Further, important perceptual audio descriptors present in the frequency domain (Wang et al. 2000). There-

Table 3.20: Number of dominant color features extracted -A comparison

S.No	Duration (in minutes)	Total DCDs extracted		Reduction (in %)
		Spatial DCD	Spatio-temporal DCD	
1	1-10	3198	1284	59.849
2	11-20	19902	9680	51.361
3	21-30	29502	16818	42.993

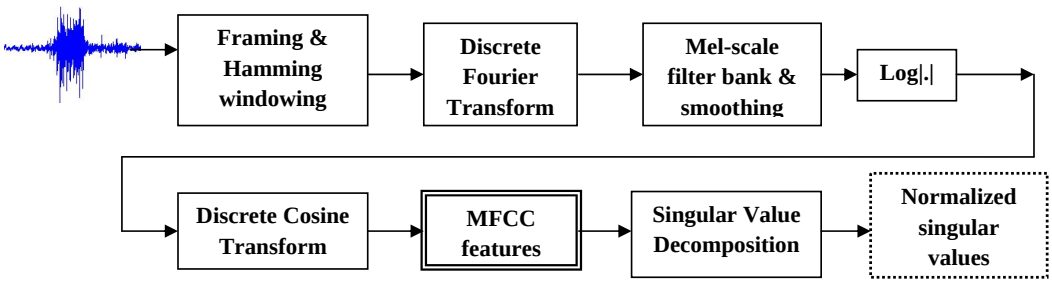


Figure 3.17: Block diagram of MFCCs extraction

fore, spectral representation of each analysis window is computed by applying Fourier transformations. Let f_i be the frame, that is partitioned into N equal segments of length M, denoted as $f_{i,q}$, where $q = 1, 2, ..., N$ (Chen et al. 2011; Özer et al. 2005). The M points (DFT) of the input audio signal AS is computed as ,

$$AS(k) = \sum_{i=0}^{M-1} f_{i,q}(i) e^{-j2\Pi ik/M}, \quad 0 \leq k \leq M-1 \tag{3.37}$$

Then a filterbank of T triangular filters is defined, which is denoted as $H_m k$, m=1,2,...,T. The log-energy spectrum present at the output of each filter is given as,

$$\Psi(k) = \ln \left[\sum_{k=0}^{M-1} |AS(k)|^2 H_m(k) \right], \quad 1 \leq m \leq T \tag{3.38}$$

Finally, the Mel-frequency cepstrum is computed as the DCT of the T filter outputs as given by,

$$c(n) = \sum_{m=1}^{T} \Psi(m) cos \left[\frac{\pi n(m-0.5)}{T} \right], \quad 1 \leq n \leq T \tag{3.39}$$

Equation (3.39) typically results in 24-40 MFCC terms. For speech signals the first 13 cepstrum coefficients are often utilized (Chen et al. 2011); hence, first 13 features are considered for the proposed CBCD task.

The MFCCs calculation in Equation (3.39) results in an F×N matrix, where F rows indicate the number of frames and N consists of 13 MFCC features extracted from a frame. Then singular value decomposition (SVD) is applied to effectively summarize the MFCC feature matrix. Precisely, the F×N matrix is decomposed as $S = U \sum V^T$, where S is F×N input matrix to be summarized, U is F×F orthogonal matrix, $\sum$ is an F×N diagonal matrix consisting of the singular values of S and V is N×N orthogonal matrix. Generally few larger singular values are utilized to provide the summarization of matrix S. The proposed CBCD framework employs six to eight singular values for the copy detection task.

In addition, normalized singular values are utilized in this study, in order to improve the robustness of the audio signature against various media manipulations such as band compression. Precisely, normalized singular values are exploited as acoustic fingerprints, which are computed as follows,

$$\delta_i = \frac{s_i - s_{min}}{s_{max} - s_{min}} \tag{3.40}$$

where δ_i is the normalized value of i^{th} singular value s_i. The s_{min}, s_{max} indicate minimum and maximum singular values respectively. Figure 3.18 shows the normalized singular value curves of the master and copied videos, in which the duplicate video is created by applying Mp3 compression at a bit rate of 64kbps. The curves plotted

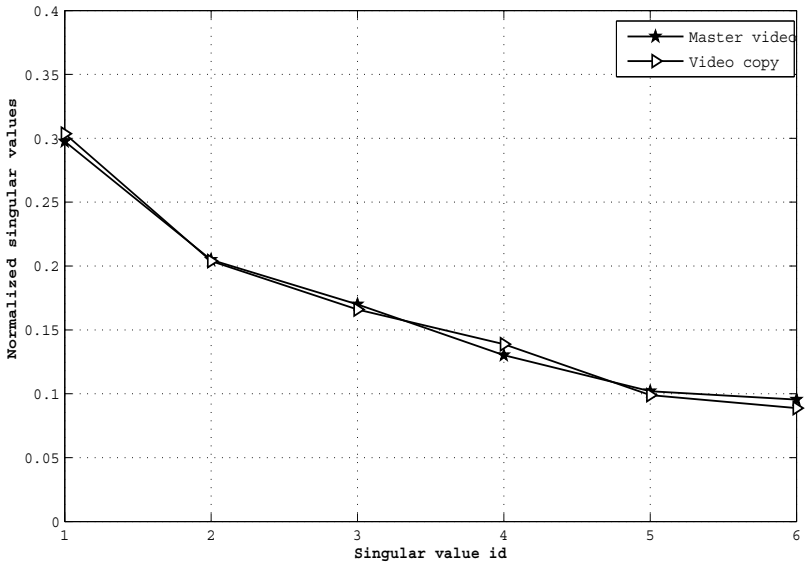


Figure 3.18: Curves showing similarity between the normalized singular values of master and duplicate video sequences.

in Figure 3.18 indicate a very high similarity between the master and duplicate video

contents; hence, prove the perceptual robustness of the proposed audio fingerprints.

3.4.3 Fusing visual-audio fingerprints

In the proposed CBCD framework, two strategies are employed for fusing visual-audio fingerprints in order to detect duplicate videos, namely combination rule and weighting factor. The purpose of the combination rule is to choose the best matching result, by integrating the independent visual and audio content similarity matches. *Further, in some duplication/copy cases, the audio content may be unavailable, either partially or completely destroyed*; hence, weight factors are set to visual-audio fingerprints, to indicate their reasonable contribution in the similarity matching task. Figure 3.19 indicates the flowchart illustrating the fusion of visual-audio fingerprints. The similarity matching of visual and acoustic fingerprints and the fusion technique are detailed below.

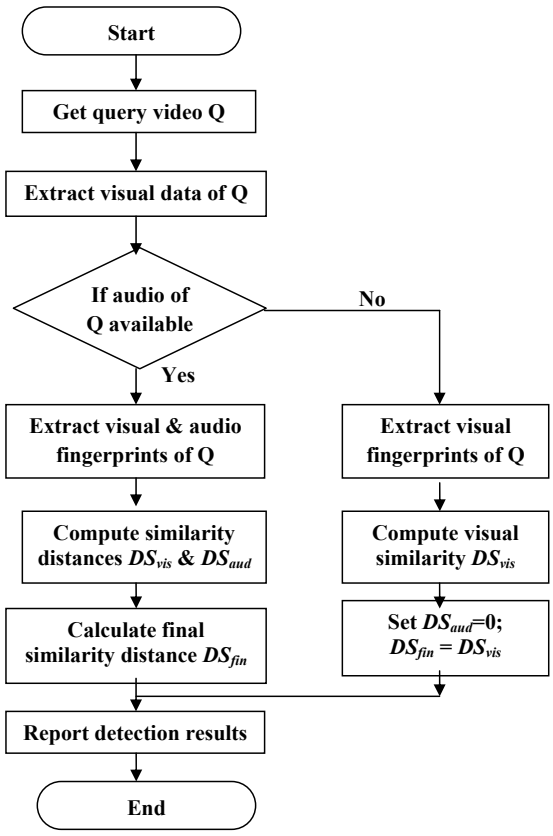


Figure 3.19: The flowchart showing fusion of visual-audio fingerprints

Similarity matching of visual fingerprints

In this step, the database is searched for similar dominant color distributions same as the input query. Let MV and QV be master and query video clips, VF_m and VF_q are their corresponding visual fingerprints which are given by,

$$VF_m = \{(c_i, p_i) \mid i = 1, 2, 3, ..., M\}, \tag{3.41}$$

$$VF_q = \{(b_j, q_j) \mid j = 1, 2, 3, ..., N\}, \tag{3.42}$$

where M, N indicate the total DCDs of MV and QV clips respectively. The similarity DS_{vis} between the visual signatures of the MV and QV is computed as specified in MPEG-7 standard (Manjunath et al. 2002) as follows,

$$DS_{vis}(VF_m, VF_q) = \sum_{i=1}^{M} p_i^2 + \sum_{j=1}^{N} q_j^2 - \sum_{i=1}^{M}\sum_{j=1}^{N} 2a_{i,j}\ p_i\ q_j \tag{3.43}$$

Here $a_{i,j}$ is the similarity coefficient between two color vectors c_i and b_j, which is computed as described in Equation (3.5) as follows,

$$a_{i,j} = \begin{cases} 1 - \frac{d_{i,j}}{d_{max}} & if\ d_{i,j} \leq T_d \\ 0 & if\ d_{i,j} > T_d \end{cases} \tag{3.44}$$

where $d_{i,j}$ is the Euclidean distance between color vectors c_i and b_j. The threshold T_d is the maximum distance used to judge whether two colors are similar or not. The distance $d_{max} = \alpha \times T_d$, where α is set to 1.2, as specified in (Deng et al. 2001). The threshold T_d mainly depends upon the distortions applied to the query video and in case of combined distortions the distance values would be higher. Therefore, the selection of T_d significantly affects the detection performance of the proposed system. In order to handle this issue, the reference dataset is tested with different T_d values ranging from 0.15-0.69. The T_d values in the range of 0.29-0.36 are yielding better results for the transformations considered in the proposed framework.

Similarity matching of audio fingerprints

Let the acoustic fingerprints AF_m and AF_q of master and duplicate video sequences are given by,

$$AF_m = \{\sigma_k|\ k = 1, 2, 3, ..., P\}, \tag{3.45}$$

$$AF_q = \{\lambda_l \mid l = 1, 2, 3, ..., R\}, \tag{3.46}$$

where P, R indicate the total singular values of MV and QV clips respectively. Then, the similarity DS_{aud} between the acoustic signatures of master and the duplicate videos is calculated using Manhattan distance measure as follows,

$$DS_{aud}(AF_m, AF_q) = \sum_{k=1}^{P} \sum_{l=1}^{R} |\sigma_k - \lambda_l| \tag{3.47}$$

Decision fusion

The similarity results from visual-audio fingerprints are combined to detect video copies using the combination rule given by,

$$DS_{fin} = W \times DS_{vis} + (1 - W) \times DS_{aud} \tag{3.48}$$

where, DS_{fin} is the final similarity distance and W is the weighting factor. In the proposed framework, the visual and audio weight factors are empirically set as 0.65 and 0.35 respectively. Thus, the final similarity distance DS_{fin} between the master video MV and query clip QV is given by,

$$DS_{fin}(MV, QV) = 0.65 \times DS_{vis} + 0.35 \times DS_{aud} \tag{3.49}$$

3.4.4 Experiments and performance evaluation

The experimental setup of the proposed CBCD system including master database and query dataset construction followed by the detection results is illustrated as follows.

Master database and query dataset construction

As described in Section 2.8, the proposed method is evaluated on TRECVID Sound & Vision data set, which is widely popular in the CBCD domain (Küçüktunç et al. 2010). Precisely, the master database comprises 40h of TRECVID-2008 Sound & Vision data, plus another 60h of TRECVID-2009 Sound & Vision data, which covers a wide variety of contents. All the video sequences are transformed into the uniform format: 352×288 pixels and 15fps. Resampling procedure is employed to synchronize the frame rates of master and query video clips. For example, 2s query clip with 60fps becomes a 240-frame sequence after implementing the resampling procedure.

Seventeen video manipulations listed in Table 3.21 are considered in the proposed CBCD framework, which commonly occur in illegal video contents. Thirty video sequences are randomly selected from the master database, where the duration of

these clips vary between 32-45 seconds. Five video clips, collected from Open Video Project are used as non-reference data, in order to test false positives. The seventeen types of transformations listed in Table 3.21 are applied to the query dataset for generating final query video clips. The resulting 595 (35×17) video sequences are used as query clips for the proposed copy detection task. Each query clip is used to detect the corresponding video sequence in the master database.

Accuracy comparison

The following seven methods are evaluated:

(1) Spatio-temporal DCD (abbreviated as 'ST-DCD');
(2) MFCCs-based signature ('MFC');
(3) Combination of methods (1) and (2)('ST-DCD+MFC');
(4) Cao and Zhu's method (2009) ('CZ');
(5) Hua et al.'s method (2004) ('HUA');
(6) Itoh et al.'s method (2010) ('IT');
(7) Hua et al.'s method+Itoh et al.'s method ('HUA+IT').

Table 3.21: Transformations used in the proposed system using DCD & MFCCs

#	Type	Description
T1	Random noise	Add 19% gaussian noise
T2	Blurring	Blurring by 26%
T3	Rotation	Rotating by 20° to 45°
T4	Brightness change	Increase brightness by 19% -25%
T5	Flip	Horizontal flip by 90°-100°
T6	Color change	Change color spectrum
T7	Pattern insertion	Insert text pattern into selected frames
T8	Moving caption	Insert moving titles into entire video
T9	Zooming	Zoom in to the frame by 17%
T10	Slow motion	Halve the video speed
T11	Fast forward	Double the video speed
T12	Mp3 compression	Change audio file format
T13	Single band comp.	Compress only specific frequency band
T14	Multiband comp.	Compress different frequency bands independently
T15	Combination of 3	15% noise, 20% blurring & 15% brightness
T16	Combination of 5	17% noise, 21% blurring, 15% brightness, rotation & pattern insertion
T17	Combination of 8	19% noise, 25% blurring, 15% brightness, color change, pattern insertion, moving caption, fast forward & 15% zoom

The methods (1),(2) and (3) include different combinations of the proposed techniques. In method (1), dominant color descriptors are extracted in a spatio-temporal

manner, using the procedure explained in Section 3.4.2 and utilized for the copy detection task. In method (2), MFCCs based audio signatures are employed for detecting video copies. In method (3), spatio-temporal dominant color features and MFCCs-based audio fingerprints are integrated with weighting factors, to detect duplicate video clips.

Cao and Zhu's method (2009) is based on YCbCr components of images. First, mean and the sum of weighted mean of YCbCr values are computed to generate image signatures. Then the image signatures and temporal order of the images are used to construct signatures of video sequences. In this method, L1-norm distance is used to calculate the similarity between two video clips.

Hua et al.'s method (2004) uses Ordinal measure, which is a widely popular visual fingerprint in the CBCD domain (Kim and Vasudev 2005). It is implemented as follows: Video frames are partitioned into 3×3 blocks and the corresponding average intensities of blocks are computed. From the block intensities, a 9-D ordinal signature reflecting the frame's relative intensity distribution is computed. Then the temporal shape of ordinal signatures is computed and sequence shape similarity algorithm is employed to detect duplicate video sequences.

Itoh et al.'s method (2010) uses acoustic power based fingerprints for the CBCD task and it is implemented as follows: First power of audio signals are calculated using sliding window scheme and acoustical power envelopes are generated by using power vs time sequences. Then the significant points indicating local minimum/maximum values are extracted from the power envelopes and matched using dynamic programming for the copy detection task.

In method (7), for the purpose of comparison, methods (5) and (6) are combined, to detect duplicate videos, which is implemented as follows: Ordinal measure is utilized as a visual signature, which is specified in method (5). Audio fingerprints extracted from acoustic power features are employed, as specified in method (6). Visual and audio fingerprints are matched separately using L1-norm distance, in order to measure the similarity between two video contents.

Detection accuracy

Table 3.22 lists the precision and recall rates of seven compared methods for T1-T6 transformations. The transformations include random noise, blurring, rotation, brightness change, flip and color change. The bold font indicates the highest precision and recall scores in the table. Method (3) provides better results compared to the reference methods for all seven transformations in terms of higher PR rates.

Table 3.22: Copy detection results for T1-T6 transformations

Attacks		**ST-DCD**	**MFC**	**ST-DCD+**	**CZ**	**HUA**	**IT**	**HUA**
Type	**Metric**	**(1)**	**(2)**	**MFC(3)**	**(4)**	**(5)**	**(6)**	**+IT(7)**
T1	P	64.18	90.64	**91.59**	45.68	60.43	80.64	82.91
	R	70.54	88.31	**89.04**	40.91	58.09	75.42	76.55
T2	P	60.92	90.51	**91.35**	51.53	58.13	79.68	81.41
	R	62.37	89.59	**90.56**	57.11	57.46	71.53	70.65
T3	P	71.68	91.86	**93.83**	67.34	51.65	82.37	84.09
	R	62.59	90.75	**90.76**	62.28	49.72	84.44	85.39
T4	P	74.45	89.91	**93.51**	50.16	62.91	72.86	79.16
	R	72.16	90.88	**91.69**	53.55	61.49	74.93	76.66
T5	P	79.36	91.36	**92.97**	68.04	48.59	79.66	79.98
	R	74.18	88.45	**89.10**	61.95	50.47	76.08	77.46
T6	P	34.02	90.54	**90.59**	49.62	50.51	80.52	82.73
	R	30.47	89.35	**89.56**	50.38	48.68	80.06	81.64

Methods (2) and (6) generally perform well in terms of good PR rates. This is because, audio fingerprints are less affected by visual attacks. The recall rate of Method (1) declines sharply for color change type. This is because color spectrum changes might alter DCD's property substantially. However, because of the integrated usage of MFCCs, proposed method (3) gains more robust performance than the reference methods in this transformation category.

Cao and Zhu's method (2009) scores well for rotation and flip types in terms of good precision rates. However, it fails to score better PR rates for random noise (T1) type, because luminance (mean YCbCr) values vary widely after applying gaussian noise. Hua et al.'s method (2004) gives poor precision value for flipping (T5) type. This is because, block intensities are much affected by mirroring the contents, which may significantly change ordinal signatures.

Method (7) yields better PR rates for all T1-T6 types compared to all other methods, except method (3). However, proposed method (method (3)) outperforms method (7) by improving the detection accuracy in terms of up to 15% enhancement in PR rates. Though the spatio-temporal DCDs and MFCCs-based audio signatures have their own benefits and constraints, they balance each other very well; hence, the integration of these two robust features results in consistent performance of method (3) compared to other methods. The detection accuracy of method (3) provides a good evidence to support this viewpoint.

Table 3.23 shows the detection results of seven compared methods for T7-T12

types, including pattern insertion, moving caption, zooming, slow motion, fast forward and mp3 compression. Cao and Zhu's method (2009) yields low precision rates for zooming and pattern insertion types. This is due to the limited capability of global descriptors, which are less robust against region based attacks.

Table 3.23: Copy detection results for T7-T12 transformations

Attacks		**ST-DCD**	**MFC**	**ST-DCD+**	**CZ**	**HUA**	**IT**	**HUA**
Type	**Metric**	**(1)**	**(2)**	**MFC(3)**	**(4)**	**(5)**	**(6)**	**+IT(7)**
T7	P	69.38	89.46	**90.52**	49.05	54.61	68.92	81.35
	R	65.49	85.37	**88.04**	51.43	53.35	64.84	70.27
T8	P	66.05	87.28	**90.06**	51.89	55.10	66.30	71.86
	R	60.37	86.05	**87.42**	51.48	52.85	65.85	69.12
T9	P	72.34	90.53	**91.73**	50.54	59.18	79.16	83.49
	R	70.63	88.14	**89.92**	56.33	50.78	74.58	81.61
T10	P	64.57	79.52	**82.54**	54.19	60.84	64.52	68.94
	R	62.45	73.91	**79.68**	51.75	60.79	61.08	63.27
T11	P	71.28	74.68	**79.24**	50.61	65.15	69.35	70.86
	R	69.37	72.43	**75.54**	55.38	62.23	64.16	66.19
T12	P	88.62	70.66	**90.13**	71.55	72.06	59.34	75.13
	R	85.05	67.51	**85.92**	70.93	70.45	47.23	77.46

Hua et al.'s method (2004) performs poorly for pattern insertion and moving caption types in terms of recall rates. The reason is that, their method generates very different video signatures for master and query clips after inserting text patterns or moving captions. Even the shape similarity scheme cannot compensate for the large differences between the query and master video signatures; hence results in poor recall rates. For slow motion (T10) and fast forward (T11) types, methods (1)-(3) utilize resampling technique to synchronize the master and query video contents; hence these methods achieve high detection accuracy compared to other reference methods. On the other hand, for slow motion and fast forward types, methods (4)-(7) match the master and query sequences directly without resampling.

The recall rate of Itoh et al.'s method (2010) declines sharply for mp3 compression type. This is because, the acoustical power envelopes vary substantially after applying mp3 compression. Yet, the proposed methods (methods (2) and (3)) using MFCCs are less affected in this category. Method (7) yields nearly similar accuracy rates for all T7-T12 types compared with the proposed method(3). However, mp3 compression has a limited impact on MFCCs, hence method(3) reduces false positive rate effectively and achieves good accuracy compared to method(7).

Table 3.24 lists the copy detection results of seven compared methods for T13-T17 types including single, multi band compressions and combination of multiple visual attacks. Cao and Zhu's method (2009) performs good for single and multiband compression types; however, it fails to score better results for combined attacks. This is because, adding the gaussian noise and changing color spectrum would substantially vary luminance values.

Table 3.24: Copy detection results for T13-T17 transformations

Attacks		**ST-DCD**	**MFC**	**ST-DCD+**	**CZ**	**HUA**	**IT**	**HUA**
Type	**Metric**	**(1)**	**(2)**	**MFC(3)**	**(4)**	**(5)**	**(6)**	**+IT(7)**
T13	P	87.34	75.28	**90.71**	75.64	77.65	56.26	81.69
	R	82.92	76.24	**84.46**	72.91	79.52	52.79	80.91
T14	P	85.43	73.46	**89.63**	70.33	76.51	50.33	79.34
	R	81.88	71.05	**80.86**	67.85	72.97	48.81	76.88
T15	P	57.46	85.64	**87.38**	50.11	54.09	78.86	80.62
	R	51.26	81.27	**82.22**	45.20	51.12	81.94	81.47
T16	P	45.12	80.59	**80.78**	38.51	40.56	71.34	75.67
	R	43.84	77.16	**77.39**	41.62	42.81	70.89	72.62
T17	P	31.76	79.44	**80.05**	28.99	30.94	70.25	76.92
	R	32.58	72.38	**72.89**	30.34	31.55	66.53	69.85

It is observed that, the performance of Hua et al.'s method (2004) is severely degraded in T16 and T17 types in terms of low precision rates. This is because, noise and pattern insertions generate different video fingerprints for master and query video clips; hence, more number of false positives are retrieved from the master dataset, which results in low precision rates. Itoh et al.'s method (2010) generally scores well for all five transformations except for multiband compression type in terms of low recall rate. This is because, different local minimum/maximum significant points are produced from the power envelopes of query and master contents after applying multiband compression type.

Among all the methods, method (3) and (7) achieve better accuracy rates for all T13-T17 types. However, because of the combined utilization of spatio-temporal dominant color features and cepstral features, method (3) is more accurate and provides better PR rates compared to method (7). The accuracy rates of method (3) shown in Table 3.21 are evident to support this viewpoint.

The experimental results indicate that, for all seventeen transformations, method (3) consistently outperforms all the reference methods. The integrated usage of two complementary features namely, spatio-temporal DCDs and MFCCs to detect du-

plicate videos, is the exact reason for the effective performance of method (3). To summarize, the experimental results prove that, the combination of two complementary features in method(3) not only enhances the detection performance, but also widens the coverage to more number of video transformations.

3.5 CBCD System Using Motion Activity and Spectral Descriptors

It is known that, motion activity features contribute an essential information about a video content. On the other hand, it is stated in the CBCD literature that, copy detection using only visual features may not be efficient against a wide variety of transformations (Küçüktunç et al. 2010). Therefore, exploiting multimodal features for the copy detection task, first enhances the detection accuracy and then consequently extends the coverage to more number of video modifications. By considering these aspects, this thesis contributes a novel CBCD system, which employs visual fingerprints derived from motion activity features and acoustic fingerprints extracted from spectral descriptors, in order to identify the illegal video sequences. Specifically, the contributions of the proposed copy detection system are given by,

- Novel copy detection system using visual-audio features is introduced, compared to existing visual feature-based CBCD schemes.
- An algorithm for computing the spatial distribution of motion activity in terms of number of active regions is presented.
- Informative *Motion Activity (MA)* and *Spectral Descriptive (SD) words* are computed, which efficiently describe a given video content.
- Similarity matching using Clustering to speed up the matching task.

The framework of the proposed CBCD system, including the extraction of motion activity features and audio spectral descriptors followed by the fingerprints fusion is illustrated as follows.

3.5.1 Proposed CBCD system using motion & audio features

The proposed copy detection framework is shown in Figure 3.20 and the relevant notations are described in Table 3.25.

Initially, motion activity and audio spectral features are extracted from the master and pirate video sequences. More precisely, motion activity features include number

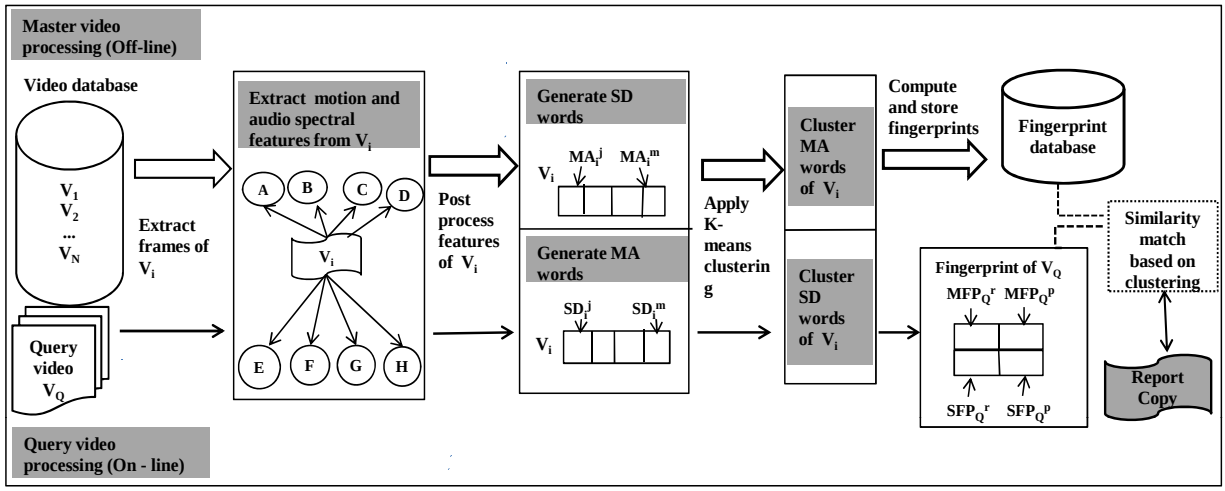


Figure 3.20: Proposed CBCD framework using motion activity & audio features

Table 3.25: Description of notations used in Figure 3.20

Notation	Description	Notation	Description
N	Number of master videos	G	Audio spectral roll-off value
V_i	i-th master video in database	H	Spectral flux of audio signal
A	Intensity of motion activity	MA_i^j	j-th MA word of V_i, j ={1,..,m}
B	Number of active regions	SD_i^j	j-th SD word of V_i, j ={1,..,m}
C	Dominant direction of activity	SFP_Q^r	r-th spectral fingerprint of V_Q
D	Mean motion vector magnitude	MFP_Q^r	r-th motion fingerprint of V_Q
E	Audio spectral centroid	m	Number of MA/ SD words of V_i
F	Energy of spectrum	p	Number of video signatures of V_Q

of active regions, motion intensity, dominant direction of activity and the standard deviation of motion vector magnitude of a frame; while the audio spectral descriptors include spectral centroid, energy, roll-off and flux. Then the resultant motion activity and spectral features are further processed and the corresponding *Motion Activity (MA)* and *Spectral Descriptive (SD) words* are computed.

MA words comprehensively represent overall motion activity, whereas SD words summarize the audio profile of video sequences. To obtain the low-dimensional representation of MA and SD words of video sequences, K-means clustering algorithm is utilized. The resulting cluster centroids are considered as fingerprints of video sequences. After this step, clustering based pruned search is performed and the copy detection results are reported.

3.5.2 Video fingerprints extraction

The proposed CBCD framework jointly exploits the motion activity features and audio spectral signatures for detecting the duplicate video clips, which are extracted as follows.

Motion activity features extraction

In any video content, the motion activity spans from high to low levels. As mentioned in Section 3.2.1., MPEG-7 Motion activity descriptor captures pace of action or intensity of activity in a video segment (Jeannin and Divakaran 2001). As described in Equation (3.11), the motion activity descriptor includes the following four attributes given by, motion intensity (I), dominant direction of activity (Dir), spatial distribution of activity ($Spatial$) and temporal distribution of activity ($Temporal$).

Motion Intensity (I): This attribute represents an effective temporal description of a video shot in terms of different intensity levels (Sun et al. 2001). In the proposed framework, SMV of blocks are utilized for computing motion intensity, as specified in Equations (3.12) and (3.13). MPEG-7 standard defines motion activity values ranging from 1-5, based upon the respective SMV values (Jeannin and Divakaran 2001), in which the quantization thresholds are recommended mainly for MPEG-1 videos, as shown in Table 3.9. However, the reference video database includes different file formats such as MPEG-1 and MPEG-4. Therefore, different threshold values are experimented and finally the threshold values given in Table 3.26 are utilized for computing the motion intensity in the proposed copy detection system.

Table 3.26: New quantization thresholds used in proposed CBCD task

Activity value	Range of SMV
1	$0 \leq$ SMV <2.5
2	$2.5 \leq$ SMV < 9.7
3	$9.7 \leq$ SMV <16.1
4	$16.1 \leq$ SMV <24.4
5	$24.4 \leq$ SMV

Dominant Direction of Activity (*Dir*): In the proposed CBCD framework, the approximate dominant directions of motion activity are computed in order to improve the robustness of the copy detection task. Specifically, in the proposed system, the direction vector (Dir), which indicates the total amount of motion activity in four

major directions including left, right, down and up is computed using the Equations (3.16)- (3.20) respectively. The highest value of Dir provides the dominant direction of motion activity in a given frame.

Spatial Distribution of Activity (*Spatial*): This attribute represents, whether the activity is confined to one region or spread across multiple regions (Savakis et al. 2003). As mentioned in Section 3.2.1., the partition of a frame into $k \times k$ regions, plays an important role in predicting the actual no. of active regions in the given frame. However, smaller values of k eliminates important semantic content, while larger values of k increases computational load. To tackle this problem, the data set is experimented with different k values ranging from 2 to 5. Maximum accuracy rate (85.36%) is achieved at $k = 3$; therefore, spatial distribution of activity of frames is computed by segmenting the frames into 3×3 regions. The Algorithm 3.1 described in Figure 3.5 is enhanced and described in Algorithm 3.4 of Figure 3.21, which accurately computes the number of active regions. More precisely, Spatial Activity Matrix (SAM) and Mean Motion Distribution (MMD) calculations are elaborately explained in Algorithm 3.4 of Figure 3.21, which computes number of active regions in a frame.

Further, the direct processing of extracted raw motion activity features is computationally expensive; hence, K-means clustering is employed to obtain the compact representation of the resultant motion activity features. Furthermore, motion vectors provide sufficient feature description, when they are captured at lower frame rates (Tasdemir and Cetin 2010). Hence, experiments are performed with different frame rates ranging from 4 to 10 and in this proposed scheme 5 fps is used for extracting motion features because of its high detection performance.

Spectral features extraction

From the down sampled audio signal, 11.60ms windows with an overlap factor of 86% using Hamming window function (Roopalakshmi and Reddy Sep-2011) are generated, as described in Section 3.3.1. Then, spectral features such as centroid, roll-off, energy and flux are extracted from the short term power spectrum of audio signals, as specified in Equations (3.23)-(3.26) respectively.

Fingerprint matching

In the proposed system, motion activity and spectral features of video files are grouped into clusters using K-means clustering. Let R_k and Q be the k^{th} reference and query video clips; and mp_r and mp_q be their corresponding motion activity features. In sim-

Algorithm 3.4: Computing Number of Active Regions in a Frame

1: Calculate Spatial Activity Matrix (SAM) of frame F using the equation,

$$SAM(F) = \begin{cases} mv(i,j) & if\ mv(i,j) \geq AMV \\ 0 & otherwise \end{cases} \tag{3.50}$$

where $mv(i,j)$ is the motion vector magnitude of block (i,j), such that i = { 1, 2, 3, ..., M} and j ={ 1, 2, 3, ..., N}. This SAM computation of F retains only high activity blocks of F.

2: Segment $SAM(F)$ into 3×3 non overlapping blocks of size W×H, where W= $\frac{M}{3}$ and H= $\frac{N}{3}$. The motion activity of k-th region R_k of F is computed as,

$$R_k(F) = \sum_{x=1}^{W} \sum_{y=1}^{H} B_m(x,y) \tag{3.51}$$

where k ∈ {1, 2, 3, ..., 9} , B_m is the m-th block of R_k and m ∈ {1, 2, 3, ..., W×H}.

3: Extract mean motion distribution (MMD) of R_k of F as follows:

$$MMD(R_k(F)) = \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} B_m(x,y)}{W \times H} \tag{3.52}$$

4: Sort the MMD values of all regions of a frame in the ascending order.

5: Regions with higher MMD values ($MMD \geq \alpha \times AMV$, where α is set as 2.4) are treated as active regions of a given frame.

Figure 3.21: Algorithm to compute number of active regions in a frame

ilarity matching tasks, performance of the comparative Manhattan distance measure is better compared to the simple absolute distance measure. Therefore, the similarity Sim_{mo} between motion activity features of R_k and Q segments is computed using comparative Manhattan distance measure as given by,

$$Sim_{mo}(R_k, Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|mp_r(i) - mp_q(j)|}{|mp_r(i)| + |mp_q(j)|} \tag{3.53}$$

where m and n indicate size of motion fingerprints of R_k and Q respectively. Then the resultant Sim_{mo} scores are compared against the Confidence Measure(CM_1) for obtaining the matching results. The reference database is experimented with different confidence thresholds ranging from 0.50 to 0.75 to reduce false positive rates. Better detection results are obtained for 0.60 threshold, thus it is set as CM_1 in the proposed copy detection task.

Let sp_r and sp_q be audio spectral fingerprints of videos R_k and Q respectively. Different distance measures are tested on the experimental dataset and the results proved better for squared Euclidean distance measure. Thus, the similarity Sim_{sp} between spectral features of R_k and Q is computed as,

$$Sim_{sp}(R_k, Q) = \sum_{i=1}^{a} \sum_{j=1}^{b} (sp_r(i) - sp_q(j))^2 \tag{3.54}$$

where a and b indicate size of spectral fingerprints of R_k and Q respectively. The Sim_{sp} scores are evaluated against the confidence threshold (CM_2), which is set as 0.69 based on experimental results.

The final similarity score $Final_{ss}$ between R_k and Q is computed as,

$$Final_{ss} = \begin{cases} 1 & if\ Sim_{mo} \geq CM_1 \&\ Sim_{sp} \geq CM_2 \\ 0 & otherwise \end{cases} \tag{3.55}$$

Based upon $Final_{ss}$ scores, the copy detection results are computed.

3.5.3 Experimental setup and results

The experimental setup of the proposed copy detection framework, including reference and query dataset construction followed by the detection results are described below.

Reference database and query dataset construction

As mentioned in Section 2.8, TRECVID-2009 Sound & Vision dataset is utilized for evaluating the performance of the proposed copy detection framework. More precisely, the reference video database includes 200 hours of video covering a wide variety of content. The motion vectors are efficient, when they are captured at lower frame rates (Tasdemir and Cetin 2010); hence, different frame rates ranging from 4-10 are experimented, and 6fps is utilized in the proposed framework for extracting the motion features because of its high detection accuracy. Table 3.27 lists the visual and audio transformations used in the proposed CBCD system.

Table 3.27: List of transformations considered in the proposed CBCD system using motion and audio features

Type	Category	Description
T1	Fast forward	Double the video speed
T2	Slow motion	Halve the video speed
T3	Color change	Changing color spectrum
T4	Blurring	Blurring by 20%
T5	Brightness change	Increase brightness by 25%
T6	Noise addition	Add 15% random noise
T7	Pattern insertion	Insert text pattern into selected frames
T8	Moving caption insertion	Insert moving titles into entire video
T9	Cropping	Crop top & bottom frame regions by 15% each
T10	Picture-inside-picture	Insert smaller resolution picture into selected frames
T11	Mp3 compression	Change audio file format
T12	Single band compression	Compress only specific frequency band
T13	Multi band compression	Compress different frequency bands independently
T14	Combination of 3	Cropping by 18%, 20% of noise & moving caption

In the experiments, 45 video sequences are selected from the reference database and fourteen types of transformations listed in Table 3.27 are applied to those 45 videos, in order to generate query video clips. The resulting 630 (45×14) video sequences are treated as query clips for the proposed copy detection task, where the duration of these clips vary between 30-45 seconds. Each duplicate video is used to identify the corresponding video sequence in the master database.

Copy detection results and discussion

The following five methods are implemented for performance evaluation:

(1) Ordinal measure (Hua et al. 2004) (abbreviated as "OM");
(2) Tasdemir et al.'s method (Tasdemir and Cetin 2010) ("KT");
(3) Motion activity features ("MA");
(4) Audio spectral descriptors ("SD");
(5) Combination of motion activity and audio spectral features ("MA+SD").

The methods-(3), (4) and (5) include different combinations of the proposed techniques. More Specifically, in method (3), motion activity features including motion intensity, dominant directions and spatial distribution of activity are used for the copy detection task. In method (4), four spectral descriptors including spectral centroid, signal energy, spectral roll-off and spectral flux are considered for detecting video

copies. In method (5), both the proposed motion activity and spectral features are combined and evaluated for identifying the illegal video sequences.

Ordinal measure (method (1)) is one of the widely used global signature in the CBCD literature (Hua et al. 2004; Kim and Vasudev 2005), which is extracted as follows. Each video frame is partitioned into 4×4 blocks and normalized average intensity of blocks are calculated. The blocks are ranked in ascending order according to their resultant intensity values and ranking order of a block is known as the frame's ordinal measure. Euclidean distance metric is used to calculate the similarity between ordinal measure values of master and query video sequences.

Tasdemir et al.'s method (2010) is based on mean motion vector magnitudes of video frames, which is implemented as follows: First, frames are sampled at a rate of 5 frames/sec and motion vector magnitudes of macro blocks are extracted. Then, normalized average motion vector magnitude of frames are computed and stored as video signatures. The similarity between motion vector magnitudes of master and query clips are calculated using L2-norm distance measure.

Table 3.28 lists the detection results of five compared methods for T1-T5 transformations, which demonstrate the improved performance of method (5)(by 36.19%) when compared to the reference methods. For T2 transformation, method (4) yields

Table 3.28: Copy detection results (in %) for T1-T5 transformations

Transformations		**OM (1)**	**KT (2)**	**MA (3)**	**SD (4)**	**MA+SD (5)**
Fast forward (T1)	P	60.10	62.71	65.85	97.89	99.81
	R	61.54	69.64	84.37	96.37	97.03
	FM	60.81	65.99	73.96	97.12	98.40
Slow motion (T2)	P	71.35	71.87	75.00	90.02	90.02
	R	59.18	70.35	87.80	79.48	92.76
	FM	64.69	71.10	80.89	88.56	96.24
Color change (T3)	P	59.26	60.01	63.63	99.39	92.45
	R	67.79	69.27	84.83	99.40	97.62
	FM	63.23	64.30	72.71	99.39	98.79
Blurring (T4)	P	56.86	55.81	57.14	99.69	99.92
	R	79.14	72.56	92.30	92.35	91.46
	FM	66.15	63.09	70.58	99.84	99.95
Brightness change (T5)	P	56.93	70.15	82.85	90.14	91.18
	R	70.19	68.09	75.00	89.92	91.93
	FM	62.86	69.10	78.72	90.02	95.79

poor recall rate compared to method (5). The reason is, spectral features are much affected by temporal attacks such as slow motion. Method (2) yields poor preci-

sion rate for blurring (T4) transformation, because lot of false positives are retrieved from the data set. The global descriptive nature of ordinal measure results in better performance for blurring (T4) transformation when compared to method (2).

Table 3.29 indicates the detection results of five compared methods for T6-T10 transformations. Table 3.29 results demonstrate the enhanced performance of method (5) by 35.02%, compared with the reference methods. Method (1) performs poor for noise addition transformation, because random noise severely affects intensity values of blocks. Methods (2) and (3) also score poor PR rates for noise addition transformation due to the noisy nature of raw motion vectors. However methods (4) and (5) using spectral descriptors are less affected in this category and thus provide better detection results. For cropping attacks, Ordinal measure scores low results, because the surrounding black borders on frame regions noticeably increase the false positive rates.

Table 3.29: Copy detection results (in %) for T6-T10 transformations

Transformations		**OM (1)**	**KT (2)**	**MA (3)**	**SD (4)**	**MA+SD (5)**
Noise addition (T6)	P	41.69	42.17	50.00	88.56	92.68
	R	40.48	41.18	46.15	89.72	94.74
	FM	41.07	41.66	47.99	89.13	92.79
Pattern insertion (T7)	P	79.68	79.82	82.85	99.09	99.96
	R	80.24	79.47	85.29	98.80	98.85
	FM	79.95	79.64	84.05	98.94	99.97
Moving caption (T8)	P	70.64	74.58	82.50	97.42	98.25
	R	71.15	72.94	84.61	98.86	99.92
	FM	70.89	73.75	83.54	98.13	99.95
Cropping (T9)	P	68.29	65.83	80.00	99.00	99.98
	R	53.58	69.98	82.75	92.66	93.50
	FM	60.04	67.84	81.35	95.72	96.63
Picture-inside-picture (T10)	P	61.54	60.17	95.19	94.44	96.04
	R	43.67	66.73	74.54	90.26	99.01
	FM	51.09	63.28	85.41	92.30	99.50

Table 3.30 shows the detection results of five compared methods for T11-T14 transformations. Ordinal measure gives very poor detection results for T14 transformation (3 combined). The reason is, the global descriptors are less robust against region based attacks such as cropping. Motion activity features are less affected by audio transformations such as Mp3 and band compressions; hence, Method(3) provides better detection results for T11-T13, when compared to that of method (4). However, method (5) outperforms method(4) by yielding better PR rates for T11-T14 trans-

formations; joint utilization of motion and audio fingerprints for CBCD task, is the exact reason for this better performance of method(5). So, Table 3.30 results prove the improved accuracy of method (5) by 30.18% compared to the reference methods.

Table 3.30: Copy detection results (in %) for T11-T14 transformations

Transformations		**OM (1)**	**KT (2)**	**MA (3)**	**SD (4)**	**MA+SD(5)**
Mp3 compression (T11)	P	73.65	79.91	83.72	67.22	99.17
	R	72.58	68.29	70.51	57.29	97.35
	FM	73.11	73.64	76.54	61.85	98.25
Single-band comp. (T12)	P	80.06	79.62	83.62	73.44	98.37
	R	73.64	75.36	80.61	50.28	97.82
	FM	76.71	77.43	82.08	59.69	98.09
Multi-band comp. (T13)	P	66.28	69.16	85.05	70.36	91.74
	R	61.15	61.19	81.64	52.22	93.82
	FM	63.61	64.93	83.31	59.94	92.76
Combination of 3 (T14)	P	58.33	60.93	80.65	98.96	99.01
	R	51.29	64.74	79.83	98.21	99.29
	FM	54.58	62.77	80.23	98.58	99.64

Tables 3.28-3.30 results indicate that, the Method(4) using audio spectral descriptors is scoring very good results compared to the method(3), which is based on motion activity features. Specifically, method(4) achieves noticeably improved PR rates for transformations such as Color change and Blurring when compared to the methods(3) and (5). In other words, Table 3.28-3.30 results are favoring the audio spectral features for the CBCD task compared to the motion activity signatures. However, this observed phenomenon might be wrong due to these reasons:

i) First of all, audio spectral descriptors are not much affected by visual attacks such as color change and blurring; hence, method(4)('*SD*') performs well for all five transformations compared to the methods (3) and (5) ('*MA+SD*').

ii) on the other hand, motion activity signatures are greatly influenced by attacks such as Fast forward and slow motion, hence methods(3) and (5), employing motion activity descriptors provide poor scores compared to that of method (4).

iii) However, Mp3 and band compressions significantly alter the audio spectral descriptors, hence method (3) achieves poor PR rates for T11-T13 transformations, compared to the PR rates of the methods (3) and (5) respectively.

iv) Though motion activity features and acoustic signatures have their own benefits as well as limitations, they complement each other very well.

v) Therefore, the integrated utilization of motion activity and audio signatures for the CBCD task, not only improves the detection accuracy, but also covers more number of video transformations; The promising results of method (5) for all T1-T14 transformations provide good evidence to support this viewpoint.

Computational cost comparison

To evaluate the effectiveness of proposed method, experiments are conducted on a PC with 2.8GHz CPU and 3 GB RAM, where the code is implemented in MATLAB. The total computational cost of all five methods including signatures extraction and similarity matching are shown in Table 3.31. The computational costs are evaluated, based on detecting a 35s query clip within 50 hours of master database. The

Table 3.31: Comparison of computational cost

Computational Cost	**OM (1)**	**KT (2)**	**MA (3)**	**SD (4)**	**MA+SD (5)**
Signature extraction	170.95	223.04	187.98	111.56	196.41
Signature matching	57.43	42.02	34.38	1.27	35.68
Total cost	228.38	265.06	222.36	112.83	232.09

total computational cost of method (5) is slightly high compared to method (1). Although method (4) is the most cost effective method, its detection results are poor for audio transformations, when compared to that of method (5). Thus results prove that, method (5) significantly improves detection accuracy and widens the coverage to more number of transformations at the cost of slight increase in computational time. If feature extraction and matching procedures are implemented in parallel, then computational time of proposed scheme can be substantially reduced.

The experimental results prove that the proposed techniques improve detection accuracy by 30-35% compared to reference methods. Method (5), which combines motion activity and audio features, provides consistently good performance for all fourteen types of video transformations. The reason for the improved performance of proposed method is, the integrated usage of robust spectral and spatio-temporal motion activity features for the copy detection task.

3.6 Summary

This chapter discusses scholarly contributions towards the content-based video copy detection (CBCD) problem, in which content-based features such as visual, motion

activity and audio fingerprints are utilized for detecting video copies. More precisely, first this chapter introduces two CBCD schemes, that employ compact and computationally efficient visual fingerprints derived from Dominant Color Descriptors (DCDs) of MPEG-7 standard. Some of the significant contributions of the proposed CBCD systems are simple DCDs extraction scheme and adaptive signature pruning mechanisms. Though the two proposed CBCD schemes (here, DCDs-based schemes), are providing good PR rates; yet, they are less effective against transformations such as Color change and Camcording. To tackle this problem, DCDs based features could be integrated with motion activity or acoustic fingerprints, which in turn considerably improves the detection accuracy.

Second, this chapter proposes a novel copy detection system, which integrates different attributes of MPEG-7 Motion Activity Descriptor such as motion intensity, dominant direction and spatial distribution of activity for detecting illegal video sequences. Describing the spatio-temporal motion activity of a video sequence with the help of various attributes of Motion Activity Descriptor is one of the important contribution of the proposed copy detection system. However, motion activity features are less effective against transformations such as fast forward. Therefore, the joint utilization of motion activity features with visual as well as audio descriptors, might enhance the robustness of the copy detection task.

As mentioned in Section 2.1.3, state-of-the-art CBCD techniques are employing only visual features of videos for detecting video copies. However, audio content is an important information source of a video sequence, which is less affected in most of the illegal captures, compared to the visual data. To handle these issues, this chapter thirdly introduces, two CBCD methods, which utilize acoustic fingerprints derived from MFCC's and spectral descriptors for identifying the duplicate video sequences. Computing compact spectral descriptive words and clustering-based pruned similarity matching are some of the significant contributions of the proposed (audio fingerprints-based) CBCD techniques. However, audio signatures are less effective against modifications such as Mp3 and band compressions.

Further, if audio is available, then the joint utilization of visual-audio fingerprints for the CBCD task, not only enhances the copy detection performance, but also extends the coverage to more number of video attacks. Based on this aspect, this chapter fourthly introduces a robust CBCD framework, which employs visual fingerprints derived from DCDs and audio signatures extracted from MFCCs for detecting illegal videos. RGB-Feature image computation, spatio-temporal DCDs extraction and fusion strategies are some of the major contributions of the proposed CBCD framework.

Furthermore, this chapter fifthly introduces a novel CBCD system, which integrates motion activity features and audio spectral descriptors for detecting pirate video sequences. Computing number of active regions in a frame and motion activity words are some of the principal contributions of the proposed CBCD system. The two proposed CBCD frameworks (here, visual-audio fingerprints based methods), noticeably improve the detection accuracy and subsequently address more number of video transformations. However, due to the utilization of multimodal features, the fingerprint extraction cost of the proposed methods may be slightly high. To tackle this problem, fingerprints extraction and matching tasks could be implemented in parallel fashion, which in turn may considerably reduce the computational cost.

Experiments evaluated on different datasets such as TRECVID and Open Video Project datasets indicate the consistent performance of the proposed CBCD schemes compared to the reference methods against a wide variety of video transformations.

Related Publications

Conference/Symposium Publications

1) R. Roopalakshmi and G. Ram Mohana Reddy, *Recent Trends in Content-Based Video Copy Detection*, in proc. of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, pp. 1-5, Dec'2010. Available: `http://dx.doi.org/10.1109/ICCIC.2010.5705802`.

2) Roopalakshmi, R. and Reddy, G.R.M. *Compact and Efficient CBCD Scheme Based on Integrated Color Features*, in proc. of International Conference on Recent Trends in Information Technology (ICRTIT), Anna University, Chennai, India, pp. 880-883, June'2011.
Available: `http://dx.doi.org/10.1109/ICRTIT.2011.5972370`.

3) R. Roopalakshmi and G. Ram Mohana Reddy, *Efficient Video Copy Detection Using Simple and Effective Extraction of Color Features*, in proc. of International Conference on Advances in Computing and Communications (ACC-2011), Kochi, India, pp. 473-480, July'2011.
Available: `DOI:10.1007/978-3-642-22726-4_49`.

4) R. Roopalakshmi, G. Ram Mohana Reddy, *A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA*, in proc. of Second International Conference on Ambient Systems, Networks and Technologies (ANT-2011), Niagara Falls, Canada, 5, pp. 149-156, Sep'2011.
Available: `DOI:10.1016/j.procs.2011.07.021.`

5) Roopalakshmi, R. and Reddy, G.R.M. *A Novel CBCD Approach Using MPEG-7 Motion Activity Descriptors*, in proc. of IEEE International Symposium on Multimedia (ISM-2011), University of California, USA, pp. 179-184, Dec'2011. Available: http://dx.doi.org/10.1109/ISM.2011.36.

6) R. Roopalakshmi and G. Ram Mohana Reddy, *Towards a New Approach to Video Copy Detection Using Acoustic Features*, in proc. of IEEE 5th International Conference on Internet Multimedia Systems Architecture and Applications (IEEE IMSAA-2011), Indian Institute of Information Technology Bangalore (IIIT-B), India, pp. 1-5, Dec'2011.
Available: http://dx.doi.org/10.1109/IMSAA.2011.6156336.

7) R. Roopalakshmi and G. Ram Mohana Reddy, *Content-Based Video Copy Detection Using Motion Activity and Acoustic Features*, in proc. of International Symposium on Signal Processing and Intelligent Recognition Systems(SIRS-2014), Indian Institute of Information Technology and Management-Kerala (IIITMK), India, pp. 491-504, March'2014.
Available: DOI:10.1007/978-3-319-04960-1_43

Book Chapters

1) R. Roopalakshmi and G. Ram Mohana Reddy, *Efficient Video Copy Detection Using Simple and Effective Extraction of Color Features*, published in Springer Book titled, 'Advances in Computing and Communications', CCIS, Vol. 193, Part IV, Pages 473-480, 2011. ISSN: 1865-0929. Available:
http://link.springer.com/chapter/10.1007/978-3-642-22726-4_49.

2) R. Roopalakshmi and G. Ram Mohana Reddy, *Content-Based Video Copy Detection Using Motion Activity and Acoustic Features*, published in Springer Book titled, 'Advances in Intelligent Systems and Computing', Vol. 264, Pages 491-504, 2014. ISSN: 2194-5357. Available:
http://link.springer.com/chapter/10.1007/978-3-319-04960-1_43.

Journal Articles

1) R. Roopalakshmi, G. Ram Mohana Reddy, *A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA*, published in Elsevier Procedia Computer Science Journal, Vol. 5, Pages 149-156, 2011. ISSN: 1877-0509.
Available: http://dx.doi.org/10.1016/j.procs.2011.07.021

Chapter 4

Video Copy Tracking/Registration Methods

Tracking piracy requires pirate video detection followed by the exact frame alignments of the master and pirate video sequences, in order to estimate the geometric distortions and illegal capture location in a theater. This thesis describes the scholarly contributions towards the video copy registration problem in this chapter. Precisely, this chapter attempts to solve the shortcomings of the existing registration methods mentioned in Section 2.2.2., by presenting various video copy registration frameworks. The proposed registration methods exploit content-based multimodal finferprints for obtaining the spatial as well as temporal alignments of the pirate video with the master sequence, which are detailed below.

4.1 Temporal Registration of Video Copies Using Visual-Audio Features

This chapter first targets the temporal registration of a pirate video with the master sequence, by introducing a new temporal alignment scheme, which utilizes visual-audio fingerprints. More precisely, the main contribution of the proposed registration scheme is, to provide accurate frame-to-frame mappiings of the two video sequences by employing compact motion profile derived from motion vector magnitudes and audio profile extracted from MFCCs (Park 2010) . Further, the proposed framework also introduces a novel selection algorithm for selecting the most similar segment of master sequence with the help of segmentation-based dynamic programming technique, which noticeably reduces the frame matching cost. Furthermore, a new frame matching scheme exploiting multimodal features is contributed, which significantly reduces false

frame matches. The proposed temporal registration framework including motion and acoustic profiles extraction followed by frame matching based on multimodal fingerprints is illustrated below.

4.1.1 Proposed temporal registration framework

The block diagram of the proposed registration framework is shown in Figure 4.1, which comprises of two phases: First, compact motion and acoustic profiles are derived from the master and pirate video sequences. Second, the resultant temporal signatures of the two video sequences are aligned using dynamic programming, to achieve accurate frame-to-frame matches.

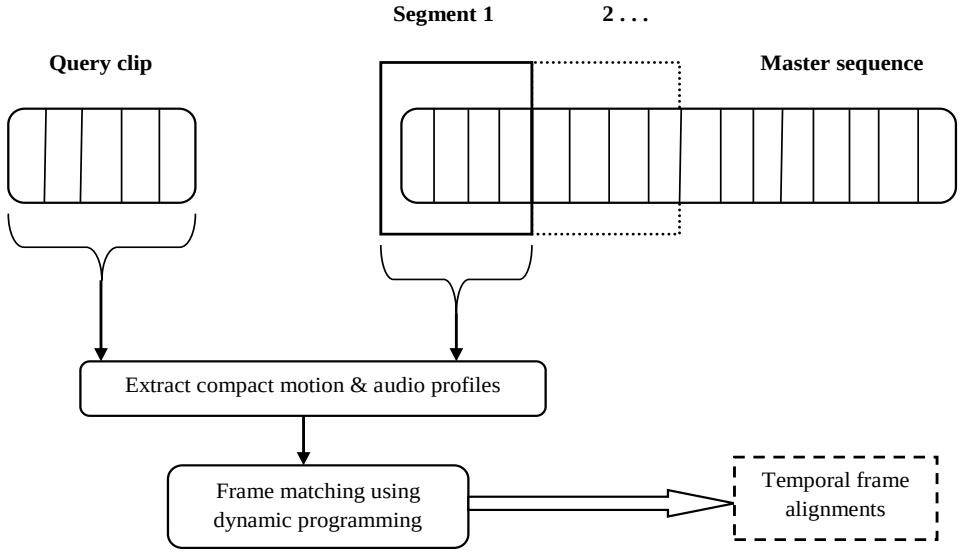


Figure 4.1: Proposed temporal registration framework using multimodal features

More precisely, when a duplicate video is given, the master sequence is divided into non-overlapping segments of size equal to the copy clip. Then, the similarity between the query clip and the windowed segment is computed using the 1-D signatures derived from motion and acoustic profiles of video sequences. The windowed segment with minimum dissimilarity score is indicated as the *candidate segment* of the master video and it is further analyzed using dynamic programming to get temporal frame-to-frame alignments of the two video contents.

Problem formulation: Let $M = \{x_i |\ i = 1, 2, ..., m\}$ be the master video sequence, where x_i is i^{th} frame of the master sequence. Let $Q = \{y_j |\ j = 1, 2, ..., n\}$ be the copy clip, where y_j is the j^{th} frame of the video copy video and $m >> n$. Here Q is derived

from M, after applying different video transformations such as blurring, rotation, scaling and Mp3 compression. Here, the goal is to determine the exact location of the subsequence $R = \{r_k |\ k = 1, 2, ..., i + n - 1\}$ in M, such that Q matches M and as a result frame-to-frame matches of Q and R can be obtained.

Fingerprints extraction

Direct comparison of feature sequences of two video contents is computationally expensive, since the features are multi-dimensional in nature. Therefore, in the proposed system, a video segment is compactly denoted using 1-D temporal signatures, which are easy to compute as well as robust against various video modifications. The compact temporal signatures including motion and acoustic profiles of video sequences are computed as follows.

Compact motion profile extraction: Motion vectors are popular temporal features; hence they are widely used in different video applications such as video summarization (Divakaran et al. 2001) and segmentation (Koprinska and Carrato 2001). However, as mentioned in Section 1.5., raw motion vectors are noisy in nature; therefore huge amount of information is needed for describing the motion content of a video. In addition, *motion vector magnitudes describe the temporal information of a video content, yet they fail to illustrate spatial distribution of motion activity in the given video sequence.* In order to tackle these issues, the proposed framework computes the compact motion profile of the video sequence, by integrating the temporal motion information and spatial distribution of motion activity. Specifically, the two attributes of MPEG-7 Motion Activity descriptor namely, motion intensity (I) and spatial distribution of activity (*Spatial*) as illustrated in Section 3.2.1 are exploited, for obtaining the motion profile of the video sequences. More specifically, as described in Equation (3.13), the Standard deviation of Motion Vector magnitude (SMV) of macro blocks are utilized to compute the motion intensity in a video frame. Further, spatial distribution of motion activity indicating the number of active regions in a frame, is computed using the Algorithm 3.4 specified in Figure 3.21. Then, the two resultant motion activity features are combined into 1-D motion signatures with the help I-order Z-curves and consequently denoted as motion profile of the video sequence.

Compact acoustic profile extraction: MFCCs are highly robust and discriminative features, thus they are widely used in video parsing and indexing applications (Tsekeridou and Pitas 2001). Further, MFCCs consider nonlinear property of the

human hearing system with respect to different frequencies, hence they are popularly used in automatic speech recognition systems (Wang et al. 2000).

In the proposed framework, MFCCs are computed using the discrete cosine transform of the log amplitude Mel-frequency spectrum, as illustrated in Section 3.5.2. However, raw MFCCs are noisy and may contain redundant data. To handle this discrepancy, MFCC variance values are employed to generate 1-D acoustic profile of video contents.

Introduction of dynamic programming: Dynamic programming is an effective recursive technique, which is widely popular in sequence alignments and comparison methods (Sankoff 2000). Precisely, the given two feature sequences can be optimally aligned using dynamic programming as follows:

a) *Computing minimum score matrix:* 2-D score matrix computation is needed for specifying the optimal alignment between the two sequences. Specifically, an element $SM(i,j)$ of score matrix SM provides minimum matching cost to match the subsequences [0,1,.., i] with [0,1,.., j], which can be recursively computed as,

$$SM(i,j) = Min \begin{cases} SM(i-1,j-1) \\ SM(i,j-1) + W_h \\ SM(i-1,j) + W_v \end{cases} + D(i,j) \tag{4.1}$$

where W_h, W_v are the penalties associated with horizontal and vertical directions respectively. The $D(i,j)$ is the difference between two feature sequences associated with the elements i and j.

b) *Determining the optimal alignment path:* A trace-back step starting from the diagonal element to the top left element is performed to compute the optimal frame-to-frame matches.

Segmentation-based dynamic programming: The computational complexity of dynamic programming to align two sequences of size a and b is $O(ab)$; hence, if the sequence size increases, the computational complexity also increases. In order to solve this problem, only the frame alignments between the pirate clip and the *candidate segment* are computed instead of the complete master sequence. Precisely, the *candidate segment* of the master sequence is selected using the Algorithm 4.1, specified in Figure 4.2, and it is aligned with the pirate video, so that the temporal frame alignments of two video sequences can be obtained.

Algorithm 4.1: Candidate Segment Selection

a: Segment the master sequence into consecutive non-overlapping blocks of length equal to the copy video.

b: Compute compact motion and audio profiles for each segment.

c: Let a master sequence $MS = \{S_i | i = 1, 2, ..., m\}$, where S_i is the i^{th} segment and m is the total segments of the master video. Here, each segment S_1 is represented using compact motion and audio signatures as,

$$S_1 = \{mf_i \cup af_j\},\ i = [1 : n];\ j = [1 : p] \tag{4.2}$$

where mf_i is the i^{th} motion based signature and af_j is the j^{th} MFCCs based signature of S_1 respectively.

d: Let the copy video CV is compactly represented using 1-D motion and acoustic signatures as,

$$CV = \{qmf_k \cup qaf_r\},\ k = [1 : n];\ r = [1 : p] \tag{4.3}$$

where qmf_k is the k^{th} motion based signature of CV and qaf_r is the r^{th} MFCCs based signature of CV.

e: The dissimilarity ($dsim$) between the query clip and the k^{th} segment of master sequence is computed using Manhattan distance as follows,

$$dsim(S_k, CV) = \sum_{i=1}^{n} |mf_i^k - qmf_i| + \sum_{j=1}^{p} |af_j^k - qaf_j| \tag{4.4}$$

where S_k is the k^{th} segment of MS. n and p indicate the size of motion and audio feature sequences of video contents respectively.

f: Select the segment with lowest $dsim$ value ($dsim$ <=threshold) as the candidate segment of master sequence. The threshold value is set as 0.38, after executing experiments for different values varying between 0.30 and 0.60.

Figure 4.2: *Candidate segment* selection algorithm

Frame alignments using visual-audio fingerprints

Once the *candidate segment* is selected, then the motion as well as audio feature sequences of query and candidate segments are matched separately using dynamic programming in order to achieve accurate frame-to-frame alignments, which is illustrated below.

Motion features based frame matching: Let CS be the candidate segment of master sequence and QS be the query segment. Let mf_k and qmf_k be the motion profiles of segments CS and QS respectively, such that $k = \{1, 2, ..., n\}$. The dissimilarity (Dis_{mot}) between the motion profiles of CS and QS segments is computed using comparative Manhattan distance measure as follows,

$$Dis_{mot}(CS(i),\ QS(i)) = \frac{|(mf_{(i)} - qmf_{(i)})|}{|(mf_{(i)})| + |(qmf_{(i)})|} \tag{4.5}$$

where $i = \{1, 2, ..., n\}$ and n indicates total motion signatures. Then, the score matrix SM is computed using Equations (4.1) and (4.5). After this step, the optimal alignments between CS and QS segments are computed and consequently the resultant frame-to-frame matches based on motion signatures (FM_{mot}) are calculated.

MFCCs based frame matching: Let af_k and qaf_k be the MFCCs based signatures of CS and QS segments respectively such that, $CS \in \{\ af_k|\ k = 1, 2, ..., p\}$ and $QS \in \{\ qaf_k|\ k = 1, 2, ..., p\}$, where p indicates total audio signatures. Then, the dissimilarity (Dis_{aud}) between audio signatures of CS and QS is computed using squared Euclidean distance as follows,

$$Dis_{aud}(CS(j), QS(j)) = |(af_{(j)} - qaf_{(j)})^2| \tag{4.6}$$

where $j = \{1, 2, ..., p\}$. After this step, the optimal alignments between audio profiles of CS and QS segments are computed using Equations (4.1) and (4.6). Then, the resultant frame-to-frame mappings based on audio signatures (FM_{aud}) are calculated.

Decision fusion: Final frame matches ($Final_{fm}$) between the query and candidate segments are computed as,

$$Final_{fm} = |FM_{mot} \cap FM_{aud}|. \tag{4.7}$$

where $Final_{fm}$ indicates the frames mapped by both the motion as well as audio fingerprints of CS and QS segments respectively.

4.1.2 Experimental setup

The proposed temporal registration framework is evaluated on TRECVID-2008 and 2009 Sound and Vision data sets. Precisely, the video database includes totally 100 hours of video (50 hours of 2008 data + 50 hours of 2009 data) covering a wide variety of content. All the video clips are converted into uniform format: 352×288 pixels and

5 frames/sec. Table 4.1 lists the different types of video transformations considered in the proposed registration framework such as geometric, temporal, filtering, audio and combined transformations. 50 video clips are randomly selected from the master

Table 4.1: Transformations considered in the proposed registration scheme

Transformation Category	Type	Description
Geometric	Rotation	Rotating by 15° to 25°
	Cropping	Crop top & bottom regions by 20% each
	Flipping	Horizontal flip by 20°-80°
Temporal	Fast forward	Double the video speed
	Slow motion	Halve the video speed
Pattern	Pattern insertion	Insert text pattern into selected frames
	Picture-in-picture	Insert smaller resolution picture
	Moving caption	Insert moving titles into entire video
Filtering	Blurring	Blurring by 28%
	Noise addition	Add 15% gaussian noise
	Contrast change	Increase contrast by 20%
Scaling	Zooming in	Zoom in to the frame by 13%
	Resolution change	Change frame resolution to 150×120 pixels
Audio	Mp3 compression	Change audio file format
	Single band comp.	Compress only specific frequency band
	Multi band comp.	Compress different frequency bands
Combined	3 combined	Cropping by 18%, 20% of noise & moving caption

video database, where the duration of these clips vary from 20-52 seconds. Seventeen types of video transformations listed in Table 4.1 are applied to the 50 selected video clips to generate the query dataset. The resulting 850 (50×17) video sequences are treated as query video clips for the proposed temporal registration task.

Overview of methods evaluated: The following six methods are implemented for evaluating the registration performance:

(1) The motion profile based matching without sliding window ('MV');
(2) The motion features based matching with sliding window ('MV+SW');
(3) MFCCs based matching without sliding window ('MFCC');
(4) MFCCs based matching with sliding window ('MFCC+SW');
(5) Chupeau et al.'s method (2006) ('CH');
(6) Motion profile+ MFCCs+ sliding window ('ALL');

The methods (1)-(4) and (6) evaluate different combinations of the proposed techniques. Methods (1) and (3) utilize different video fingerprints (namely motion profile

and MFCCs) to perform temporal alignment of two video sequences. Methods (2) and (4) are implemented to measure the effect of sliding window technique for the proposed registration task.

In method (1), the motion profile of the query clip is matched with the corresponding signatures of the entire master sequence (i.e. query clip is mapped with all segments of the master video). In method (2), sliding window mechanism is used to map the motion features of the query segment with the respective features of the candidate segment. 1-D MFCC signatures of query video are matched with the acoustic fingerprints of the entire master video sequence in method (3). In method (4), sliding window approach is utilized to match the query MFCC features with the respective features of the candidate segment.

Chupeau et al. (2006) employed color histograms for computing frame alignments between the query and master video sequences, which is executed as follows: color histograms (of size 512 bins) are obtained from consecutive video frames. Then, Euclidean distance between color histograms of consecutive frames are used as temporal fingerprints of video sequences.

Method (6) is implemented to assess the performance of proposed registration framework, that exploits multimodal features for aligning frames. In method (6), both the motion and MFCC signatures of query clip are matched separately with the corresponding features of the candidate segment in order to get accurate frame-to-frame alignments.

4.1.3 Registration results and discussion

The registration performance of six compared methods for different video transformations are discussed as follows.

Geometric and scaling transformations: Table 4.2 shows the registration accuracy of six compared schemes/methods for geometric and scaling categories, which include rotation, cropping, flipping, zooming in and resolution change transformations. Methods (3), (4) and (6) generally perform well, when compared to methods (1), (2) and (5), because audio signatures are less affected visual attacks.

There is a slight improvement in the registration accuracy (by 3.2%) of method (2), compared to that of method (1). The reason for this enhancement is, when the sliding window scheme is employed, false positive rate is reduced. Similarly, method (4) slightly enhances the registration accuracy compared to method (3)(by 1%), due to the inclusion of sliding window technique, which decreases false positives.

Method (6) performs better for all six transformations by improving the regis-

Table 4.2: Perfectly registered frames (in %) for geometric and scaling transformations

Transformations		**MV**	**MV+SW**	**MFCC**	**MFCC**	**CH**	**ALL**
Category	**Type**	**(1)**	**(2)**	**(3)**	**+ SW(4)**	**(5)**	**(6)**
Geometric	Rotation	54.45	58.67	91.82	91.82	58.83	93.29
	Cropping	53.59	59.31	90.71	90.85	49.62	92.71
	Flipping	46.72	50.77	90.63	91.69	50.07	94.68
Scaling	Zooming in	52.61	52.99	91.56	92.18	48.85	92.49
	Resolution	57.61	59.45	89.57	90.37	49.26	93.18

tration accuracy (up to 41.9%) compared to the reference methods. Integration of both motion and audio features for the registration task, is the exact reason for its improved performance. On the other hand, Chupeau et al.'s method yields poor results for flipping and zooming transformations. The reason for the poor performance of method (5) is, the limited capabilities of color histograms against region-based transformations.

Temporal and caption transformations: Table 4.3 shows the registration accuracy of six compared methods for temporal and caption based categories, which include slow motion, fast forward, pattern insertion, picture-in-picture and moving caption transformations. Method (5) gives poor results for caption based transformations. This is because, inserting text patterns would substantially change the histogram based signatures. However, the proposed methods using MFCC features (methods (3), (4) and (6)) are less affected by this category of transformations.

Table 4.3: Perfectly registered frames (in %) for temporal and caption transformations

Transformations		**MV**	**MV+SW**	**MFCC**	**MFCC**	**CH**	**ALL**
Category	**Type**	**(1)**	**(2)**	**(3)**	**+ SW(4)**	**(5)**	**(6)**
Temporal	Slow motion	53.08	54.96	65.71	66.77	51.63	90.83
	Fast forward	45.15	48.27	62.93	62.93	50.74	88.75
Caption	Pattern ins.	61.57	65.15	90.53	90.62	45.71	91.03
	Pic-in-pic	49.57	53.64	91.78	92.46	48.94	92.85
	Moving cap.	53.02	55.25	89.96	90.05	46.68	91.94

It is observed that the method (6), which combines MFCC and motion features for frame matching, significantly improves registration accuracy by 40-45% for all five transformations listed in Table 4.3. Specifically, for fast forward transformation, method (6) performs well and significantly increases registration accuracy up to 43.6% when compared to the reference methods.

Audio and filtering transformations: Table 4.4 shows the registration performance of six compared methods for audio and filtering categories. Audio category includes mp3, single band and multi band compressions, while filtering category includes blurring, noise addition and contrast change transformations.

Table 4.4: Perfectly registered frames (in %) for audio, filtering & combined types

Transformations		**MV**	**MV+SW**	**MFCC**	**MFCC**	**CH**	**ALL**
Category	**Type**	**(1)**	**(2)**	**(3)**	**+ SW(4)**	**(5)**	**(6)**
Audio	mp3 comp.	75.64	75.94	56.66	57.64	60.23	90.46
	Single band	78.24	79.24	61.16	61.16	62.82	92.38
	Multi band	76.06	76.06	61.63	61.97	61.56	91.57
Filtering	Blurring	57.70	59.31	78.59	79.38	62.77	90.34
	Noise	52.91	56.18	85.34	85.98	56.98	92.49
	Contrast	62.74	59.88	84.74	82.62	51.35	91.37
Combined	3 combined	41.68	45.16	80.67	82.56	42.85	89.56

The registration accuracy of only MFCC based methods (method (3) and (4)) degrade slightly for audio transformations. This is because the spectral descriptors are much affected by single and multi band compressions. The motion features are much affected by filtering attacks such as noise addition, which in turn reduces the registration rates of methods (1) and (2) for filtering transformations. The Table 4.4 results demonstrate the improved performance of method (6) (up to 33.3%) for all seven transformations compared to the reference method.

Computational cost comparison: To evaluate the effectiveness of the proposed method, the code is implemented in MATLAB using a PC with 3GB RAM and 2.8GHz CPU. The total computational cost of all six methods including signatures extraction and matching are shown in Table 4.5. The computational costs are measured using 24s query clip and 2944s master sequence for temporal registration. Precisely, method (1)-(6) take 234.63s, 177.58s, 150.36s, 103.54s, 182.09s and 173.06s respectively to register a query clip of duration 24s with the master sequence.

The signature matching cost of method (2) is reduced drastically (nearly 96.5%) when compared to that of method (1). The reason for this drastic reduction is, in method (2) only the candidate segment motion features are matched with the query clip features using sliding window scheme. Thus, method (2) reduces the computational time by 25% compared to method (1). There is a huge reduction (nearly 97.8%) in the fingerprint matching cost of method (4), when compared to that of method (3). This is because, in method (4) the MFCC features of query video are matched

Table 4.5: Comparison of computational cost (in seconds)

Computational Cost	**MV (1)**	**MV+SW (2)**	**MFCC (3)**	**MFCC + SW(4)**	**CH (5)**	**ALL (6)**
Signature extraction	176.95	175.57	103.98	102.49	156.41	171.39
Signature matching	57.68	2.02	46.38	1.06	25.68	1.67
Total cost	234.63	177.58	150.36	103.54	182.09	173.06

only with that of the candidate segment instead of the complete master video. Hence, method (4) reduces the computational cost by 34% when compared to method (3).

The total computational cost of method (6) is slightly high compared to methods (2)-(4). Although method (4) is the most cost effective method; however, its registration results are poor for audio transformations, when compared to that of method (6). Thus Table 4.5 results prove that, method (6) significantly improves detection accuracy by 25.6% and extends the coverage to more number of transformations at the cost of slight increase in computational time.

The experimental results demonstrate the improved registration accuracy of the proposed methods compared to the reference method. The reason is, the integration of visual and acoustic features provide accurate temporal registration with reasonable robustness against a wide variety of video transformations. In this way, method (6) consistently provides better performance for all seven categories of video transformations compared to the reference methods.

4.2 Spatio-Temporal Registration Framework Using Visual Features

Spatio-temporal frame alignment of the illegal video with the master content is prerequisite, so as to estimate the geometric distortions and camcorder capture location in a theater. Therefore, followed by the temporal registration scheme, this chapter concentrates on spatio-temporal alignment of the pirate video with the master sequence. Precisely, a novel spatio-temporal registration framework is contributed, which employs robust visual signatures derived from SURF (Bay et al. 2008) descriptors. Further, the proposed framework utilizes 1-D SURF signatures extracted from SURF key points for the temporal alignment task, which are compact compared to the current multi-dimensional SURF signatures (Roth et al. 2010; Zhang et al. 2010). Furthermore, sliding window based dynamic programming technique is utilized to decrease the fingerprint matching cost of the proposed framework. The

proposed spatio-temporal registration framework including temporal and geometrical frame alignments are described as follows.

4.2.1 Proposed registration scheme using visual features

The proposed spatio-temporal registration framework is shown in Figure 4.3, which comprises two phases: temporal and geometric frame alignments. In the first phase, when a query video is presented, the master sequence is segmented using a sliding window of size equal to pirate clip. After this step, the similarity between each windowed segment and the query video is computed using 1-D SURF signatures. Then, the windowed segment with minimum distance score, is indicated as *Most similar segment* of the master sequence. Consequently, optimal frame-to-frame alignments of the pirate video with the *Most similar* segment are computed using dynamic programming technique.

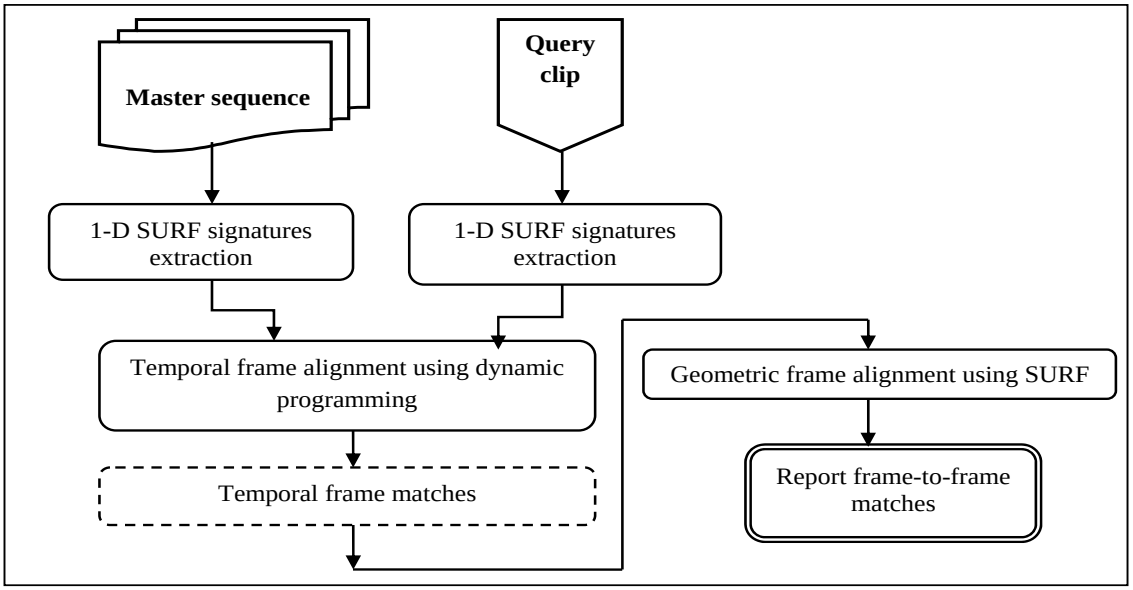


Figure 4.3: Proposed spatio-temporal registration framework using visual features

In the second phase, SURF descriptors of temporally mapped frames are aligned by means of enough key points, to obtain robust geometrical alignments. Temporal and geometric alignments of the pirate video with the *Most similar* segment are detailed below.

Temporal alignment of frames

In the proposed framework, 1-D SURF signatures are employed for implementing the temporal registration task described as follows.

Fingerprints extraction: SURF is an interest point based feature (Bay et al. 2008), which is popularly employed in the CBCD domain to identify duplicate video clips (Roth et al. 2010; Zhang et al. 2010). SURF descriptor represents each key/interest point using a higher dimensional feature vector, which is typically 64 integers/interest point. However, each frame may contain multiple interest points; therefore, there could be too much of information to index and search. Further, *direct comparison of SURF descriptors across all frames would be computationally expensive.* On the other hand, robust visual features illustrating both temporal and spatial contents are needed to obtain accurate frame-to-frame alignments.

In order to tackle these problems, the proposed system computes 1-D SURF signature by combining both the spatial and temporal information. More specifically, a video frame is divided into $n \times n$ regions and the 1-D SURF signature is calculated as the mean value of region-wise count of SURF key points of a frame. However, the segmentation of a frame into $k \times k$ regions, plays an important role in predicting the registration accuracy. In this research work, k is set as 3, after executing experiments for different k values ranging from 2-5.

Sliding window-based dynamic programming: The computational complexity of dynamic programming to map two sequences of size N and M is $O(NM)$; hence if sequence size increases, then the performance of the algorithm decreases. In order to overcome this discrepancy, frame matching between the query video and the *Most similar* segment is performed instead of the entire master sequence. Figure 4.4 shows Algorithm 4.2, which is used to select the *Most similar* segment of the master sequence.

Temporal frame alignments using dynamic programming: Let CS be the *Most similar* segment of the master sequence and QS be the query clip. Let SF_{cs}^{k} and QSF^{k} be the 1-D SURF signatures of segments CS and QS respectively as given by,

$$CS \in \{SF_{cs}^{k}\}_{k=1}^{n} \tag{4.12}$$

$$QS \in \{QSF^{k}\}_{k=1}^{n} \tag{4.13}$$

Algorithm 4.2: Most Similar Segment Selection

1: Segment the master sequence into overlapping blocks of length equal to the query clip.

2: Extract 1-D SURF signatures for each segment.

3: Let a master sequence

$$MS = \{s_1, s_2, s_3, .., s_m\}, \tag{4.8}$$

where s_i is the i^{th} segment and m is the total segments of MS. Here, the 1-D SURF signatures of s_i are denoted as,

$$s_i \in \{SF_i^k\}_{k=1}^n, \tag{4.9}$$

where SF_i^k is k^{th} visual signature of segment s_i.

4: Let QS be a query clip and the 1-D SURF signatures of QS are denoted as,

$$QS \in \{QSF^k \mid k = 1, 2, ..., n\} \tag{4.10}$$

where QSF^k is k^{th} visual signature and n is total SURF signatures of QS.

5: The similarity sim_{seg} between QS and the segment s_i is computed using Manhattan distance as follows,

$$sim_{seg}(s_i, QS) = \sum_{k=1}^{n} |SF_i^k - QSF^k| \tag{4.11}$$

6: A master segment with minimum sim_{seg} value (i.e. $sim_{seg} \leq$ threshold) is selected as the *Most similar* segment of MS. In this work, the threshold is set as 0.48, after implementing experiments for different values ranging from 0.30-0.60.

Figure 4.4: *Most similar* segment selection algorithm using sliding window

The distance between the visual signatures of CS and QS is computed using comparative Manhattan distance as follows,

$$dist_{surf}(CS(j), QS(j)) = \frac{|SF_{cs}^j - QSF^j|}{|SF_{cs}^j| + |QSF^j|} \tag{4.14}$$

where $j = [1 : n]$ and n is the total SURF signatures of video segments. Then score matrix SM is calculated using Equations (4.1) and (4.14). After this step, the optimal alignment path is determined and Temporal Frame Alignments (TFA) based

on SURF signatures is calculated as,

$$TFA \in \{\{cv_x, qv_y\}, \quad 1 \leq x \leq n, \quad 1 \leq y \leq n\} \tag{4.15}$$

Here, cv_x and qv_y indicate the frame matches of *Most similar* segment and query video respectively.

Geometric alignment of frames

Performing geometric alignment across all temporally aligned frames is not feasible due to computational load. Furthermore, all video frames may not provide essential key points to enable accurate geometric registration.

In order to tackle these issues, a small set of representative frames are utilized for the geometric registration framework. The SURF descriptors and the score matrices computed for the temporal registration provide important guidelines to select the representative frames. More specifically, frame pairs with lower distance score are considered and mapped in terms of their key points, so as to provide accurate pixel correspondences of frames. Two control points are mapped, only if the squared Euclidean distance between their feature vectors is minimum. Figure 4.5 shows the sample *Most similar* segment and query clip frames, which are geometrically mapped in terms of their key/interest point pairs. Here, query video is generated by applying random noise transformation.

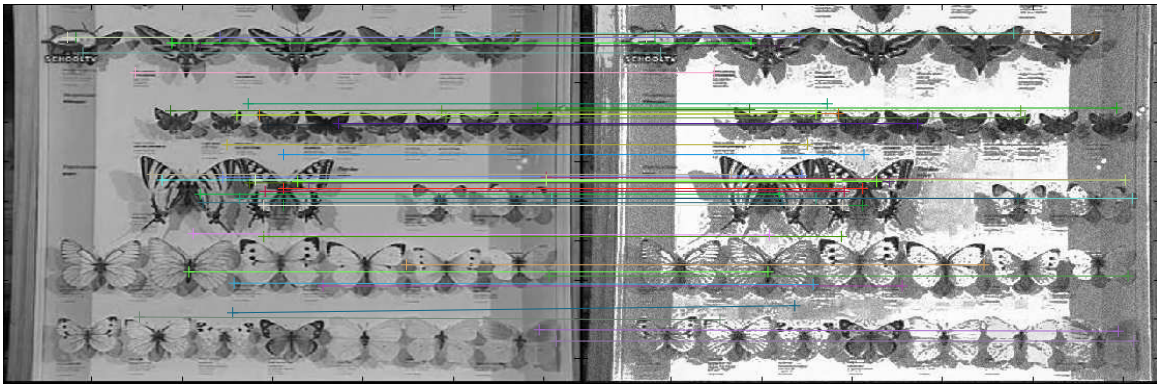


Figure 4.5: Pairs of matched interest points of *Most similar* segment(left) and pirate (right) video frames. Here, random noise transformation is applied for creating the pirate video.

4.2.2 Experimental setup and results

Master database and query dataset construction: The proposed spatio-temporal registration framework is validated on 100h of TRECVID-2009 Sound and Vision data, plus another 30h of real data comprising camcorded copies of master video sequences. All the video clips are converted into uniform format:352×288 pixels and 15fps using resampling technique. Table 4.6 lists the video transformations used in the proposed framework, which cover most of the manipulations specified in TRECVID-2009 copy detection task.

Table 4.6: Transformations considered in the proposed framework using visual features

#	Category	Description
T1	Rotation	Rotating by 15°-20°
T2	Random noise	Add 20% gaussian noise
T3	Blurring	Blur by 21%
T4	Brightness change	Increase brightness by 15%
T5	Cropping	Crop top & bottom regions by 25% each
T6	Picture-in-picture	Insert smaller resolution picture
T7	Zoom in	Zoom in to the frame by 18%
T8	Slow motion	Halve the video speed
T9	Fast forward	Double the video speed
T10	Pattern insertion	Insert text pattern into selected frames
T11	Moving caption	Insert moving titles into entire video
T12	3 combined	20% cropping, 15% noise & moving caption
T13	5 combined	17% noise, 20% blurring, 17% brightness, cropping & pattern insertion

From the master database, 45 video clips of duration 20-45s are randomly selected and transformations listed in Table 4.6 are applied to generate the query dataset. In addition, 48 camcorded copies of duration 35-115s are generated from 25 master video sequences. Thus, the resulting 633 ((45×13)+ 48) video sequences are treated as query clips for the proposed temporal registration task. Geometric registration is implemented on a set of 32 representative frames selected from the temporally aligned query and candidate segments.

Evaluated methods: The accuracy of the following three methods are evaluated:

(1) 1-D SURF signatures (abbreviated as 'SURF');
(2) Chupeau et al.'s method (2006) ('CHE');
(3) 1-D SURF signatures + sliding window ('ALL');

The methods (1) and (3) evaluated different combinations of the proposed techniques. In method (1), 1-D SURF signatures of the query video are mapped with respective signatures of the entire master sequence (i.e. query clip is matched with all segments of the master sequence).

Chupeau et al.'s method (2006) uses color histograms for computing frame-to-frame correspondences between the query and master video sequences. In this method, the distance between color histograms of successive frames are used as temporal fingerprints of video sequences. Method (3) uses sliding window mechanism to align SURF signatures of the query segment with the respective features of the *Most similar* segment instead of the entire master sequence. The *Most similar* segment of the master video is selected using the algorithm explained in Figure 4.4.

Temporal registration results: Table 4.7 shows the temporal registration accuracy of three compared methods for T1-T7 types. The results are denoted in terms of *percentage of perfectly Matched Frames* (denoted as '**MF**') and *Average Distance between true and estimated frame indexes* (indicated as '**AD**').

Table 4.7: Temporal registration results for T1-T7 types.

Attacks	**SURF (1)**		**CHE (2)**		**ALL (3)**	
	MF	**AD**	**MF**	**AD**	**MF**	**AD**
Rotation	75.18	2.0	56.69	3.1	81.18	1.1
Random noise	86.31	1.9	51.46	1.9	88.64	1.2
Blurring	76.76	2.3	59.34	2.5	80.15	1.2
Brightness	83.18	2.8	58.61	2.8	85.98	1.7
Cropping	79.54	3.4	34.28	3.6	82.59	1.2
Picture-in-picture	78.26	2.1	36.14	2.9	82.64	1.5
Zoom in	74.42	2.2	61.37	2.5	77.25	1.1

Method (3) scores better results for all seven transformations and improves registration accuracy up to 31.5% compared to the reference method. The joint utilization of robust SURF signatures and the sliding window scheme is the exact reason for the enhanced performance of method (3). In addition, method (3) yields more accurate results compared to method (2), as the AD values are always less than two.

On the other hand, Chupeau et al.'s method (2006) scores poor results for cropping and picture-in-picture types in terms of low MF and high AD rates. This is because, cropping introduces black borders on top and bottom regions; therefore, very different video signatures are generated for master and query segments.

Table 4.8 lists the temporal registration performance of the three compared methods for T8-T13 types. Method (3) generally performs well for all eight transformations and enhances the registration accuracy (by 34.53%) compared to the reference method.

Table 4.8: Temporal registration results for T8-T13 types.

Attacks	**SURF (1)**		**CHE (2)**		**ALL (3)**	
	MF	**AD**	**MF**	**AD**	**MF**	**AD**
Slow motion	87.76	5.1	50.28	4.2	88.11	3.1
Fast forward	85.45	4.8	54.67	5.4	87.34	4.7
Pattern insertion	78.63	1.4	45.19	2.6	80.10	1.2
Moving caption	66.36	1.5	48.31	2.1	69.86	1.3
3 combined	85.49	5.2	50.28	5.2	88.42	3.9
5 combined	79.61	4.6	40.59	5.1	82.68	2.9

For pattern insertion and 5 combined types, the MF rates of method (2) decline sharply. This is because, inserting patterns/captions substantially changes histogram bin values. In case of combined types, histogram signatures of query and candidate segments are severely affected by the addition of gaussian noise and insertion of text patterns. Yet, the proposed methods (1) and (3) using SURF signatures are less affected by this category.

Geometric registration results: Table 4.9 shows the geometric registration results of the proposed method for different transformations such as rotation, cropping and combined types. The registration results are denoted in terms of mean and maximum pixel distances between the geometrically mapped candidate and query segment frames. The registration results of the proposed method is very efficient, because the mean pixel distance is always less than one. The robust nature of powerful SURF descriptors is the correct reason for this accurate geometric alignments.

Computational cost comparison: The proposed method is evaluated in MATLAB using a PC with 2.8GHz and 3GB RAM. Table 4.10 indicates the total computational cost of all three methods including fingerprint extraction and frame matching costs. The costs are measured by implementing the frame alignment of a 315s query clip with the 2493s master sequence.

The frame matching cost of method (3) is drastically reduced compared to the other two methods. The reason is, in method (3) query segment features are mapped only with the corresponding *Most similar* segment signatures instead of the entire

Table 4.9: Geometric registration results in terms of mean & maximum pixel distances

Attacks	**Mean dist.**	**Max dist.**
Rotation	0.912	1.592
Random noise	0.713	1.373
Blurring	0.777	1.346
Brightness change	0.810	1.368
Cropping	0.735	1.345
Picture-in-picture	0.628	1.349
Zoom in	0.665	1.160
Pattern insertion	0.681	1.283
Moving caption	0.625	1.346
3 combined	0.854	1.459
5 combined	0.881	1.496

Table 4.10: Comparison of computational cost (in seconds)

Computational Cost	**SURF (1)**	**CHE (2)**	**ALL (3)**
Fingerprint extraction	176.95	166.41	176.39
Frame matching	47.68	45.68	1.47
Total cost	224.63	212.09	177.86

master sequence; hence false positives are removed effectively and consequently frame matching cost is considerably reduced.

4.3 Spatio-Temporal Registration Framework Using Visual-Audio Features

As mentioned in Section 2.2.3, if audio content is available, then the joint exploitation of visual-audio fingerprints for the alignment task, significantly enhances the registration accuracy. Therefore, followed by spatio-temporal alignment using visual features, this chapter also contributes a new spatio-temporal registration framework, which utilizes multimodal fingerprints for obtaining accurate frame alignments of the pirate and master video sequences. The main contributions of the proposed spatio-temporal registration framework are given by,

- A new spatio-temporal registration framework is presented by exploiting visual signatures extracted from SURF key points and audio fingerprints derived from spectral centroid features.

- A novel SURF-based visual-profile is introduced, which is compact (1-D) compared to the existing multi-dimensional SURF-based fingerprinting methods (Roth et al. 2010; Zhang et al. 2010). Roth et al. (2010) used 16-D SURF descriptors, while Zhang et al. (2010) utilized 64-D SURF signatures for the copy detection task.
- Robust acoustic features are also employed for the temporal registration task, which considerably enhance the registration accuracy, compared to the existing registration schemes (Chupeau et al. 2006; Baudry et al. 2010).
- An algorithm for selecting the *candidate segment* of the master video sequence is proposed, using sliding window based Dynamic Time Warping (DTW) technique (Rabiner and Juang 1993), which substantially decreases the frame matching cost.
- A multimodal frame matching scheme is introduced for aligning the acoustic and visual fingerprints, which noticeably reduces false frame matches.
- *Principal frames* extraction algorithm is presented, which extracts the most similar frames from the temporally aligned candidate and pirate video segments, in order to implement the geometric registration task.

The proposed spatio-temporal registration framework including temporal alignment of frames followed by multimodal frame matching and geometric frame alignments is illustrated as follows.

4.3.1 Proposed spatio-temporal registration framework

Problem formulation: The proposed registration framework is formulated as follows: Let $PS = \{p_i | i = 1, 2, ..., n_p\}$ be a pirate video sequence with n_p frames, where p_i is i^{th} copy frame; and let $MS = \{m_j | j = 1, 2, ..., n_m\}$ be a master sequence with n_m frames, where m_j is j^{th} master frame and $n_m >> n_p$. Here PS is derived from MS after applying video attacks such as noise, cropping, camcording, caption insertion, blurring and so on. The proposed framework selects a subsequence of MS denoted as a candidate segment $CS = \{m_j, m_{j+1}, .., m_{j+n_c-1}\}$ with n_c frames, using a sliding window scheme. Here, the objective is to spatio-temporally map the duplicate and master video sequences and as a result exact frame alignments of CS and PS can be obtained.

Spatio-temporal registration framework using visual-audio features:

The proposed registration framework is given in Figure 4.6, which comprises two stages namely, temporal and geometric frame alignments.

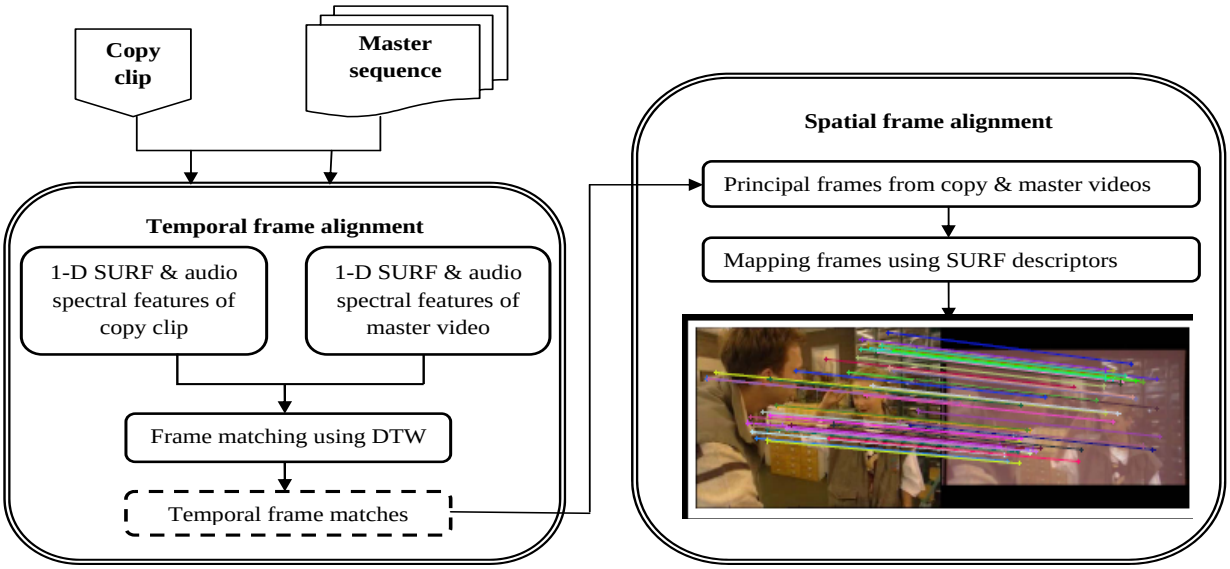


Figure 4.6: Spatio-temporal registration framework using visual-audio features

In the first stage, when a query video is presented, then the master sequence is scanned with a sliding window of size equal to query clip. In this step, similarity between the video copy and the windowed segment is computed using their temporal signatures extracted from SURF control points and spectral centroid descriptors. The windowed sequence having minimum distance score is selected and indicated as the *candidate segment*. After this step, audio-visual fingerprints of the candidate and pirate video segments are mapped separately using DTW technique and consequently mapping results are combined, in order to get temporal frame alignments.

In the second stage, from the temporally mapped and candidate and query video segments, a set of most/highly similar frames are selected and denoted as *principal frames* of two video sequences. The resultant *principal frames* are matched using their SURF descriptors by means of enough interest points, in order to obtain accurate spatial frame alignments.

4.3.2 Temporal alignment of frames

In the proposed framework, compact visual profile extracted from SURF signatures and acoustic profile derived from spectral centroid features are jointly exploited for obtaining the accurate temporal frame alignments, which is detailed below.

1-D Visual profile extraction

In the proposed registration framework, SURF interest points-based signatures are employed to extract the visual profile of video contents. SURF is a scale and rotation invariant descriptor (Bay et al. 2008); hence it is widely used in the CBCD literature to identify pirate video clips (Roth et al. 2010; Zhang et al. 2010; Yang et al. 2008). As mentioned in Section 4.2.1., SURF descriptor represents each key point by means of a multi-dimensional feature vector, normally 64 integers/interest point. Since each frame contains multiple SURF key points; hence, too much of information needs to be processed. Moreover, direct comparison of SURF feature descriptors across all frames is computationally expensive. On the other hand, existing multi-dimensional SURF-based fingerprinting methods utilize only spatial content of frames (Roth et al. 2010; Zhang et al. 2010; Yang et al. 2008). However, to create robust visual fingerprints of the given video sequence, both the temporal as well as spatial content of frames need to be considered.

To tackle these problems, in the proposed framework, a video clip is compactly indicated using 1-D SURF signatures extracted from SURF key points, which efficiently illustrate the spatio-temporal information of frames. More precisely, a video frame is segmented into $n \times n$ regions and 1-D SURF signatures are calculated as *the mean of differences between region-wise count of SURF key points of consecutive frames*. Figure 4.7 illustrates computation of 1-D SURF signatures from the sample frames on a 3×3 partition.

The partition of a frame into $n \times n$ regions, plays an active role in predicting the registration performance and computational cost. Smaller values of n increase the computational burden, whereas larger values of n decrease the robustness of the proposed system.

To handle this discrepancy, experiments are evaluated for different values of n ranging between 2-8 and the corresponding registration results are compared. More specifically, the proposed scheme is experimented on a dataset including 112 query clips and 198 reference video sequences videos, where the query clips vary between 18-35 seconds. Figure 4.8 indicates the average registration accuracy achieved for different n values and concludes that the maximum accuracy (91.8%) is achieved at n=3. Thus, the value of n is set as 3 in the consequent experiments, which provides the best balance of effectiveness and robustness.

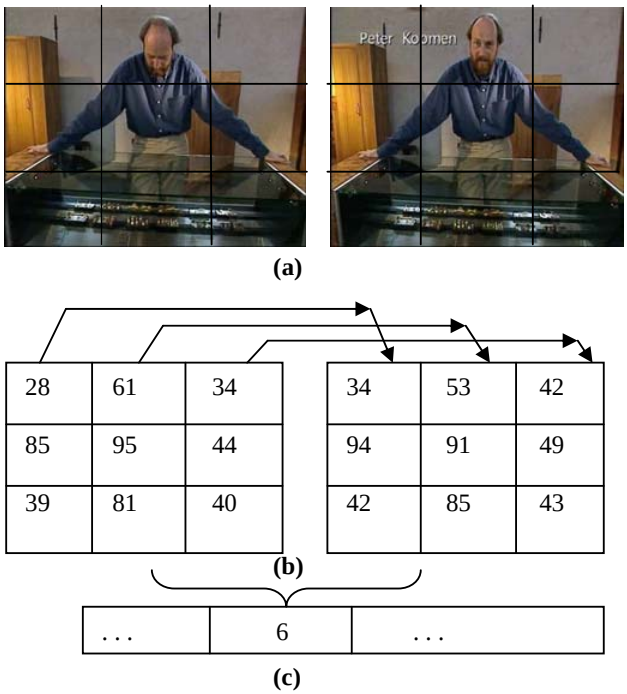


Figure 4.7: 1-D SURF signature extraction: (a) Video frames partitioned into 3 ×3 regions. (b) Region-wise count of SURF key points. (c) Computing 1-D signature with time series

1-D Acoustic profile extraction

As mentioned in Section 3.3.1., spectral centroid is an important timbral descriptor, which specifies the center of gravity of the signal spectrum (Rabiner and Juang 1993; Park 2010). Precisely, centroid describes brightness of a sound signal and it is a highly robust spectral feature (West 2008) ; hence, it is widely popular in speech recognition applications (Eronen and Klapuri 2000). Furthermore, the most significant perceptual audio descriptors exist in the frequency domain (Li et al. 2003; Jie et al. 2009). Due to these reasons, 1-D spectral centroid signatures are utilized to describe the acoustic information of video sequences, which are computed as follows:
As described in Section 3.3.1., first the audio signal is down sampled and consequently segmented into 11.60ms windows using Hamming windowing, where the window overlap factor is 80% (Roopalakshmi and Reddy Sep-2011). Then, from the power spectrum of the audio signal, the Spectral Centroid descriptor (SC) is computed using the frequency distribution values as specified in Equation (3.23). As compared with (Roopalakshmi and Reddy Sep-2011), the proposed framework employs absolute values of the spectral centroid features for the registration task. Further, normalization is

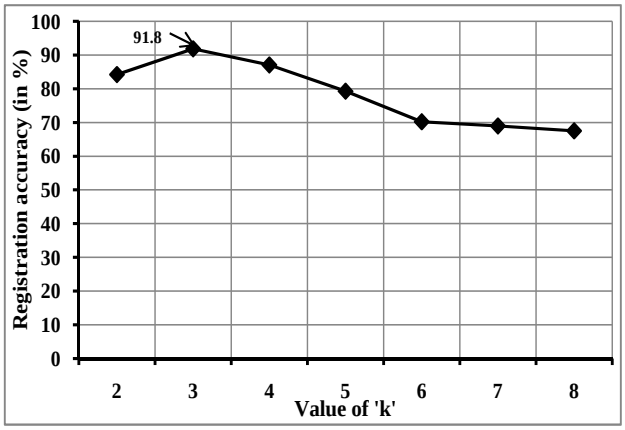


Figure 4.8: 'k' versus registration accuracy

applied to the resultant signatures in order to improve the robustness of the proposed framework.

Introduction to Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) is extremely effective in synchronizing two time-dependent sequences, since it minimizes shifting effects in time by allowing elastic transformation of sequences (Rabiner and Juang 1993; Müller 2007). Therefore, DTW technique is extensively employed in a broad range of applications such as speech recognition (Senin 2008), sequence alignment and information retrieval (Müller 2007).

Given two time-dependent feature sequences $X = \{x_i | 1 \leq i \leq N\}$ of length N and $Y = \{y_j | 1 \leq j \leq M\}$ of length M. A local cost measure C indicating the distance between x_i and y_j is formulated as,

$$C(x_i, y_j) = Dist(x_i, y_j) \tag{4.16}$$

where $Dist$ denotes Manhattan distance measure in the proposed framework. To find an alignment of X and Y, a warping path $W = \{w_1, w_2, ..., w_L\}$ with $w_l = (x_l, y_l) \in [1 : N] \times [1 : M]$ for $l \in [1 : L]$ needs to be computed. The accumulated Path Cost PC associated with W of sequences X and Y is defined as,

$$PC_W(X, Y) = \sum_{l=1}^{L} C(x_{i_l}, y_{j_l}) \tag{4.17}$$

The objective of DTW technique is to find an optimal warping path of sequences X and Y having minimal path cost among all possible warp paths (Müller 2007), which

is denoted as,

$$DTW(X,Y) = W_{op} = min\{(PC_W(X,Y)) \mid W \in P^{N\times M}\} \tag{4.18}$$

where W_{op} is the optimal warping path and $P^{N\times M}$ indicates the set of all possible warping paths. The optimal warping path $W_{op} = \{wp_1, wp_2, .., wp_L\}$ with $wp_l = (x_l, y_l) \in [1:N] \times [1:M]$ for $l \in [1:L]$. The accumulated path cost (PC_{dtw}) of $DTW(X,Y)$ is denoted as,

$$PC_{dtw}(X,Y) = \sum_{l=1}^{L} C(x_{n_l}, y_{m_l}) \tag{4.19}$$

Let $D(N,M)$ is the global cost matrix of size N×M. DTW technique calculates the warping path W_{op} based on dynamic programming (Müller 2007) in three steps as follows,

---

1: Initialization:

$D(1,1) = 0$;

First column: $D(i,1) = \sum_{k=1}^{i} C(x_k, y_1), i \in [1:N]$;

First row: $D(1,j) = \sum_{k=1}^{j} C(x_1, y_k), j \in [1:M]$;

2: Recursion:

All other elements of $D(i,j)$ are recursively computed as,

$$D(i,j) = min\{(D(i-1,j-1), D(i-1,j), D(i,j-1)) + C(x_i, y_j)\} \tag{4.20}$$

where $i \in [1:N]$ and $j \in [1:M]$.

3: Termination:

Once the entire D matrix is computed, backtracking is done to determine the optimal alignments starting from W_{op} = (M,N) to W_{op}=(1,1).

In this research study, the optimal warping path W_{op} specifying the alignment of sequences X and Y, satisfies the following conditions:

---

a) Endpoint constraints:

For the warping path W_{op}, starting point is $wp_1 = (1,1)$ and ending point is $wp_L = (N,M)$.

b) Monotonicity constraints:
In order to preserve temporal continuity, the warping function is monotonically increasing as given by,
$x_1 \leq x_2 \leq ... \leq x_L$ and $y_1 \leq y_2 \leq ... \leq y_L$.

c) Local continuity constraints:
This category constraints the slope of the warping path by means of limiting long jumps in the alignment of X and Y sequences. Normally, *the possibility of huge changes in the feature sequences of consecutive frames is very low* and thus the step size condition is formulated as,
$wp_{l+1} - wp_l \in \{(1,0),(0,1),(1,1)\}$ for $l \in [1 : L-1]$.

Sliding window based DTW

The computational complexity of DTW algorithm to match two sequences of size M and N is $O(MN)$; hence if sequence size increases, then the complexity of the algorithm also increases. To deal with this issue, frame alignments between the query clip and the *candidate* segment are computed instead of the complete master sequence. Algorithm 4.3 given in Figure 4.9 explains the steps used to select the *candidate segment* of the master sequence.

4.3.3 Multimodal frame matching

In this step, the audio-visual signatures of the two video sequences are aligned separately and the resultant matches are combined into final temporal alignments. More precisely, the multimodal frame matching scheme is implemented as follows.

Frame matching using visual signatures: Let CS be a candidate segment of the master sequence with n_c frames and PS be a pirate sequence with n_p frames. Let VF be the visual fingerprint of CS such that, $CS \in \{VF_i |\ 1 \leq i \leq n_{vf}\}$ with n_{vf} signatures. Consider QVF is the visual fingerprint of PS such that $PS \in \{QVF_j |\ 1 \leq j \leq n_{qvf}\}$ with n_{qvf} signatures. The proposed framework assumes that the length of candidate sequence is equal to size of the query clip; hence, $n_{vf} \simeq n_{qvf}$. The cost measure C_{vis} indicates the dissimilarity between two visual signatures, which is computed using comparative Manhattan distance metric as follows,

$$C_{vis}(CS_k, PS_k) = \frac{|(VF_k - QVF_k)|}{|(VF_k)| + |(QVF_k)|}, \quad 1 \leq k \leq n_{vf}. \tag{4.25}$$

Algorithm 4.3: Selection of the Candidate Segment

1: Divide the master sequence into overlapping segments of length equal to the query clip.

2: Extract 1-D visual and audio profiles for each segment as described in Section 4.3.2.

3: Let a master sequence MS be,

$$MS \in \{S_i |\ 1 \leq i \leq m\}, \tag{4.21}$$

where S_i is the i-th segment and m is total segments of MS. Here, each segment S_i of MS can be represented as,

$$S_i \in \{(V_i^k \cup A_i^r) \mid 1 \leq k \leq n,\ 1 \leq r \leq p\} \tag{4.22}$$

where V_i^k is k-th feature vector of visual fingerprint of S_i and n indicates total feature vectors. Here, A_i^r is r-th vector of audio fingerprint of S_i and p represents number of feature vectors.

4: Let a pirate sequence PS is compactly represented as,

$$PS \in \{(QV^k \cup QA^r) \mid 1 \leq k \leq n_q,\ 1 \leq r \leq p_q\} \tag{4.23}$$

where QV^k is the k-th feature vector of visual fingerprint of PS and n_q is total vectors. Here, QA^r is r-th vector of audio fingerprint of PS and p_q indicates total feature vectors.

5: Compute the segment similarity Seg_{sim} between S_k of MS and PS using DTW as follows,

$$Seg_{sim}(S_k, PS) = PC_{dtw}(V_k, QV) + PC_{dtw}(A_k, QA) \tag{4.24}$$

where PC_{dtw} represents the accumulated path cost of optimally warped visual sequences (i.e. V_k and QV) and audio feature sequences(i.e. A_k and QA) respectively.

6: Select S_i having lowest Seg_{sim} value (i.e. distance score) as a *candidate* segment of the master sequence for further comparison.

Figure 4.9: Selection of the *candidate segment* using visual-audio signatures

After this step, the optimal frame mappings between the visual signatures of two video sequences are computed using DTW technique. The resultant frame matches FM_{vis} based on visual signatures is formulated as,

$$FM_{vis} = \{\{cv_i, pv_j\} \mid 1 \leq i \leq n_c,\ 1 \leq j \leq n_p\} \tag{4.26}$$

where *cv* and *pv* indicate the matching frames of candidate and pirate video sequences respectively. Figure 4.10 shows the frame alignments of copy and candidate feature sequences in terms of global cost matrix D and the optimally warped path. The dark strips in matrix D indicate the high similarity between the two video contents.

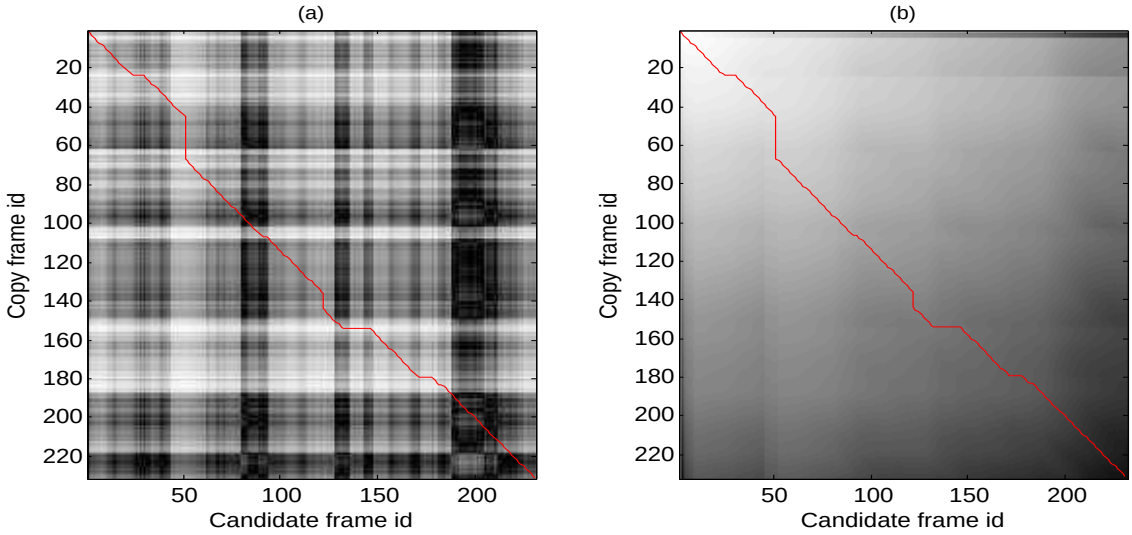


Figure 4.10: Frame alignments of copy and candidate feature sequences. (a) global cost matrix D, darker regions indicate high similarity (b) optimally warped path

Frame mapping using acoustic signatures: Let SF be the spectral centroid-based audio fingerprint of CS such that $CS \in \{ SF_m, |1 \leq m \leq n_{sf}\}$ with n_{sf} signatures. Let QSF be the audio fingerprint of PS such that $PS = \{QSF_m|1 \leq m \leq n_{qsf}\}$ with n_{qsf} signatures (in this study, $n_{sf} \simeq n_{qsf}$). The cost measure C_{aud} indicating the difference between two audio fingerprints is calculated using squared Euclidean distance measure as follows,

$$C_{aud}(CS_k, PS_k) = |(SF_k - QSF_k)^2|, \ 1 \leq k \leq n_{sf} \tag{4.27}$$

Then the optimal warping path illustrating the frame alignments of SF and QSF signatures is computed using DTW algorithm. The resultant frame matches FM_{aud} based on audio spectral signatures is formulated as,

$$FM_{aud} = \{\{cs_i, ps_j\}| \ \ 1 \leq i \leq n_c, \ 1 \leq j \leq n_p\} \tag{4.28}$$

where *cs* and *ps* indicate the matching frames of candidate and pirate sequences respectively.

Decision fusion: Frames mapped by both the visual and audio signatures are considered as final frame matches of two video contents, which is given by,

$$FM_{final} = \{\{FM_{vis}\} \cap \{FM_{aud}\}\} \tag{4.29}$$

where FM_{final} provides frame-to-frame alignments of CS and PS sequences respectively. The main benefit of the proposed multimodal frame matching technique is that, since only frames with similar visual and audio features are mapped, it considerably decreases false frame matches. In addition, the proposed matching technique significantly enhances registration accuracy which is evident in Section 4.3.6.

4.3.4 Geometric alignment of frames

As mentioned in Section 4.2.1., geometric mappings across all temporally aligned frames is not feasible due to computational load. Further, all video frames may not provide necessary key points to enable accurate geometric registration.

In order to tackle this problem, a small set of highly similar frames denoted as *principal frames* are employed for implementing the geometric registration task. The SURF descriptors and DTW optimal paths computed for temporal registration task provide significant guidelines for selecting the *principal frames*. More specifically, *principal frames* are extracted from temporally aligned candidate and pirate feature sequences using Algorithm 4.4, which is described in Figure 4.11. The resultant *principal frames* are characterized by a list of interest points and their associated SURF descriptors.

Two control points are matched only if the squared Euclidean distance between their feature vectors is minimum. On the other hand, blind comparison of all feature vectors of two frames is computationally expensive and may lead to false correspondences. In order to solve this discrepancy, feature vectors with minimum feature distances are computed and mapped in terms of their descriptors to provide accurate pixel correspondences of frames.

4.3.5 Experimental setup

The proposed framework is evaluated on three different datasets, namely TRECVID sound & vision data, CC_WEB_VIDEO dataset [1] and a set of real data consisting of camcorded copies of master video sequences.

[1]CC_WEB_VIDEO: Near-Duplicate Web Video Dataset. `http://vireo.cs.cityu.edu.hk/webvideo/`

Algorithm 4.4: Principal Frames Extraction

1: Let the optimal warping path W_{op} specifies the alignment of two feature sequences VF and QVF, such that $W_{op} = \{wp_1, wp_2, .., wp_L\}$ with $wp_l = (x_l, y_l) \in [1 : N] \times [1 : M]$ for $l \in [1 : L]$. Here, VF and QVF represent the visual fingerprints of candidate and pirate sequences respectively.

2: Consider a cost vector W^c_{op} representing the feature distances in terms of cost in each entry of optimal path W_{op} as follows,

$$W^c_{op} = \{wp^c_1, wp^c_2, .., wp^c_L\}, \tag{4.30}$$

where wp^c_1 indicates the cost given in entry wp_1 and so on.

3: Sort the W^c_{op} vector to generate the sorted list of costs represented in DTW path.

4: Lower cost values in the W^c_{op} vector indicate highly similar frames; hence select frame pairs corresponding to lower cost values in W^c_{op} as *principal frames*.

Figure 4.11: Principal frames extraction

Master video database and query dataset construction

TRECVID dataset: TRECVID sound & vision data is a benchmark dataset, which covers a wide variety of contents including science news, reports, documentaries and educational programming. The TRECVID master database comprises approximately 110h of sound & vision data used in TRECVID-2009 copy detection task, plus another 80h of sound & vision data used in TRECVID-2008 copy detection task. The proposed framework transforms the entire video data into the following uniform format: 352×288 pixels and 15fps. It is not necessary to utilize every frame in a video sequence for registration; hence, when a copy clip is given with a different frame rate, it is resampled to 15fps, in order to synchronize it with the master sequence. For example, a 5-second copy clip with 60fps becomes a 240-frame sequence after performing the resampling procedure.

In case of piracy, normally users capture videos by using camcorders and distribute them with some modifications. Thus, most of the pirate videos suffer from distortions such as camcording, photometric variations (lighting changes), editing operations (pattern insertions), frame rate changes, format changes (mp3 format), cropping, rotation attacks and so on; hence, in this context the fifteen types of transformations listed in Table 4.11 are considered in the proposed framework for generating the query dataset.

Table 4.11: Transformations used in the proposed framework

#	Category	Description
T1	Zoom in	Zoom in to the frame by 19%
T2	Slow motion	Halve the video speed
T3	Fast forward	Double the video speed
T4	Pattern insertion	Insert text pattern into selected frames
T5	Moving caption	Insert moving titles into entire video
T6	Rotation	Rotating by 10° to 12°
T7	Random noise	Add 10% gaussian noise
T8	Blurring	Blur by 13%
T9	Brightness change	Increase brightness by 10%
T10	Cropping	Crop top & bottom regions by 20% each
T11	Picture-in-picture	Insert smaller resolution picture into frames
T12	3 combined	Cropping by 15%, 10% of noise & moving caption
T13	5 combined	14% noise, 11% blurring, 14% brightness, cropping & pattern insertion
T14	Mp3 compression	Change audio file format
T15	Single band compression	Compress only specific frequency band

Precisely, from the TRECVID master database, 50 video clips are randomly selected and Table 4.11 transformations are applied to produce the query clips. The resulting 750 (50×15) video sequences of duration 20-35s are used as query clips for the proposed temporal registration task.

CC_WEB_VIDEO dataset: CC_WEB_VIDEO dataset includes video collections from video sharing websites and search engines such as YouTube, Google Video and Yahoo! Video. The CC_WEB_VIDEO master database includes 24 most viewed and top favorite videos provided by CC_WEB_VIDEO collection. The representative snapshots of all 24 master videos are shown in Figure 4.12. From the CC_WEB_VIDEO collection, duplicate and near-duplicate videos ranging from 15 to 25 are retrieved for each of the master video. In total, the CC_WEB_VIDEO query dataset includes approximately 600 video files with two different classes of distortions namely formatting and content distortions. Formatting distortions include changes in frame rate, bit rate, encoding format and frame resolution. Photometric variations (lighting changes), editing variations (e.g., logo insertions) and content modifications such as addition of unrelated frames with different content are categorized into content distortions type.

Camcorded copies: To assess the performances of the proposed framework against camcorder captured videos, the proposed scheme is evaluated on a dataset of 30 master videos and their camcorded versions. 75 camcorded copies of master videos

Figure 4.12: Snapshots of 24 master videos of CC_WEB_VIDEO collection

ranging from 1.55 to 15 minutes are generated for the registration task. The quality of camcorded copies varies from clean copies to heavily modified ones with a large amount of lighting, cropping and compression distortions.

4.3.6 Evaluation results and discussion

Overview of the evaluated methods

The following six methods are implemented for evaluating performance:

(1) The SURF signatures based matching (abbreviated as 'SURF');
(2) The spectral centroid features based matching ('SC');
(3) SURF and spectral features without sliding window ('SURF+SC');
(4) SURF and spectral signatures with sliding window ('ALL');
(5) Chupeau et al.'s method (2006)('CHE');
(6) Baudry et al.'s method (2010)('BA');

The methods (1)-(4) evaluated different combinations of the proposed techniques. Methods (1) and (2) use different video signatures (namely SURF and spectral centroids) to implement the temporal registration of two video contents. Methods (3) and (4) are implemented, to see the effect of sliding window scheme for the proposed registration task.

In method (1), 1-D visual signatures of the pirate clip are matched with the respective features of the entire master sequence (i.e. query clip is matched with all

segments of the master sequence). In method (2), 1-D spectral signatures of the query clip are mapped with the acoustic profile of the complete master sequence. Method (3) utilizes both 1-D SURF and spectral centroid signatures for temporally registering the video contents. In this method, visual-audio fingerprints of the query clip are separately aligned with the corresponding features of the entire master sequence. In method (4), the sliding window mechanism is employed to align multimodal signatures of the copy clip with the corresponding features of the candidate segment, instead of the entire master sequence. The candidate segment of the master sequence is selected using Algorithm 1, as explained in Figure 4.9.

Chupeau et al.'s method (2006) utilizes color histograms for calculating frame-to-frame correspondences between pirate and master contents. It is implemented as follows: color histograms of size 512 bins are extracted from consecutive video frames. A sequence of distances (Euclidean distance) between color histograms of successive frames are utilized as temporal fingerprints of videos and dynamic programming is applied to achieve temporal registration of frames.

Baudry et al.'s method (2010) is one of the latest methods, that uses fingerprints based on wavelet coefficients for temporally registering the query and master video sequences. In this method, first the difference between successive frames is computed and transformed into wavelet coefficients. Then, the resultant coefficients are hierarchically encoded and temporal frame alignments are computed using dynamic programming.

Temporal registration results

The registration performances of six compared methods tested on different datasets against various types of video transformations are discussed as follows.

Registration results for TRECVID dataset: Table 4.12 shows the temporal registration results of six compared methods in terms of percentage of perfectly Matched Frames ('MF') for T1-T7 transformations. The bold font indicates the highest MF scores in the table.

The performance of spectral centroid-based methods (methods (2),(3) and (4)) is superior compared to the other methods for T1-T7 types. This is because, applying transformations on the visual content would not affect acoustic features substantially. Method (4) slightly improves the registration accuracy (by 1%) compared to that of method (3), because of the incorporation of sliding window scheme, which reduces false positives. Though 1-D SURF and spectral centroid signatures have their own constraints, they balance each other very well; hence, their integrated usage in a

Table 4.12: Registration results for T1-T7 types. **MF:**% of perfectly matched frames

Attacks	**SURF (1)**	**SC (2)**	**SURF+ SC(3)**	**All (4)**	**CHE (5)**	**BA (6)**
	MF	**MF**	**MF**	**MF**	**MF**	**MF**
Zoom in	71.9	92.7	93.2	**93.2**	55.8	69.8
Slow motion	78.1	79.5	89.9	**90.4**	60.0	68.8
Fast forward	84.7	85.6	91.0	**91.0**	59.8	61.2
Pattern insertion	81.7	92.5	93.7	**94.2**	54.8	54.0
Moving caption	88.0	93.8	93.8	**93.8**	50.7	62.7
Rotation	84.6	92.8	95.2	**95.2**	68.9	59.8
Random noise	89.7	92.4	94.7	**94.8**	64.2	51.2

sliding window manner noticeably improves the registration accuracy. The improved results of method (4) shown in Table 4.12 prove this view point.

On the other hand, Chupeau et al.'s method yields poor results for moving caption and pattern insertion types in terms of low MF rates. This is because, inserting patterns or adding captions noticeably changes color histogram properties. The MF rate of Baudry et al.'s method declines sharply for random noise type. The reason is, adding random noise might alter the wavelet coefficients substantially, which leads to false fingerprints.

Table 4.13 lists the temporal registration accuracy of six compared methods for T8-T15 types in terms of MF rates.

Table 4.13: Registration results for T8-T15 types. **MF:**% of perfectly matched frames

Attacks	**SURF (1)**	**SC (2)**	**SURF+ SC(3)**	**ALL (4)**	**CHE (5)**	**BA (6)**
	MF	**MF**	**MF**	**MF**	**MF**	**MF**
Blurring	82.7	91.5	93.5	**94.2**	53.8	55.8
Brightness	90.0	95.8	95.9	**95.9**	62.0	59.7
Cropping	80.0	90.0	92.4	**92.4**	44.8	50.5
Picture-in-pic	75.4	92.3	92.9	**93.0**	39.7	42.0
3 combined	88.7	92.6	94.2	**94.4**	50.9	53.6
5 combined	89.1	92.7	92.8	**93.1**	53.9	44.8
Mp3	90.8	78.9	90.6	**90.6**	86.6	89.0
Single band	93.7	75.6	94.6	**94.7**	85.7	87.0

Method (4) generally performs well for all eight types and improves the MF rates

(up to 15%) compared to the reference methods. Method (4) slightly enhances the registration accuracy (by 1%) compared to method (3). The reason for this improvement is, when the sliding window scheme is utilized, query features are matched only with that of candidate segment and thus false positive rate is reduced. The MF rate of Chupeau et al.'s method is severely decreased for cropping and picture-in-picture types. This is because, cropping introduces black borders on top and bottom regions, that might generate very different signatures for master and query clips. In case of picture-in-picture type, insertion of picture produces different signature pattern for the query video compared to the original file.

On the other hand, Baudry et al.'s method yields poor MF rates for picture-in-picture and 5 combined types. In picture-in-picture type, there exists a discrepancy between the wavelet coefficients extracted from master and query videos, because of the insertion of a picture. This discrepancy leads to mismatches and thus reduces the accuracy of method (6). In case of 5 combined type, the wavelet coefficients vary widely after applying noise, cropping and pattern insertions and hence a lot of mismatches are retrieved. The accuracy of method (2) is sharply reduced for mp3 and single band compression types. Audio spectral features are much affected by these two types and hence MF rates decline sharply. Yet the proposed methods using SURF features (methods (1), (3) and (4)) are less affected by these two types.

Although the SURF and spectral features have their own advantages and limitations, they complement each other by their different characteristics; hence, the combination of local and spectral features not only improves the registration accuracy, but also widens the coverage to more number of transformations. The promising results of method (4) provide good evidence for supporting this viewpoint.

On the other hand, Table 4.12 and 4.13 results indicate that, the audio features-based methods (here methods (2)-(4)) generally score better registration results compared to the visual feature-based method (i.e. method (1)). In other words, the registration results are concluding that the audio features are performing better than spatial/SURF signatures; hence, acoustic fingerprints are to be preferred for pirate video registration task. However, this observed phenomenon is certainly not true, due to the reasons given below:

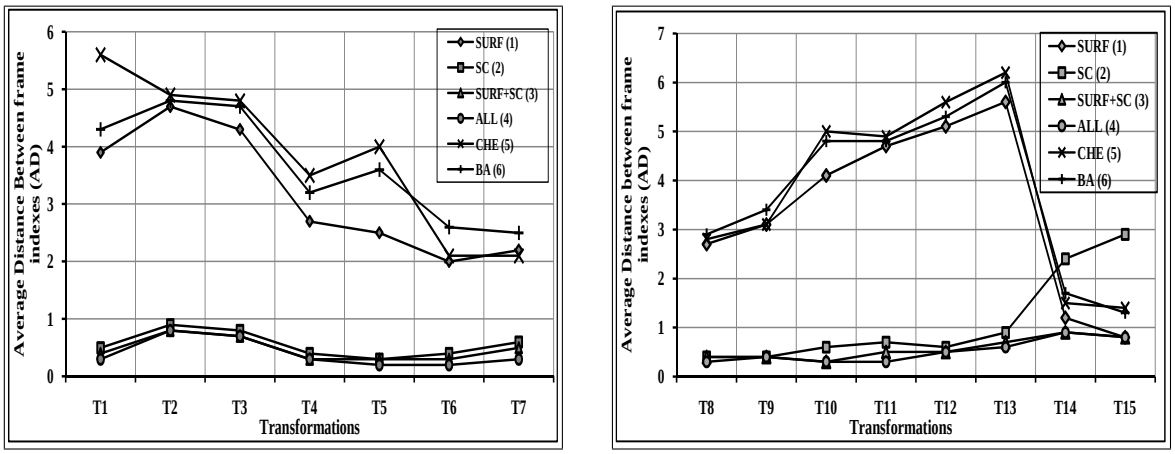
- ⋆ Most of the transformations considered in the proposed registration framework, fall under visual category of attacks (i.e. 13 out of 15 attacks are of visual type). It is well known that, audio features are not much affected by visual transformations such as blurring, noise, pattern insertion, moving caption and picture-in-picture. Therefore, acoustic fingerprints are performing better com-

pared to the visual features.

⋆ If more number of audio transformations are applied to the query video clips, then automatically the performance of the audio fingerprints might degrade to a greater extent.

⋆ Generally, for any video copy detection and tracking system, the performance mainly depends, not only on the selection of visual/audio fingerprints, but also the experimental setup as well as the type of transformations applied to the pirate video sequences. Precisely, visual fingerprints may achieve superior results against audio transformations and vice versa.

Figure 4.13(a) shows the registration results of six compared methods for T1-T7 types, in terms of Average Distance between true and estimated frame indexes ('AD'). The curves indicate the better performance of spectral signature based methods (methods (2), (3) and (4)), compared to other methods, because their AD rates are always less than 1. It is clear that, method (4) yields lowest AD rates and significantly improves accuracy compared to other methods. The combined utilization of robust visual and acoustic features in a sliding window manner is the exact reason for the enhanced performance of method (4).



(a) (b)

Figure 4.13: Comparison of AD curves for different transformations: (a) T1-T7 (b) T8-T15

Figure 4.13(b) indicates the registration results of six compared methods for T8-T15 types in terms of their AD rates. The curves show the superior performance of

method (4), compared to other methods because its AD rates are always less than 1. For T12 and T13 types, only visual features based methods (methods (1),(5) and (6)) indicate poor results in terms of higher AD rates. However, spectral signature based methods (methods (2),(3) and (4)) are less affected by this category.

Computational cost comparison

Table 4.14 shows the total time costs of methods (1)-(6), which includes signature extraction and frame matching costs. The program is executed in MATLAB and run on a PC with 2.8GHz CPU and 3GB RAM. The costs are measured by implementing frame alignment of a 298s query clip with a 3041s master sequence.

Table 4.14: Computational cost comparison (in seconds)

Process	**SURF (1)**	**SC (2)**	**SURF +SC(3)**	**ALL (4)**	**CHE (5)**	**BA (6)**
Signature extraction	68.1	21.1	90.1	90.1	47.4	59.8
Frame matching	108.0	87.4	95.0	4.1	66.8	74.1
Total cost	176.1	108.5	185.1	94.2	114.2	133.9

The signature extraction cost of method (4) is higher (up to 47%), compared to two reference methods. Interestingly, the frame matching cost of method (4) is noticeably reduced (up to 94%) compared to methods (5) and (6). This is because, in method (4) query clip signatures are aligned only with the corresponding candidate segment features instead of the entire master sequence. Thus in method (4), the usage of sliding window scheme significantly reduces the total time cost (up to 32%) and yields the lowest computational cost.

Registration results for CC_WEB_VIDEO dataset

Table 4.15 lists the registration results of five compared methods for first 12 master videos of CC_WEB_VIDEO dataset in terms of percentage of Incorrectly matched Frames ('IF'). The bold font indicates the lowest IF scores in the table.

In case of the first master video, the visual content is affected by distortions such as encoding format change and logo insertions; hence only visual feature based methods (methods (1), (4) and (5)) yield higher IF rates. For the second master video, few unrelated frames are added with same acoustic information; hence, method (2) leads to lot of false matches. However, characteristics of SURF and spectral features

Table 4.15: Registration results for 1-12 Master videos. **IF:** % of incorrectly matched frames

#	**Video Name**	**SURF (1)**	**SC (2)**	**SURF+ SC (3)**	**CHE (5)**	**BA (6)**
		IF	**IF**	**IF**	**IF**	**IF**
1)	The lion sleeps..	15.7	13.9	**10.4**	28.1	26.4
2)	Evolution of dance	24.4	42.3	**23.6**	29.5	30.4
3)	Fold shirt	27.5	14.6	**12.3**	38.4	40.9
4)	Cat massage	20.2	-	**20.2**	39.3	25.0
5)	Ok go here it..	40.5	34.0	**23.1**	55.6	53.9
6)	Urban ninja	38.2	46.6	**35.5**	38.2	39.4
7)	Real life Simpsons	41.4	52.6	**39.1**	43.3	42.2
8)	Free hugs	39.1	25.6	**20.6**	42.0	40.2
9)	Where the hell is Matt	21.6	13.3	**11.2**	29.1	34.2
10)	U2 and green day	12.6	14.9	**10.4**	28.1	21.6
11)	Little superstar	41.3	38.5	**33.6**	44.1	42.0
12)	Napoleon dynamite..	31.5	46.1	**27.5**	37.1	33.2

complement each other and hence method (3) improves accuracy and yields lowest IF rate for the second video.

In case of fourth master video, acoustic information is removed and captions are inserted to create the query videos; hence, method (2) leads to null matches. However, method (3) scores lowest IF rates, because of the robust nature of SURF-based visual signatures. For the fifth video, Chupeau et al. (2006) and Baudry et al. (2010) methods score poorly in terms of higher IF rates. The reason is, visual descriptors might be affected substantially due to the application of photometric and formatting variations such as color, lighting, frame rate and resolution changes. For the ninth video, Baudry et al.'s method gives highest IF rate compared to other methods. This is because, editing and encoding format changes widely vary wavelet coefficients and lead to lot of false positives.

Table 4.16 lists the registration accuracy of five compared methods for 13-24 master videos of CC_WEB_VIDEO dataset in terms of IF rates. Among all the methods, method (3) yields more accurate results against various types of formatting and editing attacks, due to the combined usage of visual and acoustic features.

Chupeau et al.'s method performs well for 22^{nd} and 13^{th} master videos but not as well for the 23^{rd} and 24^{th} master videos. This is because, color histograms are robust against lighting changes that are applied to the former videos, while the latter videos

Table 4.16: Registration results for 13-24 Master videos. **IF:** % of incorrectly matched frames

#	**Video Name**	**SURF (1)**	**SC (2)**	**SURF+ SC (3)**	**CHE (5)**	**BA (6)**
		IF	**IF**	**IF**	**IF**	**IF**
13)	I will survive Jesus	9.2	10.5	**7.2**	19.2	15.4
14)	Ronaldinho ping	12.5	11.1	**11.0**	20.9	19.5
15)	White and Nerdy	15.6	10.2	**10.0**	25.6	21.8
16)	Korean karaoke	26.0	35.5	**24.1**	32.5	31.0
17)	Panic at the disco...	18.3	17.3	**15.0**	22.6	25.5
18)	Bus uncle	27.2	42.6	**25.7**	28.5	32.6
19)	Sony Bravia	20.7	40.2	**30.1**	31.5	33.3
20)	Changes Tupac	35.1	20.6	**18.5**	40.7	38.2
21)	Afternoon delight	12.5	11.2	**11.1**	19.6	20.5
22)	Numa Gary	14.5	40.2	**14.1**	18.6	16.2
23)	Shakira hips dont..	40.3	36.2	**33.5**	41.3	42.5
24)	India driving	49.3	25.6	**23.1**	52.8	50.9

suffer from combined lighting and editing attacks.

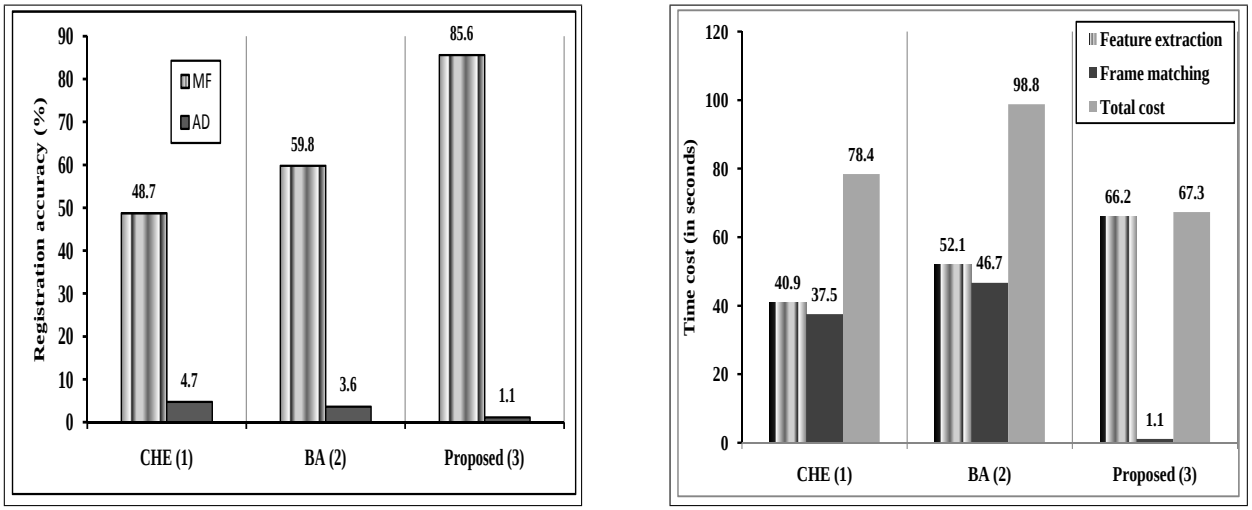
On the other hand, Baudry et al.'s method yields less accurate results for 24^{th} video in terms of higher IF rate. The reason is, 24^{th} video is modified by editing differences such as overlay text and addition of borders around frames, which in turn noticeably vary wavelet coefficients.

Registration results for camcorded videos

In the subsequent experiments, for comparison purpose the following three methods are evaluated:

(1) Chupeau et al.'s method (2006)('CHE');
(2) Baudry et al.'s method (2010)('BA');
(3) SURF+SC+sliding window for matching ('Proposed');

Figure 4.14(a) lists the registration results of three compared methods in terms of MF and AD rates. The proposed method gives extremely good results and improves the MF rates (up to 44%), compared to two reference methods. Although SURF and audio features have their own limitations, they balance each other; hence the integrated utilization of visual and acoustic features significantly improves the registration accuracy. The promising results of method (3) against heavily modified camcorded copies of master videos support this view point.



(a) (b)

Figure 4.14: Comparison of accuracy and time cost(a)MF & AD rates(b)Total time

Among all the three methods, the proposed method yields more accurate results, because the AD rate is lesser than the reference methods. Chupeau et al. (2006) and Baudry et al. (2010) methods score poor MF rates compared to the proposed method. The reason is, heavy cropping and compression distortions might substantially alter visual descriptors such as color histograms and wavelets coefficients-based signatures.

Figure 4.14(b) shows the total time costs of methods (1)-(3), which includes feature extraction and frame matching costs. The costs are measured by implementing the frame-to-frame mapping of a 215s query sequence with 2493s master sequence. Although the feature extraction cost of proposed method is higher, its frame matching cost is lower (by 97.5%) compared to the reference methods. This is because, query clip features are aligned only with the candidate segment instead of the entire master sequence. Thus in the proposed method usage of sliding window scheme noticeably reduces the total time cost up to 46.8% and provides the lowest computational cost.

Geometric registration results

Table 4.17 shows the geometric registration results of the proposed method for different video transformations in terms of mean and maximum pixel distances. Although the query video (i.e., camcorded version of the master video) is modified by heavy cropping, lighting and compression attacks; still the proposed method provides more accurate results in terms of low pixel distances. The spatial registration performance

of the proposed method is very efficient, because the mean pixel distance is always less than one. The robust nature of powerful SURF descriptors is the exact reason for this enhanced performance of the proposed method.

Table 4.17: Geometric registration results

Attacks	**Mean distance**	**Maximum distance**
Zoom in	0.60	1.20
Pattern insertion	0.62	1.30
Moving caption	0.62	1.24
Rotation	0.85	1.62
Random noise	0.63	1.30
Blurring	0.62	1.18
Brightness change	0.59	1.17
Cropping	0.63	1.41
Picture-in-picture	0.85	1.67
3 combined	0.62	1.12
5 combined	0.64	1.29
Camcording	0.69	1.23

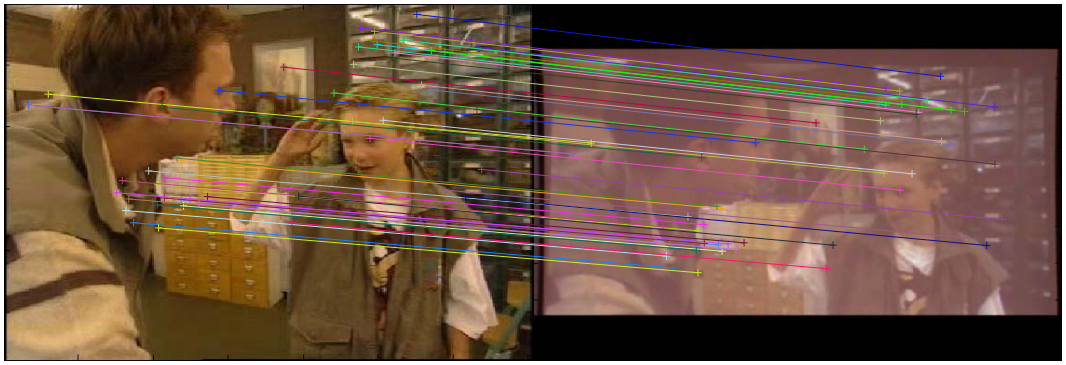


Figure 4.15: Pairs of matched interest points of candidate(left) and query(right) video frames; here, query is camcorded copy of the master video

For illustration purpose, temporally aligned master and query sequences are considered, which consists of 984 and 375 frames respectively. 74 principal frames are selected from the temporally aligned video segments using Algorithm 2 described in Figure 4.11 and utilized for the geometric alignment task. Figure 4.15 shows the geometrical mapping of the sample candidate and query frames, in which the extracted control points are highlighted with crosses. Here, query video is generated as the camcorded version of the master video.

The experiments conducted on different datasets demonstrate that, the proposed method consistently outperforms the reference methods for different types of transformations. The proposed registration framework achieves promising results in terms of higher MF and lower AD rates, by integrating visual and acoustic features for the registration task. Frame matching in a sliding window manner is another good characteristic of the proposed method which proves that, effective performance can be achieved with lowest computational cost, though the feature extraction cost is higher.

4.4 Summary

This chapter describes the scholarly contributions towards the video copy registration problem, which target accurate frame-to-frame alignments of the pirate video with master sequence.

More specifically, first this chapter contributes a new temporal registration scheme, that employs multimodal fingerprints derived from MFCCs and motion activity features for achieving temporal frame alignments of the pirate video with the master content. Followed by the temporal registration scheme, a robust spatio-temporal alignment framework is introduced, by employing SURF signatures in order to get accurate frame-to-frame alignments of the two video sequences. Though SURF descriptors are widely used in computer vision domain, introducing compact (1-D) SURF-based fingerprints is one of the major contributions of the proposed registration methods. Further, the proposed spatio-temporal visual signatures (SURF-based) are efficient when compared to the existing multidimensional SURF fingerprints, which define only the spatial content.

On the other hand, spatio-temporal alignment of a large master video with the small pirate clip is quite challenging in terms of computational cost. Therefore, identifying the most similar segment of the master video for the registration task is highly beneficial in terms of computational complexity. Due to this reason, the proposed registration frameworks also contribute, the most similar segment as well as most similar frames selection algorithms for enhancing the registration performance.

Inclusion of audio signatures in the alignment task, significantly improves the registration accuracy. Therefore, this chapter also presents a novel spatio-temporal registration framework, that utilizes new visual fingerprints derived from SURF key points and audio signatures extracted from spectral centroid features for obtaining accurate frame alignments of the two video sequences. To the best of our knowledge, this is the first attempt, which proposes a spatio-temporal registration framework

using multimodal features for obtaining the accurate frame alignments of the pirate video with the master sequence. The proposed spatio-temporal registration framework is evaluated on three different datasets, namely TRECVID sound & vision data, CC_WEB_VIDEO dataset and a set of real data comprising camcorded versions of master videos. Extensive evaluations on different datasets prove the efficiency and effectiveness of the proposed framework compared to the reference methods against various video transformations.

Related Publications

Conference Publications

1) R. Roopalakshmi and G. Ram Mohana Reddy, *Robust Features for Accurate Spatio-Temporal Registration of Video Copies*, in proc. of IEEE International Conference on Signal Processing and Communications (SPCOM-2012), Indian Institute of Science (IISc), Bangalore, India, pp. 1-5, July'2012.
   Available: `http://dx.doi.org/10.1109/SPCOM.2012.6290006`.

Journal Articles

1) R. Roopalakshmi, G. Ram Mohana Reddy, *A Novel Spatio-Temporal Registration Framework for Video Copy Localization Based on Multimodal Features*, published in **Elsevier Signal Processing** Journal, Vol. 93, Issue 8, Pages 2339-2351, Aug'2013. ISSN: 0165-1684.
   Available: `http://dx.doi.org/10.1016/j.sigpro.2012.06.004`.
2) R. Roopalakshmi and G. Ram Mohana Reddy, *Robust Temporal Registration Scheme for Video Copies Using Multimodal Features*, submitted to Springer Multimedia Systems.

Chapter 5

Geometric Distortions Estimation Framework

Followed by video copy registration, estimating the geometric distortions in a pirate video is prerequisite, in order to approximate the position of pirate in a movie theater during the illegal capture. Therefore, this thesis illustrates the scholarly contribution towards the geometric distortion estimation problem, in this chapter. Specifically, this chapter attempts to address the issues of existing distortion estimation techniques as described in Section 2.3.3, by contributing a new framework for geometric distortions estimation, which employs visual and acoustic features. More specifically, the proposed framework estimates geometric distortions in video copies, by incorporating novel visual fingerprints derived from SURF interest points and robust audio signatures extracted from MFCCs of video contents, which is described below.

5.1 Estimating Geometric Distortions in Video Copies

As illustrated in Section 2.3.3, state-of-the-art techniques are exploiting only visual features of videos for estimating the geometric distortions in watermarked video sequences; while no efforts are made towards acoustic features and non-watermarked video contents. To handle these issues, this chapter proposes a novel distortion model estimation framework, which exploits visual-audio fingerprints for the estimation task. Precisely, the main contributions of the proposed framework are given by,

- ⋆ A novel visual fingerprint denoted as *Compact Spatio-Temporal (CST) SURF* signature is introduced, which describes the spatial and temporal content of frames, when compared to the existing multi-dimensional SURF fingerprinting methods (Zhang et al. 2010; Yang et al. 2008).

- ⋆ The proposed framework exploits both the visual-audio signatures to achieve accurate temporal alignments and hence false frame matches are noticeably reduced.
- ⋆ A new and effective algorithm, for selecting the *Most Similar (MS)* segment of the master video using bipartite matching is contributed, which considerably reduces the frame matching cost.
- ⋆ An efficient algorithm to select *stable frame pairs* of pirate video and MS segments is introduced, which results in accurate geometric frame alignments.
- ⋆ The proposed framework also exploits distance measure and nearest neighbor mapping policies to obtain robust key point pairs, which are used to estimate the geometric distortions present in the duplicate video.

The proposed distortion estimation framework including temporal and geometric frame alignments followed by distortions estimation is illustrated as follows.

5.2 Proposed Distortions Estimation Framework

The proposed framework for estimating geometric distortions is shown in Figure 5.1., which includes three stages: In the first stage, compact visual-acoustic fingerprints are extracted from the master and pirate video sequences. Then, minimum weight perfect bipartite graphs are constructed to compute the exact frame-to-frame alignments of two video contents. More Precisely, master video segment with minimum matching cost is selected by employing visual-audio fingerprints and denoted as the *Most Similar (MS)* segment of the master video. After this step, pirate video and MS segment frames are mapped to get temporal frame-to-frame matches.

In the second stage, from the temporally mapped frames, a small set of frames denoted as *stable frame pairs* are extratced from the two video sequences. The resultant frame pairs are aligned using their SURF descriptors by means of control points, in order to achieve accurate geometric frame alignments. In the third stage, the proposed scheme extracts robust key point pairs from the spatially mapped frames by applying distance measure and nearest neighbor mapping policies. Finally, geometric distortion model is estimated by employing the spatial coordinates of the resultant key point pairs and Normalized Direct Linear Transformation (DLT) algorithm (Hartley and Ziserman 2004).

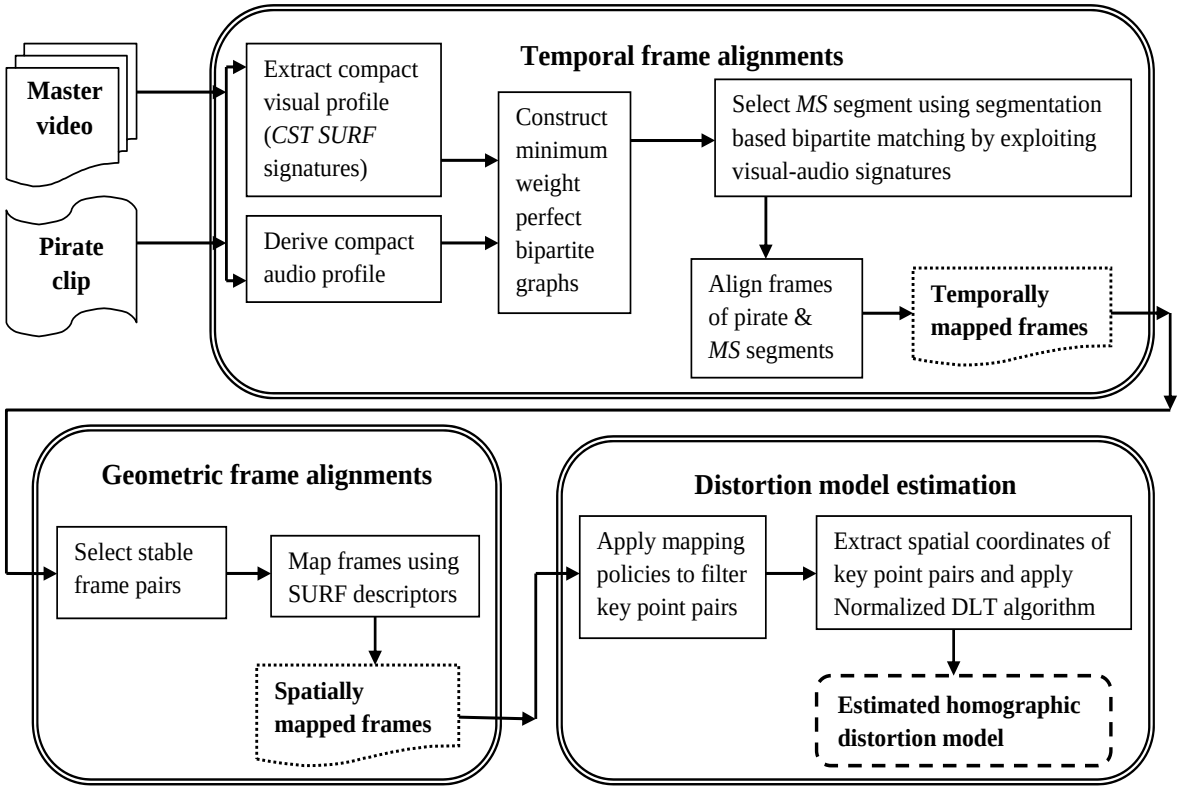


Figure 5.1: Proposed geometric distortions estimation framework

5.2.1 Temporal frame alignments

The proposed framework first computes compact visual-audio profiles of two video contents by exploiting CST SURF signatures and MFCCs respectively. The resultant feature sequences are aligned using bipartite matching in order to obtain accurate frame alignments of the pirate video with the master content, which is illustrated as follows.

Compact Visual Profile Extraction

As mentioned in Section 4.2.1, SURF is a scale and rotation invariant descriptor (Bay et al. 2008), which is popularly used in the copy detection literature to identify duplicate videos (Yang et al. 2008). In SURF, each interest point is associated with 64-D feature vectors (Bay et al. 2008). Further, each frame may contain multiple SURF interest points; hence direct comparison of SURF descriptors across all frames would be computationally expensive. On the other hand, current works employ multi-dimensional SURF fingerprints and consider only spatial content of frames (Yang et al. 2008). However, to generate a robust visual profile, spatial as well as temporal content

of frames need to be considered.

In order to solve these issues, this chapter introduces a novel visual signature, denoted as *CST SURF* signature, which effectively characterizes the spatio-temporal content of frames. Specifically, Figure 5.2 details the Algorithm 5.1, which is used to compute *CST SURF* signatures of the given video sequence. Figure 5.3 illustrates the steps involved in the computation of *CST SURF* signatures from the sample frames on a 2×2 partition.

Algorithm 5.1: *CST SURF* signatures Computation

1: Let $V = \{f_i | i = 1, 2, 3, ..., n\}$ be the video sequence comprising n frames, where f_i is i-th frame of V.

2: Segment the frame f_i into 2×2 regions, such that $f_i = \{r_i^j\}$, where $i \in [1 : n]$, $j \in [1 : 4]$.

3: Compute SURF key points of r_i^j of f_i and denote the count as Cr_i^j.

4: Calculate the differences between Cr_i^j and Cr_{i+1}^j, where $i \in$[1:n], $j \in$[1:4]. Then normalize the differences into [0:5] range.

5: Apply I-order Z-curves on normalized differences to obtain compact visual fingerprints.

6: The resultant fingerprints are denoted as Compact Spatio-Temporal (*CST) SURF* signatures.

Figure 5.2: *CST SURF* signatures computation

SURF descriptors are poor at handling illumination variations, specifically non-uniform types. However, the proposed *CST SURF* signatures consider only the differences between region-wise count of SURF key points. In addition, the resultant differences are normalized to compute *CST SURF* signatures. Therefore, the proposed fingerprints are less affected by spatially varying illumination changes and guarantee reasonable registration accuracy.

Compact Acoustic Profile Extraction

As described in Section 3.4.2., Mel-Frequency Cepstral Coefficients (MFCCs) are robust and highly discriminative spectral features; hence they are popularly used in speech recognition and multimedia content analysis applications (Rabiner and Juang 1993; Park 2010). In addition, frequency warping in MFCCs represent the perceptual features of sound signals very well. Due to these reasons, perceptually robust

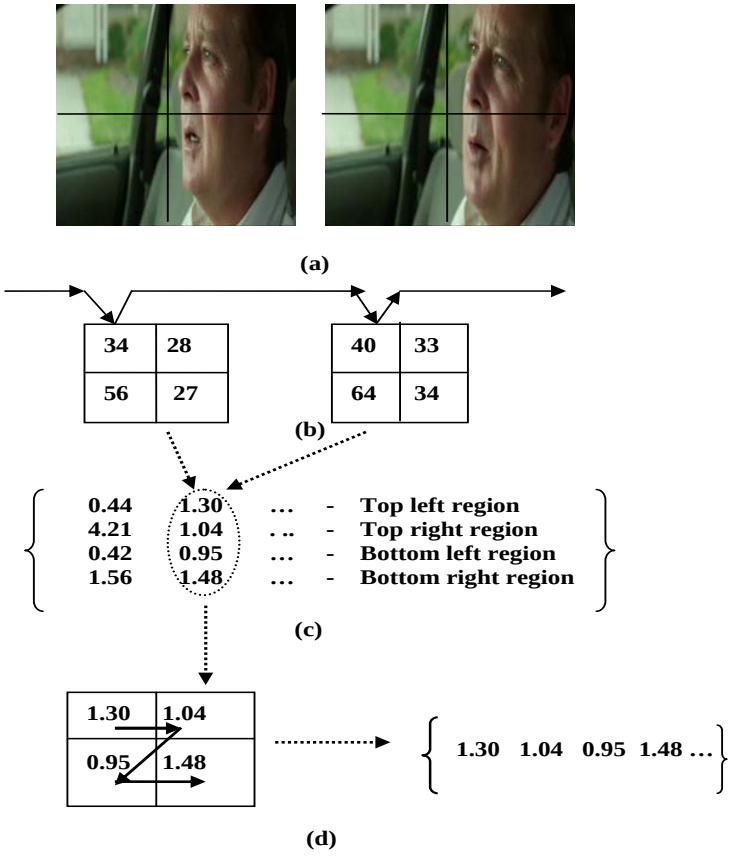


Figure 5.3: CST-SURF signatures computation. (a) Video frames partitioned into 2×2 regions. (b)Computing differences between region-wise count of SURF key points. (c) Region-wise normalized differences. (d) Applying I-order Z-curves and computing CST SURF signature with time series

MFCCs are employed to extract the compact audio profile of video sequences, which is illustrated as follows.

First an audio signal is down sampled to 22050 Hz and segmented into 11.60ms windows with an overlap factor of 70% using Hamming window function (Roopalakshmi and Reddy Sep-2011). The MFCCs calculation results in a M×N matrix, where M indicate the number of frames and N consists of 13 MFCC features of a frame. This M×N matrix can be effectively summarized using Singular Value Decomposition (SVD) technique as $A = USV^T$, where A is M×N input matrix to be summarized, S is an M×N diagonal matrix consisting of the singular values of A.

The proposed framework employs 4 to 6 singular values for extracting acoustic signatures of video contents. Further, in order to improve the robustness of audio signatures against various media distortions, normalized singular values are utilized in this study. Figure 5.4 details the steps used to compute the compact acoustic

signatures from the spectrum of audio signals.

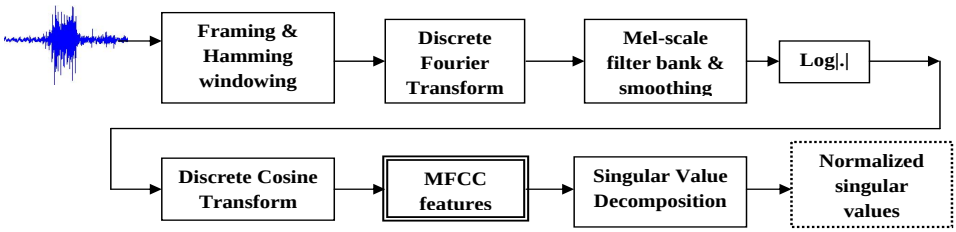


Figure 5.4: Compact acoustic signatures extraction

Introduction of Bipartite Matching

In order to obtain the temporal frame alignments of two video sequences, the frame alignment problem is reduced into a graph theoretic problem called as *Minimum Weight Perfect Bipartite Matching (MWPBM)* (Chartrand 1977; Goemans 2007), which is illustrated below.

A graph $G = (V, E)$ is said to be bipartite, if the vertex set V can be partitioned into two sets $V1$ and $V2$, such that no edge in E has both endpoints in the same set of bipartition. A matching $M \subseteq E$ is a collection of edges such that every vertex V is incident to at most one edge of M. A matching M is said to be *perfect* if its cardinality is equal to $|V1| = |V2|$. Given a cost C_{ij} for all $(i, j) \in E$, the cost of matching M denoted as $C(M)$ is given by,

$$C(M) = \sum_{(i,j)\in M} C_{ij} \tag{5.1}$$

Minimum Weight Perfect Bipartite Matching (MWPBM) technique computes a perfect matching M_{min} with minimum matching cost $C(M)_{min}$ which is formulated as,

$$C(M)_{min} = min\{C(M)\} \tag{5.2}$$

$$M_{min} = \{M_{ij}|\ C(M_{ij}) == C(M)_{min}\},\ (i, j) \in E \tag{5.3}$$

In other words, the goal of MWPBM is to find a perfect matching M, which minimizes $C(M)$, for every $i \in V1$, $j \in V2$.

Frame Alignments Using MWPBM

The proposed framework constructs a weighted bipartite graph and computes frame alignments of two video contents using MWPBM as detailed below.

Definition 1 (Video as set of frames): Let a master video MV is represented as a group of frames: $MV = \{mf_1, mf_2, ..., mf_{n_{mv}}\}$, while a pirate video PV is denoted as a set of frames: $PV = \{pf_1, pf_2, ..., pf_{n_{pv}}\}$. Here, n_{mv} and n_{pv} indicate number of master and pirate video frames respectively.

Definition 2 (Weighted bipartite graph): A weighted bipartite graph is said to be a bipartite graph G such that,

$$G = \{V, E, W\} \tag{5.4}$$

where $V = \{MV \cup PV\}$, $E = \{MV \times PV\}$ and $W = \{w(i,j) | w(i,j) = Dist(mf_i, pf_j)\}$. Here, $Dist$ indicates the distance between feature sequences of video frames mf_i and pf_j respectively, which is explained below.

Let VF_{mv} be the visual signature of the master video MV such that, $MV \in \{VF^i_{mv} |\ 1 \leq i \leq n_{mv}\}$ and VF_{pv} be the visual signature of the pirate video PV such that, $PV \in \{VF^j_{pv} |\ 1 \leq j \leq n_{pv}\}$. The distance $Dist_{vis}$ between the two visual feature sequences is computed using comparative Manhattan distance as follows,

$$Dist_{vis}(VF_{mv}, VF_{pv}) = \frac{|(VF^i_{mv}) - (VF^j_{pv})|}{|(VF_{mv_i})| + |(VF_{pv_j})|} \tag{5.5}$$

where $i \in [1 : n_{mv}]$ and $j \in [1 : n_{pv}]$ respectively.

Consider AF_{mv} be the acoustic feature sequence of MV such that, $MV \in \{AF^k_{mv} |\ 1 \leq k \leq n_{mv}\}$ and AF_{pv} be the audio signature of PV such that, $PV \in \{AF^r_{pv} |\ 1 \leq r \leq n_{pv}\}$. The distance $Dist_{aud}$ between the acoustic feature sequences of master and pirate videos is computed using squared Euclidean distance as follows,

$$Dist_{aud}(AF_{mv}, AF_{pv}) = |(AF^k_{mv} - AF^r_{pv})^2| \tag{5.6}$$

where $1 \leq k \leq n_{mv}$ and $1 \leq r \leq n_{pv}$ respectively.

Figure 5.5 shows the weighted bipartite graph used to model the master and pirate video segments. Video frames form the vertices and the difference between their feature sequences provide edge weights. The frame correspondences in the bipartite graph is generated using Hungarian algorithm (Kuhn 1955).

Definition 3 (MWPBM): In the proposed framework, a perfect matching M is computed, that provides minimum matching cost for every $mf_i \in MV$ and $pf_j \in PV$ using MWPBM technique. More precisely, MWPBM computes a perfect matching M_{min} with minimum matching cost $C(M)_{min}$, which provides lowest cost for every $mf_i \in MV$ and $pf_j \in PV$. Figure 5.5 indicates the sequence of frame pairs in dotted

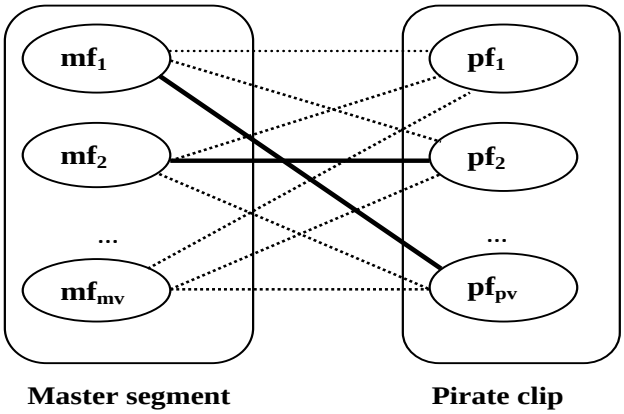


Figure 5.5: MWPBM technique to compute frame alignments. Dotted lines indicate sequence of frame pairs and darker lines represent the frame alignments with lowest possible cost.

lines and the frame alignments with the lowest possible cost in darker lines.

Segmentation-Based MWPBM

Mapping the pirate clip frames with the entire master video is computationally expensive, if the size of master video is large. More precisely, in order to align a master sequence of n_{mv} frames with the pirate clip having n_{pv} frames, MWPBM method computes an edge set $E \subseteq \{n_{mv} \times n_{pv}\}$. Hence if n_{mv} increases, the computational complexity also increases.

In order to tackle this problem, this chapter computes the *Most Similar (MS)* segment of the master video such that, the distance between the visual-audio feature sequences of MS and the pirate segments is minimum. The Algorithm 5.2 given in Figure 5.6 details the steps used to select the MS segment of the master sequence. In (Roopalakshmi and Reddy 2013), sliding window based DTW technique is utilized to compute the similar segment of the master video. However, length and overlapping factor of a sliding window play a vital role in determining system accuracy and computational cost. To solve this discrepancy, segmentation technique is exploited in the proposed framework to select the MS segment of master video. Once the MS segment is selected, then the pirate video and MS segment frames are mapped to each other, in order to obtain accurate temporal frame alignments. In this way, the joint utilization of visual-audio fingerprints significantly reduces false frame mappings.

Algorithm 5.2: *Most Similar(MS) Segment* Selection

1) Divide the master video into non-overlapping segments of length equal to the pirate clip. The last segment having total frames < length of pirate clip is padded with zero valued frames.

2) Let the master video MV be,

$$MV \in \{S_k |\ 1 \leq k \leq ns\}, \tag{5.7}$$

where S_k is the k-th segment and ns is total segments of MV. Here, each segment S_k of MV can be compactly represented as,

$$S_k \in \{\{V_{s_k}^i \cup A_{s_k}^j\} \mid 1 \leq i \leq nv_{s_k},\ 1 \leq j \leq na_{s_k}\} \tag{5.8}$$

where $V_{s_k}^i$ is i-th visual feature of S_k and nv_{s_k} indicates total visual signatures of S_k. Here, $A_{s_k}^j$ is j-th audio feature sequence of S_k and na_{s_k} represents total acoustic feature sequences of S_k.

3) Let the pirate video PV is compactly described as,

$$PV \in \{\{V_q^k \cup A_q^r\} \mid 1 \leq k \leq nv_q,\ 1 \leq r \leq na_q\} \tag{5.9}$$

where V_q^k is the k-th visual feature vector and nv_q is total visual features of PV. Here, A_q^r is r-th audio feature sequence and na_q indicates total feature vectors of PV.

4) Compute the distance between the visual-audio feature sequences of S_k and the pirate video PV, where $k \in [1 : ns]$.

5) Select the segment S_k having minimum matching cost (i.e. $C(M)_{min}$) for both the visual and audio feature sequences as the Most Similar (MS) segment of MV, which is formulated as,

$$\begin{aligned} Sim_{seg}(S_k, PV) = C(M)_{min}(V_{s_k}, V_q) + \\ C(M)_{min}(A_{s_k}, A_q) \end{aligned} \tag{5.10}$$

where $C(M)_{min}(V_{s_k}, V_q)$ indicates the cost of minimum weight perfect matching M, which aligns the two visual feature sequences V_{s_k} and V_q respectively. Here, $C(M)_{min}(A_{s_k}, A_q)$ indicates the cost of perfect matching M, which maps the two acoustic feature vectors A_{s_k} and A_q of S_k and PV segments respectively.

Figure 5.6: *Most Similar(MS) segment* selection algorithm

5.2.2 Geometric frame alignments

From the previous section, temporally aligned frame pairs of pirate and master video sequences are obtained. This section maps the resultant frame pairs by means of

their SURF descriptors to get accurate geometric frame alignments. It is not feasible to estimate the geometric distortions from all temporally aligned frames due to computational load. Further, all video frames may not provide essential control points to estimate accurate geometric distortions. On the other hand, frame misalignments may considerably degrade the distortion estimation accuracy of the proposed framework.

In order to solve these problems, the proposed framework selects a subset of temporally aligned frame pairs denoted as *stable frame pairs* for estimating geometric distortions. The edge weights and frame correspondences computed by MWPBM algorithm supply important guidelines for selecting *stable frame pairs* of two video contents. Algorithm 5.3 specified in Figure 5.7, describes the steps used to compute the *stable frame pairs* of MS and pirate video segments. Finally the resultant frame pairs are mapped by means of their SURF descriptors in order to obtain accurate geometric alignments of master and pirate video contents.

Algorithm 5.3: *Stable Frame Pairs* Selection

a) Let MS be the most similar segment with n_{ms} frames such that $MS \in \{mf_1, mf_2, ..., mf_{n_{ms}}\}$ and let PV be the pirate video segment with n_{pv} frames so that $PV \in \{pf_1, pf_2, ..., pf_{n_{pv}}\}$.

b) MWPBM computes a perfect matching M_{min} of MS and PV segments with minimum matching cost $C(M)_{min}$. Specifically, M_{min} can be represented as,

$$M_{min} \in \{\{mf_i, pf_j\} |\ 1 \leq i \leq n_{ms},\ 1 \leq j \leq n_{pv}\} \tag{5.11}$$

where mf and pf indicate the frame pairs of MS and pirate segments with minimum cost.

c) Select all the frame pairs of M_{min} satisfying the given criteria as *stable frame pairs SFP* of two video contents, which is formulated as,

$$SFP \in \{\{mf_i, pf_j\} |\ Dist(mf_i, pf_j) \leq (C(M)_{min}/2)\} \tag{5.12}$$

Figure 5.7: *Stable Frame Pairs* selection algorithm

5.2.3 Geometric distortions estimation

The geometric frame alignments stage results in a list of key point pairs extracted from pirate and master video sequences. In this section, first the stable and robust

interest point pairs from the entire set of matched key points are obtained by utilizing two filtering policies. Then the proposed framework exploits spatial coordinate values of resultant key point pairs and Normalized DLT algorithm to estimate the geometric distortions in the pirate video, which is illustrated below.

Key points Filtering

Blind comparison of all feature descriptors of a frame pair is not feasible due to the computational cost. Further, direct comparison of all SURF descriptors of frame pairs leads to false mappings. In order to tackle these discrepancies, this chapter employs two policies for selecting robust key point pairs, which are described below.

Policy 1: The mapping criteria of two key points is determined using Squared Euclidean distance measure as follows,

$$dis_{des}(d_i, d_j) = |(d_i - d_j)^2| \tag{5.13}$$

where d_i, d_j indicate the SURF descriptors derived from the MS and pirate video frames; and dis_{des} represents the distance between two feature descriptors d_i and d_j respectively. Two key points are matched only if their $dis_{des} \leq$ threshold value, which is set as 0.40 in experiments.

Policy 2: The proposed framework employs nearest neighbor mapping strategy to filter out false key point matches. For each key point, the first and second nearest neighbors with minimum distances are listed. Then the role of frame pairs is inverted and the nearest neighbors are computed. If same correspondence is obtained for a given key point pair in both the lists, then the resultant key points are matched and retained; else the key point pair is discarded. After this step, the spatial locations of resultant key point pairs are employed for estimating the geometric distortions in the pirate video.

Distortion Estimation

As mentioned in Section 2.3.1., in case of a camcorder capture in a theater the resulting images are coupled with severe geometric distortions, because the camcorder capturing axis is not perpendicular to the screen. The resultant geometric distortions can be described by perspective projection, which models the imaging process of a pinhole camera. Projective transformation transforms a square into an arbitrary quadrilateral, in which distance between the points and angle between the lines are not preserved (Hartley and Ziserman 2004).

Let $x_1 = (x_1, y_1, 1)^T$ be the homogeneous vector, which represents spatial coordinate values of a key point in MS segment frame and $x_2 = (x_2, y_2, 1)^T$ be the homogeneous vector that represents coordinate values of an interest point in the pirate segment frame. A planar projective transformation is a linear transformation on homogeneous 3-vectors represented by a non-singular 3×3 matrix as (Hartley and Ziserman 2004),

$$x_2 = \mathbf{H}x_1, \ where \ \mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{5.14}$$

Here **H** is a homogeneous matrix. Amongst the nine elements of **H**, there are eight independent ratios and thus a projective transformation has eight degrees of freedom (DOF) (Hartley and Ziserman 2004). Thus four 2-D to 2-D point correspondences from MS and pirate segment frames are required to estimate the 8-parameter homographic matrix **H**.

In this study, the coefficients of **H** are determined using estimated point correspondences and Normalized DLT algorithm (Hartley and Ziserman 2004). DLT is a simple linear algorithm used to determine the solution for **H**, which is not invariant to similarity transformations of the image. Precisely, the results of DLT algorithm depends on the coordinate frame in which the points are expressed. Normalized DLT algorithm provides a solution to this problem, by including initial data normalization procedure in the basic DLT algorithm. In this way, Normalized DLT algorithm computes the coefficients of **H**, which is invariant to arbitrary choices of scale and coordinate origin. The data normalization of Normalized DLT algorithm includes these steps: a) The points are translated so that their centroid is at the origin; b) Then the points are scaled such that the average distance from the origin is equal to $\sqrt{2}$ and c) This transformation is applied to each of the two images independently. After this normalization step, DLT algorithm is applied to determine the coefficients of **H**.

Lee et al. (2010) used corner points of two video frames and DLT algorithm to estimate the homographic matrix **H**. The proposed framework utilizes spatial locations of stable and robust key points for estimating **H**. Furthermore, the framework employs Normalized DLT algorithm to estimate **H**, which is more accurate compared to DLT algorithm (Hartley and Ziserman 2004). In order to derive a linear solution

for **H**, Equation (5.14) is expressed in terms of vector cross product as,

$$X_i' \times \mathbf{H}X_i = 0 \tag{5.15}$$

where $X_i' = (x_i', y_i', w_i')$ and if the j-th row of the matrix **H** is denoted by h_j^T, then we can write,

$$\mathbf{H}X_i = \begin{pmatrix} h_1^T X_i \\ h_2^T X_i \\ h_3^T X_i \end{pmatrix} \tag{5.16}$$

The cross product in (5.16) is given explicitly as,

$$X_i' \times \mathbf{H}X_i = \begin{pmatrix} y_i' h_3^T X_i - w_i' h_2^T X_i \\ w_i' h_1^T X_i - x_i' h_3^T X_i \\ x_i' h_2^T X_i - y_i' h_1^T X_i \end{pmatrix} \tag{5.17}$$

Since $h_j^T X_i = X_i^T h_j$ for $j \in \{1, 2, 3\}$, this gives a set of three equations in the entries of **H**, which may be written in this form,

$$\begin{bmatrix} 0^T & -w_i' X_i^T & y_i' X_i^T \\ w_i' X_i^T & 0^T & -x_i' X_i^T \\ -y_i' X_i^T & x_i' X_i^T & 0^T \end{bmatrix} \; h = 0, \; h = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} \tag{5.18}$$

Though there are three equations in (5.18), only two of them are linearly independent, since these equations include homogeneous vectors (Hartley and Ziserman 2004), which is written as,

$$A_i h = 0 \tag{5.19}$$

where A_i is the 2×9 matrix of (4.18) and $i \in \{1, 2, 3, 4\}$. Given four 2-D to 2-D point correspondences, the rows of A_i are arranged into a single matrix A as,

$$Ah = 0, \; where \;\; A \; = \begin{pmatrix} A1 & A2 & A3 & A4 \end{pmatrix}^T \tag{5.20}$$

In order to obtain the solution of (5.20), SVD of A is computed as,

$$A = UMV^T \tag{5.21}$$

where M is the diagonal matrix containing singular values in the descending order, U is an orthogonal matrix of size 8×8 and V is an orthogonal 9×9 matrix. The last column of V provides the values of h, from which matrix **H** can be determined.

5.2.4 Performance evaluation and results

The proposed algorithm is evaluated on a real data set consisting of popular movies and their camcorded copies. Table 5.1 lists the movies that constitute the master data set of the proposed framework and Figure 5.8 shows the snapshot examples of master videos.

Table 5.1: Master dataset

#	**Movie Title**
1	Aliens Vs Avatars
2	Alvin And The Chipmunks
3	God Bless America
4	Intruders
5	Journey2 The Mysterious Island
6	Mission Impossible: Ghost Protocol
7	Titanic

Figure 5.8: Snapshot examples of the master dataset

The query set consists of totally 84 camcorded copies with different distortions and the duration of copies vary between 40s to 3min. Precisely, the quality of camcorded versions range from clean copies to heavily modified ones with a large amount of noise, zooming in and perspective distortions. Further, resampling procedure is utilized in order to synchronize master and pirate video sequences, where the frame rate is set as 25 frames/sec.

Temporal registration results

The following six methods are evaluated:

(1) CST SURF signatures (abbreviated as 'SU');
(2) Compact audio signatures from MFCCs ('MF');
(3) SU + MF without segmentation ('SU+MF');
(4) SU + MF + segmentation ('SU+MF+SE');
(5) Chupeau et al.'s method (2006) ('CH');
(6) Chen and Stentiford's method (2008) ('CS');

The methods (1)-(4) evaluated different combinations of the proposed techniques. Methods (1) and (2) employ visual and acoustic fingerprints (namely SURF and MFCCs) for implementing the temporal registration task. Methods (3) and (4) are executed in order to prove the effect of segmentation scheme in the proposed framework.

In method (1), CST SURF signatures of the pirate video are temporally mapped with the corresponding features of the entire master sequence. The compact acoustic fingerprints of pirate clip are matched with the respective audio signatures of the master sequence in method (2). In method (3) visual-audio fingerprints of the pirate clip are separately aligned with the corresponding features of the entire master sequence. Method (4) temporally aligns the multimodal signatures of the pirate video with the MS segment instead of entire master sequence.

Chupeau et al. (2006) used color histograms for computing temporal frame alignments, which is executed as follows: From consecutive frames, histograms of size 512 bins are generated. Then distances between color histograms of successive frames and dynamic programming are utilized to get temporal frame alignments. Chen and Stentiford (2008) employed ordinal measure for matching pirate and master video sequences, which is widely popular in Content-based video copy detection literature. In this method, a video frame is divided into 2×2 grids and the corresponding temporal ordinal measure is computed based on the grey values of ordinal ranking matrix, to obtain precise temporal localization of frames.

Figure 5.9 shows the temporal registration results of six compared methods in terms of percentage of Perfectly Matches Frames (PMF). The performance of methods (2)-(4) is superior compared to other methods in terms of higher PMF rates. This is because, methods (2)-(4) employ MFCCs, which are robust against visual content-based distortions such as zooming in, noise and camcording. Method (4) slightly enhances PMF rates (by 1.6%) when compared to method (3). This is because, when the segmentation scheme is employed, pirate clip features are matched only with the MS segment features instead of the entire master sequence; hence lot of false positives are reduced. Method (4) scores well in terms of higher PMF rates and improves the registration accuracy up to 19.3% compared to the reference methods. Although

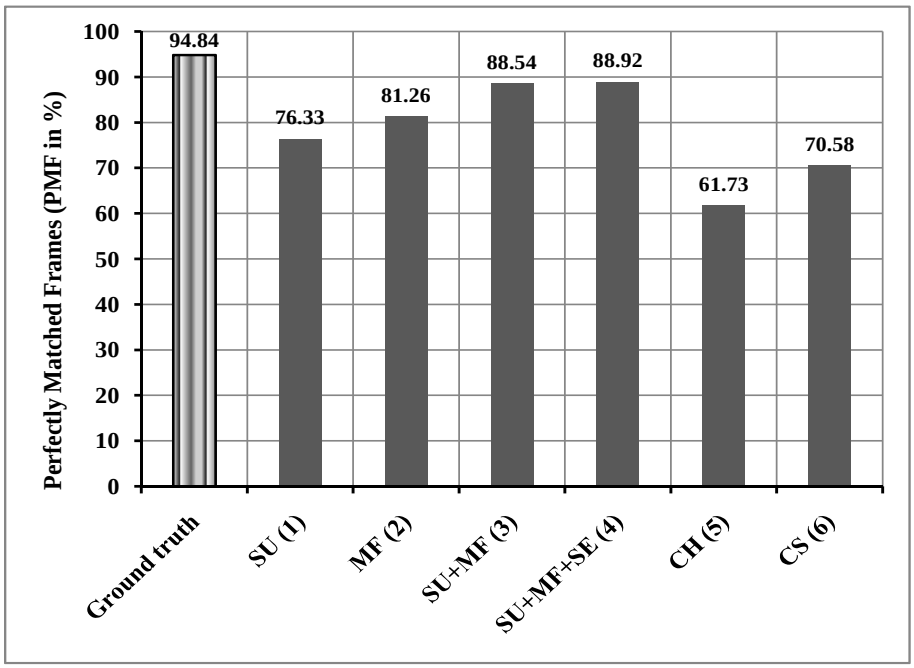


Figure 5.9: Temporal alignment results of methods (1)-(6). **PMF rates**: % of perfectly matched frames

SURF and MFCCs have their own advantages and limitations they complement each other very well; hence, their combination along with the segmentation scheme significantly improves the registration results. The enhanced results of method (4) shown in Figure 5.9 provide a good evidence to support this view point.

On the other hand, Chupeau et al.'s method achieves poor results in terms of low PMF rates. The reason is, noise and camcording distortions noticeably change the histogram properties, hence registration accuracy is substantially reduced. Chen and Stentiford's method scores low PMF rates, since camcording distortions alter the ordinal measure substantially, which leads to lot of false frame matches. Further, cropping introduces black borders on bottom and top regions, which generates different ordinal signatures for pirate and master sequences.

Figure 5.10 shows the temporal registration results of six compared methods in terms of Average Distance between true and estimated frame indexes (AD). Figure 5.10 results indicate the superior performance of method (4) compared to two reference methods in terms of lower AD rates. Methods (1),(5) and (6) score poor results in terms of higher AD rates. This is because, the methods using only visual signatures are much affected by transformations such as noise, cropping and camcording. However, MFCC signatures based methods (methods (2)-(4)) are less affected by these transformations; hence they score better AD rates.

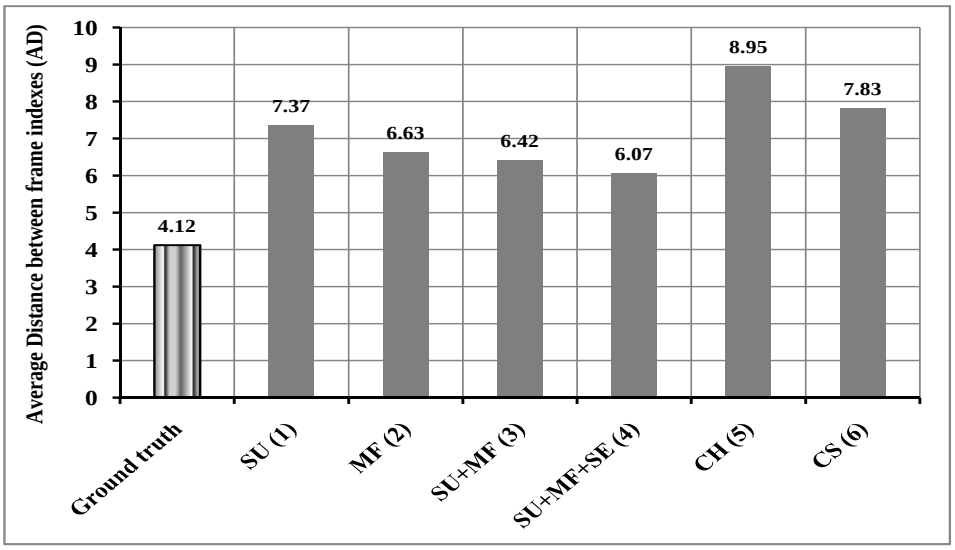


Figure 5.10: Temporal alignment results of methods (1)-(6). **AD rates**: average distance between true & estimated frame indexes

Computational cost comparison

The proposed framework is executed in MATLAB and run on a PC with 2.8GHz CPU and 3GB RAM. Table 5.2 shows the total computational costs of methods (1)-(6), including fingerprinting extraction and frame alignment costs. The time costs are measured by executing temporal frame alignment of a 108s query clip with a 4862s master sequence.

Table 5.2: Comparison of Computational Cost

Task	**SU(1)**	**MF (2)**	**SU+MF(3)**	**SU+MF +SE(4)**	**CH(5)**	**CS(6)**
Fingerprints extraction	46.10	25.81	58.63	58.63	39.42	40.86
Frame alignment	33.64	11.27	27.80	8.26	30.35	29.17
Total cost	79.74	37.08	86.43	66.89	69.77	70.03

The fingerprint extraction cost of methods (3) and (4) are slightly higher compared to the two reference methods. This is because, methods (3) and (4) exploit both the visual and acoustic features. However, the frame alignment cost of method (4) is considerably reduced (up to 72.5%) when compared to the two reference methods. The reason for this drastic reduction is, in method (4) the multimodal features of the pirate segment are mapped only with respective features of the MS segment rather

than the entire master sequence. Thus in the proposed method, the usage of segmentation scheme decreases the total time cost significantly and enhances effectiveness of the proposed framework.

Geometric Alignment Results

In the subsequent experiments, for comparison purpose the following three methods are evaluated:

(1) Chupeau et al.'s method (2006)('CH');
(2) Chen and Stentiford's method (2008)('CS');
(3) SU +MF+ segmentation ('Proposed');

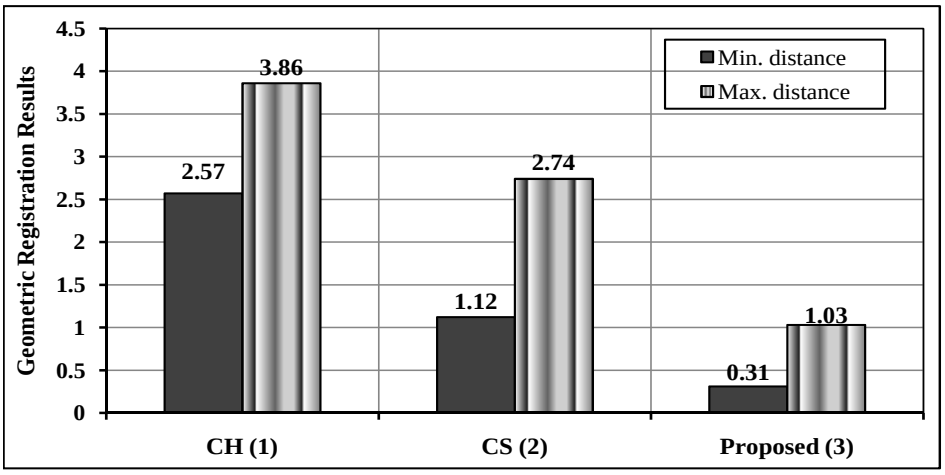
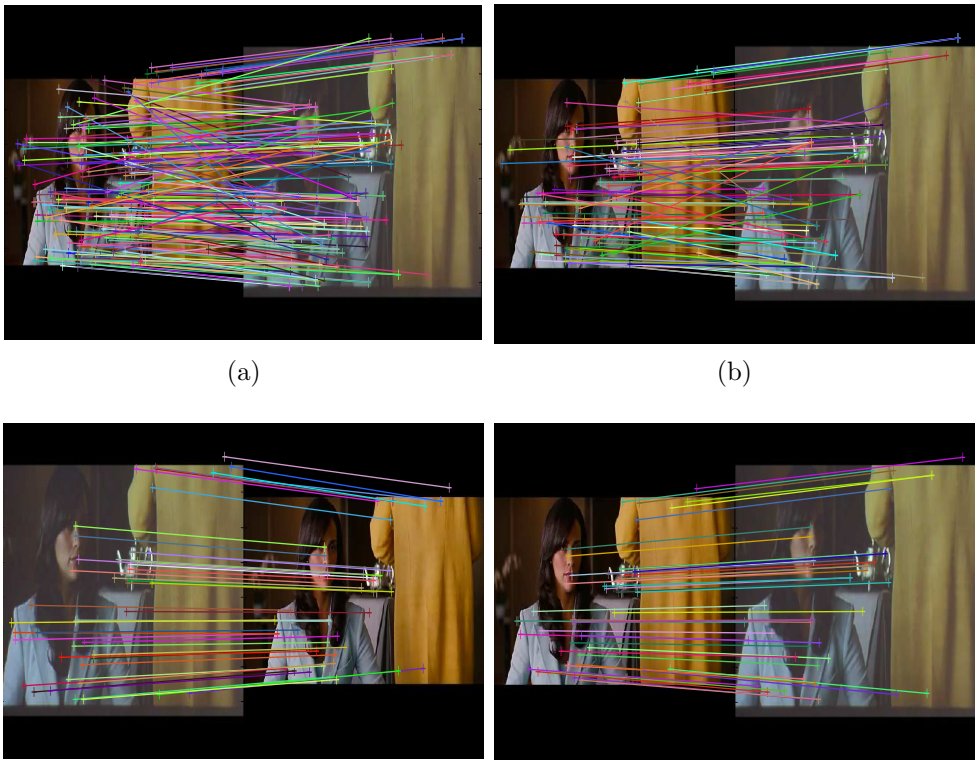


Figure 5.11: Geometric alignment results: Minimum & maximum pixel distances

Figure 5.11 indicates the geometric registration results of three compared methods in terms of minimum and maximum pixel distances. The performance of the proposed method is very accurate, because the minimum pixel distance$<$ 0.5. Though the pirate video is heavily modified by camcording, yet the proposed framework gives more accurate results in terms of lowest pixel distances. The reason is, utilization of *stable frame pairs* and their robust SURF descriptors for the geometric alignment task.

Distortion Estimation Results

For illustration purpose, spatio-temporally aligned frame pairs shown in Figure 5.12 are considered. Initially 302 and 197 key points are extracted from MS and pirate segment frames respectively, which is shown in Figure 5.12(a). Policy 1 described in



(a) (b)

(c) (d)

Figure 5.12: Sample frames from MS (left) and pirate (right) segments. (a) All key point pairs. (b) After applying Policy 1. (c) After applying Policy 2. (d) Final list of matched key point pairs

Section 5.1.4 is applied and hence the number of key point matches of the two video segments are reduced from 302 to 244 and 197 to 140 respectively, which is given in Figure 5.12(b(b). After this step, the proposed framework applies Policy 2 in order to retain the stable key point pairs, which in turn decreases the key point matches from 140 to 86 and 244 to 153 respectively, as shown in Figure 5.12(c). Finally, 44 robust and stable matched key point pairs shown in Figure 5.12(d) are extracted and spatial coordinate values of resultant key point pairs are employed to estimate the distortion model. Precisely, geometric distortions in the pirate video frame is estimated and represented using homogeneous matrix $\mathbf{H}$ as given by,

$$\mathbf{H} = \begin{pmatrix} 0.4081 & 0.1467 & -8.4936 \\ -0.1087 & 0.7377 & -20.5361 \\ -0.0007 & 0.0007 & 0.5645 \end{pmatrix} \tag{5.22}$$

$$GT = \begin{pmatrix} 0.5062 & 0.0161 & -10.9197 \\ -0.0182 & 0.6012 & -13.3173 \\ -0.0002 & 0.0000 & 0.6206 \end{pmatrix} \tag{5.23}$$

where (5.23) indicates ground truth values (GT). The estimation results clearly demonstrate the accuracy of the proposed framework, since the estimated distortions slightly deviate from the ground truth values.

5.3 Summary

This chapter describes the scholarly contribution towards the geometric distortion estimation problem, by presenting a new framework, which exploits visual and audio features. To the best of our knowledge, this is the first attempt, which employs visual-audio fingerprints for estimating geometric distortions in pirate video sequences.

Specifically, the proposed distortion estimation framework jointly utilizes novel visual fingerprints derived from SURF key points and robust audio signatures extracted from MFCCs of video sequences for the estimation task. Further, presenting algorithms to select *Most Similar (MS)* segment of the master video as well as *Stable Frame Pairs* of two video sequences, are some of the main contributions of the proposed framework. Furthermore, the proposed distortion estimation framework is evaluated on a real dataset consisting of 7 popular movies and their camcorded versions. Experimental results demonstrate the promising results of the proposed framework compared to the two reference methods.

However, frame misalignments may degrade the distortion estimation accuracy of the proposed framework. To handle this issue, a sub-set of temporally aligned frame pairs denoted as *Stable Frame Pairs* are utilized for estimating the geometric distortions. Though, robust key point pairs matching algorithms such as Least Median of Squares or RANSAC may enhance the estimation accuracy; yet the enhancement would be small. The possible reason could be the ratio of inliers and outliers, which is one of the major constraint for these algorithms.

Related Publications

Journal Articles

2) R. Roopalakshmi and G. Ram Mohana Reddy, *A Framework for Estimating Geometric Distortions in Video Copies Based on Visual-Audio Fingerprints*, Published in **Springer Signal, Image and Video Processing (SIViP)** Journal, Vol.7, Issue 1, Jan'2013. ISSN: 1863-1703. Available: `http://link.springer.com/article/10.1007/s11760-013-0424-7`.

Chapter 6

Case Study: Pirate Position Estimation Framework

Followed by geometric distortions estimation, the resultant visual-acoustic fingerprints can be efficiently employed for locating the position of the pirate in a movie theater. From another perspective, combating camcorder piracy requires forensic tracking systems to track the movie pirate, who is responsible for the illegal capture. Due to these reasons, this thesis describes the scholarly contribution towards the pirate position estimation problem in this chapter. Precisely, the current chapter illustrates a forensic tracking framework employing visual-acoustic fingerprints, for investigating the position of the pirate in a movie theater. More precisely, the proposed framework first determines the camcorder optical axis to the screen perpendicular by redefining the theater projective geometry and consequently estimates the location of the pirate in a movie theater, which is detailed as follows.

6.1 Estimating the Position of the Pirate

Fighting camcorder piracy needs not only the identification of the theater and show time information, but also the estimation of camcorder location in a theater from which a illegal recording was made, in order to find out the pirate as well as limit the number of pirate suspects. On the other hand, *addressing camcorder piracy, through forensic frameworks for identifying the movie pirate, is certainly not the aim of this thesis*. Instead of that, this research study attempts to highlight the capability of video fingerprints towards the pirate position estimation problem. In other words, current research work tries to prove that, the illegal capture location in a theater could be approximated, by performing in-depth analysis of geometric distortions and

theater projective geometry. In order to validate this view-point, In-Theater experiments are conducted and evaluated in this thesis. Therefore, this thesis contributes a new forensic tracking framework exploiting visual-audio fingerprints, for estimating the location of the pirate in a theater, by keeping the assumptions described in Section 2.4.1. More precisely, the main contributions of the proposed forensic tracking framework are listed below:

- ⋆ The proposed framework demonstrates that the visual-audio fingerprints extracted from the master and the duplicate video sequences could be utilized for finding the position of the pirate in a movie theater. *This is a brand-new application of the video fingerprinting technique, which helps to find out, where the pirate was during illegal recording and consequently restricts camcorder piracy.*
- ⋆ Precisely, the proposed forensic tracking framework first introduces spatio-temporal registration and geometric distortions estimation of the pirate video with the master sequence by employing multimodal signatures irrespective of presence/absence of watermarks in the video sequences. More specifically, in the proposed framework, the spatio-temporal registration scheme introduced in (Roopalakshmi and Reddy 2013) is extended, in order to estimate the position of pirate in a movie theater. Further, a *stable key point pairs selection* algorithm is presented, which efficiently extracts the most similar key point pairs from the temporally aligned master and pirate video sequences.
- ⋆ To the best of our knowledge, this is the first attempt, which exploits both the visual and acoustic fingerprints for estimating the location of the pirate in a theater, when compared to the conventional watermarking based forensic tracking methods.

The proposed forensic tracking method including spatio-temporal frame alignments and geometric distortions estimation followed by the pirate position approximation is illustrated as follows.

6.1.1 Scenario for identifying a movie pirate

A scenario shown in Figure 6.1 is considered for the purpose of identifying the movie pirate, which is defined as follows: (1) The pirate illegally records a movie using a camcorder in a theater and uploads the illegal content on the Internet. (2) In any anti-piracy strategy, copy identification is the first step; hence, here Content-Based video Copy Detection (CBCD) technique is employed, which detects the best matching master video for the given pirate clip. (3) After this step, a conventional

watermarking system such as (Haitsma and Kalker 2001) is used to determine the theater and the show time at which the illegal camcorder captures are made. (4) Then, the proposed position estimation framework estimates the location of the pirate in the movie theater in terms of specific seat information.

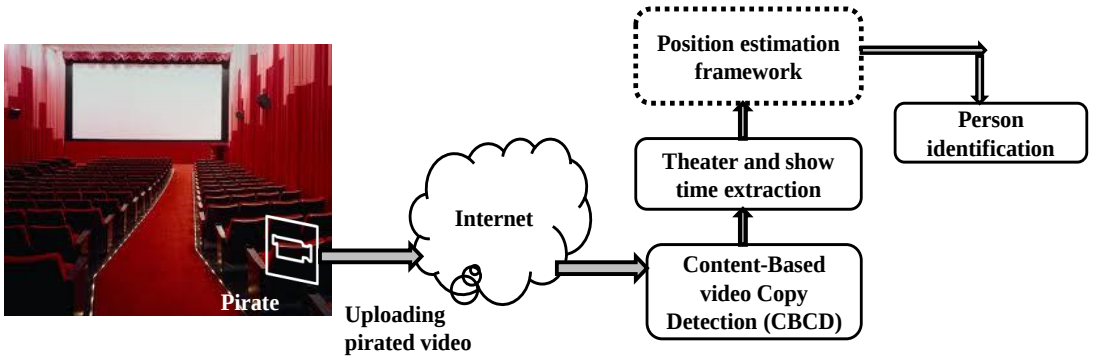


Figure 6.1: Scenario for identifying a movie pirate

More precisely, after the copy detection task, spatio-temporal alignments and estimation of distortion model between pirate and master video contents are prerequisites, in order to approximate the capture location in a theater. Therefore, the proposed position estimation framework computes spatio-temporal frame alignments of the source movie and pirate video contents by exploiting visual-audio fingerprints. Then, the proposed framework estimates the geometric distortions in the pirate video in terms of the projective matrix. Consequently, the camcorder optical axis to the screen perpendicular is determined by redefining the theater projective geometry and eventually the position of the pirate in the theater is estimated by the position estimation framework. (5) Finally, an electronic ticketing system may be used to identify the exact person who illegally captured the movie. In this way, the proposed position estimation framework restricts the number of piracy suspects and helps to identify the pirate. This chapter work focuses on the position estimation framework shown in Figure 6.1 with dotted lines, which is an essential component of this scenario.

6.1.2 Proposed pirate position estimation framework

The block diagram of the proposed position estimation framework is shown in Figure 6.2, which consists of three stages: In the first stage, the source movie and illegal video contents are registered in order to obtain accurate frame-to-frame mappings, such that the resultant frame pairs are both temporally and spatially aligned. Visual-audio fingerprints are employed to get spatio-temporal frame alignments of two video

contents using the registration scheme proposed in (Roopalakshmi and Reddy 2013).

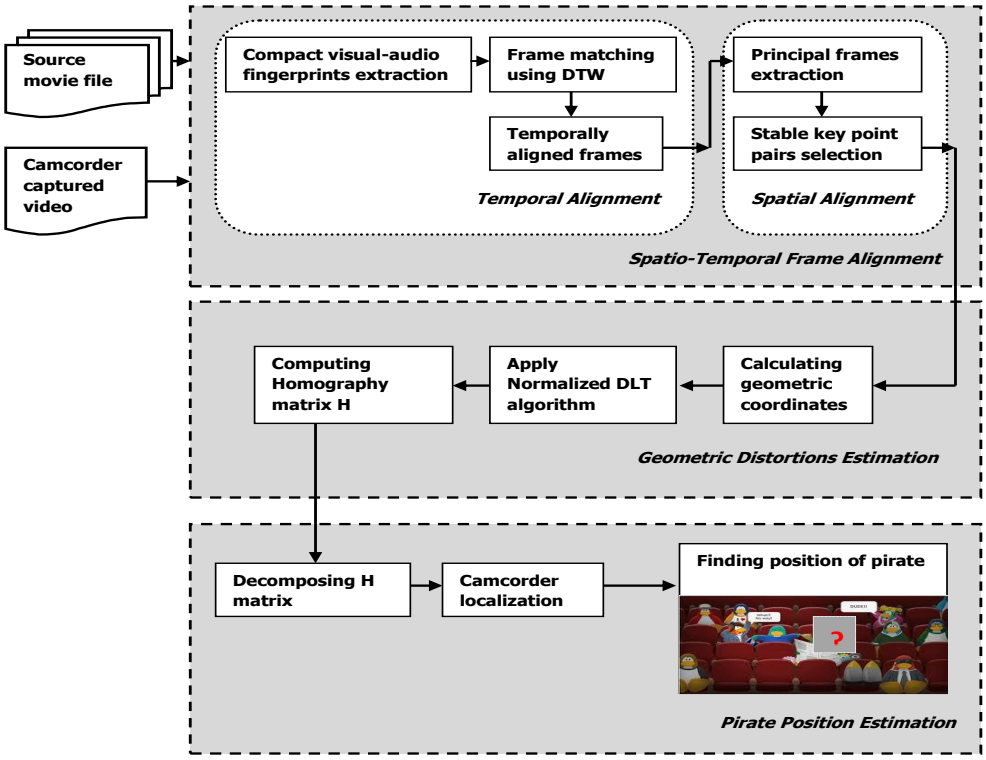


Figure 6.2: Block diagram of the proposed position estimation framework

More precisely, the two video sequences are compactly represented by exploiting visual signatures extracted from SURF interest points and acoustic fingerprints derived from spectral centroid features. Then a multimodal frame matching scheme based on Dynamic Time Warping (DTW) is utilized to temporally align the visual and acoustic feature sequences. From the temporally aligned video contents, most similar frames denoted as *principal frames* are selected, which are further analyzed in order to obtain *stable key point pairs* of two video sequences.

In the second stage, the geometric coordinates derived from key point pairs of two video contents are employed to estimate the geometric distortions. Specifically, the homographic matrix of projective geometry is determined by using the estimated point-to-point correspondences and the Normalized DLT algorithm (Hartley and Zisserman 2004). In the third stage, theater projective geometry is redefined and the coefficients of the homographic matrix are utilized to estimate the camcorder capture location. Precisely, camcorder optical axis is determined by using the translation and rotation parameters, obtained from the homographic matrix. After this step, the

camcorder capture location is estimated by computing intersection of the camcorder viewing axis and the theater seating plane.

6.1.3 Spatio-temporal frame alignments

At the beginning of this stage, first the source movie is scanned with a sliding window of length equal to the captured video clip. The similarity between the windowed source movie sequence and the captured video clip is measured based on their temporal signatures derived from SURF interest points and spectral centroid signatures. The windowed source sequence with minimum dissimilarity is selected and denoted as the *candidate segment* of source movie sequence. The details of the *candidate segment* selection algorithm is given in (Roopalakshmi and Reddy 2013).

After this point, visual-audio fingerprints of the *candidate segment* and the captured video are matched separately using DTW technique. The resultant matching results are fused in order to obtain temporal frame-to-frame alignments of two video contents. Then *principal frames* are selected from the temporally aligned candidate and pirate video segments, using the algorithm described in (Roopalakshmi and Reddy 2013).

As described in Section 4.3.5, the accuracy of the proposed spatio-temporal registration framework is assessed by performing evaluations on three different datasets, namely TRECVID sound & vision data, CC_WEB_VIDEO dataset and a set of real data consisting of camcorded copies of master video files. The proposed spatio-temporal registration framework achieves promising results in terms of better *MF* and *AD* rates compared to the reference methods.

In (Roopalakshmi and Reddy 2013), the *principal frames* of source movie and captured video clips are mapped using their SURF descriptors to obtain accurate spatial alignments. However, blind comparison of all SURF descriptors of two mapped frames is computationally demanding. Further, direct comparison of all descriptors of two frames may lead to lot of false matches. In order to handle these issues, the proposed scheme employs a nearest neighbor based mapping strategy to select *stable key point pairs* of temporally aligned frames, which is described below.

Stable Key Point Pairs Selection

The proposed framework attempts to filter out false key point matches by exploiting Euclidean distance measure and nearest neighbor mapping strategies. Precisely, the steps used to select the *stable key point pairs* of two mapped frames are detailed in Figure 6.3. Then the geometric coordinates of *stable key point pairs* of resultant

Algorithm 6.1: *Stable Frame Pairs* Selection

1: Let F_p and F_m be the temporally mapped frames of pirated video and master video sequences respectively.

2: Here, the pirated video frame $F_p \in \{it_p^1, it_p^2, ..., it_p^n\}$, where it_p^k indicates the k^{th} interest point and n is total key points of video frame F_p. Also, $F_m \in \{it_m^1, it_m^2, ..., it_m^k\}$, where It_m^j indicates the j^{th} interest point and k is total key points of master video frame F_m.

3: Compute the distance *dist* between each key point pair of two mapped frames using Squared Euclidean distance measure as given by,

$$dist = |((it_p^i) - (it_m^j))^2| \tag{6.1}$$

where i$\in$ [1 : n] and j$\in$ [1 : k].

4: List list the first and second nearest neighbors with minimum *dist* values for each interest point.

5: Invert the role of frame pairs F_p and F_m.

6: Repeat the steps 3-5 and compute the nearest neighbors with minimum *dist* values.

7: Select the key point pair, for which same correspondence is obtained in both the lists.

8: The resultant key point pairs mapped in F_p and F_m frames, are denoted as *stable key point pairs*.

Figure 6.3: Stable Key Point Pairs Selection Algorithm

frames are utilized to estimate the distortions model between the two video sequences.

6.1.4 Geometric distortion estimation

As described in Section 5.2.3, in case of camcorder capture in a movie theater, the camcorder viewing axis is not perpendicular to the screen and consequently the captured images undergo severe geometric distortions. The resultant geometric distortions can be well described by perspective projection, which replicates the imaging process of a pinhole camera (Hartley and Ziserman 2004). In the proposed framework, the geometric distortions in the video copy is estimated in terms of homographic matrix **H**, as illustrated in Section 5.2.3.

6.1.5 Pirate position estimation

In the previous subsection, geometric distortions in the captured movie clips are estimated by means of the homographic matrix **H**. The coefficients of **H** matrix represent translation and rotation parameters of the camcorder and camera calibration. Due to the zooming operation of camcorder, the resultant parameters of **H** matrix represent only the camcorder optical axis, but not the exact position of the camcorder.

In order to handle this issue, the proposed framework first performs camcorder localization, by determining its optical axis which is specified by translation and rotation parameters. Then in-depth analysis of theater geometry including its seating plane is carried out, in order to estimate the actual capture location. Precisely, the capture location is estimated by computing the intersection of the camcorder optical axis and seating plane of the test environment. Therefore, the projective geometry is redefined and the homographic matrix **H** is decomposed in order to estimate the position of the camcorder, which is illustrated as follows.

Camera Projective Geometry

The proposed framework introduces the notation X for the world coordinate of the screen represented by the homogeneous 4-vector $(x, y, z, 1)^T$ and X$'$ for the image coordinate of the captured video represented by a homogeneous 3-vector $(x', y', 1)^T$ as defined in Equation (5.14). Then the camera projection, which maps the world and image coordinates is expressed in terms of matrix multiplication with P for the 3×4 homogeneous camera projection matrix as follows (Hartley and Ziserman 2004),

$$X' = \mathbf{P}X \tag{6.2}$$

in which **P** indicates the camera projection matrix for the pinhole model (Camcorders) of central projection as given by,

$$\mathbf{P} = \mathbf{K}\ [\mathbf{R} \mid \mathbf{t}] \tag{6.3}$$

where $\mathbf{t} = -\mathbf{RC}$ and **K** is the camera calibration matrix. **R** is the 3×3 rotation matrix indicating the orientation of camera coordinate frame, whereas **C** denotes the coordinates of the camera origin in the world coordinate frame. The Equation (6.3) can be expressed as,

$$\mathbf{P} = \mathbf{K}[\mathbf{R}| - \mathbf{RC}] \tag{6.4}$$

The parameters contained in **K** are called internal camera parameters, while the

parameters of **R** and **C** are called the external parameters of the camera. In case of CCD cameras, the image coordinates and the principal point are measured in terms of pixel dimensions; hence, the general form of the calibration matrix **K** of a CCD camera is given by (Hartley and Ziserman 2004),

$$\mathbf{K} = \begin{bmatrix} \delta_x f & s & \\ & \delta_y f & \\ & & 1 \end{bmatrix} \tag{6.5}$$

where f is the focal length of the camcorder. δ_x and δ_y represent the pixel dimensions in the x- and y- directions respectively. s is referred as the *skew* parameter, which is zero for most of the CCD cameras, since the x- and y- axis are perpendicular to each other. The rotation matrix **R** represents rotation in each of the three x-, y- and z-dimensions. If the rotations are performed in clockwise direction from the origin, then the matrix **R** can be written as,

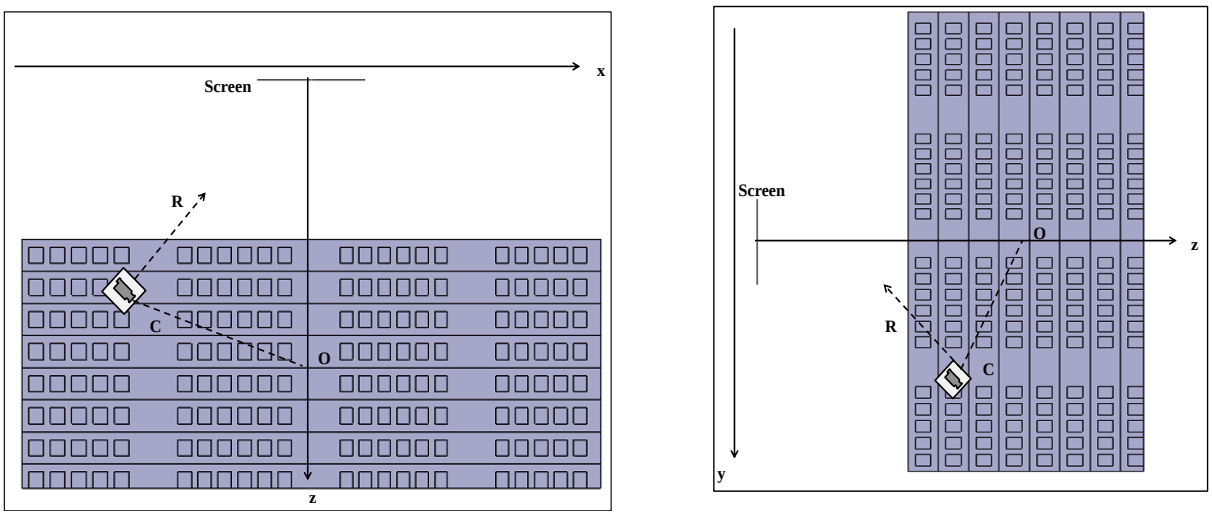
$$\mathbf{R} = \mathbf{R_x R_y R_z} \tag{6.6}$$

Thus the rotation matrix **R** is given by,

$$\mathbf{R} = \begin{bmatrix} cos\alpha_y cos\alpha_z & cos\alpha_x sin\alpha_z - cos\alpha_z sin\alpha_x sin\alpha_y & sin\alpha_x sin\alpha_z + cos\alpha_x sin\alpha_y cos\alpha_z \\ -cos\alpha_y sin\alpha_z & cos\alpha_x cos\alpha_z - sin\alpha_x sin\alpha_y sin\alpha_z & -sin\alpha_x cos\alpha_z + cos\alpha_x sin\alpha_s in\alpha_z \\ -sin\alpha_y & -cos\alpha_y sin\alpha_x & -sin\alpha_y + cos\alpha_y sin\alpha_x \end{bmatrix} \tag{6.7}$$

Projective Geometry

The 2-D view of the projective geometry consisting of a theater screen and a camcorder is indicated in Figure 6.4(a) and (b). The origin of the theater is shown as **O**, which is Z_s distance far from the screen. In Figure 6.4, the location of camcorder is indicated by means of a translation as **C** and rotation as **R**, which is specified in Equation(6.4). Computing the camera projection matrix **P** and decomposing it into $\mathbf{K}, \mathbf{R}$ and **C** leads to a trivial solution. This is because, the camera calibration matrix **K** of the camcorder, which is used for illegal capture is not known. In addition, the zooming operation of the camcorder makes the estimated distance from the origin as unreliable one. In order to tackle this problem, the proposed framework redefines the theater projective geometry to estimate the position of camcorder, which is illustrated as follows.



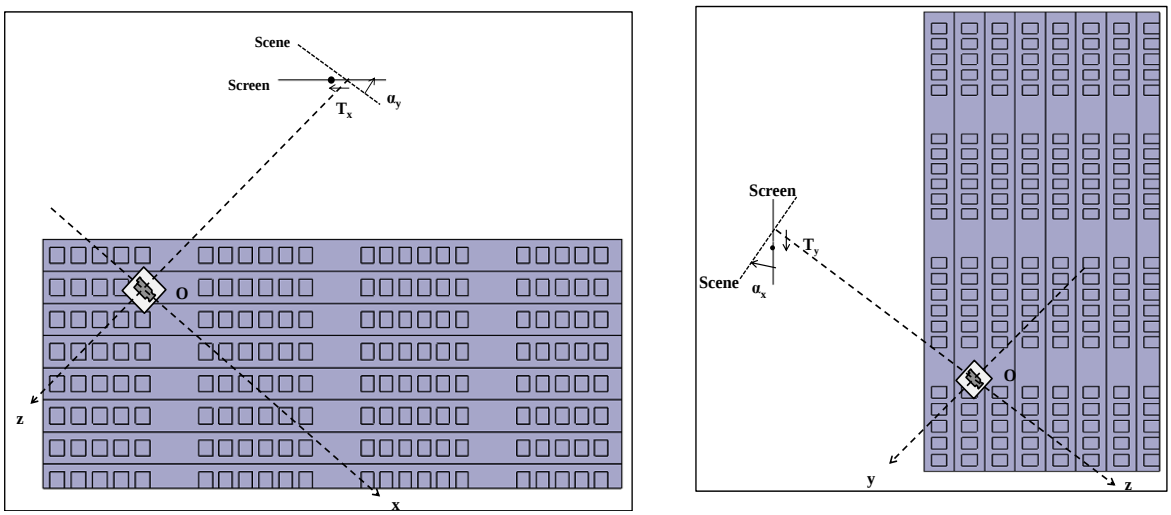
(a) (b)

Figure 6.4: 2-D view of projective geometry. (a) Top view. (b) Side view. The origin of the theater **O** is Z_s distance far from the screen. **R** is the rotation denoting the orientation of the camera coordinate frame and **C** indicates the coordinates of the camera origin in the world coordinate frame.

Redefined Projective Geometry

The projective geometry shown in Figure 6.4 is redefined, so that the camcorder is located at the origin. More precisely, displacement of the theater screen is considered instead of the camcorder displacement. The proposed scheme assumes the displacement of the theater screen from the its origin with a translation $T = (T_x, T_y, T_z)$ followed by a rotation $R = (\alpha_x, \alpha_y, \alpha_z)$. Here, T_z is equal to zero, because origin of the screen is moved along the screen plane. Also, the focal length of the camcorder f used for piracy is not known; hence the proposed framework sets the focal length f equal to the distance Z_s between the camcorder and the theater screen. In the redefined projective geometry the coordinates and the parameters are measured in pixel dimensions.

Figure 6.5(a) and (b) illustrate the redefined projective geometry in terms of top and side views. As per the redefined projective geometry, Equation (6.4) representing the camera projection matrix **P** need to be modified. Specifically, **C** in (6.4) is replaced to $(T_x, T_y, 0)$ and the camera projection matrix according to the redefined



(a) (b)

Figure 6.5: 2-D view of redefined projective geometry. (a) Top view. (b) Side view. Redefined projective geometry assumes the displacement of the theater screen instead of camcorder displacement, so that the camcorder is located at the origin.

projective geometry is given by,

$$\mathbf{P} = \mathbf{K} \begin{bmatrix} \mathbf{R}| & -\mathbf{R} \begin{pmatrix} T_x \\ T_y \\ 0 \end{pmatrix} \end{bmatrix} \tag{6.8}$$

The optical axis of the camcorder and the capture location are estimated by using Equation(6.8) as detailed below.

Camcorder Localization and Position Estimation

The camera projection matrix $\mathbf{P}$ in (6.8) is a 3×4 matrix and has 11 degrees of freedom; hence at least six point-to-point correspondences are required to compute $\mathbf{P}$ matrix. However, homographic matrix $\mathbf{H}$ with nine coefficients is already computed in the previous subsection. In addition, the theater screen is assumed as planar and hence the z-coordinates of the screen are omitted. Due to these reasons, the proposed framework utilizes H matrix to compute the position of the camcorder as follows. Equations (6.2) and (6.8) are concisely formulated as,

$$X' = \mathbf{K} \begin{bmatrix} \mathbf{R}| & -\mathbf{R} \begin{pmatrix} T_x \\ T_y \\ 0 \end{pmatrix} \end{bmatrix} X \tag{6.9}$$

Let $\mathbf{R_1}, \mathbf{R_2}$ and $\mathbf{R_3}$ are the first, second and third columns of $\mathbf{R}$ respectively, then Equation(6.9) is expressed as,

$$X' = \mathbf{K} \begin{bmatrix} \mathbf{R_1} & \mathbf{R_2} & -\mathbf{R_3} \begin{pmatrix} T_x \\ T_y \\ 0 \end{pmatrix} \end{bmatrix} X \tag{6.10}$$

Equation(6.10) has the same form as Equation(6.2), hence the matrix in Equation(6.2) is decomposed as follows,

$$\mathbf{H} = \mathbf{K} \begin{bmatrix} \mathbf{R_1} & \mathbf{R_2} & -\mathbf{R_3} \begin{pmatrix} T_x \\ T_y \\ 0 \end{pmatrix} \end{bmatrix} X \tag{6.11}$$

Nine equations shown in Equation(6.12) are obtained from (6.11), by decomposing $\mathbf{H}$.

$$\begin{aligned}
h_{11} &= \frac{\delta f cos\alpha_y cos\alpha_z}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{12} &= \frac{\delta f(cos\alpha_x sin\alpha_z - cos\alpha_z sin\alpha_x sin\alpha_y)}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{13} &= \frac{\delta f\{T_x(cos\alpha_y cos\alpha_z) + T_y(sin\alpha_z cos\alpha_x - cos\alpha_z sin\alpha_y sin\alpha_x)\}}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{21} &= \frac{f cos\alpha_y sin\alpha_z}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{22} &= \frac{f(cos\alpha_x cos\alpha_z + sin\alpha_x sin\alpha_y sin\alpha_z)}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{23} &= \frac{-f\{T_x(sin\alpha_z cos\alpha_y) + T_y(cos\alpha_x cos\alpha_z + sin\alpha_x sin\alpha_y sin\alpha_z)\}}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{31} &= \frac{-sin\alpha_y}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{32} &= \frac{-cos\alpha_y sin\alpha_x}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x} \\
h_{33} &= \frac{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x}{Z_s - T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x}
\end{aligned} \tag{6.12}$$

The term h_{33} is a scale factor; hence the normalized representation of matrix $\mathbf{H}$ is obtained by dividing all matrix entries with h_{33}. For simplification purpose, the proposed framework considers δ as pixel aspect ratio of the camcorder. In addition, the term $(-T_x sin\alpha_y + T_y cos\alpha_y sin\alpha_x)$ in h_{33} is supposed to be small compared with Z_s and it is approximated with Z_s. By eliminating the numerator f and the denominator

Z_s, the four equations in (6.12) are rewritten to obtain the four unknowns $\delta, \alpha_x, \alpha_y$ and α_z as follows,

$$\begin{aligned} h_{11} &= \delta cos\alpha_y cos\alpha_z \\ h_{12} &= \delta(cos\alpha_x sin\alpha_z - cos\alpha_z sin\alpha_x sin\alpha_y) \\ h_{21} &= cos\alpha_y sin\alpha_z \\ h_{22} &= cos\alpha_x cos\alpha_z + sin\alpha_x sin\alpha_y sin\alpha_z \end{aligned} \tag{6.13}$$

In the next step, using $\delta, \alpha_x, \alpha_y$ and α_z, T_x and T_y are computed by solving the two equations given by,

$$\begin{aligned} h_{13} &= \delta\{T_x(cos\alpha_y cos\alpha_z) + T_y(sin\alpha_z cos\alpha_x - cos\alpha_z sin\alpha_y sin\alpha_x)\} \\ h_{23} &= -\{T_x(sin\alpha_z cos\alpha_y) + T_y(cos\alpha_x cos\alpha_z + sin\alpha_x sin\alpha_y sin\alpha_z)\} \end{aligned} \tag{6.14}$$

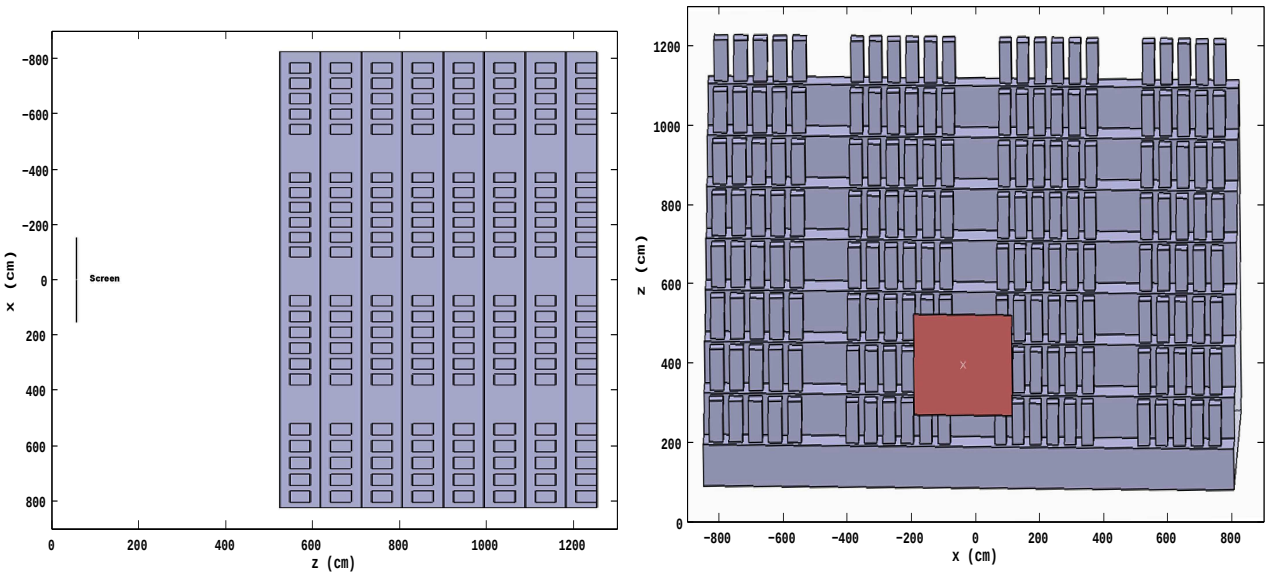
By solving the equations in (6.14), T_x and T_y values are obtained in pixel dimensions. The conversion of T_x and T_y from pixel dimensions to centimeters is formulated as,

$$T_x{}' = \frac{h_{ts}}{h_{cv}} T_x \quad ; \quad T_y{}' = \frac{h_{ts}}{h_{cv}} T_y \tag{6.15}$$

where h_{ts} represents the height of the theater screen in centimeters and h_{cv} stands for height of the captured video in pixels. After this point, the camcorder optical axis is fully specified by the translation $T(T_x{}', T_y{}')$ and rotation $R(\alpha_x, \alpha_y, \alpha_z)$ parameters. Finally, the intersection of the camcorder optical axis with the theater seating plane gives the estimated position of the camcorder in the theater.

6.1.6 In-theater experiments

To evaluate the estimation accuracy of the proposed framework, experiments are conducted in a large-scale test environment, that is an auditorium with 176 seats. The auditorium is about 17.83m wide and 12.69m long with 8 seating rows divided into 4 sections. The screen in the auditorium is about 3.26m wide and 2.44m long. So, the movie clips projected on the screen are displayed as 3.26m and 2.44m in the vertical and horizontal directions respectively. The auditorium is having its own slope between the seating rows; hence, the construction of the auditorium is measured in order to determine the seating plane of our test environment. Figure 6.6(a) and (b) indicate the top view and front view of the test environment. Ten different seats denoted as *a-j* are arbitrarily selected for camcorder capture, which spread over the entire seating plane of the test environment.



(a) (b)

Figure 6.6: (a) Top view of the test environment. The auditorium is 17.83m wide and 12.69m long with 176 seats and a screen size of 3.26m × 2.44m. (b) Front view of the test environment. The test environment has 8 seating rows divided into 4 sections.

The HD-resolution clips of two popular movies namely, '*Journey2 The Mysterious Island*' and '*Alvin And The Chipmunks-Chipwrecked*', were projected on to the screen. Figure 6.7 shows the snapshot examples of camcorder captured video clips.

From each location, different video clips ranging between 18-67 seconds were captured using SONY DCR-SR20 camcorder and stored in 640×480 format. Then source and captured video clips are spatio-temporally aligned as described in Section 6.1.3, in order to estimate the geometric distortions and camcorder capture locations.

6.1.7 Estimation accuracy evaluation and discussion

Figures 6.8 to 6.16 indicate the experimental results of approximating the suspicious seats where the camcorder capture is done. Figure 6.8(a)-(d) show the top view of first four actual seats 'a-d' and the estimated camcorder capture locations. More precisely, in Figure 6.8(a) the actual camcorder location (here seat 'a') is indicated as a colored dark line (*purple color*), while colored dotted lines (*pink color*) represent the estimated capture locations.

The top view of actual and estimated positions of the next four capture locations are shown in Figure 6.9(a)-(d). Specifically, the actual seats 'e-h' are indicated as colored dark lines, while the respective estimated positions are indicated as colored dotted lines in Figure 6.9(a)-(d). Figure 6.10(a)-(b) indicate the top view of two seats

Figure 6.7: Snapshot examples of camcorder captured video clips

'i-j' and their estimated camcorder capture locations. Precisely, the seats 'i-j' are represented as colored dark lines and the corresponding estimated capture locations are indicated as colored dotted lines in Figure 6.10(a)-(b).

Figure 6.10(c)-(d) show the isometric view of first two actual camcorder locations ('a-b') and the respective estimated camcorder positions. Precisely, Figure 6.10(c) depicts the seat 'a' as a colored dark line, while the estimated capture locations are shown as colored dotted lines. The actual seats c-f and the corresponding estimated positions are indicated in Figure 6.11(a)-(d) as snapshots of isometric view of the test environment. Figure 6.12(a)-(d) show the isometric view of the test environment, which represents the actual seats 'g-j' and the respective estimated camcorder recording locations. Specifically, The actual capture location is indicated as a colored dark line, whereas colored dotted lines represent the estimated camcorder positions as shown in Figure 6.12(a)-(d).

Figure 6.13 combines five seats *a-e* and the respective estimated camcorder locations shown in Figure 6.8(a)-(d) and Figure 6.9(a) into one graph. More precisely, actual seats are indicated as *plus* symbols, while estimated camcorder positions are shown as *dashed circles*. In Figure 6.13, the dotted circle boundaries show the largest error from the actual ones.

Five actual seats *f-j* and the corresponding estimated positions of the camcorders given in Figure 6.9(b)-(d) and Figure 6.10(a)-(b) are collectively represented in Figure 6.14. Specifically, in Figure 6.14, the actual seats are shown as *plus* symbols, while the respective estimated positions are denoted as *dashed circles*. In Figure 6.14, the dotted circle boundaries are set to indicate the largest error from the actual positions.

Figure 6.15 indicates the five actual seats *a-e* and their estimated results in terms

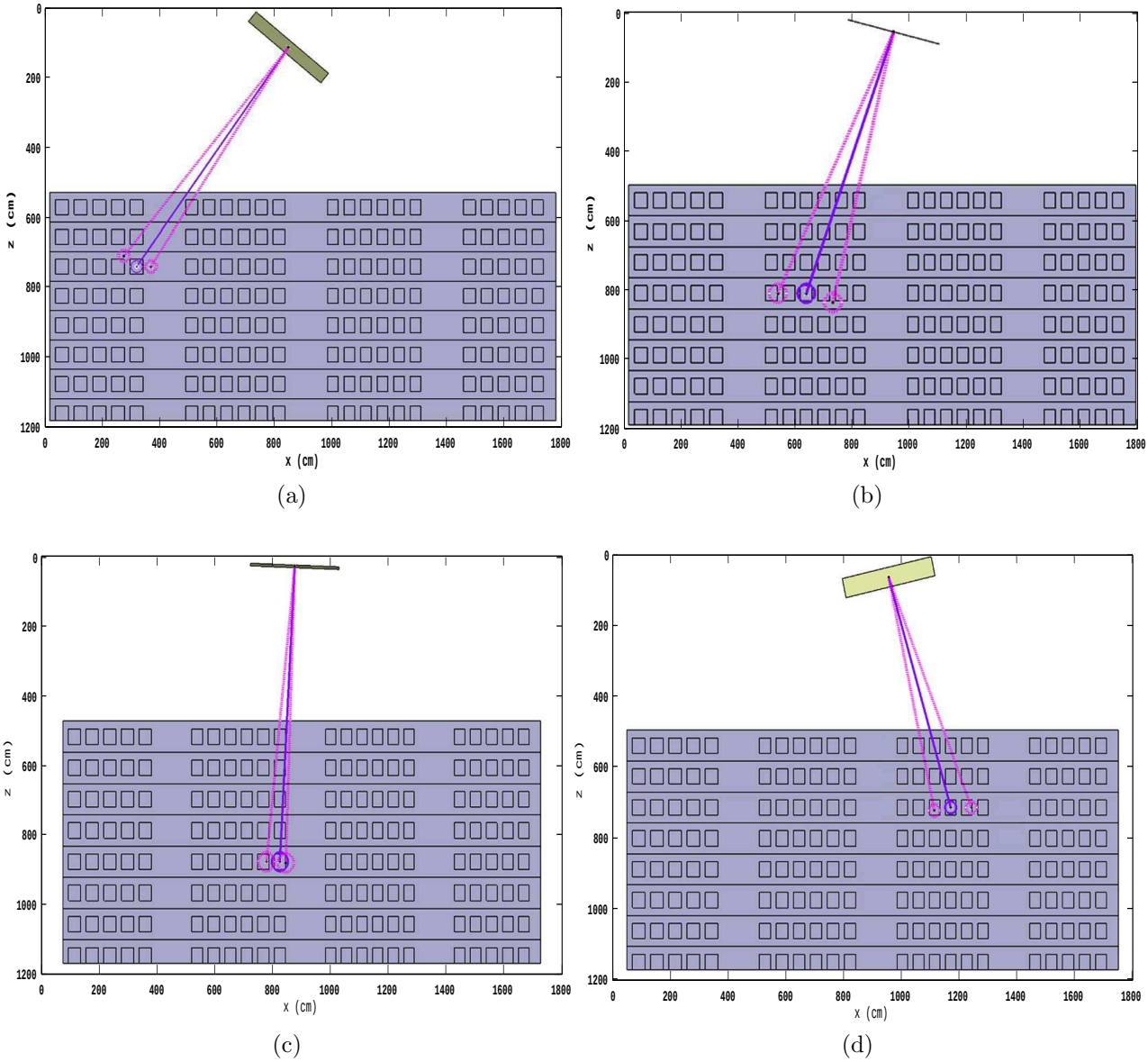
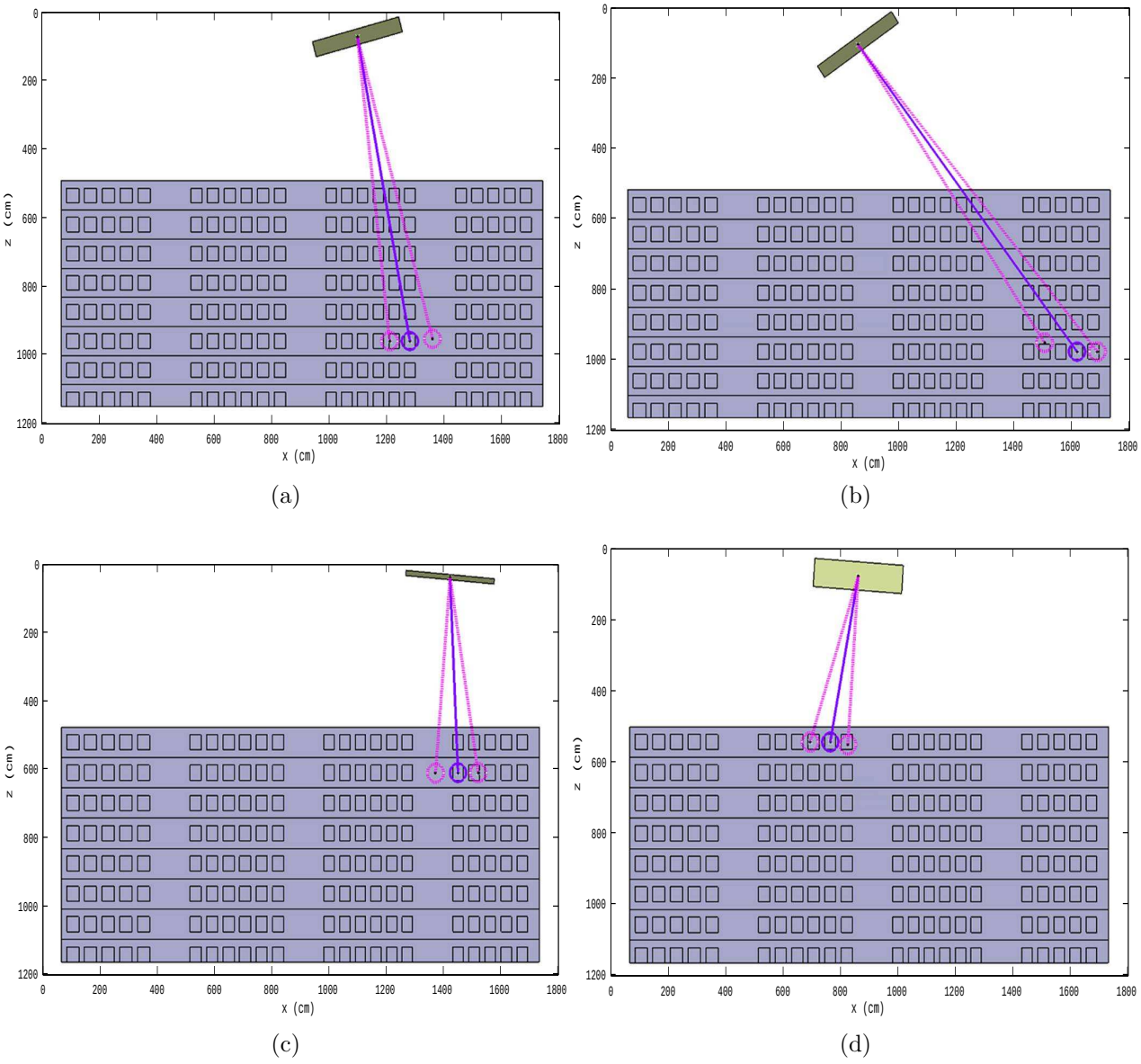


Figure 6.8: (a)-(d): Top view of actual seats *a-d* and the respective estimated positions of the camcorder in the *x-z* plane of the test environment. The intersected positions on the seating plane are determined according to the interior construction of the auditorium. The actual capture location is indicated as a colored dark line, while the colored dotted lines represent the estimated camcorder positions.

of all *x*-, *y*- and *z*- coordinate values in 3-D plots. Precisely, in Figure 6.15, the actual capture locations *a-e* are indicated as *star* symbols, while estimated camcorder positions are represented as *circle* symbols. Figure 6.16 indicates all *x*-, *y*- and *z*-coordinate values of actual camcorder locations *f-j* and their estimated results in 3-D plots. Specifically, in Figure 6.16, the *star* symbols are used to represent original seats whereas *circles* are used to indicate estimated positions.



(a) (b)

(c) (d)

Figure 6.9: (a)-(d): Top view of actual seats *e*-*h* and the corresponding estimated locations of the camcorder in the *x*-*z* plane of the test environment. The actual capture location is indicated as a colored dark line and colored dotted lines represent the estimated camcorder positions.

Table 6.1 illustrates the numerical analysis of estimation results in terms of statistical measures expressed in centimeters, which include actual position, mean estimates, mean absolute error and standard deviation along the principal axis. The mean absolute error of the estimated errors for ten positions is (38.25, 22.45, 11.11) cm and the standard deviation of the estimation errors for all ten capture locations is (22.26, 12.97, 7.29) cm respectively.

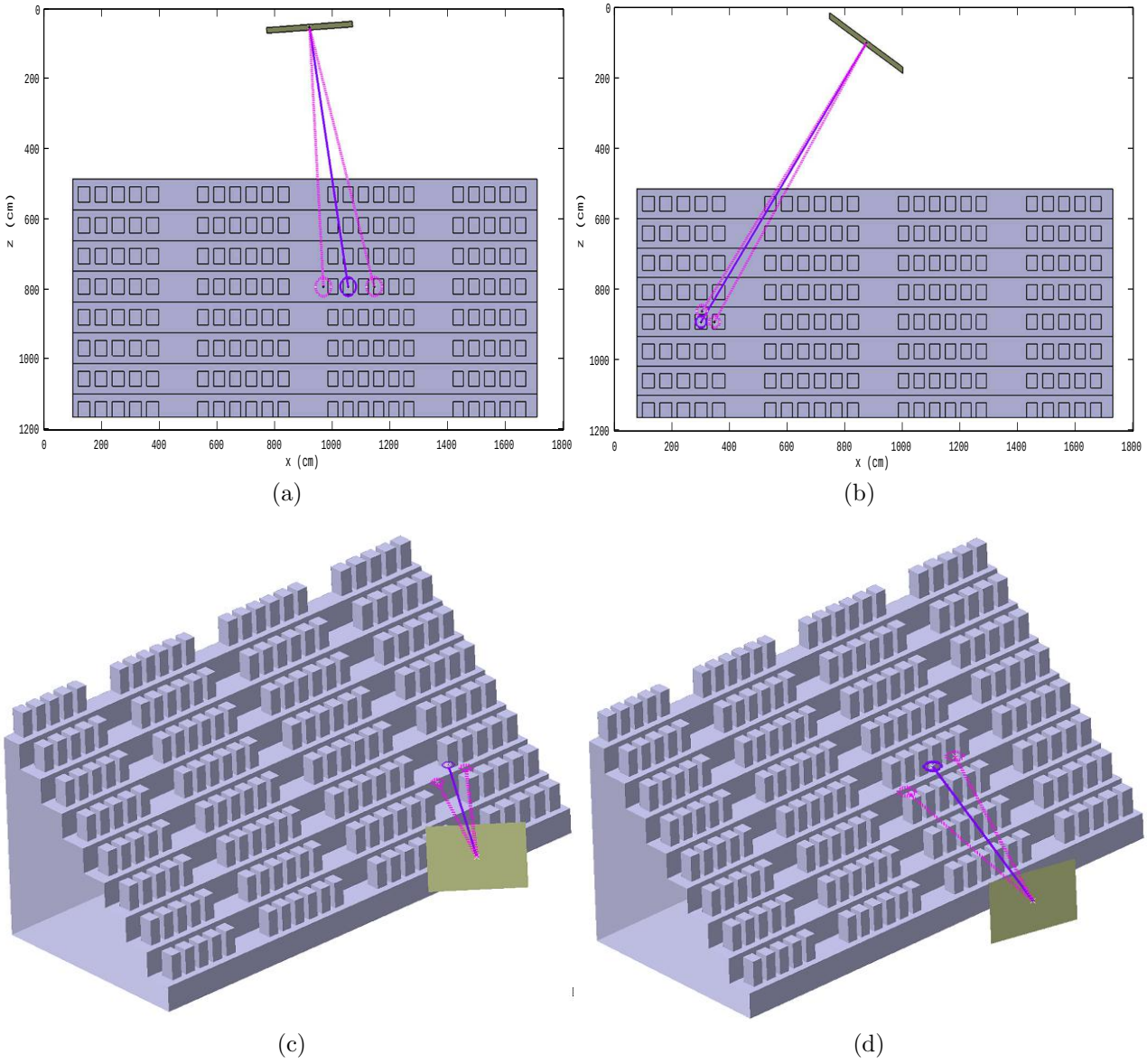


Figure 6.10: (a)-(b): Top view of actual seats *i*-*j* and the corresponding estimated locations of the camcorder in the *x*-*z* plane of the test environment. (c)-(d): Isometric view of actual seats *a*-*b* and respective estimated positions of the camcorder in the *x*-*z* plane of the test environment. The actual capture location is indicated as a colored dark line and colored dotted lines represent the estimated camcorder positions.

From Table 6.1 results, it is observed that, the mean width error ranges from 2.5 cm to 63.5 cm, while mean depth error ranges from 0.6 cm to 30.1 cm. This estimation accuracy is quite satisfactory, as the distance between two seats in a row is about 35 cm and the distance between two rows is about 100 cm. The proposed algorithm utilizes visual as well as acoustic fingerprints of the source movie sequence and the camcorder captured video clip for estimating the camcorder positions. More precisely, from

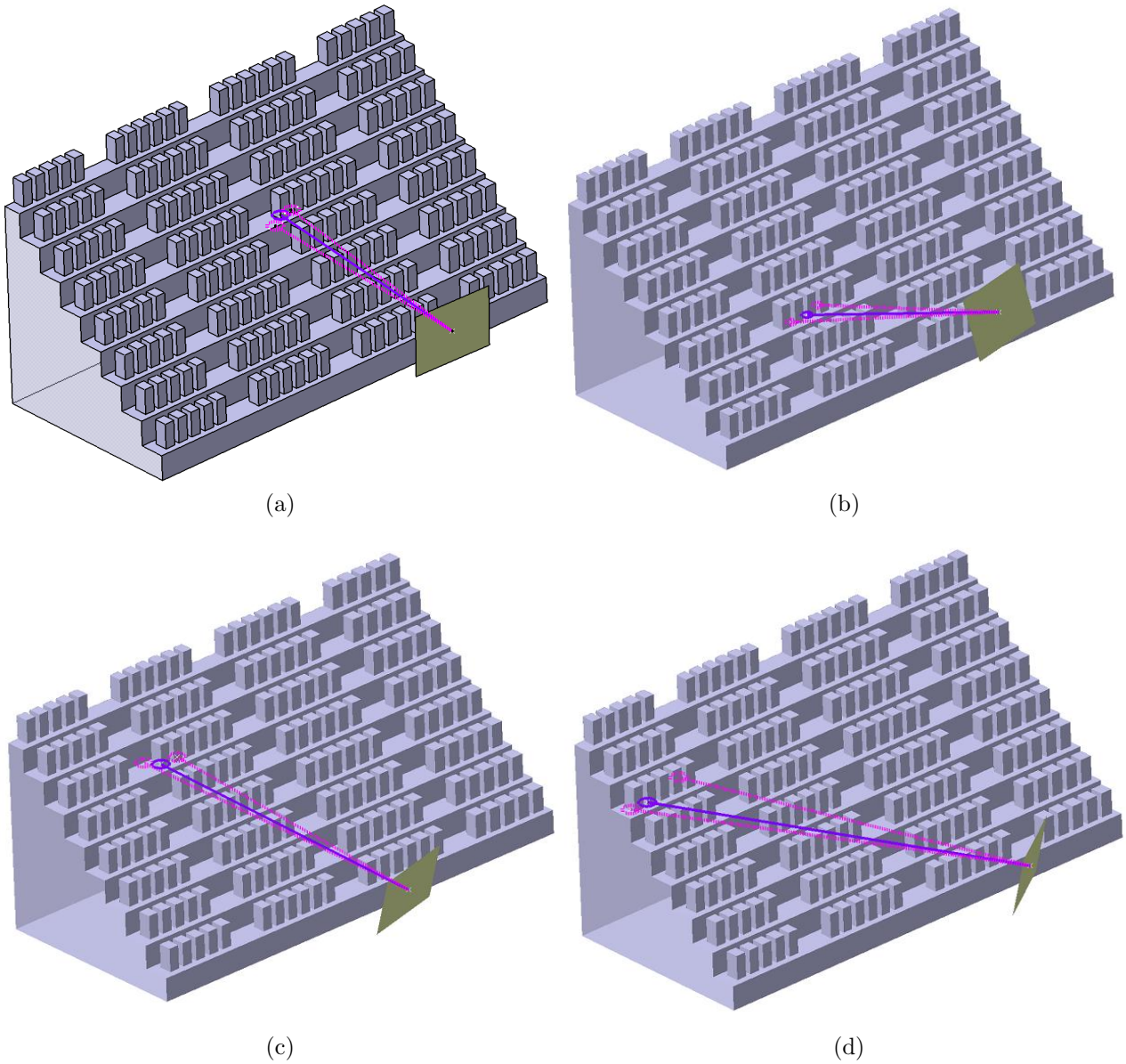
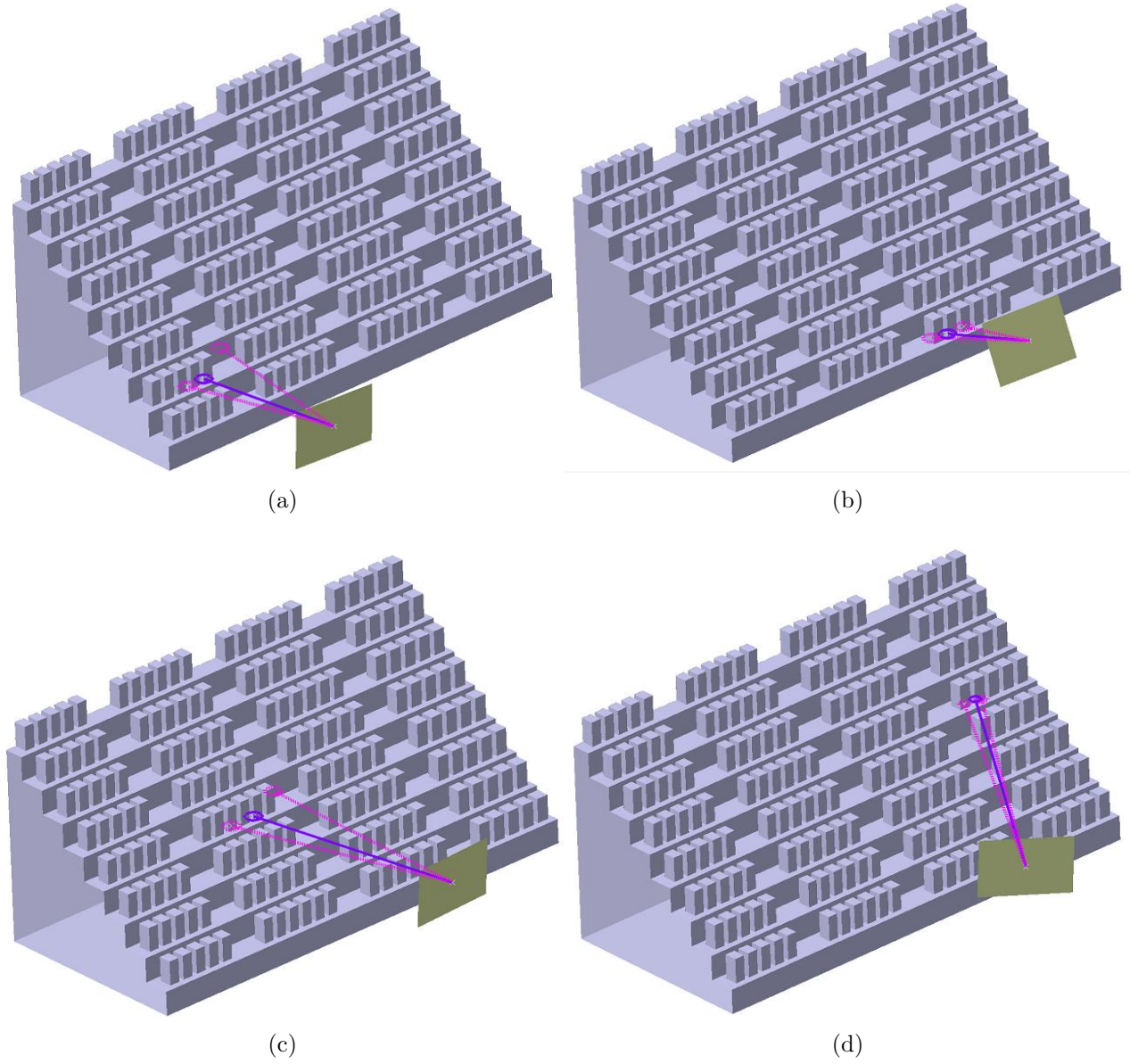


Figure 6.11: (a)-(d): Isometric view of actual seats *c*-*f* and respective estimated positions of the camcorder in the *x*-*z* plane of the test environment. The intersected positions on the seating plane are determined according to the interior construction of the auditorium. The actual capture location is indicated as a colored dark line and colored dotted lines represent the estimated camcorder positions.

the spatio-temporally aligned master and pirate video contents, only the geometric coordinates of SURF descriptors are employed to estimate the distortion model and camcorder locations. As per the definition of 2-D projective geometry, the resultant geometric distortions are described by a non-singular 3×3 homographic matrix.

Specifically, the proposed estimation framework exploits the homographic matrix **H** for localizing the camcorder. As a result of this camcorder localization, the pro-

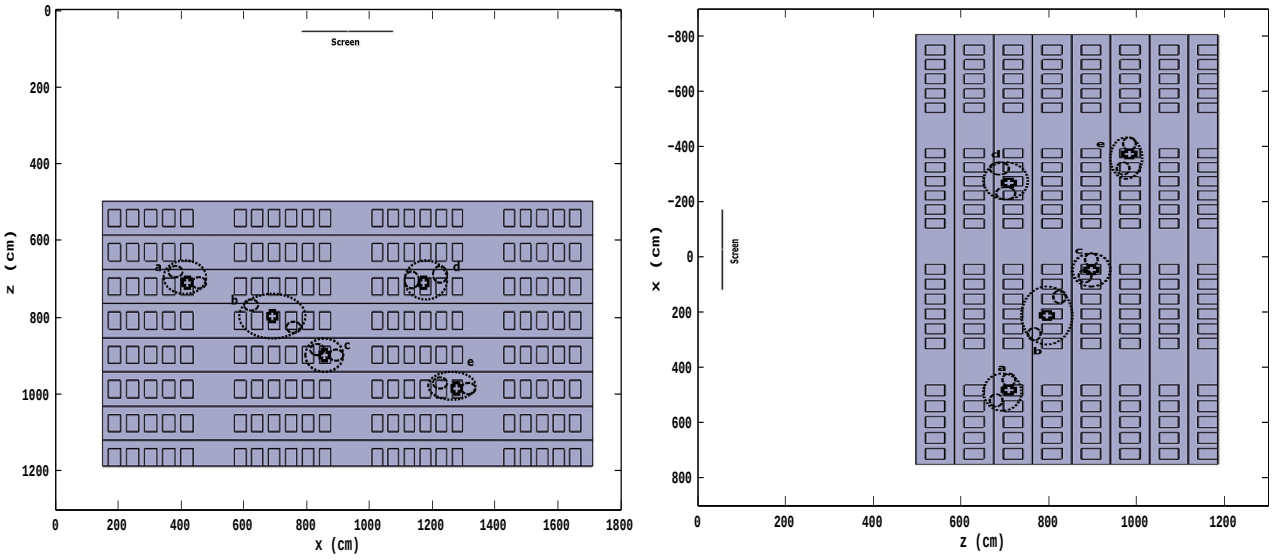


(a) (b)

(c) (d)

Figure 6.12: (a)-(d): Isometric view of actual seats *g-j* and corresponding estimated capture locations in the *x-z* plane of the test environment. The actual capture location is indicated as a colored dark line and colored dotted lines represent the estimated camcorder positions.

posed algorithm provides camcorder optical axis which has only the origin and the direction, but not the magnitude values. In the proposed framework, the magnitude values are obtained by computing the intersection between the seating plane and the camcorder optical axis. More precisely, the proposed scheme estimates the origin and the direction of the camcorder optical axis, but not the exact position of the camcorder. Therefore, if the estimated origin/direction have an error, then the er-



(a) (b)

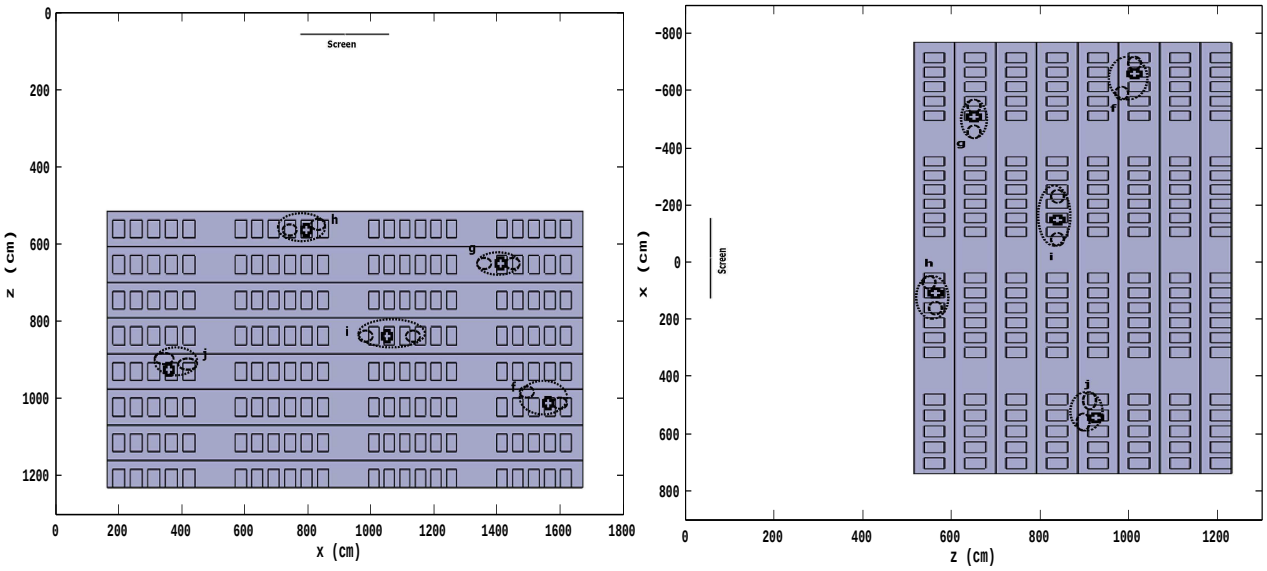
Figure 6.13: **+: Actual position**, **○: estimated positions**. (a)-(b): Top views of five actual seats *a*-*e* and the respective estimated camcorder positions in the *x*-*z* plane of the test environment. Dashed circles show the estimated camcorder locations, while the dotted circles show the largest error from the actual positions.

Table 6.1: Statistical analysis of camcorder position estimates (in cm)

Seats	**Actual position**			**Mean estimates**			**Mean abs. error**			**Std. deviation**		
	x	**y**	**z**	**x**	**y**	**z**	**x**	**y**	**z**	**x**	**y**	**z**
a	513.5	144.5	745	485.5	194.9	743	28	50.4	2.0	16.1	29	1.2
b	293.2	74.8	874.4	356.7	75.1	845	63.5	0.3	29.4	33.5	0.2	16.9
c	55.3	67.7	945	69.2	49.6	946.9	13.9	18.1	1.9	18.2	10.4	2
d	199.9	207.1	754.7	262.5	220	745	62.6	12.9	9.7	31.4	7.5	5.6
e	207.8	240	1036.8	156.8	197.4	1045	51	42.6	8.2	26.4	24.6	4.7
f	789.5	240.7	1014.9	847.5	173	1045	58	67.7	30.1	29.1	39	17.4
g	54.3	45.3	646.1	67.5	50	645	13.2	4.7	1.1	19.1	2.7	0.6
h	78.7	227.7	553.8	141.7	233	544	63	5.3	9.8	31.7	3.2	5
i	148.2	34.4	848	150.7	49	847.4	2.5	14.6	0.6	2.5	8.5	1.2
j	618.6	63.7	945	591.8	71.6	926.7	26.8	7.9	18.3	14.6	4.6	18.3

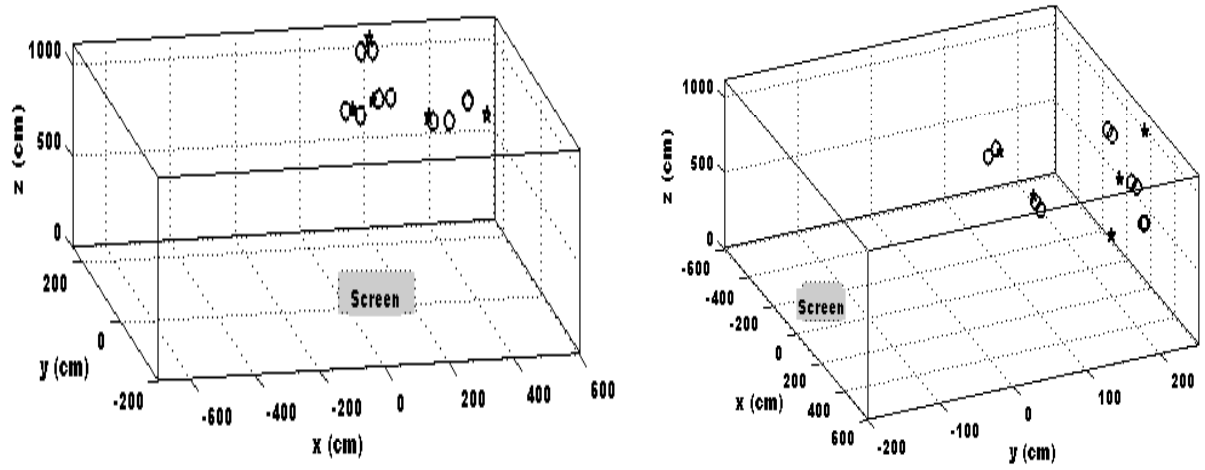
ror is proportional to the magnitude also, which may vary the accuracy of position estimation results.

The estimation results are very promising, because the proposed position estimation framework employs only content-based visual and audio fingerprints for estimating camcorder positions without embedding any watermarks. Therefore, the proposed framework can achieve satisfactory performances in digital cinema applica-



(a) (b)

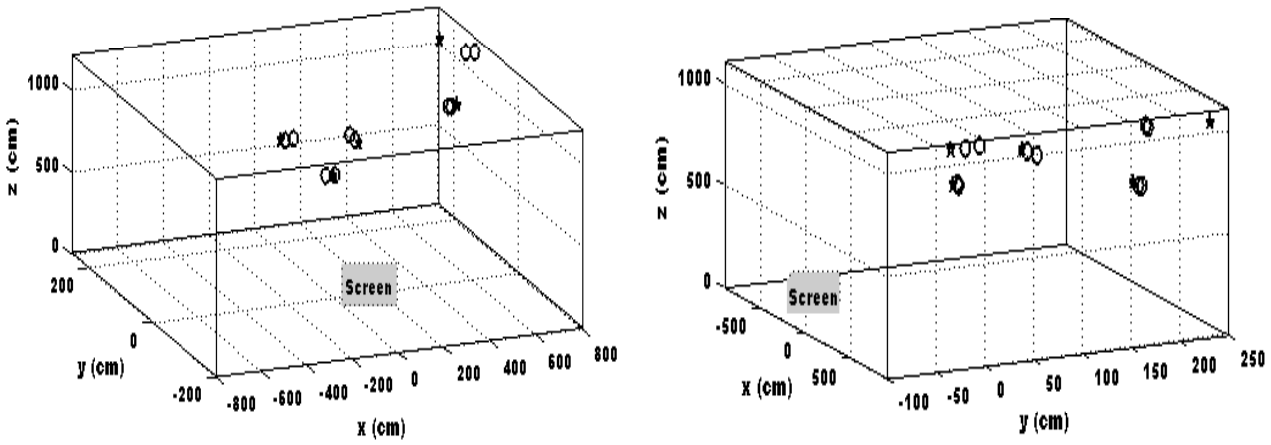
Figure 6.14: **+: Actual position**, **○: estimated positions**. (a)-(b): Top views of five actual seats *f-j* and the corresponding estimated locations in the *x-z* plane of the test environment. Dashed circles indicate the estimated camcorder positions, while the dotted circles represent the largest error from the actual ones.



(a) (b)

Figure 6.15: **★: Actual positions**, **○: estimated positions**. Five actual seats *a-e* and the respective estimated camcorder positions of the test environment in 3-D plots. (a) 3-D view 1 (b) 3-D view 2

tions. However, as per the digital cinema standards, the aspect ratio of the theater screen varies between 1.85:1 to 2.35:1. But the screen aspect ratio of the test environment is about 1.33:1. Therefore, if the experiments are conducted in a real theater,



(a) (b)

Figure 6.16: ★**: Actual positions**, ○**: estimated positions**. Five actual positions *f-j* and the corresponding estimated camcorder positions of the test environment in 3-D plots. (a) 3-D view 1 (b) 3-D view 2

then the position estimation accuracy can be substantially improved.

6.2 Summary

This chapter illustrates the scholarly contribution towards the pirate position estimation problem, by introducing a forensic tracking framework for investigating the illegal capture location in a movie theater. Precisely, the proposed framework tracks the position of the pirate in a theater by employing visual-audio fingerprints. More precisely, first the proposed framework achieves spatio-temporal alignments of the source movie and the illegal video by exploiting visual-audio fingerprints. Then, it analyzes the geometric distortions in the pirate video and computes the projective matrix. After this step, the camcorder optical axis to the screen perpendicular is calculated by redefining the theater projective geometry and consequently the position of the pirate is estimated. In this way, the proposed framework demonstrates that the visual-audio fingerprints extracted from the master and duplicate video sequences could be exploited successfully for finding the illegal capture location in a theater. Further, the *stable key point pairs* selection algorithm, which efficiently extracts the most similar key point pairs from the temporally aligned frames, is one of the main contributions of the proposed framework.

Strictly speaking, The proposed position estimation methodology is a brand-new application of the video fingerprinting technique, which helps to find out where the

pirate was during illegal recording and subsequently restricts camcorder piracy. To validate this view point, In-Theater experiments are carried out and evaluated. More specifically, experiments are conducted in a large-scale test environment with 176 seats and ten arbitrary locations are employed for camcorder captures. The statistical analysis of position estimation results demonstrate the satisfactory performance of the proposed framework. Specifically, the mean absolute error of estimation results is (38.25, 22.45, 11.11) cm and the standard deviation of the estimation errors is (22.26, 12.97, 7.29) cm respectively. In addition, the position estimation results in terms of top, isometric and 3-D views of actual and estimated camcorder locations demonstrate the reasonable performance of the proposed framework in the test environment. The proposed framework could be used for applications such as sensor forensics, which attempts to identify the acquisition device that illegally captured the movie.

Related Publications

Journal Articles

1) R. Roopalakshmi and G. Ram Mohana Reddy, *Estimating the Position of the Pirate Using Content-Based Visual-Audio Fingerprints*, submitted to Springer Signal, Image and Video Processing.

Chapter 7

Conclusion and Future Work

Due to the massive growth of on-line publishing activities, pirated videos/movies are proliferating on the Internet and causing huge piracy issues. Therefore, duplicate video detection and tracking techniques are essential in order to restrict piracy as well as copyright issues. Although video copy detection and registration techniques are studied from the past several years, yet illegal video analysis is still a challenging problem due to the constraints such as computational cost and fingerprint size. All the work in this thesis is directed towards introducing efficient techniques for restricting piracy, that involves video copy detection, registration of video copies, geometric distortions estimation and approximation of the pirate location in a movie theater. Specifically, this research work employs video fingerprints derived from content-based features such as visual, audio, motion activity and multimodal signatures for achieving the above mentioned tasks.

The first set of contributions of this thesis target at Content-Based video Copy Detection (CBCD), by introducing novel video fingerprints for detecting video copies. More precisely, this thesis first introduces two CBCD schemes, which employ visual fingerprints derived from Dominant Color Descriptors (DCDs) for identifying illegal video sequences. However, state-of-the-art CBCD schemes employ only visual features for detecting video copies; hence, new copy detection schemes are proposed by utilizing acoustic features such as MFCCs and spectral descriptors. Further, this research study presents an efficient copy detection method by integrating different attributes of Motion Activity descriptor such as motion intensity and dominant direction of activity for detecting illegal videos. On the other hand, the integrated exploitation of visual-audio features for the CBCD task, not only enhances the detection accuracy; but also widens the coverage to more number of video modifications. Based on this aspect, this thesis proposes two robust CBCD frameworks, by jointly exploiting the

visual-audio features such as DCDs, MFCCs, audio spectral descriptors and motion activity features for identifying the duplicate video sequences. In future work, the proposed CBCD techniques can be further enriched to address the following issues:

- ⋆ Robustness against video transformations such as changing the foreground or background content, camcording and combined visual-audio attacks can be enhanced. For the foreground or background change attacks, if the local features such as SURF and SIFT are jointly utilized with audio signatures, then the detection accuracy can be considerably improved. In case of camcording attacks, normally the camcorded copies suffer from distortions such as zooming in, cropping and brightness changes. On the other hand, robustness against combined attacks is not fully achievable via visual fingerprints. Therefore, robust frameworks employing global and local visual features along with acoustic signatures are needed to deal with camcording and combined visual-audio transformations.

- ⋆ In CBCD systems, high-dimensional reference databases need to be compared in order to identify the duplicate videos. Therefore, popular indexing and similarity matching algorithms such as locality sensitive hashing (LSH) can be employed to achieve faster detection of video copies.

- ⋆ The computational cost of the proposed CBCD systems can be further reduced, if the fingerprints extraction and similarity matching tasks are executed in parallel computing paradigm.

- ⋆ In the proposed copy detection methods, it is assumed that the video copy is derived from a single master video. However, if the duplicate video contains portions of multiple master videos, then efficient similarity matching techniques are required to ensure the best matching video.

- ⋆ While exploiting audio-visual features for detecting video copies, the fusion techniques play a vital role in determining the copy detection accuracy. Therefore, suitable fusion schemes at different levels such as decision-level and feature-level can be adapted for combining the audio and visual features, which may improve the copy detection performance.

The second set of contributions of this thesis attempt to address the video copy registration problem, by presenting robust registration schemes for achieving accurate frame alignments of the pirate clip and the master video sequences. Precisely, this thesis first proposes a temporal registration scheme, which employs multimodal fingerprints for obtaining temporal frame alignments of the pirate video with the master

content. Further, this research work contributes a robust temporal-geometric alignment scheme, by employing SURF signatures. However, inclusion of audio signatures in the registration task, significantly enhances the system accuracy. Therefore, this study proposes a novel spatio-temporal registration framework, which exploits visual-audio fingerprints for obtaining accurate frame alignments of the master and copied video sequences. Extensive evaluations on three different datasets demonstrate the consistent performance of the proposed framework against different video editing and transformations. There is a scope for further improvement in the methods proposed for video copy registration, which are given below:

- ⋆ Efficient representation of a video sequence using compact fingerprints is the major challenge in duplicate video registration schemes. Hence, further investigation can be carried out on new video features and fingerprint matching techniques.

- ⋆ For temporal alignment of frames, well-known pairwise sequence alignment algorithms in bio-informatics such as Basic Local Alignment Search Tool (BLAST) technique can be adapted, which ensure optimal mappings between two feature sequences.

- ⋆ The scalability of the proposed registration methods can be extended by increasing size of the datasets used for experiments. Specifically, analysis using large-sized camcorded datasets would be interesting to address the scalability issues.

- ⋆ Selecting subset of representative/key frames from the temporally synchronized video sequences is necessary to improve the geometric registration accuracy. Therefore, robust key frame filtering techniques can be adapted, which provide representative frames having higher distribution of control points.

The third set of contributions of this thesis target at geometric distortions estimation problem, by presenting a distortion estimation framework for computing the distortion model. Precisely, this thesis presents a novel framework for estimating geometric distortions in video copies, which incorporates visual fingerprints derived from SURF signatures and audio signatures extracted from MFCCs. The fourth set of contributions of this thesis attempt to address the movie pirates identification problem, by introducing a forensic tracking framework to investigate the location of the pirate in a movie theater. Specifically, this research work investigates a forensic tracking framework, which employs visual-audio fingerprints to estimate the position of the

pirate in a movie theater irrespective of presence/absence of watermarks. In-Theater experimental results demonstrate the satisfactory performance of the proposed framework. Further investigation on geometric distortions estimation and pirate position approximation can be carried out as listed below:

- ⋆ The accuracy of the distortion estimation approach can be enhanced, by employing robust estimation algorithms such as Random Sample Consensus (RANSAC) or Least Median Squares, which map greatest number of point pairs between two images.

- ⋆ As per the digital cinema standards, aspect ratio of the theater screen varies between 1.85:1 to 2.35:1. However, the screen aspect ratio of the test environment is about 1.33:1. Further, the test environment comprises a flat screen, whereas actual theaters need screen curvatures for perfect projection. For these reasons, if experiments are conducted in a real-theater, then the position estimation accuracy of the proposed forensic framework can be substantially improved.

- ⋆ The proposed position estimation framework can be further extended to applications such as sensor forensics, to verify whether two video clips are captured by the same camcorder/acquisition device or not.

To summarize, this thesis attempts to provide efficient solutions for combating Internet as well as camcorder piracy. Precisely, this research work introduces video copy detection and tracking schemes, which prevent downloading and distribution of illegal contents on the Internet and thereby restrict Internet piracy. In addition, this thesis proposes a forensic tracking framework, which attempts to estimate the illegal camcorder capture locations, so that the camcorder piracy can be controlled.

Bibliography

Anguera, X., Obrador, P., and Oliver, N. (2009). "Multimodal Video Copy Detection Applied to Social Media". *Proc. ACM Int. conf. WSM'09*. 57-64.

Baudry S., Chupeau B., and Lefebvre F. (2009). "A framework for video forensics based on local and temporal fingerprints". *Proc. IEEE Int. Conf. on Image Processing ICIP'09*. 2889-2892.

Baudry S., Chupeau B., and Lefebvre F. (2010). "Adaptive Video Fingerprints for Accurate Temporal Registration". *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP'10*. 1786-1789.

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. J. V. (2008). "Speeded-up robust features (surf)". *Computer Vision Image Understanding*. 110:346–359.

Benini, S., Xu, L. Q., and Leonardi, R. (2005). "Using lateral ranking for motionbased video shot retrieval and dynamic content characterization". *Proc. Content-Based Multimedia Indexing (CBMI05)*.

Boreczky, J. S., and Wilcox, L. D. (1998). "A hidden Markov model framework for video segmentation using audio and image features". *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-98)*. 3741-3744.

Burka, Z. (2010). "Perceptual Audio Classification Using Principal Component Analysis". *M.S. Thesis*.

Cao, Z., and Zhu, M. (2009). "An Efficient Video Copy Detection Method Based on Video Signature". *Proc. Int. Conf. on Automation and Logistics*. 851-859.

Chaisorn, L., Sainui, J., and Mander, C. (2010). "A Bitmap Indexing Approach for Video Signature and Copy Detection". *Proc. 5th IEEE Conf. on Industrial Electronics and Applications*. 1996-2001.

Chartrand, G. (1977). "Introductory graph Theory". *Courier Dover Publications*.

Chen, L., and Stentiford, F. W. M. (2008). " Video sequence matching based on temporal ordinal measurements". *Pattern Recognition Letters*. 29: 1824-1831.

Chen, N., Xiao, H. D., and Wan, W. (2011). "Audio hash function based on non-negative matrix factorization of mel-frequency cepstral coefficients". *IET Information security*. 5: 19-25.

Cheng H. (2003). "Temporal Registration of Video Sequences". *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP'03*. 489-492.

Cheng H. (2004). "A Review of Video Registration Methods for Watermark Detection in Digital Cinema Applications". *Proc. IEEE International Symposium on Circuits and Systems (ISCAS'04)*. 704-707.

Cheng H., and Isnardi M.A. (2003). "Spatial, temporal and histogram video registration for digital watermark detection". *Proc. IEEE Int. Conf. on Image Processing ICIP'03*. 735-738.

Cheung, S. C., and Zakhor, A. (2000). "Estimation of web video multiplicity". *Proc. Soc. Photo-Optic. Instrum. Eng.- Internet Imaging*. 3964:34-36.

Chiu, C. Y., Chen, C. S., and Chien, L. F. (2008). "A framework for handling spatiotemporal variations in video copy detection". *IEEE Trans. Circuits and Sys. for Video Tech.*. 18:412-417.

Chiu, C. Y., Wang, H. M., and Chen, C. S. (2010). "Fast min-hashing indexing and robust spatio-temporal matching for detecting video copies". *ACM Transactions on Multimedia Comp. ,Commns. and App.*. 6:1-23.

Cho, H. J., Lee, Y. S., Sohn, C. B., Chung, K. S., Oh, S. J. (2009). " A Novel Video Copy Detection Method Based on Statistical Analysis". *Proc. Int. Conf. on Multimedia & Expo*. 1736-1739.

Chiu, C. Y., and Wang, H. M. (2010). "Time-Series Linear Search for Video Copies Based on Compact Signature Manipulation and Containment Relation Modeling". *IEEE Trans. Circuits and Sys. for Video Tech.*. 20:1603-1613.

Cheung, S. C., Zakhor, A. (2000). "Estimation of web video multiplicity". *in Proc. Soc. Photo-Optic. Instrum. Eng.*. 3964:34–36.

Chupeau B., Oisel L., and Jouet P. (2006). "Temporal video registration for watermark detection". *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. ICASSP'06*. 157-160.

Chupeau B., Massoudi A., and Lefebvre F. (2007). "Automatic Estimation and Compensation of Geometric Distortions in Video Copies". *Proc. SPIE, Visual Communication and Image Processing*. 6508.

Chupeau, B., Massoudi, A., and Lefèbvre, F. (2008). "In-theater piracy: finding where the pirate was". *Proc. SPIE, Security, Forensics, Steganography & Watermarking of Multimedia Contents X*. 6819.

Cieplinski, L. (2001). "MPEG-7 color descriptors and their applications". *in Proc. Lecture Notes in Computer Science*.

Cui, P., Wu, Z., Jiang,S., and Huang, Q. (2010). "Fast copy detection based on slice entropy scattergraph". *Proc. IEEE Int. Conf. Multimedia & Expo*. 1236-1241.

Delannay, D., Delaigle, F., Demarty, H., and Barlaud, M. (2001). "Compensation of

Geometrical deformations for Watermark Extraction in Digital Cinema Applications". *Proc. SPIE Electronic Imaging*. 4314: 149-157.

Delannay D., Roover C., and Macq B. (2003). "Temporal alignment of video sequences for watermarking systems". *in Proc. of IS & T SPIE 15th Annual Symp. on Electronic Imaging*. 5020:481-492.

Deng, Y., Manjunath, B. S., Kenney, C., Moore, M. S., and Shin, H. (2001). "An efficient color representation for image retrieval". *IEEE Transactions on Image Processing*. 10:140-147.

Divakaran, A., Regunathan, R., and Peker, K. A. (2001). "Video summarization using descriptors of motion activity: A motion activity based approach to keyframe extraction from video shots". *Journal of Electronic Imaging*. 10:909-916.

Douze, M., Jgou, H., and Schmid, C. (2010). "An Image-Based Approach to Video Copy Detection With Spatio-Temporal Post-Filtering". *IEEE Transactions on Multimedia*. 20:257-266.

Eronen, A., and Klapuri, A. (2000). "Musical instrument recognition using cepstral coefficients and temporal features". *in Proc.of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2000*. 1753-1756.

Esmaeili, M. M., Fatourechi, M., and Ward, R. K. (2011). "Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting" *IEEE Transactions on Information Forensics and Security*. 6:213-226.

Goemans, M. X. (2007). "Lecture notes on bipartite matching" *Massachusetts Institute of Technology*.

Gu, J., Lu, L., Cai, R., Zhang, H. J., and Yang, J. (2004). " Dominant Feature Vectors Based Audio Similarity Measure" *in Proc. of PacificRim conference on Multimedia(PCM'04)*. 890-897.

Gupta, V., Varcheie, P., D., Z., Gagnon, L., Boulianne, G. (2012). " Content-based video copy detection using nearest-neighbor mapping" *in Proc. of 11th Int. Conf. on Inf. Sciences, Signal proc. and their Appl.*. 918-923.

Haitsma, J., and Kalker, T. (2001). " A watermarking scheme for digital cinema" *in Proc. of Int. Conf. on Image Processing*. 487-489.

Hampapur, A., Hyun, K., and Bolle, R. M. (2001). "Comparison of sequence matching techniques for video copy detection" *SPIE Storage and Retrieval for Media Databases*. 4676: 194-201.

Hartley, R. I., and Zisserman, A. (2004). "Multiple View Geometry in Computer Vision" *Cambridge Univ. Press*.

Hoad, T. C. and Zobel, J. (2006). "Detection of video sequence using compact signatures" *ACM Transactions on Information Systems*. 24: 1-50.

Hua, X. S., Chen, X., and Zhang, H.J. (2004). "Robust video signature based on

ordinal measure" *in Proc. of IEEE Int. Conf. on Image Processing (ICIP)*. 685-688.

Indyk, P. (2000). "High-dimensional computational geometry" *Ph.D. Dissertation*. Stanford University.

Itoh, Y., Erokuumae, M., Kojima, K., Ishigame, M., and Tanaka, K. (2010). "Time-space Acoustical Feature for Fast Video Copy Detection" *in proc. of 2010 IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*. 487-492.

Jeannin, S., and Divakaran, A. (2001). "MPEG-7 visual motion descriptors" *IEEE Transactions on Circuits and Systems for Video Technology*. 11:720-724.

Jiang, S., Su, L., Huang, Q., Cui, P., and Wu, Z. (2013). "A rotation invariant descriptor for robust video copy detection" *The Era of multimedia*. Springer Press. 557-567.

Jie, T., Gang, L., and Jun, G. (2009). "Improved Algorithms of Music Information Retrieval based on Audio Fingerprint" *in Proc. of Third Int. Symp. on Intelligent Information Technology Application Workshops*. 367-371.

Joly, A., Buisson, O., and Frelicot, C. (2007). "Content-based copy retrieval using distortion- based probabilistic similarity search" *IEEE Transactions on Multimedia*. 9:293-306.

Junee, R. (2009). "Zoinks! 20 Hours of Video Uploaded Every Minute!". `http://www.youtube.com/blog?entry=on4EmafA5MA`.

Kashiwagi, T., and Oe, S. (2007). "Introduction of Frequency Image and applications" *in Proc. of SICE Annual Conference*. 584-591.

Kim, C. and Vasudev, B. (2005). "Spatiotemporal sequence matching for efficient video Copy detection" *IEEE Transactions on Circuits and Systems for Video Tech.*. 15:127-132.

Kim, H., Lee, J., Liu, H., and Lee, D. (2008). "Video Linkage: Group Based Copied Video Detection" *in Proc. of CIVR08*. 685-688.

Kim, J., and Nam, J. (2009). "Content-based Video Copy Detection using Spatio-Temporal Compact Feature" *in Proc. of Int. Conf. on Advanced Communication Technology*. 1667-1671.

Koprinska, I. and Carrato, S. (2001). "Temporal video segmentation: a survey" *Elsevier Signal Processing:Image Communication*.

Küçüktunç, O., Baştan, M., Güdükbay, U., and Ulusoy, O. (2010). "Video copy detection using multiple visual cues and MPEG-7 descriptors" *Elsevier Journal of Visual Communication and Image Representation*. 21: 838-849.

Kuhn, H. (1955). "The Hungarian Method for the Assignment problem" *Naval Research logistics*. 2: 83-97.

Law-To, J., Chen, L., Joly, A., Laptev, I., Buission, O., Gouet-Bbrunet, V., Bouje-

maa, N., and Stentiford, F. (2007). "Video copy detection: a comparative study" *In Proc. of ACM Int. Conf. on Image and Video Retrieval (CIVR)*. 371-378.

Lee Y.Y., Kim C., and Lee S. (2009). "Video frame matching algorithm using dynamic programming" *Proc: SPIE and IS & T Journal of Electronic Imaging*. 18:1.

Lee, M.J., Kim, K.S., and Lee, H.K. (2010). "Digital Cinema Watermarking for Estimating the Position of the Pirate" *IEEE Trans. Multimedia*. 12:605-621.

Lefèbvre, F., Chupeau, B., Massoudi, A., and Diehl, E. (2009). "Image and Video Fingerprinting:Forensic Applications" *in Proc.of SPIE and IS & T, Journal of Electronic Imaging*. 7254.

Lei, Y., Luo, W., Wang, Y., and Huang, J. (2012). "Video sequence matching based on the invariance of color correlation" *IEEE Trans. Circuits & Sys. for Video Tech.*. 22: 1332-1343.

Li, T., Ogihara, M., and Li, Q. (2003). "A Comparative Study on Content- Based Music Genre Classification" *in Proc. of SIGIR-03*.

Lian, S., Nikholaidas, N., and Sencar, H., T. (2010). "Content-based video copy detection- A survey" *in Proc. of Springer Intel. multimedia analysis for security app.*. 253-273.

Liu, L., Lai, W., Hua, X. S., and Yang, S. Q. (2007). "Video histogram: A novel video signature for efficient web video duplicate detection". *Advances Multimedia Modeling*. 4352:94-103.

Lloyd, S. P. (1982). "Least Squares Quantization in PCM". *IEEE Transactions on Information Theory*. 28:129-137.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant key points". *International Journal of Computer Vision*. 60: 91-110.

Manjunath, B. S., Salembier, P., and Sikora, T. (2002). "Introduction to MPEG-7 - Multimedia Content Description Interface". *John Wiley and Sons*.

Müller, M. (2007). "Information retrieval for music and motion". *Springer press*. XVI, Edition I.

Nakashima, Y., Tachibana, R., and Babaguchi, N. (2009). " Watermarked movie soundtrack finds the position of the camcorder in a theater". *IEEE Trans. on Multimedia*. 11:443-454.

Naphade, M. Y. M., and Yeo, B.L. (2000). "A novel scheme for fast and efficient video sequence matching using compact signatures". *in Proc. SPIE, Storage and Retrieval for Media Databases*. 3972:564-572.

Natsev, A., Hill, M., and Smith, J. R. (2010). "Design and Evaluation of an Effective and Efficient Video Copy Detection System". *in Proc. IEEE Int. Conf. on Multimedia & Expo ICME'10*. 1353-1358.

Özer, H., Sankur, B., Memom, N. and Anarim, E. (2005). "Perceptual audio hashing

functions". *EURASIP Journal of Applied signal procesing*. 12: 1780-1793.

Park, T. H. (2010) "Introduction to digital signal processing-Computer musically speaking". *World scientific Press*.

Pikrakis, A., Giannakopoulos, T., and Theodoridis, S. (2008). "An overview of speech/music discrimination techniques in the context of audio recordings". *in Proc. IEEE Int. Conf. on Multimedia & Expo ICME'10*. 120:81-102.

Poullot, S., Crucianu, M., and Buisson (2008). "Scalable mining of large video databases using copy detection". *in Proc. of ACM International Conference on Multimedia (MM'08)*. 61-70.

Rabiner, L. and Juang, B. H. (1993). "Fundamentals of Speech Recognition". *Prentice Hall Signal Proc. Series*.

Ren, H., Lin, S., Zhang, D., Tang, S., and Gao, K. (2009). "Visual Words based Spatiotemporal Sequence Matching in Video Copy Detection". *in proc. of IEEE Int. conf. on Multimedia & Expo ICME'09*. 1382-1385.

Ren, J., Chang, F., Wood, T., and Zhang, J., R. (2012). "Efficient Video Copy Detection via Aligning Video Signature Time Series". *in proc. of ACM ICMR'12*.

Roopalakshmi, R., and Reddy, G. R. M. (2010). "Recent Trends in Content-Based Video Copy Detection". *in proc. of IEEE Int. Conf. on Computational Intelligence and Computing Research (ICCIC'10)*. 1-5. `Doi:http://dx.doi.org/10.1109/ICCIC.2010.5705802`.

Roopalakshmi, R., and Reddy, G. R. M. (2011). "A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA". *Elsevier Procedia Computer Science*. 5:149-156.

Roopalakshmi, R., and Reddy, G. R. M. (2011). "Efficient Video Copy Detection Using Simple and Effective Extraction of Color Features". *Springer-Verlag, Lecture Series - LNCS CCIS 193*. 4:473-480.

Roopalakshmi, R., and Reddy, G. R. M. (2013). "A Novel Spatio-Temporal Registration Framework for Video Copy Localization based on Multimodal Features". *Elsevier Signal Processing*. Vol. 93, 8:2339-2351. ISSN: 0165-1684. Available:`http://dx.doi.org/10.1016/j.sigpro.2012.06.004`.

Roth, G., Laganiere, R., Lambert, P., Lakhmiri, and Janati T. (2010). "A Simple but Effective Approach to Video Copy Detection". *in proc. of Canadian Conf. Computer and Robot Vision*. 63-70.

Roytman, E., and Gotsman, C. (1995). "Dynamic Color Quantization of Video Sequences". *IEEE Transactions. Visualization and Computer Graphics*. 1: 274-286.

Sankoff, D. (2000). "The early introduction of dynamic programming into computational biology". *Journal of Bioinformatics*. 16: 41-47.

Saracoğlu, A., Esen, E., Ateş, T. K., Acar, B.O., Zubari, Ozan, E. C., özalp, E.,

Alatan, A. A., Çiloglu, T. (2009). "Content Based Copy Detection with Coarse Audio-Visual Fingerprints". *in proc. of Seventh Int. Workshop on Content-Based Multimedia Indexing (CBMI.* 213-218.

Sarkar, A., Ghosh, P., Moxley, E., and Manjunath, B. S. (2008). "Video fingerprinting: features for duplicate and similar video detection and querybased video retrieval". *in Proc. of SPIE Multimedia Content Access: Algorithms and Systems II.*

Sarkar, A., Singh, V., Ghosh, P., Manjunath, B. S., and Singh, A. (2010). "Efficient and Robust Detection of Duplicate Videos in a Large Database". *IEEE Transactions on Circuits and Systems for Video Technology.* 20:870-885.

Savakis, A., Sniatala, P., and Rudnicki, R. (2003). "Real-time video annotation using MPEG-7 motion activity descriptors" *in Proc. of MIXDES'03.*

Schmid, C., and Mohr, R. (1997). "Local grayvalue invariants for image retrieval". *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 19:530-535.

Schoeffmann, K., Boeszoermenyi, L.(2011). "Video Sequence Identification in TV Broadcasts". *Springer Advances in Multimedia Modeling.* 129-139.

Senin, P. (2008). "Dynamic Time Warping Algorithm Review". *Information and Computer Science Dept., University of Hawaii.*

Shivakumar. (1999) "Detecting digital copyright violations on the Internet". *Ph.D. Dissertation*, Stanford University.

Sun, X., Ajay, D., and Manjunath, B. S. (2001). "A motion activity descriptor and its extraction in compressed domain". *in Proc. of IEEE Pacific-Rim Conf. Multimedia.* 450-453.

Park, T.H. (2010). "Introduction to digital signal processing- Computer musically speaking". *World scientific Press.*

Tang, J., Liu, G., and Guo, J. (2009) "Improved Algorithms of Music Information Retrieval based on Audio Fingerprint". *in proc. Third Int. Symp. on Intelligent Inf. Tech. App. Workshops.*

Tasdemir, K., and Cetin, A. E. (2010) "Motion vector based features for content based video copy detection". *in proc. of IEEE Int. Conf. on Pattern Recognition'10.* 3134-3137. DOI:10.1109/ICPR.2010.767.

Park, T.H. (2010). "Introduction to digital signal processing- Computer musically speaking". *World scientific Press.*

Tian, Y., Jiang, M., Mou, L., Fang, X., and Huang, T. (2011) "A multimodal video copy detection approach with sequential pyramid matching". *in proc. of 18th IEEE Int. Conf. on Image Processing.* 3629-3632.

Tsekeridou, S., and Pitas, I. (2001) "Content-Based Video Parsing and Indexing Based on AudioVisual Interaction". *IEEE Transactions on Circuits and Systems for Video Technology.* 11:522-535.

Uchida, Y., Takagi, K., and Sakazawa, S. (2012) "Fast and accurate content-based video copy detection using bag-of-global visual features". *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP'12*. 1029-1032.

Wang, Y., Liu, Z., and Huang, J. C. (2000) "Multimedia Content Analysis using both audio and visual cues". *IEEE Signal Processing Magazine*. 12-36.

Wei, S., Zhao, Y., Zhu, C., Xu, C., and Zhu, Z. (2011) "Frame Fusion for Video Copy Detection". *IEEE Trans. Circuits Sys. Video Tech.*. 21:15–28.

West, K. (2008) "Novel techniques for Audio Music Classification and Search". *Doctoral Thesis*.

Wu, X., Hauptmann, A. G., and Ngo, C. (2007) "Practical elimination of near-duplicates from web video search ". *in Proc. of Int. Conf. on Multimedia*. 218-227.

Wu, S., and Zhao, Z. (2012) "A Multi modal content-based copy detection approach ". *in Proc. of 8th Int. Conf. computational Intelligence & security*. 280-283.

Xu, Z., Ling, H., Zou, F., Lu, Z., Li, P., and Wang, T. (2009) "Fast and robust video copy detection scheme using full DCT coefficients". *in Proc. of IEEE Int. Conf. on Multimedia & Expo*. 434-437.

Xu, J., Bai, Q., Gu, Y., Tung, K.H.A., Wang, G., Yu, G., Zhang, Z. (2012) "EU-DEMON: A System for Online Video Frame Copy Detection by Earth Movers Distance". *in Proc. of IEEE 28th Int. Conf. on Data Eng.*. 1233-1236.

Yang, N. C., Chang, W. H., Kuo, C. M., Li, T. S. (2008) "A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval". *Elsevier journal of Visual Communication and Image Representation*. 19: 92-105.

Yang, G., Chen, N., and Jiang, Q. (2012) "A robust hashing algorithm based on SURF for video copy detection". *Elsevier Computers & Security*. 31: 33-39.

Zhai, Y., Shah, M. (2005) "Tracking news stories across different sources". *in Proc. 13th annual ACM int. conf. on Multimedia*. 2-10.

Zhang, Z., and Zou, J. (2010) "Compressed Video Copy Detection Based on Edge Analysis". *in Proc. 2010 IEEE Int. Conf. on Information and Automation*. 2497-2501.

Zhang, Z., Cao, C., Zhang, R., and Zou, J. (2010) "Video Copy Detection Based on Speeded Up Robust Features and Locality Sensitive Hashing". *in Proc. IEEE Int. Conf. on Automation and Logistics*. 13-18.

Zhu, S., Yan, J., and Liu, Y. (2009) "Improving Semantic Scene Categorization by Exploiting Audio-Visual Features". *in Proc. of Fifth Int. Conf. on Image and Graphics*. 435-440.

Publications

List of Publications/Communications Based on Thesis:

International Journals

1) R. Roopalakshmi, G. Ram Mohana Reddy, *A Novel Spatio-Temporal Registration Framework for Video Copy Localization Based on Multimodal Features*, published in **Elsevier Signal Processing** Journal, Vol. 93, Issue 8, Pages 2339-2351, Aug'2013. ISSN: 0165-1684.
Available: `http://dx.doi.org/10.1016/j.sigpro.2012.06.004`.

2) R. Roopalakshmi and G. Ram Mohana Reddy, *A Framework for Estimating Geometric Distortions in Video Copies Based on Visual-Audio Fingerprints*, Published in **Springer Signal, Image and Video Processing (SIViP)** Journal, Vol.7, Issue 1, Jan'2013. ISSN: 1863-1703.
Available: `http://link.springer.com/article/10.1007/s11760-013-0424-7`.

3) R. Roopalakshmi, G. Ram Mohana Reddy, *A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA*, Published in **Elsevier Procedia Computer Science** Journal, Vol. 5, Pages 149-156, 2011. ISSN: 1877-0509.
Available: `http://dx.doi.org/10.1016/j.procs.2011.07.021`

4) R. Roopalakshmi and G. Ram Mohana Reddy, *Estimating the Position of the Pirate Using Content-Based Visual-Audio Fingerprints*, submitted to Springer Signal, Image and Video Processing.

5) R. Roopalakshmi and G. Ram Mohana Reddy, *Robust Temporal Registration Scheme for Video Copies Using Multimodal Features*, submitted to Springer Multimedia Systems.

Book Chapters

1) R. Roopalakshmi and G. Ram Mohana Reddy, *Efficient Video Copy Detection Using Simple and Effective Extraction of Color Features*, published in Springer Book titled, 'Advances in Computing and Communications', CCIS, Vol. 193,

Part IV, Pages 473-480, 2011. ISSN: 1865-0929. Available:
http://link.springer.com/chapter/10.1007/978-3-642-22726-4_49.

2) R. Roopalakshmi and G. Ram Mohana Reddy, *Content-Based Video Copy Detection Using Motion Activity and Acoustic Features*, published in Springer Book titled, 'Advances in Intelligent Systems and Computing', Vol. 264, Pages 491-504, 2014. ISSN: 2194-5357. Available:
http://link.springer.com/chapter/10.1007/978-3-319-04960-1_43.

Conference Publications

1) R. Roopalakshmi and G. Ram Mohana Reddy, *Recent Trends in Content-Based Video Copy Detection*, in proc. of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, pp. 1-5, Dec'2010.
Available: http://dx.doi.org/10.1109/ICCIC.2010.5705802.

2) Roopalakshmi, R. and Reddy, G.R.M. *Compact and Efficient CBCD Scheme Based on Integrated Color Features*, in proc. of International Conference on Recent Trends in Information Technology (ICRTIT), Anna University, Chennai, India, pp. 880-883, June'2011.
Available: http://dx.doi.org/10.1109/ICRTIT.2011.5972370.

3) R. Roopalakshmi and G. Ram Mohana Reddy, *Efficient Video Copy Detection Using Simple and Effective Extraction of Color Features*, in proc. of International Conference on Advances in Computing and Communications (ACC-2011), Kochi, India, pp. 473-480, July'2011.
Available: DOI:10.1007/978-3-642-22726-4_49.

4) R. Roopalakshmi, G. Ram Mohana Reddy, *A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA*, in proc. of Second International Conference on Ambient Systems, Networks and Technologies (ANT-2011), Niagara Falls, Canada, 5, pp. 149-156, Sep'2011.
Available: DOI:10.1016/j.procs.2011.07.021.

5) Roopalakshmi, R. and Reddy, G.R.M. *A Novel CBCD Approach Using MPEG-7 Motion Activity Descriptors*, in proc. of IEEE International Symposium on Multimedia (ISM-2011), University of California, USA, pp. 179-184, Dec'2011.
Available: http://dx.doi.org/10.1109/ISM.2011.36.

6) R. Roopalakshmi and G. Ram Mohana Reddy, *Towards a New Approach to Video Copy Detection Using Acoustic Features*, in proc. of IEEE 5th Interna-

tional Conference on Internet Multimedia Systems Architecture and Applications (IEEE IMSAA-2011), Indian Institute of Information Technology Bangalore (IIIT-B), India, pp. 1-5, Dec'2011.
Available: http://dx.doi.org/10.1109/IMSAA.2011.6156336.

7) R. Roopalakshmi and G. Ram Mohana Reddy, *Robust Features for Accurate Spatio-Temporal Registration of Video Copies*, in proc. of IEEE International Conference on Signal Processing and Communications (SPCOM-2012), Indian Institute of Science (IISc), Bangalore, India, pp. 1-5, July'2012.
Available: http://dx.doi.org/10.1109/SPCOM.2012.6290006.

8) R. Roopalakshmi and G. Ram Mohana Reddy, *Content-Based Video Copy Detection Using Motion Activity and Acoustic Features*, in proc. of International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS-2014), Indian Institute of Information Technology and Management-Kerala (IIITMK), India, pp. 491-504, March'2014.
Available: DOI:10.1007/978-3-319-04960-1_43

Brief Bio-Data

Roopalakshmi
Research Scholar
Information Technology Department
National Institute of Technology Karnataka Surathkal
P.O.Srinivasanagar
Mangalore 575025

Permanent address

Roopalakshmi, $W/0$ Nagendran V,
No-737, 9th cross, Vidyamanya Nagara,Vishwaneedam Post,
Bangalore-560091,
Karnataka.
Phone: 09972246013
Email: roopanagendran2002@gmail.com

Qualification

- M.Tech. Computer Science, Visvesvaraya Technological University (VTU), Karnataka, 2008.