Volume 33 Number 1 March 2009 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

Multimedia Information System Security

Guest Editors:

Shiguo Lian

Dimitris Kanellopoulos

Giancarlo Ruffo

1977

EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering the European computer science and informatics community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Editorial:
Special Issue on Multimedia Information System Security

1 Introduction

With the rapid progress in information technology and an enormous amount of media (e.g. text, audio, speech, music, image and video) appearing over networks, guaranteeing multimedia information system security is becoming increasing important. Several pivotal challenges include copyright protection, integrity verification, authentication, access control, privacy protection, etc. As a consequence, the subject of multimedia information system security has attracted intensive research activities in academy, industry and also government.

With the recent advances in network and multimedia technology, the applications in commercial scenario become increasing crucial. There is an increasing trend in multimedia systems including multimedia distributed computing, multimedia databases, multimedia communications, etc. For example, in multimedia distributed computing, the multimedia content is delivered from the central service provider to the individuals using various techniques/applications such as video-on-demand, IPTV and P2P content distribution. In these applications, piracy is becoming a critical issue. Solutions are needed to protect the copyright of multimedia content. During the past decades, various techniques have been reported for secure multimedia information system such as key management, multimedia encryption, authentication, digital watermarking, digital fingerprinting, secure data mining, access control and digital rights management. These techniques are able to protect multimedia content's confidentiality, integrity, ownership, traitor traceability. In addition, in different networks such as Internet, 3G wireless, DVB-H and P2P, different secure protocols and algorithms are required to provide the system's security. All these topics are in active development.

This special issue of the Informatica Journal invited authors to submit their original work that communicates current research on multimedia information system security regarding both the novel solutions and future trends in the field. In this special issue, we have 7 papers, which can demonstrate advanced works in the field including covert communication in multimedia carriers, secret information analysis from multimedia content, copyright protection, multimedia content encryption, secure mobile multimedia communication, and secure multimedia content indexing or retrieval.

2 The papers in this special issue

In the first paper, entitled "Recent advances in multimedia information system security" we survey techniques and tools used for multimedia information system security. In addition, we present the latest research progress in the field as well as hot research topics such as Trusted Computing, security in network, and security of content sharing in social networks, privacy-preserving data processing, multimedia forensics, intelligent surveillance and steganography.

Steganography is a hot topic belonging to covert communication techniques, which hides secret information into multimedia content and thus sends it to receivers. Only the receiver partnered with the sender can extract the secret information from multimedia carrier. The third party can only detect whether the multimedia content is suspicious and then decide whether to remove it. Steganography faces two threats, i.e., steganalysis and unstable transmission. The former one denotes the technique to detect the presence of secret information based on statistical abnormality caused by information hiding. The latter one means the unstable transmission that causes transmission errors or losses to multimedia content. Although some steganalysis techniques have been proposed, their detection performances are still not satisfied. One important reason is that the methods can only detect certain statistical abnormality, such as the changes in histogram or pixel correlation.

In the second paper, S. Geetha, Siva S. Sivatha Sindhu and N. Kamaraj propose a method to distinguish the plain media and stego media by detecting content independent statistical features. In particular, they designed a feature classification technique, which is composed of two steps: training step and detection step. In the training step, the feature classifier is trained by the database composed of both plain images and stego images (generated by using different information hiding methods). Then, in the detection step, the given image is decided by the classifier automatically. Compared with existing schemes, their scheme does not depend on the steganographic methods.

For steganography, it is a challenge to resist the unstable transmission. Since the transmission errors or losses often make the secret information unrecoverable.

In the third paper, X. Zhang, S. Wang and W. Zhang present a new steganography method that aims to resist the active warden or poor channel conditions. In their method, the secret information is decomposed into a number of shares, and then embedded into different cover images respectively. The embedding efficiency and imperceptibility are improved by the proposed share embedding method. Thus, even a part of stego images are lost during transmission, most of the shares can still be extracted from the remaining stego images, and the shares can be used to recover the secret information. The experiments and analysis show the scheme's practicability.

Digital watermarking is regarded as a potential solution for copyright protection, which embeds such copyright information as content producer, content owner or content receiver into multimedia content. Visible watermarking denotes the watermarking technique that

embeds copyright information imperceptibly. Since it does not affect the commercial value of multimedia content, visible watermarking is preferred. However, watermark detection is still a challenging topic when the original multimedia content is not accessible and the marked content is degraded.

The fourth paper by H. Malik proposes a blind watermark detection method for spread spectrum watermarking. This method regards the problem of watermark detection as a blind source separation problem, and thus uses independent component analysis to estimate the embedded watermark. Since in spread spectrum watermarking, the embedded watermark and the multimedia content are mutually independent and obey non-Gaussian distribution, the proposed detection method outperforms existing correlation-based blind detection methods. The experiments on audio clips are given to show the proposed method's good detection performances.

Multimedia encryption has being emphasized with the popular applications of multimedia content in human being's daily life. Due to such properties as large volumes and real time interaction, selective encryption is preferred for multimedia encryption, which encrypts only some significant parameters in the compressed multimedia stream while leaves other parameters unchanged. Additionally, some synchronization information, e.g., syntax information in the compressed stream, is not encrypted in order to keep the stream's error resistance. Generally, the multimedia content encrypted by selective encryption is still playable. Thus, the intelligibility of the played content is in close relation with the encryption method's security. Till now, few works have been done to assess the quality of the encrypted multimedia content.

In the fifth paper, Y. Yao, Z. Xu and S. Liu propose an assessment method based on neighborhood similarity. Firstly, the neighborhood similarity is defined and cipher images' features are analyzed. Then, the objective visual security metric is defined based on the neighborhood similarity. In experiments, the cipher videos encrypted by different algorithms are assessed with the objective metric. The experimental results show that the objective metric is consistent with the human perception, and can be used to assess the encryption method's visual security automatically.

Digital rights management (DRM) becomes more and more important for protecting the usage of multimedia content. Generally, for a multimedia service system, certain business model is firstly defined, then, the protection means are proposed to support the model. Till now, some practical DRM systems have been reported for securing applications in Internet, wireless mobile network or broadcasting network. However, more and more new applications arise with the development of network technology and multimedia technology, and the corresponding DRM solutions are expected.

In the sixth paper, M. Furini proposes a secure solution for the pervasive video lectures. In this solution, the video chapters are partitioned into two parts, i.e., the pre-defined lesson and the other lesson. The videos in the former one are in clear, while the videos in the latter one are encrypted and the corresponding encryption key is hidden in the videos in the former one. Thus, only the mobile player being able to extract the encryption key correctly can play the videos in the second part successfully. The prototype implementation shows the scheme's feasibility.

In decentralized and distributed system as peer-to-peer multimedia sharing, there are some security issues. Among them, secure multimedia indexing and retrieval is a challenge. In the last paper, W. Allasia, F. Gallo, M. Milanesio and R. Schifanella propose a decentralized, distributed and secure communication infrastructure for indexing and retrieval of multimedia contents with associated digital rights. Firstly, the existing works about Distributed Hash Table (DHT) is introduced, and some security threats are pointed out. Then, the secure DHT layer is presented, and the secure protocols are proposed. Additionally, the feasibility of proposed architecture is shown with a prototype implementation. This scheme is based on structured P2P networks and allows complex queries using standard MPEG-7 and MPEG-21 multimedia metadata. This scheme is expected to attract more researchers in this field.

The list of the papers follows:

- S. Lian, D. Kanellopoulos and G. Ruffo. Recent advances in multimedia information system security.
- S. Geetha, Siva S. Sivatha Sindhu and N. Kamaraj. Detection of stego anomalies in images exploiting the content independent statistical footprints of the steganograms.
- X. Zhang, S. Wang and W. Zhang. Steganography combining data decomposition mechanism and stego-coding method.
- H. Malik. Blind watermark estimation attack for spread spectrum watermarking.
- Y. Yao, Z. Xu and J. Sun. Visual security assessment for cipher-images based on neighborhood similarity.
- M. Furini. Secure, portable, and customizable video lectures for e-learning on the move.
- W. Allasia, F. Gallo, M. Milanesio and R. Schifanella. Indexing and retrieval of multimedia metadata on a secure DHT.

Acknowledgments

The guest editors wish to thank Prof. Anton P. Zeleznikar (Editor-in-Chief of the Informatica Journal) and Prof. Matjaz Gams (Managing Editor) for providing the opportunity to edit this special issue on Multimedia Information System Security. We would also like to thank the authors for submitting their works as well as the referees who have critically evaluated the papers within the short stipulated time. Finally, we hope the reader will share our joy and find this special issue very useful.

S. Lian, D. Kanellopoulos and G. Ruffo
Guest Editors

Recent Advances in Multimedia Information System Security

Shiguo Lian
France Telecom R&D (Orange Labs) Beijing
2 Science Institute South Rd., Haidian District, Beijing 100080, China
E-mail: shiguo.lian@orange-ftgroup.com; sglian@gmail.com

Dimitris Kanellopoulos
Educational Software Development Laboratory (ESDLab)
Department of Mathematics,
University of Patras, Greece
E-mail: d_kan2006@yahoo.gr

Giancarlo Ruffo
Department of Computer Science
University of Turin, Italy
E-Mail: ruffo@di.unito.it

Keywords: multimedia system, security, encryption, authentication, digital rights management, forensic, watermark, biometric, copy detection, data mining

Received: September 1, 2008

A multimedia communication system enables multimedia data's generation, storage, management, distribution, receiving, consuming, editing, sharing, and so on. In such systems, there are various security issues, which must be considered such as eavesdropping, intrusion, forgery, piracy and privacy, etc. Until now, various security solutions for multimedia communication systems have been reported, while few works have surveyed the latest research advances. This paper gives a thorough review to multimedia information system security. It introduces a general architecture of multimedia information system, and investigates some security issues in multimedia information systems. It reviews the latest security solutions such as Digital Rights Management (DRM), confidentiality protection, ownership protection, traitor tracing, secure multimedia distribution based on watermarking, forgery detection, copy detection, privacy-preserving data mining, secure user interface, intrusion detection and prevention. Moreover, the paper presents some hot research topics such as Trusted Computing, steganography, security in network or service convergence, security of content sharing in social networks, privacy-preserving data processing, multimedia forensics and intelligent surveillance in multimedia information system security. It is expected to benefit readers by providing the latest research progress, advising some research directions and giving a list of references about multimedia system security.

Povzetek: Prispevek podaja pregled novejših pristopov pri zagotavljanju varnosti multimedijskih informacijskih sistemov.

1 Introduction

Nowadays, multimedia information applications such as mobile TV, on-line chatting, digital library, videoconference etc. [1][2] become more and more popular in human being's life. Generally, a multimedia communication system enables the generation, management, communication and consuming of multimedia data such as texts, images, audios, videos, animations, etc. Multimedia communication systems can be classified into various types. For example, according to the content, they can be classified into web systems, audio systems, audio-visual systems, etc. According to the communication infrastructure, they can be classified into broadcasting systems, multicasting systems, peer-to-peer (P2P) systems, etc. According to the service mode, they can be classified into living systems, content-on-demand systems, file downloading systems, etc.

Security protection is an important issue for multimedia information systems [2], which aims to protect the multimedia content, service interaction and user privacy, etc. For example, the content related to commercial secret needs to be protected against unauthorized users, the payment interactions between the user and the seller are sensitive to the third party, and the user profiles are private and should not be published. Until now, various techniques and tools have been proposed for multimedia information system security. Most of them focus on multimedia content security, secure interaction, and on privacy protection.

It is really difficult to give a thorough review on multimedia information system security because of two reasons. First, multimedia information system's diversity leads to the complexity and diversity of security issues and the corresponding protection means. Secondly, various new multimedia systems arise timely, which bring new security issues and solutions. Thus, few works have been done to review the state-of-art of multimedia information system security, and the existing works focused only on a certain kind of multimedia system. For example, in [3][4], Thuraisingham described security and privacy issues for multimedia database management systems including access control for multimedia database management systems, security policies and security architectures for such systems, and privacy problems resulted from multimedia data mining. However, only the secure database management system is focused, while some other systems, the key techniques and their advances are not introduced.

This paper aims to give a thorough review to multimedia information system security. It introduces a general architecture of multimedia information system, investigates some security issues in multimedia information systems, reviews the latest security solutions (their research advances and open issues), and presents some hot topics in multimedia information system security. It is expected to provide the latest research progress, advise some research directions, and give a list of references about multimedia system security to researchers.

The rest of the paper is organized as follows. In Section 2, the general architecture of multimedia information system is introduced, and the security issues are presented in Section 3. In Section 4, the latest technical solutions are reviewed in detail, including Digital Rights Management (DRM), confidentiality protection, ownership protection, traitor tracing, secure multimedia distribution based on watermarking, forgery detection, copy detection, privacy-preserving data mining, secure user interface, and intrusion detection and prevention. Then, in Section 5 some hot topics are proposed including trusted computing, steganography, the security in network or service convergence, security of content sharing in social networks, privacy-preserving data processing, multimedia forensics and intelligent surveillance. Finally, in Section 6, some conclusions are drawn.

2 General architecture of multimedia information system

Now, there are various multimedia information systems [2][3]. A general architecture of multimedia information system (Figure 1) is composed of several parts, i.e., multimedia content generation, content storage, content distribution, content receiving, and content consuming, editing and sharing. The architecture includes most of the steps from content generation, communication to consuming.

Multimedia content generation: It denotes the process to generate multimedia content produces (TV program, film, music, flash, web, etc.). Generally, various devices are used in this process, such as digital camera, Digital Video, audio recorder, etc. For some payable services, the multimedia content is usually generated by professional producers, while, for user generated content sharing there is no limitation to producers.

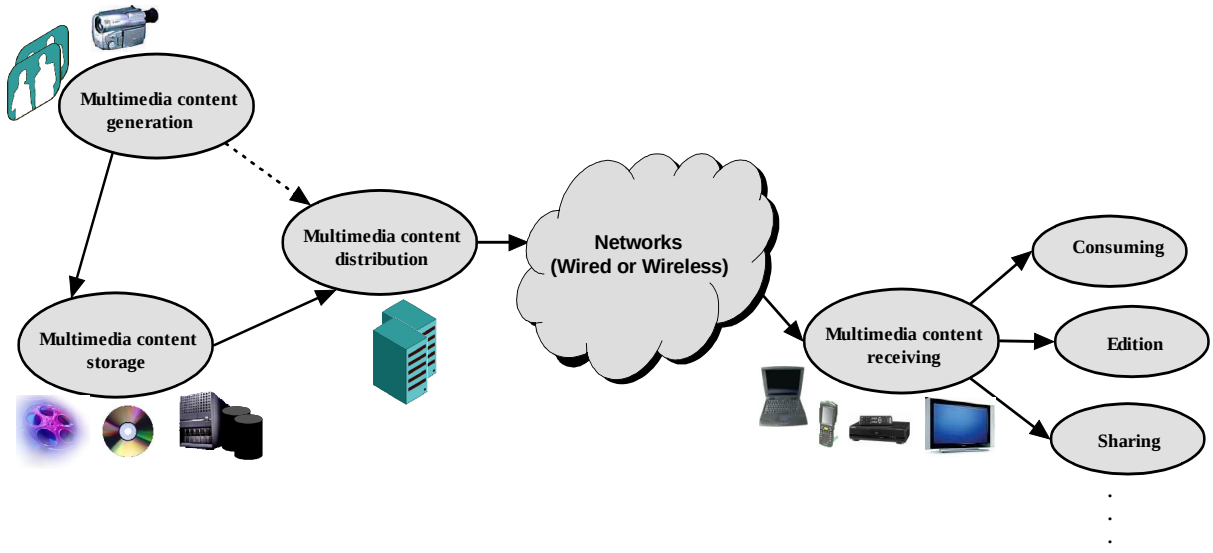


Figure 1: General architecture of multimedia information system

Multimedia storage and management: Multimedia content is often stored or backed up before being transmitted. For example, the film is recorded in the cinefilm, the music is burned into the Compact Disc, and the videos or web data are buffered in the computer servers. Generally, to make the data access or management more convenient, multimedia contents and their index information are ordered and stored in

the database, while some search or data mining techniques can be used. Additionally, to save the storage cost, multimedia content is often compressed and then stored.

Multimedia distribution: Multimedia services such as mobile TV, Internet TV, Online Music, and Content Sharing in Social Networks become more and more popular in human being's life. Among them, multimedia distribution acts as the key technique, which transmits the content from one user to others. There are two kinds of multimedia content to be transmitted, i.e., real-time content and stored content. The former one denotes the generated content without delayed storage, such as live TV or telephone call. The latter one denotes the content stored in the database, such as the video clips for video-on-demand, music segments, web data, etc.

Distribution networks: In multimedia distribution, the network is the core component. Now, various networks have been developed and popularly used. According to the physical communication channel used, they can be classified into wired networks and wireless networks. According to the communication modes, they can be classified into unicasting, broadcasting, and multicasting networks. According to the logical structure, they can be classified into client-server based networks and distributed networks (peer-to-peer networks, sensor networks, etc.). According to the application environments, they can be classified into Internet, GSM/GPRS, Satellite broadcasting, etc.

Multimedia receiving: The client receives multimedia content with terminals such as PC, TV set, mobile phone, telephone, etc. Depending on the service, certain terminal may be used. For example, PC is used for Internet access, mobile phone for short message sending, TV set for TV program watching, and telephone for calling and chatting. However, the functional difference between the terminals becomes more and more fuzzy. For example, PC is also used for chatting, and mobile phone for TV consuming and Internet access.

Multimedia consuming, editing and sharing: After receiving multimedia content, the user consumes it by decoding and playing it with terminals. He can request and browse the Web over the PC, watch TV programs over the TV set, and listening music through his mobile phone. According to the rights assigned to the user, the multimedia content may be downloaded, edited or even shared with others. For example, the user uses a PC to download a video clip from Internet, shorten it by cutting some frames and send it to his friends by peer-to-peer sharing networks. In another example, the user uses his cell phone to download a song, consume it, and then send it to his friends through Multimedia Message Sending. These operations depend on the terminal's functionalities and the content's properties

3 Security issues

In multimedia communication, security issues [3], which are generated from the transmitted information's sensitivities, should be considered. For example, the information may be related to military forbiddance, commercial secret or personal privacy. Only some authorized users can access this kind of information, and any action aiming to make the information released is regarded as the attack. With respect to the complexity of the information system, there are various threats. Some of them are described below.

3.1 Eavesdropping

Eavesdropping is the act of surreptitiously listening to a private conversation. This is commonly thought to be unethical. Eavesdropping can be done over telephone lines (wiretapping), email, instant messaging, and other methods of communication considered private. Wiretapping, also named telephone tapping, is the monitoring of telephone and Internet conversations by a third party, often by covert means. The wire tap received its name because, historically, the monitoring connection was applied to the wires of the telephone line being monitored and drew off or tapped a small amount of the electrical signal carrying the conversation. Now, eavesdropping is extended to the attack that steals the information from any network or device. Generally, it can be classified into two types: passive eavesdropping and active eavesdropping. The former one attempts only to observe the flow and gain knowledge of the information it contains, while the latter one attempts to alter the data or otherwise affect the flow of data.

3.2 Intrusion

Intrusion [5] denotes unwanted attempts at accessing, manipulating, and/or disabling of computer systems, mainly through a network such as the Internet. These attempts may take the form of attacks, as examples, by crackers, malware and/or disgruntled employees. Generally, the intrusion behavior can be classified into several types, i.e., network attacks against vulnerable services, data driven attacks on applications, host based attacks such as privilege escalation, unauthorized logins and access to sensitive files, and malware (viruses, trojan horses, and worms). Among them, the first one makes use of the apparent weakness of multimedia service systems to make them out of work, the second one adopts the iterated data requests and interactions to paralyze the applications, the third one uses the host role (with more rights than the authorized role) to steal information, and the fourth one denotes the executable program that can access, steal or tamper the software or hardware data.

3.3 Forgery

Forgery is the process of making, adapting, or imitating objects, with the intent to deceive. A forgery

is essentially concerned with a produced or altered object (multimedia content, user information, etc.). For example, in the 18th century, Europeans were curious about what North America looked like and were ready to pay to see illustrations depicting this faraway place. Some of these artists produced prints depicting North America, despite many having never left Europe. Recently, in some photo competitions, some presented photos were found non-initial produces and having been tampered with editions. Today, more and more famous drawings are imitated in order to be sold with a high price. Additionally, some attackers attempt to personate the authorized users to enjoy multimedia services.

3.4 Piracy

Copyright infringement (or piracy) is the unauthorized use of material that is covered by copyright law, in a manner that violates one of the copyright owner's exclusive rights, such as the right to reproduce or perform the copyrighted work, or to make derivative works. Especially for electronic and audio-visual media, unauthorized reproduction and distribution is occasionally referred to as piracy. Generally piracy behavior can be classified into two types, i.e., unauthorized access and unauthorized distribution. The former one denotes the unauthorized users access the multimedia content, while the latter one means that the users redistribute the accessed multimedia content to other unauthorized users. For example, the unlawful downloading of copyrighted material and sharing of recorded music over the Internet in the form of MP3 and other audio files is more prominent now than since before the advent of the Internet or the invention of MP3, even after the demise of Napster and a series of infringement suits brought by the American recording industry. Additionally, promotional screener DVDs distributed by movie studios (often for consideration for awards) are a common source of unauthorized copying when movies are still in theatrical release. Movies are also still copied by someone sneaking a camcorder into a movie theater and secretly taping the projection, although such copies are often of lesser quality than copied versions of the officially released film.

3.5 Privacy

Privacy is sometimes related to anonymity, the wish to remain unnoticed or unidentified in the public realm. When something is private to a person, it usually means there is something within them that is considered inherently special or personally sensitive. The degree to which private information is exposed therefore depends on how the public will receive this information, which differs between places and over time. Privacy can be seen as an aspect of security - one in which trade-offs between the interests of one group and another can become particularly clear. Almost all countries have laws which in some way limit privacy. An example of this would be law concerning taxation, which normally requires the sharing of information about personal income or earnings. In multimedia information systems, some personal information is private, such as user login information, subscribe information, user profile, and interaction records. Additionally, in some social networks, such as User Generated Content sharing networks, users can produce or post some multimedia content that is shared with other users. The user generated contents may be also private.

4 Technical solutions

To solve the security issues in multimedia information systems, some proposed technical solutions can realize confidentiality protection, ownership protection, traitor tracing, forgery detection, media source identification and copy detection, etc. The typical techniques include Digital Rights Management (DRM), multimedia encryption, digital watermarking, digital fingerprinting, multimedia forensics, privacy-preserving data mining, secure user interface, intrusion detection, etc. Hereafter, we describe these techniques.

4.1 Digital Rights Management (DRM)

The role of DRM [6] in distribution of content is to enable business models whereby the consumption and use of content is controlled. As such, DRM extends beyond the physical delivery of content into managing the content lifecycle. When a user buys content, he may agree to certain constraints, e.g., by choosing between a free preview version or a full version at cost, or he may agree to pay a monthly fee. DRM allows this choice to be translated into permissions and constraints, which are then enforced when the user accesses the content.

4.1.1 Scopes of DRM

Generally, a DRM system takes into consideration the following aspects: data to be protected, communication scenarios to be protected, access rights to be protected, the supported business model, and core techniques.

- **Data to be protected:** Image, video, audio, game, software, text, etc.
- **Communication scenarios to be protected:** Fixed-line services (telephone, visual telephone, etc.), Mobile/wireless services (broadcasting, Short Message Sending, Mobile TV, etc.), and Internet - based services (video-on-demand, P2P, music downloading, etc.).
- **Access rights to be protected:** access control (access to the content), usage restriction (play, playback, preview, copy, etc.), Inter-terminal copy (transfer to another terminal) and physical copy (copy the physical support), etc.
- **The supported business model:** Free, Subscription, pay per view, pay per time, etc.

- **Core techniques:** Right Expression Language, Identification and Authentication, Encryption/Decryption, Copy Protection, Access Control, etc.

4.1.2 General architecture

The general architecture of a DRM system is shown in Figure 2. It is composed of 4 main components: *Content Encoder, Content Server, Rights Issuer* and *User*. Content Encoder secures the content, Content Server distributes the content, Rights Issuer handles the license generation and assignment, and User reads the content. The typical DRM system works this way:
(1) The Content Encoder encrypts the multimedia content with the key and packages the content and key in certain format;
(2) The User requests the service from Rights Issuer and Content Server;
(3) The Content Server sends the encrypted content to the User through the manners, e.g., Internet, DVD/CDROM, email, Instant Message, Peer-to-Peer, etc.
(4) The Rights Issuer charges the service fees and distributes the rights (containing the key) to the User.
(5) The User decrypts and decodes the content with the key and consumes the content with the corresponding rights.

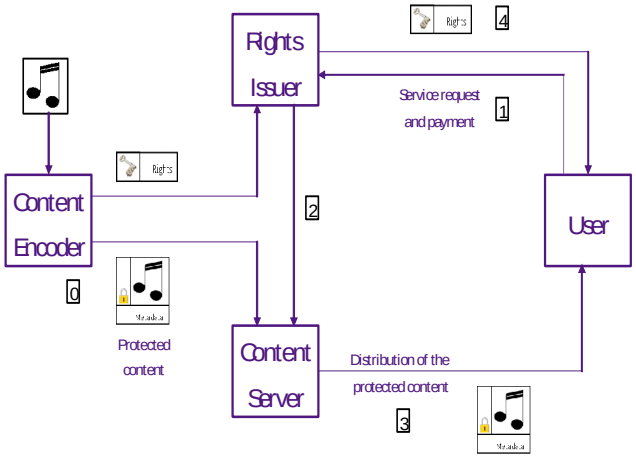


Figure 2: General architecture of a DRM system

4.1.3 Typical DRM systems

Previously, there have been various DRM systems for the corresponding services and networks. For example, the ISMACryp (Internet Stream Media Alliance Cryp) [7] is used for Internet streaming, Open Mobile Alliance DRM (OMA DRM) [8] for GSM/GPRS network, DVB-H Content Protection and Copy Management (DVB-CPCM) [9] for DVB-H network, and, High-bandwidth Digital Content Protection System (HDCP) [10], Certified Output Protection Protocol (COPP) [11] and Digital Transmission Content Protection (DTCP) [12] is used for home networks or devices. Henceforth, these DRM systems are briefly described.

ISMACryp ISMACryp [7] provides the digital rights management means for Internet Streaming Media. It defines the encryption and authentication for MPEG-4 [13] data streams. The ISMACryp specification does not mandate a cipher but AES in Counter Mode, the default encryption and authentication transform in the specification, is the de facto cipher [7] used by ISMACryp implementations. ISMACryp is different from traditional secure protocols (SRTP and IPsec). Whereas ISMACryp encrypts MPEG-4 access units (that are in the RTP payload), SRTP encrypts the whole RTP payload, and IPsec encrypts packets at network level. Thus, ISMACryp implements full end-to-end protection. Till now, there are two versions for ISMACryp, i.e., ISMA Encryption and Authentication Version 1.1 (ISMACryp 1.1) and ISMA Encryption and Authentication Version 2.0 (ISMACryp 2.0). ISMACryp 1.1 only addresses MPEG-4 codecs, such as H.264/AVC and AAC. Differently, ISMACryp 2.0 is compatible with any codec that can be stored in files of the MP4 family. This includes H.263 video and AMR audio, as widely used in mobile networks. It is compatible with any IP-network and can operate on any device, including mobile handsets.

OMA DRM The scope of OMA DRM [8] is to enable the controlled consumption of digital media objects by allowing content providers have some abilities, for example, to manage previews of DRM Content, to enable super distribution of DRM Content, and to enable transfer of content between DRM agents. The OMA DRM specifications provide mechanisms for secure authentication of trusted DRM agents, and for secure packaging and transfer of usage rights and DRM Content to trusted DRM agents. Before content is delivered, it is packaged to protect it from unauthorized access. A content issuer delivers DRM Content, and a rights issuer generates a Rights Object. The content issuer and rights issuer embody roles in the system. OMA DRM makes a logical separation of DRM Content from Rights Objects. DRM Content and Rights Objects may be requested separately or together, and they may be delivered separately or at the same time. For example, a user can select a piece of content, pay for it, and receive DRM Content and a Rights Object in the same transaction. Later, if the Rights Object expires, the user can go back and acquire a new Rights Object, without having to download the DRM Content again. The OMA DRM defines the format and the protection mechanism for DRM Content, the format (expression language) and the protection mechanism for the Rights Object, and the security model for management of encryption keys. The OMA DRM also defines how DRM Content and Rights Objects may be transported to devices using a range of transport mechanisms, including pull (HTTP Pull, OMA Download), push (WAP Push, MMS) and streaming. Any interaction between network entities, e.g. between rights issuer and content issuer, is out of scope.

DVB-H CPCM CPCM [9] is a system for Content Protection and Copy Management of commercial digital

content delivered to consumer products and networks. CPCM manages content usage from acquisition into the CPCM system until final consumption, or export from the CPCM system, in accordance with the particular usage rules of that content. Possible sources for commercial digital content include broadcast (e.g., cable, satellite, and terrestrial), Internet-based services, packaged media, and mobile services, among others. CPCM is intended for use in protecting all types of content, i.e., audio, video and associated applications and data. CPCM provides specifications to facilitate interoperability of such content after acquisition into CPCM by networked consumer devices. CPCM is only concerned with content after it has been acquired. The fundamental boundaries of control within CPCM are the local environment, and the Authorized Domain (AD). The AD is defined as a distinguishable set of DVB CPCM compliant devices, which are owned, rented or otherwise controlled by members of a single household. Content is bound to its Usage State Information (USI) which describes how it can be consumed, copied or exported relative to this Authorized Domain. This concept is fundamentally different from today's CA and DRM techniques, which normally operate on a single device basis.

HDCP High-bandwidth Digital Content Protection (HDCP) [10] is a form of Digital Rights Management (DRM) developed by Intel Corporation to control digital audio and video content as it travels across Digital Visual Interface (DVI) or High-Definition Multimedia Interface (HDMI) connections. HDCP's main target is to prevent transmission of non-encrypted high definition content. Generally, three methods are adopted to meet the goal. Firstly, the authentication process disallows non-licensed devices to receive HD content. Secondly, the encryption of the actual data sent over DVI or HDMI interface prevents eavesdropping of information. It also prevents "man in the middle" attacks. Thirdly, key revocation procedures ensure that devices manufactured by any vendors who violate the license agreement could be relatively easily blocked from receiving HD data. HD DVD, Blu-ray Disc and DVD players (with HDMI or DVI connector) use HDCP to establish an encrypted digital connection. If the display device or in the case of using a PC to decrypt and play back HD-DVD or Blu-ray media, the graphics card (hardware, drivers and playback software) does not support HDCP, then a connection cannot be established. As a result, a black picture and/or error message will likely be displayed instead of the video content.

COPP COPP [11] is a device driver technology used to enable High-bandwidth Digital Content Protection (HDCP) during the transmission of digital video between applications and high-definition displays. COPP is a Microsoft security technology for video systems that require a logo certification. For security drivers are authenticated and protected from tampering to prevent unauthorized high-quality recording from the video outputs. COPP control signals are also encrypted. Certified Output Protection Protocol (COPP) enables an application to protect a video stream as it travels from the graphics adapter to the display device. An application can use COPP to discover what kind of physical connector is attached to the display device, and what types of output protection are available. Protection mechanisms include HDCP, Copy Generation Management System - Analog (CGMS-A) and Analog Copy Protection (ACP). COPP defines a protocol that is used to establish a secure communications channel with the graphics driver. It uses Message Authentication Codes (MACs) to verify the integrity of the COPP commands that are passed between the application and the display driver. COPP does not define anything about the digital rights policies that might apply to digital media content. Also, COPP itself does not implement any output protection systems. The COPP protocol simply provides a way to set and query protection levels on the graphics adapter, using the protection systems provided by the adapter.

DTCP DTCP [12] is a DRM technology that aims to restrict "digital home" technologies including DVD players and televisions by encrypting inter-connections between devices. In theory this allows the content to be distributed through other devices such as personal computers or portable media players, if they also implement the DTCP standards. DTCP is one link in an end-to-end solution using licensing terms and conditions to enforce copy protection. It is based on well-known cryptographic algorithms and techniques and suitable for implementation on PCs and Consumer Electronics devices. The content is encrypted by DTCP source devices prior to output, and the device outputs are limited to "compliant" devices. The renewability capabilities enhance long-term integrity of system through device revocation. This system's implementation needs the supports from IT industry, CE industry, content distributors, content industries and conditional access providers.

4.1.4 Hot topics and open issues in DRM

Although various DRM systems have been proposed recently, there are still some open issues that constitute hot research topics.

DRM for P2P system P2P networks differ from traditional server-client networks. In P2P networks, the content is transmitted directly from one peer to another without going through a server, and the content is used by an indeterminate number of users. Thus, for the P2P DRM system, the distribution of user rights is a challenge. Some works [14] attempts to get the tradeoff between security and content sharing efficiency by introducing the super peer that is responsible of rights issuing and user management. However, some other issues are still pending: deployment of rights management functions, preventing the distribution of copyright infringing content, preventing the registration of illegal content, and taking into account secondary distribution in managing.

Domain-based DRM. The concept of domain is firstly introduced in DVB-H CPCM, which denotes the set of DVB CPCM compliant devices. In fact, it is suitable for many scenarios with local content sharing, such as home

networks [15], campus networks, family networks, and social networks. Thus, the content or rights can be shared with the users during the domain, and even inter-domains. To realize these functionalities, some works need to be done, including the domain's modeling, interaction protocols, and business models.

Interoperable DRM. Most of the existing DRM systems focus on certain services or networks, and there are some conflictions between different DRM systems. From another perspective, popular ubiquitous multimedia services often request users or devices access different services or networks. Therefore, it is imperative to implement the interoperability between different DRM systems. The Digital Media Project (DMP) [16] proposes the interoperable DRM and defines the architecture and components that are compatible with existing DRM systems. However, the diversity of services, networks and DRM functionalities makes it a challenge.

4.2 Confidentiality protection

Multimedia encryption is the key technique to protect the content confidentiality, which transforms multimedia content into an unintelligible form. Generally, the content is encrypted with a cipher controlled by the key that aims to resist the eavesdropping attack.

4.2.1 Performance requirements of multimedia encryption

Due to multimedia content's properties (high redundancy, large volumes, real time interactions, and packaged into certain format) multimedia encryption algorithms often have some specific requirements, i.e., security, efficiency, compression ratio, format compliance and supporting direct operations.

- Security is the basic requirement of multimedia content encryption. Different from text/binary encryption, multimedia encryption requires both cryptographic security and perception security. The former one refers to the security against cryptographic attacks [17], and the latter one means that the encrypted multimedia content is unintelligible to human perception [18].
- Since real-time transmission or access is often required by multimedia applications, multimedia encryption algorithms should be efficient so that they don't delay the transmission or access operations.
- Multimedia encryption algorithms should not change compression ratio or at least keep the changes in a small range. This is especially important in wireless or mobile applications, in which the channel bandwidth is limited.
- Multimedia data are often encoded or compressed before transmission, which produces the data streams with some format information, such as file header, time tamp, file tail, etc. Encrypting the data except the format information will keep the encrypted data stream format-compliant.
- In some applications, it will save some cost to operate directly on the encrypted multimedia data. For example, the encrypted multimedia data can be recompressed, the bit-rate of the encrypted multimedia data can be controlled, the image block or frame can be cut, copied or inserted, etc.

4.2.2 Typical multimedia encryption algorithms

According to these requirements, partial encryption is often focused, which encrypts only parts of multimedia content, while leaving the other parts unchanged. According to the type of multimedia data, the existing partial encryption algorithms can be classified into audio encryption, image encryption and video encryption.

Audio encryption. Audio data are often encoded before being transmitted in order to save transmission bandwidth. Thus, audio encryption is often applied to the encoded data. For example, encrypting only the parameters of Fast Fourier Transformation (FFT) during speech encoding process [19] encrypts the speech data. In decryption, the right parameters are used to recover the encrypted data. For MP3 music, only the sensitive parameters of MP3 stream are encrypted, such as the bit allocation information [20]. For encrypting only few data, this kind of encryption algorithm is often of high efficiency.

Image encryption. A straightforward partial encryption algorithm for images is bit-plane encryption [21]. That is, in an image, only several significant bit-planes are encrypted, while the other bit-planes are left unchanged. By reducing the encrypted bit-planes, the encryption efficiency can be improved. However, the security cannot be confirmed, especially against replacement attacks. In replacement attacks, the encrypted data part is replaced by other data, which may make the encrypted image intelligible. For compressed images, the algorithms based on DCT and wavelet codecs attract more researchers, such as JPEG or JPEG2000. The algorithm proposed in [22] encrypts only some significant bit-planes of DCT coefficients, which obtains high perception security and encryption efficiency. The algorithm proposed in [23] encrypts only the significant streams in the encoded data stream, which is selected according to the progressiveness in space or frequency. Generally, no more than 20% of the data stream is encrypted, which obtains high efficiency. Another algorithm [24] encrypts different number of significant bit-planes of wavelet coefficients in different frequency bands, which obtains high security in human perception and keeps secure against replacement attacks.

Video encryption. Since video data are of larger volumes compared with image or audio data, video data are often compressed in order to reduce the bandwidth. Generally, the compressed video data stream is composed of such parts as format information, texture information and motion

information. Thus, according to the data parts, video partial encryption algorithms are classified into several types: format information encryption, frame encryption, texture encryption, and both motion vector and texture encryption.

Format information encryption. Since format information helps the decoder to recover the multimedia data, encrypting the format information will make the decoder out of work [25]. However, it is not secure in cryptographic viewpoint to encrypt only format information. This is because the format information is often in certain grammar, which can be broken by statistical attacks. These algorithms change the format information, and thus the encrypted multimedia data cannot be displayed or browsed by a normal browser.

Frame encryption. In such video codec as MPEG1/2/4, the frame is often classified into three types: I-frame, P-frame and B-frame. I-frame is often encoded directly with DCT transformation, while P/B-frame is often encoded by referencing to adjacent I/P-frame. Thus, I-frame is the referenced frame of P/B-frame. Intuitively, encrypting only I-frame will make P/B-frame unintelligible. However, experiments [25] show that this is not secure enough. The reason is that some macroblocks encoded with DCT transformation in P/B-frame are left unencrypted. For some videos with smart motion, the number of such macroblock is high enough to make the encrypted video intelligible. As an improved method, SECMPEG algorithm [26] encrypts all the macroblocks encoded with DCT transformation in I/P/B-frame. In these algorithms, the motion information is left unencrypted, which makes the motion track still intelligible.

Texture encryption. The coefficients in DCT or wavelet transformation determine the intelligibility of the multimedia data. Encrypting the coefficients can protect the confidentiality of the texture information. For example, the algorithm proposed in [27] encrypts the signs of DCT coefficients. In AVC codec, the intra-prediction mode of each block is permuted with the control of the key [28], which makes the video data degraded greatly. These algorithms are efficient in computing, but not secure enough. Firstly, the motion information is left unencrypted, which makes the motion track still intelligible. Secondly, the video can be recovered in some extent by replacement attacks [18].

Both motion vector and texture encryption. To keep secure, it is necessary to encrypt both DCT/wavelet coefficient and motion vector. Considering that these two kinds of information occupy many percents in the whole video stream, they should be encrypted partially or selectively. Generally, two kinds of partial encryption method are often used, e.g., coefficient permutation and sign encryption. For example, the algorithm [29] permutes coefficients or encrypts the signs of coefficients and motion vectors in DCT or wavelet transformation, and the algorithm [30] encrypts the DCT coefficients and motion vectors in AVC codec with sign encryption. These algorithms encrypt both the texture information and motion information, and thus, obtain high security in human perception. Additionally, the partial encryption operation, such as coefficient permutation or sign encryption, is often of low cost, which makes the encryption schemes of high efficiency. Furthermore, these partial encryption algorithms keep the file format unchanged.

4.2.3 Open issues in multimedia encryption

During the past decades, many algorithms have been reported for multimedia content encryption and they satisfy various applications. However, there are still some open issues.

Security analysis Due to the difference between multimedia encryption and traditional encryption, the security analysis methods may also be different. Taking partial encryption for example, not only such cryptographic attacks but also some ciphertext-only attacks (aims to recover the content intelligibility) should be considered, such as replacement attacks [31] and other unknown attacks. To the best of our knowledge, it is still difficult to get a suitable metric on the intelligibility of multimedia content, which increases the difficulty to analyze a partial encryption algorithm.

Communication-compliant encryption In practice, it is difficult to avoid transmission error or delay in multimedia communication. Thus, making the encrypted data robust to transmission error is necessary. Till now, few solutions have been reported. The segment encryption [30] is a potential solution. In this algorithm, the data stream is partitioned into segments, and then encrypted segment by segment, with each segment independently encrypted. However, it is still difficult to determine the size of the segment because segment size often contradicts the security.

Format independence or format compliance In order to keep format compliance, the encryption algorithm often varies with the compression codec. In some cases, format independence is more attractive. Taking Digital Rights Management (DRM) for example, it is required that all the content encoded with various codecs should be protected equally. Thus, there is a contradiction between format compliance and format independence.

4.3 Ownership protection

Watermarking technique [32] is used to protect multimedia content's ownership, which embeds the ownership information (e.g., the producer's name or ID) into multimedia content by modifying the content slightly. Later, the ownership information can be extracted and used for authentication. Generally, invisible watermarking that embeds the ownership information imperceptibly is often used for ownership protection, and it is also investigated here.

4.3.1 Performance metrics of watermarking

A good watermarking algorithm satisfies some performance metrics such as imperceptibility, robustness, capacity, security, oblivious detection, etc.

- *Imperceptibility* means that the watermarked content has no perceptual difference with the original one. It is also named 'transparency' or 'fidelity' and makes sure that the watermarked copy is still of high quality and commercial value.
- *Robustness* refers to the ability for the watermark to survive such operations including general signal processing operations (filtering, noising, A/D, D/A, re-sampling, recompression, etc.) and intentional attacks (rotation, scaling, shifting, transformation, tampering, etc.).
- *Capacity* denotes the maximal data volumes that can be embedded in multimedia content. Considering of transparency, the capacity of each approach is not infinite, which is in relation with the carrier content.
- The *security* against various attacks should be considered when constructing a watermarking algorithm. Generally, some encryption operations are introduced to watermarking algorithms in order to keep secure.
- *Oblivious detection* means that the detection process needs not the original copy. It is also named blind detection. On the contrary, non-blind detection means that the original copy is required by the detection process.

4.3.2 Existing watermarking algorithms

During the past decades, many watermarking algorithms have been reported, which can be classified by different methods. According to the carrier media type, they can be classified into image watermarking, video watermarking, audio watermarking, text watermarking and software watermarking. Among them, image watermarking embeds information in images based on images' redundancy and the perceptual property of human's eyes. For example, the algorithm proposed in [33] embeds information in perceptual imperceptible bits. Video watermarking makes use of temporal information besides spatial information compared with image watermarking. For example, the algorithm [34] embeds information in I-frames, which is similar to image watermarking, while the one [35] embeds information in motion vectors. Audio watermarking adopts audio redundancy and human psychoacoustic model to hide information. For example, the algorithm [36] uses the echo property to hide information. Text watermarking [37] hides information by slightly adjusting such textures as vertical distance, horizontal distance or font. This kind of watermarking is only suitable for texts but not for other data. Software watermarking [38] is used to protect software copyright and it is often classified into two categories: static watermarking and dynamic watermarking. Static watermarking embeds certain sentences in the source code, but does not change the software's functionality. Dynamic watermarking designs some dynamic information in the software, such as dynamic data structure or dynamic implementation tracking.

According to the embedding method, watermarking algorithms can be classified into three categories: replacement-based watermarking, modulation-based watermarking and encoding-based watermarking. Replacement-based watermarking replaces some parts of the cover work with watermarking. For example, the LSB method [39] replaces the least significant bits with the transmitted message directly. This kind of embedding method obtains high capacity, but it is seldom robust to attacks. Modulation-based watermarking [32] modulates the cover work with the message, and can realize either blind detection or non-blind detection. It is often robust to some attacks. Compared with the above two methods, encoding-based watermarking hides information by encoding some parts of the cover work. For example, Patchwork method [40] encodes watermarking into the relation between block pairs, the authentication watermarking [41] encodes watermarking into the relation between pixels. There are also some other algorithms such as histogram-based algorithm [42] and salient-point algorithm [43]. Although they are robust to some attacks, these encoding-based watermarking methods should be carefully designed based on the cover work's properties, and the capacity is often limited.

According to the embedding domain, watermarking can be embedded in either temporal domain, spatial domain or frequency domain. Taking video watermarking for example, the watermark can be embedded in the frame-pixels, the motion vectors or the DCT coefficients, which obtains different performances. Spatial domain watermarking embeds information in pixels directly, such as the LSB method [39] and the perceptual model based methods [44][45]. Generally, these methods are often not robust to signal processing or attack, although they are efficient in computing. Frequency domain Watermarking is embedded in transformation domain, such as DCT transformation [41], wavelet transformation [46], etc. Compared with the watermarking in spatial domain, the one in frequency domain obtains some extra properties in robustness and imperceptibility. Additionally, the embedding can be done during compression, which is compatible with international data compression standard. Temporal domain Watermarking is embedded in temporal information. For example, in audios, echo property is used to hide information, which is named echo hiding [36]. In videos, the temporal sequence is partitioned into static component and motive component, with information embedded into motive component [47]. Considering that human's eyes are more sensitive to static component than to motive one, embedding in motive component can often obtain higher robustness. However, error accumulation or floating makes the watermarked videos blurred in some extent, which should be improved by error compensation.

4.3.3 Open issues and hot topics in watermarking research

Watermarking is still not widely used in practical services because of some unsolved issues. Additionally, there are some interested topics need to be investigated.

Security of watermarking algorithms. It is still not confirmed whether watermarking algorithm should be kept secret against attackers. If so, it is different from cryptography system, and needs special security evaluation metrics. Otherwise, most of the existing watermarking algorithms are vulnerable to attacks.

Robustness to mixed operations. Some algorithms may be robust to certain operations, while they are still difficult to resist mixed operations, such as camera capture that consists of noise, rotation, transformation, scaling, etc.

Watermarking special carriers. For different carriers, the performance requirement is also different. For example, high-definition data permit little degradation, which needs the watermarking algorithm with high fidelity or transparency. Similarly, for such carrier as digital map or database, the watermarking should exit in everywhere of the carrier (each region in the map or each recorder in the database). These special requirements encourage the research of novel watermarking algorithms.

Web-based applications. Based on its potential applications in identification and authentication, watermarking may be applied in web-based applications such as access control, web monitoring, web-based identification or web based data linking, etc. Compared with traditional methods, watermarking does not make up extra storage, and is sometime imperceptible.

4.4 Traitor tracing

Multimedia content distribution often faces such a problem, i.e., a customer may redistribute the received content to other unauthorized customers. The customer who redistributes the content is called the traitor. This typical problem often causes great profit-losses of content provider or service provider. As a potential solution, digital fingerprinting [48] is recently reported and studied. It embeds different information, such as Customer ID, into multimedia content, produces a unique copy, and sends the copy to the corresponding customer. If a copy is spread to unauthorized customers, the unique information in the copy can be detected and used to trace the illegal redistributor.

4.4.1 Collusion attacks

The most serious threat to watermarking-based fingerprinting is collusion attack. That is to say, several attackers fabricate a new copy through combining their unique copies in order to avoid the tracing. Attackers intend to remove the embedded fingerprinting by making use of the slight difference between different copies. This kind of attack is often classified into two categories: linear collusion and nonlinear collusion. Among them, linear collusion means to average, filter or cut-and-paste the copies, while nonlinear collusion means to take the minimal, maximal or median pixels in the copies. Generally, five kinds of collusion attacks are considered, i.e., averaging attack, linear combinatorial collusion attack (LCCA) [49], min-max attack, negative-correlation attack and zero-correlation attack.

4.4.2 Existing digital fingerprinting algorithms

Since the past decade, finding new solutions resisting collusion attacks has been attracting more and more researchers. The existing fingerprinting algorithms can be classified into three categories, i.e., orthogonal fingerprint, coded fingerprint and warping-based fingerprint.

In orthogonal fingerprinting [50], the unique information (also named fingerprint) to be embedded is the vector independent from each other. For example, the fingerprint can be a pseudorandom sequence, and different fingerprint corresponds to different pseudorandom sequence. The orthogonal fingerprint can resist most of the proposed collusion attacks, which benefits from the orthogonal property of the fingerprints. According to the property of orthogonal sequence, such detection method as correlation detection is still practical although there is some degradation caused by collusion attacks. For example, the algorithm [50] produces orthogonal fingerprinting for each customer, the fingerprinting is then modulated by the cover video, and correlation detection is used to determine the ownership or colluders from the copies. For each copy, correlation detection obtains a big correlation value that determines the customer who receives the copy. For the colluded copy (e.g., averaging between N copies) the correlation value becomes R/N, which is smaller than the original correlation value R. Thus, if the correlation value R/N is still no smaller than the threshold T, the fingerprint can still be detected, otherwise, it cannot. In fact, the correlation value decreases with the rise of colluders. That is because the fingerprint is cross-affected by each other. In order to improve the detection efficiency, some detection methods are proposed, such as recursive detection (tree-based or correlation based) [51].

Fingerprinting can be carefully designed in codeword form, named coded fingerprinting [52][53], which can detect the colluders partially or completely. Till now, two kinds of encoding methods are often referenced: the Boneh-Shaw scheme [52] and the combinatorial design based code [53]. Boneh-Shaw scheme is based on the Marking Assumption, i.e., only the different bits are changed by colluders, while the same bits can not be changed. By designing the primitive binary code, at least one colluder can be captured out of up to c colluders. And it can support more customers if it is extended to outer code. Differently, in combinatorial design based anti-collusion scheme, the fingerprint acts as a combinatorial codeword. The combinatorial codes have the following property: each group of colluders' fingerprint produces unique codeword that determines

all the colluders in the group. The codeword is constructed based on combinatorial theory, such as AND-ACC (anti-collusion codes) or BIBD [53]. Compared with orthogonal fingerprinting, the coded fingerprinting has some advantages. Firstly, the embedding method is not only limited to additive embedding, some other existing embedding methods are also usable. Secondly, the correct detection rate does not depend on the number of colluders. However, with respect to LCCA attacks, the coded fingerprinting is not so robust. That is because the linear operation may remove the fingerprint information and make the fingerprint bit undetectable. In desynchronized fingerprinting [54], the multimedia content (e.g., image or video) is desynchronized imperceptibly with some geometric operations in order to make each copy different from others. This kind of fingerprinting aims to make collusion impractical under the condition of imperceptibility. That is, to de-synchronize the carrier. Thus, the colluded copy is perceptible (generates perceptual artifacts). These de-synchronization operations include random temporal sampling (video frame interpolation, temporal re-sampling, etc.), random spatial sampling (RST operations, random bending, luminance filtering or parameter smoothing) or random warping. In the warping-based fingerprinting, the original video copy is warped under the control of customer ID, which produces different copies with slight degradation. In collusion attacks, the colluded copy is degraded so greatly that it can not be used in high definitional applications. Additionally, the more the colluders, the more the degradation. According to this case, warping-based fingerprinting makes collusion attacks unpractical, and thus is secure against collusion attacks. However, in this scheme, the compression ratio is often changed because of the pre-warping operations. Additionally, it is a challenge to support large number of customers by warping the content imperceptibly.

4.4.3 Open issues and hot topics in digital fingerprinting

It is worth mentioning that the digital fingerprinting based traitor tracing is still a new topic, and there are some open issues.

Collusion-resistance and the supported users. The existing digital fingerprinting algorithms can not get the good tradeoff between collusion-resistance and the supported users. Some new fingerprinting algorithms with high robustness to collusion attacks and good performance in efficiency can be expected. Foexample, some fingerprinting codes based on random sequences [55][56] are reported having good performances although they are still not used in multimedia content.

Efficient fingerprinting embedding In secure multimedia content distribution based on digital fingerprinting, where (at sender side, in the mediate or at receiver side) to do the embedding operation is related to the system's efficiency [57]. Additionally, for different networks, e.g., unicasting, broadcasting, multicasting and P2P, different fingerprint embedding scheme will be considered.

Fingerprinting and DRM The fingerprinting based traitor tracing scheme can be combined with existing Digital Rights Management (DRM) systems in order improve the traceability. Where and how to introduce fingerprinting operations are still open issues.

4.5 Secure multimedia distribution based on watermarking

Considering that fingerprinting technology produces different media copy to different customer, it is easily implemented in unicast network. In contrast, it is difficultly implemented in broadcast or multicast network. The key points to be confirmed are the security and the efficiency. Till now, some distribution schemes based on digital fingerprinting have been proposed. These schemes can be classified into three types as shown in Figure 3.

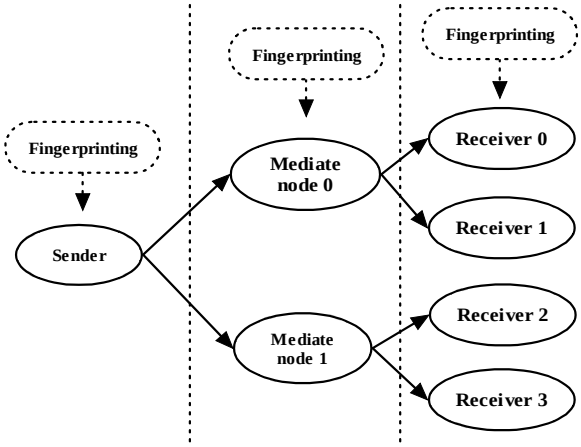


Figure 3: Watermarking - based multimedia distribution schemes

The first one [58] embeds the fingerprint and encrypts the fingerprinted media at the server side, and decrypts the media content at the customer side. In this scheme, for different customer, the media data should be fingerprinted differently, which increases the server's loading, and is not suitable for the applications with large number of customers. The second one [59] embeds the fingerprint and encrypts the fingerprinted media by the relay node, and decrypts the media at the customer side. This scheme reduces the server's loading greatly. However, the fingerprinting or encryption operation in relay node makes the network protocol not compliant with the original one. The third one [60] encrypts the media data at the server side, and decrypts the media and embeds the fingerprint at the customer side. This scheme reduces the server's loading greatly. However, for the decryption and fingerprinting operations are implemented at the customer side, the means to confirm the security is the key problem. To decrypt the media and embed the fingerprint independently is not secure, because the decrypted media data may be leaked out from the gap between the decryption operation and fingerprinting operation, as shown in Figure 4.

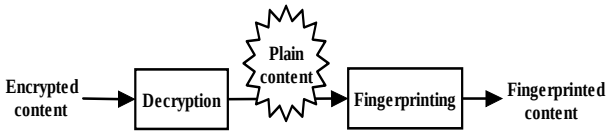


Figure 4: Leakage of multimedia content

4.5.1 Embedding Fingerprint at Sender Side

Straightforwardly, multimedia content is firstly fingerprinted, then encrypted, and finally distributed by the sender. These schemes [58] obtain high security since they forbid customers fingerprinting multimedia content. However, they need to transmit different copy to different customer, which cost much time or transmission bandwidth. To solve this problem, some improved schemes have been reported, such as broadcasting encryption and partial fingerprinting.

Broadcasting Encryption. In the broadcasting encryption based method [61], media data are partitioned into segments, each segment is watermarked into two copies, and all the segments are encrypted and distributed. At the receiver side, a key is used to select one segment from the couple segments, and different key selects different segments that produce different media copy. In this scheme, traditional ciphers can be used, which keeps the system's security. However, the key disadvantage is that double volumes need to be transmitted.

Partial Fingerprinting. In the partial fingerprinting method [62], media data are partitioned into two parts, e.g., encryption part and fingerprinting part. Among them, the former one is encrypted and broadcasted to different customers, while the latter one is fingerprinted and unicasted to each customer. Since only a small part of multimedia content is unicasted, the time cost or bandwidth cost can be reduced greatly. The difficulty is to make the two kinds of communication modes work together simultaneously.

4.5.2 Joint Fingerprint Embedding and Decryption (JFD)

To obtain a tradeoff between security and efficiency, some schemes [63][64][65][66] are proposed to joint fingerprint embedding and decryption (JFD). In these schemes, the fingerprint is embedded into media content during decryption process, which produces the fingerprinted media copy directly, thus avoids the leakage of plain media content and improves the security of embedding fingerprint at the customer side.

Chamleon Method. The Chamleon method [63] firstly encrypts the media data at the server side, then distributes the media data, and finally, decrypts the data by modifying the least significant bits under the control of different decryption key. Here, the encryption and decryption processes use different key tables, respectively. It was reported that the scheme is time efficient and secure against cryptographic attacks. However, for different customers, different key tables should be transmitted, which cost bandwidth. Additionally, the least significant bits are not robust to signal processing, such as recompression, additive noise, filtering, etc.

Kundur's Method. The JFD scheme proposed in [64] firstly encrypts the media data partially at the server side, then distributes the data, and finally, decrypts the data by recovering the encrypted parts selectively. The position of the unexplored parts determines the uniqueness of a media copy. Here, the DCT coefficients' signs are encrypted. The scheme is robust to some operations including slight noise, recompression and filtering, while the imperceptibility can not be confirmed, the encrypted media content is not secure in perception and the security against collusion attacks cannot be confirmed.

Lian's Method. The scheme proposed in [65] encrypts media data at the server side by encrypting the variable-length code's index, and decrypts media data at the customer side by recovering code's index with both decryption and fingerprinting. The scheme is secure against cryptographic attacks, while the robustness against some operations including recompression, filtering and adding noise, cannot be confirmed.

Lemma's Method. The scheme proposed in [66] encrypts media data at the server side by partial encryption, and decrypts media data at the customer side with a new key stream. The scheme is robust against signal processing, which benefits from the adopted watermarking algorithms, while the security against cryptographic attacks cannot be confirmed. Additionally, the transmission of key stream costs much time and space.

4.5.3 Open issues in this topic

Clearly, the watermarking-based multimedia distribution is still a new topic, and there are some open issues, which must be considered.

Security of partial fingerprinting. For the schemes embedding fingerprinting at sender side, the multiple content copies need to be transmitted from the sender to the receiver. Partial fingerprinting can reduce the repeated transmission costs. However, the security of partial fingerprinting scheme needs to be investigated.

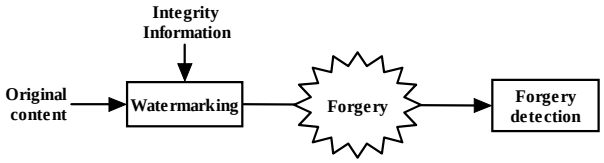
Interference between fingerprinting and encryption. In JFD schemes, the suitable fingerprinting operation and encryption operation should be considered in order to reduce the interference between them. For example, the homomorphic encryption and fingerprinting operations are potential.

Key distribution. In these schemes, different customers use different keys to decrypt the content. Some means are required to realize secure and efficient key distribution or exchange. This depends on the application environment, such as unicasting network, multicasting network, broadcasting network or P2P network, etc. Additionally, some other research directions are also attractive such as efficient broadcasting encryption, partial encryption based content distribution, combined encryption and watermarking, and so on.

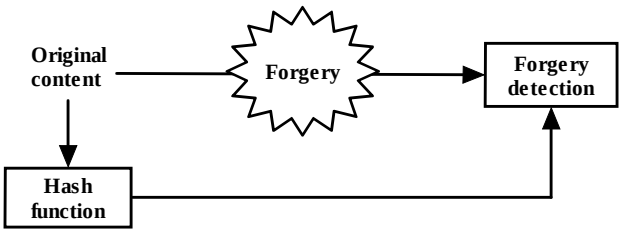
4.6 Forgery detection

4.6.1 General forgery detection schemes

Forgery detection aims to tell whether the multimedia content is authentic without forgery operations. Till now, several means have been proposed to detect forgeries, i.e., watermarking based scheme [32], perceptual hash based scheme [67], and multimedia forensic based scheme [68]. In the first one, as shown in Figure 5(a), the watermark information (e.g., integrity flag or ownership) is embedded into multimedia content imperceptibly. This embedding operation is realized during multimedia content generation (e.g., in the camera). In forgery detection, the embedded information is extracted from the operated multimedia content and compared with the original information. The comparison result tells whether the image is forged or not and even localizes the tampered regions. There are two apparent disadvantages for this scheme. Firstly, the embedding operation is needed during media generation, which is often unavailable in practical applications. Secondly, the information embedding operation degrades multimedia content's quality, which is not permitted in some applications.



(5a) Watermarking-based forgery detection scheme



(5b) Hash-based forgery detection scheme

(5c) Multimedia forensic-based forgery detection scheme

Figure 5: Various forgery detection schemes

In the second one, as shown in Figure 5(b), the perceptual hash function is applied to multimedia content, which generates the hash value composed of a certain-length string. The hash value is stored by the authenticator. In forgery detection, a new hash value is computed from the operated multimedia content, and compared with the stored one. The comparison result tells whether the content is forged or not. Similar with the watermarking based scheme, the hash based scheme realizes the hash computing operation during multimedia content generation, e.g., in the camera. The difference is that the hash value does not change the multimedia content. Similarly, the disadvantage is that the hash computing operation implemented during media generation is not always available in practical applications.

In the third one, as shown in Figure 5(c), the intrinsic features are extracted from the operated multimedia content, then, the features' properties are analyzed and compared with a common threshold, and the comparison result tells whether the content is forged or not. The extracted intrinsic features have apparent difference between the natural one and the forged one, and the difference can be distinguished by the threshold. The core technique in this scheme is to extract the distinguishable features. Generally, it depends on the model of the forgery operation, and different features will be extracted for different forgery operations. Different from the former two schemes, the forensic based scheme does not need the original media content, and neither changes the quality of media content.

4.6.2 Existing forgery detection methods based on multimedia forensics

Under the condition without pre-processing, only forensic methods can be used to detect the forgery. These forensic methods extract the features that can distinguish the original media and operated one. Generally, there are some forgery detection methods based on special features, i.e., correlation feature, double compression feature, light feature, and media statistical feature, and some methods with special functionalities, e.g., duplication detection.

Correlation based detection. Some correlations between adjacent temporal or spatial sample pixels are often introduced during multimedia content generation or content operations. From the multimedia content, these correlations can be detected and used to identify forgeries, e.g., resample [69] and color filter array interpolation [70]. These forgery detection methods can detect either the image's authentic or the image regions' authentic. However, they work with the assumption that the images are firstly interpolated during image generation and then modified during image forgery. Thus, they are not robust again re-interpolation attacks.

Double compression detection. In multimedia content forgery, the edition software often stores the content in compression formats, e.g. JPEG or MPEG2. Thus, the forged multimedia content may be recompressed. Intuitively, recompression introduces different distortions compared with once compression, which can be used to detect whether the multimedia content is recompressed [71]. Thus, if an image is recompressed, the probability of forgery is increased. The method's disadvantage is the vulnerability to attacks. For example, if the modified JPEG image is cropped before being saved in JPEG format, the periodicity of distortion is difficult to be detected.

Light property based detection In practice, the captured picture conforms to certain light direction. Suppose there is only a point light corresponding to the picture's scene, then the estimated light directions for all the objects in the picture should intersect to a point. Differently, if one object in the picture is tampered, the estimated light direction of the object will be inconsistent with other objects' light directions. Thus, by detecting the light directions, the image forgery can be detected [72]. This method can detect either the image's authentic or the image objects' authentic. Its computational complexity and robustness depend on the adopted light direction estimation methods.

Feature-based detection. With respect to the different generation processes of a natural image and the forged image, there are some specific image features, which can exhibit the difference such as the high-order statistical feature, the sharpness/blurriness, and feature fusion and classifier fusion. These features can be used to detect forgeries [73]. This method can detect whether a media region is forged or not. The challenge is how to design the optimal strategy for selecting the features or classifiers.

Duplication detection In multimedia forgery, duplication is one of the often used tampering operations. Till now, there are various duplication detection methods, i.e., direct detection and segmentation based detection. Direct detection means to detect the duplicated regions directly without any information about the regions, e.g., DCT domain sorting [74]. Segmentation based detection means to detect the duplicated objects after segmentation [75]. These methods can detect an image object's authentic, including the object deletion, healing or duplication.

4.6.3 Open issues in forensic based forgery detection

According to the above investigation, multimedia forensic-based forgery detection is still at the beginning, and there are some open issues.

Detection accuracy For the existing forgery detection methods, the correct detection accuracy is still not good enough for practical applications. One important reason is the natural media's diversity. There are so many media sources and media contents that make it difficult to design a fixed classifier or decision threshold.

Counter attacks Sometimes, it is easy to detect a single forgery operation. However, in practice, there are often more than one forgery operations on the same media content, which may reduce the detection accuracy because of the interference between different operations. Additionally, some high-level attackers may make the forgery by considering of the means to counter the forgery detection methods. In the existing detection methods, the attacks are seldom considered.

Video Forgery With the advancement of video edition software, video forgery becomes more and more popular. But video forgery detection is seldom considered. Compared with image forgery, video forgery has an additional dimension, i.e., the temporal space. Thus, some new detection methods focusing on video forgery are expected.

Test bed. Up till now, no public tests are done for forgery detection, which delays the steps to practical applications. It is caused by the shortage of a general test bed. The test bed will contain a multimedia content database and be capable of evaluating various performances, including detection accuracy, robustness and security.

4.7 Copy detection

Recently, the concept of content-based copy detection (CBCD) [76] has been proposed as an alternative means of identifying illegal media copies. Given an image registered by the owner, the system can determine whether near-replicas of the image are available on the Internet or through an unauthorized third party. If it is found that an image is registered (i.e., it belongs to a content owner), but the user does not have the right to use it, the image will be deemed an illegal copy. The suspect image is then sent to the content owner for further identification and a decision about taking legal action against the user. Image copy detector searches for all copies of a query image, and is different from content-based image retrieval (CBIR) [77] that searches for similar images. Thus, it is not usually feasible to apply existing CBIR techniques to CBCD because they may cause a considerable number of false alarms. For CBCD, the key challenge is to extract the suitable features that can obtain a good tradeoff between discriminability and robustness. The discriminability denotes the ability to distinguish different media contents. The robustness refers to the ability to survive such operations as cropping, noising, contrast changing, zoom, insertion, etc. Generally, the extracted features are compared with the registered ones, whose distance tells the repetition.

4.7.1 Existing copy detection algorithms

According to the methods that extract features, the CBCD algorithms can be classified into two types: global feature-based algorithms, and local feature -based algorithms. For example, the algorithm in [78] extracts the global features from wavelet transformed coefficients and colour space, the one in [79] extracts the ordinal measure of DCT coefficients from the whole image, the one in [80] uses elliptical track division strategy to extract features from all the elliptical track blocks, and the one in [81] uses a sliding window to extract the block's relationship with its neighbouring blocks. These global feature-based algorithms often obtain good discriminability, while bad robustness. For example, they are not robust to such operations as block cropping. Differently, local feature-based algorithms have better robustness. For example, the algorithm in [82] computes many descriptors for each image, in which, each descriptor corresponds to one image block, and the algorithm in [83] extracts the key points from

each image part. They can still identify the content even when it is tampered (e.g., cropped or modified) greatly. Their disadvantage is the high computational complexity, and the research challenge is how to determine the block size.

4.7.2 Open issues

The CBCD research is at the beginning, and there are some open issues. Firstly, the operations to be resisted need to be defined, and thus, the various CBCD algorithms can be evaluated. Secondly, the tradeoff between discriminability and robustness needs to be investigated. No existing algorithms can obtain good level in both the performances. For example, the algorithms with local features have more computational costs but present interesting results in term of robustness, while the ones with global features are very efficient for small transformations. Thirdly, for practical applications (e.g., over Web), some means should be taken into consideration to enable the algorithm running in real-time. For example, the feature extraction and comparison operations need to be fastened, and a larger set of queries should be built to support large number of users.

4.8 Privacy-preserving data mining

Privacy preserving data mining [84] is a novel research direction in data mining and statistical databases. The main consideration in privacy preserving data mining is two fold. First, sensitive raw data, e.g., identifiers, names and addresses, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge, which can be mined from a database by using data mining algorithms should also be excluded. In privacy preserving data mining, the main objective is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process.

4.8.1 Existing privacy-preserving data mining techniques

In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. Needless to say that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include perturbation, blocking, aggregation or merging, and sampling. In the privacy preservation technique, selective modification that leaves some information unchanged is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. Until now, various techniques have been reported, which can be classified into three types: heuristic-based techniques, cryptography-based techniques and reconstruction-based techniques.

Heuristic-based techniques (e.g., adaptive modification) modify only selected values that minimize the utility loss rather than all available values. The values in data mining include association rule, classification rule, etc. For example, the association rule is confused by the centralized data perturbation [85] or by the centralized data blocking [86], and the classification rule is confused by the centralized data blocking [87]. These techniques are often efficient in implementation, while their security depends on the adopted modification methods.

Cryptography-based techniques use secure multiparty computation to realize private data mining. In secure multiparty computation, two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The key-point here is how to conduct such a computation while preserving the privacy of the inputs. Thus, a computation is secure if at the end of the computation, no party knows anything except its own input and the results. For example, the method is proposed to mine private association rules from vertically partitioned data [88], the method to mine private association rules from horizontally partitioned data [89], and the method to induct private decision tree from horizontally partitioned data [90]. These techniques can obtain enough security benefiting from secure computation. But, actually, because of the nature of this solution methodology, the data in all of the cases that this solution is adopted, is distributed among two or more sites.

Reconstruction-based techniques firstly perturb the data, and then reconstruct the data's distributions at an aggregate level in order to perform data mining. Thus, the original distribution of the data is reconstructed from the randomized data. For example, the numerical data (e.g., individual records) can be perturbed and then reconstructed by estimation [91], and the binary or categorical data (e.g., association rules) are randomized and reconstructed in [92]. For these techniques, the successful reconstruction operations determine the performance of data mining.

4.8.2 Open issues

Obviously the privacy-preserving data modification results in degradation of the database performances, such as the confidential data protection, the loss of mining functionality and the communication cost. Till now, no existing techniques outperform all the others on all the performances. For example, cryptography-based techniques have better confidential data protection, while they limit some mining functionalities and needs high cost for information exchange between different sites. As a result, better techniques are expected to obtain the good tradeoff between different performances.

4.9 Secure user interface

In multimedia information systems, various secure user interface methods [93] have been widely used, which prevent or authorize the users to access services. According to the information carriers, they can be classified into three types, i.e., possession-based method, knowledge-based method and biometrical method. Possession-based method uses the specific "token", such as a security tag or a card, to realize authorization. Knowledge-based method uses a code or password to authenticate the users. Biometrical method uses the biometrics to identify specific people by certain characteristics. Biometrics can overcome the weakness in traditional authentication systems that use tokens, passwords or both. Weakness, such as sharing passwords, losing tokens, guessable passwords, forgetting passwords and a lot more, were successfully targeted by biometric systems. Biometric characteristics can be divided in two main classes: physiological characteristics and behavioural characteristics. Among them, the former ones are related to the shape of the body, such as fingerprint, face, hand, iris and DNA, while the latter ones are related to the behavior of a person such as signature, keystroke dynamics and voice. The typical biometrical methods being widely used include fingerprint recognition, face recognition, hand geometry, iris recognition, speaker recognition, etc.

4.9.1 Biometric system

A biometric system is often composed of the following components [93]: template storage, sensor, pre-processing, feature extractor, template generator, matcher and application device. Generally, it works as following steps. First, the templates corresponding to persons are stored in a database. Then, the sensor is used to acquire all the necessary data, i.e., face photo, fingerprint photo, speech samples, etc. The acquired data are then pre-processed to remove sensor artifacts, e.g., removing background noise. Then, some features are extracted from the sensor data, and used to create a template. The obtained template is then passed to a matcher that compares it with other existing templates, and the comparison result will be output for driving the application device. Generally, a biometric system can provide two functions [94], i.e., verification and identification. The former one authenticates its users in conjunction with a smart card, username or ID number. The biometric template captured is compared with that stored against the registered user either on a smart card or database for verification. The latter one authenticates its users from the biometric characteristic alone without the use of smart cards, usernames or ID numbers. The biometric template is compared to all records within the database and a closest match score is returned.

4.9.2 Existing biometric systems

For biometric systems, their performances are often measured by two errors, i.e., False Rejection Rate (FRR) and False Acceptance Rate (FAR) [95]. FRR denotes the probability to reject the match mistakenly, while FAR denotes the one to accept the match mistakenly. Now, some biometric systems can obtain good performances. For example, the face recognition system in [96] can get the 1% FAR and 10% FRR, the fingerprint recognition system in [97] can get 1% FAR and 0.1% FRR, and the hand geometry recognition system in [98] can get 2% FAR and 0.1% FRR. Additionally, some other performances are also cared, including the ease of acquisition for measurement, the permanence against aging, the authentication speed, and ease of use of a substitute, etc. It is reported that, among the various biometrics (e.g., face, fingerprint, hand geometry, keystrokes, hand veins, iris, retinal scan, signature, voice, facial thermograph, odor, DNA, gait and ear canal), fingerprint, hand geometry, hand veins, iris, facial thermograph and ear canal can obtain the better tradeoff in various performances.

4.9.3 Open issues

A biometric system cannot always give provable results because of biometrics' complex properties such as variability. A solution is multi-mode authentication [94] that means either to combine several biometrics or to combine various authentication methods (i.e., possession-based method, knowledge-based method and biometrical method). Additionally, cancellable biometrics is also a hot topic, which is a way to inherit the protection and the replacement features of biometric data more seriously [99]. It is very essential for protecting biometrics in storage or in processing state.

4.10 Intrusion detection and prevention

In multimedia information system, intrusion detection [100] is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. The system performing automated intrusion detection is called an Intrusion Detection System (IDS). An IDS can be either host-based, if it monitors system calls or logs, or network-based if it monitors the flow of network packets. Modern IDSs are usually a combination of these two approaches. When a probable intrusion is discovered by an IDS, typical actions to perform would be logging relevant information to a file or database, or generating an email alert. These automatic actions can be implemented through the interaction of Intrusion Detection Systems and access control systems such as firewalls. If intrusion detection takes a preventive measure without direct human intervention, then it becomes an intrusion-prevention system (IPS).When an attack is detected, it can drop the offending packets while still allowing all other traffic to pass. Generally, it is a network security device that monitors network and system activities for malicious or unwanted behavior and can react, in real-time, to block or prevent those activities. For example, a host-based IPS (HIPS) [101] is one where the intrusion-prevention

application is resident on that specific IP address, usually on a single computer. Differently, Network-based IPS (NIPS) [102] will operate in-line to monitor all network traffic for malicious code or attacks. Now, there are exist three kinds of NIPS, i.e., Content-Based IPS (CBIPS) that inspects the content of network packets for unique sequences, detects and prevents known types of attack such as worm infections and hacks, Protocol Analysis based IPS that natively decodes application-layer network protocols and evaluates different parts of the protocol for anomalous behavior or exploits, or Rate-Based IPS (RBIPS) that monitors and learns normal network behaviors and intends to prevent Denial of Service attacks.

The intrusion detection or prevention technology is immature and dynamic. For example, the accuracy and adequacy of IDS signatures cannot be determined [103]. The proprietary nature of the signatures for most commercial intrusion detection systems makes a detailed discussion of their accuracy and adequacy difficult. This may be adequate for very simple attacks, but are probably inadequate for sophisticated, multi-stage attacks. Additionally, it is necessary to identify unknown modes of attack continuously. Generally, intrusion detection systems can match patterns of behavior that represent signatures of known attacks, while difficult to recognize new attack strategies. The adaptive approaches are expected to solve this problem. Furthermore, intrusion detection systems could provide evidence to support prosecution in court but do not. With the rapidly growing theft and unauthorized destruction of computer-based information, the frequency of prosecution is rising, and it is urgent to use computer forensics to analyze the evidence provided by intrusion detection or prevention systems.

4.11 Performance comparison of different technical solutions

Actually, different technical solutions solve different security issues. The ten solutions mentioned above and their targeted security issues are listed in Table 1. As it can be seen, no solution can solve all the security issues. Thus, in practice, more than one solution is used together. For example, to provide a secure video-on-demand service over Internet, both Digital Rights Management technique and intrusion detection and prevention technique are used. Whether to or how to compound them together depends on the corresponding multimedia service system and its performance requirements.

Table 1: Targeted security issues of different technical solutions

Technical solutions	Targeted security issues				
	Eavesdropping	*Intrusion*	*Forgery*	*Piracy*	*Privacy*
Digital Rights Management	√			√	√
Confidentiality protection	√				
Ownership protection				√	
Traitor tracing				√	
Secure multimedia distribution based on watermarking	√			√	
Forgery detection			√		
Copy detection				√	
Privacy-preserving data mining					√
Secure user interface		√	√		
Intrusion detection and prevention		√			

5 Open issues and hot topics

In the previous section, we reconsidered some typical solutions for multimedia information system security. However, there are some other solutions, which focus on special applications. As various multimedia information systems are emerging, some new solutions are expected. Hereafter, we describe some of the interesting trends.

Trusted computing

Trusted Computing (TC) [104] is recently proposed to construct a fully trusted system. It aims to solve the security issues caused by software-only means, which enforces the trusted behavior by loading the hardware

with a unique ID and key. With Trusted Computing, the computer will consistently behave in specific ways, and those behaviors will be enforced by hardware and software. Generally, it will have the following components, i.e., endorsement key, secure input and output, protected execution, sealed storage, and remote attestation. Although it may cause consumers to lose anonymity in online interactions, it is regarded as a possible enabler for future versions of mandatory access control, copy protection, and digital rights management.

Steganography

Steganography [105] is used as a covert communication method, which hides secret information into multimedia content and thus sends it to receivers imperceptibly. Only the receiver partnered with the sender can extract the secret information from multimedia carrier. Different from watermarking, the third party can only detect whether the multimedia content is suspicious or not (i.e., steganalysis), and then decide whether to remove it. Better steganography algorithms are expected to resist the latest steganalysis methods.

Security in network or service convergence

Ubiquitous multimedia services are becoming more and more popular, and often converge several networks or services together. The challenge includes not only the exchange of network protocols, the bit-rate adaptation of multimedia content and the compliance of user terminals but also the security architecture covering all the involved networks. Interoperable DRM is not the only solution. The recent work reported in [106] shows that some potential application scenarios need to be investigated.

Security of content sharing in social networks

Nowadays, content sharing social networks enrich human being's life. Some typical networks include Blog, Video Blog, P2P sharing platforms, etc., where users can upload or post multimedia content freely. However, it is noted that, more and more unhealthy contents arise in these networks, e.g., the content related to legality, sex, privacy, piracy or terror. To detect, distinguish or prevent these contents' distribution is a new topic [107]. Some content analysis and classification techniques need to be used together with existing security solutions.

Privacy-preserving data processing

Privacy-preserving data mining addresses the necessary to protect privacy in data retrieval. However, it is now also urgent in other fields, such as multi-party interaction, remote diagnosis, content distribution [108], etc. Thus, the new protocols or operations need to be investigated, which aims to the new applications. The new technique, named signal processing in encryption domain [109], attempts to give a general solution by adopting homomorphic encryption and signal processing operations. It is still at the beginning.

Multimedia forensics

In Section 4, we review the forensicbased forgery detection techniques. However, multimedia forensics includes more valuable techniques, e.g., media source distinguish technique and device identification technique [110]. The former one denotes the technique to distinguish the devices (camera, scanner, cell phone, computer, etc.) that generate the multimedia content by investigating the content's properties. Differently, the latter one not only distinguishes the device type, but also identifies the device itself. These techniques may be useful to support prosecution in court.

Intelligent surveillance

Surveillance is now widely used in public security. With the increase of distributed surveillance cameras and collected data volumes, intelligent surveillance [111] becomes more and more urgent, as it processes the multimedia data automatically to extract usable information. The typical intelligent processing techniques include object tracking, activity analysis, crime detection, face extraction, etc. Generally, various basic techniques are required, such as video segmentation, semantic analysis, machine learning, etc.

6 Conclusions

In this paper, we introduced a general architecture of multimedia information system and addressed some important security issues. We reviewed some typical technical solutions, and proposed some hot research topics. The paper is expected to provide valuable directions to researchers working in multimedia information system security. Due to the diversity of multimedia information systems, the security issues and solutions are various and it is difficult to be included in one paper. Therefore, in this paper we considered only some important security issues such as: eavesdropping, intrusion, forgery, piracy and privacy, and reviewed only some typical solutions such as Digital Rights Management (DRM), confidentiality protection, ownership protection, traitor tracing, secure multimedia distribution based on watermarking, forgery detection, copy detection, privacy-preserving data mining, secure user interface, and intrusion detection and prevention. As more and more multimedia information systems arise, new security issues will be generated. We are inclined to the fact that the research community will solve emerging security issues by proposing novel approaches. For this reason, in the near future we will update this survey by adding the new issues and solutions.

7 Acknowledgement

The work was partially supported by Crypto project through the grant code of ILAB-PEK08-006

8 References

[1] M. C. Angelides and S. Dustdar. *Multimedia Information Systems* (The Springer International Series in Engineering and Computer Science), by Publisher: Springer, June 30, 1997.

[2] S. M. Rahman. *Design and Management of Multimedia Information Systems: Opportunities and Challenges*, Publisher: IGI Global, April 16, 2001.

[3] B. Thuraisingham. Security and privacy for multimedia database management systems, *Multimedia Tools and Applications*, 33(1): 13-29, April 2007.

[4] B. Thuraisingham. *Multimedia systems security*. Proceedings of the 9th *ACM workshop on Multimedia & Security*, Dallas, Texas, USA, Pages: 1-2, 2007.

[5] *Intrusion detection system*. http://en.wikipedia.org/wiki/Intrusion_detection_system.

[6] *Digital Rights Management*. http://en.wikipedia.org/wiki/Digital_rights_management.

[7] ISMACryp 2.0 (ISMA Encryption & Authentication Specification 2.0). http://www.isma.tv/.

[8] Open Mobile Alliance, Digital Rights Management 2.0 (OMA DRM 2.0), 03 Mar 2006.

[9] *Digital Video Broadcasting Content Protection & Copy Management* (DVB-CPCM), DVB Document A094 Rev. 1, July 2007.

[10] *HDCP (High-bandwidth Digital Content Protection System),* http://en.wikipedia.org/wiki/HDCP.

[11] *COPP (Certified Output Protection Protocol),* http://msdn2.microsoft.com/en-us/library/Aa468617.aspx

[12] *DTCP (Digital Transmission Content Protection),* http://en.wikipedia.org/wiki/DTCP

[13] W. Li. Overview of fine granularity scalability in MPEG-4 video standard, *IEEE Transactions on Circuits and Systems for Video Technology*, 11(3): 301-317, 2001.

[14] J. Y. Sung, J. Y. Jeong, and K. S. Yoon, "DRM Enabled P2P Architecture," 2006 *International Conference on Advanced Communication Technology (ICACT2006)*, pp. 487-490, 2006.

[15] J.-P. Andreaux, A. Durand, T. Furon, and E. Diehl, "Copy Protection System for Digital Home Networks," *IEEE Signal Processing Magazine*, March 2004, pp.100-108.

[16] *DMP - Digital Media Project* (http://www.dmpf.org/)

[17] R. A. Mollin, *An Introduction to Cryptography*. CRC Press. 2006.

[18] S. Lian, J. Sun, G. Liu, and Z. Wang. Efficient video encryption scheme based on advanced video coding, *Multimedia Tools and Applications*, Springer, 38(1): 75-89, 2008.

[19] S. Sridharan, E. Dawson, and B. Goldburg. Fast Fourier transform based speech encryption system. *IEE Proceedings of Communications, Speech and Vision*, 138(3): 215-223, 1991.

[20] L. Gang, A. N. Akansu, M. Ramkumar, X. Xie. Online Music Protection and MP3 Compression. In Proc. Of *Int. Symposium on Intelligent Multimedia, Video and Speech Processing*, May 2001, pp.13-16.

[21] M. Podesser, H. P. Schmidt, and A. Uhl, "Selective bitplane encryption for secure transmission of image data in mobile environments," In CD-ROM Proceedings of the *5th IEEE Nordic Signal Processing Symposium* (NORSIG 2002), Tromso-Trondheim, Norway, October 2002.

[22] R. Pfarrhofer and A. Uhl. Selective image encryption using JBIG. In *Proceeding of 2005 IFIP Conference on Communications and Multimedia Security*, pp. 98-107, 2005.

[23] R. Norcen and A. Uhl, "Selective encryption of the JPEG2000 bitstream," IFIP International Federation for Information Processing, *LNCS 2828*, pp. 194-204, 2003.

[24] S. Lian, J. Sun, D. Zhang, and Z. Wang. A selective image encryption scheme based on JPEG2000 Codec. *2004 Pacific-Rim Conference on Multimedia (PCM2004)*, Springer *LNCS, 3332*, 65-72, 2004.

[25] I. Agi and L. Gong. An empirical study of MPEG video transmissions. In *Proceedings of the Internet Society Symposium on Network and Distributed System Security*. San Diego, CA, Feb. 1996, pp. 137-144.

[26] L. Tang. Methods for encrypting and decrypting MPEG video data efficiently. In Proceedings of the Fourth *ACM International Multimedia* Conference (ACM Multimedia'96). Boston, MA, November 1996, pp. 219-230.

[27] C. Shi, and B. Bhargava. A fast MPEG video encryption algorithm. In Proceedings of the *6th ACM International Multimedia Conference*. Bristol: UK, September, 1998, pp. 81-88.

[28] J. Ahn, H. Shim, B. Jeon, and I. Choi. Digital Video Scrambling Method Using Intra Prediction Mode. 2004 Pacific-Rim Conference on Multimedia (PCM2004), Springer, *LNCS Vol.3333*, pp.386-393, November 2004.

[29] W. Zeng and S. Lei. Efficient frequency domain selective scrambling of digital video. *IEEE Trans on Multimedia*, 5(1): 118 –129, March 2003.

[30] S. Lian, Z. Liu, Z. Ren, and H. Wang. Secure advanced video coding based on selective encryption algorithms. *IEEE Transactions on Consumer Electronics*, 52(2): 621-629, 2006.

[31] S. Lian. *Multimedia Content Encryption: Techniques and Applications*. Auerbach Publication, Taylor & Francis Group, 2008.

[32] I. J. Cox, M. L. Miller and J. A. Bloom, *Digital Watermarking*, Morgan-Kaufmann, San Francisco, 2002.

[33] M. Wu, E. Tang, and B. Liu, Data hiding in digital binary images, In Proc. *IEEE Int'l Conf. on Multimedia and Expo*, Jul 31-Aug 2, 2000, New York, NY, 393-396.

[34] S. Bounkong, B. Toch, D. Saad, and D. Lowe, ICA for watermarking digital images, *Journal of Machine Learning Research*, 4(7-8): 1471-1498, 2004.

[35] Y. Bodo, N. Laurent, and J. Dugelay, Watermarking video, hierarchical embedding in motion vectors, *IEEE International Conference on Image Processing*, Spain, 14-17 Sept. 2003, vol. 2, pp. 739-742.

[36] D. Gruhl, A. Lu, and W. Bender, *Echo Hiding, Pre-Proceedings: Information Hiding*, Cambridge, UK, 1996, pp. 295-316.

[37] N. F. Maxemchuk, and S. H. Low. Performance comparison of two text marking methods, *IEEE Journal on Selected Areas in Communications*, 16(4): 561-572, May 1998.

[38] C. S. Collberg, and C. Thomborson, "Software watermarking: models and dynamic embeddings," In Proc. *ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL99),* San Antonio, Texas, 1999, pp. 311-324.

[39] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, A digital watermark, *Proc. of the IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 86–90, Austin, Texas, Nov. 1994.

[40] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding, *IBM systems Journal*, 35(3-4): 313-316, 1996.

[41] D. Ye, Y. Mao, Y. Dai, and Z. Wang. A multi-feature based invertible authentication watermarking for JPEG Images, Proceedings of the *3rd International Workshop on Digital Watermarking (IWDW2004)*, Seoul, Korea, Oct. 2004, pp. 152-162.

[42] D. Coltue, and P. Bolon, "Watermarking by histogram specification," In Proc. SPIE Electronic Imaging'99, *Security and Watermarking of Multimedia Contents*, San Jose, 1999, pp. 252-263.

[43] P. M. Rongen, M. B. Macs, and C. Overveld, Digital image watermarking by salient point modification, In Proc. *SPIE Electronic Imaging'99, Security and Watermarking of Multimedia Content*, San Jose, 1999, vol. 3657, pp. 273-282.

[44] C. I. Podilchuk, and W. Zeng, "Image-adaptive watermarking using visual models," *IEEE Journal of Selected Areas in Communication*, 16(4):525-539, 1998.

[45] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney. Robust audio watermarking using perceptual masking. *Signal Processing*, 66(3): 337-355, 1998.

[46] M. J. Tsai, K. Y. Yu, Y. Z. Chen. Joint wavelet and spatial transformation for digital watermarking. *IEEE Trans. on Consumer Electronics*, 2000, 46(1): 241~245.

[47] H. Joumaa, F. Davoine, "An ICA based algorithm for video watermarking," In *Proc. 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005*), Vol. 2, pp. 805-808.

[48] M. Wu, W. Trappe, Z. J. Wang, and R. Liu, Collusion-resistant fingerprinting for multimedia. *IEEE Signal Processing Magazine*, March 2004, 21(2): 15-27.

[49] Y. Wu, "Linear Combination Collusion Attack and its Application on an Anti-Collusion Fingerprinting," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005* (ICASSP '05). March 18-23, 2005, Vol. 2, pp. 13-16.

[50] A. Herrigel, J. Oruanaidh, H. Petersen, S. Pereira, and T. Pun, "Secure copyright protection techniques for digital images," In *Second Information Hiding Workshop* (IHW), LNCS 1525, Springer-Verlag, 1998, pp. 169-190.

[51] Z. J. Wang, M. Wu, W. Trappe, and K. J. R. Liu, "Group-oriented fingerprinting for multimedia forensics," *EURASIP Journal on Applied Signal Processing*, 2004(4) 2153-2173, 2004.

[52] D. Boneh, and J. Shaw, Collusion-secure fingerprinting for digital data, *IEEE Trans. Inform. Theory*, 44 (5): 1897-1905, Sept. 1998.

[53] W. Kim, and Y. Suh, "Short N-secure fingerprinting code for image," *2004 International Conference on Image Processing*, 2004, pp.2167-2170.

[54] Y. Mao, and M. K. Mihcak, "Collusion-resistant international de-synchronization for digital video fingerprinting," *IEEE Conference on Image Processing*, 2005, vol. 1, pp. 237-240.

[55] N. Hayashi, M. Kuribayashi, and M. Morii, Collusion-resistant fingerprinting scheme based on the CDMA-technique, Second International Workshop on Security (IWSEC 2007), LNCS 4752, pp. 28–43, 2007.

[56] G. Tardos, Optimal Probabilistic Fingerprint Coding, in: Proceedings of the *35th Annual ACM Symposium on Theory of Computing*, 2003, pp. 116– 125.

[57] S. Lian. Traitor tracing in mobile multimedia communication. In "*Handbook of Research on Mobile Multimedia*" (2nd edition), edited by Ismail Khalil Ibrahim. IGI Global, 2008.

[58] D. Simitopoulos, N. Zissis, P. Georgiadis, V. Emmanouilidis and M. G. Strintzis. Encryption and watermarking for the secure distribution of copyrighted MPEG video on DVD, *ACM*

Multimedia Systems Journal, *Special Issue on Multimedia Security*, 9(3), 217-227, Sep. 2003.

[59] I. Brown, C. Perkins, and J. Crowcroft. Watercasting: Distributed watermarking of multicast media. In Proceedings *of International Workshop on Networked Group Communication*, Springer-Verlag LNCS, 1736, pp. 286–300, 1999.

[60] J. Bloom, "Security and rights management in digital cinema," in Proc. *IEEE Int. Conf. Acoustic, Speech and Signal Processing*, Vol. 4, pp. 712-715, 2003.

[61] R. Parnes and R. Parviainen, "Large scale distributed watermarking of multicast media through encryption," in Proc. IFIP Int. Conf. Communications and Multimedia Security Issues of the New Century, pp. 149-158, 2001.

[62] H. V. Zhao, K. J. R. Liu, Fingerprint multicast in secure video streaming. *IEEE Transactions on Image Processing*, 15(1), 12-29, 2006.

[63] R. Anderson and C. Manifavas. Chamleon – a new kind of stream cipher. *LNCS*, *Fast Software Encryption*, Springer-Verlag, pp. 107-113, 1997.

[64] D. Kundur and K. Karthik, "Video fingerprinting and encryption principles for digital rights management," *Proceedings of the IEEE*, 92(6): 918-932, 2004.

[65] S. Lian, Z. Liu, Z. Ren, and H. Wang, "Secure Distribution Scheme for Compressed Data Streams," 2006 *IEEE Conference on Image Processing* (ICIP 2006), Oct 2006, pp. 1953-1956.

[66] A. N. Lemma, S. Katzenbeisser, M. U. Celik, and M. V. Veen. Secure watermark embedding through partial encryption. *Proceedings of International Workshop on Digital Watermarking (IWDW 2006)*, Springer LNCS, 4283, 433-445, 2006.

[67] C.-Y. Lin and S.-F. Chang. A robust image authentication algorithm surviving JPEG lossy compression. In *SPIE Storage and Retrieval of Image/Video Databases*, Vol. 3312, pp. 296–307 (1998).

[68] H. T. Sencar and N. Memon, Overview of state-of-the-art in digital image forensics, Book chapter, Part of Indian Statistical Institute Platinum Jubilee Monograph series titled *'Statistical Science and Interdisciplinary Research*,' World Scientific Press (2008).

[69] A. C. Popescu and H. Farid. Exposing digital forgeries by detecting traces of re-sampling, *IEEE Trans. Signal Processing*, 53(2): 758-767 (2005).

[70] A. C. Popescu and H. Farid. Exposing digital forgeries in color filter array interpolated images, *IEEE Trans. Signal Processing*, 53(10): 3948-3959 (2005).

[71] W. Wang, and H. Farid. Exposing digital forgeries in video by detecting double MPEG compression, *Proceedings of the 9th workshop on Multimedia & security* (*MM&Sec'06)*, September 26–27, 2006, Geneva, Switzerland (2006), pp. 35-42.

[72] M. K. Johnson and H. Farid. Exposing digital forgeries by detecting inconsistencies in lighting, Proc. of *ACM Multimedia Security Workshop* (2005), pp. 1-10.

[73] B. Sankur S. Bayram, I. Avcibas and N. Memon. Image manipulation detection. *Journal of Electronic Imaging* – 15(4), 041102 (17 pages) (2006).

[74] J. Fridrich, D. Soukal, and J. Lukas. Detection of copy-move forgery in digital images. In *Proceedings of 2003 Digital Forensic Research Workshop (2003)*, Cleveland, OH, USA, 2003, http://www.ws.binghamton.edu/fridrich/Research/copymove.pdf.

[75] H. Farid. Exposing digital forgeries in scientific images. *ACM MM&Sec'06*, September 26–27, 2006, Geneva, Switzerland, pp. 29 - 36.

[76] A. Joly and O. Buisson. Discriminant local features selection using efficient density estimation in a large database, in Proc. *ACM Int. workshop on Multimedia information retrieval*, pp. 201–208, New York, 2005.

[77] L. Amsaleg and P. Gros. Content-based retrieval using local descriptors: Problems and issues from a database perspective, *Pattern Anal. Appl.*, 4(2-3): 108–124, 2001.

[78] E. Y. Chang, C. Li, J.-Z. Wang, P. Mork, and G. Wiederhold. Searching near-replicas of images via clustering, in Proc. *SPIE: Multimedia Storage and Archiving Systems* IV, 1999, vol. 3846, pp. 281–92.

[79] C. Kim. Content-based image copy detection, *Signal Processing: Image Communication*, 18(3): 169–184, 2003.

[80] M. Wu, C. Lin, and C. Chang. Image copy detection with rotating tolerance, In CIS 2005, Part I, *LNAI 3801*, Springer, pp. 464-469, 2005.

[81] M.-N. Wu, C.-C. Lin, C. Chang. A robust content-based copy detection scheme, *Fundamenta Informaticae* 71(2-3): 351–366, IOS Press, 2006.

[82] S. A. Berrani, L. Amsaleg, and P. Gros. Robust content-based image searches for copyright protection, in Proc. *ACM Int. Workshop on Multimedia Databases*, pp. 70–77, 2003.

[83] J.-H. Hsiao, C.-S. Chen, L.-F. Chien, M.-S. Chen. A new approach to image copy detection based on extended feature sets. *IEEE Trans. Image Process*, 16(8): 2069-2079, August 2007.

[84] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis. State-of-the-art in privacy preserving data mining, *SIGMOD Record*, 33(1): 50-57, 2004.

[85] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, andE. Dasseni. Association rule hiding, *IEEE Transactions on Knowledge and Data Engineering*, 16(4): 434-447, 2004.

[86] Y. Saygin, V. Verykios, and C. Clifton, Using unknowns to prevent discovery of association rules, *SIGMOD Record,* 30(4): 45–54, 2001.

[87] L. Chang and I. S. Moskowitz, Parsimonious downgrading and decision trees applied to the inference problem, In Proceedings of the 1998 *New Security Paradigms Workshop* (1998), 82–89.

[88] J. Vaidya and . Clifton, Privacy preserving association rule mining in vertically partitioned data, In the *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), pp. 639–644.

[89] M. Kantarcioglou and C. Clifton, Privacy-preserving distributed mining of association rules on horizontally partitioned data, In Proceedings of the *ACM SIGMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery* (2002), 24–31.

[90] Y. Lindell and B. Pinkas, Privacy preserving data mining, In *Advances in Cryptology - CRYPTO 2000* (2000), 36–54.

[91] D. Agrawal and C. C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms, In Proceedings of the *20th ACM Symposium on Principles of Database Systems* (2001), 247–255.

[92] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, Privacy preserving mining of association rules, In Proceedings of the *8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining* (2002), pp. 217-228.

[93] *Biometrics*. http://en.wikipedia.org/wiki/Biometrics.

[94] A. K. Jain, A. Ross, and S. Pankanti. Biometrics: A tool for information security. *IEEE Transactions On Information Forensics and Security*, 1(2): 125- 143, June 2006.

[95] M. Krause and H. F. Tipton. "Characteristics of Biometric Systems". In Handbook of Information Security Management. *CRC Press, pp. 39-41*.

[96] P. J. Philips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone. Face recognition vendor test 2002: Overview and Summary. http://www.frvt.org/DLs/FRVT_2002_Overview_and_Summary.pdf, March 2003.

[97] C. Wilson, A. R. Hicklin, H. Korves, B. Ulery, M. Zoepfl, M. Bone, P. Grother, R. J. Micheals, S. Otto, and C. Watson, *Fingerprint vendor technology evaluation 2003: summary of results and analysis report*, NIST Internal Rep. 7123, Jun. 2004.

[98] E. Kukula, and S. Elliott, *Implementation of Hand Geometry at Purdue University's Recreational Center: An Analysis of User Perspectives and System Performance*, IEEE 2005, The 39th Annual 2005 International Carnahan Conference on Security Technology (CCST'05), 11-14 Oct. 2005, pp. 83-88.

[99] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems, *IBM systems Journal*, 40(3): 614-634, 2001.

[100] NIST SP 800-31, *Intrusion Detection Systems*.(Online) http://csrc.nist.gov/-publications/nistpubs/index.html

[101] Study by Gartner. *Host-based intrusion prevention systems (HIPS) update: Why antivirus and personal firewall technologies aren't enough.* (online) http://www.gartner.com/-teleconferences/attributes/attr_165281_115.pdf

[102] Study by Gartner. *Magic quadrant for network intrusion prevention system appliances, 1H08.(online)* http://www-935.ibm.com/-services/us/iss/pdf/esr_magic-quadrant-for-network-intrusion-prevention-system-appliances-1h08.pdf

[103] J. Allen, A. Christie, W. Fithen, J. McHugh, J. Pickel, and E. Stoner. *State of the Practice of Intrusion Detection Technologies*. TECHNICAL REPORT, CMU/SEI-99-TR-028, ESC-99-028, January 2000.

[104] R. Shane, "The Trusted Computing Platform Emerges as Industry's First Comprehensive Approach to IT Security", 2006. (online) https://www.trustedcomputinggroup.org/news/Industry_Data/IDC_448_Web.pdf.

[105] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker. *Digital Watermarking and Steganography,* 2nd Edition, The Morgan Kaufmann Series in Multimedia Information and Systems, 2007.

[106] S. Lian. Digital Rights Management for the Home TV based on scalable video coding. *IEEE Transactions on Consumer Electronics*, 54(3): 1287-1293, August 2008.

[107] *Social Networking Security. (online)* http://www.infosec.co.uk/files/KEY_22_1215_Social_Croft.pdf.

[108] S. Lian, Z. Liu, Z. Ren, and H. Wang. Commutative encryption and watermarking in compressed video data. *IEEE Circuits and Systems for Video Technology*, 17(6): 774-778, June 2007.

[109] *Signal Processing in the Encrypted Domain* (SPEED). (online) http://www.speedproject.eu/.

[110] J. Lukas, J. Fridrich and M. Goljan. Digital camera identification from sensor pattern noise, *IEEE Trans. Inf. Forensics and Security,* 1(2): 205-214 (2006).

[111] W. Hu, T. Tan, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man and Cybernetics-Part C: Applications and Reviews*, 34(3): 334–352, 2004.

Detection of Stego Anomalies in Images Exploiting the Content Independent Statistical Footprints of the Steganograms

S. Geetha , Siva S. Sivatha Sindhu
Faculty, Department of Information Technology
Thiagarajar College of Engineering
Madurai-625 015, Tamil Nadu, India.
E-mail: sgeetha@tce.edu

N. Kamaraj
Department of Electrical and Electronics Engineering
Thiagarajar College of Engineering
Madurai-625 015, Tamil Nadu, India.
E-mail: nkeee@tce.edu

Keywords: image steganalysis, content independent distortion measures, genetic-X-means classifier

Received: September 1, 2008

Steganography which facilitates covert communication creates a potential problem when misused for planning criminal activities. Its counter measure steganalysis is focused on detecting (the main goal of this research), tracking, extracting, and modifying secret messages transmitted through a subliminal channel. In this paper, a feature classification technique, based on the analysis of content independent statistical properties, is proposed to blindly (i.e., without knowledge of the steganographic schemes) determine the existence of hidden messages in an image. To be effective in class separation, the genetic-X-means classifier was exploited. For performance evaluation, a database composed of 5600 plain and stego images (generated by using seven different embedding schemes) was established. Based on this database, extensive experiments were conducted to prove the feasibility and diversity of our proposed system. Our main results and findings are as follows:

1. *a 80%+ positive-detection rate.(promising rate for a blind steganalyzer)*
2. *The removal of content dependency from features enhances the discriminatory power of the classifier.*
3. *Universal, blind steganalyzer. (not limited to the detection of a particular steganographic scheme)*
4. *Detection of stego images with an embedding rate as low as 5% of the maximum payload.*

Povzetek: Opisana je metoda iskanj skritih sporočil v slikah.

1 Introduction

Steganography has been known and used for a very long time, as a way to establish covert communication between parties, by embedding the secret message in another, apparently innocuous, document. The goal of steganography is to communicate as many bits as possible without creating any detectable artifacts in the cover-object. Although steganography is an ancient subject, its modern formulation is often given in terms of the *prisoner's problem* by Simmons in 1983 [1]. In today's digital world, this has taken a new facet, however, and it must be approached in a spanking new view.

Due to the proliferation of the digital media and the easy accessibility to Internet, development of new technologies for network based multimedia systems and advanced multimedia services have been intensified. Many of the multimedia processing operations like editing, storage, transmission, and access of multimedia are easily done by any subject. Early methods exploited cryptography for secure transmission, to prevent unauthorized access and tampering of secret messages. However, the encrypted form may attract special attention of network warders and is thus not fully secret. Current information hiding techniques are developed to deceive warders by embedding messages into multimedia in an imperceptible manner, but still maintain their original formats and quality. Unlike cryptography, where the goal is to secure communications from an eavesdropper, steganographic techniques strive to hide the very presence of the message itself from an observer.

Steganography may provoke negative effects in the outlook of personal privacy, business activity, and national security. The scandalous can abuse the technique for planning criminal activities. For example, commercial spies or traitors may thieve confidential trading or technical messages and deliver them to competitors for a great benefit by using hiding techniques. Terrorists may also use related techniques to cooperate for international attacks (like the 9/11 event in the U.S.) and prevent themselves from being traced. Some others may even think of the possibility of conveying a computer virus or Trojan horse programs via data hiding techniques. Thus, it raises the concerns of enhancing warders' capability and lessening these

negative effects by developing the techniques of "steganalysis".

It should be noted that the primary goal of steganography is to set up a subliminal communication channel in a completely undetectable manner. In this context, "*steganalysis*" refers to the set of techniques that are designed to distinguish between cover-objects and stego-objects. Even though nothing might be gleaned about the contents of the secret message, when the existence of hidden message is known, revealing its content is not always necessary. Just disabling and rendering it useless will defeat the very purpose of steganography. This implies that the warder should be capable of discriminating suspicious objects from a large number of innocuous ones (i.e., the so-called passive steganalysis [3], [5]). In contrast to passive steganalysis, the goal of active steganalysis is to retrieve, modify, and even fabricate the embedded messages for destroying or interfering with covert communications and rendering hidden data useless. Applications of steganalysis then include, for example, an inlet/outlet content-monitoring program that inspects and intercepts suspected multimedia data transmitted on the network. In addition, steganalysis techniques can also be utilized to evaluate the security of covert communication channels under construction.

On the outset, deciding whether the cover media contains any secret message embedded in it or not is essential to steganalysis. Although it is uncomplicated to inspect suspicious objects and extract hidden messages by comparing them to the original versions, the restricted portability and accessibility of original cover-signals generally make blind steganalysis more attractive and reasonable in many practical applications. Blindness is meant to analyze stego-data without knowledge of the original signal and without exploiting the embedding algorithm. Hence, detecting the existence of hidden information becomes quite difficult and complex without exactly knowing which embedding algorithm, hiding domain, and steganographic keys were used. Apart from these issues, a steganalysis algorithm is required to possess other properties such as low complexity and low classification risk. A low-complexity algorithm makes the system capable of inspecting objects at a high throughput rate. An algorithm of low classification risk generally makes tradeoffs between costs resulting from missing errors (i.e., false negative) and from false alarms (i.e., false positive). This motivates our current research: devising a content independent feature-based algorithm to classify multimedia objects as bearing hidden data or not. Our objective is not to extract the hidden messages or to identify the existence of particular information (as it is in watermarking applications), but only to determine whether a multimedia object was modified by information hiding techniques. Once classified, the suspicious objects can then be inspected in detail by any particular data embedding/retrieving algorithms. This pre-process would particularly contribute to save time in active steganalysis.

As is well known, steganography and watermarking constitute two main applications of information hiding techniques. Though both applications share many common principles in data embedding/extraction schemes, they differ in some criteria, such as robustness, embedding capacity, requirement of original messages, etc. In certain scenarios, content owners might need to determine the existence of hidden watermark in a multimedia object, when the authentication program fails to extract or match the targeted watermarks (due to inversion attack, geometric attacks, de-synchronization attacks etc.). In a possibly negative viewpoint, users may use this steganalytic feature to identify the existence of watermarks in an object. To summarize, steganalysis has promising applications to detect both the steganographic and watermarking schemes.

Our research starts with the analysis and categorization of existing image hiding algorithms. This approach is based on the extension of the fact that hiding information in digital media requires alterations of the signal properties that introduce some form of degradation, no matter how small. These degradations can act as signatures that could be used to reveal the existence of a hidden message. For example, in the context of digital watermarking, the general underlying idea is to create a watermarked signal that is *perceptually identical but statistically different* from the host signal. A decoder uses this statistical difference in order to detect the watermark. However, the very same statistical difference that is created could potentially be exploited to determine if a given image is watermarked or not. The addition of a watermark or message leaves unique artifacts, which can be detected using the various distortion metrics i.e., Image Quality Measures (IQM) [4]. This paper extends the work in [4] and focuses on selecting the content independent features as potential evidences in revealing the presence of hidden messages. We intend to prove that this removal of content dependency enhances the sensitivity of the steganalyzer.

To blindly classify hiding status of an image, we propose an algorithm in which a set of image distortion metrics are defined and utilized to determine the existence of covert channels in the spatial or transformation domain or not. A systematic image database was constructed for algorithm evaluation and a genetic-X-means classifier [42] [43] was trained based on these evaluated features.

2 Steganography vs. steganalysis race

Conventional approaches to data hiding within images can be categorized into spatial or transform (e.g., DCT, DWT, Ridgelet etc.) domains [5]. Least Significant Bit (LSB) addition [6],[7],[8] or substitution [10], [11] method is the most popular hiding technique. These techniques operate on the principle of tuning the parameters (e.g., the payload or disturbance) so that the difference between the cover signal and the stego signal is little and imperceptible to the human eyes. Yet, computer statistical analysis is still promising to detect such a distinction that human beings are difficult to perceive. Some tools, such as StegoDos, S-Tools, and

EzStego, provide spatial-domain-based steganographic techniques [2], [5].

There were some spatial-domain steganalytic algorithms [12], [13], [14], [15], [4], [16] developed to be against the above steganographic schemes. Fridrich *et al.* [14] proposed a steganalysis technique based on the fact that bit planes in typical images are more or less correlated so that the LSB plane can be estimated from the other seven ones. This estimation becomes less reliable as the content of the LSB bit plane is further randomized. Kong *et al.* [16] proposed to evaluate the image complexity, following a statistical filter, to determine the existence of secret messages or not, based on the phenomenon that randomization of the LSB bit plane content becomes heavier after information hiding. Sanjay Kumar et al., in [9] discuss an active steganalysis where the estimation about the hidden message length is made. The proposed algorithm reduces the initial-bias, and estimates the LSB embedding message ratios by constructing equations with the statistics of difference image histogram.

Chandramouli *et al.* [12], [13] had ever assumed a Gaussian variation model for LSB disturbances, proposed a maximize *a posterior* (MAP) detector, and analyzed the maximum embedding capacity under which a steganalyst cannot detect the presence of hidden data with a desired probability. Unfortunately, neither detailed implementation of this MAP detector was given nor realistic experiments were reported in their work. Besides, their analysis was restricted to the Gaussian modelling of embedding disturbances. Gokhan Gul et al., in [46] briefly describe PQ and propose singular value decomposition (SVD)-based features for the steganalysis of JPEG-based PQ data hiding in images. They show that JPEG-based PQ data hiding distorts linear dependencies of rows/columns of pixel values, and proposed features can be exploited within a simple classifier for the steganalysis of PQ. Andrew A. Ker [18] proposes more accurate attacks on LSB embedding through a weighted stego image detector for finding the sequential image replacement.

Hiding can also be performed in the transform domain, e.g., DCT [19], [20], [21], [22], [23], [24], [25] or DWT domain [23], [26]. Regardless of which domain, "significant" transform coefficients are often selected to mix with secret/perturbing signal in a way such that information hiding or watermarking is transparent to human eyes. For instance, Cheng *et al.* [24] proposed an additive approach to hiding secret information in the DCT and DWT domains. Wu *et al.* [25] proposed a two-level data embedding scheme, in principle of additive spread spectrum and spectrum partition, for applications in copy control, access control, robust annotation, and content-based authentication. There exist some tools, such as J-Steg and Outguess, providing this category of steganographic techniques [5], [15].

Some steganalytic methods [14], [27], [28] were proposed in the DCT domain. Manikopoulos *et al.* [27] applied the differences in the coefficients of the block DCT transforms of the original as features to the detection of block DCT-based steganography in gray-scale images. The model utilizes statistical pre-processing, over an observation region of each image that generates feature vectors over the regions. These vectors are then fed into a simple neural network classifier. Fridrich *et al.* [14] described that a modified image block will most likely become saturated (i.e., at least one pixel with the gray value 0 or 255) in a JPEG-format stego-image after information hiding. If no saturated blocks can be found, there will be no secret messages therein. Otherwise, a spatial-domain steganalytic method [14] mentioned earlier can be used to analyze these saturated blocks. In [28], the author modelled the common steganographic schemes as a linear transform between the cover and stego images, which can be estimated after at least two copies of a stego image were obtained. This is similar to a blind source separation problem that can be solved by using the independent component analysis (ICA) [29] technique. In [30], a steganalytic scheme was devised to deal with information hiding schemes mixing a secret and a cover signal in an addition rule. The phenomenon, that the center of mass of the histogram characteristic function in a stego image moves left or remains the same to that of the cover image, was observed and exploited to distinguish stego images from plain ones.

It is noticed that most of the steganalytic schemes were designed either in specific operating domain, or even for particular steganographic algorithm. Building a universal steganalytic system is, up to now, a challenging exercise.

In [31], Wen et al. has modelled a universal steganalyzer that operates to distinguish stego images from clean images using two features only namely gradient energy and statistical variance of the Laplacian parameter. The system lacks the ability to strongly attack a wavelet based stego systems. But that can be solved by using a feature that is more sensitive to such embedding strategy. In [44] Der et al. proposes an universal steganalysis scheme that focuses on the differences of statistical features formed by embedding algorithms and applies a support vector machine to distinguish the stego-image from suspicious images. Even though many steganalytic systems have been developed, each system only identifies a subset of the available embedding methods and with varying degrees of accuracy. Benjamin in [45] applies Bayesian model averaging to fuse multiple steganalysis systems and identify the embedding used to create a stego JPEG image.

There are several fundamental questions one may ask:

Which features contribute more to the discriminating power of the universal steganalyzer?

Until what point does steganalysis performance improve with the number of features used? These questions are all related to a crucial ingredient of any blind steganalyzer.

Avcibas *et al.* [4] proposed a concept that any image will incur quality degradation after smoothing or low-pass filtering and this degradation (reacting on image quality) depends on the type of the test image, especially

in categories of with or without embedded information. That is, by observing quality difference between a test image and its smoothed version, it is possible to discriminate images with and without hidden messages. They hence utilized a regression analysis with several quality measuring operators for steganalysis. They have analyzed 26 image quality metrics for the purpose of discrimination. All features are not equally valuable to the learning system. Furthermore, using too many features is undesirable in terms of classification performance due to the curse of dimensionality [29]: one cannot reliably learn the statistics of too many features given a limited training set. Hence, we need to evaluate the features' usefulness and select the most relevant ones.

However it was discovered that removing the inherent content dependency in distortion measures as calculated in [4] is beneficial. So we propose a novel method to remove this content dependency from distortion measurements. These content-independent measurements are then used to build a classifier to differentiate cover-signals and stego-signals. The experimental results justify how the proposed technique enhances the discriminatory power of the features used in the classifier.

3 Effect of removing content dependence features

This paper quantifies steganalysis task in the information-theoretic prescription context of data hiding i.e., hiding in independent and identically distributed Gaussian host samples [3]. It is quite common to choose the embedding signal i.e., message to be conveyed as a zero mean, white Gaussian process with finite variance. It is known from information theory that a Gaussian signal is the best choice for a Gaussian channel. Since most image steganography methods conveniently assume the image pixel distribution and common transform coefficient distribution to be Gaussian, the choice of secret message as Gaussian is justified.

The prospects of certain image quality metrics in envisaging the presence of watermarking and steganographic signals within an image is described in [4]. The presence of the steganographic artifact can then be put into evidence by recovering the original cover signal, or alternatively, by de-noising the suspected stego-signal. The steganalyzer can directly apply a statistical test on the denoising residual, $x - \hat{x}$, where $\hat{x}$ is the estimated original signal. This residual must also correspond to the artifact due to embedding of a hidden message. Notice that, even if the test signal does not contain any hidden message, the de-noising step will still yield an output, whose statistics can be expected, however, to be different from those of a true embedding. There subsists a motive to utilize more than one distortion measure, in order to investigate different quality aspects of the signal, which could be brunt during data hiding manipulations. In pursuing such a task, there is often the risk that the variability in the signal content itself surpasses the detector from the alterations. Thus, it is required that, whatever features are selected, the detector responds only to *the induced distortions,* which is Gaussian distributed [3], during data hiding and not be confused by the statistics of the signal content. Moreover, the original signal apparently will not be available during the testing stage. Therefore, some reference signal must be created that is common to both the training and testing stages.

In [4], a denoised version of the given signal is used as the reference. Anyway, this self-referencing, which is creating a reference signal via its own denoised version, is obviously a content-dependent scheme. The classifier performance can be inferior as it responds to both the signal content based statistics and to the distortions stimulated by data embedding operation. To eliminate this content dependency, it is recommended to use a single reference signal that is common to all signals to be tested. Thus, a content independent reference signal and its altered versions according to the type of data embedding are employed.

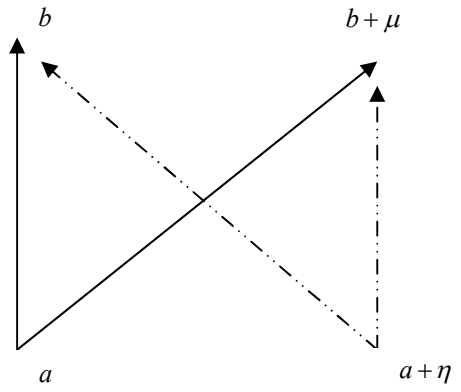


Figure 1: Signal vectors: original signal a, its embed-ded version $a+\eta$, reference signal b and its embed-ded version $b+\mu$.

Let a denote a test signal and $a+\eta$ be its stego version, and similarly, let b and $b+\mu$ indicate the reference signal and its stego version. Besides, let us consider a generic distortion functional $Distortion(x, y)$ between the signals x and y. For example, for the mean-square distortion, one simply has $Distortion(x, y) \triangleq E[(x-y)^2]$, with E being the expectation operator. The detector operates on the basis of the statistical differences of the distortions. This implicitly ensues two assumptions. First, data embedding leads to additive distortion, that is, the altered signals can be represented as $a+\eta$ and $b+\mu$. Second, the additive distortions of the test and reference signals should not be mutually orthogonal, that is, $E\{\eta, \mu\} \neq 0$. This assumption was indirectly justified by analysis of variance (ANOVA) [4] and the test results given in the experimental results section.

It is to be shown that self-referencing, as employed in [4], causes content-dependent distortion. Let $\Re$ be the specific operation by which the reference signal is generated; for example, in [4], denoising operation has

been used $b = \Re(a) = denoise(a)$. The outcomes of this operation are given by $a \xrightarrow{\Re} \Re(a)$ and $a+\eta \xrightarrow{\Re} \Re(a+\eta)$, respectively, for (1) gnal and its stego version. To illustrate the point, for the case of the mean-square distortion, one obtains

$$Distortion(a+\eta,\Re(a+\eta)) - Distortion(a,\Re(a)) = E[\Re(a+\eta)^2 + 2a\eta + \eta^2 - 2(a+\eta)\Re(a+\eta) + 2a\Re(a) - \Re(a)^2]$$

which is content dependent, because the signal terms a and $\Re(a)$ survive in the difference of distortion functions. The above difference should be some function of only the distortion term and should not contain a or any of signal derived from it, to ensure content independence and to be a real indicator of data embedding effects.

We propose an alternative way and suggest to consider an unique signal b as a reference signal. Then the distortion metrics can be measured between a and $a+\eta$, using b and $b+\mu$ as reference signals. The relationship of these signals and the distortion *vis-à-vis* the reference signals b and $b+\mu$ illustrated in Fig. 1. In this figure, the length of the vector $\overrightarrow{ab}$ is simply equal to $Distortion(a,b)$. The distance between the tips of the vectors $\overrightarrow{ab}$ and $\vec{a}(b+\mu)$ is $d = Distortion(a,b) - Distortion(a,b+\mu)$ and simi-larly $d' = Distortion(a+\eta,b) - Distortion(a+\eta,b+\mu)$, denotes the distance between the tips of the dashed pair of vectors.

For the case of mean-square distortion it follows that

$$d = E[(a-b)^2 - (a-b)^2 + 2(a-b)\mu - \mu^2] = E[2(a-b)\mu - \mu^2] \quad (2)$$

$$d' = E[(a+\eta-b)^2 - (a+\eta-b)^2 + 2\mu(a+\eta-b) - \mu^2] = E(2\mu(a-b) + 2\mu + \eta - \mu^2] \quad (3)$$

To remove the content dependency it is enough that we calculate the difference between d and d'.

$$D = 2E[\mu\eta] \quad (4)$$

The same effect of eliminating content dependency can be shown with another distortion metric, the correlation coefficient given by $Distortion(x,y) = E[(xy)]$.

Here $d = E[ab] - E[a(b+\mu)] = -E[a\mu]$ and

$$d' = E[(a+\eta)b] - E[(a+\eta)(b+\mu)] = -E(a\mu] - E[\eta\mu] \quad (5)$$

The removal of content dependency can be shown as the difference between d and d' like $D = d' - d = -E[\eta\mu]$. (6)

4 Design of the steganalyzer

This paper mainly concentrates on designing a blind steganalyzer that can distinguish between a clean image and an adulterated image, using an appropriate set of content independent IQMs. Objective image quality measures are based on image features, a functional of which, should correlate well with subjective judgment, that is, the degree of (dis)satisfaction of an observer [32]. Objective quality measures have been utilized in coding artifact evaluation, performance prediction of vision algorithms, quality loss due to sensor inadequacy etc. [33]. In [4] they have extensively studied the use of image quality measures specifically as a steganalysis tool, that is, as features in detecting watermarks or hidden messages.

4.1 Content Independent Image Quality Metrics (CIIQMs) as features

A good IQM should be accurate, consistent and monotonic in predicting quality. In the context of steganalysis, *prediction accuracy* can be interpreted as the ability of the measure to detect the presence of hidden message with minimum error on average. Similarly, *prediction monotonicity* signifies that IQM scores should ideally be monotonic in their relationship to the embedded message size or watermark strength. Finally, *prediction consistency* relates to the quality measure's ability to provide consistently accurate predictions for a large set of watermarking or steganography techniques and image types. This implies that the spread of quality scores due to factors of image variety, active warden or passive warden steganography methods should not eclipse the score differences arising from message embedding artifacts. In order to understand how these metrics measure up to the above desiderata [4] resorted to analysis of variance (ANOVA) techniques. The ranking of the goodness of the metrics was done according to the F-scores in the ANOVA tests to identify the ones that responded most consistently and strongly. In the final analysis a list of IQMs is obtained that are sensitive specifically to steganography effects, that is, those measures for which the variability in score data can be explained better because of some treatment rather then as random variations due to the image set.

The stego-detector we develop is based on analysis of a number of *relevant but content independent* IQMs. The idea behind detection of watermark or hidden message presence is to obtain a consistent distance metric for images containing a watermark or hidden message *vis-à-vis* those without, *with respect to a common reference*. The reference processing should possibly include a general signal common to both testing and training. Our approach differs from [4] in using a random signal as the common reference signal rather than using a denoised signal.

The quality metrics exploited in [4] are categorized into six groups according to the type of information they use. The categories used are:

Pixel Difference-based Measures: Mean square error, Mean absolute error, Modified infinity norm, L*a*b perceptual error, Neighborhood error and Multi-resolution error.

Correlation-based Measures*:* Measures based on correlation of pixels, or of the vector angular directions like Normalized cross correlation, Image fidelity, Czenakowski correlation, Mean angle-magnitude similarity and Mean angle similarity.

Edge-based measures: Measures based on the displacement of edge positions or their consistency across resolution levels like Pratt edge measure and Edge stability measure.

Spectral distance-based Measures: Measures based on the Fourier magnitude and/or phase spectral discrepancy on a block basis like Spectral phase error, Spectral phase-magnitude error, Block spectral magnitude error, Block spectral phase error and Block-spectral phase-magnitude error.

Proposed Algorithm: *CIIQM Based Steganalyzer*

Phase: *Learning*

Input: *A database of images*

Output: *A knowledge base capable of discriminating between a clean and a stego image*

1. **Image data base construction**: Prepare an image data base containing clean images and stego images generated out of difference embedding schemes.
2. **Removal of content dependency:** A single random reference signal that is common to all the signals is selected for evaluating IQMs
3. **IQM evaluation:** The various IQMs mentioned in Section 5 like Pixel Difference-based Measures, Correlation-based Measures, Edge-based measures, Spectral distance-based Measures and Context-based Measures are evaluated, between the test signal and the common reference signal.
4. **Genetic Algorithm based feature selection**: The content independent features that are sensitive to data embedding operation are selected based on genetic search strategy. Genetic-X-means algorithm is described in the listing 2.
5. **Data set formation**: A data set is formed out of the selected features.
6. **Training:** The steganalyzer is subjected to learning by applying the X-means algorithm over this data set and a knowledge base is constructed.
7. **System Ready**: The steganalyzer system is now ready for universal blind steganalysis.

Phase: *Detecting*

Input: *A test signal which is to be categorized as a clean or stego bearing image.*

Output: *Categorization of the signal as clean or stego-bearing*

1. **Network Daemon:** It monitors traffic and channelises the multimedia data to the IIU.
2. **Image Identification Unit:** This is used for identification of the image data files. It is achieved by observing the header information of each and every incoming data packet. Various image files being identified by this component are .BMP, .GIF, .TIFF, .PNG etc.
3. **Common reference signal selector:** The same signal chosen in the learning phase is chosen to evaluate the CIIQMs.
4. **CIIQM Evaluator:** The content independent IQMs selected in Step 4 of the learning phase are evaluated against the same common reference signal chosen in the learning phase.
5. **Genetic-X-means Clustering Engine**: The derived feature vector is given to the X-means classifier engine for diagnosis. Based on the knowledge constructed in learning phase this component decides whether the document is adulterated or untouched and identifies the specific steganographic technique used. The algorithm is given in the Listing 2.
6. **Actioner:** Actioner's role is to take necessary actions when a marked document is detected. When an adulterated file is detected exactly, the actioner does one of the following operations. 1. Warn the system administrator 2. Warn the end user 3. Kill the specific application, which executed that image file 4. Prevent the end user from running any further application.5. Extract the hidden information from/in the image file. Case 2, 3, 4 & 5 can be achieved locally at the client workstation.

Listing 1. Framework of the CIIQM based Steganalyzer.

Algorithm **:** *Genetic-X-means algorithm applied to image steganalysis.*

Input **:** *Training set D={(x1,y1),(x2,y2),…:(xn,yn)}, Lower bound=α ,Upper bound=β .*

Output **:** *The clustered model with the maximum BIC (Bayesian Information Criterion) Score, the respective value of K and K centroid parameters.*

1. *Genetic_Feature_Selection()*

Initialize K= α,

2. **Improve Params**: Run *direct k-means* to convergence, with the features selected using genetic search algorithm

Improve Structure: Add new centroids where needed by applying SPLITLOOP system as in Gaussian Mixture model identification [43].

For the locally evolved model Mj, evaluate the BIC Score locally evolved, with k=1 and k=2 :

$$BIC(M_j) = i_j(D) - \frac{p_j}{2}\log R$$

where $i_j(D)$ is the log-likelihood of the data according to the j-th model and taken at the maximum-likelihood point, p_j is the number of parameters in M_j and $R = |D|$.

Sustain the model M_j having greater BIC score.

Record the parameters of the model evolved, like: K, BIC score of the entire model, K Centroid values.

8. If K< β, Goto 2.

Listing 2: Pseudo code for Genetic-X-Means algorithm applied to Image Steganalysis.

Context-based Measures: Measures based on penalties for various functional of the multidimensional context probability like Rate distortion measure, Hellinger distance, Generalized Matusita distance and Spearman rank correlation.

4.2 Choice of genetic-X-means paradigm

According to whether the steganalysis is based on supervised or unsupervised learning, stego-anomaly detection schemes can be classified into two categories: unsupervised stego-anomaly detection and supervised stego-anomaly detection. In supervised strategy, profiles of clean/stego files are established by training using a labelled dataset. Unsupervised detection uses unlabelled to identify anomalies. The main drawback of supervised detection is the need to label the training data, which makes the process error-prone, costly and time consuming. Unsupervised anomaly detection addresses these issues by allowing training based on an unlabelled dataset and thus facilitating online learning and

Algorithm **:** *direct K-means()*

Inputs:

I = {i1,i2,…ik} (Stego/Clean instances to be clustered)

n (Number of Clusters)

θ *(Threshold value as a stopping criterion)*

Outputs:

C = {c1,c2,…,ck} (Cluster centroids)

m: $I \to C$ *(Cluster membership)*

Initialize k prototype $(w_1, w_2 \cdots w_k)$ such that $w_j = i_l$, $j \in \{1,2,...,k\}$, $l \in \{1,2,...,n\}$.

Each cluster C_j is associated with prototype w_j

Repeat

For each input vector i_l, where $l \in \{1,2,...,n\}$ do

$m(i_j) = \arg\min_{k \in \{1..n\}} distance(i_j, c_k)$

For each cluster C_j ,where $j \in \{1,2,...,k\}$ do

Update the prototype w_j to be the centroid of all samples currently in C_j so that $w_j = \sum_{i_l \in C_j} i_l / |C_j|$ and $m(i_j) = \arg\min_{k \in \{1..n\}} distance(i_j, c_k)$

Compute the error function

$$E = \sum_{j=1}^{K} \sum_{i_l \in C_j} |i_l - w_j|^2$$

Until $E < \theta$ or cluster membership no longer c hanges.

Listing 3: Pseudo code for direct k-means algorithm.

improving detection accuracy. By facilitating online learning, unsupervised approaches provide a higher potential to find new attacks. By removing the need of labelling, unsupervised detection creates a greater potential for accurate detection.

Clustering is the organization of data patterns into groups or clusters based on some measure of similarity. When applying clustering techniques for steganalysis, determining the number of clusters is a difficult issue since the data hiding algorithm is unknown. The general approach and current practice assume that data instances are always divided into two categories: normal clusters and anomalous clusters. However, this assumption need not always true in practice. The number of clusters is not supposed to be determined in advance. When data instances include only normal behavioral data, the assumptions will lead to a high false alert rate and a vice-versa case when data instances include only stego patterns. In order to achieve an efficient and effective detection, we propose in this paper, a new unsupervised stego-anomaly detection framework which consists of a clustering algorithm, named X-means and a CIIQM feature extraction based on genetic search. X-means

Algorithm : *Genetic_Feature_Selection()*

Input :
Encoded binary string of length 26 (one bit for each IQM), number of generations, and population size, Cross over probability Pc, Mutation Probability Pm.

Output :
A set of selected features.

1. Initialize the population randomly

2. $W1 = 10^4$, $W2 = 0.4$,

3. N = total number of records in the training set
4. For each chromosome in the new population

5. Apply uniform crossover operator to the chromosome with a probability of Pc.6. Apply mutation operator to the chromosome with a probability of Pm.

7. Evaluate Fitness = *W1 * Accuracy + W2 * Zeros*

8. Select the top best 50% chromosomes into new population using Tournament Selection operator.

9. If number of generations is not reached, go to step 4.

Listing 4: Pseudo Code for Feature Selection by Genetic-Search Strategy.

algorithm extends appropriately k-means with some evolutionary steps, integrates the capability of determining automatically the optimal number of clusters for a set of data, and thus addresses the limitation of traditional clustering based intrusion detection approaches.

5 Proposed algorithm for CIIQM genetic-X-means image steganalyzer model

The proposed algorithm for genetic-X-means based image steganalysis system is provided here. This can be set up in the network of the corporate sectors. The multimedia traffic (image, video, image, text, HTML pages etc.,) is keenly monitored by the system. Whenever the entry of image documents is sensed, the steganalyzer is triggered. The system consists of two main stages. They are 1. Learning Stage 2. Detecting Stage.

6 Experimental topology

In our experiments, the discrimination performance of content independent features is analyzed first. Then the classification performance of our steganalyzer under the prepared test image set is reported. Besides, the impacts of embedding rate and mismatch between the training and test sets (e.g., modified by using different embedding schemes) on the classification rate are also explored.

6.1 Preparation of test images and schemes

The design of experiments is important in evaluating our steganalytic algorithm. The key considerations include the following.

1) First, from the point of "generalization", the proposed content independent image features and associated classifier should be capable of identifying the existence of hidden data which are possibly generated by using various kinds of embedding methods, regardless of steganography or watermarking, and regardless of spatial or transform-domain operations.

2) Second, in outlook of "performance", the classifier should, on the one hand, detect hidden data as likely as possible (regardless of how transparent the embedded secret information is), and on the other hand, keep false alarms to as few as possible for plain images.

3) Third, in view of "robustness", the classifier should be capable of differentiating the effect of ordinary image processing operations (such as filtering, enhancement, etc.) from that of data embedding.

On the grounds of the above considerations, six published methods based on two types of principles, LSB embedding and spread spectrum, were chosen for evaluation.

scheme **#1**: Digimarc [34]
scheme **#2**: PGS [35]
scheme **#3**: Cox *et al.*'s [22]

scheme **#4**: S-Tools [36]

scheme **#5**: Steganos [37]
scheme **#6**: JSteg [38]
scheme **#7:** Kim *et al.'s* [39]

They can be further categorized into:

1) steganography (#4, #5, #6) or watermarking (#1,#2,#3) purpose;

2) spatial (#2, #4, #5), or transform (#1, #3, #6) domain operation.

For further testing and to verify the effectiveness of the features selected, we select an extra scheme based on the wavelet domain:

3) scheme **#7**: Kim *et al.*'s method [39].

It is expected that the difference between a cover image and its stego version can be easily detected when more secret messages are embedded. Hence the capacity of the payload of a steganography scheme should be taken into account in evaluating the detection capacity of a steganalytic classifier. To depict this, the embedding rate (ER) characterizing a scheme which is defined as the ratio between the number of embedded bits and the number of pixels in an image, is used.

- Mean square error
- Median block weighted spectral distance
  HVS based L2
- HVS normalized absolute error
- Weighted spectral distance
- Cross correlation.

Table I: content independent distortion measures selected by genetic search

The wavelet-based steganography scheme #7 was used to test our steganalytic scheme, although the proposed features are trained only on the spatial and the DCT domains.

To test the performance of the proposed method, our cover image dataset consists of 200 with a dimension of 256 X 256 8-bit gray-level photographic images, including standard test images such as Lena, Baboon, and also images from [40]. Our cover images contain a wide range of outdoor/indoor and daylight/night scenes, including nature (e.g., landscapes, trees, flowers, and animals), portraits, manmade objects (e.g., ornaments, kitchen tools, architectures, cars, signs, and neon lights), etc. Some of the sample images are shown in Fig. 2. This database is augmented with the stego versions of these images using the above mentioned seven schemes, at various embedding rates. Also a separate image set was generated by applying the image processing techniques like JPEG compression (at several quality factors), low-pass filtering, image sharpening etc. Our generation procedure is aimed at making even contributions to database images from different embedding schemes, from original or stego, and from processed or non-processed versions, so that the evaluation results can be more reliable and fair. Three different ERs are attempted for each scheme in generating the database like (#1) 5% (#2) 10% (#3) 20% of the maximum payload capacity prescribed by the techniques. The entire database contains 200*4*7=5600 (No. of images * No. of varying ER - 3 ER + 1 for clean set* No. of schemes evaluated) images on the whole.

6.2 Content independent features selection

Applying the proposed methodology and the algorithm, the content dependency was removed and the six measures as in Table I are selected after removing the content dependency from the signal.

6.3 Feature discrimination capability

Before proceeding to evaluate the performance of the classifier, discrimination capability of the proposed features is to be analyzed. The experiment involves breaking of different steganographic or watermarking strategies, which may adapt extremely different techniques for embedding ranging from LSB substitution to embedding inside the wavelet co-efficient.

Hence the feature set formed has to be normalized before feeding into the classifier for training to achieve a uniform semantics to the feature values. A set of normalized feature vectors as per the data smoothing function [41],

$$\breve{f}_i = \frac{f_i - f_i^{\min}}{f_i^{\max} - f_i^{\min}} \ , \qquad (7)$$

are calculated for each seed image to explore relative content independent feature variations after and before it is modified. $\breve{f}_i$, $f_i^{\min}$ and $f_i^{\max}$ represents the i[th] feature vector value, the corresponding feature's minimum and maximum value respectively.

6.4 Genetic-X-means classifier

In the sequel, the model is incorporated in Java JGAP [42] and the algorithm described in section [6] is implemented as per the framework proposed. The classifier was trained and evaluated by using 4800 images out of the whole database, excluding those generated by using scheme #7 (employed as the test images to see how the proposed features behave when there is a mis-match between the operation domains). Here, two-thirds (3200) of images were randomly chosen as the training set and the others (1600 images) act as the validation set.

Before evaluation, some performance indices are first defined.

- Positive detection (PD)—classifying the stego images correctly.

- Negative detection (ND)—classifying the non-stego images correctly.
- False positive (FP)—classifying the presence of secret information for non-stego images.
- False negative (FN)—bypassing or ignoring the presence of hidden information in stego images.

The classification and error rates obtained by using different values are listed in Table II. Results show that the average classification rate does not change much (from 79.5% to 86.67%). We are interested in analysing the detectability of proposed features and classifier against embedding schemes of different applications or principles. Table III lists classification and error rates to see differentiation in performances between: 1) six targeted embedding schemes; 2) steganographic or watermarking applications; 3) spatial or DCT operation domain; and 4) types of processed non-stego images. We also analyzed the ND rates for the original, smoothed, sharpened, and JPEG-compressed non-stego images. It is found that our system has a better performance in recognizing the plainness of JPEG-compressed images. The higher ND rate for JPEG-compressed images is beneficial to real applications, since most images will be compressed in the JPEG form.

	PD		**ND**		**Classification Rate(PD+ND)/2**	
Scheme	**IQM**	**CIIQM**	**IQM**	**CIIQM**	**IQM**	**CIIQM**
DigiMarc	80%	85.63%	80%	84.97%	80%	85.30%
PGS	80%	81.02%	90%	92.31%	85%	86.67%
Cox	80%	84.96%	60%	72.30%	70%	78.63%
S-Tools	90%	93.31%	60%	78.04%	75%	85.68%
Steganos	80%	87.63%	60%	75.54%	70%	81.59%
Jsteg	70%	84.97%	70%	74.02%	70%	79.50%
	FP		**FN**		**Error Rate (FP+FN)/2**	
Scheme	**IQM**	**CIIQM**	**IQM**	**CIIQM**	**IQM**	**CIIQM**
DigiMarc	20%	14.37%	20%	15.03%	20%	14.70%
PGS	10%	19.98%	20%	7.69%	15%	13.84%
Cox	20%	15.04%	40%	27.70%	30%	21.37%
S-Tools	10%	6.69%	40%	21.96%	25%	14.33%
Steganos	20%	12.37%	40%	24.46%	30%	18.42%
Jsteg	30%	15.03%	30%	25.98%	30%	20.51%

Table 2: Performance comparison of the classifiers.

Differentiation categories		PD rate
Schemes	#1	85.3%
	#2	86.67%
	#3	78.63%
	#4	85.68%
	#5	81.59%
	#6	79.5%
Applications	Watermarking	83.53%
	Steganography	82.25%
Operation domain	Spatial	84.64%
	DCT	81.14%
	DWT	86.32%
Differentiation categories		ND rate
Type of processed non-stego images	Original	82.34%
	JPEG-compressed	89.40%
	Smoothed	83.50%
	Sharpened	58.10%

Table 3: Average pd/nd rates for performance differentiation between different target schemes, different applications, different operation domains, and different types of nonstego images.

As for the detectability between different embedding schemes, we compare scheme #4 to #5 and scheme #1 to #3. Basically, embedding schemes #4 and #6 are similar in some aspects (both are in the spatial-domain, but for different applications), but the pixel change will be less for scheme #4 when embedding "0." Accordingly, we got a higher PD rate for scheme #4 than for scheme #5.

6.5 Influence of embedding rate

In this experiment, the images at various payload capacities were selected to see the influence on detectability. The ERs for the six embedding schemes were tried at 5%, 10% and 20% of the maximum hiding capacities in their proposed versions. The experimental results are listed in Table V, which depicts that the average PD rate still remains above 83.38% for 20%, 78.78% for 10% and 74.47% for 5% of maximum payload capacity. The results for steganographic schemes are more promising than for the watermarking schemes, as the steganographic schemes carry more hidden data than those of watermarking schemes, which makes the measured features more distinguishable for detection. The results reveal that clearly, our proposed content independent features and genetic-X-means classifier still yield reasonable results for stego images of less ER.

6.6 Detection with mismatch between the training and test sets

Here we evaluate the performance when images modified by using different kinds of hiding schemes are employed for training and testing. Denote the training set and the test set as S_L and S_T respectively.

First, we created S_L^1 by including stego images generated by using the steganographic schemes #1 and some processed plain images. On the other hand, S_T^1 is constituted of stego images produced by using the watermarking schemes #2, #3 and other processed plain images. Essentially, S_L^1 and S_T^1 were made disjoint and consist of 400 and 600 images, respectively. We also evaluate the detection performances in presence of other mismatches In the second case, we interchanged the roles of S_L^1 and S_T^1 to form another two sets, S_L^2 and S_T^2, i.e., $S_L^2 \leftarrow S_T^1$ and $S_T^2 \leftarrow S_L^1$. Similarly S_L^3 includes stego images created using schemes #4 and #5. S_T^3 constitutes the stego images processed by using scheme #6. S_L^4 and S_T^4 represent the reversed role of S_L^4 and S_T^4 sets. Another set S_L^5 and S_T^5 contains the stego images created by employing the schemes #1,#2,#4,#5 and tested on schemes #3 and #6 respectively. Their interchanged sets are S_L^6 and S_T^6.The experimental results are listed in Table IV, which reveals that the average classification rate is 83.35%. Noticeable is that the PD rate for S_T^2, S_T^4 and S_T^6 is much higher than that for S_T^1, S_T^3 and S_T^5. The reason is that the steganographic schemes that embed in the spatial domain reveal more statistical evidence than the ones which hide in the transform

Data Group	PD	ND	Classification rate	FP	FN	Error rate
S_L^1, S_T^1	70.71%	86.69%	78.7%	29.29%	13.31%	21.3%
S_L^2, S_T^2	75.74%	87.08%	81.41%	24.26%	12.92%	18.59%
S_L^3, S_T^3	79.28%	88.07%	83.68%	20.72%	11.93%	16.33%
S_L^4, S_T^4	92.13%	83.37%	87.75%	7.87%	16.63%	12.25%
S_L^5, S_T^5	73.74%	89.69%	81.72%	26.26%	10.31%	18.29%
S_L^6, S_T^6	91.25%	82.48%	86.87%	8.75%	17.52%	13.14%

Table 4: Classification rates obtained when characteristics of the training and test sets mismatch to each other.

Schemes	Classification rate			Error rate		
	5% of maximum payload	**10% of maximum payload**	**20% of maximum payload**	**5% of maximum payload**	**10% of maximum payload**	**20% of maximum payload**
#1	80.40%	83.20%	85.30%	19.60%	16.80%	14.70%
#2	79.80%	83.10%	86.67%	20.20%	16.90%	13.33%
#3	70.11%	73.50%	78.63%	29.89%	26.50%	21.37%
#4	76.30%	79.22%	85.68%	23.70%	20.78%	14.32%
#5	71.20%	78.33%	81.59%	28.80%	21.67%	18.41%
#6	69.80%	74.55%	79.50%	30.20%	25.45%	20.50%
#7	73.66%	79.54%	86.32%	26.34%	20.46%	13.68%

Table 5: Classification and error rates for test sets at various embedding rate.

domain. This makes the measured content independent features more distinguishable for detection. After examining Table IV, it was found that S_T^2 , S_T^4 and S_T^6 will gain a high PD rate than S_T^1 , S_T^3 and S_T^5 . Hence, we have a conjecture that characteristics (e.g., ER, type of applications, or operating domain) of the test stego image may play an important role on PD rates, but mismatch between the training and test sets might not be so significant.

6.7 Application on a completely new steganography scheme

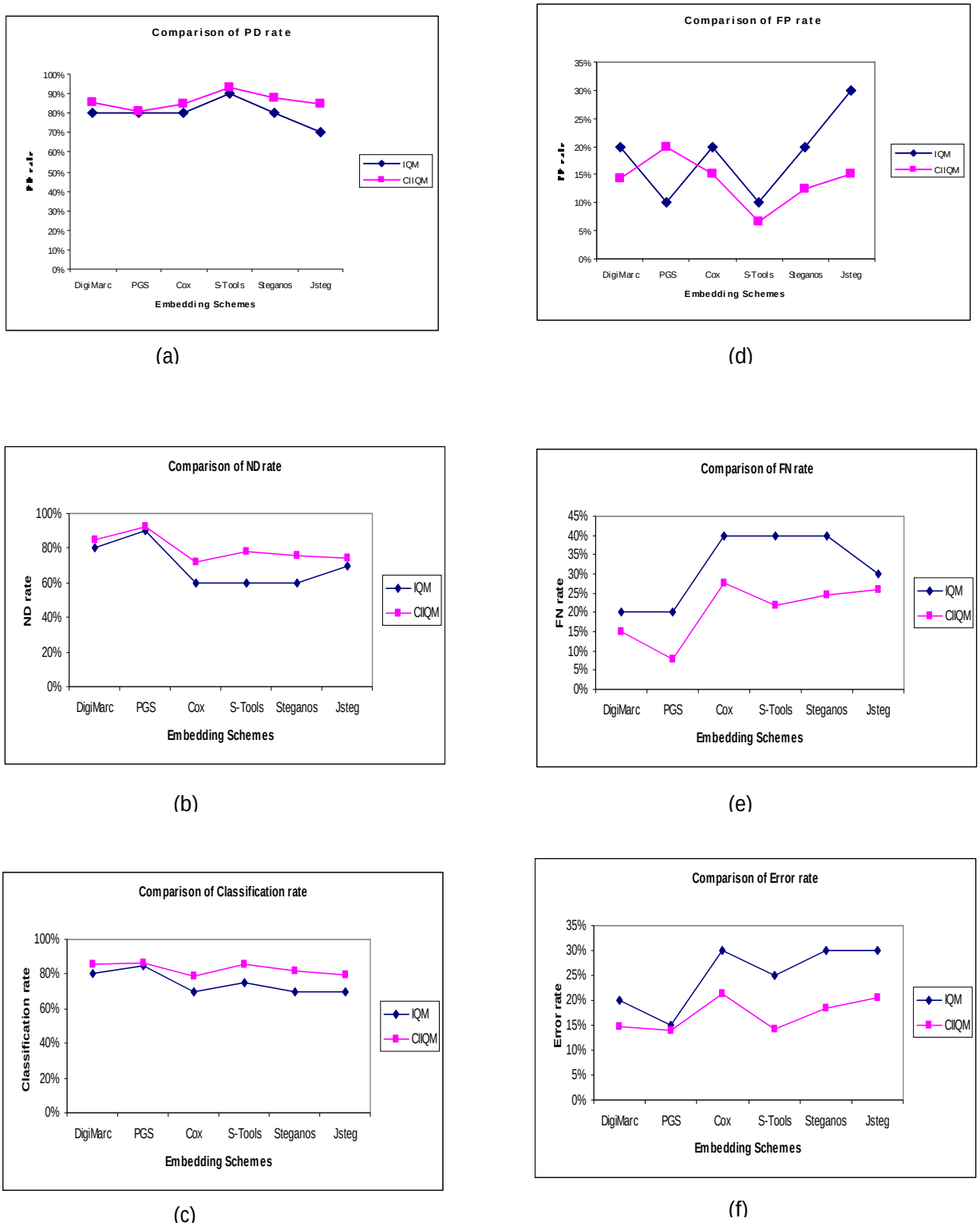


Figure 2: Performance comparison curves depicting a) Positive detection rate b) Negative detection rate c) Classification rate d) False positive rate e) False negative rate f) Error rate

In order to show that the system is dynamic i.e., adaptable to detect any new steganographic technique, the system was tested on scheme #7, which is based on the wavelet-domain techniques.. It was found that the PD rate against scheme #7 is 86.32% as given in Table V. This proves that the identified content independent IQMs are sensitive to detect even any new stego systems. To accommodate the identification of more hiding schemes, other kinds of image features should be explored further.

7 Discussion and conclusion

Recently, information hiding techniques find its applications in several fields, e.g., watermarking, copyright protection, steganography, fingerprinting, digital rights management (DRM), etc. At one end there are much research works focusing on addressing the various edges of data embedding techniques like enhancing the transparency, robustness and capacity. On the other end it is, however, interesting to detect the existence of hidden data resulting from any kind of

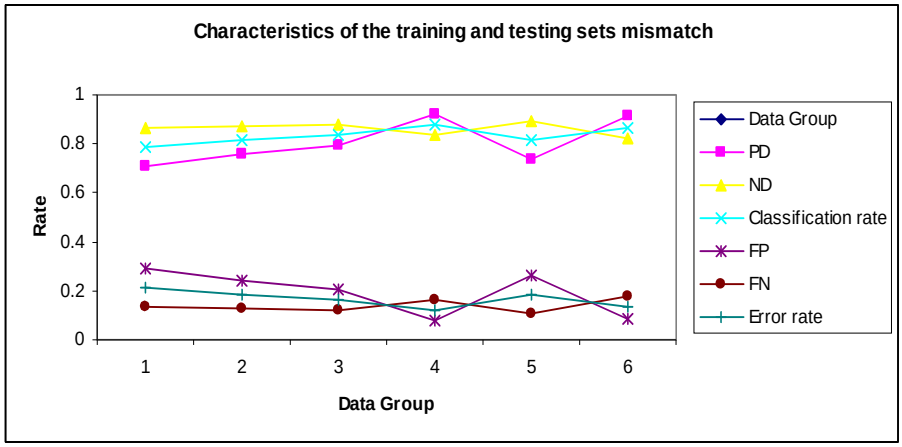


Figure. 3. Detection with Mismatch between the Training and Test Sets.

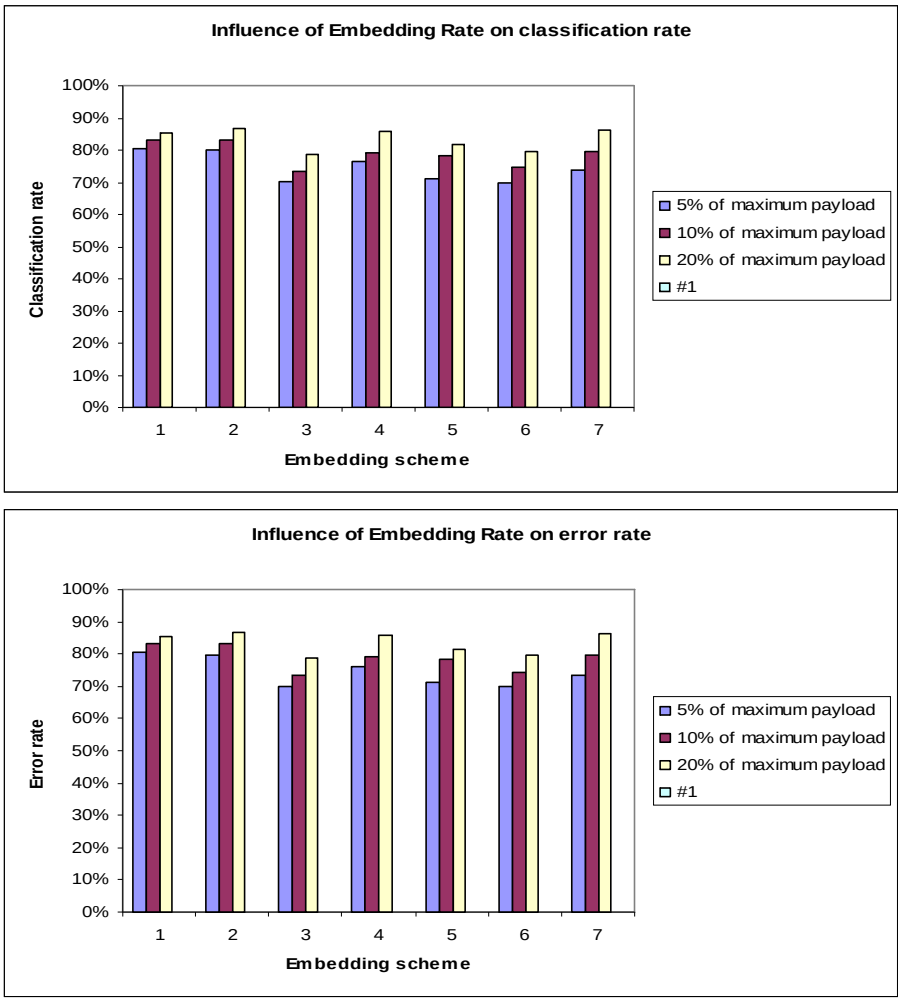


Figure 4. Influence of Embedding rate on performance of the steganalyzer.

embedding scheme, known as the "steganalysis." We have presented a rationale for a content independent distortion metrics based model for blind image steganalysis i.e., without knowledge of the kind of steganographic schemes and shown evidence using systematic experimental results. In our experiments, a database composed of processed plain images and stego images generated by using seven embedding schemes was utilized to evaluate the performance of our proposed features and classifier. Removal of content dependency from the measurements enhanced the classifier's discriminatory power and proved to be useful, especially for steganographic data embedding, where the incurred distortions are much less pronounced than in watermarking.

Table VI summarizes and compares characteristics of our proposed method with those of several other previous works in literature. In the table, not-reported (NRP) represents null information provided by the original work. For clarity, several key points are collected as follows.

1. It seems that the classification performance is necessarily proportional to the removal of content dependency from the features, heavily dependent on the ER and number of embedding schemes under tests.
2. Similar to [31], [36], and [43], our system is operated blindly and not restricted to the detection of a particular steganographic scheme (such as LSB or spread spectrum).
3. A nonlinear classifier that is easy to adapt to non-separable classes is adopted in both [33] and our system. However, the system introduced in [33] was only dedicated to the detection of spread spectrum scheme and few test images were used.
4. Our training and test database collects a larger number of stego and non-stego image samples, that were generated by using different steganographic schemes (seven kinds), different embedding rates (5% –20% maximum payload), and different image processing (low-pass filtering, sharpening, JPEG compression). This diversity makes our system more approximate to real applications.
5. The average classification rate (83%, including the PD and ND rates) for our proposed system is superior to [31] in blind steganalysis research.

To make our system more practical, future work could include the following.

a. Fitting the proposed system to classify compressed images or videos.
b. Identifying the type of steganographic algorithm utilized to generate the stego image and locating the image regions exploited to hide secret messages (active steganalysis). After these, we may be able to locate, retrieve, and analyze the embedded messages to infer the conveyed information.
c. Improving the performance as well as the scalability of the blind steganalyzer using appropriate fusion techniques

8 Acknowledgement

This work was supported by grants from National Technical Research Organization of Government of India, as a part of "Smart and Secure Environment". The authors sincerely thank the Management, Principal and Head of the Department of Information Technology of Thiagarajar College of Engineering, Madurai, India, for their support and encouragement. Authors would like to thank the anonymous reviewers for the constructive comments which helped to improve the clarity and presentation of the paper.

Steganalytic Systems	**[13]**	**[14]**	**[27]**	**[16]**	**[4]**	**[30]**	**[31]**	**Proposed**
Number of features	1	4	164	2	10	1	2	5
Domains of Feature Extraction	Spatial	Spatial DCT	DCT	Spatial	Spatial DFT	DFT	Spatial DCT	Spatial DCT DWT
Training/Classifier	Yes/Linear	Yes/Linear	Yes/Neural	Yes/Linear	Yes/Linear	Yes/Linear	Yes/Neural	Yes/Genetic-X-means
Targeted embedding scheme	LSB	Arbitrary	Spread spectrum	LSB	Arbitrary	Arbitrary	Arbitrary	Arbitrary
Number of test schemes	1	6	1	1	6	3	6	7
Payload of stego images	0.65 bpp	>0.05bpp	0.016 bpp	>0.05 bpp	>0.01 bpp	1 bpp	0.01-2.66 bpp	>0.01 bpp
Size of training database	NRP	331	28	NRP	12	20	1716	3200
Number of test images	NRP	NRP	14	80	10	4	572	1600
Average PD rate	NRP	NRP	0	97%	72.08%	96.2%	80.28%	86.25%
Average ND rate	NRP	NRP	0	NRP	NRP	94.8%	79.56%	79.53%
Side information constraint for classifier	No	No	Average PDF of selected plain images.	No	No	No	No	No

Table 6: Summarization of previous works and our proposed system. *NRP-Not Reported

9 References

[1] G. J. Simmons (1984). The prisons' problem and the subliminal channel. *Proc. Advances in Cryptology (CRYPTO'83)*, pp. 51–67.

[2] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn (1999). Information hiding—A survey. *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078.

[3] R. Chandramouli (2002). A mathematical approach to steganalysis. *Proc. SPIE*, vol. 4675, pp. 14–25.

[4] I. Avcibas, N. Memon, and B. Sankur (2003). Steganalysis using image quality metrics. *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 221–229.

[5] S. Katzenbeisser and F. A. P. Petitcolas.(2000). *Information Hiding Techniques for Steganography and Digital Watermarking*. Norwood, MA, Artech House.

[6] W. Bender, D. Gruhl, N. Morimot, and A. Lu (1996). Techniques for data hiding. *IBM Syst. J.*, vol. 35, no. 3/4, pp. 313–336.

[7] N. Nikolaidis and I. Pitas (1998). Robust image watermarking in the spatial domain. *Signal Process.*, vol. 66, pp. 385–403.

[8] L. M. Marvel, C. G. Boncelet Jr., and C. T. Retter (1999). Spread spectrum image steganography. *IEEE Transactionsoin Image Processing*, vol. 8, no. 8, pp. 1075–1083.

[9] Sanjay Kumar Jena, G.V.V. Krishna (2007). Blind Steganalysis: Estimation of Hidden Message Length. *International Journal of Computers, Communications & Control,* vol. II 2007.

[10] T.-S. Chen, C.-C. Chang, and M.-S. Hwang (1998). A virtual image cryptosystem based upon vector quantization. *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1485–1488.

[11] Y. K. Lee and L. H. Chen (2000). High capacity image steganographic model. *Proc. Inst. Elect. Eng., Vis. Image Signal Processing*, vol. 147, no. 3, pp. 288–294.

[12] R. Chandramouli and N. Memon (2000). A distribution detection framework for watermark analysis, *Proc. ACM Multimedia*, pp. 123–126.

[13] R. Chandramouli and N. Memon (2001). Analysis of LSB based image steganography techniques. *Proc. Int. Conf. Image Processing*, pp. 1019–1022.

[14] J. Fridrich and M. Goljan (2002). Practical steganalysis of digital images-state of the art. *Proc. SPIE*, vol. 4675, pp. 1–13.

[15] S. Voloshynoskiy, A. Herrigel, Y. Rytsar, and T. Pun (2002). StegoWall: Blind statistical detection of hidden data. *Proc. SPIE*, vol. 4675, pp. 57–68.

[16] X. Kong, T. Zhang, X. You, and D. Yang (2002). A new steganalysis approach based on both complexity estimate and statistical filter.*Proc. IEEE Pacific-Rim Conf. on Multimedia*, vol. LNCS 2532, pp. 434–441.

[17] Gökhan Gül, Ahmet Emir Dirik, and Ismail Avcıbas (2007). Steganalytic Features for JPEG Compression-Based Perturbed Quantization. *IEEE Signal Processing Letters*, vol. 14, No. 3, pp. 205-208.

[18] Andrew D. Ker. (2007). A Weighted Stego Image Detector for Sequential LSB Replacement. Proc. 2007 International Workshop on Data Hiding for Information and Multimedia Security attached to IAS 07. IEEE Computer Society Press.

[19] W.-N. Lie, G.-S. Lin, and C.-L. Wu (2000). Robust image watermarking on the DCT domain. *Proc. IEEE Int. Symp. Circuits and Systems*, pp. I228–I231.

[20] J. Huang and Y. Q. Shi (1998). Adaptive image watermarking scheme based on visual masking. *Electron. Lett.*, vol. 34, no. 8, pp. 748–750.

[21] T. Ogihara, D. Nakamura, and N. Yokoya (1996). Data embedding into pictorial with less distortion using discrete cosine transform. *Proc. ICPR'96*, pp. 675–679.

[22] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon (1997). Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687.

[23] C. I. Podilchuk and Z. Wenjun (1998). Image-adaptive watermarking using visual models. *IEEE J. Select. Areas Commun.*, vol. 16, no. 4, pp. 525–539.

[24] Q. Cheng and T. S. Huang (2001). An additive approach to transform-domain information hiding and optimum detection structure. *IEEE Transactions on Multimedia*, vol. 3, no. 3, pp. 273–284.

[25] F. P. Gonzalez, F. Balado, and J. R. H. Martin (2003). Performance analysis of existing and new methods for data hiding with known-host information in additive channels. *IEEE Transactions on Signal Process.*, vol. 51, no. 4, pp. 960–980.

[26] Y.-S. Kim, O.-H. Kwon, and R.-H. Park (1999). Wavelet based watermarking method for digital images using the human visual system. *Electronic Letters.*, vol. 35, no. 6, pp. 466–468.

[27] C. Manikopoulos, Y.-Q. Shi, S. Song, Z. Zhang, Z. Ni, and D. Zou(2002). Detection of block DCT-based steganography in gray-scale images. *Proc. 5th IEEE Workshop on Multimedia Signal Processing*, pp. 355–358.

[28] R. Chandramouli (2002). A mathematical approach to steganalysis. *Pro SPIE*, vol. 4675, pp. 14–25.

[29] R. O. Duda, P. E. Hart, and D. G. Stork (2001). *Pattern Classification*. New York: Wiley-Interscience.

[30] J. J. Harmsen and W. A. Pearlman (2003). Steganalysis of additive noise modelable information hiding. *Proc. SPIE*, pp. 21–24.

[31] Wen-Nung Lie and Guo-Shiang Lin (2005). A Feature-Based Classification Technique for Blind Image Steganalysis. *IEEE Transactions on Multimedia*, vol. 7, no. 6.

[32] S. Daly (1993). The visible differences predictor: An algorithm for the assessment of image fidelity. *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, pp. 179–205.

[33] C. E. Halford, K. A. Krapels, R. G. Driggers, and E. E. Burroughs (1999). Developing operational performance metrics using image comparison metrics and the concept of degradation space. *Opt. Eng.*, vol. 38, pp. 836–844.
[34] PictureMarc, Embed Watermark, v 1.00.45, Digimarc Corp.
[35] M. Kutter and F. Jordan. JK-PGS (Pretty Good Signature). [Online]. Available: http://ltswww.epfl.ch/~kutter/watermarking/JK_PGS.html, Last retrieved 10, September 2008.
[36] A. Brown. S-tools version 4.0. [Online]. Available: http://members. tripod.com/steganography/stego/s-tools4.html, Last retrieved 11, September 2008.
[37] Steganos II Security Suite.. [Online]. Available: http://www.steganos.com/english/steganos/download.htm, Last retrieved 12, September 2008.
[38] J. Korejwa. Jsteg shell 2.0. [Online]. Available: http://www.tiac.net/users/korejwa/steg.htm, Last retrieved 13, September 2008.
[39] Y.-S. Kim, O.-H. Kwon, and R.-H. Park, "Wavelet based watermarking method for digital images using the human visual system," *Electron. Lett.*, vol. 35, no. 6, pp. 466–468, 1999.
[40] Images. [Online]. Available: http:// www.cl.cam.ac.uk/~fapp2/watermarking/benchmark/image_database.html. Last retrieved 10, September 2008.
[41] Fang Min A Novel Intrusion Detection Method Based on Combining Ensemble Learning with Induction-Enhanced Particle Swarm Algorithm *IEEE Third International Conference on Natural Computation (ICNC 2007)* .
[42] http://jgap.sourceforge.net/, Last retrieved 20, September 2008.
[43] Dan Pelleg and Andrew Moore (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters.*ICML 2000.*
[44] Der-Chyuan Lou, Chih-Lin Lin, and Chiang-Lung Liu (2007). Universal steganalysis scheme using support vector machines. *Optical Engineering*. vol. 46, 117002.
[45] Benjamin Rodriguez, Gilbert Peterson and Kenneth Bauer (2008). Fusion of Steganalysis Systems Using Bayesian Model Averaging. *IFIP International Federation for Information Processing* Springer Verlag, 2008.

Steganography Combining Data Decomposition Mechanism and Stego-coding Method

Xinpeng Zhang, Shuozhong Wang and Weiming Zhang
School of Communication and Information Engineering, Shanghai University, China
E-mail: {xzhang,shuowang,weimingzhang}@shu.edu.cn

Keywords: steganography, data decomposition, embedding efficiency

Received: September 1, 2008

A novel steganographic scheme based on data decomposition and stego-coding mechanisms is proposed. In this scheme, a secret message is represented as a sequence of digits in a notational system with a prime base. Each digit block is decomposed into a number of shares. By using stego-coding technique, these shares are then embedded in different cover images respectively. In each cover, a share is carried by a group of cover pixels and, at most, only one pixel in the group is increased or decreased by a small magnitude. That implies a high embedding efficiency, and therefore distortion introduced to the covers is low, leading to enhanced imperceptibility of the secret message. A further advantage of the scheme is that, even a part of stego-images are lost during transmission, the receiver can still extract embedded messages from the surviving covers.

Povzetek: Predstavljena je nova steganografska metoda.

1 Introduction

Steganography is a branch of information hiding that aims to send secret messages under the cover of a carrier signal. While many steganographic methods have been proposed for various types of cover media in recent years, techniques of steganalysis have also rapidly developed to detect the presence of secret messages based on statistical abnormality caused by data hiding [1, 2]. Generally speaking, the more the embedded data, the more vulnerable the system will be to the steganalytic attempts. When a multimedia product is under suspicion, the channel warden may refuse to transmit it, and the source of the message can be tracked. As a countermeasure, the data-hider always tries to improve statistical imperceptibility of the hidden message.

An important technique to improve imperceptibility is to reduce the amount of alterations to be introduced into the cover for hiding the same quantity of data, in other words, to improve embedding efficiency. For example, Matrix encoding uses less than one change of the least significant bit (LSB) in average to embed l bits into 2^l-1 pixels [3]. In this way, distortion is significantly lowered compared to a plain LSB technique in which secret bits simply replace the LSB. Further, some effective encoding methods derived from the cyclic coding have been described [4], and the matrix encoding can be viewed as a special case. In [5], two methods based on random linear codes and simplex codes are developed for large payloads. Another method, termed running coding, can also be performed on a data stream derived from the host in a dynamically running manner [6]. All the above-mentioned stego-coding techniques are independent of any particular cover-bit-modification approaches. For example, if a stego-coding method is used in the LSB plane of an image, adding 1 to a pixel is equivalent to subtracting 1 from the pixel to flip its LSB for carrying the secret message. In addition, we [7] and Fridrich et al. [8] independently presented a same method with better performance, termed respectively exploiting modification direction (EMD) and grid coloring (GC for short). Using this method, $\log_2(2q+1)$ secret bits are embedded into q cover pixels and, at most, only one pixel is increased or decreased by 1. In [8], a data-hiding approach incorporating GC with Hamming-derived steganographic encoding technique is also studied, which in fact is a special case of GC. We also applied the wet paper codes to steganography to further increase embedding efficiency [9, 10 11].

Since stego-covers may be lost due to an active warden or poor channel conditions, a steganographic system capable of resisting interference is also desired to the data-hider. This paper proposes a novel steganographic scheme by introducing a data decomposition mechanism together with stego-coding techniques, such as running coding and EMD embedding methods. In this way, the secret message is inserted into a number of cover images with high embedding efficiency. Even a part of stego-images are missing, one can still extract the hidden message from the remaining covers.

The rest of this paper is organized as follows. Section 2 introduces the related stego-coding methods. The proposed scheme is described in Section 3 and 4. Then, the experimental results are shown in Section 5. Finally, we conclude in Section 6.

2 Related Stego-coding methods

In stego-coding methods, a number of patterns of cover data are used to represent a type of secret data, and the data-hider modifies the original cover data to the nearest pattern mapping the secret data to be hidden. This way, by changing a small part of cover data, a fairly large amount of secret data can be embedded. In this section, we briefly review the related techniques including running coding and EMD embedding methods.

2.1 Running coding

With running coding method [6], each secret bit is represented by a series of consecutive cover bits, and each available cover bit also relates to several consecutive secret bits. In other words, the secret message is embedded as a data stream, and each cover-bit-alteration is used to embed several consecutive secret bits.

Assume that the secret message to be hidden contains K bits: $[x_1, x_2, \ldots, x_K]$, and the available LSB for carrying the secret message are $[b_{1,1}, b_{1,2}, \ldots, b_{1,T}; b_{2,1}, b_{2,2}, \ldots, b_{2,T}; \ldots; b_{K,1}, b_{K,2}, \ldots, b_{K,T}]$, where T is an integer power of 2 ($T = 2^t$). A binary generating matrix **G** sized $(t+1) \times T$ is first constructed. Denote the elements in **G** as $g(i, j)$, where $1 \le i \le t+1$ and $1 \le j \le T$. Assign all the elements in the first row as '1' and make all the 2^t columns in **G** mutually different. For example, the generating matrix of the 4th running coding is

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \tag{1}$$

According to the original host data and the generating matrix **G**, calculate

$$y_v = \sum_{i=1}^{t+1} \sum_{j=1}^{T} \left[g(i, j) \cdot b_{v-i+1, j} \right] \bmod 2, \quad v = 1, 2, \ldots, K \tag{2}$$

where $b_{v-i+1,j} = 0$ if $v-i+1 \le 0$. The data-hider can use a small number of alterations in these host bits to make the values of y_vs equal to the secret bits x_vs. Let

$$z_v = \begin{cases} 0, & \text{if } x_v = y_v \\ 1, & \text{if } x_v \ne y_v \end{cases} \tag{3}$$

Orderly arrange all z_vs to form a vector $\mathbf{Z} = [z_1, z_2, \ldots, z_K]^T$, and divide **Z** into a set of sub-vectors in the following way:

1. Scan the vector **Z** from the beginning to the end;
2. If the encountered bit is '0', define this '0' as a sub-vector containing only one element;
3. If the bit is '1', define this '1' together with the following t bits as a sub-vector with a length $(t+1)$. Obviously, the sub-vector in this case must be identical to one of the columns in **G**.

That means **Z** is segmented into a sequence of sub-vectors, each being either a column of the generating matrix **G** or a single zero. According to (2), flipping the value of host bit $b_{v,j}$ will change the value of y_{v+i-1} if $g(i, j) = 1$ ($1 \le i \le t+1$, $1 \le j \le T$). Thus, we can modify only one host bit to change the values of several y_vs. Assume that a sub-vector $[z_v, z_{v+1}, \ldots, z_{v+t}]^T$ is same as the j-th column of **G**. By flipping the value of host bit $b_{v,j}$, the data-hider may make $[y'_v, y'_{v+1}, \ldots, y'_{v+t}]$ identical to $[x_v, x_{v+1}, \ldots, x_{v+t}]$, where $[y'_v, y'_{v+1}, \ldots, y'_{v+t}]$ are obtained from the modified host bits according to (2). This way, the secret data can be embedding using a small number of bit-alterations.

2.2 EMD embedding

EMD embedding [7] is an alternative method for inserting secret data into a certain cover image with a high embedding efficiency. Using this method, each symbol in notational system with an odd base will be carried by a group of pixels, and, at most, only one pixel is increased or decreased by 1.

Denote a secret symbol in notational system with an odd base $(2q+1)$ as s, and the gray values of pixels in a group as $g_1, g_2, \ldots, g_q$. Calculate the extraction function f as a weighted sum modulus $(2q+1)$

$$f(g_1, g_2, \ldots, g_q) = \sum_{i=1}^{q} (g_i \cdot i) \bmod (2q+1) \tag{4}$$

Consider the vector $[g_1, g_2, \ldots, g_q]$ as a hyper-cube in q-dimensional space. The extraction function must have the following two properties: 1) values of the extraction function on all hyper-cubes fall in the interval [0 $2q$], and 2) the values of f on any hyper-cube and its $2q$ neighbors are mutually different. This implies that a symbol in the $(2q+1)$-ary notational system can be carried by a pixel-group, and, at most, only one pixel will be increased or decreased by 1. If the symbol s equals the extraction function of the original corresponding pixel-group, no modification is needed. When $s \ne f$, calculate $u = s - f \bmod p$. If u is no more than q, increase the value of g_u by 1, otherwise, decrease the value of g_{p-u} by 1.

For example, considering an original pixel-group [137 139 141 140] with $q = 4$, $f = 3$ and a corresponding symbol 4 in 9-ary notational system, a data-hider can calculate $u = 1$, so he can increase the gray value of the first pixel by 1 to produce the stego-pixels [138 139 141 140]. If the symbol to be hidden is 0, $u = 8$ can be calculated and the gray value of the forth pixel will be decrease by 1 to yield [137 139 141 139].

3 Data embedding procedure

In this proposed scheme, a secret message is firstly represented as a series of shares according to a data

decomposition mechanism and various indices, and the shares corresponding to different indices are respectively inserted into different cover images. Then, a generalized running coding or EMD method is employed to keep stego-induced distortion at a low level, and redundancy in the shares ensures that one can recover the original secret message from a part of stego-covers.

3.1 Data decomposition

At the beginning, a data-hider converts a secret message into a digit sequence in a notational system with an odd and prime base p, such as 3, 5, 7, 11, etc. If the secret message is a binary stream, it can be segmented into many pieces, each having L_1 bits, and the decimal value of each secret piece is represented by L_2 digits in a p-ary notational system, where

$$L_1 = \lfloor L_2 \cdot \log_2 p \rfloor \tag{5}$$

For example, the binary message (1001 1101 0110) can be rewritten as (14 23 11) in 5-ary notational system when $L_1 = 4$ and $L_2 = 2$. Thus, the rate of redundancy in the digit sequence

$$R_{\mathrm{R}} = 1 - \frac{L_1}{L_2 \cdot \log_2 p} < \frac{1}{L_1 + 1} \tag{6}$$

With large L_1 and L_2, R_{R} is very close to 0, therefore can be ignored. So, the secret message is regarded as a digit sequence in p-ary notational system in the following discussion.

Then, the data-hider segments the secret digit sequence into a series of blocks, each of which contains m digits. Denote the number of blocks as K, and the block as $\{d_{k,1}, d_{k,2}, \ldots, d_{k,m}\}$ (k = 1, 2, …, K). Inspired by [12], decompose each secret block into n shares, $\{s_{k,1}, s_{k,2}, \ldots, s_{k,n}\}$, in the following way,

$$\begin{bmatrix} s_{k,1} & s_{k,2} & \ldots & s_{k,n} \end{bmatrix} = \begin{bmatrix} d_{k,1} & d_{k,2} & \ldots & d_{k,m} \end{bmatrix} \cdot \mathbf{A} \tag{7}$$

where $m \le n \le p$,

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ a_1 & a_2 & a_3 & \cdots & a_n \\ a_1^{\,2} & a_2^{\,2} & a_3^{\,2} & \cdots & a_n^{\,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_1^{\,m-1} & a_2^{\,m-1} & a_3^{\,m-1} & \cdots & a_n^{\,m-1} \end{bmatrix} \tag{8}$$

and the symbol "·" in (7) is a multiplication operator with a modulus p. We call a_1, a_2, …, and a_n as indices. All indices lie between [0 p–1] and are mutually different. For example, assuming p = 5, n = 4, m = 3, and

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 4 & 0 & 1 \\ 4 & 1 & 0 & 1 \end{bmatrix} \tag{9}$$

a digit block {2, 4, 1} can be represented as 4 shares: 4, 4, 2, and 2. Note that the shares are also within the p-ary notational system.

Collect all shares and divide them into n sets $\{s_{1,1}, s_{2,1}, \ldots, s_{K,1}\}$, $\{s_{1,2}, s_{2,2}, \ldots, s_{K,2}\}$, …, $\{s_{1,n}, s_{2,n}, \ldots, s_{K,n}\}$, each of which contains K shares. Then, the n share-sets and their corresponding indices will be embedded into n cover images, respectively. Since the indices are within [0 p–1], they can also be regarded as symbols in the p-ary notational system. In other words, each cover image will be used to conceal (K+1) symbols in the p-ary notational system, $s_{1,t}, s_{2,t}, \ldots, s_{K,t}$ and a_t (t = 1, 2, …, n).

3.2 Generalized running coding

In order to improve steganographic imperceptibility, we use stego-coding technique to lower the distortion caused by data embedding. As mentioned above, running coding in [6] is only suitable for binary data-hiding system. This subsection generalizes the running coding method, so that the secret symbols in the p-ary notational system can be carried by a sequence of gray-pixel-value of cover image. Actually, for each cover image, either generalized running coding or EMD embedding can be employed to embed the shares and index.

In the generalized running coding method, each secret symbol in the p-ary notational system is represented by a series of consecutive cover values, and each cover value also relates to several consecutive secret symbols. Thus, a data-hider can modify a selected cover value to embed several secret symbols, so that the distortion introduced into the cover signal is significantly reduced, which also means the data-hiding efficiency is increased.

For convenience, we denote the (K+1) symbols in the p-ary notational system to be embedded into a certain cover as $[x_1, x_2, \ldots, x_{K+1}]$. Pseudo-randomly select (K+1)·T pixels in cover image according to a secret key, and denote the gray-levels of them as $[h_{1,1}, h_{1,2}, \ldots, h_{1,k+1}; h_{2,1}, h_{2,2}, \ldots, h_{2,k+1}; \ldots; h_{T,1}, h_{T,2}, \ldots, h_{T,k+1}]$. That means the number of host values is T times of that of secret symbols.

3.2.1 The case of $T = p^t$

Firstly, we discuss the case that T is an integer power of p ($T = p^t$). Inspired from [6], construct a generating matrix **G** sized (t+1) × T. Denote the elements in **G** as $g(i, j)$, and assign them according to the following principle,

1. All elements are integers within [0, p–1].
2. All elements in the first row are 1.
3. All p^t columns in **G** are different.

For example, when p=3 and k=9,

$$\mathbf{G}=\begin{bmatrix}1&1&1&1&1&1&1&1&1\\0&1&2&0&1&2&0&1&2\\0&0&0&1&1&1&2&2&2\end{bmatrix} \quad (10)$$

From the original host data and the generating matrix **G**, calculate

$$y_v=\sum_{i=1}^{t+1}\sum_{j=1}^{T}\left[g(i,j)\cdot h_{v-i+1,j}\right] \bmod p, \quad v=1,2,\ldots,K+1 \quad (11)$$

where $h_{v-i+1,j}=0$ if $v-i+1\le 0$. That means the value of y_v is determined by $(t+1)\times T$ host values $h_{v-t,1}$, $h_{v-t,2}$, …, $h_{v-t,T}$, $h_{v-t+1,1}$, $h_{v-t+1,2}$, …, $h_{v-t+1,T}$, …, $h_{v,1}$, $h_{v,2}$, …, $h_{v,T}$. Similarly, we will modify a small number of host values to make each y_v equal to the corresponding secret x_v. Let

$$z_v=x_v-y_v \bmod p, \quad v=1,2,\ldots,K+1 \quad (12)$$

Arrange all the z_v to form a vector $\mathbf{Z}=[z_1, z_2, \ldots, z_{K+1}]^{\mathrm{T}}$, and then divide **Z** into a set of sub-vectors in the following way:

1. Scan the vector **Z** from the beginning to the end;
2. If the encountered digit is '0', define this '0' as a sub-vector containing only one element;
3. If the encountered digit z_v is not '0', define this digit together with the following t digits as a sub-vector with a length $(t+1)$. Because all the elements in the first row of **G** are 1 and p is prime, the sub-vector in this case must be equal to product of z_v and one of the columns in **G** with modulus p.

Equation (11) indicates that any change on host value $h_{v,j}$ will affect the values of y_v, y_{v+1}, …, y_{v+t}. Thus, we can modify only one host value but embed several secret symbols. Assume that z_v is not 0 and the sub-vector $[z_v, z_{v+1}, \ldots, z_{v+t}]^{\mathrm{T}}$ equals the product of z_v and the j-th column of **G** with modulus p. Either increasing the value of $h_{v,j}$ by z_v or decreasing the value of $h_{v,j}$ by $p-z_v$ will make $[y'_v, y'_{v+1}, \ldots, y'_{v+t}]$ identical to $[x_v, x_{v+1}, \ldots, x_{v+t}]$, where $[y'_v, y'_{v+1}, \ldots, y'_{v+t}]$ are obtained from the modified host values according to (11). In this way, all secret symbols can be embedded by performing the similar operation for all sub-vectors.

Consider that, for example, a host value sequence with length 21 for carrying secret message is [40 187 99, 93 231 19, 82 78 33, 11 176 134, 56 27 121, 31 249 83, 90 111 24], and 7 secret digits in ternary system, implying p = 3, [2110100]. Because $T = 21/7 = 3^1$, construct a generating matrix **G**.

$$\mathbf{G}=\begin{bmatrix}1&1&1\\0&1&2\end{bmatrix} \quad (13)$$

From (11), the vector **Y** is [2200021], so that **Z** = $[0210112]^{\mathrm{T}}$. Append '0' to the end of **Z** and segment it into 5 sub-vectors: [0], $[21]^{\mathrm{T}}$, [0], $[11]^{\mathrm{T}}$, and $[20]^{\mathrm{T}}$. Note that appending '1' or '2' is also allowable. Since the sub-vector $[21]^{\mathrm{T}}$ is a product of 2 and the 3rd column of **G** with modulus 3, the data-hider should increase $h_{2,3}$ by 2 or decrease $h_{2,3}$ by 1. To lower the distortion, the value of $h_{2,3}$ is decreased by 1. Similarly, $h_{5,2}$ should be increased by 1, and $h_{7,1}$ decreased by 1. So, the stego-sequence [40 187 99, 93 231 18, 82 78 33, 11 176 134, 56 28 121, 31 249 83, 89 111 24] are produced. In this way, 7 symbols in ternary system are embedded by adding/subtracting 1 to/from three pixels. On the receiving side, a simple calculation of (11) can recover the embedded data, when the receiver knows the values of p, T and **G**.

A ratio between the number of embedded bits and the distortion energy caused by data hiding, E, is used to indicate the embedding efficiency. As mentioned above, a sub-vector must be '0', or contains $(t+1)$ elements and begins with a non-zero digit. Since the values of z are also uniformly distributed within $[0, p-1]$, the probability of the former case is $1/p$, while that of the later case is $(p-1)/p$. In the former case, the secret symbol has been represented and any modification is needless, while in the later case, a modification on one host value is made to embed $(t+1)$ digits. In average, $(p\cdot t-t+p)/p$ secret symbols are embedded by modifying $(p-1)/p$ host values. As p is odd, the modifications on host values are within $[(1-p)/2, (p-1)/2]$, thus,

$$E=\frac{\left[(p\cdot t-t+p)/p\right]\cdot\log_2 p}{\frac{p-1}{p}\cdot\frac{2}{p-1}\sum_{u=1}^{(p-1)/2}u^2}=\frac{(p\cdot t-t+p)\cdot\log_2 p}{2\cdot\sum_{u=1}^{(p-1)/2}u^2} \quad (14)$$

which is significantly larger than 2, the embedding efficiency of plain LSB replacement/matching method.

3.2.2 The case of $p^t < T < p^{t+1}$

If T is not an integer power of p, i.e., $p^t < T < p^{t+1}$, a generating matrix **G** sized $(t+2)\times T$ can also be constructed as follows:

1. All elements are integers within $[0, p-1]$.
2. All elements in the first row are 1.
3. All $g(t+2, j)$ are 0 where $1 \le j \le p^t$.
4. The columns in **G** are different.

For instance, when p=5 and T=8,

$$\mathbf{G}=\begin{bmatrix}1&1&1&1&1&1&1&1\\0&1&2&3&4&0&1&2\\0&0&0&0&0&1&1&1\end{bmatrix} \quad (15)$$

Similarly, y_s and z_s can be computed from (11) and (12). Vector **Z** can be segmented into sub-vectors in the following way:

1. Scan the vector **Z** from the beginning to the end;

2. If the encountered digit is 0, designate this digit as a sub-vector and denote this type of sub-vector as SV_0;

3. If the encountered digit z_v is not 0, and the vector containing z_v and the following (t+1) digits is equal to a product of z_v and one of the columns in **G** with modulus p, designate the (t+2) digits as a sub-vector, and denote this type of sub-vector as SV_1;

4. If the encountered digit z_v is not 0, and the vector containing z_v and the following (t+1) digits is not the same as the product of z_v and any column in **G** with modulus p, designate z_v and the following t digits as a sub-vector with length (t+1), and denote this type of sub-vector as SV_2. In this case, the next sub-vector must start with a non-zero digit.

Denoting the up-left sub-matrix of **G** sized (t+1) × p^t as **G'**, an SV_2 sub-vector must be equal to a product of z_v and one of the columns in **G'** with modulus p. For an SV_0 sub-vector, no modification is needed. But for an SV_1 sub-vector equal to a product of its first element z_v and the j-th column in **G** with modulus p or an SV_2 sub-vector equal to a product of its first element z_v and the j-th column in **G'** with modulus p, the data-hider should increase the value of $h_{v,j}$ by z_v or decrease it by $p-z_v$.

For example, consider 80 host values available for carrying secret message and 10 secret symbols in 5-ary notational system [2314023343]. Because T = 80/10 = 8, we can construct a generating matrix **G** as in (14). Assuming the vector **Y** = [2130204131] can be calculated according to (11), thus **Z** = $[0234324212]^T$. Append a '0' to the end of **Z** and segment it into 5 sub-vectors [0], $[23]^T$, $[43]^T$, $[242]^T$, and $[120]^T$. Following the rule of modification as described above, the data-hider should increase $h_{2,5}$ by 2, decrease $h_{4,3}$ by 1, increase $h_{6,8}$ by 2, and increase $h_{9,3}$ by 1 so as to embed the secret data.

Now we calculate the embedding efficiency. As mentioned, a sub-vector following an SV_2 must be SV_1 or SV_2. Therefore, any sequence of sub-vectors between the end of an SV_1 and the end of the next SV_1 must be in the form of {0, 0, …, 0, SV_2, SV_2, …, SV_2, SV_1}. Denote the numbers of consecutive 0s and SV_2 sub-vectors as l_0 and l_2 (l_0, l_2 = 0, 1, 2, …), respectively. In the above example, the pattern of the first 4 sub-vectors is {0, SV_2, SV_2, SV_1} (l_0 = 1, l_2 = 2). Denoting

$$\eta = T / p^{t+1} \tag{16}$$

the probability of a sub-vector sequence with l_0 '0's, l_2 SV_2 sub-vectors, and an SV_1 is

$$P(l_0, l_2) = \frac{p-1}{p^{l_0+1}} \cdot (1-\eta)^{l_2} \cdot \eta \tag{17}$$

For the sub-vector sequence, a total of [l_0 + l_2·(t+1) + t + 2] secret digits are embedded by modifying (l_2+1) host values. Thus,

$$E = \frac{\left[(p \cdot t - t + p)/p\right] \cdot \log_2 p}{\frac{p-1}{p} \cdot \frac{2}{p-1} \sum_{u=1}^{(p-1)/2} u^2} = \frac{(p \cdot t - t + p) \cdot \log_2 p}{2 \cdot \sum_{u=1}^{(p-1)/2} u^2} \tag{18}$$

3.3 Application of EMD embedding

When using EMD embedding for concealing (K+1) p-ary symbols, including K shares and an index, into a cover image, pseudo-randomly select (K+1)·q pixels according to a secret key, and divide them into (K+1) pixel-groups, each of which contains q pixels. Here,

$$q = \frac{p-1}{2} \tag{19}$$

Then, we map the (K+1) symbols to the pixel-groups in a one-by-one manner. Using EMD embedding method, each symbol in the p-ary notational system is carried by a group of pixels, and, at most, only one pixel is increased or decreased by 1. As analysed in [7], the embedding efficiency is

$$E = \frac{p \cdot \log_2 p}{p-1} \tag{20}$$

which is also significantly larger than 2, the embedding efficiency of plain LSB replacement/matching method.

Note that both generalized running coding method and EMD embedding method can be used to gain a high embedding efficiency, and the stego-coding techniques used in different covers may be different. So, an additional bit that labels the stego-coding technique used in a certain cover, e.g., '0' for generalized running coding and '1' for EMD embedding, as well as the values of p and K, should be embedded into the cover image itself. If running coding is executed, the parameter T should be also hidden in the corresponding stego-image. Actually, LSB replacement method can be used to embed the additional secret information into cover images, and the embedding positions may be determined by the secret key.

4 Data extracting procedure

As mentioned in the previous section, the secret message is embedded into n cover images, and all the n stego-images are sent through a poor channel. Assume the stego-images may be lost in the channel. If the number of received stego-images is no less than m, one can still recover the original secret message using m arbitrary stego-images.

For each received stego-image, the receiver first extracts the embedded label-bit of stego-coding

technique and the values of parameter p, K and T. If generalized running coding is used in the cover, the receiver selects $(K+1)\cdot T$ pixels according to the same secret key, and calculates the K embedded shares, $s_{1,t}$, $s_{2,t}$, …, $s_{K,t}$, and the embedded index a_t (t = 1, 2, …, n) using (11). If EMD method is used, the receiver selects $(K+1)\cdot q$ pixels according to the same secret key, and divides them into $(K+1)$ pixel-groups. Then, he calculates the extraction function of stego-pixel-groups to obtain the $(K+1)$ embedded symbols. This way, the receiver may extract a total of $K\cdot m$ shares and m indices from m received stego-images. Denote the extracted indices as $a_{t_1}, a_{t_2}, \ldots, a_{t_m}$. For each digit block, Equation (7) can be reformulated as

$$\begin{bmatrix} s_{k,t_1} & s_{k,t_2} & \cdots & s_{k,t_m} \end{bmatrix} = \begin{bmatrix} d_{k,1} & d_{k,2} & \cdots & d_{k,m} \end{bmatrix} \cdot \mathbf{A}_t \tag{21}$$

The left side is m shares extracted from different stego-images, and $\mathbf{A}_t$ is an $m\times m$ matrix made up of m columns of $\mathbf{A}$ corresponding to the extracted indices

$$\mathbf{A}_t = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ a_{t_1} & a_{t_2} & a_{t_3} & \cdots & a_{t_m} \\ a_{t_1}^{\,2} & a_{t_2}^{\,2} & a_{t_3}^{\,2} & \cdots & a_{t_k}^{\,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{t_1}^{\,m-1} & a_{t_2}^{\,m-1} & a_{t_3}^{\,m-1} & \cdots & a_{t_k}^{\,m-1} \end{bmatrix} \tag{22}$$

As well known, $\mathbf{A}_t$ is a Vandermonde matrix, and its determinant is

$$|\mathbf{A}_t| = \prod_{i>j} \left(a_{t_i} - a_{t_j} \right) \tag{23}$$

Since all a_t are mutually different, $|\mathbf{A}_t|$ can not be zero. That means $\mathbf{A}_t$ must have an inverse with the modulus p,

$$\mathbf{A}_t^{-1} = \frac{\mathbf{A}_t^*}{|\mathbf{A}_t|} \tag{24}$$

where $\mathbf{A}_t^*$ is the adjoint matrix of $\mathbf{A}_t$. So, the secret digit block can be restored by using m extracted shares and the inverse matrix of $\mathbf{A}_t$

$$\begin{bmatrix} d_{k,1} & d_{k,2} & \cdots & d_{k,m} \end{bmatrix} = \begin{bmatrix} s_{k,t_1} & s_{k,t_2} & \cdots & s_{k,t_m} \end{bmatrix} \cdot \mathbf{A}_t^{-1} \tag{25}$$

For example, for the matrix $\mathbf{A}$ in Equation (9), the indices extracted from three stego-images are 2, 4, and 1, the receiver can obtain

$$\mathbf{A}_t = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 1 \\ 4 & 1 & 1 \end{bmatrix} \tag{26}$$

and its inverse

$$\mathbf{A}_t^{-1} = \begin{bmatrix} 3 & 0 & 2 \\ 2 & 2 & 1 \\ 1 & 3 & 2 \end{bmatrix} \tag{27}$$

If the three extracted shares are respectively 4, 4, and 2, the digit block is then calculated

$$\begin{bmatrix} 4 & 2 & 2 \end{bmatrix} \cdot \mathbf{A}_t^{-1} = \begin{bmatrix} 2 & 4 & 1 \end{bmatrix} \tag{28}$$

After calculating all the digit blocks, the receiver can concatenate them to retrieve the secret message.

5 Experiment results

In the experiment, a secret message with 3.6×10^5 bits was first converted into 1.3×10^5 digits in 7-ary notational system. After segmenting the secret digit sequence into a series of blocks with length 4, each digit block was decomposed into 6 shares using Equation (7). That means p = 7, n = 6, and m = 4. Then, the 6 share-sets and their corresponding indices were embedded into 6 cover images sized 512×512. In other words, each cover image was used to conceal 3.2×10^4 symbols in 7-ary notational system. Then, we produced 6 stego-images, three of them produced by using generalized running coding with T = 8, and the rest three by using EMD method. Figure 1 shows 4 stego-images among them. In each stego-image produced by generalized running coding, the number of changed pixels was 1.5×10^4 with the modifications within [−3, 3], and the value of PSNR due to data hiding is 53.9 dB, indicating the visual imperceptibility. In each stego-image produced by EMD method, there were 2.8×10^4 pixels increased/decreased by 1, and the value of PSNR is 57.8 dB. Since only a small part of cover pixels were increased or decreased by small magnitudes, it is difficult to detect the presence of secret message. Actually, if one receives no less than four stego-imges among all the six, he can always recover the secret message using the data extracted from the received stego-images.

We also attempted to conceal the same secret message using various steganographic methods, respectively. With a plain LSB embedding method, two cover images with a size of 512×512 were required to provide sufficient LSBs for accommodating the secret data. In this case, the values of PSNR of the stego-images are 52.1 dB. When employing the original running coding and assigning the parameter T = 2, the secret message were carried by three 512×512 cover images with PSNR 52.9 dB. Alternatively, after

converting the secret message into a series of 7-ary symbols (q = 3), we exploited EMD embedding method to conceal them into four cover images. Here, PSNR due to data-hiding is 57.8 dB. When the three methods are used, all stego-images are necessary for data extraction at receiver side. Table 1 shows the performance comparison between the three methods and the proposed scheme. Note that, although the proposed scheme exploits more cover images, the steganographic distortion is lower and the secret message can be transmitted through a severe channel.

(a) (b) (c) (d)

Figure 1: Four stego-images. While (a) and (b) are produced by running coding with PSNR 53.9 dB, (c) and (d) are produced by EMD embedding with PSNR 57.8 dB

Table 1. Performance comparison between various steganographic techniques

Steganographic technique	Number of cover images	PSNR due to data embedding	Condition for data extraction
LSB method	2	52.1 dB	All stego-images must be received
Running coding (T = 2)	3	52.9 dB	
EMD embedding (q = 3)	4	57.8 dB	
Proposed scheme (Data decomposition & Stego-coding)	6	53.9 dB and 57.8 dB	Any four stego-imges among all the six are received

6 Conclusion

In the proposed steganographic scheme, a data decomposition mechanism is introduced to represent the secret message as a number of share sets, and both generalized running coding and EMD embedding methods can be employed to embed the shares into different cover images with high efficiency. This way, even though a part of stego-images are lost in a severe channel, one can still recover the hidden message from the received covers.

Two aspects deserve further study in the future. One is error correcting capability of the proposed scheme. In many applications, the receiver may obtain most stego-images with channel noise. Since there is redundancy between the share-sets embedded into different covers, the receiver can still restore the secret message when the noise is not too serious. On the other hand, since it is necessary to distribute the payloads into cover images according to their various sizes, a technique for decomposing the secret message into share-sets with different amounts is desired, i.e., a generalization of the data-decomposing mechanism should be developed.

Acknowledgement

This work was supported by the Natural Science Foundation of China (60872116, 60773079, 60502039), the High-Tech Research and Development Program of China (2007AA01Z477), and the China Postdoctoral Science Foundation Funded Project (20070420096).

References

[1] H. Wang, and S. Wang, "Cyber Warfare: Steganography vs. Steganalysis," *Communication of ACM*, **47**(10), pp.76-82, 2004.

[2] J. Fridrich, and M. Goljan, "Practical Steganalysis of Digital Images — State of the Art," in *Security and Watermarking of Multimedia Contents IV*, *Proceedings of SPIE*, **4675**, San Jose, USA, Jan. 2002, pp.1–13.

[3] A. Westfeld, "F5 — A Steganographic Algorithm," in *4th International Workshop on Information Hiding*, *Lecture Notes in Computer Science*, **2137**, Springer-Verlag, 2001, pp. 289–302.

[4] M. Dijk, and F. Willems, "Embedding Information in Grayscale Images," in *Proc. 22nd Symp. Inform. Theory in the Benelux*, The Netherlands, 2001, pp. 147-154.

[5] J. Fridrich, and D. Soukal, "Matrix Embedding for Large Payloads," in *Security, Steganography, and Watermarking of Multimedia Contents VIII*, *Proceeding of SPIE-IS&T*, **6072**, 2006, pp. 60721W1–12.

[6] X. Zhang and S. Wang, "Dynamically Running Coding in Digital Steganography," *IEEE Signal Processing Letters*, **13**(3), pp. 165–168, 2006.

[7] X. Zhang and S. Wang, "Efficient Steganographic Embedding by Exploiting Modification Direction," *IEEE Communications Letters*, **10**(11), pp.781-783, 2006.

[8] J. Fridrich, and P. Lisonek, "Grid Colorings in Steganography," *IEEE Trans. Information Theory*, **53**(4), pp. 1547-1549, 2007.

[9] X. Zhang, W. Zhang and S. Wang, "Efficient Double-Layered Steganographic Embedding," *Electronics Letters*, **43**(8), pp. 482–483, 2007.

[10] W. Zhang, X. Zhang, and S. Wang, "A Double Layered 'Plus-Minus One' Data Embedding Scheme," *IEEE Signal Processing Letters*, **14**(11), pp. 848–851, 2007.

[11] X. Zhang, W. Zhang, and S. Wang, "Integrated Encoding with High Efficiency for Digital Steganography," *Electronics Letters*, **43**(22), pp. 1191–1192, 2007.

[12] A. Shamir, "How to Share a Secret," *Communication of ACM*, **22**(11), pp.612-613, 1979.

Blind Watermark Estimation Attack for Spread Spectrum Watermarking

Hafiz Malik
Electrical and Computer Engineering Department
University of Michigan–Dearborn, Dearborn, MI 48128, USA
E-mail: *hafiz*@umich.edu, URL: http://www-personal.engin.umd.umich.edu/∼hafiz

Keywords: Spread-spectrum watermarking, independent component analysis, blind source separation, watermark estimation, detection, decoding

Received: September 18, 2008

This paper presents an efficient scheme for blind watermark estimation embedded using additive watermark embedding methods. The scheme exploits mutual independence between the host media and the embedded watermark and non-Gaussianity of the host media for watermark estimation. The proposed scheme employs the framework of independent component analysis (ICA) and poses the problem of watermark estimation as a blind source separation (BSS) problem. Analysis of the scheme shows that the proposed detector significantly outperforms existing correlation-based blind detectors traditionally used for SS-based watermarking. The proposed ICA-based blind detection/decoding scheme has been simulated using real-world audio clips. The simulation results show that the proposed ICA-based method can detect and decode watermark with extremely low decoding bit error probability (less than 0.01) against common watermarking attacks and benchmark degradations.

Povzetek: Opisana je metoda odkrivanja vodnega tiska.

1 Introduction

Digital forgeries and unauthorized sharing of digital media have emerged as a growing concern over the last decade. The widespread use of multimedia information is aided by factors such as the growth of the Internet, the proliferation of low-cost and reliable storage devices, the deployment of seamless broadband networks, the availability of state-of-the-art digital media production and editing technologies, and the development of efficient multimedia compression algorithms. Multimedia piracy has subjected the entertainment industry to enormous annual revenue losses. For example, music industry alone claims multi-million illegal music downloads on the Internet every week. It is therefore imperative to have robust technologies to protect copyrighted digital media from illegal sharing and tampering. Traditional digital data protection techniques, such as encryption and scrambling, alone cannot provide adequate protection as these technologies are unable to protect digital content once they are decrypted or unscrambled. Digital watermarking technology complements cryptography for protecting digital content even after it is deciphered [1].

Digital watermarking refers to the process of imperceptible embedding information (watermark) into the digital object (or the host object). Existing watermarking schemes based on the watermark embedding method used can be classified into two major categories:

1. *blind embedding*, in which the watermark embedder does not exploit the host signal information during watermark embedding process. Watermarking schemes based on spread-spectrum (SS) [1, 2, 3, 4, 5] fall into this category.

2. *informed embedding*, in which the watermark embedder exploits knowledge of the host signal during watermark embedding process. Watermarking schemes based on quantization index modulation [1, 6] belong to this category.

Similarity, existing watermarking schemes based on the detection method used cab be classified into two major categories:

1. *informed detector*, which assume that the host signal is available at the detector during watermark detection process, and

2. *blind detector*, which assume that the host signal is not available at the detector for watermark detection.

Although the performance expected from a given watermarking system depends on the target application area [1], but robustness of the embedded watermark and efficient detection are desirable features of a give watermarking scheme. In addition, fidelity (or imperception) of the embedded watermark is additional requirement of perception based watermarking schemes [1]. To meet fidelity requirement, the power of the embedded watermark (watermark strength) is generally kept much lower than the host signal power.

In this paper we consider additive watermark embedding model, e.g. SS-based watermarking, where the watermark signal is added to the host signal in the marking space to

obtain the watermarked signal. Existing watermark detection schemes for SS-based watermarking generally employ statistical characterization of the host signal to develop an optimal or suboptimal watermark detector [6, 7, 8]. It is important to mention that blind watermark detectors for SS-based watermarking perform poorly as the host-signal acts as interference at the blind decoder. Therefore, nonzero decoding error probability at the blind watermark decoder even in the absence of attack-channel distortion is one of the limitations of existing blind watermark detectors for SS-based watermarking schemes.

This paper presents a novel blind watermark detection method for the blind additive watermark embedding schemes[1, 2, 3, 4, 5]. The main motivation of this paper is to design a blind detector for SS-based watermarking schemes capable of suppressing host-signal interference (or improving watermark-to-host ratio) at the detector, hence improving decoding as well as detection performance. Towards this end, the proposed detector uses ICA framework by posing watermark detection problem as a blind source separation (BSS) problem. The proposed detector models the received watermarked signal as a linear mixture of underlying independent components (the host signal and the watermark). It also assumes non-Gaussianity of the host signal. Recently, we have shown in [15, 16, 17] that the watermark estimation problem for SS-based watermarking can be modeled as that of BSS of underdetermined mixture of independent sources. Therefore, the ICA framework could be used to estimate the watermark from the watermarked signals obtained using additive embedding model.

The proposed ICA-based detector first estimates the hidden independent components (i.e., the watermark and the host signal) from the received watermarked signal using the ICA framework, and then these estimated components are used to detect the embedded watermark. We present theoretical analysis to show that the proposed ICA-based detector performs significantly better than the existing watermark detectors operating without canceling the host signal interference at the watermark detector for watermark detection [6, 7]. Simulation results also show that the proposed detector in estimation-correlation based detection settings also outperforms the normalised correlation based detector (commonly used for watermark detection in SS-based watermarking community [1, 2, 3]) operating without host interference suppression. Simulation results presented in this paper are evaluated against variety of signal manipulations and degradations applied to the watermarked media. These signal degradations include addition of colored and white noise, resampling, requantization, lossy compression, filtering, time- and frequency-scaling, and StirMark for audio benchmark attacks [20, 19, 18]. The proposed ICA-based watermark detector is applicable to SS-based watermarking of all media types, i.e. audio, video and images. However, in this paper the proposed detector is tested for digital audio (which includes music and voiced speech signals only) as the host media for watermark embedding, detection, and performance analysis.

In the past ICA-based framework has been used for multimedia watermarking [9, 10, 11, 13, 14, 12]. However, existing ICA-based data-hiding schemes are either not applicable to SS-based watermarking [9, 10, 11, 13] or use an informed detection framework for watermark extraction/extraction [14, 12] therefore are not discussed in this manuscript. For example, Yu et al in [14] have proposed ICA-based watermark detector that can be used for SS-based watermarking but their detector uses the embedded watermark and a private data during watermark extraction process. Similarly, Sener et al's proposed ICA based watermark detector in [12] is also applicable to SS-based watermark detection, but their proposed detector also also requires the original watermark during watermark detection process; therefore, cannot be used for blind watermark detection/extraction applications.

Rest of the paper is organized as follow: basics of SS-based watermarking are discussed in Section 2; a brief overview of the independent component analysis theory is provided in Section 3. The proposed ICA-based watermark detector along with its decoding, detection, and maximum watermarking-rate performance analysis are described in Section 4. Simulation results for decoding bit error probability performance of the proposed ICA-based watermark detector and a correlation-based detector against different attacks and signal degradations are described in Section 5. Finally the concluding remarks along with future research directions are presented in Section 6.

2 Basics of SS-based watermarking

The SS based watermarking system can be modeled using a classical secure communication model [1], as shown in Fig. 1. In Fig. 1, $\mathbf{S} \in \mathcal{R}^n$ is a vector containing coefficients of the host signal in marking space. It is assumed that the coefficients, $S_i : i = 0, 1, \cdots, n-1$, are independent and identically distributed (i.i.d.) random variables (r.v.) with zero mean and variance σ_s^2. A watermark, $\mathbf{V}$, is generated using: (1) a message bit, $b \in \{\pm 1\}$, to be embedded into n coefficients of the host signal, (2) a key-dependent pseudo-random sequence $\mathbf{W} \in \{\pm 1\}^n$, and (3) a perceptual mask, $\alpha \in \mathcal{R}^n$, estimated based on the human auditory system (HAS) and the host signal $\mathbf{S}$, i.e. $\alpha = f(\mathbf{S}, \text{HAS})$. We further assume that the watermark sequence $\mathbf{W}$ and the host signal coefficients $\mathbf{S}$ are mutually independent. The amplitude-modulated watermark is spectrally shaped according to perceptual mask α to meet the fidelity requirement of the perception based watermarking. The watermarked signal $\mathbf{X}$ is obtained by adding an amplitude-modulated watermark $\mathbf{V} = \alpha \odot \mathbf{W}b$, here $\odot$ denotes element-wise product of the two vectors, to the host signal $\mathbf{S}$. The watermarked signal $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \mathbf{S} + \mathbf{V}, \tag{1}$$

The embedding distortion, $\mathbf{D}_e$ can be expressed as,

$$\mathbf{D}_e = \mathbf{X} - \mathbf{S}. \tag{2}$$

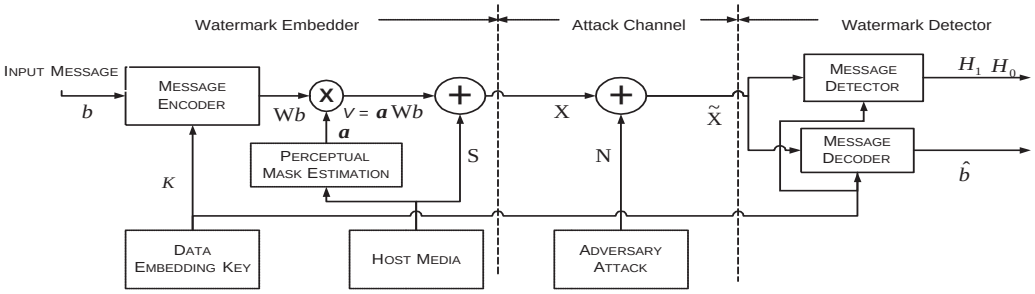


Figure 1: Perceptual based data hiding system with blind receiver

The mean-squared embedding distortion, d_e is expressed as,

$$\begin{aligned} d_e &= \frac{1}{n}E\{\|\mathbf{D}_e\|^2\} \\ &= \frac{1}{n}E\{\|\mathbf{X}-\mathbf{S}\|^2\} \\ &= \frac{1}{n}\|\alpha \odot \mathbf{W}b\|^2 \\ &= \frac{1}{n}\sum_{i=0}^{n-1}\alpha_i^2 = \sigma_v^2, \end{aligned} \tag{3}$$

where $\|\cdot\|$ represents the Euclidian norm, $E\{\cdot\}$ denotes expected value of a r.v., and σ_v^2 represents variance of the watermark $\mathbf{V}$.

The signal distortion due to an active adversary attack can be viewed as channel noise, $\mathbf{N}$, as shown in Fig. 1. The received watermarked signal at the detector, $\tilde{\mathbf{X}}$,

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{N}, \tag{4}$$

is processed for watermark detection.

The watermarking schemes based on blind additive embedding model generally use probabilistic characterization of the host signal to develop an optimal or suboptimal watermark detector (in ML sense). The statistical characterizations of real-world host signal are available in spatial domain as well as in the transform domain. For example, stationary speech samples/coefficients both in the time domain and in the DWT domain can be approximated by Laplacian distribution [21] (see Appendix A for the probability distribution function (pdf) of DWT coefficients) i.e.,

$$f_s(\tau) = \frac{\beta}{2}e^{-\beta|\tau|}, \ \ |\tau| < \infty, \tag{5}$$

where $\beta = \frac{\sqrt{2}}{\sigma_s}$

The average decoding bit error probability, P_e, under zero-channel distortion scenario, i.e. $N_i = 0$, can be calculated by assuming that

1. the watermarked sample X_i is obtained by adding a binary amplitude-modulated watermark V_i, i.e. $\alpha_i W_i b$,
2. the detector is based on Neyman-Pearson criterion,
3. no pre-processing is applied to the watermarked audio to suppress host interference,
4. W_i takes values ± 1 with probability $\frac{1}{2}$,

In addition, for performance analysis we will consider two information embedding scenarios: (1) one bit $b \in \{\pm 1\}$ of information is embedded in each coefficient of the host signal, S_i, and (2) one bit $b \in \{\pm 1\}$ of information is embedding in $|\zeta|$ coefficients of the host signal $\mathbf{S}$, where $|\zeta|$ denotes the cardinality of the selected coefficient indices set ζ.

Consider one bit embedding per coefficient, i.e. $n = 1$, case first. It has been shown in [7] that the ML decoder estimates $\hat{b} = 1$ if $\tilde{X}_0 W_0 > 0$ and an error will occur when $\tilde{X}_0 W_0 < 0$. The average P_e is given by

$$\begin{aligned} P_e &= \Pr\{\tilde{X}_0 W_0 < 0 | b = 1\} \\ &= \int_{-\infty}^{0} f_s(\tau - \alpha)\, d\tau. \end{aligned} \tag{6}$$

Assuming the Laplacian distribution model for the host, it can be shown

$$P_e = \frac{1}{2}e^{-\sqrt{2}/\lambda_0}, \tag{7}$$

where $\lambda_0 = \frac{\sigma_s}{\sigma_{v_0}}$ which is generally referred as *signal-to-watermark ratio* (SWR), when expressed in dB i.e. $SWR = 20\log_{10}\lambda$.

It can be observed from Eq. (7) that non-zero P_e is not achievable even in the absence of attack-channel distortion, and $P_e = f(\lambda)$. In addition, the value of the parameter λ determines the tradeoff between fidelity of the embedded watermark and P_e.

Consider second embedding scenario, i.e., one bit information is embedded in $|\zeta| = n$ coefficients of the host. In this case the watermarked audio is given by,

$$X_i = S_i + \alpha_i W_i b, \ i \in \zeta. \tag{8}$$

Let us assume that the watermarked signal used for detection is free of attack-channel distortion, and message symbols are equally probable. In this case, the ML decoder that minimizes the decoding error probability will assign decision regions D_- and D_+ as follow,

$$\ln\frac{f_x(\mathbf{x}|b_+)}{f_x(\mathbf{x}|b_-)} = \ln\frac{f_x(\mathbf{x}-\hat{\alpha}\mathbf{w})}{f_x(\mathbf{x}+\hat{\alpha}\mathbf{w})} \underset{D_-}{\overset{D_+}{\gtrless}} 0, \tag{9}$$

where $b_+(resp.b_-)$ represent the event that binary information $b = +1(resp.b = -1)$ is embedded in the selected indices and $\hat{\alpha}$ is the masking threshold estimated from watermarked audio. It is shown in Section 4 that the estimated of masking threshold from the unwatermarked and watermarked audio clip are very close given that attack-channel distortion induced into the watermarked audio is below certain threshold. It is therefore reasonable to assume that $\hat{\alpha} \approx \alpha$.

The ML sufficient statistic, T, assuming Laplacian pdf for the host coefficients S_i, can be written as,

$$T(\mathbf{x}|\ \mathbf{s},\ \hat{\alpha}) = \sum_{i\in\zeta} \beta\left(|X_i + \hat{\alpha}_i W_i| - |X_i - \hat{\alpha}_i W_i|\right). \quad (10)$$

If $b = 1$ was embedded, then the sufficient statistics T can be expressed as,

$$T(\mathbf{x}|\ \mathbf{s},\ \hat{\alpha}) = \sum_{i\in\zeta} \beta\left(|S_i + 2\hat{\alpha}_i W_i| - |S_i|\right). \quad (11)$$

Here the ML detector is a bit-by-bit hard decoder, i.e.,

$$\hat{b} = \text{sgn}(T). \quad (12)$$

To determine the bit error probability for this ML decoder, a statistical characterization of T is required. Here T is sum of $|\zeta|$ i.i.d. random variables. Therefore, by applying the central limit theorem (CLT), T can be approximated by the Gaussian random variable. Mean of T, $E\{T\}$ can be calculated as,

$$E\{T(\mathbf{x}|\ \mathbf{s},\ \hat{\alpha})\} = \sum_{i\in\zeta} \beta\left(E_{s,w}\left(|S_i + 2\hat{\alpha}_i W_i| - |S_i|\right)\right), \quad (13)$$

and variance,

$$E\{T\} = \sum_{i\in\zeta} \left(e^{-2\sqrt{2}/\lambda_i} + \frac{2\sqrt{2}}{\lambda_i} - 1\right), \quad (14)$$

$$\text{Var}\{T(\mathbf{x}|\ \mathbf{s},\ \hat{\alpha})\} = \sum_{i\in\zeta} \beta\left(\text{Var}_{s,w}\left(|S_i + 2\hat{\alpha}_i W_i| - |S_i|\right)\right) \quad (15)$$

$$\text{Var}\{T\} = \sum_{i\in\zeta} \left(3 - e^{-4\sqrt{2}/\lambda_i} - e^{-2\sqrt{2}/\lambda_i}\left(1 + \frac{4\sqrt{2}}{\lambda_i}\right)\right). \quad (16)$$

In this case, the P_e is given as,

$$P_e = Q\left(\frac{|E\{T\}|}{\sqrt{\text{Var}\{T\}}}\right), \quad (17)$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt$

Eq. (17) shows that the decoding error probability P_e is non-zero even in the absence of attack-channel distortion, and P_e is a function of λ. The above analysis also shows that the detection/decoding performance of a blind detector for additive embedding schemes is inherently bounded below by the host-signal interference at the detector. The main motivation behind this paper is to design a watermark detector for additive embedding schemes with an improved watermark detection, decoding, and maximum watermarking-rate performances by suppressing the host-signal interference at the blind detector. Towards this end, theory of ICA is used by posing watermark estimation for additive embedding as a BSS problem. The proposed framework first estimates embedding watermark using BSS based on ICA which is then used for detection and decoding. The fundamentals of the ICA theory are briefly outlined in the following section followed by the details of the proposed ICA-based detector.

3 Independent Component Analysis

Independent component analysis (ICA) is a statistical framework for estimating underlying hidden factors or components of multivariate statistical data. In the ICA model, the data variables are assumed to be linear or nonlinear mixtures of some unknown latent variables, and the mixing system is also unknown [23, 22]. The hidden variables are also assumed to be non-Gaussian and mutually independent. The ICA model can be considered as an extension of the principal component analysis (PCA) and factor analysis [23, 22]. In fact, ICA can be treated as non-Gaussian factor analysis, since data is modeled as a linear mixture of underlying non-Gaussian factors. The ICA framework has been used in diverse application scenarios including blind source separation (BSS), feature extraction, telecommunication, and economics [23, 22]. In the following we will review only the linear ICA framework since only that is relevant to the SS-based watermarking model. In general, the linear ICA model can be defined for noise-free as well as noisy scenarios as follows.

Noise-free ICA model: *ICA of a random vector* $\mathbf{X} \in \mathcal{R}^m$ *consists of estimating the following generative model of the data:*

$$\mathbf{X} = \mathbf{AS}, \quad (18)$$

where $\mathbf{X}$ *represents* n*-realizations of the observed* m*-dimensional random vector,* $\mathbf{S} \in \mathcal{R}^{n_1}$ *is the hidden random variables and* $\mathbf{A} \in \mathcal{R}^{m\times n_1}$ *is mixing matrix. The hidden variables,* $\mathbf{S}^{(i)}$*, in the vector* $\mathbf{S} = \left[\mathbf{S}^{(1)}, \cdots, \mathbf{S}^{(n_1)}\right]^{\mathrm{t}}$ *are assumed statistically independent.*

Noisy ICA model: *ICA of a random vector* $\mathbf{X}$ *consists of estimating the following generative model of the data:*

$$\mathbf{X} = \mathbf{AS} + \mathbf{N}, \quad (19)$$

where $\mathbf{N}$ *is* n*-realizations of an* m*-dimensional random noise, while* $\mathbf{X}$, $\mathbf{S}$, and $\mathbf{A}$ *are the same as in the noise-free model in* Eq. (18).

In this paper, we use the noisy ICA generative model to design an ICA-based watermark detector for SS-based watermarking schemes. The proposed ICA-based watermark detector attempts to estimate the embedded watermark from the watermarked signal while reducing the host-signal interference at the watermark detector. Before estimating the underlying independent components from observed data using ICA framework, the generative model should meet certain conditions to ensure the identifiability of the ICA model. The identifiability constraints defined in [22, 24, 25, 29, 26, 27] underdetermined ICA (UICA) model are outlined below:

1. *Statistical independence*: The hidden (latent) variables/sources are statistically independent.
2. *Non-Gaussianity*: At most one of the underlying independent components $\mathbf{S}^{(i)},\ i = 1, 2 \cdots n_1$, is normally distributed.

Therefore, independence and maximum non-Gaussianity are two fundamental ingredients of the UICA framework. Independence of the underlying components is one of the assumptions that is made to estimate components from the linear mixture. Note that independence of the underlying components is a stronger condition than uncorrelatedness, e.g., for the BSS problem, there might be many dependent but uncorrelated representations of the observed signals and these uncorrelated but dependent representations of the observed signals cannot separate the mixed sources [22]. Therefore, uncorrelatedness itself is insufficient to solve the BSS problem. In fact, independence implies nonlinear uncorrelatedness [22], that is, if $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are two independent components then any nonlinear transformations of these components, say, $\phi_1(\mathbf{S}^{(1)})$ and $\phi_2(\mathbf{S}^{(2)})$, are uncorrelated as well (i.e. their covariance is zero). On the other hand, if $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are assumed to be just uncorrelated then in general, the corresponding nonlinear transformations do not necessarily have zero covariance. Thus to perform ICA, a stronger form of decorrelation of the underlying components is required, that is, nonlinear decorrelation. A suitable selection of nonlinearities, i.e. $\phi_1(\cdot)$ and $\phi_2(\cdot)$, can be achieved by using tools like maximum likelihood and mutual information from estimation theory and information theory [22].

Maximum non-Gaussianity is another important requirement of ICA-based hidden components estimation [23, 22, 38, 30]. A quantity kurtosis defined in terms of the fourth-order central moment κ is generally used as a measure of non-Gaussianity of a random variable. Kurtosis of a real random variable S can be defined as,

$$\kappa = \left(E\{(S - E(S))^4\} / E^2\{(S - E(S))^2\} \right) - 3. \quad (20)$$

A normal random variable has zero kurtosis; therefore, kurtosis is a measure of the *distance* of a random variable from a Gaussian distribution. Distributions that are peakier (flatter) about the mean than a Gaussian distribution generally have positive (negative) kurtosis. Random variables with positive kurtosis, i.e. $\kappa > 0$, are generally called super-Gaussian. The Laplacian distribution is a typical example of this case. Random variables with negative kurtosis value, i.e., $\kappa < 0$ are called sub-Gaussian, e.g., the uniform distribution.

The BSS is one of the most widely explored applications of the ICA model [23, 22]. In case of BSS using ICA framework, the recovery of the underlying sources relies on the assumption that the constituent sources are mutually independent. The *cocktail party problem* is a classical example of BSS, where several people are simultaneously speaking in the same room and objective is to separate voices of different speakers using microphone recordings (in the room). In order to illustrate the idea n_1 speakers (sources) are considered here. The observation $\mathbf{X} \in \mathcal{R}^{m \times\ n}$ is generated by mixing sources $\mathbf{S} \in \mathcal{R}^{n_1 \times\ n}$ by a *mixing matrix* $\mathbf{A} \in \mathcal{R}^{m \times\ n_1}$. The static linear mixing model can be expresses as,

$$\mathbf{X}_i = \mathbf{A}\mathbf{S}_i + \mathbf{N}_i,\ i = 1, 2, \cdots,\ n \quad (21)$$

The aim of BSS is the recover the underlying sources $\mathbf{S}^{(l)},\ l = 1, 2, \cdots, n_1$ from the observation $\mathbf{X}$ only. The ICA achieves the separation relying on the assumption that the underlying sources are mutually independent. To this end the ICA framework finds a linear representation in which the underlying components are statistically independent. In other words, BSS using ICA tries to estimate the *demixing* (*separating*) matrix, $\mathbf{B} \in \mathcal{R}^{n_1 \times m}$, from the observed data $\mathbf{X}$. The estimated demixing matrix is the inverse (or generalized inverse) of mixing matrix $\mathbf{A}$, i.e., $\hat{\mathbf{B}} = \hat{\mathbf{A}}^{\dagger} = (\hat{\mathbf{A}}^T \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T$. Most of existing BSS schemes using ICA model are based on the information-theoretic framework. For example, Bell et al's [21] ICA scheme is based on the idea of information maximization, or *infomax* among the estimated independent components. P. Comon in [23] has used higher-order cumulants whereas, Gaeta et al in [28] used ML estimation framework for BSS. Many existing BSS methods are extensions of infomax, higher-order cumulants, and ML method [23, 22].

4 Proposed ICA Based Watermark Detector

The proposed ICA-based watermark detector consists of two stages: 1) watermark estimation stage, and, 2) watermark decoding and/or detection stage. The watermark estimation stage estimates watermark $\tilde{\mathbf{V}}$ from the received watermarked audio $\tilde{\mathbf{X}}$ using ICA framework, whereas, the watermark decoding (resp. detection) stage decodes (resp. detects) the embedded watermark using the ML approach. The block diagram of the proposed watermark detector is given in the Fig. 2.

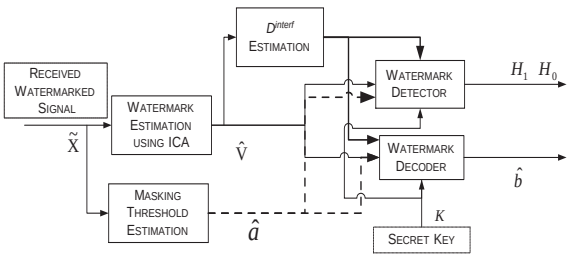


Figure 2: Block diagram of the proposed ICA-based watermark detector

In general the ICA model for BSS estimates the demixing matrix $\hat{\mathbf{B}}$ from the observed data $\mathbf{X}$, and hence the underlying independent components $\hat{\mathbf{S}}^{(i)}$. This model is ex-

tendable to the watermark estimation problem for the watermarked signal, assuming identifiability conditions of the UICA model are satisfied. To verify whether the additive embedding model (Eq. (1)) satisfies the identifiability constraints of an ICA model, rewrite Eq. (1) with $b = 1$, i.e.,

$$\mathbf{X} = \mathbf{S} + \alpha \odot \mathbf{W}.$$

The non-Gaussianity of the host signal and the watermark is the only requirement to satisfy constraints of UICA. As mentioned in Section (2) that real-world audio samples/coefficients in the time domain as well as in the DWT domain can be approximated by the Laplacian distribution (see Appendix A) and therefore if the watermark, $\mathbf{W}$, is generated based on some non-Gaussian distribution then non-Gaussianity constraint of UICA model is satisfied as well. Once the identifiability conditions of the UICA model are satisfied, the noisy ICA model can be extended to estimate the watermark from the watermarked audio generated using Eq. (1).

4.1 Watermark Estimation

For watermark estimation, the proposed watermark detector first estimates the watermark-mixing matrix $\hat{\mathbf{A}}$ which is then to estimate the underlying independent components (i.e., the host signal $\mathbf{S}$ and the watermark $\mathbf{W}$). An estimate of the watermark-mixing matrix, $\hat{\mathbf{A}}$, is usually obtained by optimizing a highly nonlinear function of the hidden sources also known as contrast function [23, 22]. The pseudo-inverse of the estimated watermark-mixing matrix $\hat{\mathbf{A}}^{\dagger}$ is applied to the observed mixture to estimate the host signal $\hat{\mathbf{S}}$ and the watermark $\hat{\mathbf{W}}$. However, as noted earlier, in the case of blind detectors for SS-based watermarking schemes, watermark estimation using ICA framework is a degenerate case, i.e., $m < n_1$. Therefore, just the estimation of watermark-mixing matrix is insufficient to separate the underlying independent components perfectly. In the case of additive embedding, the equation $\mathbf{X} = \hat{\mathbf{A}}\mathbf{S}$ has an affine set of solutions [34]. A preferred solution in this affine set is generally selected using probabilistic prior model of the independent components [39]. The performance of the proposed ICA-based watermark estimator depends on the separation quality of the separated (estimated) watermark. The separation quality of the separated source is generally measured in terms of, 1) *source-to-interference ratio* (*watermark-to-interference ratio (WIR)*, in case of watermark estimation), *source-to-noise ratio*, and 2) *source-to-artifact ratio* (for further details on these separation quality measures please see [35] and references therein). For performance analysis of the proposed ICA detector, only WIR distortion measure is considered here; therefore, the estimated watermark can be expressed as

$$\hat{V}_i = \eta_{1i}\alpha_i W_i b + S_i^{\text{interf}}, \tag{22}$$

where $\eta_{1i} \in \mathcal{R},\ 0 < \eta_{1i} \leq 1$ and S_i^{interf} is interference due to the host signal.

Let $S_i^{\text{interf}} = \eta_{2i} S_i,\ \eta_{2i} \in \mathcal{R}$, and $0 < \eta_{2i} \leq 1$ then Eq. (22) can be rewritten as,

$$\hat{V}_i = \eta_{1i}\alpha_i W_i b + \eta_{2i} S_i. \tag{23}$$

The relative distortion due to interference in the estimated watermark is defined as,

$$D^{\text{interf}} = \left(\eta_1/\eta_2\right)^2, \tag{24}$$

where $WIR = 10\log_{10}\left(D^{\text{interf}}\right)$ dB.

In general, $D^{\text{interf}} > 0$ dB for most of existing BSS schemes based on ICA framework [34, 35]. Several researchers have proposed elegant BSS algorithms based on ICA model for noisy data [38, 36, 31], these algorithms can be used for watermark estimation from the watermark audio. Among these, the FastICA for noisy data [38] is used in this paper due to its better computational and separation quality performance over existing algorithms [34].

It can be observed from Eq. (23) that the ICA stage acts as a pre-processing stage that suppresses the host interference or improves *watermark-to-host ratio*. Once estimated watermark $\hat{\mathbf{V}}$ is available, an optimal detector can be designed based on the statistics of $\hat{\mathbf{V}}$ for watermark detection (resp. decoding). It is important to notice that ICA based pre-processing stage uses constraints like mutual independence of the underlying sources, non-Gaussianity, and multichannel observation i.e. $m \geq 2$. A constrained optimization of highly nonlinear cost function e.g. $\tanh(x)$, $x\exp(-x^2)$, etc. is used to suppress the host interference in the estimated watermark [22, 23]. In addition, under practical scenarios, BSS using ICA also requires reasonably large number of data samples n to separate the underlying sources. Therefore, ICA based pre-processing to suppress host interference is inherently different from filtering based pre-processing schemes i.e., optimal linear filtering [44], wiener filtering, non-linear filtering, etc. The ICA-based pre-processing stage is to improve *watermark-to-host ratio* hence expected to improve the detection performance [41, 42]. It is however important to mention that improvement comes at the cost of higher computational power.

In the following subsections we analyze the performance of the proposed ICA-based detector in terms of three parameters: (1) detection rate in terms of false positives and true positives 4.2, (2) decoding error probability 4.3, and (3) maximum watermarking rate 4.4.

4.2 Watermark Detection: *Performance Analysis*

A watermark detector is generally characterized by two performance measures: the probability of *false alarm* P_F and the probability of *detection* P_D. The probability of detection represents the probability of deciding on the presence of a watermark when the received audio indeed contains a watermark. The probability of false alarm represents chances of deciding the presence of a watermark when in

fact the received audio does not contain a watermark. The watermark detection process can be treated as a binary decision problem in the presence or absence of attack-channel distortions.

We first consider the case where the received watermarked audio has not suffered attack-channel distortion. The estimated watermark is given by,

$$\hat{V}_i = \eta_{1i}\alpha_i W_i b + \eta_{2i} S_i,\ i \in \zeta. \tag{25}$$

In this scenario, the watermark detection can be formulated as a binary hypotheses test,

$$\begin{aligned} H_1 : \hat{V}_i &= \eta_{1i}\hat{\alpha}_i W_i b + \eta_{2i} S_i \\ H_0 : \hat{V}_i &= \eta_{2i} S_i,\ i \in \zeta. \end{aligned} \tag{26}$$

In this detection problem, the watermark $\mathbf{W}$ is the target signal and host interference, $\eta_2\mathbf{S}$ acts as additive noise. The goal of watermark detector is to determine presence or absence of the watermark in the estimated watermark $\hat{\mathbf{V}}$ based on the statistics of $\mathbf{S}$ and $\mathbf{W}$. Let us assume that statistics of unwatermarked and watermarked audio are same [43], therefore *pdfs* under each hypothesis are known. The decision rule, in this scenario is based on likelihood ratio which is given as:

$$\Lambda(\hat{\mathbf{V}}) = \prod_{\zeta} \left(\frac{f_{\hat{v}}(\hat{\mathbf{v}}|H_1)}{f_{\hat{v}}(\hat{\mathbf{v}}|H_0)} \right) \underset{H_0}{\overset{H_1}{\gtrless}} \acute{\xi} \tag{27}$$

where $\Lambda(\hat{\mathbf{V}})$ is likelihood ratio and $\acute{\xi}$ is decision threshold.

The log-likelihood is defined as,

$$\begin{aligned} L(\hat{\mathbf{V}}) &= \ln\left(\Lambda(\hat{\mathbf{V}})\right) \qquad (28) \\ &= \ln\left(\prod_{\zeta} \left(\frac{f_{\hat{v}}(\hat{\mathbf{v}}|H_1)}{f_{\hat{v}}(\hat{\mathbf{v}}|H_0)} \right) \right) \\ &= \ln\left(\prod_{\zeta} \left(\sum_{l} \left(\frac{p(b_l) f_{\hat{v}}(\hat{\mathbf{v}}|H_1)}{f_{\hat{v}}(\hat{\mathbf{v}}|H_0)} \right) \right) \right) \\ &= \ln\left(\prod_{\zeta} \left(\sum_{l} \left(\frac{p(b_l) f_{\acute{s}}(\hat{\mathbf{v}} - \hat{\alpha} \odot \mathbf{w} b_l)}{f_{\acute{s}}(\hat{\mathbf{v}})} \right) \right) \right) \\ &\underset{H_0}{\overset{H_1}{\gtrless}} \xi \qquad (29) \end{aligned}$$

where, $l \in \{\pm 1\}$, $\xi = \ln(\acute{\xi})$ and r.v. $\acute{S}_i$ is defined as $\acute{S}_i = \eta_{2i} S_i$.

In the above test, the decision threshold ξ can be minimized based on Neyman-Pearson rule, that is, maximize the P_D for a given value of P_F [41, 42].

Assuming Laplacian distribution for the host audio, the $L(\hat{\mathbf{V}})$ can be written as,

$$L(\hat{\mathbf{V}}|\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2) = \sum_{i\in\zeta} \acute{\beta}_i \left(|\hat{V}_i| - |\hat{V}_i - \hat{\eta}_{1i}\hat{\alpha}_i W_i| \right), \tag{30}$$

where $\acute{\beta}_i = \sqrt{2}/\hat{\eta}_{2i}\sigma_s$, $\hat{\eta}_1$, and $\hat{\eta}_2$ are estimates of scaling coefficients of $\mathbf{V}$ and $\mathbf{S}$ in $\hat{\mathbf{V}}$. Estimation details of $\hat{\eta}_1$, and $\hat{\eta}_2$ are discussed in Section 4.5.

The statistical characterization of $L(\hat{V})$ under hypothesis H_0 can be determined as,

$$L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2) = \sum_{i\in\zeta} \acute{\beta}_i \left(|\hat{V}_i| - |\hat{V}_i - \hat{\eta}_{1i}\hat{\alpha}_i W_i| \right), \tag{31}$$

$$L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2) \overset{\text{def}}{=} \sum_{i\in\zeta} \acute{\beta}_i\left(Z_i\right), \tag{32}$$

$$\text{where } Z_i \overset{\text{def}}{=} |\hat{V}_i| - |\hat{V}_i - \hat{\eta}_{1i}\hat{\alpha}_i W_i|. \tag{33}$$

Here $L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)$ is the sum of $|\zeta|$ statistically independent random variables that can be approximated by the Gaussian random variable based on the CLT, mean, m_0 and variance, σ_0^2 of $L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)$ is calculated as follows,

$$\begin{aligned} m_0 &\overset{\text{def}}{=} E\{L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} \qquad (34) \\ &= \sum_{i\in\zeta} \acute{\beta}_i E\{Z_i\}, \\ \sigma_0^2 &\overset{\text{def}}{=} \text{Var}\{L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} \qquad (35) \\ &= \sum_{i\in\zeta} \acute{\beta}_i^2 \text{Var}\{Z_i\}. \end{aligned}$$

Averaging Eq. (36) over W, we have,

$$E_w\{Z_i\} = |\acute{S}_i| - \frac{1}{2}\left(|\acute{S}_i| + \hat{\eta}_{1i}\hat{\alpha}_i + \left| |\acute{S}_i| - \hat{\eta}_{1i}\hat{\alpha}_i \right| \right), \tag{36}$$

$$\text{Var}_w\{Z_i\} = \frac{1}{4}\left(|\acute{S}_i| + \hat{\eta}_{1i}\hat{\alpha}_i + \left| |\acute{S}_i| - \hat{\eta}_{1i}\hat{\alpha}_i \right| \right)^2. \tag{37}$$

These equation can be rewritten as,

$$E_w\{Z_i\} = \begin{cases} |\acute{S}_i| - \hat{\eta}_{1i}\hat{\alpha}_i & |\acute{S}_i| \le \hat{\eta}_{1i}\hat{\alpha}_i \\ 0 & |\acute{S}_i| > \hat{\eta}_{1i}\hat{\alpha}_i, \end{cases} \tag{38}$$

$$\text{Var}_w\{Z_i\} = \begin{cases} |\acute{S}_i^2| & |\acute{S}_i| \le \hat{\eta}_{1i}\hat{\alpha}_i \\ \hat{\eta}_{1i}^2\hat{\alpha}_i^2 & |\acute{S}_i| > \hat{\eta}_{1i}\hat{\alpha}_i. \end{cases} \tag{39}$$

Averaging it over $\acute{S}$ we have,

$$\begin{aligned} E\{Z_i\} &= E_{\acute{s}}\left(E_w\{Z_i\}\right) \qquad (40) \\ &= \frac{1}{\acute{\beta}_i}\left(1 - e^{-\acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i} - \acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i\right), \end{aligned}$$

$$\begin{aligned} &\text{Var}\{Z_i\} = E_{\acute{s}}\left(\text{Var}_w\{Z_i\}\right) + \text{Var}_{\acute{s}}\left(E_w\{Z_i\}\right) \qquad (41) \\ &= \frac{1}{\acute{\beta}_i^2}\left(3 - e^{-2\acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i} - 2e^{-\acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i}\left(1 + 2\acute{\beta}_i\hat{\eta}_i\hat{\alpha}_i\right)\right). \end{aligned}$$

Substituting $E\{Z_i\}$, and $\text{Var}\{Z_i\}$ in Eq. (36), we have,

$$m_0 = \sum_{i\in\zeta}\left(1 - e^{-\frac{\hat{\eta}_{1i}\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}} - \frac{\hat{\eta}_{1i}\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}\right), \tag{42}$$

$$\sigma_0^2 = \sum_{i\in\zeta}\left(3 - e^{-\frac{\hat{\eta}_{1i}2\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}} - 2e^{-\frac{\hat{\eta}_{1i}\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}}\left(1 + \frac{2\hat{\eta}_{1i}\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}\right)\right). \tag{43}$$

Similarly, $L(\hat{V})$ under hypothesis H_i can be written as,

$$L(\hat{\mathbf{V}}|H_1,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2) = \sum_{i\in\zeta} \acute{\beta}_i \left(|\hat{V}_i + \hat{\eta}_{1i}\hat{\alpha}_i W_i| - |\hat{V}_i| \right) \tag{44}$$

Here $L(\hat{\mathbf{V}})$ can be approximated by a Gaussian random variable with the same set of assumptions as under hypothesis H_0. In addition, the distribution of $L(\hat{\mathbf{V}})$ under hypothesis H_1 is symmetrical to the distribution of under H_0 with respect to the origin. Therefore,

$$\begin{aligned} m_1 &\stackrel{\text{def}}{=} E\{L(\hat{\mathbf{V}}|H_1,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} \quad (45)\\ &= -E\{L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\},\\ \sigma_1^2 &\stackrel{\text{def}}{=} \text{Var}\{L(\hat{\mathbf{V}}|H_1,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} \quad (46)\\ &= \text{Var}\{L(\hat{\mathbf{V}}|H_0,\ \mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} \end{aligned}$$

Now the probability of false alarm P_F and the probability detection P_D are given as,

$$P_F = Q\left(\frac{\xi + m_1}{\sigma_1}\right), \quad (47)$$

$$P_D = Q\left(\frac{\xi - m_1}{\sigma_1}\right). \quad (48)$$

Lets define the watermark-to-noise ratio (*WNR1*) as

$$WNR1 \stackrel{\text{def}}{=} \frac{m_1^2}{\sigma_1^2}. \quad (49)$$

If we denote by $Q^{-1}(P_F)$ the value $x \in \mathcal{R}$ such that $Q(x) = P_F$ then *receiver operating characteristics (ROC)* of the proposed detector can be expressed as:

$$P_D = Q\left(Q^{-1}(P_F) - 2\sqrt{WNR1}\right). \quad (50)$$

It can be observed from Eq. (50) that the detection performance of the proposed detector is a function of WNR1. Since the proposed ICA-based detector is designed to reduce the host-signal interference before detection, therefore, the ICA-based detector is expected to perform better than the existing blind detectors [6, 3, 1] operating without reducing host signal interference. The detection performance improvement can be attributed to its host interference suppression or watermark-to-host ratio improving capability. To illustrate this notion the theoretical ROC performance of the proposed detector based on Eq. (50) for different values of host interference suppression values (or *WIR*) is given in Fig. 3. It can be observed from Fig. 3 that the proposed detector performs superior that the detector operating without host interference canceling.

4.3 Watermark Decoding: *Performance Analysis*

To evaluate performance of the proposed detector in terms of decoding error probability, let us consider watermark embedding model given in Eq. (1) and decoding framework discussed in Section 2. Consider one bit per coefficient embedding case first, that is, $X_0 = S_0 + \alpha_0 W_0$. The P_e in this case for $\hat{V}_0$ can be expressed as,

$$P_{e\_ICA} = \frac{1}{2} e^{-\{\frac{\hat{\eta}_{10}}{\hat{\eta}_{20}}\}\left(\sqrt{2}/\lambda_0\right)}. \quad (51)$$

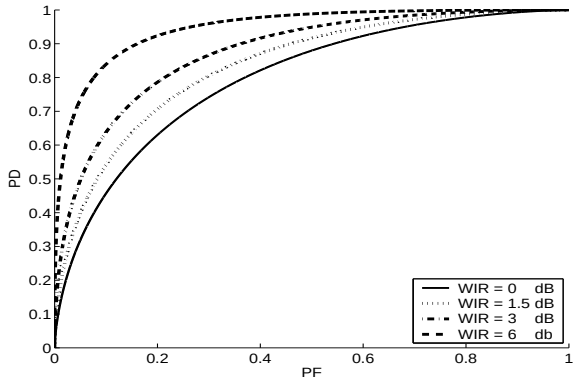


Figure 3: ROC performance of the proposed detector for different values of *WIR*, SWR = 13 dB, and one bit per $|\zeta|$ coefficients embedding, where $|\zeta| = 5$ (theoretical values)

Here Eq. (51) shows that ICA-based detector does improve decoding error performance. The decoding error performance gain for the proposed ICA-based detector over the traditional detector can expressed,

$$G = \frac{P_e}{P_{e\_ICA}} = e^{-\frac{\sqrt{2}}{\lambda_0}\{1 - \frac{\hat{\eta}_{10}}{\hat{\eta}_{20}}\}}. \quad (52)$$

It is important to mention that in general BSS using ICA have relatively small interference distortion, i.e., $\hat{\eta}_{10}/\hat{\eta}_{20}$ [34], therefore, $G \geq 0$ for $WIR > 0$dB. The performance gain ot the proposed ICA-based detector over that of the decoder given by Eq. (7) is plotted in Fig. 4. It can be observed from Fig. 4 that for a fixed value of *SWR*, decoding error probability of the proposed detector improves with the increase in the separation quality of the ICA scheme used for watermark estimation.

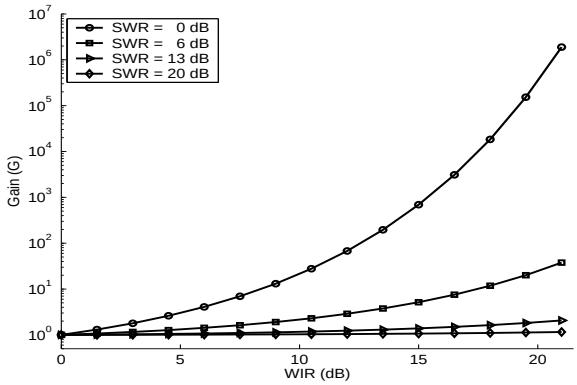


Figure 4: The decoding performance gain due to host-interference suppression at the detector (theoretical values)

Now consider second embedding scenario, that is, one bit is embedded into $|\zeta|$ coefficients of the host signal $\mathbf{S}$, i.e.,

$$X_i = S_i + \hat{\alpha}_i W_i b,\ i \in \zeta. \quad (53)$$

In this case, the estimated watermark $\hat{\mathbf{V}}$ using proposed ICA-based watermark detector, under zero attack-channel

distortion, can be expressed as,

$$\hat{V}_i = \eta_{2i} S_i + \eta_{1i}\hat{\alpha}_i W_i b,\ i \in \zeta. \tag{54}$$

For equally probable message symbols the ML decoder that minimizes the P_e will satisfy the following condition,

$$\ln \frac{f_{\hat{v}}(\hat{\mathbf{v}}|b_+, \hat{\alpha}, \hat{\eta}_1, \hat{\eta}_2)}{f_{\hat{v}}(\hat{\mathbf{v}}|b_-, \hat{\alpha}, \hat{\eta}_1, \hat{\eta}_2)} = \ln \frac{f_{\acute{s}}(\hat{\mathbf{v}} - \hat{\eta}_1\hat{\alpha}\mathbf{w})}{f_{\acute{s}}(\hat{\mathbf{v}} + \hat{\eta}_1\hat{\alpha}\mathbf{w})} > 0. \tag{55}$$

The ML sufficient statistics for Laplacian $\acute{\mathbf{S}}$ can be written as,

$$T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2) = \tag{56}$$
$$\sum_{i\in\zeta} \acute{\beta}_i \left(|\hat{V}_i + \hat{\eta}_{1i}\hat{\alpha}_i W_i| - |\hat{V}_i - \hat{\eta}_{1i}\hat{\alpha}_i W_i| \right)$$

Assuming $b = 1$, then $T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)$ can be expressed as,

$$T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2) = \sum_{i\in\zeta} \acute{\beta}_i \left(|\acute{S}_i + 2\hat{\eta}_{1i}\hat{\alpha}_i W_i| - |\acute{S}_i| \right), \tag{57}$$

$$T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2) \ \stackrel{\text{def}}{=}\ \sum_{i\in\zeta} \acute{\beta}_i Z_i, \tag{58}$$

where, $Z_i \stackrel{\text{def}}{=} \left(|\acute{S}_i + 2\hat{\eta}_{1i}\hat{\alpha}_i W_i| - |\acute{S}_i| \right)$ The ML decoder is a bit-by-bit hard decoder

$$\hat{b} = \text{sgn}(T) \tag{59}$$

To determine the P_e for the ML decoder, a statistical characterization of $T(\hat{\mathbf{V}})$ is required. As $T(\hat{\mathbf{V}})$ is sum of $|\zeta|$ i.i.d. random variables, therefore, using CLT, $T(\hat{\mathbf{V}})$ can be approximated by the Gaussian random variable, the mean and variance of T can be computed as,

$$E\{T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} \ \stackrel{\text{def}}{=}\ \sum_{i\in\zeta} \acute{\beta}_i E\{Z_i\}, \tag{60}$$

$$\text{Var}\{T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} \ \stackrel{\text{def}}{=}\ \sum_{i\in\zeta} \acute{\beta}^2 \text{Var}\{Z_i\}. \tag{61}$$

In Z_i, W and $\acute{S}$ are the only r.v.s, so averaging Z_i over r.v. W condition to the selected host indices $\acute{S}$ and $W_i \in \{\pm 1\}$ with probability $\frac{1}{2}$ we have,

$$E_w\{Z_i\} = \frac{1}{2}\left(|\acute{S}_i| + 2\hat{\eta}_{1i}\hat{\alpha}_i + \left||\acute{S}_i| - 2\hat{\eta}_{1i}\hat{\alpha}_i\right| \right) - |\acute{S}_i|, \tag{62}$$

$$\text{Var}_w\{Z_i\} = \frac{1}{4}\left(|\acute{S}_i| + 2\hat{\eta}_{1i}\hat{\alpha}_i + \left||\acute{S}_i| - 2\hat{\eta}_{1i}\hat{\alpha}_i\right| \right)^2. \tag{63}$$

rewriting the above equations, we have,

$$E_w\{Z_i\} \ = \ \begin{cases} -|\acute{S}_i| + 2\hat{\eta}_{1i}\hat{\alpha}_i & |\acute{S}_i| \le 2\hat{\eta}_{1i}\hat{\alpha}_i \\ 0 & |\acute{S}_i| > 2\hat{\eta}_{1i}\hat{\alpha}_i, \end{cases} \tag{64}$$

$$\text{Var}_w\{Z_i\} \ = \ \begin{cases} |\acute{S}_i^2| & |\acute{S}_i| \le 2\hat{\eta}_{1i}\hat{\alpha}_i \\ 4\hat{\eta}_{1i}^2\hat{\alpha}_i^2 & |\acute{S}_i| > 2\hat{\eta}_{1i}\hat{\alpha}_i. \end{cases} \tag{65}$$

Now averaging over r.v. $\acute{S}_i$, we have,

$$\begin{aligned} E\{Z_i\} &= E_{\acute{s}}\left(E_w\{Z_i\}\right) \\ &= \frac{1}{\acute{\beta}_i}\left(e^{-2\acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i} + 2\acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i - 1 \right), \end{aligned} \tag{66}$$

$$\text{Var}\{Z_i\} = E_{\acute{s}}\left(\text{Var}_w\{Z_i\}\right) + \text{Var}_{\acute{s}}\left(E_w\{Z_i\}\right) \tag{67}$$
$$= \frac{1}{\acute{\beta}_i^2}\left(3 - e^{-4\acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i} - 2e^{-2\acute{\beta}_i\hat{\eta}_{1i}\hat{\alpha}_i}\left(1 + 4\acute{\beta}_i\hat{\eta}_i\hat{\alpha}_i\right) \right).$$

Substituting $E\{Z_i\}$, and $\text{Var}\{Z_i\}$ in Eq. (61) and Eq. (61), we have,

$$E\{T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} = \sum_{i\in\zeta}\left(e^{-\frac{\hat{\eta}_{1i}2\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}} + \frac{2\hat{\eta}_{1i}\sqrt{2}}{\hat{\eta}_{2i}\lambda_i} - 1 \right), \tag{68}$$

$$\text{Var}\{T(\hat{\mathbf{V}}|\,\mathbf{s},\ \hat{\alpha},\ \hat{\eta}_1,\ \hat{\eta}_2)\} = \tag{69}$$
$$\sum_{i\in\zeta}\left(3 - e^{-\frac{\hat{\eta}_{1i}4\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}} - 2e^{-\frac{\hat{\eta}_{1i}2\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}}\left(1 + \frac{4\hat{\eta}_{1i}\sqrt{2}}{\hat{\eta}_{2i}\lambda_i}\right) \right).$$

Therefore, P_e for an ICA-based detector is given as,

$$P_{e\_ICA} = Q\left(\frac{|E\{T\}|}{\sqrt{\text{Var}\{T\}}} \right) \tag{70}$$

It can be observed from Eq. (70) that the decoding error probability of the ML decoder applied to the estimated watermark is a function of *WIR* and *SWR*. The performance of the proposed ICA-based detector given by Eq. (70) for different values of *WIR* and *SWR* is plotted in Fig. 5. It can observed from Fig. 5 that the proposed ICA-based detector perform superior than the detector operating without host-interference cancelation.

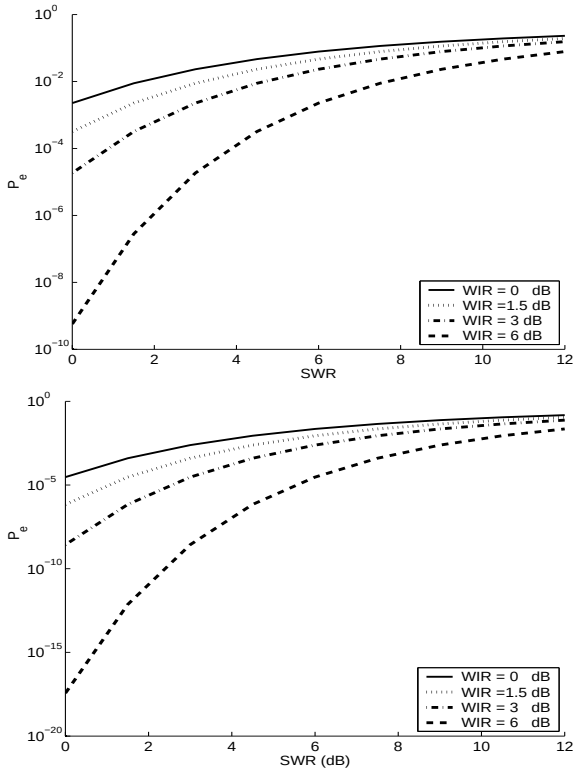


Figure 5: The P_e performance of the Proposed ICA-based Detector for different values of *WIR* and one bit per $|\zeta|$ Coefficients Embedding, i.e. $|\zeta| = 5$ (top) , $|\zeta| = 10$ (bottom) (theoretical values)

4.4 Maximum Watermarking-Rate: *Performance Analysis*

Maximum watermarking-rate (MWR) is another watermarking performance measure which indirectly depends on

the detector structure. Researchers in data hiding community have proposed various host interference suppression methods based on linear as well as non-liner filtering to improve MWR performance of a blind detector. For example, Su et al in [44] used optimal linear filtering to suppress host interference at the blind detector to improve MWR. The MWR performance of the proposed ICA-based watermark detector is evaluated for one bit per coefficient embedding case, i.e, $X_0 = S_0 + \alpha_0 W_0 b$. Let us assume the the received watermarked sample is corrupted by independent additive white Gaussian noise, with mean zero and variance $\sigma^2_{n_0}$. Here using CLT $\tilde{X}_0$ can be approximated by a Gaussian r.v. with mean zero and variance

$$\sigma^2_{\tilde{x}_0} = \sigma^2_{s_0} + \sigma^2_{v_0} + \sigma^2_{n_0}. \tag{71}$$

In this case, the estimated watermark sample, $\hat{V}_0$ can be expressed as,

$$\hat{V}_0 = \eta_{10}\alpha_0 W_0 + \eta_{20} S_0 + N_0 \tag{72}$$

Again $\hat{V}_0$ can also be approximated by Gaussian r.v. with mean zero and variance

$$\sigma^2_{\hat{v}_0} = \eta^2_{10}\sigma^2_{v_0} + \eta^2_{20}\sigma^2_{s_0} + \sigma^2_{n_0}. \tag{73}$$

The MWR of watermarking schemes based on additive embedding using blind correlation-based watermark detector can be approximated by the capacity of an additive white Gaussian noise channel, i.e.,

$$R_{Cor} = \frac{1}{2}\log_2\left(1 + \frac{\sigma^2_{v_0}}{\sigma^2_{s_0} + \sigma^2_{n_0}}\right) \tag{74}$$

Similarly, MWR using an informed detector can be expressed as,

$$R_{Informed} = \frac{1}{2}\log_2\left(1 + \frac{\sigma^2_{v_0}}{\sigma^2_{s_0}}\right) \tag{75}$$

And, MWR of the proposed ICA-based watermark detector is given as,

$$R_{ICA} = \frac{1}{2}\log_2\left(1 + \frac{\hat{\eta}^2_{10}\sigma^2_{v_0}}{\hat{\eta}^2_{20}\sigma^2_{s_0} + \sigma^2_{s_0}}\right) \tag{76}$$

Since the ICA scheme used for watermark estimation has reasonably good source separation performance [34, 35], therefore following inequality will hold,

$$\frac{\hat{\eta}^2_{10}\sigma^2_{v_0}}{\hat{\eta}^2_{20}\sigma^2_{s_0} + \sigma^2_{n_0}} \leq \frac{\sigma^2_{v_0}}{\sigma^2_{s_0} + \sigma^2_{n_0}} \tag{77}$$

$$\Rightarrow R_{ICA} \geq R_{Cor} \tag{78}$$

It can be observed from Eq. (78) that the proposed ICA-based detector performs better than the blind detector operating without suppressing the host signal interference. In addition, MWR performance of ICA-based detector is bounded below by the blind detector (0% suppression) and bounded above by an informed detector (100% suppression).

Performance analysis of the proposed ICA-based watermark detector indicates that it performs better than existing blind watermark detectors [1, 2, 3, 4, 5] operating without reducing host signal interference. This improved detection performance of ICA-based detector can be attributed to its host signal interference suppression at the detector.

4.5 Estimation of Masking Threshold, Distribution Parameter and WIR factor

This section provides details on how to estimate masking threshold, $\hat{\alpha}$, host distribution parameter, $\hat{\beta}$, and $\hat{\eta}_1$, $\hat{\eta}_2$ at the blind detector. The $\tilde{\mathbf{x}}$ is analyzed at the blind detector to estimate $\hat{\alpha}$ based on HAS. It is reasonable to assume that $\hat{\alpha}$ estimated from watermarked audio is similar to the $\hat{\alpha}$ from the corresponding an unwatermarked audio clip given that embedding and attack-channel distortion are imperceptible. To validate this assumption, we estimated $\hat{\alpha}$ from both the unwatermarked and corresponding watermarked music clips. To this end four music clips (*Pos1, Pop2, Classic*, and *Vocal*) listed in Table 1) were used. Here music clips *Pop1* and *Classic* were watermarked using FSSS based watermarking scheme proposed in [5] and *Pop2* and *Vocal*, were watermarking using audio watermarking scheme presented in [3]. Plots of the $\hat{\alpha}$ estimated from the each watermarked music clip, $\hat{\alpha}_W$ and corresponding unwatermarked music clip $\hat{\alpha}_{UW}$ are given in Fig. 6.

It can be observed from Fig. 6 that for both embedding schemes $\hat{\alpha}_W \approx \hat{\alpha}_{UW}$. Similarity between $\hat{\alpha}_W$ and $\hat{\alpha}_{UW}$, for the music clips listed in Table 1, in terms of mean squared error (MSE) (in dB) is {*Pos1, Melodic, Pop2, Classic, Vocal*} = {0.21566, 1.7321, 2.4507 , 1.7716, 0.21566}. Here watermarked music clips were generated using FSSS-based watermarking. These results shows that it is reasonable to estimated masking threshold from the watermarked audio at the blind detector.

Distribution parameter, β, can be estimated from the estimated variance $\hat{\sigma}^2_s$ of the host audio, which can be estimated from the watermarked audio available at the detector

$$\hat{\sigma}^2_s = \hat{\sigma}^2_x - \frac{1}{M}\sum_j \hat{\alpha}^2_j \tag{79}$$

where $\hat{\alpha}^2_m$ is the variance of the watermark sequence for m^{th} audio segment and M is total number of watermarked segments.

Here $\hat{\sigma}^2_x$ is estimated using sample variance, i.e.,

$$\hat{\sigma}^2_x = \frac{1}{M}\sum_j \mathbf{X}^2{}_j - \frac{1}{M^2}\left(\sum_j \mathbf{X}_j\right)^2 \tag{80}$$

It is important to mentions that if this estimate is used to calculate sufficient statistics, this will introduce additional dependence between watermark and sufficient statistics which is hard to analyze theoretically. Due to this

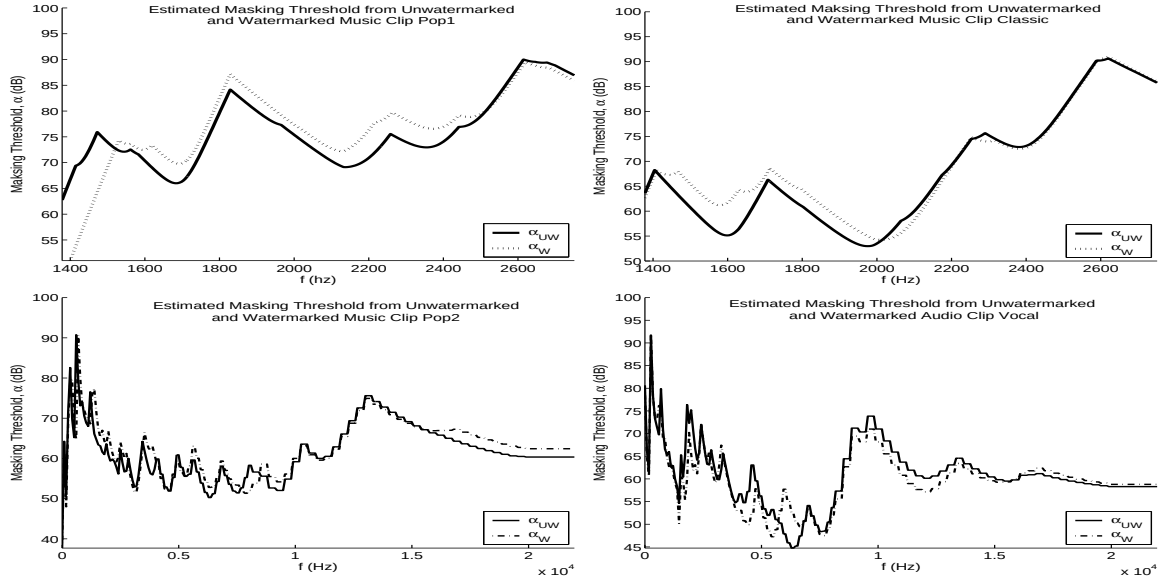


Figure 6: Plots of the estimated masking threshold from watermarked music clips $\hat{\alpha}_W$ and unwatermarked music clips $\hat{\alpha}_{UW}$

added dependence, slight variation between theoretical approximation and experimental results is expected.

The problem of estimating η_1 and η_2 is bit hard due to ambiguity in the scale and sign of the estimated sources using ICA. However, if we assume that scale and sign ambiguity of the separated sources is resolved and *WIR* factor $D^{\text{interf}} = \frac{\eta_1}{\eta_2}$ is known then, using Eq. (23) and (24), η_1 and η_2 can be estimated by simultaneously solving the following expressions,

$$\hat{\sigma}_{\hat{v}}^2 = \eta_1^2\hat{\alpha}^2 + \eta_2^2\hat{\sigma}_s^2, \eta_1^2 = D^{\text{interf}}\eta_2^2. \tag{81}$$

Here D^{interf} can calculated using separation quality measure of the ICA method used as discussed in [34, 35].

5 Simulation Results

This section provides detection performance of the proposed ICA-based watermark detector (ICAWD) and its comparison with the conventional normalized correlation watermark detector (NCWD) [1]. The proposed ICAWD can be used to detect watermark for almost all existing SS-based watermark embedding schemes [1, 2, 3, 5, 4]. However detection performance of the proposed detector is compared with Swanson et al's SS-based audio watermarking scheme [3]. Swanson et al's [3] proposed scheme used correlation based detector for watermark detection. To provide a fair performance comparison of both the proposed ICAWD and the NCWD, the proposed ICAWD is used in the estimation-correlation-based detection settings. The simulation results presented based on FSSS-based audio watermark embedding scheme presented in [5]. Detains of watermark embedding using FSSS [5] outlined here.

5.1 FSSS-based Watermark Embedding

The block diagram of the FSSS-based watermark generation and embedding used for simulations is illustrated in Fig. 7. The watermark is generated using a pseudo-random noise generator obeying non-Gaussian distribution to satisfy the non-Gaussianity requirement of the ICA model. A secret key $\mathrm{K_w}$ is used as a '*seed*' for the pseudo-random noise generator for watermark generation. In addition, same watermark is embedded in two consecutive audio segments, i.e., if watermark $\mathbf{V}$ is embedded into i^{th} audio segment then same watermark is also embedded into $(i+1)^{th}$ segment. Repeated embedding is a necessary condition of the proposed detector to separate hidden signals obeying heavy-tail distribution, especially for BSS from underdetermined linear mixtures [40, 16]. For audio watermarking using FSSS, a secret key, K_{sb} is used to select subband from watermark embedding.

5.2 Watermark Detection

The proposed modified ICA-based detector has access to the secret key K only, which is combination of K_{sb} and $\mathrm{K_w}$, i.e., $\mathrm{K} = \mathrm{K}_{sb}|\mathrm{K_w}$. The watermark detection process for FSSS-based audio watermarking under proposed detection scheme consists of watermark estimation using ICA framework followed by correlation based detection. The main steps of the detection process are outlined below:

- ***Sync Point Extraction***: The received audio signal is analyzed first to extract the set of sync points (SP) [4, 5] used to combat desynchronization attacks.
- ***Segmentation***: An audio frame consisting of n-samples is selected around each $SP_i : i = 1, 2 \cdots M$. Where M is cardinality of SP set.

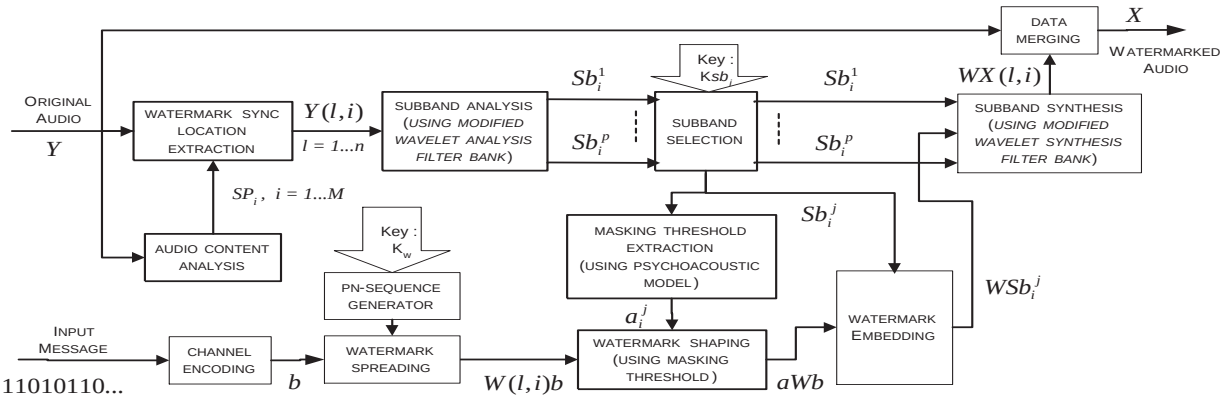


Figure 7: Block diagram of the FSSS-based watermark embedding

– ***Frame Decomposition***: Each frame is then decomposed into p–subband signals using l–level analysis filter bank described in [5] .

– ***Subband Selection***: A secret key K_{sb_i}, is used to select a subband from lower $(p-1)$–subbands of i^{th} and $(i+1)^{th}$ frame i.e. $\widetilde{\mathbf{Sb}}_i^j$, and $\widetilde{\mathbf{Sb}}_{i+1}^j$.

– ***Watermark Estimation***: The selected subband signals, i.e. $\widetilde{\mathbf{Sb}}_i^j$ and $\widetilde{\mathbf{Sb}}_{i+1}^j$ are used to estimated the embedded watermark, $\mathbf{V}$ Here, observation matrix, $\mathbf{X}$, can be expressed as, $\mathbf{X} = \left[\widetilde{\mathbf{Sb}}_i^j, \widetilde{\mathbf{Sb}}_{i+1}^j\right]^T$.

Existing BSS schemes for underdetermined mixtures based on ICA model [27, 39] to estimate watermark from the watermarked image for the proposed detector. However, in this paper, the proposed ICAWD uses the statistical ICA using mean-field approaches presented in [39] for watermark estimation from the watermarked audio. The watermark detection stage uses the correlation based similarity measure to determine the presence or the absence of the embedded watermark from the estimated sources. It is important to mention that permutation ambiguity in the estimated sources using ICA will contribute nonzero P_e due to incorrect source decoding. The error due to ambiguity in the permutation of the estimated sources is reduced by adding correlation based watermark detection (resp. decoding). However for the sake of simplicity, during analysis part in Section 4, error due to incorrect source decoding is neglected here.

– ***Information Decoding***: A binary hypothesis test is used to determine the presence or the absence of the embedded watermark in the estimated signal. For fast and reliable information decoding, normalized correlation between the estimated watermark and the key dependent watermark generated at the watermark detector are used. The normalized correlation is then compared against decision threshold, *Th*, to determine the presence or the absence of watermark. Following binary hypothesis test is used to decode binary information,

$$H_1: \quad \max \left| ncor\left(\hat{\mathbf{S}}^{(r)}, \mathbf{W}^{(q)}\right)\right| \geq Th \text{ Decode } q$$
$$H_0: \quad \text{otherwise } \textit{no watermark}$$

where *ncor*(.,.) is the normalized correlation function defined as:

$$ncor\left(\hat{\mathbf{S}}^{(r)}, \mathbf{W}^{(q)}\right) = \frac{\sum_{l=-n}^{n} \hat{S}_l^{(r)} W_{t+l}^{(q)}}{\sqrt{\sum_{l=0}^{n}\left(\hat{S}_l^{(r)}\right)^2 \sum_{l=0}^{n}\left(W_l^{(q)}\right)^2}} \tag{82}$$

where $\hat{\mathbf{S}}^{(r)}$ is the estimated signals using ICA, *Th* is the decision threshold (for our simulation results Th was set to 0.15, which corresponds to false positive rate, $P_{fp} = 10^{-4}$), $r = 1, 2, 3$, and $q \in \{0, 1\}$.

5.3 Experimental Results

To evaluate the robustness performance of the proposed ICAWD, several experimental tests were performed in which the watermarked audio is subjected to commonly encountered degradations. These degradations include addition of white and colored noise, resampling, lossy compression (MP3 Audio compression), filtering, time- and frequency-scaling, and stirmark benchmark attacks for audio [18, 20].

Decoding error probability, Pb_e, at the watermark detector is used for performance evaluation. Here Pb_e is defined as,

$$Pb_e = \left(1 - \frac{N_d}{N_e}\right) \tag{83}$$

where N_d is number of bits correctly detected and N_e number of bits embedded into the audio clip.

Block diagram of the proposed ICAWD and the traditional correlation based detector e.g. NCWD used for FSSS audio watermark detection process is given in Fig. 8. The watermark detector given in Fig. 8 acts as ICAWD when switch S is connected to terminal 1 and NCWD when S is connected to 2.

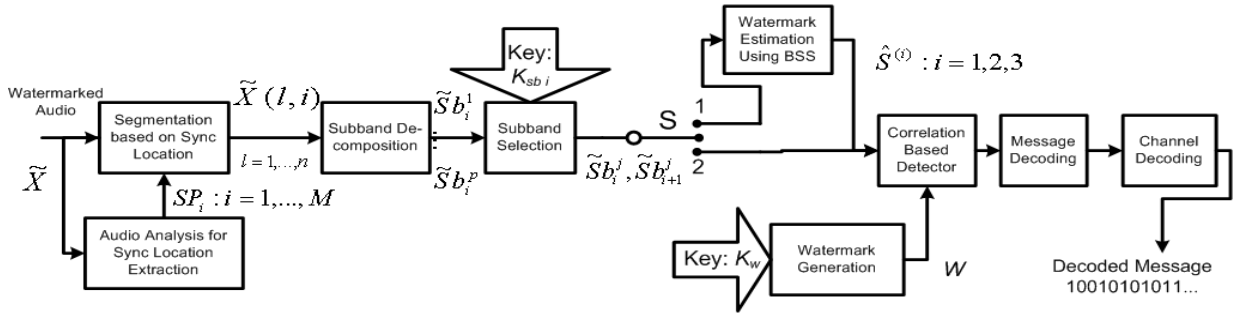


Figure 8: Block diagram of the ICAWD and the NCWD used for performance comparison

5.4 Robustness Performance

To evaluate the robustness performance of the proposed watermarking scheme we have performed several experimental tests in which the watermarked audio is subjected to commonly encountered degradations. These degradations include addition of white and colored noise, resampling, lossy compression (MPEG audio compression), filtering, time- and frequency-scaling, multiple watermarking, and StirMark benchmark attacks for audio.

The robustness performance of the proposed scheme against common degradations for the above settings is discussed next.

5.5 Data Set

Experimental results presented here are based on the data set consisting of the *sound quality assessment material* (SQAM) audio database downloaded from [45] and five audio clips listed in Table 1. All audio clips used for the performance evaluation here are based on mono audio channel sampled at 44.1 kHz with 16 bits resolution.

In our experiments, the watermarks are generated and embedded using FSSS-based audio watermarking scheme presented in [5]. A perceptual mask is estimated using method discussed in [5]. This mask is then multiplied by 200 independently generated pseudo-random sequences $\mathbf{W}$, with zero-mean and unit variance, to generate 200 independent watermarks. In case of ICAWD, the pseudo-random sequences, $\mathbf{W}$, follow Laplacian distribution, i.e.,

$$f_W(\tau) = \frac{\beta}{2} e^{-\beta|\tau|}, \quad |\tau| < \infty \tag{84}$$

where $\beta = \frac{\sqrt{2}}{\sigma_{\mathbf{W}}}$, and for the NCWD $\mathbf{W}$ follows normal distribution. These 200 random watermarks are embedded in each audio clip according to Eq. (1) that resulted 4000 watermarked audio clips. Experimental results presented in the following sections are averaged over 4000 watermarked audio clips.

5.6 Parameter Settings

Simulation results presented in this section are based on the following system settings:

- Salient point list (SP) was assumed to be available at the detector, therefore decoding bit error probability P_e presented here is due decoding bit error only.
- Audio frame size ($2^l N_1$) was set to 2^{13} for $f_s = 44.1$ kHz.
- Five-level wavelet decomposition was used, i.e. $l = 5$, therefore eight target subbands were available for watermark embedding.
- Only one subband was selected at random from eight target subbands for watermark embedding (except multiple watermark embedding case).
- Target false positive rate P_{fp} was set to 3.5×10^{-4} which corresponds to decoding threshold $Th = 0.15$ (using Eq. (42)).
- False positive bit rate, P_{fp}, was calculated by applying original (unwatermarked) music clip the proposed detector, and average false positive for the the 20 audio clips used for performance evaluation was calculated to be 2.9×10^{-4}.
- Robustness performance in terms of average decoding bit error rate was calculated without channel coding.
- In case of ICAWD, watermark repeating factor of two was used during watermark embedding process, i.e., two consecutive audio frames were watermarked with same watermark $\mathbf{w}$.

The above settings for watermark embedding using FSSS-based audio watermarking yielded *per sample embedding capacity* of 1 bit per 512 sample.

Fidelity (or transparency) performance of the embedded watermark is evaluated based on the objective degradation measure. Signal-to-watermark ratio (SWR) is used for the objective degradation here which is calculated as,

Table 1: Audio Clips used for Performance Evaluation

Singer Name, Song Title	***Genre***	***Duration*** (sec)
Back Street Boys, *I Want It That Way*$\cdots$	Pop, (Pop1)	22
L. Mangeshkar, *Kuch Na Kaho*$\cdots$	Melodic, (Melodic) (Melodic)	15
A. Bhosle, & R. Sharma, *Kahin Aag Laga*$\cdots$	Pop, (Pop2) (Pop2)	10
N. F. A. Khan, *Afreen Afreen*$\cdots$	Semi-Classic, (Classical)	20
Suzanne Vega, *Tom's diner*$\cdots$	Female Vocal, (Spoken Language)	5

$$SWR = 10\log_{10}\left(\frac{\sigma_s^2}{\sigma_\nu^2}\right) \quad (85)$$

where σ_ν^2 is calculated using Eq. (3).

The average SWR the watermark audio clips used for simulation was $Ave_{SWR} = 42.7$ (dB), $\sigma_{SWR} = 9.17$, $max_{SWR} = 74.5$ (dB), and $min_{SWR} = 21.5$ (dB). Calculated SWR from watermarked audio clips indicates that on the embedded watermark is very weak compared to the original audio.

5.7 Detection Performance

Detection performance of the proposed detector is evaluated for various audio degradations. Detection of the proposed ICAWD and its comparison with NCWD for each degradation is provided next.

5.7.1 Addition of White Noise

: White Gaussian noise ranging from zero to 200 % of the power of the audio signal was added to the corresponding watermarked audio clips. The P_e average over 4000 watermarked audio clips for ICAWD and NCWD for different SNR values are plotted in Fig. 9 which shows that the ICAWD performs better than the NCWD. Superior detection performance of ICAWD than the NCWD can be attributed to its host signal interference cancellation capability. It can be observed from Fig. 9 that for SS-based watermarking very low decoding bit error probability is achievable even in the presence of noise with 60 - 70 % power of the audio signal.

5.7.2 Resampling

To simulate resampling attack, a watermarked audio signal was down-sampled at a sampling rate of $\frac{f_s}{r_f}$ (where r_f denotes resampling factor) and then interpolated back to f_s. The watermark detection was then applied to the resulting watermarked audio clips. Average P_e for $r_f = 2, \cdots 10$, is given in Fig. 10 which shows that the proposed watermarking scheme (using ICAWD) can withstand resampling attacks with r_f value up to 5 for each watermarked audio clip, similar decoding performance is achievable for NCWD by using channel coding. Again ICAWD performs better than the NCWD and its superior detection performance can be attributed to its host signal interference suppression capability.

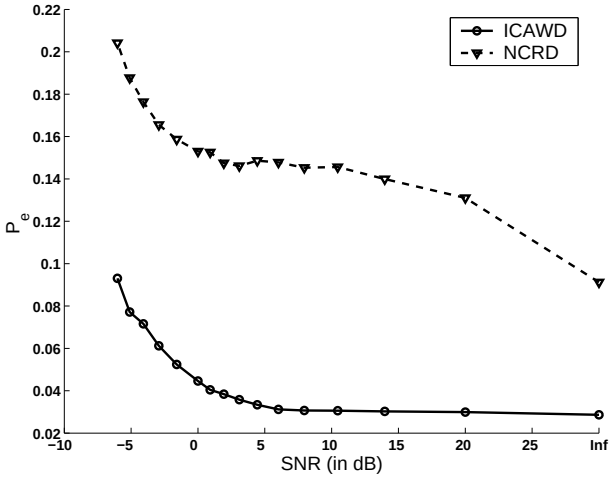


Figure 9: Detection performance comparison for AWGN attack.

5.7.3 Lossy Compression

Lossy compression for audio (e.g. MP3) is generally applied to the digital audio for multimedia applications like transmission and storage to reduce the bit rate. To test the survivability of the watermark, audio encoding/decoding was applied to the watermarked audio using ISO/MPEG-1 Audio Layer III [47] coder at bit rates 32, 64, 96, 112, 128, 192, 256, and 320 k bits/s (kbps). The average P_e for lossy compression attacks for bit rates rates 32, 64, 96, 112, 128, 192, 256, and 320 (kbps) is given in Fig. 11. It has been observed from Fig. 11 that the detection per-

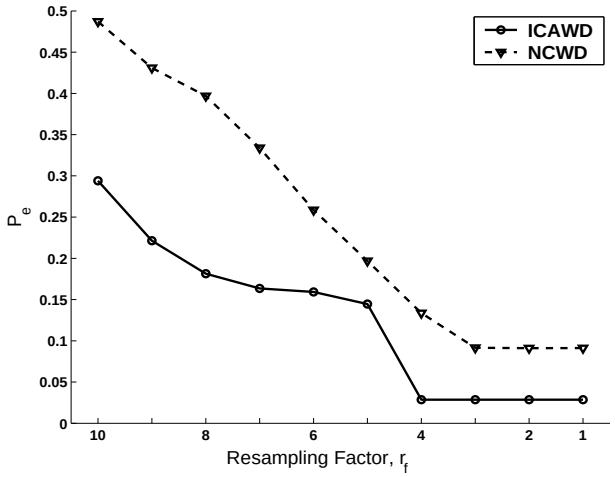


Figure 10: Detection performance comparison for resampling attack.

formance for both detectors deteriorates as the bit rate of the encoder/decoder decreases; this is due to the stronger distortion introduced by the encoder for lower bit rates. In addition, the ICAWD performs better than the NCWD.

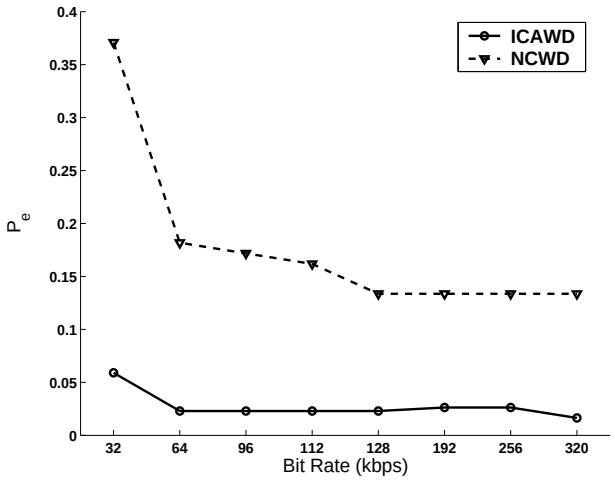


Figure 11: Detection performance comparison for MP3 compression attack

5.7.4 Addition of Colored Noise

To simulate an attack with colored noise, white Gaussian noise was spectrally shaped according to the estimated masking threshold using corresponding watermarked audio clip based on the HAS model [46, 47]. This just audible colored noise was then added to the watermarked audio signal. Average P_e for the resulting watermarked audio clips is presented in Fig. 12. It has been observed from Fig. 12 that NCWD performs poorly, this is due to increase in interference level, as the colored noised is generated with a process almost identically to that of the watermark generation. Therefore, additive colored noise acts as a second watermark interfering with the watermark to be detected. On the other hand, ICAWD is efficient in handling such attacks due to its interference cancellation ability.

5.7.5 Rescaling

Rescaling attacks include time- and frequency-scaling. Time-scaling attacks can be used to desynchronize a watermark detector for SS-based watermarking systems. To test the robustness of the proposed scheme against time-scaling attacks, the watermarked audio clips were time-scaled with time-scaling factor, TSp(n) = +(-) 1 %. The detection performance for time-scaling attack using both detection schemes, e.g., ICAWD and NCWD is given in Fig. 12.

The frequency-scaling attacks are generally used to deteriorate the detection performance of the frequency domain watermarking schemes. As the proposed watermarking scheme is also a frequency domain watermarking scheme; therefore, it is reasonable to test the robustness performance of the proposed scheme against frequency-scaling attacks as well. To simulate frequency-scaling attack, the watermarked audio clips were frequency-scaled using frequency-scaling factor, FSp(n) = +(-) 1%. The detection performance for the resulting audio clips for both detection schemes, e.g., ICAWD and NCWD is presented in Fig. 12. It can be observed from Fig. 12 that the proposed scheme can withstand rescaling attack of TS $\leq \pm 1\%$ and FS$\leq \pm 1\%$ (especially for ICAWD).

5.7.6 Filtering

To test the robustness of the proposed watermarking scheme against filtering attacks, the watermarked audio signals were subjected to lowpass filtering (LPF), highpass filtering (HPF), and bandpass filtering (BPF) attacks. The specification of filters used for the filtering attacks are,

1. Lowpass Filter: cut-off frequency: $f_c = 5$ kHz with 12 dB/octave roll-off
2. Highpass Filter: cut-off frequency: $f_c = 1000$ Hz with 12 dB/octave roll-off
3. Bandpass Filter: cut-off frequencies: $f_{clow} = 50$ Hz, and $f_{cup} = 5.5$ kHz with 12 dB/octave roll-off

Detection performance comparison of the ICAWD and the NCWD for LPF, HPF, and BPF attacks is given in Fig. 12.

5.7.7 StirMark Audio Benchmark Attacks

For StirMark for audio benchmark attack, watermarked audio clips were subjected to StirMark audio benchmark attacks. The StirMark audio benchmark software, available at [20], was used in the default parameters settings. The decoding bit error probability, P_e, averaged over 100 watermarked audio clips with the ICAWD and the NCWD, is given in Table 2. It can be observed from Table 2 that

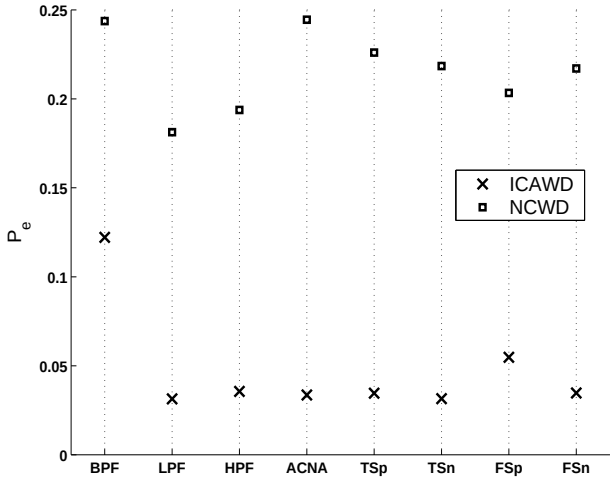


Figure 12: Detection performance comparison between the ICAWD and the NCWD for filtering (LPF, HPF, BPF), rescaling (TSp, TSn, FSp, FSn), requantization (Res), and colored noise addition (ACNA) attacks

the proposed ICAWD based scheme using exhibits superior detection performance than the NCWD. Better performance of ICAWD can be attributed to its better host signal suppression capability.

6 Conclusion

An improved watermark detector for additive embedding is presented here. The proposed watermark detector is capable of canceling the host-signal interference at the watermark detector. Bind watermark detection, lower host-signal interference at the detector, improved decoding, detections and watermarking-rate performances are the salient features of the proposed ICAWD. The proposed ICAWD can be used for SS-based watermarking for all types of multimedia data, e.g., audio, video, images, etc. The theoretical results show that the proposed detector performs significantly better than existing blind detectors. Simulation results for real-world data show that the proposed ICAWD performs much better than the traditional NCWD. Moreover, the detection performance of the proposed detector can be improved further by employing channel coding. It is important to mention that better detection performance of ICAWD comes at the cost of security, as ICAWD requires repeated embedding (at least twice) which makes embedded watermark more vulnerable to watermark estimation attacks than without repeated embedding.

References

[1] Cox, I.J., Miller, M.L., and Bloom, J.A.:(2001) *Digital Watermarking*, Morgan Kaufmann, San Francisco.

Table 2: Performance Comparison for StirMark Audio Benchmark Attacks

Decoding Bit Error Probability, P_e		
StirMark Attack NCWD	ICAWD	
addbrumm_100	0.088	0.0091
addbrumm_1100	0.088	0.0091
addbrumm_2100	0.088	0.0091
addbrumm_3100	0.1023	0.0091
addbrumm_4100	0.1257	0.0091
addbrumm_5100	0.1412	0.0091
addbrumm_6100	0.1477	0.0091
addbrumm_7100	0.1904	0.0091
addbrumm_8100	0.2228	0.0091
addbrumm_9100	0.2293	0.0234
addbrumm_10100	0.2293	0.0491
addfftnoise	1	1
addnoise_100	0.088	0.0491
addnoise_300	0.088	0.0491
addnoise_500	0.088	0.0491
addnoise_700	0.088	0.0634
addnoise_900	0.088	0.0634
addsinus	0.088	0.0634
amplify	0.088	0.0491
compressor	0.088	0.0491
copysamples	0.529	0.1749
cutsamples	0.791	0.4835
dynnoise	0.1056	0.0667
echo	0.0818	0.0667
exchange	0.1056	0.0818
extrastereo_30	0.1056	0.0818
extrastereo_50	0.1056	0.0818
extrastereo_70	0.1056	0.0818
fft_hlpass	0.1074	0
fft_invert	0.1056	0.0818
fft_real_reverse	0.1056	0.0818
fft_stat1	0.1295	0.0238
fft_test	0.1056	0.0238
flippsample	0.1281	0.0725
invert	0.088	0.0491
lsbzero	0.1056	0.0818
normalize	0.088	0.0673
rc_highpass	0.0945	0.0491
rc_lowpass	0.088	0
smooth	1	1
resample	0.1056	0
smooth2	0.1056	0
stat1	0.1056	0
stat2	0.1056	0.0818
voiceremove	1	1
zerocross	0.088	0
zeroremove	0.2759	0.0363
zerolength	0.2189	0.0607

[2] Cox, I.J., Kilian, J., Leighton, T., and Shamoon, T.:(1997) Secure Spread Spectrum Watermarking for Multimedia, *IEEE Trans. on Image Processing*, vol. 6(12), pp. 1673–1687.

[3] Swanson, M.D., Zhu, B., Tewfik, A.H., and Boney, L.:(1998) Robust Audio Watermarking using Perceptual Masking, *Signal Processing*, vol. 66(3), pp. 337–355.

[4] Wu, C.-P., Su, P.-C., and Kuo, C.-C. J.:(1999) Robust Audio Watermarking for Copyright Protection, *Proc. SPIE's* 44^{th} *Ann. Meet. Adv. Sig. Proc. Alg. Arch. Impl. IX (SD39)*, vol. 3807, pp. 387–397.

[5] Malik, H., Khokhar, A., and Ansari, R.:(2004) Robust Audio Watermarking using Frequency Selective Spread Spectrum Theory, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'04)*, Montreal, Canada, pp. 385–388.

[6] P-Gonzalez, F., Balado, F., and Hernández, J.R.:(2003) Performance Analysis of Existing and new Methods for Data Hiding with Known-Host Information in Additive Channels, *IEEE Trans. on Signal Processing*, vol. 51(4), pp. 960–980.

[7] Hernandez, J., Amado, M., and Perez-Gonzalez, F.:(2000) DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure, *IEEE Trans. on Image Processing*, vol. 9(1), pp. 55–68.

[8] Briassouli, A., and Strintzis, M.:(2004) Locally Optimum Nonlinearities for DCT Watermark Detection, *IEEE Trans. on Image Processing*, vol. 13(12), pp. 1604-1617.

[9] Noel, S., and Szu, H.:(2000) Multimedia Authenticity with ICA Watermarks, *IS&T/SPIE. Proc., Wavelet Applications VII*, vol. 4056, pp. 175–184.

[10] Serrano, F., and Fuentes, J.:(2001) Independent Component Analysis Applied to Digital Image Watermarking, *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'01)*, vol. 3, pp. 1997-2000.

[11] Toch, B., Lowe, D., and D. Saad, D.:(2003) Watermarking of Audio Signals using using Independent Component Analysis, *Proc.* 3^{rd} *Int. Conf. WEB Delivering of Music*, pp. 71-74.

[12] Sener, S., and Gunsel, B.:(2004) Blind Audio Watermark Decoding using Independent Component Analysis, *Proc.* 17^{th} *Int. Conf. Patt. Reco. (ICPR'04)*, vol. 2, pp. 875–878.

[13] Bounkong, S., Toch, B., Saad, D., and Lowe, D.:(2002) ICA for Watermarking Digital Images, *J. Machine Learning Research 1*, pp. 1–25.

[14] Yu, D., Sattar, F., and Ma, K.:(2002), Watermark Detection and Extraction using Independent Component Analysis, *EURASIP J. Applied Signal Processing*, pp. 92–104.

[15] Malik, H., Khokhar, A., and Ansari, R.:(2005) Improved Watermark Detection for Spread-Spectrum based Watermarking using Independent Component Analysis, *Proc.* 5^{th} *ACM Workshop On Digital Rights Management (DRM'05)*, Washington DC, pp. 102–111.

[16] Malik, H., Khokhar, A., and Ansari, R.:(2006) New detector for spread-spectrum based image watermarking using underdetermined ICA, *IS&T/SPIE Conf. Security, Steganography, and Watermarking of Multimedia Contents VIII '06*, vol. 6072, pp. 747–758.

[17] Malik, H., Khokhar, A., and Ansari, R.:(2006) Blind Detection for Additive Embedding Using Underdetermined ICA, *Proc.* 8^{th} *IEEE Int. Symposium on Multimedia, (ISMapos'06)*, pp. 758–761.

[18] Steinebach, M., Lang, A., Dittmann, J., and Prtitcolas, F.A.P.:(2002) StirMark Benchmark: Audio Watermarking Attacks based on Lossy Compression, *Proc. SPIE Security Watermarking Multimedia*, vol. 4675, pp. 79–90.

[19] Lang, A., Dittmann, J., Spring, R., and Vielhauer, C.:(2005), Audio watermark attacks : from single to profile attacks, *Proc. ACM Multimedia and Security Workshop (MM & Sec'05)*, New York, NY, USA, pp. 39–50.

[20] *StirMark Benchmark for Audio*, available at `http://amsl-smb.cs.uni-magdeburg.de/smfa/main.php`, accessed on June 23, 2008.

[21] Bell, A., and Sejnowski, T.:(1995) An Information Maximisation Approach to Blind Separation and Blind Deconvolution, *Neural Computation*, MIT Press Journals, vol. 7(6), pp. 1129–1159.

[22] Hyvarinen, A., Karhunen, J., and Oja E.:(2001) *Independent Component Analysis*, John Wiley & Sons.

[23] Comon, p.:(1994) Independent Component Analysis, A New Concept?, *EURASIP Signal Processing*, vol. 36(3), pp. 287–314.

[24] Cao, X.-R., and Liu, R.-W.:(1996) Genaral Approach to Blind Source Separation, *IEEE Trans. Signal Processing*, vol. 44(3), pp. 562–571.

[25] Eriksson, J., and Koivunen, V.:(2004) Identifiability, Separability, and Uniqueness of Linear ICA Models, *IEEE Signal Processing Letters*, vol. 11(7), pp.601–604.

[26] Davis, M.:(2004) Identifiability Issues in Noisy ICA, *IEEE Signal Processing Letters*, vol. 11(5), pp. 601–604.

[27] De Lathauwer, L., Comon, P., De Moor, B., and Vandewalle, J.:(1999) ICA Algorithms for 3 Sources and 2 Sensors, *Proc. IEEE Signal Processing Workshop Higher-Order Statistics (HOS'99)*, pp. 116–120.

[28] Gaeta, M., and Lacoume, J.-L.:(1990) Source Separation Without Prior Knowledge: The Maximum Likelihood Solution, *Proc. EUSIPCO'90*, pp. 621–624.

[29] Moulinesand, E., Cardoso, J-F., and Gassiat, E.:(1997) Maximum Likelihood for Blind Separation and Deconvolution of Noisy Signals using Mixture Models, *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'97)*, pp. 3617–3620.

[30] Cardoso, J.-F.:(1999) High-order Contrasts for Independent Component Analysis, *Nural Computation*, Elsevier Science, vol. 11(1), pp. 157–192.

[31] Bofill, P., and Zibulevsky, M.:(2001) Underdetermined Blind Source Separation using Sparse Representations, *EURASIP Signal Processing*, vol. 81(11), pp. 2353–2362.

[32] Zibulevsky, M., and Zeevi, Y.Y.:(2002), Extraction of a Single Source from Multichannel Data using Sparse Decomposition, *Neurocomputing*, Elsevier Science, vol. 49, pp. 163–173.

[33] Li, Y., Cichocki, A., and Amari, S.:(2004), Analysis of Sparse Representation and Blind Source Separation, *Neural Computation*, MIT Press Journals, vol. 16(6), pp. 1193–1234.

[34] Gribonval, R., Benaroya, L., Vincent, E., and Fevotte, C.:(2003) Proposals for Performance Measurement in Source Separation, *Proc.* 4^{th} *Int. Sym. Independent Component Analysis and Blind Source Separation*, pp. 763–768.

[35] Li, Y., Powers, D., and Peach, J.:(2000) Comparison of Blind Source Separation Algorithms, *Advances in Neural Networks and Applicaitons*, N. Mastorakis (Ed.), WSES, pp. 18–21.

[36] Cichocki, A., Douglas, S., and Amari, S.:(1998) Robust Techniques for Independent Component Analysis (ICA) with Noisy Data, *Neurocomputing*, Elsevier Science, vol. 22, pp. 113–129.

[37] Pajunen, P.:(1997) Blind Separation of Binary Sources With Less Sensors Than Sources, *Proc. Int. Conf. on Neural Networks (ICNN'97)*, vol. 3, pp. 1994–1997.

[38] Hyvarinen, A.:(1999) Fast Independent Component Analysis with Noisy Data using Gaussian Moments, *Proc. ISCS'99*.

[39] Hojen-Sorensen, P., Winther, O., and Hansen, L.K.:(2002) Mean-Field Approaches to Independent Component Analysis, *Neural Computation*, MIT Press Journals, vol. 14, pp. 889–918.

[40] Hansen, L.K., and Petersen, K.B.:(2003) Monoaural ICA of While Noise Mixture is Hard, *Proc. of Sym ICA and BSS (ICA2003)*, pp. 815–820.

[41] Poor, H. V.:(1994) *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, 2^{nd}-ed.

[42] Kay, S.:(1998) *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice Hall, Upper Saddle River, New Jersey.

[43] Swanson, M., Zhu, B., and Tewfik, A.:(1996) Robust Data Hiding for Images, *Proc. IEEE Digital Signal Processing Workshop*, pp. 37–40.

[44] J. Su, J., Eggers, J., and Girod. B.:(2001) Analysis of Digital Watermarks Subjected to Optimum Linear Filtering and Additive Noise, *EURASIP Signal Processing*, vol. 81, pp. 1141–1175.

[45] *SQAM - Sound Quality Assessment Material*, `http://sound.media.mit.edu/mpeg4/audio/sqam/`, accessed on June 23, 2008.

[46] Zwicker, R. E., and Fastl, H.:(1999) *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin.

[47] Noll, P.:(1997) MPEG Digital Audio Coding, *IEEE Signal Processing Magazine*, vol. 14(5), pp. 59–81.

[48] Papoulis, A., and Pillai, S.:(2002) *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 4^{th} Ed.

[49] Mallat, S.:(1989) A Theory for Multiresolution Signal Decomposition, the Wavelet Representation *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11(7), pp. 674 –693.

7 Appendix A: Statistical characterization of the wavelet coefficients of audio signals

To determine the statistical characterization of the subband coefficients of the real-world speech samples, the speech samples, $Y_i,\ i = 0, 1 \cdots n-1$ are assumed to be i.i.d. Laplacian random variable with mean zero and variance σ_y^2. The one-dimensional discrete wavelet transform (DWT) of audio signal, $\mathbf{Y}$, can be calculated using Mallat's algorithm [49]. The DWT coefficients using Mallat's algorithm [49], e.g., approximate coefficients a_k and detailed

coefficients d_k, at different scales can be expressed as,

$$a_i^{j-1} = \sum_{k=0}^{N_1-1} h_{k-2i} a_k^j \tag{86}$$

$$d_i^{j-1} = \sum_{k=0}^{N_1-1} g_{k-2i} a_k^j \tag{87}$$

where j denotes the resolution and i is the index.

Eq. (86) and (87) describe linear filtering operation using filters $\mathbf{h}$ and $\mathbf{g}$ followed by down-sampling. Here $\mathbf{h}$ and $\mathbf{g}$ are finite impulse response (FIR) quadrature-mirror filters, also known as the scaling and the wavelet filters, respectively. The scaling filter is a lowpass filter, while the wavelet filter is a highpass filter. Moreover, the top-level coefficients a^J represent the original signal $\mathbf{y}$. Eq. (86) and (87) can be expressed using a single equation,

$$S_i^{j-1} = \sum_{k=0}^{N_1-1} \delta_{k-2i} S_k^j \tag{88}$$

where δ_i is the weighting factor depending on the filter coefficients h_i and g_i, i.e. approximate coefficients or detailed coefficients, and S_i^j is the wavelet coefficient at j^{th}-level.

Here Eq. (88) states that a wavelet coefficient at an arbitrary level $j-1$, is a weighted sum of N_1 wavelet coefficients from j^{th}-level wavelet. The wavelet coefficients at $J-1$ level can be expressed as

$$S_i^{J-1} = \sum_{k=0}^{N_1-1} \delta_{k-2i} S_k^J \tag{89}$$

According to Eq. (88), each wavelet coefficient an arbitrary level $j : 1 \leq j \leq J-1$ is a weighted sum of i.i.d. r.v. (e.g. audio samples in our case), therefore, the pdf of a wavelet coefficient S_i^j at j^{th}-level, can be determined using joint characteristic function $\Phi_{s_i^j}(\omega)$. If we assume that audio sample Y_i, is a Laplacian r.v., then pdf of Y_i can be expressed as,

$$f_y(\tau) = \frac{\gamma}{2} e^{-\gamma|\tau|}, \ |\tau| < \infty \tag{90}$$

where $\gamma = \frac{\sqrt{2}}{\sigma_y^2}$

Here characteristic function of r.v. Y_i, $\Phi_{y_i}(\omega)$, can be expressed as [48],

$$\Phi_{y_i}(\omega) = \frac{\gamma_i^2}{\omega^2} \tag{91}$$

Let us consider a r.v. Z which is obtained by magnitude scaling of a r.v. Y i.e., $Z = \delta Y$, the characteristic function of Z, $\Phi_z(\omega)$, in terms of $\Phi_y(\omega)$ can be expressed as [48],

$$\Phi_z(\omega) = \Phi_z(\delta\omega) \tag{92}$$

Therefore, the characteristic function of r.v. S_i^{J-1},$\Phi_{s_i^{J-1}}(\omega)$, can be expressed as

$$\Phi_{s_i^{J-1}}(\omega) = \prod_{k=1}^{N_1} \Phi_{y_k}\left(\delta_{k-2i}\omega\right) \tag{93}$$

$$= \prod_{k=1}^{N_1} \frac{\gamma_i^2}{\left(\gamma_i^2 + \left(\delta_{k-2i}\omega\right)^2\right)} \tag{94}$$

$$= \prod_{k=1}^{N_1} \frac{\acute{\gamma}_i^2}{(\acute{\gamma}_i^2 + \omega^2)} \tag{95}$$

where $\acute{\gamma}_i = \gamma_i/\gamma_{k-2i}$ and N_1 is the length of the wavelet filter.

In order to determine the pdf of wavelet coefficients S_i^{J-1}, $f_{s_i^{J-1}}(\tau)$ characteristic function $\Phi_{s_i^{J-1}}(\omega)$ (given by Eq. (95)) is used. The pdf of a r.v. can be determined either using the uniqueness theorem or the convolution theorem [48]. The pdf of wavelet coefficients S_i^{J-1}, $f_{s_i^{J-1}}(\tau)$ using $\Phi_{s_i^{J-1}}(\omega)$ based on the convolution theorem can be expressed,

$$f_{s_i^{J-1}}(\tau) = \frac{\acute{\gamma}_i}{2} e^{-\acute{\gamma}_i|t|} \left(\sum_{k=0}^{N_1-1} c_i^k \acute{\gamma}_i^k t^k \right) \tag{96}$$

where $c_k \in \mathcal{R}$ is a real constant.

For different values of N_1, the polynomial coefficients, c_k, are given as:

$N_1 = 2, c_0 = c_1 = \frac{1}{2}$, and $N_1 = 3, c_0 = c_1 = \frac{3}{8}$, and $c_1 = \frac{1}{8}$ and so on.

According to the Eq. (95) and (96), as the pdf of wavelet coefficients at j_{th} level, $f_{s_i^{j-1}}(\tau)$, is a obtained by convolving the pdf of r.v. Y_i, therefore, based on the CLT, the pdf of the subband coefficients move to words Gaussianity as value of N_1 increases or in other words, pdf of wavelet coefficients at coarser level is closer to the Gaussianity than higher level coefficients. This is because at coarser level, for each wavelet coefficient more audio samples contribute in the weighted-sum equation (given by Eq. (89)) than higher level coefficients.

In order to provide evidence in support of this model, a 4–level DWT decomposition of an arbitrary frame of the music clip *I Want It That Way*$\cdots$ by *Backstreet Boys*, using '*Daubechies*-8' decomposition filter, is given in Fig. 13. The pdf (based on histogram approximation) of corresponding wavelet coefficients at different levels is plotted in Fig. 13. This is clear from Fig. 13 that the higher level, wavelet coefficients exhibit non-Gaussian distribution and distribution moves towards Gaussianity for coarser coefficients due to longer weighted-sum effect at the coarser level.

Therefore, the pdf of each subband coefficient (at higher level) of the host signal, S_i, can be approximated by Laplacian distribution, which is given as,

$$f_s(\tau) = \frac{\beta}{2} e^{-\beta|\tau|} : |\tau| < \infty \tag{97}$$

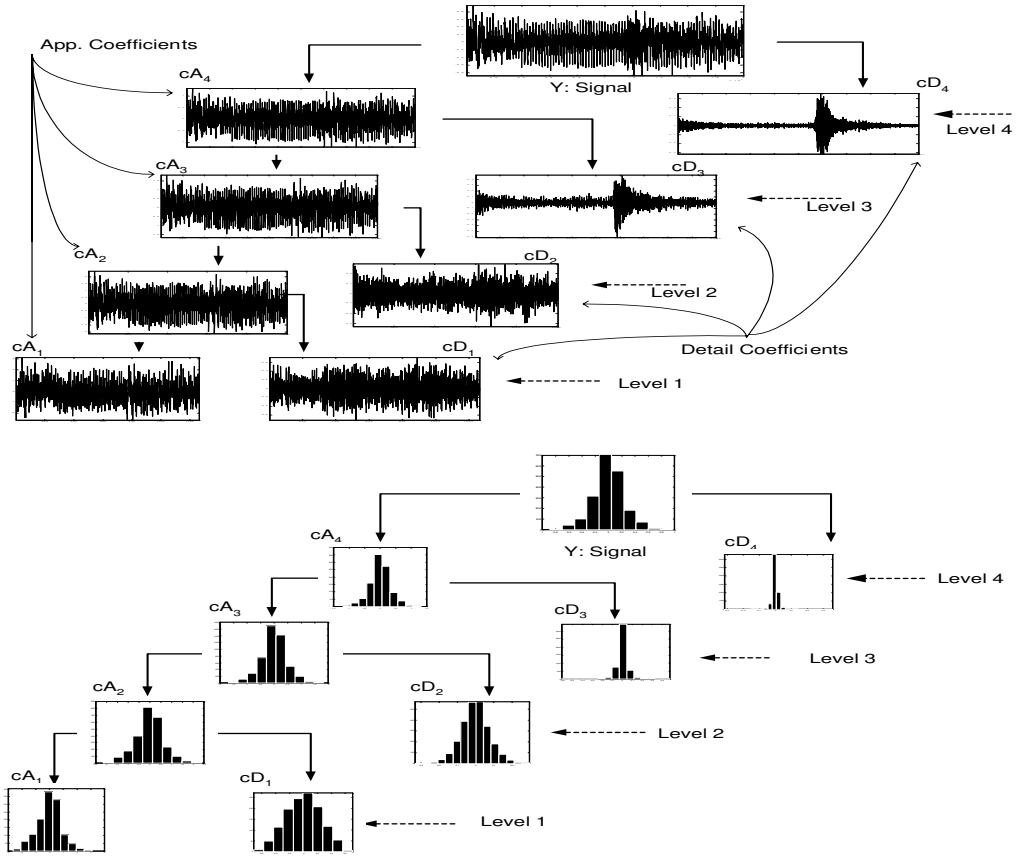


Figure 13: Plots of empirical distribution based histogram approximation of detailed and approximate coefficients at each level of 4-Level wavelet decomposition of an audio signal, $\mathbf{y}$

where $\beta = \frac{\sqrt{2}}{\sigma_s^2}$

Visual Security Assessment for Cipher-Images Based on Neighborhood Similarity

Ye Yao, Zhengquan Xu and Jing Sun
Liesmars, Wuhan University, Wuhan 430079,
Hubei, China
E-mail: xuzq@whu.edu.cn

Keywords: cipher-images, visual security, objective assessment, neighborhood similarity

Received: September 7, 2008

In the recent decades, many practical algorithms have been put forward for images and videos encryption. However, there is no objective security assessment algorithm or calculation index has been proposed at present. According to the differences of pixel value and neighborhood distribution between cipher-images and original images, we present a visual security assessment algorithm based on neighborhood similarity. The experiment result shows that the scheme can provide an objective assessment which is match up to subjective assessment, and is also suitable for the security assessment of cipher-images produced by other selective encryption algorithms.

Povzetek: Analizirana je ocena varnosti kodiranja slik.

1 Introduction

With the rapid development of information and network technology, the acquisition and transmission of visual media have been developed at a higher speed than ever. The visual media has been extensively applied to many key departments and fields which are closely related to people's livelihood as well as national security. As a result, the security of visual media (images & videos) is becoming more and more important. In the recent decades, many practical algorithms have been put forward for images and videos encryption[1].

Compressed bitstreams of images and videos become to be cipher-bitstreams when they are encrypted by selective encryption algorithms [1] that can maintain bitstream format compatibility. If cipher-bitstreams are directly inputted to standard decoder and are decoded without decryption, the images we get are called ***cipher-images***.

Compared with the original images (in Fig.1(a)), the pixel value and neighborhood distribution of the cipher-images (in Fig.1(c)) all have been changed. However, for the same original image, different encryption algorithms produce different cipher-images (in Fig. 1(d~f)), which have different changing degrees of pixel value and neighborhood distribution, and then make the ***unrecognizable degree*** much different. The higher the unrecognizable degree of the cipher-images is, the less visual information the attacker will get, which can make the attack more difficult, and the security level of the relevant encryption algorithm should be higher. Therefore, when evaluating and comparing the encryption algorithms of visual media, we need to consider the unrecognizable degree of cipher-image, namely ***visual security***.

Visual security of cipher-images has attracted a lot of attention. However, no researcher put forward systematic research results in visual security assessment means on cipher-images, and no objective assessment algorithm has been proposed. Current researchers of visual media generally give the cipher-images decoded from the cipher-bitstreams at first, and then make a subjective assessment for the unrecognizable degree of the cipher-images. Subjective assessment for the visual security of the cipher-images can be influenced by measurement

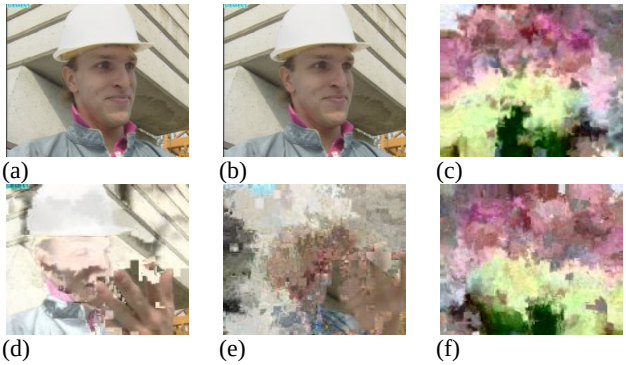


Figure 1: Original images and cipher images. (a) is the original image, (b) is the reconstruction image, (c) is the cipher-image, (d~f) are cipher-images of the image produced by different encryption algorithms.

[1] Supported by the National Basic Research Program of China (Grant No. 2006CB303104) and the National Natural Science Foundation of China (Grant No.40871200/D010702)

environment and subjective sensation. What's more, because of low speed and high cost, it is not very feasible in practical application.

According to the differences of pixel value and neighborhood distribution between cipher-images and original images, we present a visual security assessment algorithm based on neighborhood similarity. This assessment algorithm can assess how much the video information in cipher-videos is distorted, shuffled, and unrecognized, and provide objective assessment compliant with subjective assessment.

2 Background

2.1 Selective encryption algorithms

Visual media have large volume of data with complicated syntax. General data encryption algorithm can not provide direct encryption to protect visual media data. The initial algorithms of visual media encryption protect images by means of shuffling and scramble. As the development of visual media codec technology for images and videos, visual media encryption algorithms have been widely studied since late 1990s, and many achievements have been reported. Among these algorithms, selective encryption algorithms [1,2,3] have attracted more and more attention.

The Selective Encryption Processes with Visual Security Assessment is shown in Fig.2. Selective encryption algorithm encrypts the bits in the compressed bitstreams that are the most critical to image reconstruction. By reducing the amount of data that need to be encrypted, selective video encryption provides a perfect solution to lightweight video encryption. Furthermore, some selective video encryption algorithms generate the encrypted bitstreams that are still compliant with standard syntax

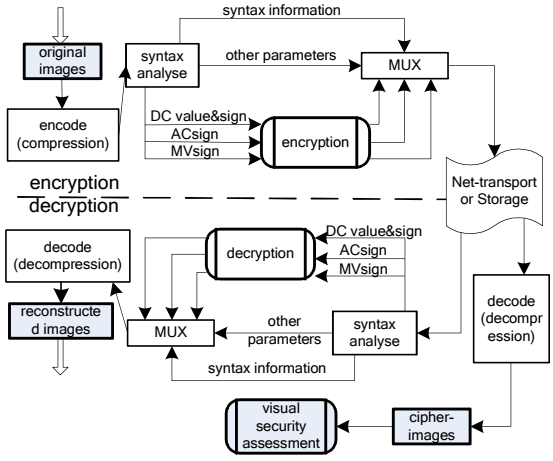


Figure 2: Selective Encryption Processes with Visual Security Assessment.

format. The critical bits of bitstreams do not contain syntax information, such as markers and headers, so the cryptographic bitstreams can keep full bit-level compliance to standard syntax. It is very important to keep bitstreams compliance after encryption for many applications. A standard player will work properly (does not crash) when it decodes these compliant bitstreams of cipher-videos.

Typical selective encryption algorithms [4,5,6,7,8,9,10] include encrypting DCT Coefficients, motion vector, and sign bits of those, and so on. For different kinds of key data encrypted, these cipher-images have different visual security. For example, images without encryption of the motion vectors can be clearly recognized the motion information of people or objects in videos; images without encryption of DC coefficient of low frequency component can be recognized the approximate luminance information; images without encryption of AC coefficient of high frequency component can be recognized the outline information.

2.2 Visual security assessment methods

Visual security assessment is a necessary part of the performance analysis on the image and video encryption algorithms. Security analysis of the encryption algorithm is commonly needed for evaluating and comparing the performance of encryption algorithms. Performance analysis based on cryptanalysis can prove the complexity for the attacker deciphering the encryption algorithm in theory, but can not provide the visual security degree of the cipher-images. In order to develop an objective assessment algorithm on visual security degree of visual media, current security assessments methods of image and video encryption were deep studied and divided into three kinds: assessment based on cryptographic analysis, assessment based on subjective evaluation, and assessment based on video quality assessment.

Assessment based on cryptographic analysis quantitatively analyzes the possibility of deciphering the cipher visual media through the use of cryptanalysis theory. Reference [11] gives the possibilities of ciphertext-only attack, known-plaintext attack and chosen-plaintext attack during the security analysis of encryption algorithms. It also quantitatively gives the compute complexity of ciphertext-only attack through the use of exhaustive method. Reference [12] presents two quantitative cryptanalytic findings on the performance of ciphers against plaintext attacks based on a general model of permutation-only multimedia ciphers. In different perspectives, other references [13,14] also use the cryptanalysis to analyze the possibility and complexity for the attacker to decipher the encryption algorithms successfully. These types of assessments, which process security analysis of encryption algorithms by means of cryptanalysis, are extensively adopted by the most of performance analysis of visual media encryption algorithms.

Assessment based on subjective evaluation process security analysis to the encryption algorithm

subjectively through judging the unrecognizable degree of the cipher-images which decoded from the cipher-bitstreams directly. After introducing the encryption algorithm, reference [15] directly presents six cipher-images of three video test sequences to prove that encryption algorithm can distort the visual information of images, and that cipher-images are unrecognizable to meet the need of visual security. Similarly, reference [16] presents more graphics of cipher-images, and analyzed the unrecognizable degree of the cipher-images subjectively, then compared the visual security of cipher-images getting from different encryption methods. Subjective assessment for the visual security of the cipher-images can be influenced by the measuring environment and subjective sensation. What's more, because of low speed and high cost, it is not very feasible for security analysis only based on the subjective assessment in practical application. Cryptanalysis is needed for security assessment. The subjective assessments are combined with the cryptanalysis, which have already been extensively applied in visual media security evaluation, especially in security evaluation of image encryption algorithm.

Assessment based on video quality assessment theory is a kind of objective security evaluation method, but for which there are few studies or applications so far. There are many video quality assessment methods currently, among which the method based on peak signal noise ratio (PSNR) is widely applied for easy implement and low computational complexity. At present, a few papers of visual media encryption analyze the cipher-images' unrecognizable degree with PSNR value when evaluating the encryption algorithm. Reference [17] presents cipher-images for subjective assessment, and analyze the security degree of the cipher-images according to the cipher-images' PSNR value at the same time. The reference points out that the lower the PSNR value is, the more different between the cipher-images and the original images there will be, and the lower the intelligibility degree of the cipher-image is, so the better the security level is. Reference [18] gives the cipher-images and the change curve of PSNR value, and analyze the recognizable degree of the cipher-images getting from different kinds of encryption algorithms.

Security level evaluation of cipher-images is different from video quality assessment of video codec because they have different research objects and goals. Video quality assessment is a method to measure the distorted degree of loss compression in video codec. It only reflects the accumulation value of error between original image and reconstruction image, which is adopted to assess images that have little difference after compressed and reconstructed. The aim of visual security assessment is to assess how much video information in cipher-videos is distorted, shuffled, and unrecognized. That is to say, visual security assessment has an emphasis on the evaluation of the unidentifiable degree of cipher-images. Cipher-images have many changes not only in pixel value but also in spatial distribution. Therefore, visual security assessment is different from video quality assessment, and the traditional video quality assessment algorithms are not appropriate to evaluate the security level of cipher-videos, and thus it needs to present new objective assessment methods.

3 The proposed assessment scheme

Video image consists of many structured pixels, and there're different levels of brightness value and chromatic value among the pixels. The neighboring pixels present spatial continuous distribution of brightness and chroma. Human visual system could comprehend the continuous distribution of brightness and chroma, and get content information in the images. We consider the continuous characteristic of brightness and chroma in images, and name it as ***neighborhood similarity***.

In Fig.1, the values of pixels in parts of the original image (a) are very close to each other, such as the hat, the wall in background, the face, the clothes, etc. Pixels in these areas have a strong neighborhood similarity. However, when the image is encrypted, the regular spatial distribution of the pixels in these areas is distorted, which results in the decrease of the neighborhood similarity between the neighboring pixels, and the whole image (c) becomes unrecognizable. This paper proposes the definition of neighborhood similarity according to the similarity of the neighborhood distribution of pixels in images. The characteristic that the neighborhood similarity of video images will decrease when the video images are encrypted can provide a way to assess the visual security of cipher-images objectively.

3.1 Definition of neighborhood similarity

***Definition A*:** Let (i, j) and $(i+\alpha, j+\beta)$ denote two pixel in one image with distance as (α, β), and their pixel values are $g_{i,j}$ and $g_{i+\alpha,j+\beta}$. Let the positive constant m denote the difference of the pixel value (Only consider the brightness of pixel value.). Let

$$g(i,j,\alpha,\beta)=\begin{cases}1 & \left|g_{i,j}-g_{i+\alpha,j+\beta}\right|\le m\\ 0 & \left|g_{i,j}-g_{i+\alpha,j+\beta}\right|>m\end{cases} \quad (1)$$

, then the two pixels are ***Similar*** if $g(i,j,\alpha,\beta)=1$; otherwise the two pixel are ***not Similar*.**

***Definition B*:** to calculate whether the points of the $(2d+1)^2$ number on $[-d,+d]$ are similar to the center point (i, j), and to accumulate and normalize the results, then obtain

$$f(i,j) = \sum_{\alpha,\beta\in[-d,+d]} g(i,j,\alpha,\beta) \Big/ (2d+1)^2 \quad (2)$$

. We call $f(i,j)$ the ***Neighborhood Similarity*** of the pixel point (i,j) on the rectangle with radius d.

***Definition C*:** For an image with width M and height N, let the positive constant m denotes the difference of the pixel values, then count the similarity degree of each pixel respectively, and accumulate and normalize the results, get

$$count_m = \sum_{i,j\in[M,N]} f(i,j) \Big/ (M*N) \quad (3)$$

. We call $count_m$ m-level ***Neighborhood Similarity Degree*** of this image.

3.2 Rectangular radius d and pixel value difference m

According to the definition of neighborhood similarity, the value of neighborhood similarity relate to not only spatial distribution states of image's pixel but also the value of d and m. Different value of d and m can lead to different precision.

As to rectangular radius d, big value can get good precision of neighborhood similarity, and result in good description of visual security in theory. However, the computational complexity of neighborhood similarity will increase obviously, when rectangular radius d is numerically larger. For visual security assessment of visual media encryption, we propose that the rectangular radius d have value of 3 or 4, because many video coding algorithms adopt DCT transform with size of 8x8, and such size rectangular radius d represent the pixel boundary of images during video codec. In this paper, rectangular radius d is fixed to 3.

Different images have different spatial distribution of pixels, and different quantity of information. Such difference also exists in different region of the same image. For example, in the original image (a) of Fig.1, the pixel values in the hat region change slightly and in the background wall change more, but in the person face change the most. Therefore, it can assign pixel value difference m to different value to satisfy different precision requirement. Through analysis of some video test sequence, three ranks of pixel value difference m are designed.

1) High precision: High precision value of m represents the similarity of the image area in which the change of pixel value is smooth. For example, in the original image (a) of Fig.1, the hat region pixel value changes slightly, so we can set m a smaller value to get a higher precision. Selecting the pixel point (171, 45), and setting rectangular radius d as 3, can get $(2d+1)^2$ pixel points for which where their pixel values are shown in Fig.3(a).

Analyzing the pixel values, it can be seen that the value of pixel point (171, 45) is 236, and the value of most pixel points around (171, 45) are 235 or 236 with a difference span which does not exceed 2. Due to the discussion above, assigning pixel value difference m to 2 will assure us a higher precision.

2) Medium precision: Medium precision value of m represents the similarity of most image regions. For example, in the original image (a) of Fig.1, the pixel value of the eyes (shown in Fig.3(d)), the nose and the background wall are closely similar to neighbor pixel on a changing trend, so we can set a medium value to get a medium precision. Selecting the pixel point (151,146), and setting rectangular radius d as 3, can get $(2d+1)^2$ pixel points for which their pixel values are shown in Fig.3 (b).

Analyzing the pixel values, we can see that the value of pixel point (151,146) is 64. The values of the pixel points around (151,146) distribute in a wider span. The pixels in left eye of foreman distribute in the range of [61,71], so the pixel value difference m can be set to 5. Due to the discussion above, assigning pixel value difference m to 5 will assure us a medium precision.

236	236	236	236	236	236	235
236	236	236	237	236	236	235
236	235	235	236	235	235	236
236	235	236	236	235	236	237
235	235	236	236	235	237	235
236	236	236	237	236	236	235
235	236	236	234	235	235	235

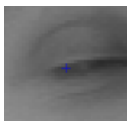
(a) High precision (m=2)

103	105	106	106	101	96	85
99	94	87	77	72	68	66
75	64	63	63	61	63	65
70	68	66	64	63	63	65
77	73	74	71	70	71	74
87	82	79	79	84	84	85
98	96	97	98	98	99	98

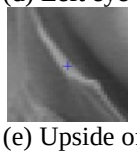
(b) Medium precision (m=5)

159	152	128	89	79	80	76
153	157	146	119	86	78	76
130	158	155	142	116	85	77
108	146	160	151	140	110	87
89	120	156	156	144	136	110
86	96	129	154	154	146	135
83	85	105	134	156	160	155

(c) Low precision (m=10)



(d) Left eye of foreman



(e) Upside of right collapsible

Figure 3: Neighborhood Characteristic of image.

3) Low precision: Low precision value of m represents the similarity of the image area in which the change of pixel value is large. As the areas at the top of the right collapsible shown in Fig. 3(e), the pixel points distribute in a strip. The changes of pixel value within that area are in a wider span, and there are big differences with the pixel points outside these areas. Selecting the pixel point (141,247), and setting rectangular radius d as 3, can get $(2d+1)^2$ pixel points for which their pixel values are shown in Fig.3 (c).

Analyzing the pixel values, it can be seen that the value of pixel point (141,247) is 151. The changes of

pixel value in that area are in a wider span, but the pixel points distribute at the edge of the collar distribute in one strip. The pixel values distribute between [140,160], so the pixel value difference m can be set to 10. Due to the discussion above, assigning pixel value difference m to 10 will assure us a low precision. It is necessary to set pixel value difference m to a bigger value in order to describe neighborhood similarity of complex images.

For different images, we should select pixel value difference m according to different precision demand. High precision pixel value difference m (m=2) is suitable for simple images, which have less quantity of information. Low precision pixel value difference m (m=10) is suitable for complex images, which have more quantity of information. But for many pictures, some regions are smooth and their structures are simple, other regions have much more texture information. We can adopt weighted neighborhood similarity to describe such images, and adjust weighted factor to meet different demand of visual security assessment for different kinds of images.

***Definition D*:** For an image, suppose three level neighborhood similarity are $count_{m=2}$, $count_{m=5}$, $count_{m=10}$ respectively, and the weighted factor is $a+b+c=1$. We define

$$count = a*count_{m=2} + b*count_{m=5} + c*count_{m=10} \quad (4)$$

, then we call $count$ ***Weighted Neighborhood Similarity Degree*** of this image.

3.3 The calculation and comparison of neighborhood similarity degree

For an image or a video frame with width M and height N, its Neighborhood Similarity calculation process is as follows:

1) Choose appropriate rectangular radius d, and pixel value difference m based on 3.2 section's analysis.

2) According to ***Definition B,*** calculate $f(i,j)$ the Neighborhood Similarity of each pixel point (i,j) on the rectangle with radius d by Eq.2.

3) According to ***Definition C,*** accumulate $f(i,j)$ of each pixel (i,j) on the rectangle with radius d, and then get Neighborhood Similarity Degree of the image or the video frame by Eq.3.

For different cipher-images by using different encryption algorithms, we can obtain their objective assessment results on visual security by comparing their Neighborhood Similarity Degree. The larger the Neighborhood Similarity Degree of the image is, the smaller the distorted degree. And the higher the recognizable degree is, the lower the visual security of the corresponding encryption algorithm.

For video sequences, we calculate the Neighborhood Similarity Degree of each frame, and get the curses of Neighborhood Similarity Degree by using different encryption algorithms. Because of different encryption algorithms to be used, different curves of the neighborhood similarity can be obtained, and the smaller the Neighborhood Similarity Degree of the image is, the higher the visual security of the used encryption algorithm. Through observation and comparison of changes in the curves, we can determine the visual security degree of the encryption algorithms.

4 Experiment and results

In this section, through the analysis of different cipher video sequences, three examples are provided for the real applicability of the proposed assessment scheme.

4.1 Visual security assessment for key data encryption

We take the cipher-images of MPEG4 as an example in this paper, and introduce objective assessments of visual security based on neighborhood similarity. The cipher-images of MPEG4 are generated by the selective encryption algorithms which can keep the format compatibility of the bitstreams. There are already many research results [1,3,19,20] on selective encryption algorithm for the MPEG4 compressed bitstreams. Among these results, the proposed key data that can be selectively encrypted includes: DC coefficient sign, DC value, AC coefficient sign, and motor vector sign. Separately encrypting the four types of key data, we can get cipher-images in different recognizable degree. Using the method based on neighborhood similarity we proposed in this paper, we can obtain objective assessment of visual security to these four types of cipher-images.

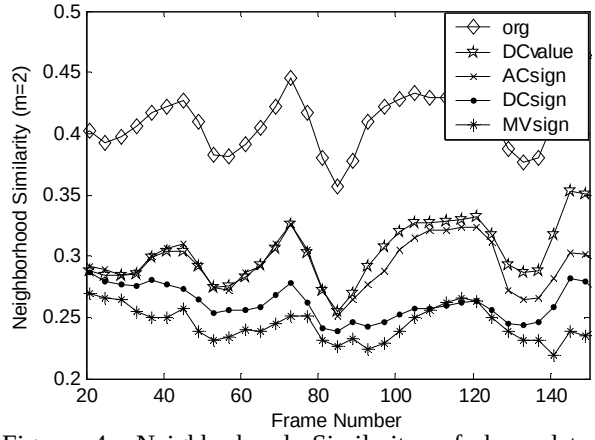


Figure 4: Neighborhood Similarity of key data encryption.

The neighborhood similarity curves of the cipher-images generated by separately encrypting the four types of key data are shown as Fig.4. The topside line (org) of the curve represents the neighborhood similarity of the original image. The lower the value of the neighborhood similarity degree is, the more the image's pixel distribution is distorted, and the better

the visual security level will be. According to the curves of the Fig.4, the cipher-images generated by encrypting MV sign have the best visual security, while the DC value encrypted has the lightest impact. Motion vectors include all the motion information of the video sequence, the motion information after encryption makes the reconstruction of the cipher-images refer to the wrong macro blocks, which has the greatest influence on the recognizable degree of the cipher-images. Based on the above analysis, it can be seen that the cipher-images generated from MV sign encrypted bitstreams have the best visual security. The objective assessments are consistent with the subjective evaluation, and also with the theoretical analysis.

4.2 Visual security assessment for multi-level encryption

By the combine encryption of the four key data, we can realize multi-level encryption to meet the needs of different security and the application of different processing capability. The multi-level encryption algorithm proposed in reference [3], and VEA algorithm, MVEA algorithm and RVEA algorithm proposed in reference [19] all can realize the multi-level encryption. This paper applied the algorithm proposed in reference [19] to encrypt MPEG4 compressed bitstreams. The combination put forward in this reference is as follows: the first level, encrypt AC sign; the second levle, add the encryption of DC value and DC sign; the third level, add MV sign encryption. Different level of encryption leads to different recognizable degree of the cipher-images. By calculating the neighborhood similarity degree, we can get objective assessment of visual security to different security level of cipher-images.

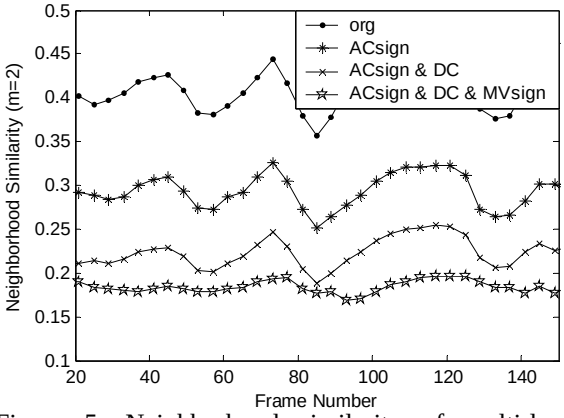


Figure 5: Neighborhood similarity of multi-level encryption.

The neighborhood similarity curves of the cipher-images generated from multi-level encryption of foreman video sequence are shown as Fig.5. We can infer from the curves that with the increase of the key data encrypted, the neighborhood similarity of the cipher-images gradually drops. The cipher-images of the third level encryption that encrypted all the key data have the lowest neighborhood similarity, which reflects the best visual security. However, the original image (org) has no change on the distribution of the pixels, so it has the highest neighborhood similarity. The objective assessments are consistent with the subjective evaluation, and also with the theoretical analysis in theory.

4.3 Visual security assessment for several cipher videos

Encrypting several video test sequences respectively with multi-level selective encryption algorithm, we can get cipher video sequences of different security level. Separately calculating the neighborhood similarity of every frame of the cipher video sequences, and then computing the mean of all the frames' neighborhood similarity, we can get the mean neighborhood similarity of each cipher video sequences. By comparison of the mean neighborhood similarity of the three security level of cipher video sequences, we get the objective visual security assessment result of the multi-level encrypted video sequences.

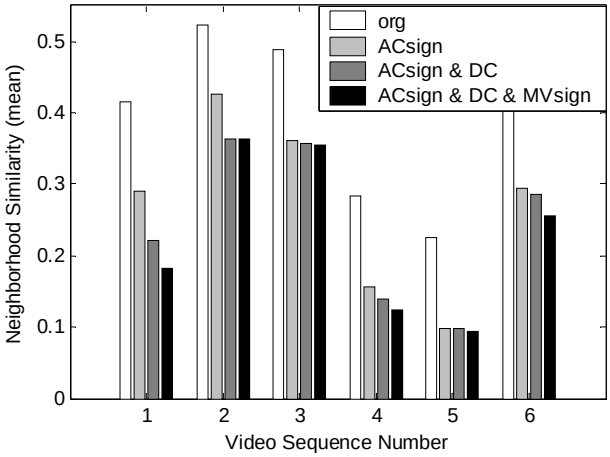


Figure 6: Neighborhood similarity of several cipher videos.

Using bar charts, the mean values of neighborhood similarity for the six multi-level encrypted video sequences are shown as Fig.6. The six video sequences are: foreman, mother&daughter, news, tempete, mobile, hallmoniter. Inferred from the bar charts, with the key data encrypted adding, the neighborhood similarity of each cipher video sequence gradually decreased. The height of the bars basically reflects relative value of neighborhood similarity. Except for sequence 2, 3 and 5, the distinction degrees of the visual security assessment is better for the other video sequences. There is no significant difference between the neighborhood similarity of the third level and that of the second level of video sequence 2 (mother&daughter), that's because the motion information of this video sequence is comparatively less, and the third level cipher-bitstreams have no significant difference from the second level after adding the MV sign encryption. Distinction degrees of

video sequence 3 and 5 are not ideal, which shows the objective algorithm proposed in this paper needs to improve.

5 Performance analysis

In this section, we give performance analysis of the proposed assessment scheme, such as computational complexity, time complexity and applicability, and so on. On applicability analysis, different evaluation results were analyzed in detail in Section 4. As can be seen from the analysis, the changes of neighborhood similarity of cipher video sequences in the curve can be a very good reflection of the changes of visual security.

On the computational complexity and time complexity, we also did experiment and analysis. For the CIF (352 * 288) size of the video sequences, the calculation time of the neighborhood similarity for each frame is not more than one second. Furthermore, from Fig. 2, we can see that the proposed assessment of visual security is independent of the process of video codec, and it also independent of encryption and decryption. Visual security assessment scheme can be made an independent module. For cipher video sequences by using different encryption algorithms, we can obtain objective evaluation results by off-line analysis. As a result, computational complexity and time complexity will not affect the application of the proposed objective visual security assessment scheme. And the research on visual security should be focused on the applicability of assessment algorithms.

6 Conclusions and future work

Visual security is a very important target of security assessment in the field of video encryption, which has direct relation to the attacker's comprehension degree of the cipher-images. The more information the attacker gets from the cipher-images, the faster the unauthorized decryption will be. Security analysis of video encryption now mostly put emphasis on the security analysis of the encryption algorithm, but not on the visual security assessment method. Till now, an applicable objective security assessment algorithm or calculation index has not been proposed. Therefore, it need to present new objective security assessment method. On one hand, based on the analysis of encrypting different key data, it can guide us to design good combinations of key data encryption, and then design selective encryption algorithms of high visual security. On the other hand, it can be applied to evaluation the visual security of the encryption algorithms and provide some references on performance analysis. It can be believed that the study will be very significant to further research on video encryption.

We present an objective method based on neighborhood similarity to carry out visual security assessment in this paper. It takes cipher-images as the research object, which is independent of video encryption algorithms. Therefore, it can be made an independent module based on off-line analysis, and its computational complexity and time complexity will not affect its application. The detail analysis in Section 4 verifies that our proposed assessment method is efficient to evaluation the visual security of the encryption algorithms. Our next step is focused on the research of the availability of evaluation, especially on the extension of its applicability.

Visual security assessment is a brand-new research topic. After we present the objective assessment method based on neighborhood similarity, all the relevant research will begin soon. The method proposed in this paper will probably be further optimized and improved, and new objective assessment algorithms would be developed based on the characteristic of cipher-images.

References

[1] Wen Jiangtao, Severa Michael, Zeng Wenjun, Luttrell Maximilian, etc. A format compliant configurable encryption framework for access control of Video. IEEE Tran. Circuits & Systems for Video Technology, 2002, Vol. 12, 545-557.

[2] Howard Cheng and Xiaobo Li. Partial Encryption of Compressed Images and Videos. IEEE Transactions on Signal Processing, v 48, n 8, Aug, 2000, 2439-2451.

[3] Yuan chun, Zhong yuzhuo, Yang Shiqiang. Composite Chaotic Pseudo-Random Sequence Encryption Algorithm for Compressed Video. Tsinghua Science and Technology. 2004, Vol.9, No.2, 234-241.

[4] Iskender Agi and Li Gong. An empirical study of secure MPEG video transmission. Proceedings of the Internet Society Symposium on Network and Distributed Systems Security. San Diego, CA, 1996. 137-144.

[5] Tang Lei. Methods for encrypting and decrypting MPEG video data efficiently. Proceedings of the Fourth ACM International Multimedia Conference (ACM Multimedia 96'). Boston, MA, 1996. 219-230.

[6] Ali Saman Tosum and Wuchi Feng. Efficient multilayer coding and encryption of MPEG video streams. IEEE International Conference on Multimedia and Expo. New York, 2000. 119-122.

[7] Lintian Qiao and Klara Nahrstedt. Is MPEG encryption by using random list instead of zigzag order secure. IEEE International Symposium on Consumer Electronics. Singapore, 1997, 226-229.

[8] Shi C G, Bhargava B. A fast MPEG video encryption algorithm. Proceedings of the 6th ACM International Multimedia Conference. Bristol, 1998, 81-88.

[9] Changgui Shi, Shengyih Wang, Bharat Bhargava. MPEG video encryption in realtime using secret key cryptography. Proceedings of

the International Conference of Parallel and Distributed Processing Techniques and Applications (PDPTA99'). Las Vegas, Nevada, 1999, 2822-2828.

[10] JuiCheng Yen, Juning Guo. A new MPEG encryption system and its VLSI architecture. IEEE Workshop on Signal Processing Systems. Taipei, 1999. 430-437.

[11] Shiguo Lian, Zhongxuan Liu, Zhen Ren, Zhiquan Wang. Selective Video Encryption Based on Advanced Video Coding. PCM 2005, Part II, LNCS 3768, 281-290.

[12] Shujun Li, Chengqing Li, Guanrong Chen, Nikolaos G. Bourbakis, Kwok-Tung Lo. A general quantitative cryptanalysis of permutation-only multimedia ciphers against plaintext attacks. Signal Processing: Image Communication, 2008, vol. 23, no. 3, 212-223.

[13] Yuan Li, Liwei Liang, Zhaopin SU, Jianguo Jiang. A New Video Encryption Algorithm for H.264. ICICS 2005, 1121-1124.

[14] Yuanzhi Zou, Tiejun Huang, Wen Gao, Longshe Huo. H.264 Video Encryption Scheme Adaptive to DRM. IEEE Transactions on Consumer Electronics, Vol.52, No.4, NOVEMBER 2006, 1289-1297.

[15] Jinhaeng Ahn, Hiuk Jae Shim, Byeungwoo Jeon, Inchoon Choi. Digital Video Scrambling Method Using Intra Prediction Mode. PCM 2004, LNCS 3333, 386-393.

[16] Sang Gu Kwon, Woong Il Choi, Byeungwoo Jeon. Digital Video Scrambling Using Motion Vector and Slice Relocation. ICIAR 2005, LNCS 3656, 207-214.

[17] Shiguo Lian, Zhongxuan Liu, Zhen Ren, Haila Wang. Secure Advanced Video Coding Based on Selective Encryption Algorithms. IEEE Transactions on Consumer Electronics, Vol.52, No.2, MAY 2006, 621-629.

[18] Thomas Stutz, Andreas Uhl. On Efficient Transparent JPEG2000 Encryption. MM&Sec'07, September 20-21, 2007, Dallas, Texas, USA.

[19] Bharat Bhargava, Changgui Shi, Sheng-Yih Wang. MPEG Video Encryption Algorithms. Multimedia Tools and Applications, 2004, Vol.24, No.1, 57-79.

[20] Wenjun Zeng, Shawmin Lei. Efficient Frequency Domain Selective Scrambling of Digital Video. IEEE Transactions on Multimedia, 2003, Vol.5, No.1, 118-129.

Secure, Portable, and Customizable Video Lectures for E-learning on the Move

Marco Furini
Department of Social, Cognitive and Quantitative Sciences
University of Modena and Reggio Emilia
Via Allegri 9 - 42100 Reggio Emilia, Italy
E-mail: marco.furini@unimore.it

Keywords: e-learning, secure video lectures, media in education

Received: September 27, 2008

The production of video lectures for the mobile scenario is becoming popular, as the pervasiveness of mobile technologies is making learning independent of time and space. In such a scenario several challenges need to be addressed, and the contribution of this paper is MOLE, an architecture that produces secure, portable, and customizable video lectures for generic mobile devices. Video lectures are produced with a format that ensures play out compatibility in most mobile devices, and with a security mechanism that protects contents from un-authorized usage. Furthermore, a video lecture is organized in such a way that a learner can adapt the lesson development to his/her learning needs by locally interacting with the system, which means that people with different needs may use the same video lecture file. A prototype implementation of MOLE shows its feasibility. The production of secure, portable, and customizable video lectures may help expanding mobile learning as generic mobile devices may be used as learning tools.

Povzetek: Video predavanja za e-učenje.

1 Introduction

Mobile learning, the combination of mobile computing and e-learning, is expected to expand and to evolve dramatically over the next few years. Exploiting the pervasiveness of mobile technologies it is possible to make learning independent of time and space so as to make mobile learning a real opportunity. As a result, the production of video lectures is becoming more and more important, as video is one of the most powerful media to present information and students find video materials very compelling [1].

To make mobile learning a success, several challenges are still open and need to be addressed: mobile learning may favor technologically advanced students; the large variety of learning devices my cause lectures to be encoded in several formats; being a digital media, video lectures might be subject of intellectual properties and copyright issues [2,3].

In this paper we propose MOLE (MObile LEarning), an architecture to produce *secure, portable*, and *customizable* video lectures for the mobile environment. MOLE aims at producing video lectures for generic mobile devices, i.e., devices with different computational and storage characteristics, from simple video players to PDAs, from iPods to smart phones. Security is a key feature in nowadays mobile scenario, as video lectures may contain copyrighted materials. Therefore it is necessary to protect such material from un-authorized usage. Since the mobile scenario is filled with devices that have limited computational resources (e.g., cellphones), the usage of complex security mechanisms may be a burden for most devices. For this reason, MOLE is designed with a security mechanism that is light enough to be used over generic mobile devices. Portability is another important feature, as people are equipped with a large variety of devices. In such a scenario it is not reasonable to produce several versions of the same video lecture so as to meet the characteristics of students' devices. MOLE aims at producing video lectures with a format that can be played over the majority of mobile devices. Similarly, in addition to the heterogeneousness of mobile devices, the mobile scenario is filled with students who have different learning needs. Also in this case, it is not reasonable to produce, for the same subject, several video lectures with different learning levels. However, content adaptation is very important in learning as it allows a better supporting of learners with different skills and motivations [4,5], and, if not provided, remote students would feel frustrated and would tend to drop courses [6,7]. For this reason, MOLE organizes the contents of a video lecture in such a way that it contains several learning levels, giving the student the opportunity to tailor the lesson to his/her learning needs.

The feasibility of MOLE is tested through a developed video lecture player. Results show that MOLE produces video lectures for the mobile scenario with

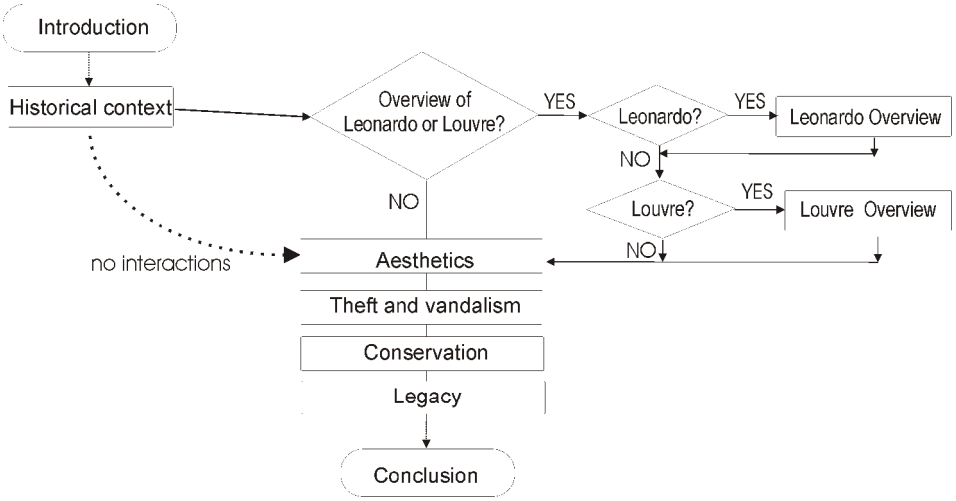


Figure 1: Classroom interaction during a lesson related to Leonardo's artwork The Mona Lisa. The lesson can take different directions, according to students' request. If interactions are not provided, the lesson has only one direction.

characteristics that potentially transform any mobile device into a learning tool.

The remainder of this paper is organized as follows. In Section 2 we briefly overview proposals in the field of mobile learning; Section 3 presents details of the MOLE architecture, whereas its feasibility investigation is discussed in Section 4. Conclusions are drawn in Section 5.

2 Related work

The effects of mobile technologies in learning have been investigated under different prospectives [6]: new models for teaching and learning (e.g., [8,9]); effects on the design process and on student experience (e.g., [10,11]); effectiveness and costs of using mobile devices in education (e.g., [12]); new tools for distance learning (e.g., [13,14,15,16]); adaptation of learning contents to mobile devices (e.g., [17,18,19]). Since our approach deals with portability, security, and content adaptation, in the following we focus on approaches that investigate the same issues.

Marinelli and *Stevens* [20] focus on customization; they propose segmenting a video lesson into small chunks, which are stored on a video server. When a student access to the video lesson, depending on his/her choices, the most appropriate video chunks are streamed. This approach requires a data transfer rate able to stream the video content. Unfortunately, while effective for wired environments, the streaming of a video lecture may be problematic in the mobile scenario: data transfer rate may be insufficient (e.g., in rural area, where high speed wireless networks are not still available), not to mention that data traffic is usually metered and paid by megabytes. Furthermore, this approach requires a play out device with communication facilities, and thus it cuts out a considerable number of mobile devices (e.g., Ipods). *Liu* and *Choudary* [21] focus on the data transfer problem and propose propose encoding video lectures with different quality levels and with different compression ratios, so as to meet different transfer data rate and different mobile devices. Although ameliorating the problem of streaming video lectures in a mobile scenario, this approach requires a device with communication facilities. *Kung* and *Wu* [22] focus on customization too, and they propose integrating synchronous and asynchronous learning systems to provide content adaptation to student's needs. *Weippl* [23] focuses on security and analyzes the weaknesses inherent to mobile devices.

The novelty of MOLE is that streaming is not used, mobile devices do not need communication facilities, and security is guaranteed with a light security mechanism.

3 The MOLE architecture

In this section we present details of MOLE (MObile LEarning), the architecture designed to produce *secure, portable*, and *customizable* video lectures for the mobile scenario.

Before presenting details of the architecture, let us consider a simple lesson which we'll refer to in the remainder of this paper. The lesson is about *Leonardo's artwork The Mona Lisa*. Briefly, it involves different topics: historical context, aesthetics, theft and vandalism, conservation, and legacy. Since the part related to the historical context contains topics explained to students (e.g., Leonardo history, Louvre museum) in previous lessons, a classroom lesson usually develops according to whether the students have clear these topics or not. If not, the teacher usually spends some minutes in reviewing them, without entering into the (already explained) details. Hence, the classroom lesson may develop differently, as depicted in Figure 1: a pre-defined lesson path goes from *historical context* to *aesthetics*, but due to students/teacher interactions, other paths are possible. These additional paths represents different learning levels that may be used by students to adapt the video lecture to their learning needs.

In most of distance learning systems, if a student has not clear a subject explained in previous lessons, he/she has to stop playing out the current video lecture, has to

browse the library to find out the right video lecture, has to browse the video to find the points of interest and, finally, has to re-watch it. Needless to say, it is likely that most students will continue watching the video lecture even though they don't clearly have the picture of a particular subject. Due to the importance of content adaptation MOLE allows simulating the classroom scenario also when students and teacher are distributed in time and space.

The idea is to store several possible lesson paths into a single video lecture file and on defining points where a student can modify the lesson development by interacting with the system. This means that all the different lesson paths are stored in a single file.

Figure 2 shows the architecture of MOLE: a video lecture is produced by first recording/encoding the educational material; then the material is organized through a text-based script; finally, the material is stored in a multimedia container and protected against un-authorized usage. At the student-side, once the video lecture is downloaded, the first step is the removal of the protection, so as to allow the retrieval of all the contents (educational material and content organization). Then, the video lecture is played out, and according to the student's interactions, the content of the video lecture is adapted to the student's learning needs.

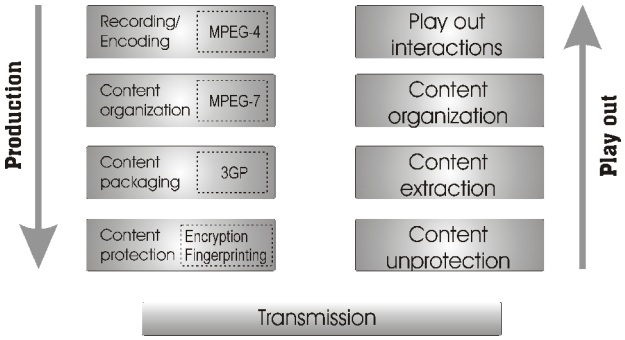


Figure 2: The MOLE architecture to produce and protect video lectures. To guarantee large compatibility in most mobile devices, video lectures are produced with well-know standards mechanisms (dotted box).

3.1 Production of the video lecture

3.1.1 Encoding of the audio/video material

To reduce both the download time and the storage space, a video lecture should have low bit rate, while providing a good video quality. To this aim, MOLE uses MPEG-4, a standard designed to encode audio/video contents with high quality at low bitrates [24]. Since a detailed description of this standard goes beyond the scope of this paper, here, we simply highlight that MOLE encodes video with MPEG-4 Part 2 specifications, and audio with MPEG-4 Part 3 specifications. Thanks to the exceptional performance and quality of Part 2 and Part 3, and to the large presence of MPEG-4 players in mobile devices, MOLE produces video lectures that are limited in size and playable over most modern mobile devices.

3.1.2 Organization of educational material

Since there is no way to know in advance the different learning needs of the people who will watch the video lecture, it is necessary to include, inside the video lecture, points where students can virtually interact with the teacher (and actually interact with the system) so as to provide different learning paths. The organization of the educational material is done through a text-based description which is called *lesson script* and is produced according to the teacher's experience.

The description virtually divides a lesson into several video chapters. Similarly to the IEEE's learning object model, a video chapter contains a portion of the lesson that can be used once or several times during the lesson play out, depending on the defined script and on the student requests. However, it is worth noting that all the video chapters are part of a single video file and hence there is no need for the student to get multiple files or to be on-line to receive the most appropriate video stream. In fact, it is the player that uses the lesson script to jump from one video chapter to the other depending on the student's interaction.

MOLE defines four different types of video chapter (Figure 3), with the following meaning.

- **Initial**. It's the chapter that begins the lesson; it has only one possible outgoing lesson direction and only one initial chapter per lesson is allowed;
- **Interactive**. It's the chapter that allows a student to virtually interact with the teacher in order to modify the lesson development. It has multiple possible incoming and outgoing lesson directions. Several interactive chapters per lesson are allowed;
- **Sequential**. It's the chapter that simply plays a portion of the lesson. It has multiple possible incoming lesson directions, but only one

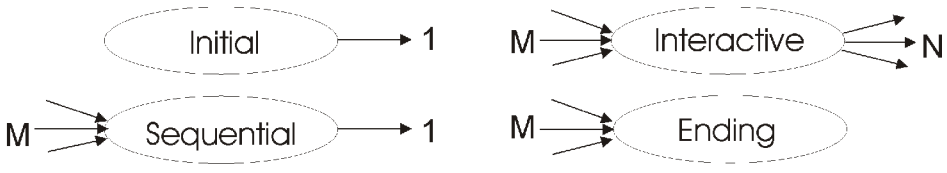


Figure 3: Types of video chapter that may compose a video lecture. The number of possible incoming and outgoing lesson directions characterize the typology.

possible outgoing direction. Several sequential chapters per lesson are allowed;

- **Ending**. It's the chapter that ends the lesson. It has multiple possible incoming lesson directions. Only one ending chapter per lesson is allowed.

Thanks to interactive video chapters, a student can modify the lesson development according to his/her learning needs. For instance, Figure 4 presents three possible lesson developments: student \#1 only requires an overview of *Leonardo*, student \#2 does not require any material overview, whereas student \#3 requires *Louvre* overview. This simple example shows that the organization of the contents allows having different lesson path inside the same video lecture.

Once the material has been divided into several chapters, the actual lesson description takes place and specifies: i) all the information that describe a video chapter (e.g., beginning and duration time, possible textual information associated, etc.), and ii) the points where a student can modify the lesson development.

Different description languages may be used to produce the actual lesson description (e.g., XML, SCORM, MPEG-7), all of them with pros and cons. For the sake of clarity, in the following we present examples described based on the MPEG7-MDS standard [25]. This standard is composed of metadata structures and is used to produce a description of the spatial layout of different media objects (e.g., audio, video) and of the temporal order in which these objects will be played out.

```
<VideoSegment>
  <label>"Z"</label>
  <RelatedMaterial>
    <MediaLocator><MediaUri>
      http://.../mm/monalisa.pdf
     </MediaUri></MediaLocator>
   </RelatedMaterial>

   <MediaTime>
      <MediaTimePoint>
        00:22:30
      </MediaTimePoint>
      <MediaDuration>
        00:03:30
      </MediaDuration>
   </MediaTime>
   <TextAnnotation>
      <FreeTextAnnotation>
         The Mona Lisa painting
         can be found at page 356.
      </FreeTextAnnotation>
   </TextAnnotation>
</VideoSegment>
```

Table 1: The usage of MPEG7-MDS to describe a portion of a video lesson.

The description is text-based and uses tags (in the form of `<tag [attribute=value]>`) to define properties of a media object (or of a part of it). For instance, Table 1 shows a video chapter description related to *Mona Lisa*: it begins after 22 minutes and 30 seconds and last 3 minutes and 30 seconds. Additional information, like a text description `textannotation` tags) or related material (`RelatedMaterial` tag), can be attached to any chapter.

To define the set of possible points where a student can modify the lesson development, MOLE uses a table called *scene transition table*. Each entry of this table is uniquely identified with a Video Chapter Identifier (VCI) number and includes the possible question asked by the teacher to begin the interaction with students and a set of possible destinations the lesson can take based on the student's answer. For instance, the entry related to video chapter *Y* (Fig. 4) has two possible destinations (chapters *Y.1* or *Z*), whereas the video chapter *X* only has a single chapter destination (chapter *Y*). In this way, a student can get get a lesson tailored to his/her needs, by simply selecting the most appropriate video chapter for his/her learning process. Note that also the scene transition table is described with textual information.

3.1.3 Protection of the video lecture file

To protect contents from illegal usage or from un-authorized modifications, MOLE is equipped with a license-based security mechanism (a.k.a. Digital Right Management, DRM for short) that discloses interactive materials only to authorized students.

The basic idea of the security mechanism is to wrap a media file with encryption, code authentication and

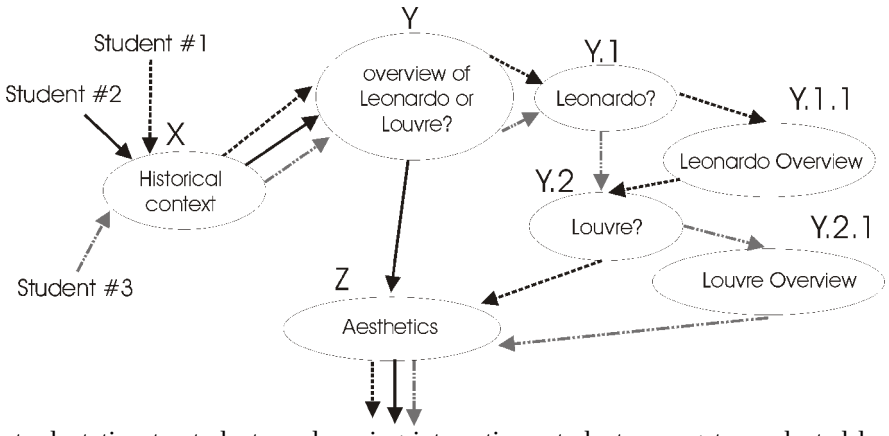


Figure 4: Content adaptation to student needs: using interactions students can get an adapted lesson. Here, three different lesson developments are shown.

information hiding, as depicted in Figure 5. MOLE gives the choice to provide the single pre-defined lesson path in clear so that every ordinary player can access to and can play out the single pre-defined lesson path. This feature may be useful to those organizations that want to give students the opportunity to appreciate the educational material so as to tempt them to buy the right to play out all the lesson paths. However, it is worth noting that the single pre-defined lesson path in clear is an option and is not mandatory (the single pre-defined lesson path may be encrypted as well as all the other lesson paths). In particular, MOLE organizes the video material as follows:

- All the video chapters that compose the single pre-defined lesson path are stored in the first part of the video file. These chapters might be in clear (so that every ordinary player can play them out) or might be encrypted with a symmetric technique;
- All the other chapters are encrypted with a symmetric technique and are stored in the second part of the file;
- The first and the second part of the video file are separated by 60 seconds of blank video, so that if the pre-defined lesson is in clear, an ordinary player will not produce an immediate play out error when trying to play out the encrypted second part of the video file.

The encryption/decryption key is hidden inside the first part of the video, so that only players able to retrieve the key can play out the video lecture;

Before presenting details of how the second part of the video file is encrypted, for the sake of clarity, let us review how a license-based mechanism usually works: first, the content provider generates a symmetric key and then encrypts the media content (the second part of the audio stream and the the lesson script in our case). After the encryption, the symmetric key is hidden inside the media file and a license file is generated with all the information needed by the decoder to play out the media file (users rights, positions of the media file where to find the hidden key, etc.). Finally, the license file is encrypted with an asymmetric key technique so as to bind the license file to the owner of the private key. License acquisition can be done in a transparent way, as it happens today with several DRMs (e.g., Microsoft DRM 10).

Figure 5 shows details of the MOLE security mechanism. The key used to encrypt the second part of the lesson file and the lesson description (α in our case) is hidden into the first part of the media file using a watermarking technique that spreads the hidden key without compromising the media content (e.g., [26]). To spread the information, this technique generates the so-called *watermarking key* (WKey) and uses it to hide the information. This key is stored inside the license file, along with information used to check data integrity. To guarantee data integrity the content provider uses a hash function H to compute the hash value of the encrypted video stream (i.e., $HV=H(E_\alpha(video\_stream))$) and of the lesson script (i.e., $HD=H(E_\alpha(lesson\_script))$) and stores them inside the license file. These values are also watermarked inside the media file and inside the lesson description, as they will be used by the video lecture player to check the integrity of the stream and of the lesson script.

Finally, the license file is encrypted with a public/private key, so that only who owns the private key can decrypt it. We recall here that the private key is usually given during the sign-up procedure and is stored in the device repository key, which is hidden with suitable software engineering techniques.

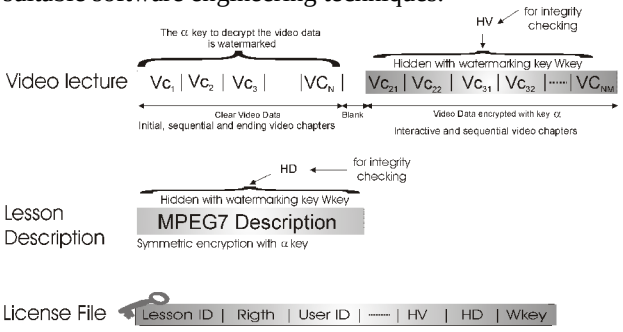


Figure 5: Security mechanism: the first part of the lesson file might be in clear or encrypted, whereas the second part of the lesson file is encrypted. The decryption key is hidden in the first part of the file. The lesson description is encrypted with the same key used to encrypt the second part of the lesson file. The license file is encrypted with private/public key and contains information to decrypt the lesson file.

3.1.4 Packaging of educational contents

Video material and lesson description need to be stored in a single multimedia container. MOLE uses 3GP, a multimedia container designed to handle multimedia contents in a mobile environment [27]. It has been defined by the Third Generation Partnership Project (3GPP), which provides worldwide standard specifications for multimedia contents over 3rd generation cellular networks. A 3GP file can contain material encoded with different schemes like H.263, MPEG4, MP3, and is supported by the majority of multimedia mobile devices. The file structure is composed of data structures called boxes, which are hierarchically organized. Each box is identified with a tag and contains a media object (e.g., audio, video), which can be actual media data or simple metadata (information to describe the media properties). MOLE stores the video material in a `trak` box and the lesson description in a `udta` box. For details about the 3GP file structure we refer the readers to [27].

3.2 Play out of the lesson content

To enjoy the full features of the lesson file, an enhanced player is necessary. As shown in Figure 2, the player is in charge of removing the protection so as to access to

educational material. In particular, the player is in charge of: i) checking the integrity of the video file, ii) decrypting and playing out the lesson file, iii) interacting with the student and jumping from one chapter to another depending on the student's choices.

To check the integrity of the video file as well as of the lesson script, the player first uses the student's private key (note that private keys are usually stored into the device repository) to decrypt the license file. Once decrypted, the license file provides, among other information, the watermarking key and the values to check the integrity. At this point, the player retrieves *HV* and *HD* from the lesson file and the lesson description and compares them against the values retrieved from the license file. If the integrity check fails, the play out is interrupted, otherwise the player retrieves the hidden α key and begins the lesson play out.

While playing out the video lecture, the player retrieves the video chapter information from the MPEG-7 description (i.e., chapter label, media time, etc.) and from the scene transition table (i.e., question and destinations, if any). If the chapter is interactive, depending on the option selected by the student, the player jumps to the video chapter in order to continue with the lesson.

4 MOLE prototype implementation

In this section we present a prototype implementation of MOLE and a security evaluation.

4.1 Video lecture production

To evaluate MOLE we produce a video lecture encoded with MPEG-4 at 15 frames per second, with a resolution of 176x144 pixels (common display resolution in most mobile devices) and a bitrate of 64 kbps; audio is MPEG-4 encoded (Part 3, commonly known as AAC-LC), two channels at 128 kbps. The overall bitrate of the video lecture is of 192 kbps. Content organization is specified with MPEG7-MDS, and content protection is achieved with the MOLE security mechanism. All the contents are then stored on a 3GP multimedia container.

4.2 Player implementation

Using the Nokia Prototype SDK 4.0 for Java ME, we develop a player able to play out video lectures produced with MOLE. Figure 6(a) shows the play out of a non interactive chapter, whereas Figure 6(b) displays that the play out of an interactive chapter changes the menu option. When the *ask* button is pressed, the player presents a multiple choices menu (Figure 6(c)), where a student can select his/her preferred subject. Once selected, the player jumps to the associated video chapter.

(a) (b) (c)

Figure 6: Lesson play out: (a) sequential chapter; (b) interactivity activates the *ask* button; (c) selection of a particular topic.

4.3 Feasibility Analysis

A first investigation analyzes whether the security mechanism can be sustained by current mobile devices. In fact, security operations like decryption use complex mathematical operations and can be problematic to mobile devices with limited processing power. For this reason, we analyze the decryption processing cost in relation to recent released PDAs (processors speed that ranges between 126 and 624 MHz) and cellphones (processors speed around 200 MHz)[1], in order to investigate the feasibility of our approach.

Video stream and lesson description are encrypted with symmetric technique. *Ravi* et *al*. [28] showed that a 206MHz SA-1110 processor can sustain a decryption rate of 1.8 Mbps per second, when fully dedicated to the task, and a decryption rate of 180 kbps when only 10% of the computational resource is dedicated. Hence, the decryption workload of the video stream (192 kbps) can be sustained by new generation PDAs and the cellular phones.

In order to handle interactions, the lesson script has to be decrypted during the play out of the first two chapters (in fact, since the first chapter is not interactive, the play out of the first two chapters is guaranteed). Since the description of a single video chapter is very short, it is very likely that the decryption workload of the lesson script can be sustained. For instance, 100 video chapter descriptions (as the one reported in Table 1) can be decrypted in just 2 seconds (considering a decryption rate of 180kbps). Since it is reasonable to assume that the first two video chapters last much more that two seconds, and that a lesson does not have thousands of video chapters, it is safe to assume that the decryption workload of the lesson script can be sustained by recent PDA and cellphones.

The license file is encrypted with asymmetric technique, which uses more intensive mathematical operations than a symmetric technique. A study performed in [29] showed that an asymmetric decryption takes 2.63 ms for 1 KB of data, using a 100 MHz processor. Again, since this file is usually very small (around 2-4 KB), portable devices can sustain the decrypting of the license file without causing an excessive delay to the content material play out.

A second investigation analyzes the storage space required by the proposed approach. In fact, before beginning the play out, the video lecture has to be

[1] For instance, the Nokia N90 serie is equipped with a 220 MHz processor, whereas the iPhone has a 620 MHz processor as described at: http://www.engadget.com/2007/07/01/iphone-processor-found-620mhz-arm/

entirely downloaded, and the storage space required might be problematic to mobile devices. Since the overall bitrate of the video lecture is of 192 kbps, a one-hour video lecture requires around 86 MBytes. Even though a video lecture file may contain several lesson paths, it is reasonable to assume that a video lecture file requires space on the order of hundreds of MBytes. Recent released PDAs and cellphones are equipped with an internal memory storage that is on the order of GBytes[2], or are equipped with a slot where users can insert SD memory card (which are becoming very popular for their limited cost). Therefore, it is safe to assume that the space storage is not a burden for the proposed approach.

A final investigation analyzes whether the security goal is achieved and under which conditions. By assuming the existence of cryptographically secure hash functions and of a secure symmetric/asymmetric key encryption scheme, sharing of a video lecture and/or a license file is useless. In fact, one can successfully share it if he/she can capture both the video and the script description, but to break the protection, the adversary must learn the watermarking key of the destination player. As previously mentioned, this is hard with no knowledge of the the watermarking key. Furthermore, also alteration of the content rests on the security offered by the watermarking scheme. In fact, alterations are possible only if new integrity parameters (i.e., the hash value of the video lecture and of the lesson script) can be stored.

It is worth mentioning that security of digital material is a problem that admits no final and provably strong solution. Briefly stated, the reason is that the security mechanism works in an untrusted environment (the playout device is under the user's control). Readers can refer to [30] for a broad and interesting survey paper on effectiveness of security mechanisms. It also worth mentioning that watermarking techniques are also subjects of a never ending debate: The biggest threat faces the removal or alteration of the marks. This is typically obtained using multiple copies of the same file, containing different fingerprints. According to [31], by averaging these copies the fingerprint is altered. In general, *Schonberg* and *Kirovski* [32] claim that watermarking is not secure with current technology. However, *Shan He* and *Min Wu* [33] are much more optimistic about the security provided by watermarking techniques. Far from trying to settle the dispute, we simply note that all the current security mechanisms are based on fingerprinting schemes.

Note that to increase the security of the proposed approach, one could use stronger cryptographic tools (e.g, public-key certificates, Internet-based procedure), but since our mechanism is designed for devices with limited computational resources, the usage of such tools is avoided for performance reasons.

[2] For instance, Nokia N16, as well as the iPhone, has 16GB of internal memory.

5 Conclusions

In this paper we presented MOLE, an architecture that produces secure, portable, and customizable video lectures for the mobile scenario. The novelties introduced by MOLE are: i) video lectures are produced in a format that is widely accepted by modern mobile devices; ii) a video lecture contains several lesson paths so as to meet different learning levels requirements; iii) contents are protected against un-authorized usage by using a security mechanism whose computational lightness ensures an easy usage over current mobile devices.

These characteristics allow MOLE to produce video lectures for most mobile devices causing these to be considered as learning tools. As a result, we think MOLE is a candidate approach to ease the process of learning across time and space.

References

[1] K. D. Kelsey (2000), Impact of communication apprehension and communication skills training on interaction in a distance education course, *Journal of Applied Communications*, Vol. 84, No. 4, pp. 7–92.

[2] R. Benlamri, J. Berri, Y. Atif (2006), A framework for ontology-aware instructional design and planning, *Journal of E-Learning and Knowledge Society,* Vol. 2, No. 1, pp. 83–96.

[3] J.R. Corbeil, M.E. Valdes-Corbeil (2007), Are You Ready for Mobile Learing?, *Educase Quarterly*, November 2007, pp. 51–58.

[4] H. S. Keng Siau, F. F.-H. Nah (2006), Use of a classroom response system to enhance classroom interactivity, *IEEE Transaction on Education,* Vol. 49, No. 3, pp. 398–403.

[5] M. C. Wang, G. D. Haertel, H. J. Walberg (1992), What influences learning? A content analysis of review literature, *Journal on Educational Resources,* Vol. 84, No. 1, pp. 30–43.

[6] R. Y.-L. Ting (2005), Mobile learning: Current trend and future challenges, in: *Proceedings of the IEEE International Conference on Advanced Learning Technologies* 2005, IEEE Computer Society.

[7] P. F. Whelan (1997), Remote access to continuing engineering education (RACeE), *IEE Engineering Science And Education Journal,* pp. 205–211.

[8] A. P. Massey, V. Ramesh, V. Khatri (2006), Design, development, and assessment of mobile applications: the case for problem-based learning, *IEEE Transactions on Education,* Vol. 49, No. 2, pp. 183–192.

[9] M. Sharples (2000), The design of personal mobile technologies for lifelong learing, *Computers & Education* Vol. 34, pp. 177–193.

[10] M. Berry, M. Hamilton (2006), Mobile computing, visual diaries, learning and communication: changes to the communicative ecology of design students through mobile computing, in: ACE ’06:

Proceedings of the 8th Austalian conference on Computing education, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 2006, pp. 35–44.

[11] R. Nachmian (2002), A research framework for the study of a campus-wide web-based academic instruction project, *Internet and Higher Education* Vol. 5, No. 3, pp. 213–229.

[12] J. Traxler (2003), m-learning: Evaluating the effectiveness and cost, in: *Proceedings of the 2nd Annual MLEARN Conference*, 2003, pp. 70–71.

[13] J. T. Black, L.W. Hawkes (2006), A prototype interface for collaborative mobile learning, in: IWCMC '06: *Proceeding of the 2006 international conference on Communications and mobile computing*, ACM Press, New York, NY, USA, 2006, pp. 1277–1282.

[14] Y. Zhang, S. Zhang, S. Vuong, K. Malik (2006), Mobile learning with bluetooth-based e-learning system, in: IWCMC '06: *Proceeding of the 2006 international conference on Communications and mobile computing*, ACM Press, New York, NY, USA, 2006, pp. 951–956.

[15] M. Virvou, E. Alepis (2005), Mobile educational features in authoring tools for personalized tutoring, *Computers & Education,* Vol. 44, pp. 53–68.

[16] L. F. Motiwalla (2007), Mobile learning: A framework and evaluation, *Computers & Education*, Vol. 49, No. 3, pp. 581–596.

[17] Y.-K. Wang (2004), Context awareness and adaptation in mobile learning, in: *Proceedings of the 2nd international workshop on wireless and mobile technologies in education*, IEEE Press, 2004.

[18] A. Syvanen, R. Beale, M. Sharples, M. Ahonen, P. Lonsdale (2005), Supporting pervasive learning environments: Adaptability and context awareness in mobile learning, in: *Proceedings of the 2005 IEEE International Workshop on Wireless and Mobile Technologies in Education*, IEEE Computer Society, 2005.

[19] Y. Cao, T. Tin, R. McGreal, M. Ally, S. Coffey (2006), The athabasca university mobile library project: increasing the boundaries of anytime and anywhere learning for students, in: IWCMC '06: *Proceeding of the 2006 international conference on Communications and mobile computing*, ACMPress, New York, NY, USA, pp. 1289–1294.

[20] D. Marinelli, S. M. Stevens (1998), Synthetic interviews: the art of creating a 'dyad' between humans and machine-based characters, in: *Proceedings of the 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, 1998.

[21] T. Liu and C. Choudary (2007), Scalable Coding and Wireless Streaming of Lecture Videos for Mobile Learning, *Advanced Technology for Learning*, Vol. 4, No. 2, 2007

[22] H.-Y. Kung, M.-Y. Wu (2005), The design and implementation of an adaptive mobile learning mechanism, in: *Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies* (ICALT05), 2005.

[23] E.R. Weippl (2007), Security Considerations in MLearning: Threats and Countermeasures, *Advanced Technology for Learning*, Vol. 4, No. 2.

[24] MPEG4, Overview of the MPEG- 4 Standard, Research report, MPEG Group, [on-line] Available at http://www.chiariglione.org/mpeg/standards/-mpeg-4/mpeg-4.htm (2002).

[25] MPEG7 home page, in: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[26] S. Cheng, H. Yu, Z. Xiong (2002), Enhanced spread spectrum watermarking of MPEG-2 AAC audio, in: *Proceedings of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 4, Orlando, FL, USA, 2002, pp. 3728–3731.

[27] RFC 3839: MIME type registrations for 3rd Generation Partnership Project (3GPP) Multimedia files, 2004.

[28] S.Ravi, A. Raghunathan, P. Kocher, S. Hattangady (2004), Security in embedded systems: Design challenges, *ACM Transactions on Embedded Computing Systems*, Vol. 3, No.3, pp. 461–491.

[29] C. Mclvor, M. McLoone, J. V. McCanny (2003), Fast montgomery modular multiplication and RSA cryptographic processor architectures, in: *Proceedings of Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 2003, pp. 379–384.

[30] K. Biddle, P. England, M. Peinado, B. Willman (2002), The darknet and the future of content distribution, in: *Proceedings of the ACM Workshop on Digital Rights Management*, Washington, DC, USA, 2002.

[31] S. He, M. Hu (2004), Performance study of ECC-Based Collusion-Resistant Multimedia Fingerprinting, in: *Proceedings of the 38th Conferences on Information Sciences and Systems, Princeton*, NJ, USA, 2004, pp. 827–832.

[32] D. Schonberg, D. Kirovski (2004), Fingerprinting and Forensic Analysis Of Multimedia, in: *Proceedings of the 12th annual ACM international conference on Multimedia*, New York, NY, USA, 2004, pp. 788–795.

Indexing and Retrieval of Multimedia Metadata on a Secure DHT

Walter Allasia and Francesco Gallo
Research and Innovation Department
EURIX Group
26, Via Carcano, Torino, Italy
E-mails: allasia@eurixgroup.com, gallo@eurixgroup.com

Marco Milanesio and Rossano Schifanella
Computer Science Department
University of Torino
185, Corso Svizzera, Torino, Italy
E-mails: milane@di.unito.it, schifane@di.unito.it

Keywords: multimedia metadata, digital rights, secure distributed hash table, peer-to-peer

Received: August 31, 2008

This paper proposes a decentralized, distributed and secure communication infrastructure for indexing and retrieving multimedia contents with associated digital rights. The lack of structured metadata describing the enormous amount of multimedia contents distributed on the the web leads to simple search mechanisms that usually are limited to queries by title or by author. Our approach is based on structured peer-to-peer networks and allows complex queries using standard MPEG-7 and MPEG-21 multimedia metadata. Moreover, security aspects limit the development of general purpose real applications using a peer-to-peer routing infrastructure for sharing digital items with an associated license. Accordingly, we propose a framework made up of a secure Distributed Hash Table layer based on Kademlia, including an identity based scheme and a secure communication protocol, providing an effective defense against well known attacks.

Povzetek: Predstavljen je sistem za učinkovito indeksiranje in doseganje digitalnih vsebin.

1 Introduction

Nowadays the growing of digital items exchanged on the web increases the need of their accurate description. We can define metadata as the description of the data. Even if it is possible to share multimedia items, it is very difficult or impossible to search them without appropriate description provided by content metadata. Usually people making use of web-sharing systems do not provide detailed metadata information, which in most cases is only limited to the title or the author. This lack of information determines the growth of unstructured information. Using metadata it is possible to structure the information and thus, on one side, to enhance and enrich the information related to a content and, on the other, to search and retrieve digital items. It is clear that more detailed are the metadata, more complex is the structure which they are inserted on.

Moreover, in order to reach a common understanding of metadata, it is important to adopt standards. The adoption of MPEG-7 [1] for describing metadata related to the digital items and of MPEG-21 [2] for describing metadata related to a governed content (i.e., with an associated license), as proposed in this paper, is a common approach used by an increasing number of scientific communities. The use of the standards mentioned above can improve the expressiveness of the query language for the multimedia items and can make governable the content distribution.

The enormous amount of media available on the web promotes the adoption of completely decentralized infrastructures, such as peer-to-peer (P2P) content sharing systems, that minimize the impact of a single point of failure fostering scalability, reliability and efficiency. Unfortunately, such approaches introduce a large spectrum of security flaws that limit the adoption in a real scenario. In fact, if it is true that digital contents are growing up very fastly especially in such distributed environments, it must be noticed that such systems usually offer poor functionalities for indexing and retrieving structured information. The main issue comes from the flat indexing space that affects these systems: the lack of a central entity offering a complete representation of complex information (i.e., the set of the metadata characterizing the digital items) results in a poorly expressive query language (e.g., parsing of the query string and pattern matching). Moreover most of these topologies are not providing any kind of content government and in the worst case they are not taking into account any digital rights associated to the exchanged resources.

We propose a decentralized, distributed and secure

communication infrastructure for the indexing and the retrieval of governed as well as ungoverned multimedia contents. Our approach, based on Distributed Hash Tables, allows complex queries to the system by means of complex multimedia metadata indexing. Moreover, the sharing of digital items on the basis of the associated license (either free or not), enables the usage of the P2P routing infrastructure for real applications, where a particular care has to be devoted to security aspects.

The main contributions of our work are summarized in the following:

- a decentralized scheme to index and retrieve structured metadata related to multimedia contents,
- a policy to manage digital rights expressed by MPEG-21 Rights Expression Language (REL) [3] profiles that enables the governed sharing of digital items along with the protection of the intellectual property,
- a secure structured overlay network that assures the basic security functionalities providing an effective defense against well known attacks,
- a Java-based prototype implementation that shows the feasibility of our approach.

The remainder of this paper is structured as follows. Section 2 presents an overview of the related studies available in the literature, while Section 3 describes the general model developed for the proposed framework. The indexing and retrieving schemes are discussed in Section 4, while Section 5 introduces a secure structured overlay infrastructure built on top of Kademlia that provides a defense against well known attacks. Moreover, Section 6 presents a Java-based implementation of the proposed system. Finally, Section 7 discusses some concluding remarks and future works.

2 Related Works

In this Section we focus on a general overview about the building blocks that compose the proposed framework. For network topology we adopted a structured P2P network based on a Distributed Hash Table, described in Section 2.1, where the fundamental properties are briefly discussed. In Section 2.2 an overview of security concerns related to DHTs, based on the available literature, is presented.

For metadata representation we adopted the MPEG-7 [1] and MPEG-21 [2] standards, which are outlined in Section 2.3. Concerning the governed content management, we adopted the solutions developed by the *Digital Media Project* (DMP) [4]. Accordingly, Section 2.3.1 describes the overall architecture of *Chillout* [5], the reference software implementation of the ISO/IEC 23000-5 (Media Streaming Application Format) standard.

2.1 Distributed Hash Tables

Distributed Hash Tables (DHTs) [6, 7, 8] are a class of distributed algorithms that provides the same functionality of a traditional hash table, by making available the mapping between a *key* and a *value*. DHTs are typically designed to scale to large numbers of nodes and to handle continual node arrivals and failures. The basic functionality provided by a DHT is the `lookup(key)` operation that returns the identifier of the node responsible for the `key`. In a DHT, nodes and objects are assigned with random identifiers (called node IDs and keys, respectively) from a large ID space. Given a message and a key, the DHT routes the message to the node with the node ID that is numerically closest to the key in a logarithmic number of hops with respect to the size of the network. In order to route a message, each node maintains a local routing table that contains information on a logarithmic subset of the entire system, granting scalability to the structured P2P system.

Even if DHTs offer a very good level of scalability and robustness, they suffer also from various drawbacks. First of all, in order to locate the node that stores a key, one needs to know in advance the exact identifier, but this cannot be always assumed at the application level. This phenomenon is known as the *exact match lookup* problem. As a consequence, distributed applications based on structured peer-to-peer overlay networks have to set up an interface to communicate with the P2P network providing the keys used for both routing messages and searching resources. A typical solution allows the insertion of meta-information and meta-keys extracted from the query string (such as in eMule[1] with Kademlia support).

Another relevant issue arises when a new node joins the network: it needs to know at least one living peer that is contacted in order to gather the necessary information to build the peer's state and the related routing table. Obviously, this node (called *bootstrap node*) represents a single point of failure: if it is off-line, the oncoming node can not enter correctly the system. However, usually the new peer holds a list of existing peers and it contacts each of them until an on-line node is reached. Of course, the presence of the bootstrap node raises also some security issues since the correctness of information provided is necessary to ensure a valid join mechanism. Furthermore, in most DHTs every information is replicated and cached in the system, to improve reliability and performance: this leads to the problem of balancing the trade-off between consistency and communication overhead between peers that need to update their cache. Finally, the DHT paradigm assumes that all peers equally participate to the system without any difference in terms of bandwidth, computational power or resource availability of nodes. In such a scenario, it is possible that low-capacity peers act as a bottleneck in terms of system performance.

[1] http://www.emule-project.net. Last visited: 15 Nov 2008.

2.2 Security Flaws on DHTs

Recently, a lot of effort has been put on securing DHTs [9] and the applications built on them. The usual robustness and efficiency of a DHT-based system can be overwhelmed by the malicious behavior of groups of peers that do not follow properly the DHT protocol. Examples of attacks that a DHT-based application has to face with can be divided regarding the targets: the overlay routing (e.g., eclipse attack, sybil attack, churn based attacks, and adversarial routing) or the applications (e.g., DDoS, attacks on Data Storage). In the following we will give a brief overview of all the attacks which are tipically carried against DHTs.

2.2.1 Threats

A popular family of attacks is known as *routing poisoning*. As active nodes' routing tables are maintained and renewed through a push-based approach (i.e., unsolicited messages, such as the publication of route tables of neighboring nodes or lookup messages sent from unknown nodes, supply an information that is used to update table's entries), it is possible for a malicious peer to inject random routing data into victim nodes, (e.g., during bootstrapping).
When carried against nodes, a particular form of routing poisoning is the so-called *eclipse attack* which aim is to separate a set of victim nodes from the rest of the overlay network, mediating most overlay traffic and effectively eclipsing correct nodes from each other's view. When carried against the stored contents on a DHT (i.e., making inaccessible the values of the DHT), the Eclipse attack is called *node insertion attack*: a vast number of nodes marked with identifiers numerically close to the target identifier are initiated, intercepting thus most of the lookup requests and answering with fake contents or not replying at all, effectively hiding the content.

Since typically there exists no verifiable link between the participating entity (human user or machine) and its identity (the nodeId), it is possible for any entity to show multiple identities to the system. The generation of multiple identities under a single entity is called Sybil attack and it undermines the redundancy property of a P2P system, because it enables the gathering of a large number of nodes on few machines, centralizing unsafely many keys' responsibilities and content replicas. The Sybil entities are usually exploited to increase the effectiveness of other attacks (e.g., Eclipse, DDoS) without needing huge computational resources or without the help of other colluding entities.

An index poisoning based attack [10] consists in inserting corrupted contents among the storages of a group of index nodes. A corrupted content might be something not related to the key for which it was stored, or even a fake information, like a reference to the wrong source. An attacker can make a bogus content highly visible by flooding fictitious records under 'strategic' indexes (e.g., among nodes responsible for "hot" keys), flushing legitimately stored content. In file-sharing applications, the most similar attack is the *content pollution*, that inserts on the DHT fake meta-data (i.e., meta-data that should be correct but that point to corrupted resources).

A distributed denial of service attack consists in inducing a large number of nodes of the overlay to generate a huge amount of messages to be sent to a target entity located internally or externally the P2P network. It can be achieved with a redirect technique [11], carried out through an index poisoning attack. In file-sharing systems, the attacker can insert meta-data related to a very popular content, pointing to the target IP address as a source of such a file: the victim will be overflowed by connection requests until the 'polluted' content will be kept in index nodes' storage.

Concluding this overview on the attacks, it is worth notice that some studies [12] show that in the Kad network at least half of the network is prone to a Man In The Middle attack. To avoid this, communicating must be sure about the integrity of messages and about the identity of the sender. An authenticated channel between endpoints can instantly exclude a third malicious entity.

2.2.2 Defenses

Most of the overlay routing attacks countermeasures are given in terms of routing protocol changes or access control policies. An exhaustive overview of the commonly used distributed access control mechanisms is given in [13], stressing the difference between the different threshold signatures. Authors underline that the use of RSA for generating keys in a distributed environment leads to an high communication and computation overhead, particularly harmful for mobile and ad-hoc networks. Saxena et al. [14] developed an identity-based group admission control technique that overcomes the drawbacks of previous certificate-based approaches, presenting ID-GAC (ID-based Group Admission Control), based on the threshold version of BLS signature scheme, an identity-based mechanism since the membership token used to prove membership is derived from the group member's identity. The use of a super singular elliptic curve influences the overall cost of the scheme. The whole scheme comes along with a distributed membership revocation mechanism based on the membership revocation lists.

A possible approach to locate Sybil nodes is periodically sending a different challenge to each node: requiring a high computational effort to be solved, one machine cannot solve a challenge for each Sybil node it hosts within a specified short time interval. The main issue within this approach is the difficulty to practice it in an heterogeneous domain [15]. A central authority that assigns a certified nodeId only after a user registration process might limit this phenomenon, because the time required to the creation of a new node would be considerably longer.
An exhaustive overview of the different behaviors of peers in the KAD network is given in [16]: among other results, it's clear that node identifiers are not necessarily persistent as was assumed in previous works. In [17], authors con-

sider the vulnerability of KAD against Sybil Attack and point out that a solution is to prevent a peer from choosing its own ID and avoiding a peer to obtain a large number of IDs. For doing so, they sketch out a centralized solution that makes it impossible for an attacker to obtain arbitrary KAD IDs: a central agent binds the KAD ID to a cellular phone number.

Sybil Attack is also the core of the work in [18] and [19]. In the first work, a resistant routing strategy is introduced on a variant of Chord, assuring that lookups are performed using a diverse set of nodes, and thus that at least a subset of the nodes involved in the lookup process is not malicious. As a consequence, the lookup process makes forward progress, not only converging fast to the destination, but also minimizing the number of trusted bottlenecks: when choosing the next node in the path, the variant will take into account the sources of information about the previous hops, and strive to avoid relying on a single trusted bottleneck. In [19] an admission control system for structured P2P systems is given. The system constructs a tree-like hierarchy of cooperative admission control nodes, from which a joining node has to gain admission. The admission control system is implemented by the nodes, and it examines joining nodes via client puzzles. The burden of self-organization and admission control is placed on the peer-to-peer nodes themselves. For this reason, the computational load of these activities must be low. Analysis shows that these costs are vanishingly small for all nodes in the network. Admission Control System (ACS) defends against Sybil attacks by adaptively constructing a hierarchy of cooperative admission control nodes. A node wishing to join the network is serially challenged by the nodes from a leaf to the root of the hierarchy. Nodes completing the puzzles of all nodes in the chain are provided a cryptographic proof of the examined identity.

A tool which could effectively combat the content pollution and the index poisoning attacks is the use of credentials, bound to the content, provided by the owner of the content during the insertion phase: if the content is bound to the identity of an owner, when a fake resource is found, it is possible to trace back to content creator. If the application implements a reputation system, it could be possible to penalize or even to ban a malicious node.
Credentials and reputation systems can also be used against DDoS: as it would be too costly to oblige replica nodes to verify the authenticity of each inserted content, it is necessary to adopt a reputation system so that peers who have made incorrect insertions are recognized as soon as possible and banned from the network.
Against the Eclipse attack, an anonymous auditing technique is proposed in [20], but still it is shown to be ineffective against Node insertion attack: the introduction of a third party trusted certification service that assigns randomly generated certified identifiers to nodes seems to be an effective solution to prevent this attack.

S/Kademlia [21] is a secure key-based routing protocol based on Kademlia [22] that has a high resilience against common attacks by using parallel lookups over multiple disjoint paths, limiting free nodeId generation with crypto puzzles and introducing a reliable sibling broadcast, needed to store data in a safe replicated way. In order to make Kademlia more resilient they suggest limiting free nodeId generation by using crypto puzzles in combination with public key cryptography, extending the Kademlia routing table by a sibling list, reducing the complexity of the bucket splitting algorithm and allowing a DHT to store data in a safe replicated way, and finally a lookup algorithm which uses multiple disjoint paths to increase the lookup success ratio.

In [23] periodic routing table resets, unpredictable identifier changes and a rate limit on routing table updates are given, in order to make attackers unable to entrench themselves in any position that they acquire in the network, and also to make them unable to fix an appropriate strategy for addressing some specific nodes. Authors propose also a practical defense against the eclipse attack, extending the Bamboo DHT[2].

A distributed node ID generation scheme would limit the rate in which an attacker can obtain IDs. The former authors of Pastry [24] require prospective nodes to generate a private/public key pair such that the hash of the public key has the first p bits equal to zero [25]. They also suggest binding the IP address of the node with its ID. To overcome the possibility of an attacker to accumulate node IDs they suggest periodically to invalidate node IDs and using different setting for the hash initialization. However, this would require legitimate nodes to obtain new IDs every time this happens. Authors show how the use of secure routing can be reduced by using self-certifying application data.

Finally, an admission control framework suitable for different flavors of peer groups nd match them with appropriate cryptographic techniques and protocols is presented in [26].

2.3 Multimedia Metadata Representation

MPEG-7 [1], formally named Multimedia Content Description Interface, provides a rich set of standardized tools to describe multimedia contents. It mainly focuses on description of the digital items, without considering how and where this information is used. In particular, the MPEG-7 descriptions of content may include (1) information describing the creation and production processes of the content (director, title, short feature movie), (2) information related to the usage of the content (copyright pointers, usage history, broadcast schedule), (3) information of the storage features, (4) on spatial, temporal or spatio-temporal components or about low level features (colors, textures, sound timbres, melody description) and many others.

MPEG-7 standard has been included in several metadata language, such as ODRL (Open Digital Rights Language[3]) and has been coupled with other important TV ontologies

[2] http://bamboo-dht.org. Last visited: 15 Nov 2008.
[3] http://www.w3.org/TR/odrl. Last visited: 15 Nov 2008

(e.g., TVAnytime RMPI [27]). Concerning digital rights, MPEG-7 provides a standard XML schema and the metadata to define conditions for accessing the content (including links to a registry containing intellectual property rights data and price) and additional information about the content (copyright pointers, usage history, broadcast schedule). An MPEG-7 Query Format reached the Final Committee Draft, during the MPEG meeting on October 2007. Moreover several query frameworks based on MPEG-7 are still under investigation [28].

MPEG-21 [2] differs from MPEG-7 because it aims to define a normative open framework to be used by all the players in the delivery and consumption value chain. This framework will provide an open market to content creators, producers, distributors and service providers. The goal of MPEG-21 is the definition of a standard technology needed to support users in order to exchange, access, consume, trade and otherwise manipulate digital items in an efficient, transparent and interoperable way. In particular, part 5 of MPEG-21 defines a Rights Expression Language (REL) to be used in the description of customized rights applied to any digital item, since it is seen as a machine-readable language that can declare rights and conditions defined in the Rights Data Dictionary (also standardized by MPEG-21). Rights metadata are expressed by means of MPEG-21 REL, which describes the license associated to a specific resource, along with several available rights (*play*, *copy*, *modify*, *print*, etc.). According to the schema shown in Figure 1 [29] we can imagine the license as made up of an issuer (with multiplicity 0 or 1), an undefined number of grants (multiplicity 0 or more), and a principal (multiplicity 0 or 1). The issuer is the owner of the rights associated to a given content (eventually coincident with the creator or distributor of the resource) and can assign a given right (e.g., the authorization to copy or modify the content) to the principal. For example, in the wide commonly used *CreativeCommons* [30] licenses the principal is not specified since this kind of license is intended for everyone.

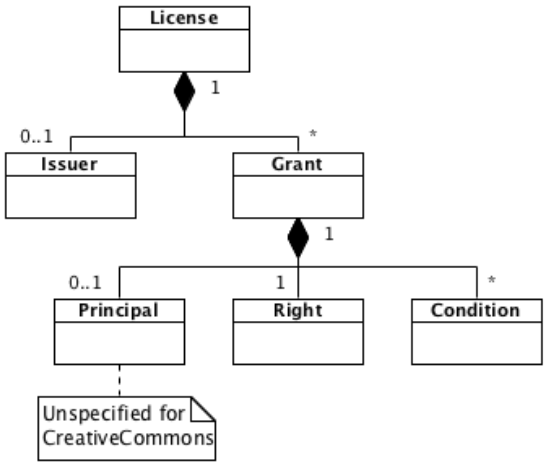


Figure 1: Simplified diagram of a REL license.

2.3.1 Chillout

Chillout [5] is the reference software of the Digital Media Project (DMP) [4]. DMP is a no profit organization that has recently approved a version 3.0 of its specification, called Interoperable DRM Platform (IDP-3.0). Chillout is also the reference implementation of ISO/IEC 23000-5 Media Streaming Application Format [31], addressing the distribution of governed content over streaming channels. The most important technologies adopted by Chillout are: (a) a data structure capable of hosting different data types accompanying a resource (e.g., audio, video, image, text, etc.), (b) a content identification system, (c) a set of technologies for content protection, (d) the Rights Expression Language, (e) a file format for storing digital items and resources and (f) a technology to transmit digital items in streaming mode.

Two file formats for managing digital contents are used as depicted in Figure 2: DCI (DMP Content Information) and DCF (the DMP Content File) [32]. The DCI is a standard XML-based format which is intended mainly to express the license metadata and is compliant with two MPEG-21 REL profiles: the Open Access Content (OAC) profile [33], for expressing equivalent CreativeCommons licenses, and the Dissemination And Capture (DAC) profile [34], mapping the TV Anytime RMPI [35] licenses, used in the broadcasting domain. The specification of the DCI allows also to include the MPEG-7 representation for the content. The DCF file has been conceived as a container of the DCI and the resources as well and we extract a subset of the metadata contained in the DCF for indexing. The resources can be stored within the DCF file or can be referred to by means of pointers.

3 Model

In the previous Sections we have presented the building blocks (secure DHT, MPEG multimedia metadata and Chillout) that have been used in our solution in order to create a prototype system that is able to *share governed contents on P2P networks*, where *share* here means the possibility to publish, index, search, retrieve and consume a digital item and *governed* refers to the fact that each digital content distributed on such system is governed according to its associated license. It is worthwhile pointing out that a DRM system could use the proposed solution as the underlying software to manage (create, index, retrieve) govenred contents, demanding to an other application software placed on top of it to manage or not the associated digital rights. This solution allows also the integration of the proposed prototype with proprietary DRM solutions, where the content representation is based on MPEG standards. Moreover, despite the common feeling about P2P networks in relationship with abuse or violation of digi-

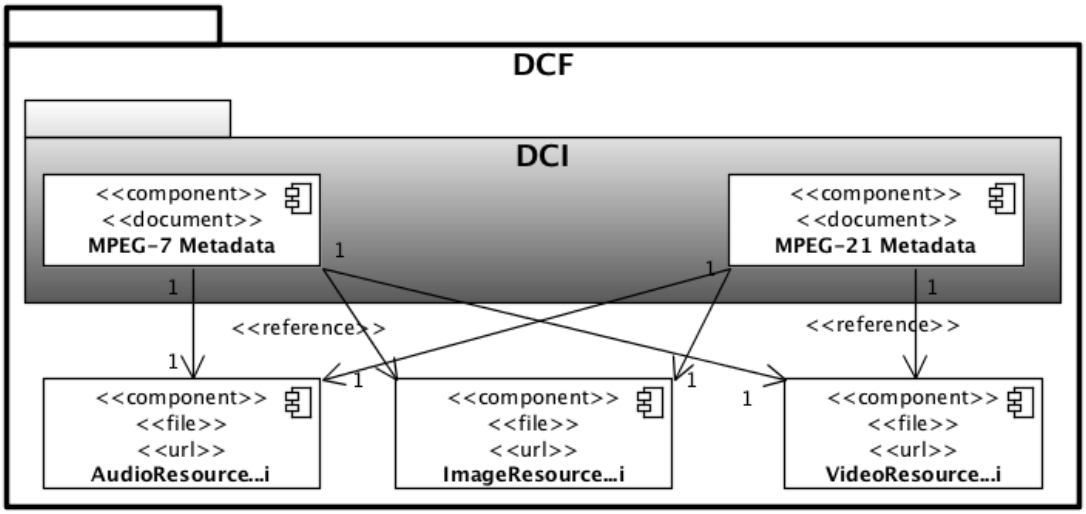


Figure 2: DCF and DCI structure diagram.

tal rights and intellectual property rights in general, mainly due to the sharing of copyrighted or otherwise licensed content, the software solution proposed in this paper proofs that it is possible to have content government on these popular networks and is also possible to make them secure.

We make use of the MPEG-7 standard for expressing the metadata related to the digital content itself, describing the user metadata (e.g., title or author) as well as the metadata describing the content as visual descriptors (e.g., ScalableColor or HedgeHistogram). We have adopted the MPEG-21 standard for expressing licenses because MPEG-21 REL provides several profiles for specific environments and purposes (broadcasting, mobile applications,...), which guarantee high interoperability with other rights languages and therefore it is able to express most of the possible licenses. As described in Section 2.3.1 Chillout can manage governed content using MPEG-7 and MPEG-21 representation, which is contained into a DCI structure, specified by ISO/IEC 23000-5. Hence the proposed solution is able to share on a secure DHT the DCF files and to index the metadata stored in the DCI files.

As shown in Figure 3, our approach is made up of three logic layers:

- *User Interaction Layer*, where the several user software components communicate with the application layer providing and consuming digital contents.

- *Application Layer*, which is in charge for extracting the information to be indexed and for communicating with the DHT layer in order to index the related keys.

- *Overlay Layer*, which is responsible for the management of the overlay network and the routing of messages.

On top of the layered architecture, the *User Interaction Layer* describes the way an entity can interact with the application exploiting the provided functionalities. As shown by the components depicted in Figure 3, a user device can play different roles:

- *Content Creator*, which is the component responsible for the creation of governed content (in DCF format), making use of user resources and the associated licenses (expressed in the DCI file).

- *Content Provider*, which is the component responsible for providing governed contents that can be created by the same user as well as by others.

- *Player (End User)*, which is the component that can consume the resources according to the associated licenses. When the user asks the system to consume a resource, it recognizes which rights are guaranteed to the current user (e.g., copy, play, modify, distribute) and can enforce them.

The *Application Layer* is made up of three main components: *Retrieving*, *Indexing* and *Exchanging*, as shown in Figure 3. The *Retrieving component* provides functionalities for (a) extracting a defined subset of MPEG-7 and MPEG-21 metadata from the DCF file associated to the governed resource, (b) computing the identifiers associated with the extracted metadata, (c) querying the underlaying DHT with the computed identifiers and (d) collecting and merging the lookup results.

The *Indexing component* provides functionalities for (a) extracting a defined subset of MPEG-7 and MPEG-21 metadata from the DCF file associated to the governed resource, (b) computing the identifiers associated with the extracted metadata, (c) inserting the governed resource in a

storage layer and (d) inserting the relative mappings in the DHT.

A user can search for resource related metadata (e.g., the title in MPEG-7), license related metadata (e.g., the issuer in MPEG-21 REL) or a combination of the two. A detailed description of the *Retrieving* and *Indexing* components can be found in Section 4.

The *Exchanging component* communicates with the *Transport component* by mean of two socket connections (represented in Figure 3 using UML 2.0 [36] conventions), one for exchanging metadata information that are basically DCI documents and the other for exchanging the real digital content, for example as byte array. This communication is asynchronous and completely separated. The user can make use of the metadata exchanging component looking for several resources and can decide to download only one of them, making use of the other exchanging component, the one for accessing the actual contents.

Finally, the *Overlay Layer* is made up of the *DHT* and the *Transport* components. The former is responsibile for the DHT management and is described in Section 5 while the latter is responsible for exchanging/downloading the contents between peers and also for exchanging the full metadata available in the DCF and contained in the DCI. In order to provide a system open to further extensions, we layered the application core functionalities of the DHT component under a *facade design pattern* which can be considered as a bundle of interfaces widely used by different DHT implementations. This choice improves the system flexibility, allowing the choice of other DHT implementations with no (or only minor) changes.

4 Indexing and Retrieving

A goal of the proposed approach is to provide a fully distributed system that exploits the scalability, resiliency, and efficiency properties of DHTs in order to index and retrieve audiovisual contents through their descriptions.

In all DHTs, to each node and resource, an identifier computed from the same space is given. This means that there is no way, starting from an identifier, of knowing if this is the index of a node, of a resource or anything else. What can be exploited within these distributed algorithms is the key consistency and the collision avoidance of the used hashing functions. Once computed several identifiers, at an application level it will be possible to address any domain specific data structure complexity. The DHT layer allows, in any case, the convergence of the routing mechanism and the scalability of the system, as the diameter of the system will never be bigger than $O(logN)$, for N nodes in the network.

Accordingly, it is clear that there is a discrepancy between metadata representation and the way in which information are stored in a DHT-based infrastructure. In the first case, the content is described by a structured XML-based formalism, e.g., MPEG-7 and MPEG-21 documents, in the second case, the information are codified in a flat set of $\langle key, value \rangle$ relations. To fill up this gap, we proposed an iterative indexing and retrieving scheme [37] that is similar to the hierarchical indexing scheme described in [38].

Let's consider a generic audiovisual content R that is associated with a set of metadata. Such metadata are extracted during the indexing phase from the MPEG-7 and/or MPEG-21 documents related to R. As described in Section 2.3 both MPEG-7 and MPEG-21 standards contain a large spectrum of data describing multimedia contents and digital rights. Therefore, it is evident that to index the complete metadata knowledge could represent a very expensive computational and spatial cost. To lighten the load of each node, we decided to not index all the metadata: we chose a subset of the overall tags, used for a first step of the query process. To refine the result set it is possible to query locally the retrieved resources against the complete schema through well-established approaches. This hybrid strategy can lead to a good trade-off between efficiency, scalability and query expressiveness.

In more details, given the resource R, we define the subset of medatata $M_R = \{m_0, m_1, ..., m_n\}$ where the generic m_i has the form $m_i = (tag_0|[attribute_0]|\dots|tag_m|[attribute_m]|value)$. In other words, each metadata item is composed by chaining the tags and the optional tags attributes with the corresponding value. For sake of simplicity, assume a MP3 audio file that has to be indexed on the system. Figure 4 shows the MPEG-7 document related to the song. In this scenario, we can describe the title by way of the metadata item $m_i = (title|songtitle|Times\ Like\ These)$ or the genre through $m_j = (genre|name|Acoustic\ \ Rock)$. After the selection of the metadata, we compute the identifier id_R (calculated applying the hashing function $hash()$ of the resource itself) and the set of identifiers $I_{M_R} = \{id_{m_0}, id_{m_1}, ..., id_{m_i}\}$ where the generic id_{m_i} is equal to $hash(m_i)$. Each identifier must reference id_R, in order to allow metadata based queries. We insert on the DHT a set of $\langle key, value \rangle$ pairs in the form $\langle id_{m_i}, id_R \rangle, \forall\ id_{m_i} \in I_{M_R}$, along with the relation $\langle id_R, R \rangle$. This basic scheme allows a user to retrieve the resource associated to a metadata m_i. In fact, during the retrieving phase the system calculates the identifier id_{m_i} and, by means of a $lookup(id_{m_i})$, the related resource identifier id_R. Then, R itself is obtained. For instance, let's suppose that a user wants to find the song entitled *"Times Like These"*. She submits the request to the system that computes m_i following the rules depicted above, then it calculates id_{m_i} and, by means of a $lookup(id_{m_i})$, it retrieves the identifier of the corresponding resource. At last, a subsequent lookup is able to get (directly or indirectly) the requested resource.

A key aspect of our approach is the ability to index and retrieve audiovisual items with associated digital rights. We can divide contents between governed and ungoverned. Ungoverned items do not have licenses associated and the keywords to be indexed are just MPEG-7 elements. Gover-

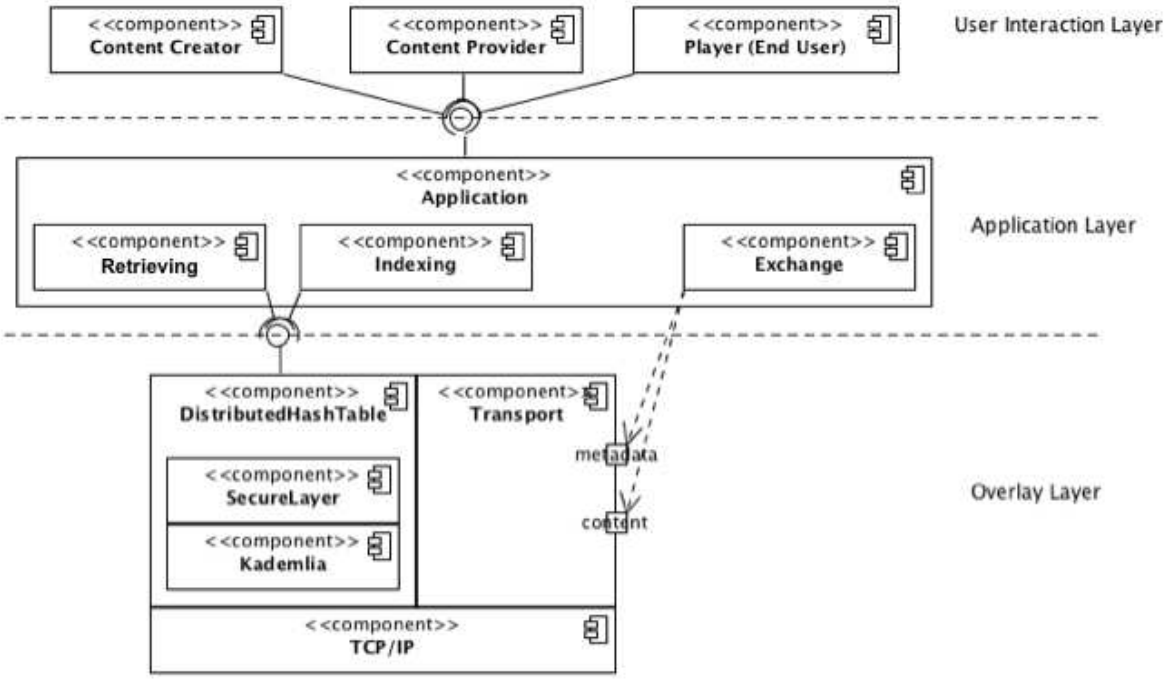


Figure 3: System overview

```
<?xml version="1.0" encoding="UTF-8"?>
<Mpeg7  xmlns="urn:mpeg:mpeg7:schema:2001"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001
     http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-7_schema_files/mpeg7-v2.xsd">
     <Description xsi:type="CreationDescriptionType">
         <CreationInformation id="jj-2005-onon-track-01">
             <Creation>
                  <Title type="songTitle">Times Like These</Title>
                  <Title type="albumTitle">On and On</Title>
                  <Creator>
                       <Role href="urn:mpeg:mpeg7:RoleCS:2001:PERFORMER"/>
                       <Agent xsi:type="PersonType">
                            <Name>
                                      <FamilyName>Johnson</FamilyName>
                                      <GivenName>Jack</GivenName>
                            </Name>
                       </Agent>
                  </Creator>
                  <CreationCoordinates>
                       <Date><TimePoint>2003</TimePoint></Date>
                  </CreationCoordinates>
             </Creation>
             <Classification>
                  <Genre href=" urn:id3:cs:ID3genreCS:v1:80"><Name>Acoustic Rock</Name></Genre>
             </Classification>
        </CreationInformation>
   </Description>
</Mpeg7>
```

Figure 4: Example of a MPEG-7 description for a MP3 audio file.

ned items have a license and we have defined the following structure to be indexed: for each right described in the license we index three MPEG-21 REL tags: *issuer*, *right*, *principal* (see Section 2.3). Although typical licenses contain one or more grants, we assume in the following a single issuer and a single principal for each right and for every

grant expressed in the license we index the bundle of issuer, right and principal linking the associated content. Hence, the DHT contains the indexes of the general purpose metadata and in addition, for governed resources, the bundle of grants linking the digital item.

Let's consider again the resource R with an associated MPEG-21 REL license as shown in Figure 5. We extract the following metadata elements:

$m_{issuer} = (issuer|keyholder|keyname|value)$
$m_{right} = (grant|value)$
$m_{princip} = (principal|keyholder|keyname|value)$

For instance, we have that $m_{issuer} = (issuer|keyholder|keyname|Jack\ Johnson's\ key)$ and $m_{right} = (grant|play)$. The principal is not defined since the item is governed by a *Creative Commons License*. Afterwards, a key is calculated for the metadata, i.e., id_{issuer}, id_{right} and $id_{princip}$ respectively, and the mappings $\langle id_{issuer}, id_R\rangle$, $\langle id_{princip}, id_R\rangle$, and $\langle id_{right}, id_R\rangle$ are put on the DHT as explained in the MPEG-7 scenario.

In order to allow complex queries, all possible combinations of those three metadata are inserted. In other words, we derive the following relations:

$\langle hash(m_{issuer}|m_{right}), id_R\rangle$
$\langle hash(m_{right}|m_{princip}), id_R\rangle$
$\langle hash(m_{issuer}|m_{right}|m_{princip}), id_R\rangle$

In this scenario, a user could search for "all the digital items issued by someone", or could submit composite queries like "all the contents with a grant of copy issued by someone", beyond looking for titles and authors.

In summary we are indexing rights metadata on a structured overlay network, allowing users to search governed resources looking for specific issuers, grants or principals. It is worth noting that our system easily enables keyword-based queries like eMule with Kademlia support does. In this case, we index keywords extracted from the file name combining them to allow complex queries as described above.

5 Secure DHT Layer based on Kademlia

As previously underlined, one of the main concern that limits a broad adoption of a DHT-based content sharing platform is the security aspect. In this Section we will describe a communication protocol and an identity management scheme that provide a secure layer on which general purpose applications can be built.

The adversary model considered here is composed by nodes in the DHT system (with reference to Kademlia) that do not properly follow the protocol. We assume that a malicious node is able to generate packets with arbitrary contents (including forged source IP addresses) and, furthermore, to overhear or modify communications between other nodes.

Kademlia [22] is a structured P2P system featured by the use of a XOR metric for computing distance between points in the identifier space. In Kademlia every node has a random 160-bit *nodeId* and maintains a routing table consisting of up to 160 k-buckets. Every k-buckets contains at most *k* entries with <IP address, UDP port, NodeId> triples of other nodes, with *k* as a redundancy factor for robustness purposes. Buckets are arranged as a binary tree and nodes get assigned to buckets according to the shortest unique prefix of their nodeIds.

Kademlia combines provable consistency and performance, latency minimizing routing, and a symmetric, unidirectional topology.

The Kademlia protocol is vulnerable to all the attacks introduced in Section 2.2, even if it can mitigate the harmfulness of some of them. Nodes' identifiers are not certified and they can be generated at will on the local node, so it's possible to quickly instantiate a large number of Sybil nodes with arbitrary Ids in order to complete a node insertion attack. There is no credential associated with contents maintained in storages and no control is performed by replica nodes over the information stored in the DHT thus allowing the index poisoning and derivative attacks. There is no authentication protocol between nodes. Nevertheless, k-buckets provide resistance to certain DoS and index pollution attacks; in fact, one cannot flush nodes routing state by flooding the system with new nodes. Kademlia nodes will only insert the new nodes in the k-buckets when old nodes leave the system. Unfortunately, it is very easy to inject into a route table information relating to contacts whose identifier is very close to the victim node Id, because of the bucket splitting procedure.

Finally, it is possible to affect the lookup procedure to lead the searching node to contact a set of replica peers controlled by the attacker. The Kademlia lookup procedure for a key χ starts selecting α nodes whose ids are the nearest to the local id and sending to each of them a FIND-NODE(χ) RPC. The response to these messages are list of triples $< IP, Port, NodeId >$ that locates the contacts closest to among all the entries of the queried nodes' routing tables. The lookup initiator selects, among all received triple, α contacts whose nodeId is closest to χ and iterates the same procedure until it gets responses from the k nodes closest to χ it has seen. If a malicious node receives a FIND-NODE RPC, it responds with k triples that identify colluding nodes whose id is claimed to be close to the lookup key. The searching peer has no way for verifying messages and it will trust every response.

5.1 Protocol

In order to nullify or to reduce the impact of the DHTs' vulnerabilities, we define a framework that includes an identity based scheme and a secure communication protocol that may provide an effective defense against well known attacks. For further details refer to [39]. The proposed approach is layered on Kademlia and its architecture is based

```
<?xml version="1.0" encoding="UTF-8"?>
<license xmlns="urn:mpeg:mpeg21:2003:01-REL-R-NS"
  xmlns:mx="urn:mpeg:mpeg21:2003:01-REL-MX-NS" xmlns:m3x="urn:mpeg:mpeg21:2006:01-REL-M3X-NS">
  <grant>
     <mx:play/>
     <digitalResource licensePartId="jj-2005-onon-track-01">
         <nonSecureIndirect URI="urn:newspaper:news:2005_07_10-12H-00M"/>
     </digitalResource>
     <m3x:copyrightNotice noticeType="ShowBeforeExercise">
         <m3x:copyrightString> Written by Jack Johnson, 2005</m3x:copyrightString>
     </m3x:copyrightNotice>
  </grant>
  <issuer>
     <keyHolder>
        <info>
           <dsig:KeyName>Jack Johnson's key</dsig:KeyName>
        </info>
     </keyHolder>
  </issuer>
</license>
```

Figure 5: Example of a MPEG-21 REL license for the audio file described in Figure 4.

on the presence of a Certification Service (CS). The CS can be a centralized or decentralized authority whose task is to generate random nodeIds and to certify the link between nodeIds and users' identities by signing peculiar tokens. To accomplish this, we suppose that a classic public key cryptography scheme is used: in this section we assume that the CS is a centralized authority owner of a public key known to every Kademlia node, and holder of its private counterpart. Similarly, we assume that each user who intends to take advantage of the network services should be in possession of a key pair. The following notation is used throughout:

A, B	:	nodes
$NodeId_A$	:	A's Kademlia Id
$UserId_A$	:	A's user identifier
K_A^+, K_A^-	:	A's public and private key
K_{CS}^+, K_{CS}^-	:	CS public and private key
$Sign(m, k)$	:	message m signed with key k
$H(o)$	:	hash code of the object o
$AuthId_A$	:	node A's authenticated id
$Auth_{AB}$	:	authentication by A for B
ts, TTL	:	timestamp, time to live
$a\|\|b$	:	concatenation of strings

The proposal enhances the join procedure, the node interaction protocol and the content storage procedure defined by Kademlia. In a preliminary initialization phase a node applies to the Certification Service for a certified NodeId and for bootstrap information; since the certified NodeId has an extensive temporal validity, initialization is not executed at every bootstrap but only periodically. After the initialization, the node performs the network join procedure to take part to the overlay. In order to correctly interact with other nodes, the newly joined one must follow a communication protocol for incoming and outcoming messages; especially, the node must produce special credentials related to every content to be inserted in the DHT.

5.1.1 Initialization

Node A must obtain its own certified id, in order to interact with other peers. To this aim the node sends a request to the CS containing an identifier and its public key:

$$NodeIdReq = UserId_A, K_A^+$$

The $UserId_A$ is the identity by which user A presents himself to the network community. It is an identifier of a generic account of user A and whose validity must be verifiable by the same CS. It may be assumed that the $UserId$ is an existing and verifiable identity, (e.g., an OpenID URL or an email address), in which case the CS should initiate an interaction with an external authority (e.g., an Identity Provider, a mail server) to verify its effectiveness. Otherwise the same CS could be able to maintain user accounts and verifying the identity with a password request.

The CS makes the $UserId$ verification procedure (whose steps depend on the nature of the $UserId$ itself), and then binds the user identity with his public key and with a $NodeId$ by producing the following token:

$$AuthId_A = Sign(NodeId_A||UserId_A||K_A^+||exp_A, K_{CS}^-)$$

The $NodeId$ is randomly chosen; exp_A is a timestamp that establishes the expiration date of the signed $NodeId_A$. The CS keeps track of the association between $UserId$ and $AuthId$, so that all subsequent $NodeIdReq$ received by the same users receive in response the same $AuthId$ passed earlier, unless it is expired or close to expiration. This is a precaution to avoid the CS producing useless signatures. Then, the CS sends to the client a response message structured as follows:

$$CS \rightarrow A : AuthId_A, Sign(bootstrapList, K_{CS}^-)$$

The $bootstrapList$ is a list of triple $< NodeId, IP, port >$ that points to a set of nodes

that the CS assumes active; by contacting at least one of these nodes, the peer can join the network. The way the CS obtains the entry of bootstrap list is described in Section 5.1.5.

5.1.2 Join

Once initialization step is completed, the node may initiate the network join procedure as described by the Kademlia protocol, namely sending a lookup request for its own $NodeId$ to one of the bootstrap contacts. However, once obtained an $AuthId$, it is important that the nodes avoid contacting the certification service, unless if necessary. After making the first join using information obtained from the $bootstrapList$, each node should get in a different way a list of nodes to be contacted for subsequent join operations. For example a node can maintain its own list of trusted bootstrap nodes, or the same CS could periodically insert a signed $bootstrapList$ in the DHT, so that every active node could download it before disconnection and use it for its next join. Only if all the known nodes are off-line the CS will be contacted again to request a new $bootstrapList$.

The messages sent by the nodes during the join procedure must follow, like any other message, the protocol described in the next paragraph.

5.1.3 Nodes interaction

A node A can successfully send a RPC (join primitive included) to a node B and obtain a proper response only if both A and B observe the following communication protocol:

I $A \rightarrow B : NodeId_A, N1$

II $B \rightarrow A : NodeId_B, N2$

III $A \rightarrow B : AuthId_A, Auth_{AB}$, RPC-REQ

IV $B \rightarrow A : AuthId_B, Auth_{BA}$, RPC-RES

We call this four way exchange a $session$ between A and B. RPC-REQ and RPC-RES fields are respectively the request and response RPC defined in Kademlia; $N1$ and $N2$ are randomly generated nonces. Messages sent at steps I and II must be somehow marked differently (e.g., different opcode), to distinguish the request from the response.

Authentication tokens are structured as follows:

$$Auth_{AB} = Sign(NodeId_B||N2||H(\text{RPC-REQ}), K_A^-)$$
$$Auth_{BA} = Sign(NodeId_A||N1||H(\text{RPC-RES}), K_B^-)$$

In step III (and IV), the receiving node checks signatures (in $AuthId$ and $Auth$), expiration times validity, equalities between nonces, and equalities between $NodeId$ in step I (and II), in $Auth$, and in $AuthId$.

In steps III and IV, the receiving node performs the following controls:

1. Validity of $AuthId$ signature
2. Validity of $AuthId$ expire time
3. Validity of $Auth$ signature
4. Equality between the $NodeId$ contained in the $Auth$ and the receiver's $NodeId$
5. Equality between the nonce contained in the $Auth$ and the nonce sent previously
6. Equality between the $NodeId$ contained in the $AuthId$ and the $NodeId$ received at step I or II
7. Check of the RPC hash

Signature and expiration time validity checks on $AuthId$ demonstrate the existence of a valid and randomly generated $NodeId$, associated with an $UserId$ and with a public key; validity of signature in $AuthId$ and equality check on $NodeId$ assures that the sender is the same entity certified by $AuthId$ and that the present node is the correct recipient of the message. Equality checks on nonces in $Auth$ and the ones received previously protect against replay attacks. A's verification of $NodeId_B$ included in $AuthId_B$ assures that B is really the node that A wanted to contact; B's verification of $NodeId_A$ included in $AuthId_A$ proves that the RPC has been called by the same node that started the session. Finally, both peers execute an integrity check on the RPC hash to verify that no attacker has replaced the original RPC with a bogus one. The reader should observe that nonces are used against man in the middle attacks instead of exchanging timestamps because we cannot assume that hosts are synchronized to a common clock.

5.1.4 Content storage system

RPCs follow Kademlia's definitions, except for the store RPC. Let A be a node, owner of a content Obj. If A wants to store Obj in the DHT it locates via lookup the k nodes closest to the content key and then sends to them a store message structured as follows (suppose that B is a generic replica node):

$$A \rightarrow B : AuthId_A + Auth_{AB} + StoreRPC$$
$$StoreRPC = k||Obj||Cred$$
$$Cred = Sign(UserId_A||k||H(Obj)||ts||TTL, K_A^-)$$

$Cred$ binds the $UserId$ to the key for which the content was inserted and to the hash code of the content, so that is subsequently possible to prove that the owner had inserted the content Obj at the key k. $Cred$ includes also a timestamp and a time to live to specify the content submission time and its persistence period. During the periodic content spreading procedure, all replica nodes send store messages keeping the original credentials associated with each content. A node performing a lookup for contents related to a key χ receives all the objects marked with χ from replica nodes responsible for that key; before passing the content to the application, the node must verify the credentials signature and the object hash and must discard the object if the check fails.

If the application ascertain that the content is somehow polluted (e.g., the key that marks the content is not related

with it), it can benefit from the information included in the credentials to penalize the owner of the content. This could be simply accomplished by instructing the underlying node to blacklist the cheater user in order to refuse all the incoming requests marked with the malicious node's $AuthId$. The description of a reputation service that can manage feedbacks from the users and the details concerning a possible revocation policy for the identifiers of misbehaving users, are beyond the goal of this paper. However, it is important to say that an effective reputation manager, that can be external to the network, as well as integrated in the application, can help to exclude more rapidly the polluter from the whole network. Nevertheless, the propagation of polluted content is largely limited due to credentials' verification.

5.1.5 Bootstrap list construction

The bootstrap node selection is a problem inherent to the fully distributed nature of P2P networks. The bootstrap information acquisition process must prevent an attacker to manipulate bootstrapping information to let a victim join a malicious parallel network. Kademlia does not face the bootstrap node selection problem.

The CS maintains a list of active peers in a cache, where a generic entry stores the following information:

$$CacheEntry = (NodeId, IPaddress, UDPport, ts)$$

The CS probes nodes in the cache, controlling a DHT node, marked with a self signed $AuthId$, that runs a sequence of FIND-NODE RPCs for random generated keys. The CS adds to its cache the pointers to the nodes that replied to the FIND-NODE RPCs, then it can iterate the procedure until it gathers enough contacts for cache replacing. A least-recently cache replacement policy is implemented, except that active nodes are never removed from the list: if the cache is full then the least-recently seen node is pinged. If it fails to respond, it is replaced with a newly discovered one. Otherwise, if the least-recently seen node responds, it is moved to the tail of the list, and the new contact is discarded.

5.2 Discussion

In this section we discuss how this proposal strongly limits dangerousness of attacks described in Section 2.2.1.

Routing attacks In Kademlia, the sender contact of every incoming message is added to the route table if there is enough room in the buckets. The contacts with a nodeId close to the local id are always added to the route table due to the splitting procedure. To effectively put off a routing attack, the attacker must inject bad routing information in the target node by sending him messages that report sender ids near to the victim's id. Combined usage of $AuthId$ and $Auth$ makes the communication between nodes authenticated, so the attacker can inject only its own contact into the target route table, and because the ids are randomly chosen by the CS, the attacker cannot generate its id "ad hoc". Routing attacks (including eclipse) are unfeasible. Moreover, it is unfeasible for an attacker to hide a content marked with a given key k by way of a node insertion attack, because the malicious node cannot register a substantial number of nodes with IDs close to k: in fact, he cannot control id generation by his own.

Kademlia's lookup vulnerability is corrected by authenticated message exchange and random id generation. If the malicious FIND-NODE RPC receiver responds with a set of references to invalid nodes (i.e., devoid of $AuthId$s), the victim node is not able to contact any of them because the authentication protocol fails in signature verification. If the attacker responds with a set of valid colluding nodes, its attack results ineffective because the colluders' ids are scattered along the keyspace, so the lookup procedure proceeds properly.

Sybil attack Every user can have multiple identities (e.g., many email addresses), so a user can bind each of his identities to a different node by sending many $NodeIdRequest$ to the CS, and then he can run all those nodes on the same machine. So the Sybil attack is not completely wiped out with this scheme. Nevertheless, each node corresponds to a different user account and the node initialization requires a verification procedure for that account. If the user authentication procedure requires a human interaction it would be difficult for an attacker to create many different nodes in an automated way, actually lowering the risk of Sybil Attacks. For this reason, we strongly suggest to adopt OpenId verification methods, that redirect the user agent to an identity provider, and that returns to the CS when the submitted identity has been correctly authenticated.

Storage attacks Every storage entry in the DHT is bound with its $Cred$, created by the content owner with an unforgeable signature. A node performing a lookup operation returns to the application only those results that are bound with some $Cred$, and that has been previously verified. Therefore, the consumer application (or the human user himself) can interact with a reputation system to reward or penalize the owner of the consumed object depending on the quality of the content. The underlying node can be then instructed to exclude from network traffic those nodes whose reputation is too bad. The use of $Cred$ can contrast attacks like index poisoning, content pollution or even DDoS attacks based on redirection by punishing the malicious users who attempt these attacks.

Man in the middle attack An attacker who's able to intercept and alter the messages flowing between two

nodes has no way to act as one of the endpoints or to fool correct nodes into accepting forged messages. $AuthId$ and $Auth$ cannot be modified since they are signed, and the RPC cannot be altered or replaced because the Authenticator contains the RPC hash code. An attacker cannot effectively replay an intercepted $Auth$ because it includes a nonce which validity is limited to a node interaction session; moreover authenticators are addressee-specific, because they include the recipient node ID. Finally, the nonce based two-way authentication scheme grants protection against common interleaving attacks as Oracle session attacks, parallel attacks and offset attacks.

6 Prototype

In this Section we describe the main functionalities of the proposed application, according to the system architecture depicted in Figure 3. In the implemented prototype the main application interacts with the user components described in Section 3. As already mentioned above, we assume that the content indexed and retrieved in the P2P network is always governed, requiring the adoption of an appropriate format which is able to provide a full description of the content. We used the MPEG-7 metadata for the multimedia content representation and MPEG-21 metadata to express the digital rights. Moreover we also consider protected contents, obtained by applying encryption tools for DRM. According to Chillout reference implementation [5], we make use of the DCI and DCF formats, already discussed in Section 2. The user is able to search for a subset of relevant MPEG-7 and MPEG-21 metadata extracted from DCF and at the same time the whole DCF can be accessed in a completely distributed way, by means of a *DCFMetadataService* protocol provided by the *Transport* component (see Figure 3), implemented for building the mapping between the subset of indexed keys, inserted in a structured way, and the DCF it refers to.

We built our system upon Kademlia but, exploiting the separation level between the DHT implementation and the applications, the framework provides the possibility of using other DHTs. This level exports to upper modules the insertion of new mappings and the retrieval of the *key's root* functionalities. In order to communicate with the DHT module (see Section 2.1),we made use of Java *Future* objects for non-blocking asynchronous insertion and querying operations, that were introduced in the Java Development Kit from version 5[4]. We used a *CollectorParameter* design pattern in order to collect the results provided by the Future objects. The main reason is that the communication works asynchronously because it has to take into account the network latency and topology reconfiguration due to the peers joining and leaving the overlay. The *Transport* component is not put directly on top of the DHT (see Figure 3): they communicate in order to get the information about the two (or more, e.g., for multisource download) endpoints of the direct connection established for downloading the DCF. This module is composed by two subcomponents and the Figure 3 is showing the two socket listeners: one is responsible for transferring the resources (multimedia files) and the other is responsible for transferring the related metadata, actually a Java object which wraps the DCF file.
The *Application* component exports a set of high level functionalities in order to join/leave the system and to insert/retrieve the DCF files. As already described, it makes use of a DCI and DCF wrapper for parsing the digital content files and for extracting the metadata to be indexed.
The insertion of a content proceeds as follows: the *Content Creator* component is responsible for creating the DCI and the DCF. The user can choose one or more resources to be published in the P2P network in a single DCF file and can associate to each resource a different license, which can be completely customized for different purposes. It is worthwhile noticing that some resources in the DCF file could be also encrypted to ensure that even if they are retrieved from the P2P network the consumption of the content is possible only to the principal specified in the license. Once the DCF (or simply the DCI) is created, it can be shared on the structured peer-to-peer network. Concerning the content retrieval, the lookup operation on the DHT could be done by simple keywords or structured bundle of MPEG-21 REL tags, resulting, at low level, in the index of the content whose DCF (DCI) is fulfilling the request. The Application module contacts then the publishing sources (peers) asking for more information about the content. Every user can check the license conditions associated to a given content before downloading it. The *Transport* component communicates by means of a *DCFMetadataService* on a separate channel (socket), with a specific protocol which is able to exchange the wrapper of the DCF. In this way we can provide to the user all the available metadata related to the searched keywords and grants. The results of the query are collected by means of the *CollectorResults*, which generates a separate thread looking for the asynchronous return messages. The user can select the specific content from the result list and the *Application* component will contact the specific owner source (the <IP address,port> pair), through the *FileTransportation* component (see Figure 3) which communicates using a separate channel (socket), with a specific protocol for exchanging files. Our first approach has been the adoption of a simple file transportation but a possible improvement could come from making use of more sophisticated solutions enabling the multi-source download, as the BitTorrent [40] exchange protocol.

7 Conclusions and future works

We have described a decentralized, distributed and secure communication infrastructure for indexing and retrieving multimedia contents with associated digital rights. We have discussed a feasible approach to share digital items accord-

[4]http://java.sun.com. Last visited: 15 Nov 2008

ing to the associated license, making use of a P2P routing infrastructure based on DHT. Complex queries on standard MPEG-7 and MPEG-21 multimedia metadata are supported.

Concerning the future works, in order to express queries in a standard format, we will evaluate the use of MPEG Query Format (MPQF) [41], part 12 of the MPEG-7 standard, whose reference software implementation is currently under development by people involved in the MPEG consortium. MPQF lets us also investigate novel approaches for searching digital contents on peer-to-peer infrastructure, as *range* and *by feature* queries that could be introduced into a future prototype. Moreover, we plan to evaluate the Java implementation upon a live large-scale testbed like PlanetLab[5] in order to test the efficiency, scalability and reliability properties in a real scenario.

8 Acknowledgements

This work was partially supported by the SAPIR project[6], funded by the European Commission under IST FP6 (Contract no. 45128) and by the Ministero Italiano per l'Università e la Ricerca (MIUR) within the framework of the PRIN "PROFILES" project[7].

References

[1] MPEG-7 - ISO/IEC 15938 - Information Technology Multimedia Content Description Interfaces. `http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm`. Last visited: 15 Nov 2008.

[2] MPEG-21 - ISO/IEC 21000 - Information Technology Multimedia Framework. `http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm`. Last visited: 15 Nov 2008.

[3] MPEG-21 Rights Expression Language - ISO/IEC 21000-5 - Information Technology Multimedia Framework. `http://www.chiariglione.org/mpeg/technologies/mp21-rel/index.htm`. Last visited: 15 Nov 2008.

[4] Digital Media Project (DMP). `http://www.dmpf.org`. Last visited: 15 Nov 2008.

[5] Chillout. The Reference Software for DMP Interoperable DRM Platform. `http://chillout.dmpf.org`. Last visited: 15 Nov 2008.

[6] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of the ACM SIGCOMM '01 Conference*, San Diego, California, August 2001.

[7] P. Maymounkov and D. Mazieres. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In *IPTPS '02: Proceedings of 1st International Workshop on Peer-to-Peer Systems, Cambridge, MA, USA*, 2002.

[8] A. Rowstron and P. Druschel. Pastry: Scalable, Decentralized Object Location and Routing for Large-scale Peer-to-Peer Systems. In *Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms (Middleware), Heidelberg, Germany*, pages 329–350, 2001.

[9] Klaus Wehrle, Stefan Götz, and Simon Rieche. Distributed Hash Tables. In *Peer-to-Peer Systems and Applications*, pages 79–93, 2005.

[10] J. Liang, N. Naoumov, and K. W. Ross. The index poisoning attack in p2p file sharing systems. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, pages 1–12, 2006.

[11] D. Dumitriu, E. Knightly, A. Kuzmanovic, I. Stoica, and W. Zwaenepoel. Denial-of-service resilience in peer-to-peer file sharing systems. *SIGMETRICS Perform. Eval. Rev.*, 33(1):38–49, 2005.

[12] Moritz Steiner, Taoufik En-Najjary, and Ernst W. Biersack. A global view of kad. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 117–122, New York, NY, USA, 2007. ACM.

[13] Nitesh Saxena, Gene Tsudik, and Jeong Hyun Yi. Access control in ad hoc groups. In *HOT-P2P '04: Proceedings of the 2004 International Workshop on Hot Topics in Peer-to-Peer Systems*, pages 2–7, Washington, DC, USA, 2004. IEEE Computer Society.

[14] Nitesh Saxena, Gene Tsudik, and Jeong Hyun Yi. Identity-based access control for ad hoc groups. In *Information Security and Cryptology (ICISC)*, pages 362–379, 2004.

[15] John R. Douceur. The sybil attack. In *IPTPS '01: Revised Papers from the First International Workshop on Peer-to-Peer Systems*, pages 251–260, London, UK, 2002. Springer-Verlag.

[16] Moritz Steiner, Taoufik En-Najjary, and Ernst W. Biersack. Analyzing peer behavior in KAD. Technical Report EURECOM+2358, Institut Eurecom, France, Oct 2007.

[5] http://www.planet-lab.org. Last visited: 15 Nov 2008
[6] http://www.sapir.eu. Last visited: 15 Nov 2008
[7] http://dit.unitn.it/profiles. Last visited: 15 Nov 2008

[17] Moritz Steiner, Taoufik En-Najjary, and Ernst W. Biersack. Exploiting kad: possible uses and misuses. *SIGCOMM Comput. Commun. Rev.*, 37(5):65–70, 2007.

[18] George Danezis, Chris Lesniewski-Laas, Frans M. Kaashoek, and Ross Anderson. Sybil-resistant dht routing. In *ESORICS*, volume 3679 of *LNCS*, pages 305–318. Springer, 2005.

[19] Hosam Rowaihy, William Enck, Patrick Mcdaniel, and Thomas La-Porta. Limiting sybil attacks in structured peer-to-peer networks. Technical Report NAS-TR-0017-2005, Network and Security Research Center, Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA, 2005.

[20] A. Singh, T. W. Ngan, P. Druschel, and D. S. Wallach. Eclipse attacks on overlay networks: Threats and defenses. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, pages 1–12, 2006.

[21] I. Baumgart and S. Mies. S/Kademlia: A Practicable Approach Towards Secure Key-Based Routing. In *Proc. of P2P-NVE 2007 in conjunction with ICPADS 2007, Hsinchu, Taiwan*, volume 2, December 2007.

[22] Petar Maymounkov and David Mazières. Kademlia: A peer-to-peer information system based on the XOR metric. In *IPTPS*, pages 53–65, 2002.

[23] Tyson Condie, Varun Kacholia, Sriram Sankararaman, Joseph M. Hellerstein, and Petros Maniatis. Induced churn as shelter from routingtable poisoning. In *In Proc. 13th Annual Network and Distributed System Security Symposium (NDSS)*, 2006.

[24] Antony Rowstron and Peter Druschel. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In *IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)*, pages 329–350, November 2001.

[25] Miguel Castro, Peter Druschel, Ayalvadi Ganesh, Antony Rowstron, and Dan S. Wallach. Secure routing for structured peer-to-peer overlay networks. *SIGOPS Oper. Syst. Rev.*, 36(SI):299–314, 2002.

[26] Yongdae Kim, Daniele Mazzocchi, and Gene Tsudik. Admission control in peer groups. In *NCA '03: Proceedings of the Second IEEE International Symposium on Network Computing and Applications*, page 131, Washington, DC, USA, 2003. IEEE Computer Society.

[27] Fotis G. Kazasis, Nektarios Moumoutzis, Nikos Pappas, Anastasia Karanastasi, and Stavros Christodoulakis. Designing Ubiquitous Personalized TV-Anytime Services. In *CAiSE '03: Proceedings of the 15th Conference on Advanced Information Systems Engineering, Klagenfurt/Velden, Austria*. CEUR-WS.org, 2003.

[28] Rubén Tous and Jaime Delgado. L7, An MPEG-7 Query Framework. In *AXMEDIS '07: Proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution, Barcelona, Spain*, pages 256–263. IEEE Computer Society, 2007.

[29] Walter Allasia, Francesco Gallo, Filippo Chiariglione, and Fabrizio Falchi. An Innovative Approach for Indexing and Searching Digital Rights. In *AXMEDIS '07: Proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution, Barcelona, Spain*, pages 147–154. IEEE Computer Society, 2007.

[30] CreativeCommons. `http://creativecommons.org`. Last visited: 15 Nov 2008.

[31] MPEG-A - ISO/IEC 23000-5 - Information Technology Multimedia Application Format, Part 5: Media Streaming Application Format. `http://www.chiariglione.org/mpeg/standards/mpeg-a/mpeg-a.htm`. Last visited: 15 Nov 2008.

[32] Digital Media Project (DMP). Approved Document No. 3 - Technical Specification: Interoperable DRM Platform, Version 3.0 - 1003/GA15. `http://www.dmpf.org/open/dmp1003.zip`. Last visited: 15 Nov 2008.

[33] MPEG-21 Rights Expression Language - ISO/IEC 21000-5 Amendment 3 : the OAC (Open Access Content) profile.

[34] MPEG-21 Rights Expression Language - ISO/IEC 21000-5 Amendment 2: the DAC (Dissemination And Capture) profile.

[35] TV-Anytime. `http://www.tv-anytime.org`. Last visited: 15 Nov 2008.

[36] Unified Modeling Language. UML 2.1.1 - UML 1.4.2 (ISO/IEC 19501). `http://www.uml.org`. Last visited on Nov 15th 2008.

[37] W. Allasia, F. Gallo, M. Milanesio, and R. Schifanella. Governed content distribution on dht based networks. *Internet and Web Applications and Services, 2008. ICIW '08. Third International Conference on*, pages 391–396, June 2008.

[38] Marco Milanesio, Giancarlo Ruffo, and Rossano Schifanella. A Totally Distributed Iterative Scheme for Web Services Addressing and Discovery. In

PDCS '07: Proceedings of the 19th IASTED International Conference on Parallel and Distributed Computing Systems, Cambridge, MA, USA, pages 85–90. ACTA Press, 2007.

[39] L. M. Aiello, M. Milanesio, G. Ruffo, and R. Schifanella. Tempering kademlia with a robust identity based system. In *Proc. of 8th International Conference on Peer-to-Peer Computing 2008 (P2P'08)*, 2008.

[40] BitTorrent. `http://www.bittorrent.com`. Last visited: 15 Nov 2008.

[41] MPEG Query Format. `http://dmag.upf.edu/mpqf`. Last visited: 15 Nov 2008.

Multi-Modal Emotional Database: AvID

Rok Gajšek, Vitomir Štruc and France Mihelič
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, SI-1000 Ljubljana, Slovenia
E-mail: {rok.gajsek, vitomir.struc, france.mihelic}@fe.uni-lj.si, http://luks.fe.uni-lj.si/en/index.html

Anja Podlesek, Luka Komidar, Gregor Sočan and Boštjan Bajec
Psychological Methodology
Faculty of Arts
University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
E-mail: {anja.podlesek, luka.komidar, gregor.socan, bostjan.bajec}@ff.uni-lj.si

Keywords: multimodal database, speech recognition, emotion recognition

Received: November 5, 2008

This paper presents our work on recording a multi-modal database containing emotional audio and video recordings. In designing the recording strategies a special attention was payed to gather data involving spontaneous emotions and therefore obtain a more realistic training and testing conditions for experiments. With specially planned scenarios including playing computer games and conducting an adaptive intelligence test different levels of arousal were induced. This will enable us to both detect different emotional states as well as experiment in speaker identification/verification of people involved in communications. So far the multi-modal database has been recorded and basic evaluation of the data was processed.

Povzetek: V članku je predstavljeno snemanje večmodalne zbirke, ki vsebuje audio in video posnetke različnih čustvenih stanj.

1 Introduction

This study paper describes initial attempts to collect a multimodal emotional speech database as a part of our research under the ongoing interdisciplinary project *"AvID: Audio-visual speaker identification and emotion detection for secure communications"*. The goal of the project is to use speech and image technologies in video telecommunication systems for identification/verification and detection of the emotional state of persons involved in communication. Such a system should provide additional information about the identity and psychophysical condition displayed on the communication devices enabling more secure and credible exchange of information. Project partners are from the Faculty of Electrical Engineering – Department for Automatics from the University of Ljubljana and Jožef Stefan Institute – Department for Intelligent Systems from Ljubljana, the Faculty of Arts – Department for Psychology from the University of Ljubljana and the industrial R&D company Alpineon D.O.O. from Ljubljana.

Unfortunately most of the available speech databases with emotional speech were obtained by recording of acted different type of emotions usually from professional actors [5, 6, 11, 2, 3, 7][1] and therefore do not represent very adequately data for training and testing procedures for speaker psychophysical condition detection in real environment. To distinguish between normal and non-normal condition, and to compare speaker verification performances we also need quite a lot of speech material with normal speech, that is also usually not the case for available databases. Although we plan to perform our experiments on as much available data as possible, we also intend to obtain some data providing more realistic conditions for our task. Therefore we are planing to collect reasonable amount of audio and video recordings of spontaneous speech in normal (relaxed) and non-normal psychophysical conditions (the conditions of excitement and arousal with different valence, both positive and negative) from a representative group of speakers. Initial strategies to obtain desired speech corpora and recording setup along with the statistics of already recorded data are described in the following sections.

2 Recording strategies

In the beginning, each participant was told that the main purpose of the experiments was to examine whether dif-

[1] There are of course some exceptions as, for example, the German SmartKom database [12] which was recorded using a Wizard of Oz technique and tried to evoke different emotions in the participants.

ferent measures of his/her state could be used in an adaptive test of intelligence. The biometric measures to be indicative of his/her psychophysical state at different moments were: psychophysiological response (the electrodermal and electrocardiographic response), verbal response, and facial expression. After a written consent to participate in the study was obtained from each individual, sensors were placed on the index and the middle finger on the left hand and the audio and visual recording started. The participant was instructed to speak loudly enough and not to move. With the right hand he/she had to hold a computer mouse in order to prevent the hand from excessive movement.

To obtain recordings of speech in both neutral and changed psychophysiological state of the participant, we designed an experiment composed of four parts. In Part I, after the participant introduced himself/herself with a few words (stated the name, the place of living, age, and main occupations), photographs with neutral content were presented on the screen. The participant was instructed to describe each photograph in detail, as if he/she were describing what he/she sees to a blind person. In this part we supposedly measured his/her verbal fluency. When he/she finished with the descriptions, he/she instructed the experimenter to continue with the presentation of the next photograph.

Before the start of Part II, the participant was told that we will be assessing the efficiency of his/her verbal instructions given to a teammate in order to achieve a common and specific goal. We explained to the participant that the team, which will involve himself/herself and the experimenter, will play a computer game (Tetris) and that he/she will observe the progression of the game on the computer monitor and will be giving verbal instructions, whereas the experimenter will not be able to observe the game and will carry out his/her orders by pressing the appropriate buttons on the keyboard. If the participant had no prior experience with the game, we explained the rules of Tetris and let him/her play for a few minutes. The participant who observed the ongoing game on the screen had to lead the experimenter through the game by uttering the following four commands: Left ('Levo' in Slovene), Right ('Desno'), Around ('Okrog'), and Down ('Dol'). The goal of the team was to achieve the highest score possible. Passive commands (e.g. 'Around' instead of 'Turn it around') were chosen in order to be suitable for use also in Part III of the experiment. At the end of the game the participant had to tell the experimenter what score they had achieved, what happened during the game, and why the game ended.

The ongoing game was recorded by CamStudio screen capture program. In Part III the recorded movie of the game was played and the participant had to describe what was happening on the screen by using the same four commands as in Part II. At the end the same description of the events on the screen had to be given as at the end of Part II. The aim of Part II was to obtain positive arousal (joy, satisfaction) as well as negative arousal (frustration, anger), whereas Part III was carried out to obtain exactly the same utterances in a relaxed, non-aroused state, because in Part III the participant was just a passive observer.

At the beginning of Part IV, the participant was told that he/she will be given an adaptive intelligence test where the difficulty of the task will be chosen by the computer according to (i) the correctness of the answer in the previous task, (ii) the mental strategy used for solving the task, and (iii) the biometric measures (EDR and heart rate). We explained to her that several values will be presented on the left part of the screen: the momentary IQ value, the arrows pointing upwards when the IQ estimate was increasing and downwards when it was decreasing, the momentary values of EDR and EKG measures, and the time that remained for solving the current task. On the right part of the screen, matrices with different figures or symbols were presented with one element absent. The participant had to reason aloud about the principles of the arrangement of matrix elements in rows and columns and find the proper solution among five to six possible answers. The participants believed they had to reason aloud so that the experimenter will be able to assess their mental strategy used for solving the task. After the experimenter showed two examples of matrices and explained how the reasoning should be verbalised, 20 matrices had to be solved, some of which were very difficult or did not have a known solution. If the participant ceased to speak aloud, the experimenter encouraged her to verbalise her thoughts. After the solution was found, the experimenter clicked some buttons to input the chosen solution and the presumed category of mental strategy. He could choose among six options with which he controlled the changes in the unfounded IQ estimate. He raised the IQ value only when the correctness of the solution was obvious, the reasoning was straightforward and solution was derived quickly. In other cases the IQ value was decreased. The main purpose of decreasing the IQ value was to increase the participant's subjective stress level. Besides decreasing the temporary IQ value, the experimenter could also manipulate participant's stress level by increasing the EDR and heart rate values. In order to attract attention to the EDR and heart rate indicators during the test, the values changed its colour from black to red when a certain value was exceeded.

After Part IV was over we debriefed the participant. The experimenter explained that the temporary and final IQ scores were not valid estimates of her intelligence and that the real aim of the study was to obtain the recordings of speech in the normal, relaxed state and in the aroused, stress-induced emotional state. The participants were then asked to describe (freely) their feelings, thoughts, and involvement in each part of the experiment. In the end, some general data on participants and their speech characteristics were gathered (see Table 1).

Subject	Sex	Age	Voice type	Health	Smoking	Overall mood	Dialect	Speech peculiarities
01	M	20	baritone	normal	NO	slightly tense	Central	
02	F	37	mezzo-sop.	cold	casual	relaxed	Eastern	
03	F	19	mezzo-sop.	normal	NO	relaxed	Littoral	
04	F	25	mezzo-sop.	normal	NO	relaxed	Eastern	
05	F	21	mezzo-sop.	normal	NO	relaxed	Central	
06	F	26	mezzo-sop.	normal	YES	NA	Central	
07	M	21	bass	normal	NO	relaxed	Central	rash speech
08	F	19	soprano	normal	stopped	distracted	Eastern	
09	F	20	mezzo-sop.	normal	NO	distracted	Littoral	
10	M	28	tenor	normal	NO	relaxed	Central	
11	F	26	mezzo-sop.	normal	NO	relaxed	Eastern	
12	F	20	mezzo-sop.	normal	NO	relaxed	Lower Car.	
13	F	27	mezzo-sop.	normal	NO	relaxed	Eastern	
14	F	20	mezzo-sop.	normal	YES	relaxed	Eastern	
15	F	27	soprano	normal	NO	relaxed	Central	

Table 1: Basic participants' data relevant for recorded speech analysis.

3 Recording conditions and inventory

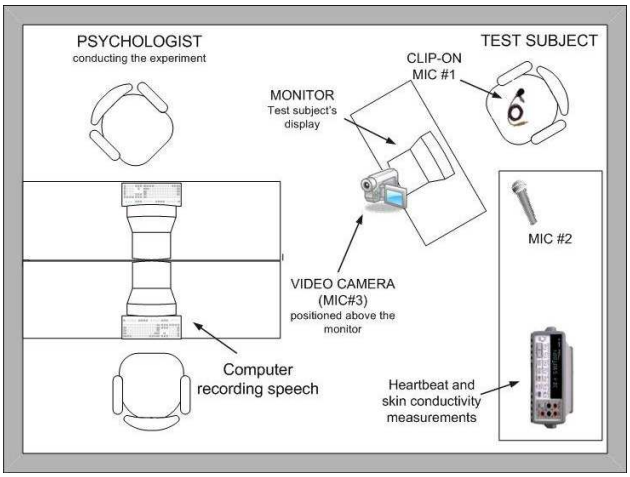


Figure 1: An outline of the recording setup.

Recordings were done in a closed room using a digital video camera and three microphones [10]. We used several microphones as in our previous GOPOLIS database [10] to enable some environment and channel normalisation tests. For the reference also some physiological sensors were used to detect changes in the heart beet rate and skin conductivity during the tests on few test objects.

Each participant was asked to position himself/herself in front of the computer monitor shown in the upper right corner of Fig. 1. Behind the monitor a digital camera mounted on a tripod was placed to capture the video recording (i.e., video as well as one channel of audio data). To ensure an appropriate quality of the captured video data the participant was seated in front of a relatively homogeneous, white background and a light source was directed towards the participant's face. This setup resulted in the recorded video sequences showing a fairly "clean", i.e., without to much shadows, frontal view of the participant's face - see Fig. 2 where the recording setup is presented.

Note that the quality of the recorded video data could have been further improved by using additional light sources or diffused light, however, as our goal was to collect a realistic database the employed setup fully sufficed for our requirements.

For capturing the audio signal two microphones were used in addition to the one integrated in the digital camera. The first, denoted as MIC #1 in Fig. 1, was attached on the participant's clothing near the chest, while the second, denoted as MIC #2 in Fig. 1, was positioned on the nearby desk. Both microphones were hooked to a computer which was used for recording and storing of the audio data.

Two people were supervising the acquisition of the database: (i) a psychologist who was in charge of the recording session and tried to induce a "non-normal" psychophysical condition in the participant by applying the strategies presented in Section 2 and (ii) a technician who overlooked the technical aspects of the acquisition process.

3.1 Video data collection

The video part of the AvID emotion database was acquired using a high-definition Sony HDR-SR11E digital handycam which captures video at a resolution of 1920×1080 pixels and a bit-rate of 16Mb/s. The video data was recorded at a frame aspect ratio of $16 : 9$ and later on archived in the AVCHD (Advanced Video Codec High Definition) format[2]. Specifications of the employed format

[2]A high-definition format jointly established by Panasonic and the Sony Corporation.

Figure 2: The recording setup.

Figure 3: A sample frame from the AvID audio–video emotional database.

for the video data can be found in Table 2.

Video signal	$1080/50i$
Pixels	1920×1080
Aspect ratio	$16:9$
Compression	MPEG4 AVC/H.264
Luminance sampling frequency	74.25 MHz
Chroma sampling format	4:2:0
Quantization	8-bit

Table 2: Specifications of the employed AVCHD format.

A high-definition camera was chosen for capturing the video data of the database for several reasons: (i) different kinds of experiments (e.g., biometric verification or identification, emotional state recognition, lip reading, etc.) can be performed on high quality video, (ii) with simple image- and video-processing techniques the quality of the video data can easily be degraded and research can be conducted on lower-quality video, and (iii) as high-definition technology is spreading with an increasing speed, it will soon find its way into peoples daily lives; with its widespread deployment the technology will also become easily affordable and, therefore, suitable for employment in low- (or medium-) cost recognition systems. A sample frame captured with Sony's HD camera in the AVCHD format is shown in Fig. 3.

3.2 Audio data collection

As mentioned above the audio signal was captured using three different microphones. Channel number one was recorded using a Sennheiser ew122-p G2 system with a clip-on microphone which transmitted the signal to the recording computer via radio waves. The microphone was pinned to the speakers chest roughly $10-20$cm away from the speaker's mouth.

The second channel was captured with a Shure PG81 microphone which was positioned approximately $30-40$ cm away from the speaker. Both microphones specify a frequency range of $40-18000$ kHz. Channels were recorded at a sampling rate of 16 kHz and 16-bit linear encoding.

The third channel was acquired from the employed video camera's built in microphones that record in Dolby Digital 5.1 and use AC-3 compression for audio storage.

4 Database description

A total of 15 native Slovenian speakers (12 female and 3 male) were recorded with one session lasting approximately an hour. After extracting only the speaker's speech from the session we got roughly a half an hour of usable audio per speaker. For the video part of the database one continuous recording was captured for each of the participants resulting in over 15 hours of high-definition video. As already mentioned in the previous section the participants were recorded in front of a white background and with frontal illumination. The average inter-ocular distance, which is the traditional measure of the size of the face in an image or video frame, is more than 150 pixels. Similar databases (uni- or multi-modal) used either for assessing biometric identification/verification or emotion recognition algorithms, such as the XM2VTS [9], the Cohn-Kanade [4] and the eNTERFACE'05 [8] databases, typically feature face images (or video sequences) with a distance of $40-60$ pixels between the left and the right eye. The AvID database is therefore suitable as the foundation for the development of recognition algorithms that make use of high resolution information.

The audio recordings were later split to shorter utterances - approximately one utterance per sentence. Transcriptions and labels describing emotional state of the speaker were made using Transcriber tool [1] and followed the LDC broadcast speech transcription conventions[3].

[3]LDC broadcast speech transcription conventions: http://projects.ldc.upenn.edu/Corpus_Cookbook/ transcription/

5 Evaluation results

Subjective reports were analysed - descriptions of the states at the beginning of the experiment and within each part of the experiment were classified into five categories (see Figure 1) where possible. Where the category is composed of two arousal levels (e.g., moderately and highly aroused), the first one reflects the prevalent state and the second reflects temporary peaks of slightly elevated stress level. Mostly the participants reported of the relaxed or slightly tense state prior to the experiment (when the sensors were attached and the procedure was explained to them). In Part I, when describing photographs, most of them reported no tension, and some reported a slight tension that later vanished. Their arousal was increased slightly while playing Tetris. Some participants reported negative emotions (e.g., irritation) in Part III due to the inability to take over the control. The majority reported that Part IV was difficult and stressful because they had troubles with verbalising their reasoning and were worried and puzzled about the calculated IQ value. This was reflected in a notable decrease of speech loudness. The change towards higher arousal during the experiment is indicated with the prevalence of darker pattern in Figure 4, whereas the transparent patterns represent a more relaxed state. It may be concluded that the experimental situations elicited the presumed levels of arousal: neutral emotional state with describing photographs and events, and arousal in playing an exciting computer game and in the situation where an individual wants to perform well under social pressure.

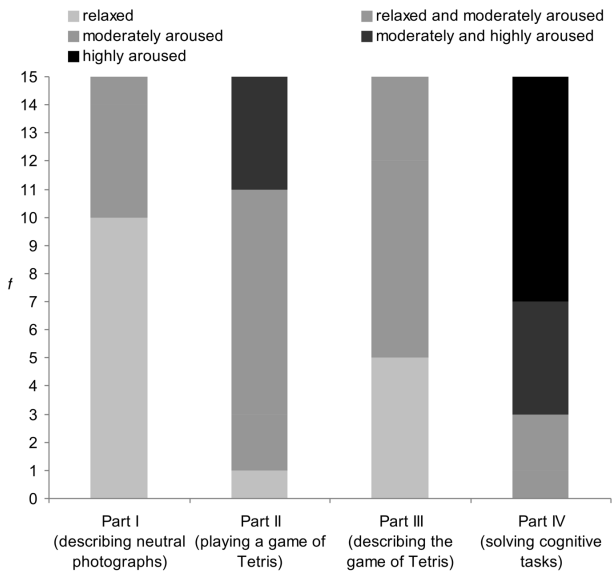


Figure 4: Levels of arousal during different parts of the experiment based on the participants' subjective report.

5.1 Future prospects

In future studies, a triangulation of different methods will be used to assess the arousal and the emotional state of the participants more systematically and objectively. To obtain additional indicators of the participant's emotional state we will use: (i) standardised instruments of subjective emotional experience, (ii) the psychophysiological measures, such as EDR and EKG, and (iii) behavioural expressions.

6 Conclusion

The goal of recording a multi-modal speech database containing different spontaneous emotions was achieved. Due to well selected experiments different levels of arousal were induced and measured by different biometric parameters: facial expression - video, verbal response - audio and psychophysical response - electrodermal and electrocardiographic response. Video and audio comprise the database, where psychophysical measures are only used to provide an objective information about the level of arousal.

Enough data was collected (which is especially important for speech research) to form a bases for future studies on speaker identification/verification, emotion recognition and spontaneous speech analysis research.

Acknowledgement

This work was supported by the Slovenian Research Agency (ARRS), development project M2-0210 (C) entitled "AvID: Audiovisual speaker identification and emotion detection for secure communications."

References

[1] C. Barras, E. Geoffrois, Z. Wu and M. Liberman (2001) Transcriber: development and use of a tool for assisting speech corpora production, *Speech Communication*, pp. 5–22.

[2] A. Battocchi and F. Pianesi (2004) DAFEX: Un Database Di Espressioni Facciali Dinamiche, *Proceedings of the SLI-GSCP Workshop "Comunicazione Parlata e Manifestazione delle Emozioni"*, pp. 1–11.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss (2005) A Database of German Emotional Speech, *Proceedings Interspeech 2005*, pp. 1–4.

[4] T. Kanade, J.F. Cohn and Y. Tian (2000) Comprehansive database for facial expression analysis, *Proceedings of the 4th AFGR'00*, pp. 46-53.

[5] LDC (1999) SUSAS (Speech Under Simulated and Actual Stress), *Proceedings of the 4th AFGR'00*, Language Data Consortium,

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78.

[6] LDC (2002) Emotional Prosody Speech and Transcripts, Language Data Consortium, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28.

[7] O. Martin, I. Kotsia, B. Macq and I. Pitas (2006) The eNTERFACEŠ05 Audio-Visual Emotion Database, *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp. 1–8.

[8] O. Martin, I. Kotsia, B. Macq and I. Pitas (2006) The eNTERFACE'05 Audio-Visual Emotion Database, *Proceedings of the 22nd International Conference on Data Engenireeng Workshops*, IEEE Computer Society.

[9] K. Messer , J. Matas, J. Kittler, J. Luettin and G. Maitre (1999) XM2VTSDB: the extended M2VTS database , *Proceedings of AVBPA'99*, pp. 72–77.

[10] F. Mihelič, J. Žganec Gros, S. Dobrišek, J. Žibert and N. Pavešić (2003) Spoken language resources at LUKS of the University of Ljubljana, *Int. J. Speech Technology*, pp. 221–232.

[11] V. Hozjan, Z. Kačič and B. Horvat (2001) Prosody feature analysis for emotion modeling, *Electrotechnical Review*, pp. 213–218.

[12] U. Turk, (2001) The Technical Processing in SmartKom Data Collection: a Case Study, *Proceedings of Eurospeech*, pp. 1541–1544.

Efficient Morphological Parsing with a Weighted Finite State Transducer

Damir Ćavar
University of Zadar, Croatia
E-mail: dcavar@unizd.hr and http://personal.unizd.hr/~dcavar/

Ivo-Pavao Jazbec and Siniša Runjaić
Institute of Croatian Language and Linguistics, Croatia
E-mail: {ipjazbec,srunjaic}@ihjj.hr and www.ihjj.hr

Keywords: weighted finite state transducer, morphological analysis, croatian

Received: October 31, 2008

This article describes a highly optimized algorithm and implementation of a deterministic weighted finite state transducer for morphological analysis. We show how various functionalities can be integrated into one machine, without sacrificing performance or flexibility, and and still maintaining applicability to various languages. The annotation schema used in this implementation maximizes interoperability and compatibility by using a direct mapping of tags from the GOLD ontology of linguistic concepts and features, providing possible extended processing scenarios.

Povzetek: Opisana je morfološka analiza za hrvaški jezik.

1 Introduction

For the majority of natural languages, the s.c. low density languages, appropriate linguistic data and language processing tools do not exist, neither enough raw language data (e.g. text or audio recordings). For some languages with much higher language resource density the appropriate language technology is missing that would help in creating necessary and valuable quantitative and qualitative linguistic information, essential not just for research purposes. Thus, even languages that do not face the low density problem, still lack crucial resources. For many languages information as for example contained in CELEX [7] is not available. Thus, for the majority of languages the distributional and quantitative models of phonetic, phonemic, morphological and syntactic properties do not exist.

In recent years the amount of available linguistic data was growing. Recordings and transcriptions, dictionaries and textual corpora build the basis for an impressive amount of empirical linguistic research, as well as language technology for various application domains. However, the resources face crucial problems. On the one hand, we see a growing number of specific purpose data, with limited value for a wide range of research and development domains, representing snapshots of a specific state of a language at a specific time, often based on easily available raw data like newspapers or books. Language change and dynamic aspects of languages require permanent creation and adaptation of existing resources. This task cannot be accomplished without the technologies, e.g. adaptive tools or appropriate machine learning algorithms.

Linguistic annotation of corpora is limited in another way. The choice of part-of-speech (PoS) tags is theory driven, and thus in general restricted to a specific view or framework, with a likely limited value for other research and development purposes.

Focusing on morphology as one of the levels of linguistic representation and grammar, corpus annotations tend to be lexeme and word-form oriented, PoS-tags for lexemes in the corpus, rather than segmentation of word-forms into morphemes and allomorphs with their particular feature annotation. Even the notion of *morphological information* is used inconsistently in the literature, e.g. associated exclusively with lexeme and PoS information. Thus existing resources, e.g. the documented Croatian morphological lexicon [14], do not provide information about the morphological structure and specific feature annotations of single morphemes, but rather word-forms and lexemes with PoS-annotation. Specific research questions, on the other hand, require detailed morphological analyses of lexical tokens in a corpus. On the basis of the Croatian Language Corpus [6], as one of our major data sources, needs to be annotated for subsequent analysis.

1.1 Central goals

Our specific goal was the development of a system that parses lexemes into morphological structure. The desired output information, as far as morphology is concerned, requires a morphological lexicon and morphological corpus annotation to include parsed lexemes on the morphological level, with annotations and explicit feature bundles associated with each single morpheme or allomorph, as shown in table 1 for the word *pročitamo* (Croatian, “to read(out)”).

Table 1: Example of a morphological parse

token	*pročitamo*		
parse	*pro*	*čita*	*mo*
	stem		inflectional
	prefix	root	suffix
	aspect	verb	1^{st}
	perfective	transitive	plural
			present

This type of a morphological parse is already simplified. Certain potentially required and theoretically motivated information is excluded. For example a hierarchical tree structure for morpheme relations is not displayed, although it might be useful to reveal scope ambiguities of semantic properties. The shown parse represents just linear segmentations that include a quasi-hierarchical dependency with for example the prefix and root being contained in the stem, as shown in table 1. In general, we expect ambiguities to occur, and in fact we are interested in all possible parses that lead to a complete analysis of a morphological complex word. For research purposes in the first stage we do not intend to disambiguate the parses.

In addition to the morphological parse, the output should ideally also contain information about the lexical lemma (i.e. base-form) and the related root lemma. Once the morphological segmentation is available, the generation of lemmata can be achieved by appending the canonical inflectional suffix to the identified base, and potentially applying the necessary allomorphic change to the root. Furthermore, for establishing associations of word-forms to semantic fields, i.e. identifying the semantic root of a complex word-form, the lemma of the root provides a useful additional annotation information. For most Slavic and Germanic languages the rightmost root in a word-form is the semantic head of a complex morpheme. Thus, the root-lemma is generated by picking the rightmost root morpheme and append to it the canonical inflectional suffix. We annotate individual word forms for both lemma types, i.e. the root- and the base-lemma. The latter is achieved by inclusion of all prefixes in the lemma formation rule that are part of the morphological base.

1.2 Scope and problems

The morphological parse and annotation is supposed to cope with raw language data from various time periods and dialects, i.e. synchronic and diachronic data, focusing for the time being on Croatian. Such textual data is problematic since e.g. different orthography standards have been and are still used. The lexical environment was and is not static, with lexical items emerging and disappearing, their semantic properties changing etc. Lexical changes occurred, some might have affected the morphological makeup of individual word-forms (including changes in paradigms), some might be related to different feature bundles associated with them.

Given these conditions, it is obvious that various domains of lexical and morphological properties and features in our particular case are still subject to ongoing research, the set of features is necessarily open and unspecified from the outset. We expect in particular semantic properties, new feature types that result from linguistic conceptual necessities, or marking of linguistic origin and cultural background to emerge during future studies, i.e. the annotations of morphemes should be extensible.

2 Technical realization

In general, the technical realization of the described annotator appears to be feasible, using a very simple, and nevertheless efficient technical solution, i.e. finite state transducers [3, 4]. Specific attention is given to efficiency and adaptability (to variants and dialects, as well as other languages). In the following we describe the algorithmic specification of a morphological parser for the Croatian standard, and synchronic and diachronic variants.

2.1 Previous approaches

Finite state methods for computational modeling of natural language morphology are wide-spread and well understood. Various commercial and open-source FSA-based development environments, libraries and tools exist for modeling of natural language morphology. A detailed discussion of their properties and application for various languages would be beyond the scope of this article. Some overview can be found in recent literature, e.g. [18, 2, 15], further links to literature and implementations can be found in the context of the OpenFst library [1].

For Croatian there are various descriptions of the formalization and computational modeling of morphology in terms of finite state methods [19, 12]. However, an implemented testable application is not available.

Some solutions that have been implemented for example for German come close to the system requirements specified above. The SMOR [17] and Morphisto [21] systems partially represent such a type of computational morphology application. An almost complete overlap of features and properties can be found in the implementation of the German morphology as described in the TAGH [11] system.

Common problems of some of the existing tools are the lack of efficient handling of ambiguities in natural languages, and in particular different code-pages for character sets. The common strategy is to replace input symbols in the machine compilation process with a fixed mapping to integer values. This implies that every input has to be transcoded using this mapping before analysis, which represents a performance penalty. Ambiguity is usually dealt with by using non-deterministic finite state models and relying on backtracking, or alternative strategies, which again comes with a performance penalty.

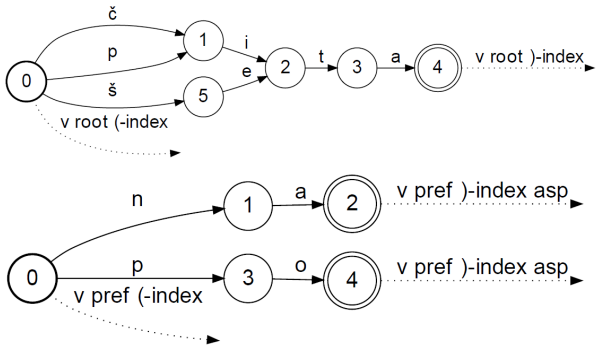
3 FST-based segmentation

While we decided to stick to the approach and implementation strategy of TAGH, we apply our own experimental libraries and development environment.[1] Following the TAGH-approach [11], we model Croatian morphology by referring exclusively to morphotactic regularities, using morpheme and allomorph sets and regular morphological rules, such that a deterministic finite state transducer (FST) can be generated.

For the compilation of a FST morpheme lists are required. In the initial modeling step morphemes are grouped on the bases of specific criteria. The main criteria are a. morphemes having the same feature specification, and b. being subject to the same morphological rules, where morphological rules are purely distributional and morphotactic, not derivational in the sense of e.g. lexical phonology [16].

Each morpheme group in the Croatian morphological parser represents one deterministic and acyclic finite state transducer (DFST), comparable to the Mealy [13] or Moore machine [5]. Every morpheme DFST emits on entry a tuple of the byte-offset in the input string, and the feature bundle that is associated with the DFSA path. In every final state the DFST emits the same tuple with a specific end-bit set. Thus morphemes are marked with a start and end index, as well as the corresponding feature bundle, representing the desired annotation. Morpheme analyses consist of a pair of emission tuples on a stack. Redundancies are avoided by limiting the placement on the stack to one occurrence only. Only complete analyses that span the complete input string are returned in the final output.

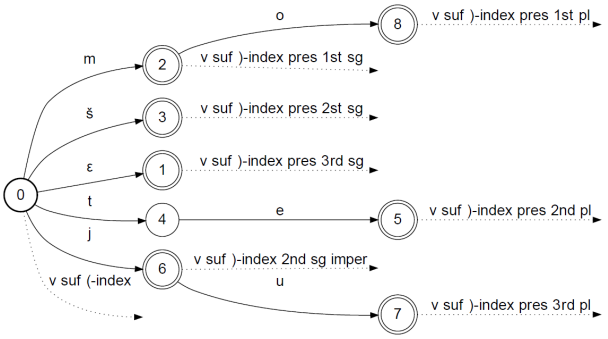
The following graph shows a simplified example of an acyclic DFST for verbal roots and for example aspectual prefixes:



To clarify the semantics of the emission arrows (dotted line), one should keep in mind that the initial emission tuple is placed on the stack only when the initial state is left, while the final emission tuple is placed there when the final state is reached.

All affixes are organized in the same way, e.g. the verbal inflectional paradigm, as in the following graph:

[1]The automata and grammar definitions we use are compatible with several existing systems and libraries.



Since the model is based on purely morphotactic distributional regularities, potential phonological phenomena are expressed using exclusively allomorphic variations, i.e. alternative sets of allomorphs. Consider for example the allomorphic variation in the case of *banka*(nominative singular of "the bank") with the dative/locative singular form *banci*. In this case the allomorph *banc* is grouped with all such allomorphs that occur in the context of the instrumental singular suffix *i*, while the root *bank* is grouped with all other nominal roots that have an allomorphic variant in this particular case. This way specific paradigms for such nouns can be defined, that lack the instrumental singular suffix for all such root morphemes, and combine with all the other case specific suffixes, and so on.

Once all morphemes are grouped into DFSTs, and the appropriate emission symbols (the annotations) are assigned to each entry and final state of the DFST, each morpheme group is assigned an arbitrary variable name, which is used in the definition of rules. A rule that makes use of the automata above could be defined as follows:

```
vAspectPref* . vAtiRoots . vInflSuf
```

This rule describes the concatenation of the DFST for the verbal aspectual prefixes, the verbal roots and the DFST for the verbal inflectional paradigm, using common regular expression notation. In this case we use the regular expression syntax as defined for the Ragel [20] state machine compiler. Additionally, the prefixes are defined as optional and potentially recursive prefixes concatenated with the verbal root DFST. This definition generates a cyclic[2] deterministic transducer.

Such a DFST emits a tuple containing the byte-offset and the corresponding annotation symbols at the initial state, and at each morpheme boundary (former initial and final states of the sub-DFSTs).

Using this approach, all lexical classes are defined as complex (potentially cyclic) DFSTs, and combined, together with the closed class items, as one monolithic DFST.

[2]Cyclicality in this particular case leads to more compact automata. In principle, the depth of recursion of such prefixes could be limited (empirically and formally), and formalized using the appropriate regular expression syntax. Independent of the theoretical question whether such type of recursion indeed exists, or is conceptually necessary, for the analyzer it is empirically irrelevant, and has no impact on its properties, except of size optimization.

The advantage of such a representation is not only that the resulting morphological representation is maximally compressed, but also that it is processed in linear time, with the identification of morpheme boundaries and corresponding feature bundles being restricted by contextual rules.

In order to cope with morphological ambiguity, this approach is extended. In principle there are two major approaches to deal with ambiguity, either one has to allow for non-deterministic automata (two different transitions with the same symbol sequence as input emit a different output tuple), or ambiguity is mapped on the emission of multiple annotation tuples. For Croatian, the latter option is used in the modeling. Every emission is a tuple of length 0 to n, such that e.g. orthographically ambiguous nominal suffixes like *a* (genitive singular or plural) are modeled as a single transition in a DFST with the final state emitting two annotation tuples that contain the specific case and number features.

3.1 Interoperability and annotation standard

Annotated language data plays an important role in various domains, be it language technology development, or linguistic research. Many different annotations were developed, for various purposes, with different goals in mind. Due to the diversity of encoding and annotation standards, current language resources face a problem related to issues of interoperability and annotation compatibility. The various different tag-sets that are used for different and particular languages tend not to be straight-forward compatible. In the same way, linguistic annotation tools do not necessarily make use of some standardized tag-set, and such a tag-set actually does not even exist. Annotation of language data, however, is an expensive task, as well as the change and adaptation of existing data to a new or specific annotation standard.

Thus, the question of annotation standards is crucial for the conceptualization and development of new language resources and language processing tools. In principle, two major options exist. Either a certain annotation standard is promoted and agreed upon, or a specific standard is chosen that maximizes interoperability and compatibility with other existing standards.

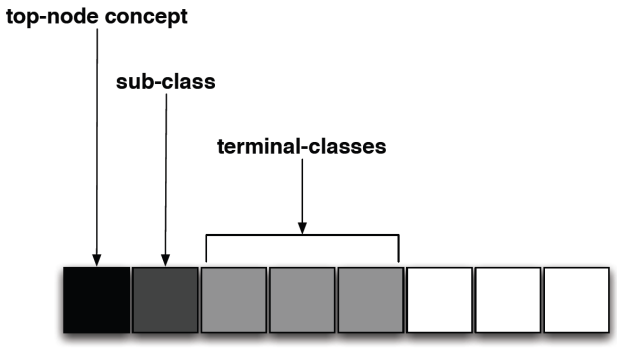
For our purposes here we decided not to promote a specific annotation standard, but rather to offer maximal interoperability in the resulting corpus annotation, as well as in the annotation tool as such, by using a tag-set that appears to be maximally compatible with existing tag-sets, as language specific as necessary, and at the same time maximally extensible. The General Ontology for Linguistic Description (GOLD) [9, 10, 8] *was originally envisioned as a solution to the problem of resolving disparate markup schemes for linguistic data*. GOLD specifies basic linguistic concepts and their interrelations, and can be used, to a certain extent, as a description logic for linguistic annotation. The current specification of GOLD is not complete, many concepts are missing, various might change, and need further specification. Nevertheless, the defined development process of GOLD handles current insufficiencies by providing language specific extensions, as well as general extensions with features necessary for the description of a wider language group.

For the purposes here, i.e. morphological and morphosyntactic annotation, the existing definition of GOLD is extended with three additional concepts. All other concepts are covered in GOLD 2008. We make use of three core concept classes in GOLD, and the necessary sub-concepts, i.e. MorphoSemanticProperty, MorphosyntacticProperty, and LinguisticExpression. The concepts defined therein relate to the notions that are expected to be emitted, i.e. morphological properties of morphemes (e.g. prefix, suffix, root), morpho-syntactic properties (e.g. case, number), and morpho-semantic properties (e.g. aspect, mood, tense).

By using the labels for concepts as defined in GOLD, we should be able to maintain maximal compatibility with other existing tag-sets. We developed example mappings to specific tag-sets, e.g. alignments to the MULTEXT-East tag-set, with the loss of features that are not defined in MULTEXT-East.

While the logic of GOLD would burden a morphological parsing algorithm, the reference to the concepts doesn't seem problematic. Representing the concepts as pure emission strings associated to the emission states, as discussed above, might decrease memory and performance benefits of a DFST-based analyzer. To maximize the performance, the GOLD-concepts and relations are mapped on a bit-vector. Encoding of the relevant concepts can be achieved with bit-vectors of less than 64 bit.

The mapping defines constants that correspond to bit-masks that are pre-compiled into the DFST. The bit-mask for example for `Genitive` might be defined as one that corresponds to set first and second bits of the terminal-class bit-field, and additionally the corresponding bits that indicate that the sub-class `CaseProperty` is set, as well as the bit for the corresponding top-node class `MorphosyntacticProperty`, as shown in the following graphic:



In a limited way, via definitions of constants and mapping of linguistic annotation in the morpheme dictionaries,

one can maintain implicatures and inheritance relations, as defined in the ontology, via bit-vector representations and appropriate bit-masks.

For the morphological analyzer this does not imply any additional processing load, i.e. the emission tuples consist of bit-vectors in form of 4-byte numerical integer values. All emitted tags are pre-compiled into the binary representation of the machine. Converting the emission tuples (i.e. individual bit-vectors) into literal string representations is achieved efficiently, once an input string is analyzed completely. This output mapping is optional such that post-processing components can consume the bit-vector representation for subsequent optimal analysis, e.g. in syntactic parsing.

3.2 Implementation

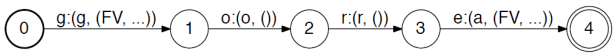
The morphological analyzer consists of two sets of code-bases. The first component converts a lexical base into a formal automaton definition. The second compiles together with the automaton definitions into a binary application.

The lexical base is kept either in database tables, spreadsheets, or textual form. The different formats allow us to maintain a minimally invasive lexical coding approach. Linguists or lexicologists are not required to learn a formal language for DFST definitions. Furthermore, they are free to use their individual way of annotation, being guided by GOLD concepts, but free to define their own, should these not be part of GOLD. The current implementation provides guidelines for the data-format, but also the possibility to use individual scripts for data conversion and annotation mappings.

The individual morpheme lists, annotations and rule definitions are compiled into Ragel [20] automata definitions, as described above. Besides rules that are related to concrete morpheme lists and the corresponding DFSTs, there are also guessing rules that define general properties of nouns, verbs and adjectives. The features that are used are mapped on bit-vectors, and C-header files with the constant literal and bit-vector mask definitions are generated.

Ragel generates a monolithic DFST as C-code, using highly efficient C-jump code (`goto`-statements), as well as a DOT-file for visualization of the resulting automaton (using e.g. Graphviz[3]). The generated code is wrapped in a C++ class that handles input and output, and controls the program logic.

In the current version the generation of the root- and the base-lemma is encoded in the emission bit-vector. One byte is reserved to mark the reverse offset for string concatenation, while two bytes are reserved to point to an element in a string array with the corresponding string that needs to be appended. The form *čitamo* would be associated with an offset of -2 and a corresponding suffix *ti*. This solution doesn't match the general declarative paradigm of FSTs, and is just temporary. In the next release the output characters of the corresponding lemma will be integrated in the emission of the transducer, associated with each single transition, as shown in the graph below. Thus every emission will be a tuple that contains tuples of output characters and optional annotation bit-vectors.



The parser expects a token list as input. The code-page of the lexical base for machine compilation has to match the input tokens. Otherwise there is no restriction on a specific code-page or character encoding, since the automaton processes strings by consuming bytes in the binary representation.

Tokens are processed sequentially. For each token, all emitted tuples are collected in a stack. Only matching start- and end-tuples are returned, if there are compatible sub-morpheme analyses that span over the complete input token length. Thus, no hypotheses of sub-morphemes are generated, and the number of irrelevant hypotheses is radically reduced.

The significant implementation features that differentiate our implementation from other solutions, are that the code-base is platform independent and open-source, based on free and open tools like GCC and Ragel. Furthermore, the fact that doesn't transcode the lexical base or the input words, it can be based on any encoding. The binary processing strategy allows even for mixed encoding of the lexical base and input tokens, without major consequences for the size and efficiency of the resulting machine.

The extension of the morphological base is kept trivial, along the lines of the requirements specified above, i.e. the necessity to be able to add newly identified morphemes or paradigms from diachronic and synchronic variants.

4 Evaluation

The evaluation version of the implementation for Croatian contains approx. 120,000 morphemes in its morpheme-base, using UTF-8 character encoding. The number of strings it can recognize is infinite, due to cyclic sub-automata. Unknown word-forms can be analyzed due to incorporated guessing rules.

For the following evaluation results we used a 2.4 GHz 64-bit Dual-Core CPU. In the evaluation version only a single core is used during runtime of the FST, while both CPU cores are used during compilation.

Compilation of the morphology requires min. 4 GB of RAM using GCC 4.2. This is expected due to the monolithic architecture, and since the Ragel-generated C-code of the transducer gets very large. The compilation process takes less than 5 minutes, using both CPU cores. The resulting binary footprint is less than 5 MB of size.

The final automaton consists of approx. 150,000 transitions and 25,000 states.

[3] See `http://www.graphviz.org/` for details.

We selected randomly 10,000 tokens with an average morpheme length of 2.5 morphemes. The parser processes in average approx. 50,000 tokens per second (real 10,000 tokens per 150 millisec.), including runtime instantiation in RAM, mapping of the analysis bit-vectors to the corresponding string representations, generation of lemmata, and output redirection to a log-file. An extension of the morpheme base has no significant impact on memory instantiation time, neither on the runtime behavior. The memory instantiation can be marginalized for a large processing sample.

The current implementation doesn't include transitional or emission-probabilities, due to missing quantitative information from training data. Once an annotated corpus is available, these weights can trivially be implemented as additional weights in the emission tuple. The described machine is not disambiguating the generated output. For disambiguation the transitional probabilities (and thus the likelihood of a given parse for one lexeme) might be useful. In general we are convinced that disambiguation necessarily has to rely on contextual information, and thus must include some sort of parser or contextual language model, i.e. be part of a more complex analysis component.

A relevant evaluation result is the coefficient of the ratio between all and relevant emissions, i.e. the percentage of relevant (possible) morpheme analyses and all generated ones. Due to certain limitations, we cannot perform such an evaluation, neither a recall evaluation on a predefined evaluation corpus. Future availability of reference corpora should enable us to provide such extremely relevant evaluation results.

For evaluation and potential application to other languages, the source code is made available on the web site `http://personal.unizd.hr/~dcavar/CroMo/`.

References

[1] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, pages 11–23. Springer-Verlag, 2007.

[2] Kenneth R. Beesley and Lauri Karttunen. *Finite State Morphology*. CSLI Publications, Stanford, April 2003.

[3] Jean Berstel. *Transductions and Context-Free Languages*. Teubner Studienbücher, Stuttgart, 1979.

[4] Jean Berstel and Christophe Reutenauer. *Rational Series and Their Languages*. EaTCS Monographs on Theoretical Computer Science. Springer-Verlag, Berlin, December 1988.

[5] Paul E. Black. Dictionary of algorithms and data structures. Online publication: U.S. National Institute of Standards and Technology, Available online, December 2004.

[6] Dunja Brozović-Rončević and Damir Ćavar. Hrvatska jezična riznica kao podloga jezičnim i jezičnopovijesnim istraživanjima hrvatskoga jezika. In *Vidjeti Ohrid*, Hrvatska sveučilišna naklada, pages 173–186, Zagreb, 2008. Hrvatsko filološko društvo.

[7] Gavin Burnage. CELEX - A guide for users. Technical report, Centre for Lexical Information, University of Nijmegen, Nijmegen, 1990.

[8] Scott O. Farrar. *An Ontology for Linguistics on the Semantic Web*. PhD thesis, The University of Arizona, Tucson, Arizona, 2003.

[9] Scott O. Farrar and D. Terence Langendoen. A linguistic ontology for the semantic web. *Glot International*, 7(3):1–4, March 2003.

[10] Scott O. Farrar, William D. Lewis, and D. Terence Langendoen. A common ontology for linguistic concepts. In N. Ide and C. Welty, editors, *Semantic Web Meets Language Resources: Papers from the AAAI Workshop*, pages 11–16. AAAI Press, Menlo Park, CA, 2002.

[11] Alexander Geyken and Thomas Hanneforth. TAGH: A complete morphology for german based on weighted finite state automata. In Anssi Yli-Jyrä, Lauri Karttunen, and Juhani Karhumäki, editors, *FSMNLP*, volume 4002 of *Lecture Notes in Computer Science*, pages 55–66. Springer, September 2005.

[12] Vjera Lopina. Strojna obrada imenične morfologije u pisanome hrvatskom jeziku. Ma thesis, Centar za postdiplomske studije Dubrovnik, Dubrovnik, October 1999.

[13] George H. Mealy. A method for synthesizing sequential circuits. *Bell System Technical Journal*, 34(5):1045—1079, September 1955.

[14] Antoni Oliver and Marko Tadić. Enlarging the croatian morphological lexicon by automatic lexical acquisition from raw corpora. In *Proceedings of LREC 2004*, volume IV, pages 1259–1262, Lisbon, May 2004. ELRA.

[15] Brian Roark and Richard Sproat. *Computational Approaches to Syntax and Morphology*. Oxford University Press, Oxford, 2007.

[16] Jerzy J. Rubach. *Cyclic and Lexical Phonology. The Structure of Polish*. Foris Publications, Dordrecht, 1984.

[17] Helmut Schmid, Arne Fitschen, and Ulrich Heid. SMOR: A german computational morphology covering derivation, composition, and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263–1266, Lisbon, Portugal, 2004.

[18] Richard Sproat. *A Computational Theory of Writing Systems*. AT&T Bell Laboratories, New Jersey, July 2000.

[19] Marko Tadić. *Računalna obradba morfologije hrvatskoga književnog jezika*. doctoral dissertation, Filozofski fakultet Sveučilišta u Zagrebu, Zagreb, Croatia, 1994.

[20] Adrian D. Thurston. Parsing computer languages with an automaton compiled from a single regular expression. In *11th International Conference on Implementation and Application of Automata (CIAA 2006)*, volume 4094 of *Lecture Notes in Computer Science*, pages 285–286, Taipei, Taiwan, August 2006.

[21] Andrea Zielinski and Christian Simon. Morphisto – an open-source morphological analyzer for german. In *Proceedings of FSMNLP 2008*, Ispra, Italy, September 2008.

SOR '09

The tenth International Symposium on Operational Research in Slovenia - SOR'09, http://www.fgg.uni-lj.si/SOR09, will take place in Nova Gorica, Slovenia, September 23-25, 2009.

SOR'09 is organized by Slovenian Society Informatika (SDI) Section of Operations Research (SOR). This symposium is the premiere scientific event in the area of operations research. It represents a continuity of nine previous symposia, which have attracted a growing number of international audience. As traditionally, SOR'09 will provide an international forum for scientific exchange at the frontiers of operations research (OR) in mathematics, statistics, economics, engineering, education, environment, computer science etc. Since OR comprises a large variety of mathematical, statistical and informational theories and methods to analyze complex situations and to contribute to responsible decision making, planning and the efficient use of the resources, we believe, that in the world of increasing complexity and scarce natural resources there will be a growing need for such approaches in many fields of our society.

The scientific program of the 10th symposium SOR'09 will consist of plenary lectures and contributed papers, where authors from different countries will present their work in the fields of the OR. The main topics of the international symposium are focused on professional aspects of OR, methods and techniques of OR, areas of application, information and computing aspects of OR.

More information about SOR’09 is available at http://www.fgg.uni-lj.si/SOR09.

The organizers of SOR’09 are looking forward to welcoming you to Nova Gorica, Slovenia and sharing with you a stimulating scientific and professional atmosphere. In addition to this, you will have the opportunity to explore the western part of Slovenia.

On behalf of SDI-SOR and Programme and Organizing Committee of SOR'09

Prof. dr. Lidija Zadnik Stirn

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 800 staff, has 600 researchers, about 250 of whom are postgraduates, nearly 400 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit an email with the manuscript to one of the editors from the Editorial Board or to the Managing Editor. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

☐ Send Informatica free of charge

☐ Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than sixteen years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering the European computer science and informatics community - scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name: .....

Title and Profession (optional): .....

.....

Home Address and Telephone (optional): .....

.....

Office Address and Telephone (optional): .....

.....

E-mail Address (optional): .....

Signature and Date: .....

Informatica WWW:

http://www.informatica.si/

Referees:

Witold Abramowicz, David Abramson, Adel Adi, Kenneth Aizawa, Suad Alagić, Mohamad Alam, Dia Ali, Alan Aliu, Richard Amoroso, John Anderson, Hans-Jurgen Appelrath, Iván Araujo, Vladimir Bajič, Michel Barbeau, Grzegorz Bartoszewicz, Catriel Beeri, Daniel Beech, Fevzi Belli, Simon Beloglavec, Sondes Bennasri, Francesco Bergadano, Istvan Berkeley, Azer Bestavros, Andraž Bežek, Balaji Bharadwaj, Ralph Bisland, Jacek Blazewicz, Laszlo Boeszoermenyi, Damjan Bojadžijev, Jeff Bone, Ivan Bratko, Pavel Brazdil, Bostjan Brumen, Jerzy Brzezinski, Marian Bubak, Davide Bugali, Troy Bull, Sabin Corneliu Buraga, Leslie Burkholder, Frada Burstein, Wojciech Buszkowski, Rajkumar Bvyya, Giacomo Cabri, Netiva Caftori, Particia Carando, Robert Cattral, Jason Ceddia, Ryszard Choras, Wojciech Cellary, Wojciech Chybowski, Andrzej Ciepielewski, Vic Ciesielski, Mel Ó Cinnéide, David Cliff, Maria Cobb, Jean-Pierre Corriveau, Travis Craig, Noel Craske, Matthew Crocker, Tadeusz Czachorski, Milan Češka, Honghua Dai, Bart de Decker, Deborah Dent, Andrej Dobnikar, Sait Dogru, Peter Dolog, Georg Dorfner, Ludoslaw Drelichowski, Matija Drobnič, Maciej Drozdowski, Marek Druzdzel, Marjan Družovec, Jozo Dujmović, Pavol Ďuriš, Amnon Eden, Johann Eder, Hesham El-Rewini, Darrell Ferguson, Warren Fergusson, David Flater, Pierre Flener, Wojciech Fliegner, Vladimir A. Fomichov, Terrence Forgarty, Hans Fraaije, Stan Franklin, Violetta Galant, Hugo de Garis, Eugeniusz Gatnar, Grant Gayed, James Geller, Michael Georgiopolus, Michael Gertz, Jan Goliński, Janusz Gorski, Georg Gottlob, David Green, Herbert Groiss, Jozsef Gyorkos, Marten Haglind, Abdelwahab Hamou-Lhadj, Inman Harvey, Jaak Henno, Marjan Hericko, Henry Hexmoor, Elke Hochmueller, Jack Hodges, John-Paul Hosom, Doug Howe, Rod Howell, Tomáš Hruška, Don Huch, Simone Fischer-Huebner, Zbigniew Huzar, Alexey Ippa, Hannu Jaakkola, Sushil Jajodia, Ryszard Jakubowski, Piotr Jedrzejowicz, A. Milton Jenkins, Eric Johnson, Polina Jordanova, Djani Juričič, Marko Juvancic, Sabhash Kak, Li-Shan Kang, Ivan Kapustøk, Orlando Karam, Roland Kaschek, Jacek Kierzenka, Jan Kniat, Stavros Kokkotos, Fabio Kon, Kevin Korb, Gilad Koren, Andrej Krajnc, Henryk Krawczyk, Ben Kroese, Zbyszko Krolikowski, Benjamin Kuipers, Matjaž Kukar, Aarre Laakso, Sofiane Labidi, Les Labuschagne, Ivan Lah, Phil Laplante, Bud Lawson, Herbert Leitold, Ulrike Leopold-Wildburger, Timothy C. Lethbridge, Joseph Y-T. Leung, Barry Levine, Xuefeng Li, Alexander Linkevich, Raymond Lister, Doug Locke, Peter Lockeman, Vincenzo Loia, Matija Lokar, Jason Lowder, Kim Teng Lua, Ann Macintosh, Bernardo Magnini, Andrzej Małachowski, Peter Marcer, Andrzej Marciniak, Witold Marciszewski, Vladimir Marik, Jacek Martinek, Tomasz Maruszewski, Florian Matthes, Daniel Memmi, Timothy Menzies, Dieter Merkl, Zbigniew Michalewicz, Armin R. Mikler, Gautam Mitra, Roland Mittermeir, Madhav Moganti, Reinhard Moller, Tadeusz Morzy, Daniel Mossé, John Mueller, Jari Multisilta, Hari Narayanan, Jerzy Nawrocki, Rance Necaise, Elzbieta Niedzielska, Marian Niedq'zwiedziński, Jaroslav Nieplocha, Oscar Nierstrasz, Roumen Nikolov, Mark Nissen, Jerzy Nogieć, Stefano Nolfi, Franc Novak, Antoni Nowakowski, Adam Nowicki, Tadeusz Nowicki, Daniel Olejar, Hubert Österle, Wojciech Olejniczak, Jerzy Olszewski, Cherry Owen, Mieczyslaw Owoc, Tadeusz Pankowski, Jens Penberg, William C. Perkins, Warren Persons, Mitja Peruš, Fred Petry, Stephen Pike, Niki Pissinou, Aleksander Pivk, Ullin Place, Peter Planinšec, Gabika Polčicová, Gustav Pomberger, James Pomykalski, Tomas E. Potok, Dimithu Prasanna, Gary Preckshot, Dejan Rakovič, Cveta Razdevšek Pučko, Ke Qiu, Michael Quinn, Gerald Quirchmayer, Vojislav D. Radonjic, Luc de Raedt, Ewaryst Rafajlowicz, Sita Ramakrishnan, Kai Rannenberg, Wolf Rauch, Peter Rechenberg, Felix Redmill, James Edward Ries, David Robertson, Marko Robnik, Colette Rolland, Wilhelm Rossak, Ingrid Russel, A.S.M. Sajeev, Kimmo Salmenjoki, Pierangela Samarati, Bo Sanden, P. G. Sarang, Vivek Sarin, Iztok Savnik, Ichiro Satoh, Walter Schempp, Wolfgang Schreiner, Guenter Schmidt, Heinz Schmidt, Dennis Sewer, Zhongzhi Shi, Mária Smolárová, Carine Souveyet, William Spears, Hartmut Stadtler, Stanislaw Stanek, Olivero Stock, Janusz Stokłosa, Przemysław Stpiczyński, Andrej Stritar, Maciej Stroinski, Leon Strous, Ron Sun, Tomasz Szmuc, Zdzislaw Szyjewski, Jure Šilc, Metod Škarja, Jiři Šlechta, Chew Lim Tan, Zahir Tari, Jurij Tasič, Gheorge Tecuci, Piotr Teczynski, Stephanie Teufel, Ken Tindell, A Min Tjoa, Drago Torkar, Vladimir Tosic, Wieslaw Traczyk, Denis Trček, Roman Trobec, Marek Tudruj, Andrej Ule, Amjad Umar, Andrzej Urbanski, Marko Uršič, Tadeusz Usowicz, Romana Vajde Horvat, Elisabeth Valentine, Kanonkluk Vanapipat, Alexander P. Vazhenin, Jan Verschuren, Zygmunt Vetulani, Olivier de Vel, Didier Vojtisek, Valentino Vranić, Jozef Vyskoc, Eugene Wallingford, Matthew Warren, John Weckert, Michael Weiss, Tatjana Welzer, Lee White, Gerhard Widmer, Stefan Wrobel, Stanislaw Wrycza, Tatyana Yakhno, Janusz Zalewski, Damir Zazula, Yanchun Zhang, Ales Zivkovic, Zonling Zhou, Robert Zorc, Anton P. Železnikar

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Vožarski pot 12, 1000 Ljubljana, Slovenia.
The subscription rate for 2009 (Volume 33) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: Dikplast Kregar Ivan s.p., Kotna ulica 5, 3000 Celje.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Franjo Pernuš)
Slovenian Artificial Intelligence Society; Cognitive Science Society (Matjaž Gams)
Slovenian Society of Mathematicians, Physicists and Astronomers (Bojan Mohar)
Automatic Control Society of Slovenia (Borut Zupančič)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Igor Grabec)
ACM Slovenia (Dunja Mladenič)

Informatica is surveyed by: Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

The issuing of the Informatica journal is financially supported by the Ministry of Higher Education, Science and Technology, Trg OF 13, 1000 Ljubljana, Slovenia.

Volume 33 Number 1 March 2009 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **33** (2009) Number 1, pp. 1–115