

Bart De Decker
Jana Dittmann
Christian Kraetzer
Claus Vielhauer (Eds.)

LNCS 8099

Communications and Multimedia Security

14th IFIP TC 6/TC 11 International Conference, CMS 2013
Magdeburg, Germany, September 2013
Proceedings



ifip



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Bart De Decker Jana Dittmann
Christian Kraetzer Claus Vielhauer (Eds.)

Communications and Multimedia Security

14th IFIP TC 6/TC 11 International Conference, CMS 2013
Magdeburg, Germany, September 25-26, 2013
Proceedings

Volume Editors

Bart De Decker

KU Leuven, Department of Computer Science, iMinds-DistriNet

Celestijnenlaan 200A, 3001 Leuven, Belgium

E-mail: bart.dedecker@cs.kuleuven.be

Jana Dittmann

Otto-von-Guericke-Universität Magdeburg

Universitätsplatz 2, 39106 Magdeburg, Germany

E-mail: jana.dittmann@iti.cs.uni-magdeburg.de

Christian Kraetzer

Otto-von-Guericke-Universität Magdeburg

Universitätsplatz 2, 39106 Magdeburg, Germany

E-mail: kraetzer@iti.cs.uni-magdeburg.de

Claus Vielhauer

Fachhochschule Brandenburg/Otto-von-Guericke-Universität Magdeburg

Magdeburger Str. 50, 14770 Brandenburg, Germany

E-mail: claus.vielhauer@fh-brandenburg.de

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-40778-9

e-ISBN 978-3-642-40779-6

DOI 10.1007/978-3-642-40779-6

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: : 2013946793

CR Subject Classification (1998): K.4.4, E.3, C.2.0, C.2, K.6.5, J.1, H.4

LNCS Sublibrary: SL 4– Security and Cryptology

© IFIP International Federation for Information Processing 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

It is with great pleasure that we present the proceedings of the 14th IFIP TC-6 and TC-11 Conference on Communications and Multimedia Security (CMS 2013), which was held in Magdeburg, Germany during September 25–26, 2013. The meeting continues the tradition of previous CMS conferences which were held in Canterbury, UK (2012), Ghent, Belgium (2011) and Linz, Austria (2010).

The Program Committee (PC) received 30 submissions, comprising 23 full papers, 6 short papers and 1 extended abstract, out of which only 5 full papers were accepted (22% acceptance rate). In this edition, we have included 11 short papers, which describe valuable work-in-progress, as well as 5 extended abstracts, which describe the posters that were discussed at the conference. Some of the latter two categories are shortened versions of original full or short paper submissions respectively, which the PC judged to be valuable contributions but somewhat premature for submission under their original category.

We are grateful to Prof. Sabah Jassim of the University of Buckingham, UK and Prof. Sachar Paulus of the Brandenburg University of Applied Sciences, Brandenburg, Germany for accepting our invitations to deliver keynote addresses, the abstracts of which can be found at the end of these proceedings.

We would also like to say a word of appreciation to our sponsor: The Advanced Multimedia and Security Lab (AMSL) of the Otto-von-Guericke University of Magdeburg. Without its financial support, it would not have been possible to attract as many young researchers.

Finally, special thanks go to the organizing committee who handled all local organizational issues and provided us with a comfortable and inspiring location and an interesting evening event. For us, it was a distinct pleasure to serve as program chairs of CMS 2013.

We hope that you will enjoy reading these proceedings and that they may inspire you for future research in communications and multimedia security.

September 2013

Bart De Decker
Jana Dittmann
Claus Vielhauer

Organization

CMS 2013 is the 14th Joint IFIP TC6 and TC11 Conference on Communications and Multimedia Security. It has been organized by the Otto-von-Guericke University of Magdeburg, Germany.

Executive Committee

Conference Chair

Claus Vielhauer

Brandenburg University of
Applied Sciences, Germany

Program Co-Chairs

Bart De Decker

KU Leuven, Belgium

Jana Dittmann

Otto-von-Guericke University Magdeburg,
Germany

Claus Vielhauer

Organizing Chair

Jana Dittmann

Organizing Committee

Jana Dittmann

Christian Kraetzer

Silke Reifgerste

Program Committee

Anas Abou El Kalam

UCA-ENSA of Marrakesh, Morocco

Eric Alata

LAAS-CNRS, France

Patrick Bas

CNRS-Lagis, Lille, France

David W. Chadwick

University of Kent, UK

Howard Chivers

University of York, UK

Isabelle Chrismnt

LORIA-University of Nancy, France

Gabriela F. Ciocarlie

Computer Science Lab, SRI International, USA

Frédéric Cuppens

Télécom Bretagne, France

Italo Dacosta

KU Leuven, Belgium

Hervé Debar

Télécom SudParis, France

VIII Organization

Sabrina De Capitani di Vimercati	Università degli Studi di Milano, Italy
Bart De Decker	KU Leuven, Belgium
Yvo Desmedt	University of Texas at Dallas, USA and University College London, UK
Lieven De Strycker	KU Leuven, Technology Campus Ghent, Belgium
Jana Dittmann	University of Magdeburg, Germany
Stelios Dritsas	Athens University of Economics and Business, Greece
Gerhard Eschelbeck	Sophos, USA
Simone Fischer-Hübner	Karlstad University, Sweden
Teddy Furon	Inria Rennes - Bretagne Atlantique, France
Jürgen Fuß	University of Applied Sciences Upper Austria, Hagenberg, Austria
Sébastien Gambs	Université de Rennes 1 - Inria/Irisa, France
Christian Geuer-Pollmann	Microsoft Research, Germany
Dieter Gollmann	Hamburg University of Technology, Germany
Jean Hennebert	University of Applied Sciences, HES-SO, Switzerland
Eckehard Hermann	University of Applied Sciences Upper Austria, Hagenberg, Austria
Jens Hermans	KU Leuven, Belgium
Andreas Humm	University of Fribourg, Switzerland
Edward Humphreys	XiSEC, UK
Christophe Huygens	KU Leuven, Belgium
Witold Jacak	University of Applied Sciences Upper Austria, Hagenberg, Austria
Sushil Jajodia	George Mason University, USA
Günter Karjoth	IBM Research - Zurich, Switzerland
Stefan Katzenbeisser	TU Darmstadt, Germany
Markulf Kohlweiss	Microsoft Research Cambridge, UK
Romain Laborde	Institut de Recherche en Informatique de Toulouse (IRIT), France
Jorn Lapon	KU Leuven, Technology Campus Ghent, Belgium
Herbert Leitold	Secure Information Technology Center (A-SIT), Austria
Javier Lopez	University of Malaga, Spain
Louis Marinou	European Network and Information Security Agency (ENISA), Greece

Keith Martin	Royal Holloway, University of London, UK
Chris Mitchell	Royal Holloway, University of London, UK
Refik Molva	Eurécom, France
Yuko Murayama	Iwate Prefectural University, Japan
Vincent Naessens	KU Leuven, Technology Campus Ghent, Belgium
Nick Nikiforakis	KU Leuven, Belgium
Chandrasekaran Pandurangan	Indian Institute of Technology, Madras, India
Günther Pernul	University of Regensburg, Germany
Alessandro Piva	University of Florence, Italy
Franz-Stefan Preiss	IBM Research - Zurich, Switzerland
Jean-Jacques Quisquater	Université catholique de Louvain, Belgium
Kai Rannenber	Goethe University Frankfurt, Germany
Pierangela Samarati	Università degli Studi di Milano, Italy
Riccardo Scandariato	KU Leuven, Belgium
Ingrid Schaumüller-Bichl	University of Applied Sciences Upper Austria, Hagenberg, Austria
Jörg Schwenk	Ruhr-Universität Bochum, Germany
Stefaan Seys	KU Leuven, Belgium
Einar Snekkenes	Gjøvik University College, Norway
Andreas Uhl	University of Salzburg, Austria
Umut Uludag	Scientific and Technological Research Council (TUBITAK), Turkey
Vijay Varadharajan	Macquarie University, Australia
Pedro Veiga	University of Lisbon, Portugal
Claus Vielhauer	Brandenburg University of Applied Sciences, Germany
Tatjana Welzer	University of Maribor, Slovenia
Andreas Westfeld	Dresden University of Applied Sciences, Germany
Ted Wobber	Microsoft Research Silicon Valley, USA
Shouhuai Xu	University of Texas at San Antonio, USA
Moti Yung	Google & Columbia University, USA
Gansen Zhao	South China Normal University, China
Ge Zhang	Karlstad University, Sweden

Reviewers

Filipe Beato	Microsoft Research, Cambridge, UK
Michael Diener	University of Regensburg, Germany
Jean-Luc Dugelay	Eurécom, France
Miltiadis Kandias	Athens University of Economics and Business, Greece

Andrea Melle	Eurécom, France
Aleksios Mylonas	Athens University of Economics and Business, Greece
Moritz Riesner	University of Regensburg, Germany
Ahmad Sabouri	Goethe University Frankfurt, Germany
Moustafa Saleh	University of Texas, San Antonio, USA
Dieter Sommer	IBM Research - Zurich, Switzerland
Fatbardh Veseli	Goethe University Frankfurt, Germany
Qingji Zheng	University of Texas, San Antonio, USA

Sponsoring Institutions

The Advanced Multimedia and Security Lab (AMSL) of the Otto-von-Guericke University of Magdeburg, Germany.

Table of Contents

Part I: Research Papers

Biometrics

Towards a Standardised Testsuite to Assess Fingerprint Matching Robustness: The StirMark Toolkit – Cross-Feature Type Comparisons	3
<i>Jutta Hämmerle-Uhl, Michael Pober, and Andreas Uhl</i>	
Achieving Anonymity against Major Face Recognition Algorithms	18
<i>Benedikt Driessen and Markus Dürmuth</i>	
Client-Side Biometric Verification Based on Trusted Computing	34
<i>Jan Vossaert, Jorn Lapon, Bart De Decker, and Vincent Naessens</i>	

Applied Cryptography

Dedicated Hardware for Attribute-Based Credential Verification	50
<i>Geoffrey Ottoy, Jorn Lapon, Vincent Naessens, Bart Preneel, and Lieven De Strycker</i>	
Decentralized Ciphertext-Policy Attribute-Based Encryption Scheme with Fast Decryption	66
<i>Y. Sreenivasa Rao and Ratna Dutta</i>	

Part II: Work in Progress

Biometrics

Security of Features Describing the Visual Appearance of Handwriting Samples Using the Bio-hash Algorithm of Vielhauer against an Evolutionary Algorithm Attack	85
<i>Andreas Hasselberg, Rene Zimmermann, Christian Kraetzer, Tobias Scheidat, Claus Vielhauer, and Karl Kümmel</i>	

Digital Watermarking, Steganography and Forensics

Video Watermarking Scheme with High Payload and Robustness against Geometric Distortion	95
<i>Huajian Liu, Yiyao Li, and Martin Steinebach</i>	
Use of Linear Error-Correcting Subcodes in Flow Watermarking for Channels with Substitution and Deletion Errors	105
<i>Boris Assanovich, William Puech, and Iuliia Tkachenko</i>	
Detecting Resized Double JPEG Compressed Images – Using Support Vector Machine	113
<i>Hieu Cuong Nguyen and Stefan Katzenbeisser</i>	
Pit Stop for an Audio Steganography Algorithm	123
<i>Andreas Westfeld, Jürgen Wurzer, Christian Fabian, and Ernst Piller</i>	
Robust Hash Algorithms for Text	135
<i>Martin Steinebach, Peter Klöckner, Nils Reimers, Dominik Wienand, and Patrick Wolf</i>	
Hardware Based Security Enhanced Direct Memory Access	145
<i>Marcel Eckert, Igor Podebrad, and Bernd Klauer</i>	

Social Network Privacy, Security and Authentication

Privacy Visor: Method for Preventing Face Image Detection by Using Differences in Human and Device Sensitivity	152
<i>Takayuki Yamada, Seiichi Gohshi, and Isao Echizen</i>	
E-Learning of IT Security Threats: A Game Prototype for Children	162
<i>Jana Fruth, Carsten Schulze, Marleen Rohde, and Jana Dittmann</i>	
Hiding Information in Social Networks from De-anonymization Attacks by Using Identity Separation	173
<i>Gábor György Gulyás and Sándor Imre</i>	
An Equivalent Access Based Approach for Building Collaboration Model between Distinct Access Control Models	185
<i>Xiaofeng Xia</i>	

Part III: Extended Abstracts

Authentication with Time Features for Keystroke Dynamics on Touchscreens	197
<i>Matthias Trojahn, Florian Arndt, and Frank Ortmeier</i>	

Visibility Assessment of Latent Fingerprints on Challenging Substrates in Spectroscopic Scans	200
<i>Mario Hildebrandt, Andrey Makrushin, Kun Qian, and Jana Dittmann</i>	
Creation of a Public Corpus of Contact-Less Acquired Latent Fingerprints without Privacy Implications	204
<i>Mario Hildebrandt, Jennifer Sturm, Jana Dittmann, and Claus Vielhauer</i>	
SocACL: An ASP-Based Access Control Language for Online Social Networks	207
<i>Edward Caprin and Yan Zhang</i>	
Watermark Resynchronization: An Efficient Approach Based on Eulerian Tours around a Robust Skeleton	211
<i>Konstantinos Raftopoulos, Klimis Ntalianis, Paraskevi Tzouveli, Nicolas Tsapatsoulis, Aleatha Parker-Wood, and Marin Ferecatu</i>	

Part IV: Keynotes

Face Recognition from Degraded Images – Super Resolution Approach by Non-adaptive Image-Independent Compressive Sensing Dictionaries	217
<i>Sabah A. Jassim</i>	
Trustworthy Software Development	233
<i>Sachar Paulus, Nazila Gol Mohammadi, and Thorsten Weyer</i>	
Author Index	249

Part I
Research Papers

Towards a Standardised Testsuite to Assess Fingerprint Matching Robustness: The StirMark Toolkit – Cross-Feature Type Comparisons

Jutta Hämmerle-Uhl, Michael Pober, and Andreas Uhl

Multimedia Signal Processing and Security Lab (WaveLab)
Department of Computer Sciences, University of Salzburg
andreas.uhl@sbg.ac.at

Abstract. We propose to establish a standardised tool in fingerprint recognition robustness assessment, which is able to simulate a wide class of acquisition conditions, applicable to any given dataset and also of potential interest in forensic analysis. As an example, StirMark image manipulations (as being developed in the context of watermarking robustness assessment) are applied to fingerprint data to generate test data for robustness evaluations, thereby interpreting certain image manipulations as being highly related to realistic fingerprint acquisition conditions. Experimental results involving three types of fingerprint features and matching schemes (i.e. correlation-based, ridge feature-based, and minutiae-based) applied to FVC2004 data underline the need for standardised testing and a corresponding simulation toolset.

1 Introduction

One of the big issues in fingerprint recognition is robustness of recognition accuracy against sample image quality degradation [1, 2]. The performance of a fingerprint recognition system is usually heavily affected by fingerprint image quality. A wide variety of factors influence the quality of a fingerprint image: Skin conditions (e.g. , dryness, moisture, dirt, cuts and bruises), sensor conditions (e.g. , dirt, noise, size), and other acquisition conditions like user cooperation or crime scene preservation in forensic settings, etc. Some of these factors are inevitable and some of them change over time. Poor quality images often result in spurious and missed features, therefore decreasing the recognition accuracy of the overall system.

However, the different levels at which fingerprint features are extracted [2] and the different feature types extracted at these levels influence the impact of quality degradations on recognition performance in various ways. Moreover, there is interplay among different types of feature extraction and acquisition technology / conditions such that it is not clear a priori which type of feature extraction is favourable under which conditions. Therefore, it is essential to provide reliable methodology to comparatively assess fingerprint recognition robustness under varying conditions.

This issue is classically tackled from two sides: First, benchmarking frameworks have been established, which facilitate a common evaluation basis with standardised

protocols for various fingerprint recognition algorithms, see *e.g.* the fingerprint verification contests (FVC [2]), independent suggestions like [3], and the BioSecure evaluation framework [1]. Second, usually these frameworks rely on the establishment of test data which are used to compare the different algorithms on a common basis. A very good example, specifically focusing onto the robustness issue, are the FVC data sets. FVC2002 (only i) & iv)) and FVC2004 data have been acquired in a way to introduce higher intraclass variation by i) putting the finger at slightly different vertical position, ii) applying low or high pressure against the sensor, iii) exaggerating skin distortion and rotation, and iv) drying or moistening fingers. For FVC2006, the population was chosen to be more heterogeneous, including manual workers and elderly people.

While the availability of these and similar datasets is a significant achievement, the data collection and database establishment is tedious work. Moreover, if additional acquisition conditions should be considered which have not been included into the original dataset, re-enrolment is required, involving complicated procedures for getting the original people back to enrolment. Also, it is hard to compare the different quality degradations from dataset to dataset (*e.g.* FVC, MCYT, BIOMET, MSU), since usually, there is no standardised manner to generate the acquisition conditions applied. Therefore, the experimental results of recognition algorithms in case applied to different datasets are hardly comparable and the results shown in many papers are difficult to interpret.

A strategy to cope with the various problems of generating natural datasets is to generate synthetic fingerprints, the SFinGe [4] being the most well known tool for doing this. The generated fingerprints have proven to be highly realistic and serve as a sensible tool to generate large datasets for benchmarking. While SFinGe also allows to apply some manipulations to the images, *e.g.* noise insertion, translations, rotations and uses a skin deformation model, a simulation of specific sensor types is not foreseen.

In the area of robust watermarking, a similar situation could be observed – while of course the notion of robustness is different in watermarking (means basically the ability of embedded data to withstand common image manipulations or unspecific attacks), the general problem was of comparable nature: Each watermarking scheme presented was evaluated on a specific dataset, where especially the types of introduced image manipulations and their respective extent to prove robustness varied from paper to paper, thus making a comparison of techniques impossible. To cope with the situation, standardised benchmark toolsets consisting of a collection of parameterisable image manipulations have been created, including StirMark [5] and CheckMark [6]. This enabled developers and authors to apply these manipulations to publicly available datasets thus making their results comparable.

In recent work [7], we have proposed to establish a standardised tool in fingerprint recognition robustness assessment, which is able to simulate a wide class of acquisition conditions, applicable to any given dataset. As an example, StirMark image manipulations have been applied to fingerprint data to generate test data for robustness evaluations. Since these manipulations can be applied to any dataset, the effect of manipulations on data originating from different sensors and acquisition conditions can be studied with respect to recognition accuracies of the algorithms used. Contrasting to previous work [7], where experiments have been restricted to fingerprint matchers of minutiae type, here we focus on fingerprint matchers relying on very different

feature types and compare the obtained results. Additionally, different distortion types are investigated as compared to [7].

In Section 2, we explain the StirMark image manipulations and discuss the interpretation of those procedures in the context of fingerprint acquisition and quality, respectively. Section 3 briefly reviews the fundamental ideas behind three very different types of fingerprint feature extraction and matching techniques which are subsequently used in experiments. Experimental results are presented in Section 4 where we shortly describe the employed FVC2004 dataset and experimental conditions with respect to evaluation protocols. Finally, we present fingerprint verification results generated on the FVC2004 dataset processed with a set of StirMark image manipulations with increasing strength. Section 5 concludes the paper.

2 The StirMark Toolkit

The StirMark Benchmark is a generic benchmark test for evaluating the robustness of digital image watermarking methods, developed by Fabien A. P. Petitcolas *et al.* [5, 8]. The basic idea behind the robustness tests in the StirMark benchmark is, that a digital watermark within an image can be attacked and possibly rendered useless, by introducing small, ideally imperceptible perturbations into the marked image. To be suitable for application in a common generic benchmark, the specific types of perturbations are pre-defined and the respective intensity is adjustable via a given set of parameters. The corresponding software is currently available “StirMark Benchmark 4.0” at <http://www.petitcolas.net/fabien/watermarking/stirmark/>. A related, also watermarking-robustness focused toolset is CheckMark [6] which could be used by analogy, however, it is less well supported.

In the following, we describe the set of StirMark image manipulations that has been selected for this study. We explain the way each manipulation is defined, how it is parameterised to achieve varying strength of the manipulation, and we discuss which realistic fingerprint acquisition condition could be modelled by applying the manipulation to fingerprint sample images. Thus, only a subset of the complete range of StirMark tests is used, which simulate “natural” perturbations – in other words, tests, whose influence on fingerprint images creates perturbed versions thereof, that resemble cases appearing in real-life fingerprint application scenarios. It has to be noted that we do not consider all manipulations even if they would be suitable candidates – *e.g.* , JPEG compression, although contained in the StirMark suite, is not applied here since there have been quite some studies focusing on the effects of JPEG compression in fingerprint recognition [9, 10]. Example images shown have been generated by applying StirMark tests with increasing intensity to a sample image taken from the FVC2004 database DB1 (see Section 4.1).

Additive Noise is introduced to the input image. The amount of noise is adjustable and can range from “none” to “completely random image”, controlled by a single parameter, ranging from 0 to 100. Fig. 1 shows examples for increasing noise content.

This test is intended to simulate noise, that might “naturally” appear in fingerprint sample images. Possible causes for this kind of noise could be actual dust on the contact area during acquisition of the imprint, graining caused by the acquisition equipment

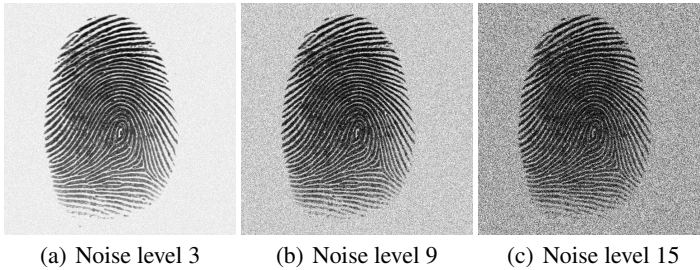


Fig. 1. Examples for the *Additive Noise* test, applied to an image from DB1 (ID 91_2)

itself (sensor noise) or any other kind of systematic error introduced during processing, transmission and/or storage of the collected images (e.g. a grainy surface the latent fingerprint has been taken off can cause noise in forensics).

Median Cut Filtering. This test applies a median cut filter to the input image. The size of the filter mask can be set (height and width of the filter take the same value and only odd-valued dimensions are accepted), the upper limit is a size of 15, thus resulting in a 15×15 filter. Fig. 2 shows examples for medium filter sizes.



Fig. 2. Examples for the *Median Cut Filtering* test, applied to an image from DB1 (ID 91_2)

The *Median Cut Filtering* test is used to simulate smudgy fingerprints, as they are common in real-life applications, for example when the fingertip is too moist during the acquisition by the scanner. The result is a certain amount of blur to the image, but additionally it also corrupts the clarity of the ridge-and-furrow structure of the imprint.

Remove Lines and Columns. This test removes rows and columns from a given image at the specified frequency k – “remove 1 line in every k lines.” It has to be noted that the line removal operation naturally also reduces the size of the output image. Fig. 4 illustrates the effect of this test when applied to fingerprint images.

This test aims to simulate errors in fingerprint images, that occasionally occur during fingerprint acquisitions, in case the scanner is not able to read the fingerprint in its entirety, but misses/skips certain lines. Especially sweep sensors are prone to this kind of complications. Two corresponding examples can be found in Fig. 3.a.



Fig. 3. Examples for distortions from actual acquisition problems



Fig. 4. Examples for the *Remove Lines* test, applied to an image from DB1 (ID 91_2)

Rotation rotates the image by a given angle, the set of angular values that will be inspected in the experiments is $\{-20^\circ, -15^\circ, -10^\circ, -5.5^\circ, -5^\circ, 7^\circ, 7.5^\circ, 13^\circ, 18^\circ, 20^\circ\}$. Examples for rotations of -15° , -5.5° , and 20° can be seen in Fig. 5.

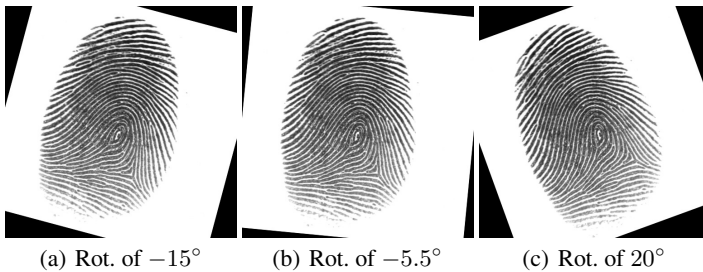


Fig. 5. Examples for the *Rotations Lines* test, applied to an image from DB1 (ID 91_2)

Rotation is a very typical, not to say – omnipresent – challenge for fingerprint matching, as in very few cases a finger will be presented twice in exactly the same orientation

to the contact area during image acquisition. Thus, this test provides the means for comparison of the rotational alignment capabilities of the various fingerprint matchers.

Affine Transformation is a generic manipulation for arbitrary affine image transformations. The user specifies the parameters a, \dots, f of the inverse transformation matrix of the form:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

The application of affine transformations to fingerprint images is intended to simulate distortions of the entire finger imprint, that can appear in real-life situations during the fingerprint acquisition, depending on the way, the finger is pressed on the contact area. As special cases, we consider *shearing* and *stretching*.

Stretching in X-direction is parameterised by setting $b = c = e = f = 0$ and $d = 1$, while configurations 1 - 8 set a to the values $\{1.035, 1.070, 1.105, 1.140, 1.175, 1.210, 1.280, 1.350\}$. Configurations 1, 5, and 8 are shown in Fig. 6.

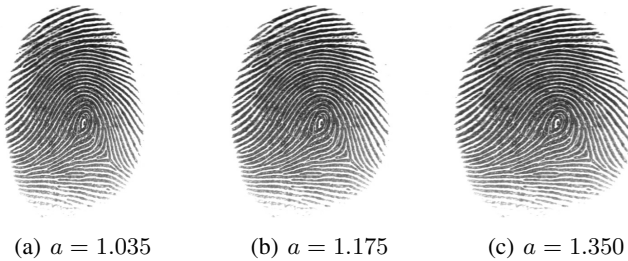


Fig. 6. Examples for the *Affine Transformations – Stretching in X-Direction* test, applied to an image from DB1 (ID 91_2)

A certain *stretching* might appear in the finger imprint, when the amount of force applied while pressing the finger on the contact area is large or larger than usual. Considering the forensic scenario, stretching of a fingerprint appears if the finger was imprinted on a soft or flexible surface.

Shearing in Y-direction is parameterised by setting $b = e = f = 0$ and $a = d = 1$, while configurations 1 - 6 set c to the values $\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. Configurations 1, 4, and 6 are shown in Fig. 7. A *shearing* effect can occur, when the force that is exercised while pressing the finger on the contact area is not exerted perpendicular to this area. For example, when the finger is presented, with the user pushing rather in direction to the upper-right corner of the sensor, than straight downwards.

Small Random Distortions. The StirMark test. Being a combination of several basic manipulations (i.e. random minor geometric distortion followed by resampling and interpolation, a transfer function to emulate analog/digital converter imperfections, global “bending”, high frequency displacement, and JPEG compression), this test originally aims to simulate a resampling process, i.e. the errors introduced when printing an image



Fig. 7. Examples for the *Affine Transformations – Shearing in Y-Direction* test, applied to an image from DB1 (ID 91_2)

and then scanning it again. This test is executed with parameter values $\{0.6, 1.0, 1.4, 1.8, 2.2, 2.6, 3.0, 3.4, 3.8, 4.2\}$, three out of which are illustrated in Fig. 8. The involved image warping is performed both on a global, as well as on a very local level, adding even more to the “natural” and “coincidental” character of the output fingerprint images.

In its character of being a combination of several different image distortions, by applying *the StirMark* test on fingerprint images, we aim to simulate an interaction of various naturally occurring image perturbations: Foremost a random warping of the ridge lines, that in real life would be caused by *e.g.* unevenly distributed pressure exercised on the contact area during acquisition, or if this contact area were to be uneven by itself. Also inaccuracies or errors introduced by the fingerprint scanner can be a source for this type of deformation (two corresponding examples are shown in Fig. 3.b).

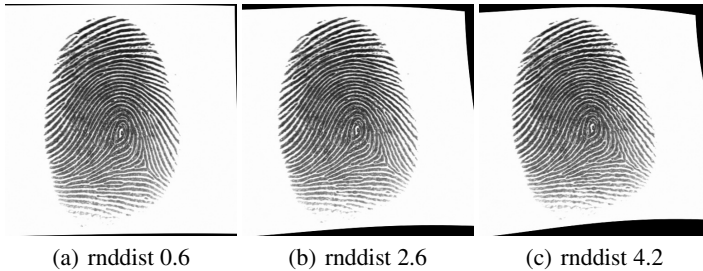


Fig. 8. Examples for the *Small Random Distortions* test, applied to an image from DB1 (ID 91_2)

3 Fingerprint Feature Types and Matching

An interesting question is to ask if a certain type of fingerprint feature extraction method has particular strong points or weaknesses when dealing with a specific type of acquisition condition. In order to get a sensible answer we will consider significantly different types of fingerprint feature extraction schemes, based on the discriminative characteristics fingerprint do contain [2]. On a global level, the overall ridge flow structure with

the embedded singular points (i.e. loops, whorls, deltas) can be perceived. Going into more detail, looking at the ridge and furrow structure in the fingerprint from a more localised point of view, then characteristics like orientation and frequency of the local ridge flow can be observed. On a local level, details of the ridge shapes themselves can be determined. The most important ones are *ridge endings* and *ridge bifurcations* which are subsumed in the term *minutiae*. Looking even closer at a fingerprint, then also diminutive intra-ridge features are detectable – the finger’s sweat pores. However, their respective pattern can only be determined in sufficiently good quality in high-resolution fingerprint images of 1000dpi and above.

Now depending on which type of features is used to determine similarity, fingerprint matching approaches can be classically categorised into one of the following classes:

Correlation-Based Matcher. These approaches use the fingerprint images in their entirety, the global ridge and furrow structure of a fingerprint is decisive. Images are correlated at different rotational and translational alignments, image transform techniques may be utilised for that purpose.

Ridge Feature-Based Matcher. Matching algorithms in this category likewise deal with the overall ridge and furrow structure in the fingerprint, yet in a localised manner. Characteristics like local ridge orientation or local ridge frequency are used to generate a set of appropriate features representing the individual fingerprint.

Minutiae-Based Matcher. The set of minutiae within each fingerprint is determined and stored as list, each minutia being represented (at least) by its location and direction. The matching process then basically tries to establish an optimal alignment between the minutiae sets of two fingerprints to be matched, resulting in a maximum number of pairings between minutiae from one set with compatible ones from the other set.

The minutiae-based approach is the most common and most widely used method for fingerprint matching. Of course, there also exist fingerprint matching algorithms, that combine some or all of the above listed techniques (termed *hybrid*), aiming to combine the particular strong points of each individual approach into a single, more precise matcher (e.g. by using minutiae alignment to compensate for rotation and translation but use ridge-based features for actual matching [11]).

As a representative of the correlation-based matcher type we use a custom implementation of the phase only correlation (POC) matcher [12]. The POC of two images is computed by calculating the normalised cross spectrum (or cross-phase spectrum) from the DFT of the two images. The POC is then obtained by taking the inverse DFT of the normalised cross spectrum. Properties like shift invariance, brightness invariance, and high immunity to noise make POC an interesting candidate for biometric matching since template alignment is eased.

The algorithm first conducts rotational alignment by computing POC for rotated fingerprints in a range of $\pm 20^\circ$ with a step-width of 1° , employing bi-cubic interpolation. The rotated version with the highest POC response is used in subsequent matching. Displacement alignment is performed according to the position of the POC peak as computed before, subsequently probe and gallery images are cropped to the common area of intersection (containing fingerprint data), as the non-overlapping regions lead

to uncorrelated noise in the POC function. Finally, a band-limited version of the POC is computed. The idea is to limit the frequency spectrum involved in matching to only those areas, that are strongly related with the actual fingerprint information – especially the inherent elliptical frequency band originating from the ridge pattern – thereby excluding the interfering components in the high frequency areas. The final matching score is then established by summing up the P highest peaks (Ito *et al.* suggest $P = 2$, while we found $P = 1$ to perform better for two out of three databases) of the band-limited POC function. It has to be noted that fingerprint enhancement as used in the subsequent algorithm [13] also improves matching results for two out of three databases and is used correspondingly.

As a representative of the ridge feature-based matcher type we use a custom implementation of the fingercode approach (FC) [14], in particular we follow improvements as suggested by Ross *et al.* [15, 11] and de Sa *et al.* [16] which avoid the usage of a circular tessellation around a core point, and partially apply the fingerprint enhancement strategy as suggested by Hong *et al.* [13].

As a first stage, normalisation is applied by pixel-wise adjusting the gray-levels to obtain an image with pre-specified mean and variance. Based on the output of a Marr-Hildreth operator, a least square estimate of the local ridge orientation in blocks of 16×16 pixels is established subsequently. A low-pass filter is used to smoothen the result which is called orientation image. The normalised fingerprint image and the orientation image are used to create the frequency image, representing the local ridge frequency. In an oriented window the *x-signature* is calculated per block by projecting the respective gray-level values of the normalised image onto the length of the window which is placed in a direction orthogonal to the ridge orientation of the block. The frequency can be determined by taking the reciprocal of the average distance between peaks in the x-signature. Interpolation of invalid blocks (i.e. those where the x-signature did not form a discrete sinusoidal-shape wave) is done with a discrete Gaussian kernel. Contrasting to the original papers, we use high frequency together with a check of an admissible dynamic range as criteria to determine blocks which actually represent useful fingerprint texture, thus declared foreground blocks to be further used in matching.

For actual feature extraction, a Gabor filter bank consisting of eight separate Gabor filters, each oriented at a different constant angle is convolved with the image, examining the varying responses of the ridges and furrow structure to the differently oriented filters. This results in eight distinct filtered images for each of which a standard deviation in a 16×16 neighbourhood is computed per pixel, the union over all eight images is called *Standard Deviation Map*. If the fingerprint image is intended for database registration (i.e. enrolment), this map is subsampled by a factor of 16 to generate ridge feature images, the union of which is called *Ridge Feature Map*. Translational alignment is achieved by computing correlations among differently displaced ridge feature images in the Fourier domain and compensating the shift identified by maximal correlation, rotational alignment is done by storing ridge feature maps of rotated fingerprint versions in an angular range of $[-20^\circ, +20^\circ]$ with a step-width of 1° , and again taking the version with the maximal correlation. The matching score is obtained by computing Euclidian distance among aligned ridge feature map entries.

As a representative of the minutiae-based matcher type we use *mindtct* and *bozorth3* from the “NIST Biometric Image Software” (NBIS) package (available at <http://fingerprint.nist.gov/NBIS/>) for minutiae detection and matching, respectively. *mindtct* generates several image quality maps and binarises the fingerprint images as a first step. Subsequently, minutiae are detected in admissible areas by detecting specified pixel patterns, followed by false minutiae removal and minutiae quality assessment. *bozorth3* is designed to be rotation and translation invariant and provides a matching score based on traversing certain inter-fingerprint compatibility tables.

A comparison of the three fingerprint recognition schemes with respect to recognition performance on the three “natural” FVC2004 databases (without StirMark manipulations being applied) is provided in the subsequent subsection.

4 Experiments

We first provide details about the employed FVC2004 data set. Subsequently, experimental results are presented and discussed, covering questions of robustness of recognition accuracy against various StirMark manipulations and in particular a comparison of the behaviour of the different feature types in that respect.

4.1 Experimental Settings

We employ three out of four databases provided for the FVC2004 [17] as shown in Table 1, each with 500dpi resolution (DB3 with 512 dpi).

Table 1. Details on the fingerprint images in the three employed FVC2004 databases and EERs for the considered fingerprint recognition schemes when applied to the original, “undistorted” sample image databases

	Sensor Type	Model	Image Size	NBIS (%)	FC (%)	POC (%)
DB1	Optical	CrossMatch V300	640 × 480	14.81	12.54	22.60
DB2	Optical	Digital Persona <i>U.are.U 400</i>	328 × 364	11.12	9.60	9.69
DB3	Thermal Sweep	Atmel <i>FingerChip</i>	300 × 480	6.68	8.98	15.07

It does not make too much sense to include the fourth dataset of FVC2004 (although it would be possible in principle of course) since it consists of synthetically generated fingerprint images (SFInGe [4] was used). In order to model real-life distortions for these data, a much more sensible way would be to apply respective distortion “operations” already during the generation process of the synthetic fingerprints instead of applying them ex post to the final data.

The procedure for performance evaluation is basically the same in all FVCs, from 2000 to 2006. We follow this specification by conducting all genuine tests and the required impostor tests for DB1, DB2, and DB3, thereby obtaining FNMR and FMR as required. Finally, equal error rate (EER) is determined and used as measure for recognition accuracy to compare different settings. Table 1 also shows the result when applying

the three considered feature types to the FVC2004 test data *without* having applied any StirMark manipulations, but already within the StirMark framework.

It can be clearly seen that the ranking of the three feature types is heavily dependent on the used dataset. Each type of feature extraction is ranked first for a single dataset, while only FC is never ranked third. Only when considering DB2, the performance is really close.

4.2 Experimental Results

Fig. 9.a shows the influence of additive noise on recognition performance considering DB2. Especially for a higher degree of noise content FC clearly shows the best robustness, POC is also better compared to NBIS for this setting, but clearly inferior to FC.

Noise Level	NBIS (%)	FC (%)	POC (%)	Noise Level	NBIS (%)	FC (%)	POC (%)
unperturbed	11.12	9.60	9.69	unperturbed	6.68	8.98	15.07
03	10.86	11.85	10.65	03	7.05	9.25	15.28
07	15.03	14.22	14.36	07	7.19	10.50	15.16
11	20.54	17.74	20.22	11	7.08	14.79	15.71
15	30.78	21.80	26.94	15	7.91	24.99	17.46
(a) DB2				(b) DB3			

Fig. 9. EERs for *Additive Noise* test

In Fig. 9.b we see that the situation changes when considering a different dataset, DB3 in this example. NBIS recognition results are hardly affected even by a significant amount of noise, entirely contrasting to the results obtained on DB2. Also, the ranking between FC and POC is swapped, POC results are also quite stable while FC recognition accuracy severely suffers from high noise content.

It is also interesting to note that on unperturbed data, FC is superior to POC, while POC is getting clearly superior under the influence of more noise being present. This effect underlines the need for systematic robustness testing in a feature-type comparative manner.

This example overall shows that even feature-type ranking results with respect to robustness achieved on a specific database cannot be generalised but need to be verified for each single dataset. This nicely illustrates the general need for systematic testing and evaluation tools.

Fig. 10.a shows robustness results with respect to median cut filtering on DB1, which is another example that ROC performance on unperturbed data cannot predict robustness properties. While NBIS is clearly superior to POC on the original data, it gets inferior when introducing significant mean cut filtering.

In Fig. 10.b very high stability of NBIS and FC against line removal is shown, while POC suffers considerably in case the amount of missing lines is increasing.

Filter Size	NBIS (%)	FC (%)	POC (%)	k	NBIS (%)	FC (%)	POC (%)
unperturbed	14.81	12.54	22.60	unperturbed	11.12	9.60	9.69
03	15.50	12.90	23.63	90	11.04	9.71	10.00
05	17.69	13.52	24.92	70	11.60	9.73	10.24
07	32.17	16.55	30.71	40	11.99	9.47	11.18
09	46.88	28.26	38.11	20	12.92	9.97	14.75
(a) Median Cut Filtering				(b) Remove Lines			

Fig. 10. EERs for robustness tests conducted on sample image databases DB1 and DB2, respectively

Rotation	NBIS (%)	FC (%)	POC (%)	Rotation	NBIS (%)	FC (%)	POC (%)
unperturbed	14.81	12.54	22.60	unperturbed	11.12	9.60	9.69
-15	13.00	14.74	24.34	-15	11.00	14.13	14.22
-5.5	12.94	12.90	22.67	-5.5	11.28	12.04	10.89
13	13.05	13.63	23.79	13	10.59	13.57	13.01
20	13.41	15.44	26.18	20	10.94	16.27	18.08
(a) DB1				(b) DB2			

Fig. 11. EERs for *Rotation* test conducted on sample image databases

One of the most important robustness issues is fingerprint rotation, since this effect is omnipresent in sample acquisition. Fig. 11 compares the results for DB1 and DB2.

The first thing to note is the excellent robustness of NBIS against rotation for both datasets. FC and POC are both affected, however, the extent of result degradation is much larger for DB2 as compared to DB1. For example, based on experimental results obtained on DB1, one would have predicted $EER \approx 12.5$ for POC on DB2 under 20° rotation, in fact we observe EER to be 18.08 which almost doubles EER as compared to not manipulated data. Again, the need for dedicated testing for each sensor type is confirmed.

Affine transformations also model a class of very important acquisition conditions. Table 2 shows the catastrophic effect of stretching in a single dimension only. No feature extraction type can handle this type of distortion in a sufficient degree. Obviously, there is need to introduce stretching robustness into feature sets.

Shearing robustness as illustrated in Fig. 12.a is shown to be much better as compared to stretching. Apart from very strong distortions, NBIS and FC can handle shearing quite well, while POC exhibits steadily decreasing EERs for an increasing amount of shearing.

Finally, robustness results against a combination of manipulations are shown in Fig. 12.b. These rather localised distortions can be handled quite well by NBIS and FC, at least up to some medium extent. POC, similar to its sensitivity against affine transformations, exhibits steadily increasing EERs for increasing strength of the distortions.

Table 2. EERs for *Affine Transformations – Stretching in X-Direction* test conducted on sample image database DB3

Configuration	NBIS (%)	FC (%)	POC (%)
unperturbed	6.68	8.98	15.07
2	7.70	10.99	23.89
4	11.13	13.90	31.72
6	14.14	17.98	37.20
8	23.38	25.79	42.69

Configuration	NBIS (%)	FC (%)	POC (%)	Factor	NBIS (%)	FC (%)	POC (%)
unperturbed	14.81	12.54	22.60	unperturbed	11.12	9.60	9.69
1	13.85	12.57	22.64	0.6	10.78	12.42	10.89
2	13.88	12.76	24.79	1.0	11.35	12.34	11.49
3	14.88	13.30	27.43	1.8	11.75	12.40	13.23
4	16.26	14.15	29.90	2.6	12.57	12.61	16.34
5	17.96	14.71	37.73	3.4	13.23	13.57	19.00
6	21.46	15.82	40.22	4.2	14.82	14.05	21.96

(a) *Shearing in Y-Direction*(b) *Small Random Distortions***Fig. 12.** EERs for robustness tests conducted on sample image databases DB1 and DB2, respectively

5 Conclusion

We have employed the StirMark benchmark testsuite to generate large scale test data to assess robustness of fingerprint recognition schemes in various acquisition conditions. Experimental results confirm a significant variability of robustness properties across different types of fingerprint feature extraction schemes **and** across different datasets considered. For example, we have observed significant impact of Median Cut Filtering and Affine transforms like Stretching and Shearing for almost all feature types and datasets, while Noise Insertion is tolerated well by some feature types (FC and POC on DB1 and NBIS and POC on DB3) but leads to considerable impact for all techniques on DB2. As compared to our previous work [7] where only minutiae-based matching schemes have been compared, we see even larger variability in this present work when comparing matching results relying on entirely different feature extraction schemes.

These results underline the need for a standardised tool in fingerprint recognition robustness assessment, which is able to simulate a wide class of acquisition conditions, applicable to any given dataset.

While we have motivated the interpretation of several image manipulations contained in the StirMark benchmark as being closely related to a wide class of fingerprint acquisition conditions (including some forensic settings), these experiments only represent a first step. In fact, the aim is to establish a benchmark explicitly designed for systematic fingerprint recognition robustness evaluations, where these current StirMark based results can serve as first guidelines to model actual fingerprint acquisition conditions more accurately.

References

- [1] Alonso-Fernandes, F., Bigun, J., Fierrez, J., Fronthaler, H., Kollreider, K., Ortega-Garcia, J.: Fingerprint recognition. In: Petrovska-Delacretaz, D., Chollet, G., Dorizzi, B. (eds.) *Guide to Biometric Reference Systems and Performance Evaluation*, pp. 51–88. Springer (2009)
- [2] Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: *Handbook of Fingerprint Recognition*, 2nd edn. Springer (2009)
- [3] Noviyanto, A., Pulungan, R.: A comparison framework for fingerprint recognition methods. In: *Proceedings of the 6th SEAMS-UGM Conference (Computer, Graph and Combinatorics)*, pp. 601–614 (2011)
- [4] Cappelli, R.: Synthetic fingerprint generation. In: Maltoni, D., Maio, D., Jain, A., Prabhakar, S. (eds.) *Handbook of Fingerprint Recognition*, 2nd edn., pp. 271–302. Springer (2009)
- [5] Kutter, M., Petitcolas, F.A.P.: Fair evaluation methods for image watermarking systems. *Journal of Electronic Imaging* 9(4), 445–455 (2000)
- [6] Meerwald, P., Pereira, S.: Attacks, applications and evaluation of known watermarking algorithms with Checkmark. In: Wong, P.W., Delp, E.J. (eds.) *Proceedings of SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents IV*, vol. 4675, pp. 293–304. SPIE, San Jose (2002)
- [7] Hämmerle-Uhl, J., Pober, M., Uhl, A.: Towards standardised fingerprint matching robustness assessment: The stirmark toolkit – cross-database comparisons with minutiae-based matching. In: *Proceedings of the 1st ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec 2013)*, Montpellier, France (to appear June 2013)
- [8] Petitcolas, F.A.P., Anderson, R.J., Kuhn, M.G.: Attacks on copyright marking systems. In: Aucsmith, D. (ed.) *IH 1998. LNCS*, vol. 1525, pp. 218–238. Springer, Heidelberg (1998)
- [9] Funk, W., Arnold, M., Busch, C., Munde, A.: Evaluation of image compression algorithms for fingerprint and face recognition systems. In: Cole, J., Wolthusen, S. (eds.) *Proceedings from the Sixth Annual IEEE Systems, Man and Cybernetics (SMC) Information Assurance Workshop*, pp. 72–78. IEEE Computer Society (June 2006)
- [10] Mascher-Kampfer, A., Stögnner, H., Uhl, A.: Comparison of compression algorithms’ impact on fingerprint and face recognition accuracy. In: Chen, C., Schonfeld, D., Luo, J. (eds.) *Proceedings of SPIE Visual Communications and Image Processing (VCIP 2007)*, vol. 6508, pp.650810–1 – 65050N–10. SPIE, San Jose (January 2007)
- [11] Ross, A., Jain, A.K., Reisman, J.: A hybrid fingerprint matcher. *Pattern Recognition* 36(7), 1661–1673 (2003)
- [12] Koichi, I., Hiroshi, N., Koji, K., Takafumi, A., Tatsuo, H.: A fingerprint matching algorithm using phase-only correlation. *IEICE Transactions on Fundamentals* E87-A(3), 682–691 (2004)

- [13] Hong, L., Wan, Y., Jain, A.: Fingerprint image enhancement: Algorithm and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8), 777–789 (1998)
- [14] Jain, A.K., Prabhakar, S., Hong, L., Pankanti, S.: Filterbank-based Fingerprint matching. *IEEE Transactions on Image Processing* 9(5), 846–859 (2000)
- [15] Ross, A., Reisman, J., Jain, A.K.: Fingerprint matching using feature space correlation. In: Tistarelli, M., Bigun, J., Jain, A.K. (eds.) *ECCV 2002*. LNCS, vol. 2359, pp. 48–57. Springer, Heidelberg (2002)
- [16] de Sá, G., de Alencar Lotufo, R.: Improved fingercode matching function. In: *SIBGRAPI*, pp. 263–272 (2006)
- [17] Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.K.: FVC2004: Third Fingerprint Verification Competition. In: Zhang, D., Jain, A.K. (eds.) *ICBA 2004*. LNCS, vol. 3072, pp. 1–7. Springer, Heidelberg (2004)

Achieving Anonymity against Major Face Recognition Algorithms

Benedikt Driessen and Markus Dürmuth

Ruhr-University Bochum

{benedikt.driessen,markus.duermuth}@rub.de

Abstract. An ever-increasing number of personal photos is stored online. This trend can be problematic, because face recognition software can undermine user privacy in unexpected ways. Face de-identification aims to prevent automatic recognition of faces thus improving user privacy, but previous work alters the image in a way that makes them indistinguishable for both computers and humans, which prevents a wide-spread use.

We propose a method for de-identification of images that effectively prevents face recognition software (using the most popular and effective algorithms) from identifying people, but still allows human recognition. We evaluate our method experimentally by adapting the CSU framework and using the FERET database. We show that we are able to achieve strong de-identification while maintaining reasonable image quality.

Keywords: Face recognition, anonymity, de-identification.

1 Introduction

The number of personal photos that is available online has been rapidly increasing over the past years.¹ This development is driven by the wide availability of (stationary and mobile) high-speed Internet, cheap electronic storage, and ubiquitous digital cameras on the one hand, and a strong trend towards social networks and managing friends online on the other hand. Recently, some services publicly announced the deployment of face recognition software on the stored images (see, e.g., [12]). Face recognition software can be beneficial for the user, as it helps finding and tagging friends in pictures. However, it can also be used to gather additional information about friendship-relations (i.e., the social graph), even relations the user deliberately did not share with everybody.

People often try to separate some groups of people, e.g., personal friends and work colleagues and reserve an online profile for personal friends. Note that Facebook considers the social graph as public information that can even be queried via a special API [13], and in general one can easily imagine external services crawling the image database. A rather drastic example includes predicting a

¹ As an example, Facebook hosted 10 billion photos in Oct. 2008, receiving about 250 million new photos a day [11]. Flickr hosted 4 billion photos in Oct. 2009 [14].

persons sexual orientation from the social graph [21]. So a person might want to hide (parts of) their social graph in order to protect such information, but still might want to post images. (We stress that automated extraction of the social graph is a much bigger problem than manual extraction, because of the large-scale extraction of information that becomes possible.)

In the past few months, the criticism of automated face recognition, and in particular Facebook as the most prominent example, has increased, e.g., from the Electronic Privacy Information Center (EPIC), which considers Facebook’s handling of personal data a violation of European privacy law (see, e.g., [8,10]).

In this work we will demonstrate a system that effectively protects the anonymity, in particular the social graph, of a user by thwarting face recognition software, while still allowing humans to identify faces and keeping the visual changes to the image small. In particular, previous work [23] mapped several “similar” faces to the same “average”-face (see Section 2 for details), thus the resulting face-images are indistinguishable for both computers *and* humans. While their approach provides strong security guarantees, it constitutes a strong visual alteration of the face images, and in particular prevents humans from distinguishing persons. We believe that this is too intrusive for wide-spread use.

1.1 Our Contribution

We propose a system to anonymize face images in a way that retains more details of the original image than previous work, thus allowing a human to still identify the person from the image, and works against all major classes of face recognition algorithms.

We exploit the fact that face recognition algorithms reduce the dimensionality of the data of the face image in order to reduce noise and improve speed. We show how to manipulate these relevant parts of the image to fool all important face recognition algorithms. We aim at a slightly weaker form of anonymity than k -anonymity (see Section 4), where an attacker’s confidence in correct identification is small. This is well suited for the two main threats we are considering: first, we want to prevent extraction of the relationships (the social graph), e.g., by evaluating who is present on a sufficiently large number of pictures of one specific person; second, we want to prevent automated tagging of people on pictures.

Of course, when humans can recognize a face, then the face will not be anonymous in a strict sense. Studies show [2] that the “price” most users are willing to pay for privacy is pretty low, so we hope that by providing a reasonable image quality, this approach will find more acceptance by users, while still preventing automated extraction of information and thus providing a reasonable level of security.

2 Eigenface-Based Face Recognition

We introduce the basic terminology, face recognition algorithms based on Principal Component Analysis (PCA), and show the basic idea how we can manipulate

these algorithms. We consider the classical Eigenface-algorithm [32], which is interesting because it constituted a breakthrough in recognition performance at the time of its invention, and still provides very competitive performance for images taken in a moderately controlled environment. Also, it forms the basis for a wide range of algorithms, including Linear Discriminant Analysis (LDA), which can be applied after PCA, and the Bayesian classifier, both of which we present also.

2.1 Brief Introduction to Face Recognition Terminology

The task of face recognition can be described in simple terms: Given a set of images of a number of (known) persons (*gallery images*), and given an image for an unknown person (*probe image*), decide which person from the gallery is shown on the probe image.

The process is usually divided into the following steps: First, one needs to find the approximate position of the face in the image, this is called *face-detection* [35] and a separate line of research. Most work on face recognition considers this job to be completed before². Second, images are *normalized*, which usually includes an affine transformation to align the eyes, histogram equalization, and sometimes masking of the background. Third, in *feature extraction* algorithm-dependent features the probe image are extracted. Representing an image by a set of features can be seen as a step in data reduction that aims at extracting a compact but discriminating description of the image. Ideally, the output of this step is at the same time robust against changes in posture, lightning, face expression, etc. Finally, the pre-processed probe image is *matched* against gallery images. Different distance functions, measuring the similarity of two images, can be used here, we will see Euclidean distance, weighted Euclidean distance, and angles in the sequel.

The output of a face recognition algorithm is a list of identifiers, where the algorithm estimates that the first identifier (e.g. name) is the most likely one, matching the subject on the probe image. The list is typically ordered by descending probability (or ascending distance, depending on the distance measure). The performance of face recognition algorithms is usually measured in *rank curves* (see Figure 1), where on the x -axis we plot ranks and on the y -axis recognition rates. When the curve passes point (x, y) this means that for a fraction of y probe images, the correct identifier was contained in the first x suggestions of the recognition algorithm.

2.2 Classical Eigenfaces

Given L training pictures $\mathbf{u}_1, \dots, \mathbf{u}_L$, written as vectors with p pixels each, we compute the (point-wise) average image $\mathbf{m} := \frac{1}{L} \sum_{k=1}^L \mathbf{u}_k$ and compute the mean-subtracted images $\mathbf{u}'_k := \mathbf{u}_k - \mathbf{m}$. Write $U = [\mathbf{u}'_1, \dots, \mathbf{u}'_L]^t$ for the matrix

² Commonly used face image databases such as the FERET database (see Section 5.2) annotate the images with the eye coordinates.

of images. Then the (empirical) covariance matrix is written as $C = U^t \cdot U$. The matrix U has size $p \times L$ and the matrix C has size $p \times p$. Let $\lambda_0, \dots, \lambda_{N-1}$ be the N largest Eigenvalues with associated Eigenvectors $\mathbf{e}_0, \dots, \mathbf{e}_{N-1}$. Using the orthonormal vectors $\mathbf{e}_0, \dots, \mathbf{e}_{N-1}$, we calculate a feature vector (in what is called *face-space*, the space spanned by Eigenvectors), by simply projecting an image \mathbf{u} on the Eigenvectors

$$s_k = \mathbf{e}_k^t \cdot (\mathbf{u} - \mathbf{m}) \quad \text{for } k = 0, \dots, N - 1, \quad (1)$$

where s_k is the k -th component of the projection of \mathbf{u} in face-space. Calculating a feature vector for every gallery image as well as the probe image, and using Euclidean distance to find the gallery image that is closest to the probe image, we get the original Eigenfaces face recognition algorithm.

Instead of Euclidean distance we can also use different distance measures, such as MahCosine, which measures the angle of the projections \mathbf{m}, \mathbf{n} of two vectors \mathbf{u}, \mathbf{v} into Mahalanobis space see [4] for details and more examples. In the following we use this measure because it outperformed other measures for purely PCA-based recognition methods [18] and in order to enhance our understanding of the effectiveness of de-identification methods in the presence of non-Euclidean distance measures.

2.3 Modifying Projections (in Image Space)

Our basic idea is to manipulate the face images in such a way that the projection onto the face-space changes, while hopefully making minimal changes.

Given the input image \mathbf{u} as a row-vector and a set of orthogonal and normalized vectors $\mathbf{e}_0, \dots, \mathbf{e}_{N-1}$ (the selected principal eigenvectors as computed by the PCA, spanning the face-space), consider Equation (1). By adding $\Delta_k \cdot \mathbf{e}_k$, a multiple of the k -th Eigenvector, to an image \mathbf{u} , we can arbitrarily change the k -th component of the projection; an easy calculation shows that

$$((\Delta_k \cdot \mathbf{e}_k^t + \mathbf{u}^t) - \mathbf{m}^t) \cdot \mathbf{e}_k = \Delta_k + s_k. \quad (2)$$

Another easy calculation shows that we can adjust several components by simply adding several Eigenfaces, because the $\mathbf{e}_1, \dots, \mathbf{e}_{N-1}$ are pairwise orthogonal.

Compared with previous work [23], the image quality of this approach is better, because we do not alter the information outside the Eigenspace.

2.4 PCA+LDA and Bayesian Face Recognition

LDA can be applied directly to the input data [1], but was found to be more effective when applied after a PCA transform [37]. While slightly different parameters are optimal for the PCA transform when used as preprocessing stage for LDA, one can see that the same techniques that we developed in Section 2.3 is applicable.

The Bayesian face recognition algorithm [19,31] is different from most other algorithms in that it breaks down face recognition into a series of classification

problems: in order to recognize a face, the algorithm iterates over all stored *persons* (not faces), and for each decides if it is the correct person or not. The central idea is that it tries to decide if the difference of two faces is in one of two classes, either *inter-personal* (Ω_I) or *extra-personal* (Ω_E), where in preprocessing the algorithm learns what are “typical variations” for the difference of two images of the same face, and for two images of different faces. In this training, again, PCA is applied to the input to reduce dimensionality, with the difference that here it is applied to the (point-wise) difference of two images. This yields Eigenfaces different from the above, but a closer look reveals that again methods similar to those from Section 2.3 can be applied.

3 Elastic Bunch Graph Mapping Face Recognition

Elastic bunch graph mapping [33] is another algorithm for face recognition that fared very well in the FERET tests [28,24]. What makes this algorithm particularly interesting is that it is fundamentally different from the previous algorithms: it is not based on PCA and is commonly classified as feature-based instead of holistic, i.e., it bases its decision on particular local features (eyes, mouth, ...) instead of a holistic view of the face.

3.1 Algorithm Description

We give a brief overview of the algorithm, for more details we refer the reader to [33,5]. A central tool for EBGGM are *Gabor wavelets*, convolution kernels which are plane waves bounded by a Gaussian envelope function. Let ψ_j be a Gabor wavelet, then the convolution at point x with the image \mathbf{u} is given by

$$J_j(x) = \int \mathbf{u}(x')\psi_j(x' - x)dx', \quad (3)$$

where $J_j(x)$ is a complex value and the index j ranges over 40 values for 8 orientations and 5 frequencies. Convolution of a fixed point of an image with Gabor wavelets of different orientation and frequency is called a *jet*; intuitively, a jet contains a reduced description of the surrounding of that point. Gabor wavelets are robust against a number of variations and are motivated by human vision research.

For the faces, one defines a set of *fiducial points*, such as pupils, corners of the eyes or mouth, and top and bottom of the ears. The nodes of these graphs are labeled with a jet. Initially, for a small set of faces, the fiducial points are extracted by hand, and the jets are computed. When presented with a new face, the information extracted above is used to automatically fit the above graph to a new face: First, the rough position of the face is determined by matching the average of all above graphs onto the probe image. Then the graph that fits best is selected, allowing for small displacements and scaling of the graph, followed by successively relaxing the graph geometry and adapting the points individually.

The graph is fitted on every image in the gallery, and the resulting vector of jets is stored. For a probe image, the closest match with a gallery image is computed as the mean of the individual jet similarities, where jet similarities are computed as “normalized vector product”.

3.2 Modifying Jets (in Image Space)

Our basic idea is that by adding appropriate multiples of a wavelet $\alpha \cdot \psi_j$ to an image \mathbf{u} at position x , we can change the value of the convolution with this particular wavelet at the specific position. This can be verified by a simple calculation.

One difference to the situation for PCA is that these changes are not independent of each other: modifying one jet value also changes other values for this jet, and several jets are close enough that other jets are influenced as well. For this reason we proceed iteratively as follows:

1. Do 150 times, over all jets and wavelets:
 - (a) Find the maximum difference between the current and target value
 - (b) Add Gabor wavelet to bridge 1/5-th of the distance
2. Over all jets and wavelets (wavelets with large radius first):
 - (a) Add Gabor wavelet to bridge 1/20-th of the distance

We established these parameters empirically and found them to work well. As for PCA-based techniques, it is not necessary to set the image to be equal to the target image, because probe and gallery images have a certain distance anyway.

4 Achieving Anonymity

Next, we describe how we utilize the approaches from the previous sections to anonymize face images.

4.1 k -Anonymity

An established definition of security against identification is *k-anonymity* [29,30], see also the notion of *anonymity sets* [25]. For face recognition, a person remains *k*-anonymous if the face recognition algorithms cannot narrow the person down to a set of less than *k* persons.

For our envisioned targets, weaker forms of anonymity are sufficient. There are two scenarios we would like to protect against: First, automated tagging of persons on uploaded pictures (note that we are not targeting the automated proposal of persons to tag, because this still contains human interaction and thus is only making tagging easier...); second, automated derivation of friendship-relations from a large set of pictures. For both scenarios, weaker privacy guarantees suffice. Along with this weaker privacy guarantee comes a large improvement in image quality (as perceived by a human), so we hope that our system will lead to more wide-spread usage of privacy-protecting systems.

4.2 Anonymizing Face Images

Here we describe our approach to face anonymization, which builds on the methods we have developed above.

1. We select a partition of the involved persons such that each set has at least k members. We choose them by picking a random face image and selecting the $k - 1$ nearest images (of distinct persons) according to a suitable distance measure (we will elaborate on this in Section 5) and we call this set a *cluster*.
2. For every cluster we project each of the k images and compute the average projection (wrt. PCA).
3. All images in a cluster are modified (as described in Section 2.3) to have the same, averaged projection. However, we may also choose to adjust the projection by a fraction $\sigma \in \mathbb{R}$ with $0 < \sigma < 1$ of the difference between an images actual projection and the average projection. The parameter σ was determined experimentally, see Section 5.4.
4. Next, all images in a cluster are modified wrt. EBGM (as described in Section 3.2) to resemble the average face for that cluster. We apply EBGM-modifications after PCA-modifications, because EBGM-modifications are local, whereas PCA-changes influence the entire image (see classification of EBGM as a local feature-approach, as opposed to PCA being a holistic approach).

A central observation is that we do not need to change the projections of PCA in face-space to the actual average, but it's sufficient to go some way in that direction. The reason is that probe image and gallery image of the same person are already quite some distant apart, so moving partially in the correct direction suffices. We will show experiments substantiating this claim.

5 Experiments

In this section we present extensive experiments that substantiate our privacy claims. We used the CSU framework of face recognition algorithms to test our results on a subset of 1000 images of the FERET database. We performed de-identification experiments for all three classes of algorithms (Eigenface-based, Bayesian, EBGM) and finally realized a synthesis of our results. The FERET database contains images of faces only. Thus our experiments omit some pre-processing steps that would be required to apply this idea in a real-world scenario, but these are well-known and add little insight to our goal of de-identification.

5.1 The CSU Framework

The CSU framework [6,4] was created at Colorado State University to facilitate the comparison of different algorithms, and is available for free for research. The framework runs on UNIX/Linux systems, the source code is available and therefore easily adaptable. The current version 5.1, published July 2010, supports the

following face recognition algorithms: (i) Classical Eigenfaces (i.e., PCA) with different distance measures, e.g. Euclidean, MahCosine, etc., (ii) LDA+PCA, also with different measures, (iii) Bayesian classification with the MAP and ML classifier, and (iv) Elastic Bunch Graph Mapping. Details about the specific implementations can be found in a series of papers, most notably [31,5,34,3]. The framework utilizes the FERET dataset and allows to easily measure the recognition performance. Furthermore, due to its modularity, the framework can be extended by new algorithms in order to benchmark these against already known methods. Although not intended, the framework can easily be adapted to allow benchmarking the de-identification performance of our algorithms.

5.2 The FERET Database

The FERET program [26,27] started 1993 and ran until 1997. It was sponsored by the Department of Defense Counterdrug Technology Development Program through the Defense Advanced Research Products Agency. Its primary mission was to develop automatic face recognition algorithms that could be employed to assist security, intelligence and law enforcement personnel. The FERET dataset was assembled to support government monitored testing and evaluation of face recognition algorithms using standardized tests and procedures. The final set of images has 3300 images from 1200 persons, with varying mimical expressions, from different dates, under semi-controlled conditions. The dataset is available for research related to face recognition.

5.3 Scope and Conduct of Experiments

For our experiments we used the FA and FB subsets of the FERET database, each containing one facial expression of 1195 subjects. We ran the experiments on a random subsets of 500 subjects each. The FA set served as our gallery, face recognition performance figures are given in terms of successively matching subjects from the probe set FB to their alternate image in FA. In our experiments, we apply our de-identification methods against both sets of images which implies that, although all gallery images have been de-identified, the face recognition system still has identifiers associated to them. This is not only a necessary prerequisite for identification, but also realistic when considering that most social networks encourage users to manually identify persons on photos, thus adding images to the gallery. Several experiments were performed to validate our de-identification approaches and identify a reasonable set of parameters. First, we tested our de-identification method for each recognition algorithm individually; the results are shown in Section 5.4 and Section 5.5. Then we combined these preliminary findings, applying both techniques at the same time, for our final results in Section 5.6.

All experiments were conducted with the default configuration of the CSU framework, the only exception concerns the normalization step. EBGM uses a pre-processing procedure which is different from the normalization required by all other recognition methods. Our experiments were conducted entirely on the

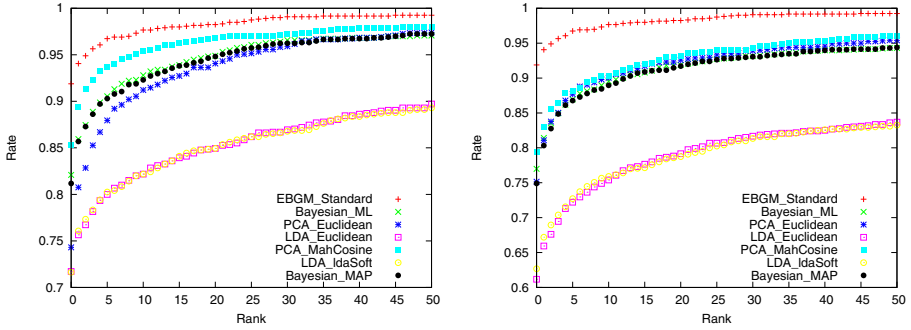


Fig. 1. Baseline performances of the algorithms (prior to de-identification), in the first image with the original normalization parameters, in the second image with identical normalization for all algorithms

data-set generated by EBGM pre-processing. The reason is our two-fold strategy for de-identification. Ideally, all PCA-related modifications are applied to images normalized for PCA-based methods, then the result is transformed back to the original image which is subsequently pre-processed for EBGM. Finally, the EBGM modifications are applied. However, this is tedious work that adds little insight to the interesting questions, so we chose to use the EBGM data-set for all methods. The impact of this strategy on recognition rates is shown in Figure 1. It displays the rank curves for all of the algorithms targeted by us. The left graph shows recognition performance (prior to de-identification) in the default configuration, i.e. where each class of algorithms operated on specifically normalized images. The right graph shows a slight degradation in recognition performance, which is due to non-optimal normalization, which seems to affect LDA-based methods most. The performance results of our de-identification methods, which are expressed as rank curves as well, are to be understood in relation to the right graph.

5.4 Experiments for Eigenface-Based Face Recognition

Experiments to determine the effectiveness of our de-identification method against Eigenface-based methods are parameterized by k , which is the size of the anonymity clusters, and $0 < \sigma < 1$ which is a factor weighting the addition of modifications. In a first series of experiments, presented in this section, we tested and validated our de-identification method against face recognition using PCA and LDA+PCA. For both, we tested different distance measures, which gives us assurance that the proposed method is robust against changes in this metric.

In a first series of experiments, we targeted each of these methods individually by building clusters of persons according to the same face recognition method, because we wanted to learn how sensitive our method is to how exactly the clusters are chosen. For our first experiment we consider de-identification with

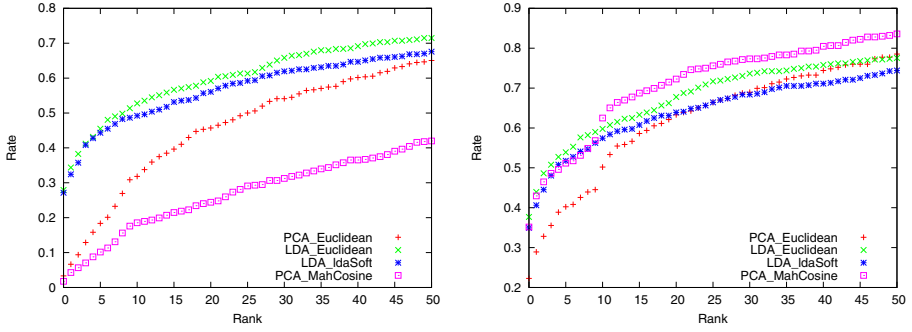


Fig. 2. Eigenface-based recognition after de-identification with $k = 10$ and $\sigma \in \{1, 0.85\}$ clusters obtained from the same measure



Fig. 3. Comparison of original image (left) with image after de-identification (right) with parameters $k = 10, \sigma = 0.85$

$\sigma = 1, k = 10$ and measure the recognition rate of all four methods, again. Comparing the left graph in Figure 2 with the original results from Figure 1, we see that the two PCA algorithms perform quite similar with an extremely low rank-0 recognition rate, and the the PCA+LDA algorithms perform better, yet still much lower than without de-identification. Also, we can see that the performance of the MahCosine distance measure, which is very accurate without de-identification, decreases disproportionately strong.

In a second series of experiments, we determined a suitable weighting factor σ . The lower we choose σ , the less an image is actually altered (thus not completely bridging the distance to the cluster’s target image) which consequently yields a better image quality. We found that de-identification with $k = 10$ and $\sigma = 0.85$ works well, and additionally this parameter balances the recognition rate for the four algorithms, as can be seen in the right graph in Figure 2. Figure 3 shows the visual effects of de-identification for $k = 10, \sigma = 0.85$ for all four recognition methods in comparison with the original image. In both cases the person is clearly recognizable. The strongest effect on the pictures can be seen at the line between the person’s hair and the background. This effect hardly affects the recognizability of a face, and can most likely be avoided by restricting the Eigenfaces to the actual face, as in the usual preprocessing for PCA. Also, the modified images look somewhat lighter than the original images.

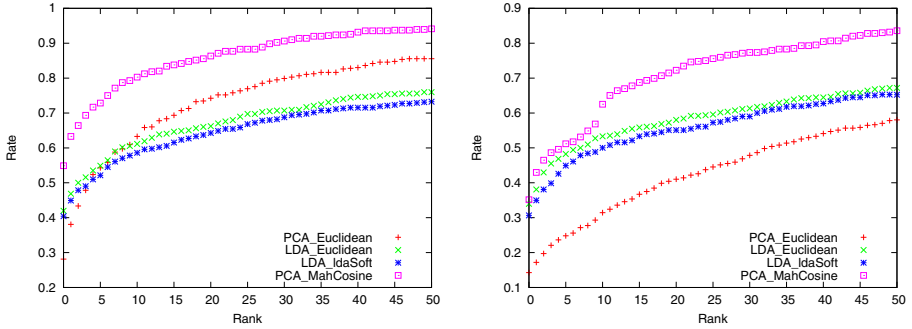


Fig. 4. Eigenface-based recognition after de-identification of the same cluster with $k = 10$, $\sigma = 0.85$, using averaging (top) and MahCosine (bottom).

In the previous experiments, we have selected the image clusters using the same distance measure as in the recognition task. This is not a realistic option in practice, so in a third series of experiments, we determined the best approach to de-identify images using a single clustering. We tested two approaches:

- Compute the four de-identified images for one subject, each grouped by one of the Eigenface-based measures, and average these pixel-wise.
- Compute clusters using a single distance measure, here we used the MahCosine measure (which performed best for $\sigma = 0.85$).

The results for both approaches are shown in Figure 4. The graph on the left shows that averaging over the four de-identified images per subject yields rather bad results, all algorithms have a rank-0 recognition rate of 40%-55%. The likely reason for this is that the clusters used to compute each of these projections were different, and the right image lies in the intersection between these. The right graph shows much better de-identification and is our preferred method. The recognition rate of the classical Eigenfaces method is worse than in the case were we *specifically* targeted this method, see Figure 2. The curves of all other three algorithms very much resemble the already known results as in the afore mentioned graph. What is more important, they are still worse than the performance of the MahCosine distance measure, which serves as our benchmark in this case.

We conclude that clustering wrt. MahCosine achieves a high degree of de-identification among all tested Eigenface-based methods while only minimally impacting identification by humans. We expect that other distance measures exhibit similar performance.

5.5 Experiments for Bayesian Face Recognition

For the Bayesian face recognition, there are two (related) classifiers: MAP operates on both intra- and extra-personal spaces, while the simpler ML classifier bases its estimate on the intra-personal space only. For most applications, the

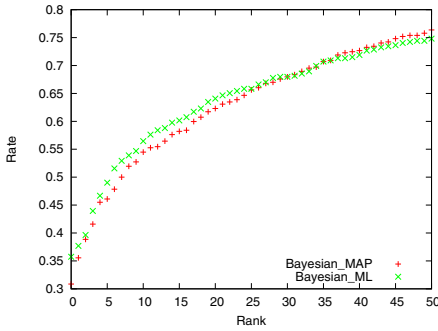


Fig. 5. Bayesian recognition after de-identification with $\sigma = 0.85$ and $k = 10$

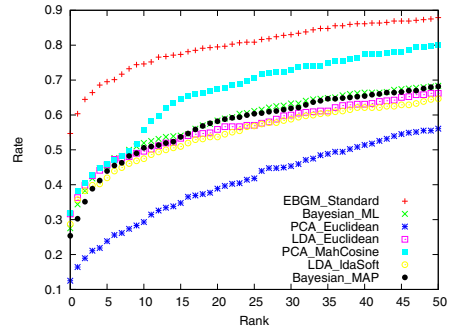


Fig. 6. Recognition rates of all algorithms after de-identification with $\sigma = 0.85$ and $k = 10$



Fig. 7. Comparison of image quality after de-identification with $k = 10$, $\sigma = 0.85$ for clusters grouped by MAP and ML

two variants provide very similar results, but our first set of experiments targets each method individually.

For the Bayesian method we have again performed de-identification as described previously. We have grouped $k = 10$ subjects into clusters, determined by their closeness according to the classification by MAP and ML. Then we have de-identified the clusters with $\sigma = 0.85$. The performance of the recognition algorithms is shown in Figure 5 and is in line with the results expected due to prior experiments for Eigenface-based algorithms.

Figure 7 shows the visual effects of the de-identification. Again we see that both subjects are clearly recognizable by humans, although the outline of the heads is blurred to a certain degree, which we attribute to variations in the pre-processing/normalization of the EBGM-method as we explained before.

5.6 Experiments for Combined Face Recognition

Finally, we combined the method of modifying PCA-based projections with our algorithm to alter EBGM-specific features. Since clustering subjects by their closeness according to MahCosine proved to be effective against Eigenface-based approaches, we decided to do the same for this experiment. We have de-identified the probe- and gallery-set with $k = 10$ and $\sigma = 0.85$ in the first step and then



Fig. 8. Image quality after de-identification for all algorithms with $k = 10, \sigma = 0.85$. The images shown represent one cluster grouped according to MahCosine, i.e., previous work maps these to the *same* average image.

modified the jets of the resulting images to resemble the average of the same set of clusters (still grouped by MahCosine).

Figure 6 shows the recognition performance of all algorithms when operating on the same set of images. We see that the EBGGM algorithm fares better in recognizing de-identified subjects than the other methods, but still only gets a 55% rank-0 recognition rate, see the discussion in the next section. Interestingly, the curves for Bayesian MAP, ML and Eigenface-based methods are very close to each other. Figure 8 shows the visual outcome of the de-identification procedure. The images are all taken from the same cluster (of 10 images total). Note that previous work (e.g., [16]) would have assigned the *same* average image to all of these images, i.e., they would be indistinguishable for humans. However, the images produced by our method are clearly distinguishable, which was the main goal of our work.

6 Discussion of Results

Face recognition algorithms work very well for images taken in a controlled environment (i.e., with regard to background, illumination, tilt, etc.). For example, the CSU implementation of the EBGGM algorithm achieves a rank-0 recognition rate of more than 90% (c.f. Figure 6), and the original EBGGM implementation fared even better. Our modifications reduce the rank-0 recognition rate to 55% for EBGGM, and below 30% for the other algorithms, which is a big improvement.

The EBGGM algorithm is, according to our experiments, harder to fool than other algorithms. Possible reasons are that EBGGM is a feature-based algorithm (i.e., it works on small patches of the face, not a holistic view of the face), and it has a very good recognition rate in general. In practice, images stored on image sharing sites are not taken in a controlled environment, and we expect that in a real environment the recognition rates will drop substantially compared to the above experiments.

Often, a corporation’s interest in collecting data and a user’s interest in privacy are diametrical. That said, using anonymization techniques will probably cause a reaction by the corporations deploying face recognition algorithms, which could eventually lead to an arms-race between both sides. This and finding a solution to the remaining issues will most likely inspire future research.

7 Related Work

Systematic research on face de-identification started with work by Newton, Sweeney, and Malin (e.g., [22,23]). Their goal was to achieve *perfect anonymity*, which makes identifying an individual from the anonymized photo provably impossible for machines (and humans). Their first approach, called “k-same”-method, computed the de-identified image either averaging the closest k images pixel-wise (“k-same-pixel”-method) or in Eigenface (“k-same-eigen”-method). While achieving strong security guarantees, the resulting image quality of both approaches was mediocre. Consequently, follow-up work improves image quality [16] and defines a more rigorous framework for the different notions of privacy protection models [15].

More recent work by Do et al. focuses on systematic de-identification methods for feature-based systems which are used in forensics [9]. In this work, the authors delude a class of image recognition algorithms (as opposed to face recognition), which can be understood as a superset of the class of algorithms we are concerned with.

A couple of ad-hoc methods such as masking parts of the face (e.g. the eyes) or blurring or pixelation of faces [7,17,20,36] were eventually tested. However, these methods are visually intrusive and target human and algorithmic recognition alike. Even worse, it was shown that their effectiveness is very limited [23], so they are not a good option.

8 Conclusion

We have shown a reliable way to de-identify face images against a wide range of currently available recognition algorithms. While not achieving the very strong notion of k -anonymity, we achieved a level of anonymity which is sufficient to counter the two most pressing problems that face recognition software poses for users of social networks: first, automated extraction of the social graph, i.e., friendship relations; second, automated tagging of people in images. At the same time we get an image quality which still allows humans to identify persons in images.

Acknowledgment. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office [26,27].

References

1. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1064, pp. 45–58. Springer, Heidelberg (1996)
2. Beresford, A.R., Kübler, D., Preibusch, S.: Unwillingness to pay for privacy: A field experiment. IZA Discussion Paper No. 5017 (2010), <http://ftp.iza.org/dp5017.pdf>
3. Beveridge, J., She, K.: Fall 2001 update to the CSU PCA versus PCA+LDA comparison (2001), <http://www.cs.colostate.edu/evalfacerec/>
4. Beveridge, R., Bolme, D., Teixeira, M., Draper, B.: The CSU face identification evaluation system user’s guide: Version 5.0 (2003), <http://www.cs.colostate.edu/evalfacerec/>
5. Bolme, D.S.: Elastic bunch graph matching. Master’s thesis, CSU Computer Science Department (2003)
6. Bolme, D.S., Beveridge, J.R., Teixeira, M., Draper, B.A.: The CSU face identification evaluation system: Its purpose, features and structure. In: Proc. International Conference on Vision Systems, pp. 304–311 (2003)
7. Boyle, M., Edwards, C., Greenberg, S.: The effects of filtered video on awareness and privacy. In: Proc. of the 2000 ACM Conference on Computer Supported Cooperative Work, CSCW 2000, pp. 1–10. ACM (2000)
8. Deutsche Welle. Facebook facial recognition raises eyebrows in Germany, EU, <http://www.dw-world.de/dw/article/0,,15144128,00.html>
9. Do, T.-T., Kijak, E., Furon, T., Amsaleg, L.: Deluding image recognition in sift-based cbir systems. In: Proc. 2nd ACM Workshop on Multimedia in Forensics, Security and Intelligence, pp. 7–12. ACM (2010)
10. EPIC.org. EPIC files complaint, urges investigation of Facebook’s facial recognition techniques (2011), <http://epic.org/2011/06/epic-files-complaint-urges-inv.html>
11. Facebook Announcement: 10 billion photos, http://www.facebook.com/note.php?note_id=30695603919
12. Facebook Announcement: Making Photo Tagging Easier, <http://www.facebook.com/blog.php?post=467145887130>
13. Facebook Inc., <http://developers.facebook.com/docs/opengraph/>
14. Flickr Blog: 4 billion photos, <http://blog.flickr.net/en/2009/10/12/4000000000/>
15. Gross, R., Sweeney, L.: Towards real-world face de-identification. In: First IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS 2007, pp. 1–8 (2007)
16. Gross, R., Sweeney, L., de la Torre, F., Baker, S.: Model-based face de-identification. In: Proc. 2006 Conference on Computer Vision and Pattern Recognition Workshop, CVPRW 2006. IEEE Computer Society (2006)
17. Hudson, S.E., Smith, I.: Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems. In: Proc. 1996 ACM Conference on Computer Supported Cooperative Work, CSCW 1996, pp. 248–257. ACM (1996)
18. Miller, P., Lyle, J.: The effect of distance measures on the recognition rates of PCA and LDA based facial recognition. Tech. rep., Clemson University (2008)
19. Moghaddam, B., Jebara, T., Pentland, A.: Bayesian face recognition. *Pattern Recognition* 33(11), 1771–1782 (2000)

20. Neustaedter, C., Greenberg, S., Boyle, M.: Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Trans. Comput.-Hum. Interact.* 13, 1–36 (2006)
21. New York Times, How privacy vanishes online (2010), <http://www.nytimes.com/2010/03/17/technology/17privacy.html>
22. Newton, E., Sweeney, L., Malin, B.: Preserving privacy by de-identifying facial images. Tech. Rep. CMU-CS-03-119, Carnegie Mellon University, School of Computer Science (2003)
23. Newton, E., Sweeney, L., Malin, B.: Preserving privacy by de-identifying facial images. *IEEE Transactions on Knowledge and Data Engineering* 17(2), 232–243 (2005)
24. Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., von der Malsburg, C.: The Bochum/USC face recognition system and how it fared in the FERET phase III test. In: *Face Recognition: From Theory to Applications* (1998)
25. Pfitzmann, A., Köhntopp, M.: Anonymity, unobservability, and pseudonymity – a proposal for terminology. In: *Workshop on Design Issues in Anonymity and Unobservability*, pp. 1–9 (2000)
26. Phillips, J.P., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(10), 1090–1104 (2000)
27. Phillips, P., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing* 16(5), 295–306 (1998)
28. Phillips, P.J., Rauss, P.J., Der, S.Z.: FERET (face recognition technology) recognition algorithm development and test results. Tech. Rep. ARL-TR-995, Army Research Laboratory (1996)
29. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In: *IEEE Symposium on Security and Privacy*. IEEE Computer Society (1998)
30. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10(5), 557–570 (2002)
31. Teixeira, M.L.: The Bayesian intrapersonal/extrapersonal classifier. Master’s thesis, Colorado State University (2003)
32. Turk, M., Pentland, A.: Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
33. Wiskott, L., Fellous, J.-M., Kruger, N., von der Mals, C.: Malsburg, C. V. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 775–779 (1997)
34. Yambor, W., Draper, B., Beveridge, R.: Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures. In: *Empirical Evaluation Methods in Computer Vision* (2002)
35. Yang, M.-H., Kriegman, D., Ahuja, N.: Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(1), 34–58 (2002)
36. Zhao, Q.A., Stasko, J.T.: Evaluating image filtering based techniques in media space applications. In: *Proc. of the 1998 ACM Conference on Computer Supported Cooperative Work, CSCW 1998*, pp. 11–18. ACM (1998)
37. Zhao, W., Krishnaswamy, A., Chellappa, R., Sweets, D., Weng, J.: Discriminant analysis of principal components for face recognition. In: *Proc. 3rd IEEE Int. Conference on Face and Gesture Recognition*, pp. 336–341 (1998)

Client-Side Biometric Verification Based on Trusted Computing

Jan Vossaert¹, Jorn Lapon¹, Bart De Decker², and Vincent Naessens¹

¹ Katholieke Hogeschool Sint-Lieven, Department of Industrial Engineering
Gebroeders Desmetstraat 1, 9000 Ghent, Belgium

`firstname.lastname@kahosl.be`

² KU Leuven, Department of Computer Science, iMinds-DistriNet
Celestijnenlaan 200A, 3001 Heverlee, Belgium

`Bart.DeDecker@cs.kuleuven.be`

Abstract. Traditionally, a user requires substantial trust in a workstation for correctly handling her credentials (e.g. password/login). Unfortunately, malware and compromised software makes them unsuitable for secure credential management. Credentials are easily stolen and the user cannot trust what is being displayed on her workstation, obstructing informed consent.

This paper presents a new solution that addresses these issues. Credentials are bound to the owner using biometrics, effectively impeding abuse through credential sharing and theft. The biometric verification is performed on the client side, preserving the privacy of the user. The solution ensures that the user is correctly informed about the pending authentication, preventing abuse by malware. To demonstrate the feasibility of our approach, a prototype was implemented.

1 Introduction

Many companies and governments are digitalizing their services. This allows users to access these services remotely. To prevent unauthorized users from gaining access and protect the integrity of these services, access control measures are enforced. These access control measures are typically enforced using digital credential technologies such as X.509 certificates, attribute based credentials or simply a username/password combination.

However, merely using digital credential technologies is not sufficient to fulfill the complex security and privacy requirements that apply in these settings. Credential technologies themselves, for instance, do not prevent users from *sharing* their credentials (e.g. digital credentials can be copied and distributed among users). Moreover, credentials can also be *abused by malicious software*. For instance, malicious software can use the credentials of the user to access personalized services, potentially without consent of the user. This impedes abuse detection and consequently credential revocation.

These issues can be tackled by binding the credentials to the owner. One way to bind a credential to its owner is by using biometric authentication to activate

the credential. Existing solutions typically use a tamperproof device such as a smart card on which the user's biometric template and credentials are stored. The user can activate these credentials by transferring her biometric scan to the tamperproof device. This binds the user to the credentials stored on the tamperproof device.

These biometric-based solutions, however, still have a number of disadvantages. They often require a significant amount of trust in the workstation. For instance, to inform the user about the pending authentication or to handle the user's biometric data. Moreover, besides a biometric scanner, these solutions require dedicated tamperproof hardware. Updating and patching the software running on these tamperproof devices is often difficult or even impossible. This hinders software security updates from being rolled-out and decreases flexibility with respect to using new biometric algorithms or credential technologies.

This paper presents a new solution for activating credentials bound to the owner by means of biometrics. In contrast to existing solutions based on tamperproof hardware, the verification is performed on the workstation in a trusted application. It, therefore, applies *Secure Virtualization Technologies* embedded in modern commodity workstations and laptops for building a Secure Execution Environment (SEE). In the prototype solution the user's credentials are stored on her mobile device and are bound to her biometrics. Both the biometric scan and its binding to the credential are verified in the SEE at the client side, trusted by both the user and the service provider. This strategy avoids leaking biometric information to, for instance, the service provider and requires no additional hardware infrastructure to be rolled out. Moreover, this solution is designed to be generic and, hence, supports the use of public, potentially untrusted workstations.

The *contribution* of this paper is threefold. First, it presents a solution for the secure verification of biometric traits on a workstation by applying Secure Virtualization Technologies. Second, access to remote services is controlled by credentials that can only be used after a successful biometric verification on the workstation. A mobile device is used to carry the user's credentials. Third, a prototype implementation of the system was realized, validating our solution. For the implementation an UHCI USB stack and a biometric scanner driver and algorithm were added to an existing framework for building SEE applications.

The rest of this paper is structured as follows. Section 2 points to related work. The used technologies are described in Section 3 and are followed by the design in Section 4. In Section 5, more information about the realization is presented. Subsequently, the system is evaluated in Section 6. Finally, we draw some conclusions in Section 7.

2 Related Work

Some credential systems such as the Identity Mixer library [8] provide all-or-nothing non-transferability to discourage users from sharing their credentials. All the credentials of the user are tied together. Hence, assuming that the user owns

at least one valuable credential, he will not be willing to share her credentials with other users. A similar approach is PKI-assured non-transferability where the credentials are bound to a valuable secret outside the system (e.g. credit card information). Whereas these systems focus on the discouragement of credential sharing, the system proposed in this paper also addresses abuse (after theft) prevention and informed consent.

Another approach to prevent digital credentials from easily being copied or shared is by embedding them inside tamperproof hardware. The system proposed in [7] leverages the DAA [6] protocol, available in Trusted Platform Modules (TPMs) of modern workstations, to bind the user’s anonymous credentials to a TPM. Similarly, smart cards are used to implement anonymous credential systems [2,21,1]. Although smart cards prevent the credentials from being copied, they do not fully prevent sharing of the credentials. Moreover, anonymous credential systems have other unsolved issues, such as their performance and correctly informing which information is being disclosed.

To tackle this, biometric authentication can be used to bind credentials to the owner. For instance, in [17,4] the *wallet with observer* architecture [9] is extended to include biometric authentication towards the observer. The user is issued a tamperproof card containing her credential and biometric template. To use the credential in the card, the holder is required to scan her biometric data. Only if the scanned data matches the template stored in the card, the credential is activated. As an example, a privacy preserving identity card has been designed [11] taking advantage of this approach. Another approach [3] uses *fuzzy extractors* [14,16] to generate a cryptographic key based on the retrieved biometric features. This key is never stored and the tamperproof device is trusted to erase the value after authentication. Hence, fresh biometric readings are required to reconstruct the cryptographic key. These systems focus on the prevention of abuse through theft and sharing, but do not fully realize the aspect of informed consent. This is especially important in case anonymous credentials are used. Tamperproof devices are also less flexible with respect to software updates and the support of complex biometric systems or credential technologies.

3 Background

TCG Trusted Computing. Nowadays, modern commodity computers are equipped with a Trusted Platform Module (TPM) [15]. This is a hardware module physically attached to the computer’s motherboard, extending the system with a set of security related features. One these features is the measurement of the state of the system. To this end, the TPM contains several Platform Configuration Registers (PCRs). These are cleared upon power up and can only be modified using the `extend` operation, performed inside the TPM. The result of this operation for a specific PCR is a new PCR value, being the hash of the current value concatenated with a new value (i.e. $PCR_n := SHA1(PCR_n || value)$).

A transitive trust model is employed: each software component, starting from the Core Root of Trust for Measurement in the BIOS, is responsible for measuring the following component in the chain before passing control. Hence, before

loading subsequent software components, the preceding component hashes the binaries of the components to be loaded and extends the result in a specific PCR. As a result, the PCRs represent the state of the system, or in other words, the loaded software configuration.

Based on this state, the TPM also supports a number of additional features. Data can be sealed with the `seal` operation and only if the system resides in the state specified during the seal operation can the data be unsealed (`unseal`). Additionally, the `quote` command returns a proof of the state (i.e. a *quote*) which a third party can verify (`verifyQuote`) asserting that the (remote) system runs in a specific (trusted) state. This functionality uses the private key and certificate of the TPM (i.e. sk_{tpm} and $cert_{tpm}$) to assert that the operation is performed by a genuine TPM. These credentials can either be generated by the hardware manufacturer or during an enroll phase.

Secure Execution Environment. While TPMs are being embedded in workstations for several years now, a more recent evolution is the adoption of SEE technologies such as Intel's Trusted Execution Technology (TXT) [10] and AMD's Secure Virtual Machine (SVM) [13]. These technologies allow the execution of measured code independently of previously executed software. The TPM specification has been extended with additional capabilities to support these new technologies.

Recent work [20] presents a framework that uses these technologies to allow developers to isolate security critical code from applications and run it in a secure environment. The main, possibly untrusted, OS is temporarily suspended after which the sensitive Piece of Application Logic (*PAL*) is securely executed. When the execution of the sensitive code is completed, the OS resumes execution. Typically, the *PAL* extends its state in the TPM with a fixed known value before releasing control back to the OS. This prevents the OS from gaining access to secrets from the *PAL* or from asserting data on behalf of the *PAL*. The framework supports data transfer between the main OS and the *PAL*. The TPM operations can be used to assert to a remote party that certain data was generated by a trusted *PAL*. The framework supports both Intel TXT and AMD's SVM technology on Windows and Linux based systems. In [5] this framework is extended to allow secure user interaction (i.e. input via the keyboard and output via the monitor).

Biometric Authentication. Biometry can be used to uniquely identify a person [19,18]. Commonly used biometric traits include a fingerprint, iris, face and voice. A special purpose sensor device is used to read the biometric trait of the user. During enrollment, a distinguishing feature set is extracted from the biometric data and stored as a *biometric template* (bt_u) of the user. During authentication, the user scans her biometric trait (`bioScan`) and the resulting feature set is *matched* to the feature set contained in the template of the user. Based on the similarity of the two sets, the authentication is either accepted or rejected (`bioVerify`).

These biometric templates can be bound to the credentials of the user. The credential based authentication is then combined with the biometric authentication. A verifier can, hereby, check that the user of the credentials is also the owner. Binding the user's biometrics to an X.509 certificate can, for instance, be done by including a cryptographic hash of the biometric template in the certificate. For other types of credentials such as passwords or anonymous credentials similar principles can be applied. The verification of the biometric binding is further denoted as `verifyBinding`.

4 Design

This section first lists the different actors in the system. Followed by the requirements and a general description of the system. Finally, a detailed description of the protocols is presented.

4.1 Roles

We assume a user U , carrying a mobile device M . The mobile device stores the user's credentials and is used as a credential vault for accessing remote services from the workstation. The workstation runs a legacy operating system (WS) and supports SEE technologies for running a trusted application (PAL). A biometric scanner is attached to the workstation.

4.2 Requirements and Adversary Model

Functional Requirements

- F_1 The system requires commodity hardware only.
- F_2 The solution does not require authentication to be bound to a particular workstation.
- F_3 The system is extensible and modular, allowing for new biometric systems or algorithms and credential technologies to be included.
- F_4 In case vulnerabilities are found in the system, software updates are easily applied.

Security and Privacy Requirements

- S_1 A credential for authenticating to a remote service can only be used by its owner.
- S_2 Malicious software cannot mislead the user into approving malicious signing transactions or authentication attempts.
- P_1 The system protects the biometric information of the user.

Adversary Model

- A_1 Regarding the secure execution environment, the same assumptions as the Trusted Computing Group [15] are made.
- A_2 The trusted application (*PAL*) is assumed to be formally verified.
- A_3 The biometric system is assumed to be secure (i.e. sufficiently low false acceptance rate).
- A_4 The mobile device is assumed to securely handle the user's credentials and biometric information. To this end, an embedded secure element can be used for managing this data.
- A_5 The service providers correctly verify the received attestations.

4.3 General Approach

A user authenticates on the workstation towards a remote service provider. The service provider requires that the user proves ownership of the used credentials, before allowing access to its services. To his end, the user's credentials are bound to her biometrics. The verification of the biometric binding between the credentials and the user is performed by a dedicated trusted application (i.e. the *PAL*). In addition, the *PAL* informs the user about the details of the pending authentication. This ensures that malware cannot mislead the user into approving malicious transactions. This is especially important in case anonymous credentials are used, as the user should give consent on the selective disclosure of attributes. The *quote* functionality of the TPM is used to assert to service provider that a trusted *PAL* properly executed the verification. The credentials and biometric data are stored on the mobile device of the user and are only released towards the trusted *PAL*, running on the workstation. This ensures that the user does not release her biometric data to malicious applications.

The *PAL* is trusted by both the user and the service provider. The user trusts the *PAL* not to release her biometric information to third parties and for informed consent regarding the pending authentication. The service provider trusts the *PAL* to correctly verify the biometric binding between the user and the used credential(s). This trust is supported by the attestation of the *PAL* towards the service provider and mobile. The state of the trusted *PAL* can be certified by a trusted third party. The *PAL* should also be open source so that independent developers can verify that the certified state correctly represents the desired functionality.

4.4 Protocols

Prerequisites. The SEE technologies on the workstation are enabled in the BIOS and the TPM has been certified (i.e. sk_{tpm} and $cert_{tpm}$ are generated). The *PAL* running on the workstation of the user has been initialized. During initialization the *PAL* generates a keypair (pk_{pal} and sk_{pal}) and *seals* it to its state resulting in the sealed object denoted as *keyStore*. The mobile device and the service provider obtain the certified PCR state of the trusted *PAL*. Hence,

these parties can verify that the application running on the workstation is indeed the intended trusted application.

A credential issuer issues credentials bound to the user’s biometrics, which are stored on the user’s mobile device. The mobile requires the user to choose a unique authentication image (img_u) that will allow the user to visually verify that the software running on the workstation is indeed trusted.

Pairing. Before authentication, the mobile device verifies that a valid trusted application is running on the workstation. As part of this protocol, the mobile retrieves the *PAL*’s public key (pk_{pal}), which is used to encrypt data addressed to the *PAL*. The mobile device stores this public key for future authentications. We denote this protocol as the *pairing protocol*. Note that this pairing can be performed immediately before the authentication in case a public workstations is used.

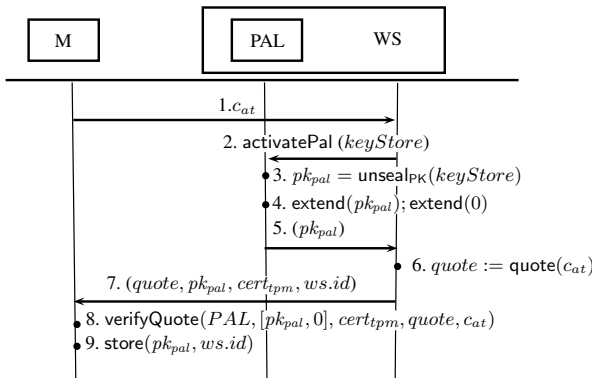


Fig. 1. The pairing protocol

Figure 1 illustrates the steps of the pairing protocol. First, the mobile sends a random attestation challenge to the workstation (1). The workstation activates the trusted *PAL* with the sealed *keystore* (2) and the *PAL* retrieves his public key from the sealed *keystore* (3). To allow attestation that this public key is indeed managed by the trusted application, the state is extended with this key (4). The *PAL* returns its public key to the workstation which resumes its execution (5). A quote operation on the state resulting from the *PAL* execution is performed using the attestation challenge (6). As the public key was extended in the state, this ensures the authenticity of the public key sent to the mobile. The challenge ensures freshness of the quote. The quote, public key, certificate and an identifier of the workstation is sent back to the mobile, which then verifies the quote (7-8). If the quote verification was successful, the public key is stored together with the workstation’s identifier (9).

Authentication. Figure 2 presents the protocol for authenticating the user towards a remote service provider. First, the user requests access to a protected resource from the remote service provider (1). The provider responds with an authentication request containing its certificate, an authentication and an attestation challenge (2). The mobile now receives the authentication challenge, certificate of the service provider and a unique identifier of the workstation (3). The mobile informs the user about the workstation on which the authentication will be performed and towards which service provider (4). If the user acknowledges, the mobile signs the authentication challenge with the user's credential and encrypts the signature, authentication certificate, biometric template and the user's unique image using the public key of the *PAL* (5-7). The encryption is sent to the workstation where it is then passed as a parameter to the *PAL*, together with the *PAL*'s sealed keys in *keyStore* (8-9). The *PAL* now unseals its private key to decrypt the encrypted data *enc* (10-11). Subsequently, the binding between the biometric template and the authentication credential is verified (12). If the verification succeeds, the user is informed about the pending authentication and requested to scan his biometric data. The user's unique image is shown to indicate to the user that the trusted environment is running (13). The user can, hence, trust all information shown on the monitor. To acknowledge the authentication, the user scans his biometric data using the biometric scanner attached to the workstation (14). As the *PAL* is in complete control of the workstation, it can directly interact with the hardware. This ensures that the data shown on the monitor and read from the biometric scanner cannot be tampered with. The *PAL* can now verify if the biometric template matches with the scanned biometric data (15). Upon successful verification, the *PAL* is assured that the user is the owner of the authentication credentials and *extends* its state with the signature, authentication certificate and the certificate of the service provider (16). The *PAL* ends its execution and returns the user's signature and certificate back to the regular OS, that resumes its operation (17). The OS now performs a *quote* operation on the state resulting from the *PAL* execution (18). This quote attests towards the service provider that the trusted *PAL* indeed verified the biometric binding (i.e. the *PAL* state is extended with the user's certificate and signature) and that the user was shown the correct information about the service provider (i.e. the *PAL* state is extended with the service provider's certificate). The resulting quote is sent to the service provider along with *sig* and *cert_u* and the certificate of the TPM (19). The service provider now verifies the quote, the user's certificate and signature (20-21). Upon success, the service provider grants access to the requested resource (22).

5 Realization

For the realization of a prototype, an off-the-shelf USB fingerprint scanner (i.e. Eikon UPEK fingerprint reader) was used as biometric reader. The user's credential is a X.509 certificate with a 1024 bits RSA key. The user's fingerprint template is bound to her authentication credential by including a cryptographic

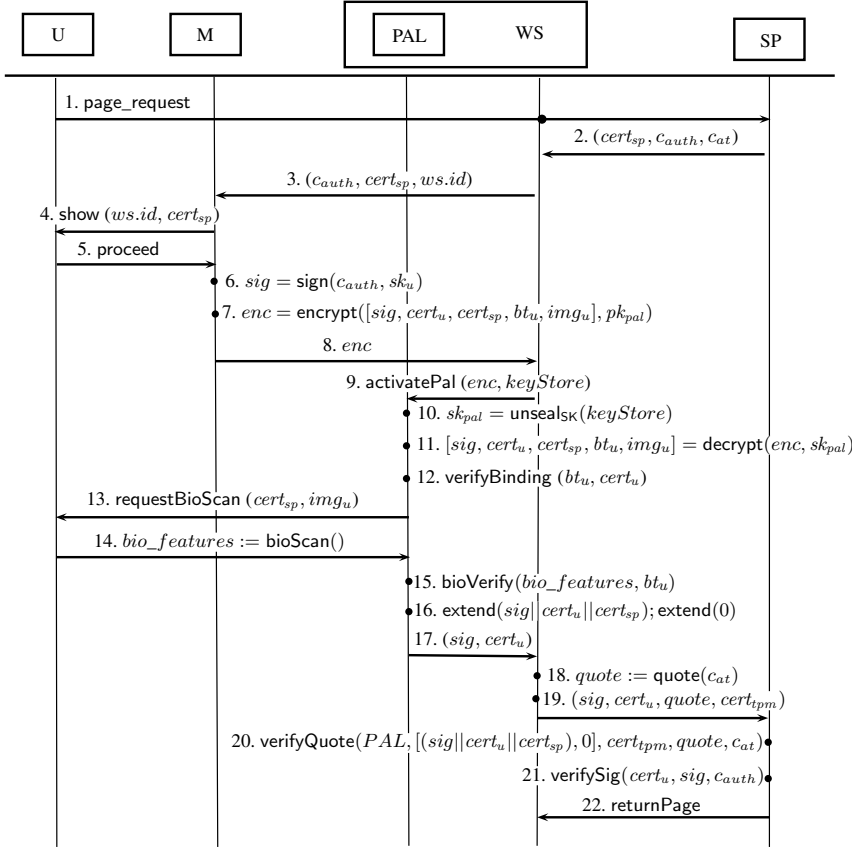


Fig. 2. The authentication protocol

hash of the fingerprint template in the X.509 certificate. Table 1 shows the hardware used for the realization of the different entities.

Table 1. The hardware platforms for the realization of the different entities

Entity	Realization
M	Samsung i9000 Galaxy S: 1GHz ARM Cortex-A8, 512MB RAM The smartphone runs Android 2.3.3
SP, PAL, WS	DELL E4200: Intel Core2 Duo U9400 @ 1.4GHz, 4GB RAM The laptop runs Ubuntu 12.04 32-bit, 3.2.0 kernel

For the functionality of the mobile device, an Android app has been developed. The service provider is implemented on an Apache Tomcat server. Spring Security is used to enforce its access control policy. A Spring authentication

module was added to handle the custom authentication protocol. Nevertheless, the focus of the prototype lies in the development of the software components on the workstation. The user accesses the service provider via the browser on the workstation. When authentication is required, the authentication request (*bioauth:authRequestData*) is forwarded to a local application on the workstation. For demonstration purposes a bidirectional network connection is setup between the local application and the mobile device. To this end, a QR code, containing the workstation’s IP address, is displayed by the local application and scanned with the mobile device. Using this IP address, the mobile connects to the workstation and uses this channel for further communication. The local application also informs the user about the progress of the authentication protocol. Note that the actual authentication is delegated to a background daemon that accepts incoming TCP/IP connections from the mobile of the user and has the required privileges to suspend the OS and start the *PAL*.

For the biometric verification, a basic USB UHCI stack and fingerprint driver for the Eikon reader were implemented. For parsing X.509 certificates and other cryptographic operation, parts of the PolarSSL library were used. When the *PAL* has finished execution, the daemon uses the TrouSerS TCG software stack to implement the *quote* operation which is used to attest the *PAL* towards the service provider.

Our *PAL* application consists of three main components. The Eikon fingerprint driver (447 lines of code) running on top of the USB UHCI stack (1001 lines of code). The implementation of the *PAL* protocol itself consists of 1040 lines of code. This adds a total of 2488 lines of code to the flicker/soft cards framework, preserving a minimal Trusted Computing Base (TCB).

Table 2 illustrates the performance of the prototype. The computations on the mobile and server have negligible impact on the overall system performance. The *PAL* and the *WS* introduce the largest overhead as they use the constrained resources of the TPM. The workstation uses the TPM for the *quote* operation, the *PAL* for the *unseal* functionality and the establishment of the *SEE*. The actual scanning of the fingerprint is not included in the measurement to avoid impact of the user interaction. The measurements of the *PAL*, however, do include the initialization of the USB stack and fingerprint driver. Once, the reader is initialized, the actual swiping is only a fraction of the total time. Although the authentication phase of the *PAL* is the most time consuming operation, the user experience isn’t degraded as some operations can be performed while the user is reviewing the authentication details on the monitor.

Table 2. The performance of the pairing and the authentication (ms)

Performance	PAL	WS	M	SP	Total
Pairing	1900	720	30		2650
Authentication	4550	740	40	8	5338

6 Evaluation

6.1 Requirements Review

This section discusses how the requirements presented in section 4.2 are realized in the design.

Functional Requirements. As the prototype illustrates, the system only requires commodity hardware. The mobile component was implemented on a smartphone and the secure virtualization technologies required on the workstation are being embedded in off-the-shelf workstations, satisfying requirement F_1 .

The system works with any workstation supporting the required SEE prerequisites. Hence, a user can use any such workstation to securely authenticate towards a remote service provider, satisfying requirement F_2 .

The *PAL* can easily be updated by distributing a new binary to the workstation and certifying the new state of the application with the mobile device of the users and the service providers. This can be managed by a trusted third party that certifies and revokes these states. The software on the mobile device and server can be updated using traditional mechanisms. This allows the integration of additional biometric and/or credential technologies and security updates to be installed, satisfying requirements F_3 and F_4 .

Security and Privacy Requirements. The authentication credentials of the user are bound to her biometrics. The *PAL* ensures the correct verification of the user's biometrics before asserting the authentication. The service provider can verify the assertion and, hence, check that the actual owner authorized the authentication, satisfying requirement S_1 .

The mobile device verifies integrity of the *PAL* running on the workstation, before sending any personal data. Moreover, the user is assured that the trusted application is running, as her personal image img_u is shown on the workstation. The user is, hence, assured that the provided information about the pending authentication is correct. Furthermore, the *PAL* binds the verification process to the service provider presented to the user. This satisfies security requirement S_2 .

In the prototype, the biometric template of the user is bound to the user's certificate by including a cryptographic hash of the template in the certificate. This prevents biometric information of the user from being leaked when the certificate is released. The mobile device only discloses the biometric template to a trusted *PAL* (i.e. the *PAL* does not reveal the fingerprint to a third party) over a secure channel. Moreover, the user can verify that a trusted *PAL* is running when scanning her fingerprint. This prevents freshly scanned biometric information of the user from being leaked, satisfying requirement P_1 .

6.2 Security and Privacy Considerations

The main focus of the system is protecting the user from software attacks, as these are the most common and scalable types of attacks. Moreover, TPMs are

not designed to be secure against hardware attacks. To mitigate the impact of hardware attacks on a TPM, its credentials can be revoked.

The *PAL* is trusted by both the user and the service provider to correctly execute the specified protocol. To limit implementation flaws that could be exploited, the functionality of the *PAL* is kept to the minimum (i.e. informed consent, USB communication and biometric verification). The small TCB decreases the chance of bugs and suggests that formal verification is possible. Moreover, as mentioned in the functional requirements review section, the *PAL* can be easily updated after which the previous version can be revoked by blacklisting its state.

Currently, the system assumes that all service providers require biometric verification. If a service provider does not require this proof of ownership, the protocol requires some minor modifications. Otherwise, malware could trick the user in signing an authentication challenge obtained from another service provider, which does not require this biometric verification. One simple solution could be to have the mobile verify the service provider's certificate and encrypt the user's signature with the contained public key.

Ideally, the biometric reader used for obtaining a biometric scan of the user is bound to the workstation on which the user is working. However, most biometric scanners are plug-and-play and can easily be removed and replaced with other hardware devices. If a fingerprint scan is eavesdropped, it can be replayed as a fresh reading. This risk can be mitigated by implementing a cryptographic protocol between the trusted application and the reader to ensure freshness of the scan. Although hard to achieve, to prevent relay attacks, the *PAL* should also be able to verify that the used biometric reader is actually attached to the workstation on which the *PAL* is running.

The system presented in this paper increases the user's privacy with respect to the biometric authentication. Nevertheless, this system is easily extended to further increase privacy by supporting anonymous credentials in which no linkable information should be released to the service provider. In the prototype, the *quote* generation requires a certificate bound to the TPM. This makes all transactions performed on a single workstation linkable. Therefore, modern TPMs also support the *DAA* protocol. This protocol allows anonymous attestation of the platform. As such, the system only leaks the state of the *PAL* and the data disclosed during the user authentication.

Note that it is even possible to support password based authentication with our system. In that case the certification authority should bind the biometric data to the login or username.

6.3 Applicability

This section discusses two possible application domains in which the system presented in this paper can be used to realize increased security compared to currently deployed systems, namely *eID systems* and *online banking*.

eID Systems typically allow the user to access a wide range of personal services. This can go from very privacy sensitive services such as online tax submission

to less privacy sensitive services such as pay-per-view news site. While these privacy sensitive services will typically only be accessed from a trusted home computer, other services might be accessed on workstations not fully trusted by the user. However, when using the same authentication credentials for the privacy sensitive and the other services, malware on an untrusted workstation could use the authentication credentials to access other services then requested by the user. The system presented in this paper prevents this type of abuse by correctly informing the user about the service which will be accessed.

Online banking enables a wide variety of banking services (e.g. viewing the status of your bank accounts, loans and execute bank transactions) via a workstation connected to the Internet. The user owns a credential with which he can log in and authorize transactions (e.g. wire transfer).

In current eBanking systems this secret is stored in an smart card (i.e. the bank card of the user). To authorize transactions, the details of the transaction are transferred to the bank card that subsequently generates an authorization response. This approach protects the credentials of the user but is vulnerable to phishing attacks. Our approach also tackles the latter as the secure execution environment application correctly informs the user about the pending transaction. It, moreover, replaces PIN authentication of the user with stronger biometric authentication.

6.4 Comparison with Existing Systems

As discussed in related work, several systems for misuse protection of credentials exist. For this comparison, the existing systems are categorized as follows. *Hardware based protection (HBP)* systems [7,2,21] rely on tamperproof hardware to prevent credentials from being digitally copied and, hence, easily abused. *Software based protection (SBP)* systems [8] rely on binding the credentials of the user to a valuable secret of the user, discouraging users to share their credentials. Finally, *biometry-based protection (BBP)* systems [17,4,9,12,11,3] embed the user's credentials in a tamperproof module that requires biometric authentication of the user before the credentials can be used. The results of the comparison are summarized in Table 3.

Informed consent allows user to (dis)approve transactions based on correct information about the transaction. The tamperproof hardware components used in HBP and BBP systems typically do not allow direct communication with the user. Therefore, other, potentially compromised, devices are required to inform the user about the transaction. In SBP systems, the credential operations are typically executed on a regular workstation. In the system presented in this paper, the trusted application has full control over the hardware of the workstation and can, therefore, use the monitor to reliably inform the user about the transaction.

The HBP and BBP both rely on tamperproof hardware components to store the user's credentials and execute the credential and biometric operations. These devices are typically resource constrained limiting the usage of computationally

intensive credential technologies such as anonymous credentials and the number of authentication credentials that can be stored inside those devices. The SBP and the system proposed in this paper are implemented on a general purpose workstation.

The HBP system implements misuse protection by embedding the credentials in a tamperproof module preventing them from being digitally copied. The SBP system discourages the user from sharing her credentials. The BBP and the approach discussed in this paper both implement misuse protection by binding the credentials to a specific user using biometrics. The solution presented in this paper, however, could easily be extended to support these credentials without checking the biometric binding.

The HBP and BBP both rely on tamperproof modules for executing credential and biometry related operations. The software installed on these modules is typically difficult to update providing less flexibility compared to the SBP and the system presented in this paper. These systems run on general purpose hardware in which updates can be part of the update infrastructure of the operating system.

Table 3. Comparison between the system proposed in this paper and existing misuse protection systems

	HBP	SBP	BBP	Our approach
Informed consent	No	No	No	Yes
Hardware resources	Constrained	Powerful	Constrained	Powerful
Protection	Copy prevention	Discourage sharing	Bio. binding	Bio. binding
Flexibility	Low	High	Low	High

7 Conclusion

This paper presents a new solution for activating credentials bound to its owner by means of biometrics. It assures that users are physically present when their credentials are used, effectively impeding credential sharing and abuse by theft. Moreover, credential abuse by malware is prevented by isolating the credential operations in a secure environment on a workstation. Apart from the hardware support available in modern commodity workstations, no additional infrastructure is required.

The system can be applied to general client-server authentication use-cases or dedicated use-cases such as electronic identity systems or eBanking. A prototype implementation demonstrates the feasibility of our system. In future research, this approach can be extended to include the verification of contextual information such as geographical data. A service provider could require that the workstation to be located in a certain country. This could be used to impede relay attacks.

Acknowledgements. This work is made possible through funding from the MobCom project, by the Flemish agency for Innovation by Science and Technology (IWT).

References

1. Bichsel, P., Camenisch, J., De Decker, B., Lapon, J., Naessens, V., Sommer, D.: Data-minimizing authentication goes mobile. In: De Decker, B., Chadwick, D.W. (eds.) CMS 2012. LNCS, vol. 7394, pp. 55–71. Springer, Heidelberg (2012)
2. Bichsel, P., Camenisch, J., Groß, T., Shoup, V.: Anonymous credentials on a standard java card. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, CCS 2009, pp. 600–610. ACM, New York (2009)
3. Blanton, M., Hudelson, W.M.P.: Biometric-based non-transferable anonymous credentials. In: Qing, S., Mitchell, C.J., Wang, G. (eds.) ICICS 2009. LNCS, vol. 5927, pp. 165–180. Springer, Heidelberg (2009)
4. Bleumer, G.: Biometric yet privacy protecting person authentication. In: Aucsmith, D. (ed.) IH 1998. LNCS, vol. 1525, pp. 99–110. Springer, Heidelberg (1998)
5. Brasser, F.F., Bugiel, S., Filyanov, A., Sadeghi, A.-R., Schulz, S.: Softer smartcards - usable cryptographic tokens with secure execution. In: Keromytis, A.D. (ed.) FC 2012. LNCS, vol. 7397, pp. 329–343. Springer, Heidelberg (2012)
6. Brickell, E., Camenisch, J., Chen, L.: Direct anonymous attestation. In: Proceedings of the 11th ACM Conference on Computer and Communications Security, CCS 2004, pp. 132–145. ACM, New York (2004)
7. Camenisch, J.: Protecting (anonymous) credentials with the trusted computing group's TPM V1.2. In: Fischer-Hübner, S., Rannenberg, K., Yngström, L., Lindskog, S. (eds.) Security and Privacy in Dynamic Environments. IFIP, vol. 201, pp. 135–147. Springer, Boston (2006)
8. Camenisch, J., Van Herreweghen, E.: Design and implementation of the idemix anonymous credential system. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, CCS 2002, pp. 21–30. ACM, New York (2002)
9. Chaum, D., Pedersen, T.P.: Wallet databases with observers. In: Brickell, E.F. (ed.) CRYPTO 1992. LNCS, vol. 740, pp. 89–105. Springer, Heidelberg (1993)
10. Intel Corporation. LaGrande technology preliminary architecture specification. Intel Publication no. D52212 (May 2006)
11. Deswarte, Y., Gams, S.: A proposal for a privacy-preserving national identity card. *Trans. Data Privacy* 3(3), 253–276 (2010)
12. Deswarte, Y., Gams, S.: The challenges raised by the privacy-preserving identity card. In: Naccache, D. (ed.) *Cryptography and Security: From Theory to Applications*. LNCS, vol. 6805, pp. 383–404. Springer, Heidelberg (2012)
13. Advanced Micro Devices. AMD64 architecture programmer's manual: Volume 2: System programming. AMD Publication no. 24594 rev. 3.11 (December 2005)
14. Dodis, Y., Reyzin, L., Smith, A.: Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In: Cachin, C., Camenisch, J.L. (eds.) *EUROCRYPT 2004*. LNCS, vol. 3027, pp. 523–540. Springer, Heidelberg (2004)
15. Trusted Computing Group. TCG TPM specification, http://www.trustedcomputinggroup.org/resources/tpm_main_specification
16. Hao, F., Anderson, R., Daugman, J.: Combining crypto with biometrics effectively. *IEEE Trans. Comput.* 55(9), 1081–1088 (2006)
17. Impagliazzo, R., More, S.M.: Anonymous credentials with biometrically-enforced non-transferability. In: Proceedings of the 2003 ACM Workshop on Privacy in the Electronic Society, WPES 2003, pp. 60–71. ACM, New York (2003)

18. Jain, A.K., Flynn, P., Ross, A.A.: Handbook of Biometrics. Springer-Verlag New York, Inc., Secaucus (2007)
19. Jain, A.K., Nandakumar, K., Nagar, A.: Biometric template security. EURASIP J. Adv. Signal Process, 113:1–113:17 (January 2008)
20. McCune, J.M., Parno, B.J., Perrig, A., Reiter, M.K., Isozaki, H.: Flicker: an execution infrastructure for tcb minimization. SIGOPS Oper. Syst. Rev. 42(4), 315–328 (2008)
21. Mostowski, W., Vullers, P.: Efficient U-Prove implementation for anonymous credentials on smart cards. In: Rajarajan, M., Piper, F., Wang, H., Kesidis, G. (eds.) SecureComm 2011. LNICST, vol. 96, pp. 243–260. Springer, Heidelberg (2012)

Dedicated Hardware for Attribute-Based Credential Verification

Geoffrey Ottoy¹, Jorn Lapon², Vincent Naessens²,
Bart Preneel³, and Lieven De Strycker¹

¹ KAHO Sint-Lieven, DraMCo Research Group,
Gebroeders de Smetstraat 1, 9000 Gent, Belgium
geoffrey.ottoy@kahosl.be
<http://www.dramco.be/>

² KAHO Sint-Lieven, MSEC Research Group,
Gebroeders de Smetstraat 1, 9000 Gent, Belgium
jorn.lapon@kahosl.be
<http://www.msec.be>

³ KU Leuven, COSIC and IBT,
Kasteelpark Arenberg 10, bus 2446, 3001 Leuven-Heverlee, Belgium
<http://www.esat.kuleuven.be/cosic/>

Abstract. Attribute-based credentials systems offer a privacy-friendly solution to access electronic services. In this field, most research has been directed into optimizing the prover operations and exploring the usability boundaries on mobile platforms like smart cards and mobile phones. This research assumes that the verification of credential proofs occur at a powerful back end. However, a broad range of (embedded) applications lack this powerful back end.

This article shows that hardware accelerators for modular exponentiations, greatly reduce the run time of applications that require credential verification in an embedded context. In addition, when verification requires a considerable amount of the total run time (i.e., communication included), the use of dual-base (simultaneous) exponentiation hardware further increases the overall performance.

All tests have been performed in a practical setup between a smart-phone and an embedded terminal using NFC communication.

Keywords: attribute-based credentials, dedicated hardware, embedded, NFC.

1 Introduction

Today, a lot of personal information is revealed through the use of electronic services. Although service providers require only little information for running their service, often many more attributes are collected. For instance, when a user needs to prove that he is older than 18 to access a gambling site, when using an electronic identity card based on certificate technology, other personal data are disclosed as well. Moreover, all actions of the same user can be linked. Attribute-based credentials [12,10,11,8] offer a more privacy-friendly solution to

access electronic services. First, they support selective disclosure. This means that a user can opt to reveal only a subset of possible properties of the attributes embedded in the credential. For instance, a credential with the user's date of birth as an attribute, can be used to prove that the owner is over 18, without disclosing the exact date of birth or other attributes. Second, some attribute-based credential systems also support accountability and/or unlinkability of transactions. The former allows to identify individuals conditionally (e.g., in case of abusive behavior). The latter means that multiple actions of the same user cannot be linked.

Two major classes of attribute-based credential systems exist. The first class [8] uses *blind signatures* [12] in order to break the link between the issuer and the user's credential. In short, the issuer signs the user's credential, without knowing the resulting signature value. As a result, when the credential is used, even if the issuer and relying party share information, it cannot be linked to the issuance phase. In order to make multiple transactions unlinkable, a batch of credentials can be issued and for each anonymous transaction, a new credential will be used. The second class, also called CL-based credentials [10,11], uses *zero-knowledge proofs* to break the link between the credential issuance and its use. During authentication, the user proves in zero-knowledge to the relying party, that she has a genuine credential (i.e., certified by the trusted issuer). Here, zero-knowledge means that, in the general case, nothing is disclosed except the fact that the credential is genuine, and the prover is the holder of the credential (i.e., she knows the corresponding private key). An implementation is available for each class: **U-Prove** [32], implements the first class, and **Identity Mixer** [17], belongs to the second class.

A major disadvantage of attribute-based credential technologies is their poor performance – both in terms of processing power and memory footprint – compared to certificate technology. For instance, in case of **Identity Mixer**, zero-knowledge proofs require multiple exponentiations by the prover and the verifier. This currently leads to unacceptable response times in many application domains. Existing research focuses on optimizing the prover operations and exploring the usability boundaries on mobile platforms like smart cards and mobile phones. This research assumes that the verification of credential proofs occurs at a powerful back end. However, selective disclosure is also very relevant in settings where a user authenticates to a terminal that is not connected with a powerful back end. For instance, a stand-alone cigarette or beverage vending machine wants to verify if the user is older than 18. Similarly, a waste disposal center wants to restrict access to local inhabitants. Therefore, users need to prove to live in a certain city before they are granted access to that location. In both situations, users do not want to release too much information to discourage extensive profiling.

This paper especially targets hardware support to accelerate the *verification* of credential attributes released during authentication. We, thereby, focus on CL-based credentials. The main contributions are threefold. First, a hardware platform is presented that supports the acceleration of a wide range of

cryptographic operations. Second, the feasibility of the platform is demonstrated by integrating it in an access control terminal with NFC communication capabilities. The terminal is applied in a scenario where a user stores an attribute-based credential on her NFC-enabled smartphone, and selectively discloses some attributes to get access. Third, the terminal is extensively evaluated. More specifically, the performance is compared to a terminal without dedicated hardware support. The effect of the number of attributes and whether they are revealed, is studied as well.

The rest of this paper is structured as follows. Section 2 and 3 respectively point to related work and offer background information. Thereafter, the case study is presented in Section 4. Section 5 focuses on the realization of the prototype. Next, the hardware platform is evaluated in Section 6. We end up with conclusions in Section 7.

2 Related Work

Attribute-based credentials based on CL signatures, require substantially more computational effort, for both proving and verifying, than traditional authentication technologies. This is mainly due to a substantial number of exponentiations during zero-knowledge proofs. Hence, performance might be a bottleneck, especially if credentials are stored on mobile carriers (like smart cards or mobile devices), and proofs are generated on those carriers. In literature, multiple implementations of attribute-based credentials in such environments have been presented. Bichsel et al. [4] presented a full Java Card implementation of attribute-based credentials. Computing the credential proof took about 7.4 s for a 1280-bit modulus, and up to 16.5 s for a 1984-bit modulus. In [3] and [33], to increase efficiency, the computation of the proof is divided between a tamper-proof smartcard and a partially trusted host. This approach, however, requires partial trust on the host. Recently, Vullers and Alpár [38] also implemented the CL-based prover protocol on a MULTOS card. The computation of the proof only takes about 1.1s (for a 1024-bit modulus) for a simple credential proof. This is an interesting result that shows that also CL-based credentials on smartcards may become practical in the near future.

In contrast to the CL-based schemes, there are also prototypes implementing U-Prove [32] attribute-based credentials, which take about 5 s for showing a credential [37]. Later, Mostowski and Vullers [20] implemented the same protocol on a MULTOS [15] card with better support for modular arithmetic, resulting in only about 0.5 s. Note that in order to preserve unlinkability, the U-Prove system requires the issuance of a new credential for each transaction, which may quickly exhaust the EEPROM of the card [4].

Unfortunately, all of these solutions evaluate and/or optimize the performance of the prover side. Our work, on the other hand, mainly focuses on the fast verification of attribute-based credentials in resource-constrained environments. It is clear that faster – and simultaneous – calculation of modular exponentiations may reduce the response times at the verifier, and hence, increase the overall

performance of an attribute-based authentication procedure. Many hardware implementations and optimizations are presented in the literature.

The most straightforward way of performing a modular exponentiation is by repeated squarings and multiplications to get the final result [6,34]. The square and multiply step can be performed either in parallel [22] or sequentially [13]. Booth encoding is also used to increase the number of “00”-combinations of exponent bits in simultaneous exponentiation [18]. This reduces the number of multiplications and thus results in a speedup. However, more memory is required to store precomputed values. Also for simultaneous exponentiations, the square-and-multiply approach can be followed. A major drawback lies in the fact that the required amount of memory increases exponentially with the number of exponentiations carried out simultaneously.

As most (if not all) exponentiation algorithms rely on multiplications, it is not surprising that many implementations of hardware multipliers have been described in the literature. Almost all of these multipliers rely on Montgomery’s algorithm [19]. This algorithm allows for very efficient hardware implementations. The algorithm’s main disadvantage is the carry propagation. Several architectures have been proposed that combine Montgomery multiplication with either a redundant radix number system [31,16] or the Residue Number System [2,28] to cope with this problem. Unfortunately, these implementations have several other drawbacks. For instance, a lot of preprocessing of the operands is required.

The best known class of Montgomery multipliers is the systolic array architecture. Nedja and Mourelle [22] have shown that for operands larger than 512 bit, the systolic array implementation improves the *time* \times *area* product over other implementations. They compared their work with that of Blum and Paar [5], which was one of the first milestones for systolic array implementations. Systolic array Montgomery multipliers have been implemented in 2-dimensional [22,21] and 1-dimensional designs [24]. The 1-dimensional variant requires less silicon to implement, but is slower than the 2-dimensional implementation.

The k -partition method is a different way to speed up Montgomery multiplication [23]. In this method, k partitions operating in radix 2^k , each of which computes a part of the total result. The fastest multiplication would execute in n/k cycles. The complexity of the partitions, however, is higher than for a standard Montgomery multiplier, but often the FPGA’s on-board multipliers are used to implement the partitions.

Montgomery multipliers can use Booth encoding [7] to replace the use of precomputed hard multiples in the processing elements. In [29], this is combined with left-shifting of the input operands –instead of right-shifting the result– to reduce the critical path. In [1,30], more flexibility is added to the design by varying the length of the processing elements and by implementing the Booth encoding in hardware rather than requiring precomputed hard multiples of the input operands.

Tenca and Koç [35] came up with a scalable pipelined design where the circuit of the Montgomery multiplier is split into “word size” processing elements. Since then, several improvements have been made to their original design [36,39].

3 Verification of CL Based Credentials Proofs

We briefly recall the protocol for the verification of a CL-based credential proof, which is part of the signature proof of knowledge. This is also the protocol that will be running on the terminal. For a full definition of the CL signature scheme, the signature proof of knowledge and the construction of the proof, we refer to Appendix A.

The verifier verifies the proof as follows:

1. Compute:

$$\hat{Z} = \left(\frac{Z}{\prod_{j \in A_r} R_j^{m_j} (A')^{2^{l_e-1}}} \right)^{-c} (A')^{\hat{e}} \left(\prod_{i \in A_h} R_i^{m_i} \right) (S^{\hat{v}'}) \pmod n \quad (1)$$

2. Verify results:

$$\begin{aligned} \hat{m}_i &\stackrel{?}{\in} \{0, 1\}^{l_m+l_\phi+l_H+1} \quad \forall i \in A_h \\ \hat{e} &\stackrel{?}{\in} \pm\{0, 1\}^{l_{e'}+l_\phi+l_H+1} \\ c &\stackrel{?}{=} H(\text{context}, A', \hat{Z}, n_1) \end{aligned}$$

The proof is rejected if any of these checks fail.

4 Realization of the Embedded Terminal

4.1 Platform Specification

To analyze attribute-based credential verification on an embedded terminal, we use an FPGA-based test platform [25,27]. This allows us to easily change the hardware configuration of the design. Furthermore, our platform implements a standard embedded design setup, so conclusions drawn with this platform can be generalized to non-FPGA-based systems.

Fig. 1 shows the structure of the embedded test terminal. It is based on a Xilinx ML605 evaluation board housing a Virtex 6 FPGA. The central controller is a MicroBlaze embedded processor running embedded Linux. Using Linux (in our case the PetaLinux distribution¹) has the big advantage that standard libraries become available on the embedded system; specifically we use GMP² for the large number arithmetic and libnfc for the NFC communication³.

An NXP PN532 development kit implements the RF front end for the NFC communication and is able to emulate different types of NFC devices like,

¹ PetaLinux software development kit information:

<http://www.xilinx.com/tools/petalinux-sdk.htm>

² GMP project website: <http://gmplib.org/>

³ libnfc project website: <http://www.libnfc.org/>

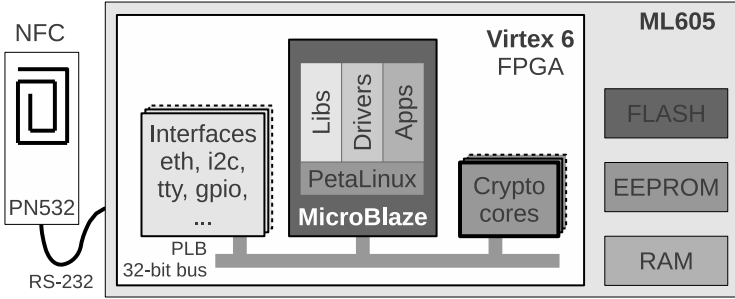


Fig. 1. Embedded test terminal setup

Mifare, ISO14443-A/B, DEP, Felica, etc. We configure the PN532 as an NFC initiator scanning for ISO14443-A tags. A Google Nexus S smartphone running CyanogenMod 9.1 in card emulation mode acts as the prover.⁴

Hardware accelerators are provided for both the hashing (SHA-256) and for the exponentiations. Furthermore, the platform offers various types of volatile and non-volatile memory as well as several types of I/O (e.g., LEDs and switches). The complete platform, both CPU and hardware cores, run at a frequency of 100 MHz.

4.2 Crypto Core Specification

The hardware accelerator for the exponentiations [26] is specifically designed to carry out dual-base exponentiations (2) simultaneously, but it can also be used to compute a single-base exponentiation or a modular multiplication. The kernel of the accelerator is a pipelined 1-dimensional systolic array Montgomery multiplier. The design and hardware documentation can be found at: http://opencores.org/projects/mod_sim_exp.

$$g_0^{e_0} \cdot g_1^{e_1} \bmod m \quad (2)$$

The lengths of the exponents can be chosen freely by the controlling software. However, the construction of the hardware requires that both exponents are of the same length in case of dual-base exponentiations. This means that a shorter exponent needs to be padded with preceding zeros to match the length of the longer exponent. It also needs to be noted that all exponents are interpreted as positive integers. Modular inverses are computed in software.

Run times for multiplication and average run times for single-base exponentiation and dual-base exponentiation can be computed by equations (3), (4), (5) respectively.

⁴ Cyanogenmod project website: <http://www.cyanogenmod.org/>

$$t_{\text{mult.}} = \left\lceil \frac{n}{s} + 2 \cdot (n - 1) \right\rceil / f_{\text{clk}} \quad (3)$$

$$t_{1\text{-exp.}} = \frac{3}{2} \cdot w_0 \cdot t_{\text{mult.}} \quad (4)$$

$$t_{2\text{-exp.}} = \left\lceil \frac{3}{2} \cdot (w_0 - w_1) + \frac{7}{4} \cdot w_1 \right\rceil \cdot t_{\text{mult.}} \quad (5)$$

With:

n the operand length [# bits]

s the pipeline stage length [# bits]

f_{clk} the clock frequency [Hz]

w_0, w_1 the length of the exponents e_0, e_1 where $w_1 \leq w_0$ [# bits]

Note that due to the bus traffic and latency introduced by the OS, an overhead of 5 to 6 ms needs to be taken into account for the exponentiations.

4.3 Implementation Details

To apply the hardware accelerator, the credential verification protocol (1) is rearranged and split into a modular product of dual-base exponentiations as presented in (6). The A' exponent ($c \cdot 2^{l_e - 1} + \hat{e}$) as well as all the modular inverses are computed in software. All multiplications are computed in hardware.

$$\hat{Z} = \underbrace{(Z^{-1})^c \cdot R_0^{\hat{m}_0}}_{(a)} \cdot \underbrace{(A')^{(c \cdot 2^{l_e - 1} + \hat{e})} \cdot S^{\hat{v}'}}_{(b)} \cdot \underbrace{\prod_{j \in A_r} (R_j^{-1})^{m_j} \cdot \prod_{i \in A_h} R_i^{\hat{m}_i}}_{\text{attribute-dependent}} \pmod n \quad (6)$$

Special attention is required when using the hardware to perform dual-base exponentiations. From equation (5), it follows that a minimal run time can be achieved when the length of the exponents differs as little as possible. That is why the verification step is rearranged and computed as shown in (6). For the attribute-based part, the revealed and hidden attributes are grouped in separate dual-base exponentiations where possible.

5 Tests and Results

In prior work, only little timing results on the verification of attribute-based credentials are available and they are often hard to compare. Nevertheless, Table 1 presents a comparison of our embedded terminal with comparable implementations available in literature. A precise comparison is difficult because of the different architectures, processor speeds and implementations. Still, the figures clearly show the value of our hardware accelerated solution with respect to general purpose devices. Note that the measurements by Dietrich [14] are performed based on the DAA verification protocol, which is closely related to the CL credential verification used in this paper.

Table 1. Timing results (in ms) for the verification of attribute-based credentials, with only a master secret, compared to prior work

n	Bichsel [3]	Dietrich [14](DAA)			This paper
	Intel Core 2-Duo	Intel Core 2-Duo	P910	Nokia 6131	
	P9600 @ 2.53GHz	T7500 @ 2.2GHz	ARM9 @ 156 MHz	ARM9 @ 229 MHz	
1024	78	40	8240	16960	124
1536	187	-	-	-	215
2048	375	110	22100	64270	-

5.1 High Level Description of Tests

As pointed out before, we are interested in the run time of the credential verification protocol on an embedded terminal. To that end, different cases are evaluated on the test platform described in Section 4. All tests are performed with a modulus length of both 1024-bit and 1536-bit. The hardware accelerator has 16-bit stages and operates at 100 MHz; this is also the clock frequency of the embedded CPU.

With these test cases, several questions are addressed:

- What is the effect of hardware acceleration on the run time? We will compare an embedded software implementation –using GMP on the MicroBlaze CPU– with an implementation using hardware offload, both using single-base exponentiations and dual-base exponentiations.
- What is the effect of the number of attributes in the credentials on the verification run time?
- What is the effect of the number of hidden/revealed attributes?
- To compare the verification run time with the total protocol run time, we will also take into account the communication overhead.

5.2 Verification Performance

As a first test, the verification of a credential with a single attribute (i.e., the master secret) is evaluated. Table 2 presents the run time of an embedded software implementation (*SW*), compared with an implementation using the hardware accelerator; both with single-base (*1-exp*) and dual-base exponentiations (*2-exp*). The table also shows the NFC communication time (i.e., the time required to transmit the proof π) as part of the overall protocol run time.

As expected, the verification with embedded software takes significantly longer than the hardware accelerated implementations. For a modulus of 1536 bits, it takes about 14 seconds or 95% of the total run time. The hardware accelerator clearly improves the performance of the verification and hence the overall run time. Moreover, the main share of the run time shifts towards the communication over NFC. The Table also shows that the communication speed is not constant, even for the same modulus lengths. Although the results are averages, this is

Table 2. Comparison of the average timing results for the credential proof verification on an embedded platform, where all computations are performed in software (*SW*), and where a hardware accelerator is used for single-base *1-exp* and dual-base *2-exp* exponentiations. The credential contains a single hidden attribute (m_0 , the master secret).

n	Implementation	NFC Communication		Verification		Hash and check		Total [ms]
		[ms]	[%]	[ms]	[%]	[ms]	[%]	
1024	<i>SW</i>	748	12	5696	88	7	0	6451
	<i>1-exp</i>	733	82	159	18	7	1	899
	<i>2-exp</i>	721	85	124	15	7	1	852
1536	<i>SW</i>	765	5	13846	95	9	0	14620
	<i>1-exp</i>	782	76	240	23	9	1	1031
	<i>2-exp</i>	768	77	215	22	9	1	992

mainly due to overhead introduced in the operating system; keep in mind that Linux is not a real-time operating system.

The tests also illustrate that a different size of the modulus has much more effect (relatively) on the verification time than on the communication time. This is the case for all three implementations. As could be expected, the time for hashing and making the necessary checks is negligible with respect to the time required for running the verification protocol.

5.3 Influence of the Number of Attributes

Second, the effect of the number of attributes in the credential is examined. Fig. 2 shows the run time for the communication (a) and the verification (b) with respect to the number of attributes in the credential. Fig. 2(b) also presents the gain in run time of dual-base exponentiations with respect to single-base exponentiations, defined as:

$$\text{Gain} = \frac{T_{1-exp} - T_{2-exp}}{T_{1-exp}} \cdot 100\%$$

Increasing the number of attributes increases both communication time and verification time. This is obvious, as the proof π will also be larger. If a single-base exponentiation accelerator is used, the verification time increases linearly with the number of attributes. However, if dual-base exponentiations are used, the verification time increases stepwise. This is due to the fact that the time for computing a dual-base exponentiation is comparable to a single-base exponentiation. Hence, based on equation 6, when verifying a credential with an odd number of attributes, our dual-base exponentiations can be used to their full extent. In contrast, in the case of an even number of attributes, a single-base exponentiation is required. In other words, increasing the number of attributes from an even number to an odd number, results in replacing a single-base exponentiation by a dual-base exponentiation, which requires only a small overhead.

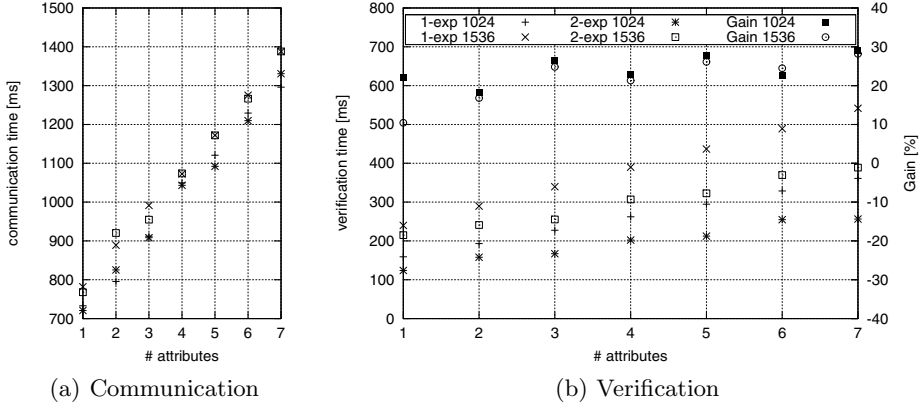


Fig. 2. Influence of the number of attributes on the run time. All attributes remain hidden.

Increasing the number of attributes from an odd number to an even number results in having an additional single-base exponentiation.

As can be seen, verifying more attributes also increases the gain for using dual-base exponentiations instead of single-base exponentiations (e.g., about 30% gain for a credential with 7 attributes). The gain is roughly the same for both 1024-bit and 1536-bit moduli.

5.4 Influence of the Number of Revealed Attributes

Finally, the influence of the number of revealed attributes on the run time has been examined. The number of revealed attributes varies from 0 to 6 (i.e., the master secret is never revealed). The communication time and verification run time is set out as well as the gain for using dual-base exponentiations (Fig. 3).

Both the communication time and verification time decrease with an increasing number of revealed attributes. This is because the exponent m_j of a revealed attribute is shorter than the exponent \hat{m}_i of a hidden attribute. Again, there is a linear relationship when single exponentiations are used and a stepwise behavior for dual-base exponentiations. Note that for our tests the size of the attribute values was 256 bits, while in reality this will often be much smaller. In fact, an attribute representing, for instance, the user’s gender, could require only a single bit. Hence, the decrease in time when releasing more attributes would become even more significant.

6 Evaluation

It is clear that for embedded applications the use of our hardware accelerator for credential verification greatly increases the feasibility in terms of run time. Compared to a powerful back end as is used in [3,14], the run times in this

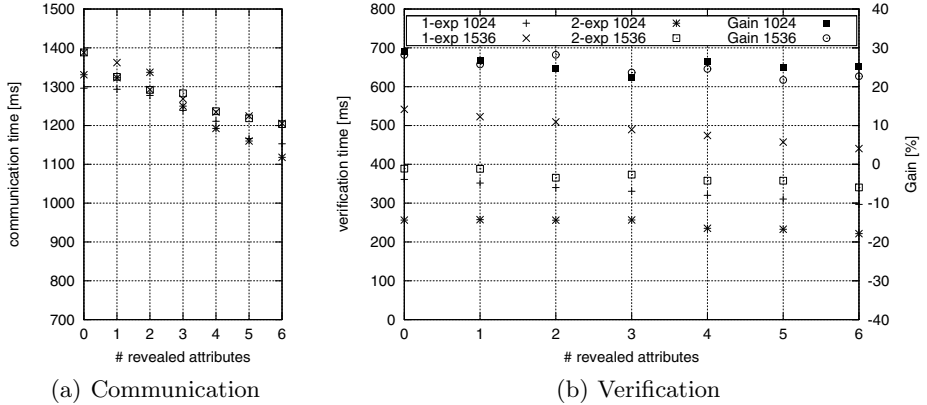


Fig. 3. Influence of the number of revealed attributes on the run time. The credential contains 7 attributes (master secret included).

embedded context are of the same order of magnitude. Thus, applications such as physical access control (e.g., opening a door) or stand-alone vending machines that support privacy, anonymity and/or unlinkability become feasible.

Note that for testing purposes, the current hardware accelerator was implemented on FPGA. For several applications where form factor or energy consumption are key requirements, this is undesirable. However, the same accelerator can be implemented on ASIC (as a first step) or a custom IC (a final product), which are better suited for commercial applications. The advantages demonstrated with our prototype implementation remain valid in these settings as well.

The current hardware accelerator is not tamper-proof and resistant against side-channel attacks. For applications that only require credential verification, this is not problematic. If however, disclosed information must not be learned by other parties than the verifier, extra measures should be taken. For instance, to prevent timing analysis (which may reveal for instance if more or less attributes are disclosed), exponentiations should be performed with a constant timing; more specifically with the worst case timing: $(2 \cdot w_0 \cdot t_{\text{mult}})$. Obviously this means a decrease in performance, which is the same for both single-base and dual-base exponentiations. Hence, multi-base exponentiations get even more attractive in this setup.⁵

The scenario presented in this paper, in which NFC is used for communication, shows that the gain between single or dual-base exponentiations is only marginal with respect to the overhead caused by the communication. The figures in this paper were obtained using NFC @ 106 kbits. Clearly, faster communication (e.g., NFC @ 424 kbit/s or Bluetooth v2.1 @ +1 Mbit/s) makes the use of our dual-base exponentiation accelerator more interesting.

⁵ Note that for an l -base exponentiation, 2^l n -bit memory locations are required.

The efficiency of the dual-base hardware accelerator is maximal in the case of exponents of the same length. However, in the prototype, the difference in the size of the exponents in dual-base exponentiation (b) of Eq. 6, is significant. A solution is to split exponentiation $S^{\hat{v}'}$ into a dual-base exponentiation $S_1^{\hat{v}'_1} \cdot S_2^{\hat{v}'_2}$ with smaller exponents (see [9,4] for more details). This, however, may require a slight modification of the prover protocol.

In this article we only examined the speedup of the verification of attribute-based credential proofs on embedded platforms. Note that for a complete setup, certificate revocation should also be supported.

7 Conclusions

In this paper, we present the use of a hardware accelerator for modular exponentiations, in order to reduce the run time of applications that require attribute-based credential verification in an embedded context. In addition, the use of dual-base (simultaneous) exponentiation hardware may further increase the overall performance. This is especially important when the overhead caused by communication can be decreased and a large number of attributes are included in the credential.

In future work, the efficiency of communication handover (e.g., from NFC to Bluetooth) could be studied, to decrease the overhead of the communication. More complex credential proofs such as range proofs and credential revocation were not implemented. Future work could be directed to find out the impact of these additional features on the performance of the verification and the requirements of the embedded platform (e.g., connectivity with a revocation server).

References

1. Amberg, P., Pinckney, N., Harris, D.M.: Parallel high-radix montgomery multipliers. In: 2008 42nd Asilomar Conference on Signals, Systems and Computers, pp. 772–776 (October 2008)
2. Bajard, J.-C., Didier, L.-S., Kornerup, P.: An rns montgomery modular multiplication algorithm. In: Proceedings of the 13th IEEE Symposium on Computer Arithmetic, pp. 234–239 (July 1997)
3. Bichsel, P., Camenisch, J., De Decker, B., Lapon, J., Naessens, V., Sommer, D.: Data-minimizing authentication goes mobile. In: De Decker, B., Chadwick, D.W. (eds.) CMS 2012. LNCS, vol. 7394, pp. 55–71. Springer, Heidelberg (2012)
4. Bichsel, P., Camenisch, J., Groß, T., Shoup, V.: Anonymous credentials on a standard java card. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, CCS 2009, pp. 600–610. ACM, New York (2009)
5. Blum, T., Paar, C.: Montgomery modular exponentiation on reconfigurable hardware. In: Proceedings of the 14th IEEE Symposium on Computer Arithmetic, pp. 70–77 (1999)
6. Blum, T., Paar, C.: High-radix montgomery modular exponentiation on reconfigurable hardware. IEEE Transactions on Computers 50(7), 759–764 (2001)

7. Booth, A.: A Signed Binary Multiplication Technique. *Quarterly Journal of Mechanics and Applied Mathematics* 4(2), 236–240 (1951)
8. Brands, S.: *Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy*. MIT Press, Cambridge (2000)
9. Brickell, E., Camenisch, J., Chen, L.: Direct anonymous attestation. In: *Proceedings of the 11th ACM Conference on Computer and Communications Security*, pp. 132–145. ACM (2004)
10. Camenisch, J.L., Lysyanskaya, A.: An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In: Pfitzmann, B. (ed.) *EUROCRYPT 2001*. LNCS, vol. 2045, pp. 93–118. Springer, Heidelberg (2001)
11. Camenisch, J., Lysyanskaya, A.: A signature scheme with efficient protocols. In: Cimato, S., Galdi, C., Persiano, G. (eds.) *SCN 2002*. LNCS, vol. 2576, pp. 268–289. Springer, Heidelberg (2003)
12. Chaum, D.: Security without identification: transaction systems to make big brother obsolete. *Communications of the ACM* 28(10), 1030–1044 (1985)
13. de la Piedra, A., Touhafi, A., Cornetta, G.: Cryptographic accelerator for 802.15.4 transceivers with key agreement engine based on montgomery arithmetic. In: *2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, pp. 1–5 (November 2011)
14. Dietrich, K.: Anonymous credentials for java enabled platforms: A performance evaluation. In: Chen, L., Yung, M. (eds.) *INTRUST 2009*. LNCS, vol. 6163, pp. 88–103. Springer, Heidelberg (2010)
15. France-Massay, T.: *MULTOS - the high security smart card OS*. Technical report, MAOSCO Limited (2005)
16. He, Y., Chang, C.-H.: A new redundant binary booth encoding for fast 2^n -bit multiplier design. *IEEE Transactions on Circuits and Systems I: Regular Papers* 56(6), 1192–1201 (2009)
17. *Specification of the Identity Mixer cryptographic library – version 2.3.2*, IBM Research – Zurich (2010)
18. Lai, Y.-P., Chang, C.-C.: An efficient multi-exponentiation scheme based on modified booth’s method. *International Journal of Electronics* 90(3), 221–233 (2003)
19. Montgomery, P.L.: Modular multiplication without trail division. *Mathematics of Computation* 44(170), 519–521 (1985)
20. Mostowski, W., Vullers, P.: Efficient U-prove implementation for anonymous credentials on smart cards. In: Rajarajan, M., Piper, F., Wang, H., Kesidis, G. (eds.) *SecureComm 2011*. LNICST, vol. 96, pp. 243–260. Springer, Heidelberg (2012)
21. Nedjah, N., de Macedo Mourelle, L.: Reconfigurable hardware implementation of montgomery modular multiplication and parallel binary exponentiation. In: *Proceedings of the Euromicro Symposium on Digital System Design*, pp. 226–233 (2002)
22. Nedjah, N., Mourelle, L.M.: Three hardware architectures for the binary modular exponentiation: sequential, parallel, and systolic. *IEEE Transactions on Circuits and Systems I: Regular Papers* 53(3), 627–633 (2006)
23. Neto, J.C., Tenca, A.F., Ruggiero, W.V.: A parallel k-partition method to perform montgomery multiplication. In: *2011 IEEE International Conference on Application-Specific Systems, Architectures and Processors, ASAP*, pp. 251–254 (September 2011)
24. Örs, S.B., Batina, L., Preneel, B., Vandewalle, J.: Hardware implementation of a montgomery modular multiplier in a systolic array. In: *Proceedings of the International Parallel and Distributed Processing Symposium*, p. 8 (April 2003)

25. Ottoy, G., Martens, J., Saeys, N., Preneel, B., De Strycker, L., Goemaere, J.-P., Hamelinckx, T.: A Modular Test Platform for Evaluation of Security Protocols in NFC Applications. In: De Decker, B., Lapon, J., Naessens, V., Uhl, A. (eds.) CMS 2011. LNCS, vol. 7025, pp. 171–177. Springer, Heidelberg (2011)
26. Ottoy, G., Preneel, B., Goemaere, J.-P., De Strycker, L.: Flexible design of a modular simultaneous exponentiation core for embedded platforms. In: Brisk, P., de Figueiredo Coutinho, J.G., Diniz, P.C. (eds.) ARC 2013. LNCS, vol. 7806, pp. 115–121. Springer, Heidelberg (2013)
27. Preneel, B., Goemaere, J.-P., Stevens, N., De Strycker, L., Ottoy, G.: Open-source hardware for embedded security. In: EDN (2013)
28. Phillips, B.: Modular multiplication in the montgomery residue number system. In: Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1637–1640 (November 2001)
29. Pinckney, N., Harris, D.M.: Parallelized radix-4 scalable montgomery multipliers. In: Petraglia, A., Pedroni, V.A., Cauwenberghs, G. (eds.) Proceedings of the 20th Annual Symposium on Integrated Circuits and Systems Design, SBCCI 2007, Copacabana, Rio de Janeiro, Brazil, September 3-6, pp. 306–311. ACM (2007)
30. Pinckney, N., Harris, D.M.: Parallelized radix-4 scalable montgomery multipliers. *Journal of Integrated Circuits and Systems*, 39–45 (2008)
31. Shigemoto, K., Kawakami, K., Nakano, K.: Accelerating montgomery modulo multiplication for redundant radix-64k number system on the fpga using dual-port block rams. In: IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, EUC 2008, vol. 1, pp. 44–51 (December 2008)
32. Christian Paquin Stefan Brands. U-Prove cryptographic specification v1.0, Microsoft Corporation (2010)
33. Sterckx, M., Gierlichs, B., Preneel, B., Verbauwhede, I.: Efficient implementation of anonymous credentials on Java Card smart cards. In: First IEEE International Workshop on Information Forensics and Security, WIFS 2009, pp. 106–110 (December 2009)
34. Sutter, G.D., Deschamps, J.-P., Imana, J.L.: Modular multiplication and exponentiation architectures for fast rsa cryptosystem based on digit serial computation. *IEEE Transactions on Industrial Electronics* 58(7), 3101–3109 (2011)
35. Tenca, A.F., Koç, Ç.K.: A scalable architecture for montgomery multiplication. In: Koç, Ç.K., Paar, C. (eds.) CHES 1999. LNCS, vol. 1717, pp. 94–108. Springer, Heidelberg (1999)
36. Tenca, A.F., Koç, Ç.K.: A scalable architecture for modular multiplication based on montgomery’s algorithm. *IEEE Transactions on Computers* 52(9), 1215–1221 (2003)
37. Tews, H., Jacobs, B.: Performance issues of selective disclosure and blinded issuing protocols on java card. In: Markowitch, O., Bilas, A., Hoepman, J.-H., Mitchell, C.J., Quisquater, J.-J. (eds.) WISTP 2009. LNCS, vol. 5746, pp. 95–111. Springer, Heidelberg (2009)
38. Vullers, P., Alpár, G.: Efficient selective disclosure on smart cards using idemix. In: Fischer-Hübner, S., de Leeuw, E., Mitchell, C. (eds.) IDMAN 2013. IFIP AICT, vol. 396, pp. 53–67. Springer, Heidelberg (2013)
39. Zhou, Y., Wang, X.: An improved implementation of montgomery algorithm using efficient pipelining and structured parallelism techniques. In: Signals and Systems Conference (ISSC 2010), pp. 7–11. IET Irish (June 2010)

A CL-Based Credentials

CL Signature Scheme. We briefly recall the CL-signature scheme, with a signer S and verifier V , for blocks of L messages as presented in [11] and implemented in the **Identity Mixer** library [17]:

$I : (pk_{Sig}, sk_{Sig}) \leftarrow \text{setup}_{\text{CL}}(l_n)$

Choose a special RSA modulus $n = pq$ of length l_n with $p = 2p' + 1, q = 2q' + 1$ where p, q, p' and q' are prime. Choose, uniformly at random $R_0, \dots, R_{L-1}, S, Z \in_R \mathcal{QR}_n$ with public key $pk_{Sig} = (n, R_0, \dots, R_{L-1}, S, Z)$ and secret key $sk_{Sig} = (p)$.

$S : (\sigma) \leftarrow \text{sign}_{\text{CL}}(m_0, \dots, m_{L-1}, sk_{Sig})$

Let l_m be a parameter defining the message space as $m_i \in \pm\{0, 1\}^{l_m}$ for $0 < i < L$. Choose a random prime e of length $l_e > l_m + 2$ and a random number $v \in_R \pm\{0, 1\}^{l_n + l_m + l_r}$, with l_r a security parameter, and compute the signature $\sigma = (A, e, v)$ on (m_0, \dots, m_{L-1}) such that $A^e \equiv \frac{Z}{R_0^{m_0} \dots R_{L-1}^{m_{L-1}} S^v} \pmod n$. The latter requires knowledge of the order of the subgroup to compute the inverse of e .

$V : (Bool) \leftarrow \text{verify}(\sigma, m_0, \dots, m_{L-1}, pk_{Sig})$

Parse σ as a tuple (A, e, v) and return true if $Z \equiv A^e R_0^{m_0} \dots R_{L-1}^{m_{L-1}} S^v \pmod n$, $2^{l_e - 1} < e < 2^{l_e}$ and $m_i \in \pm\{0, 1\}^{l_m}$ for $0 < i \leq L$ holds, else return false.

Note that to fully benefit from the privacy features provided by attribute-based credentials, they should be used in combination with anonymous communication in order to prevent linking or identification through for instance IP address, MAC address, cookies, or browser identification. Nevertheless, even without anonymous communication, attribute-based credentials are superior and more privacy-friendly than traditional authentication technologies.

Signature Proof of Knowledge. Attribute-based credential authentication based on CL signatures (e.g., as in the **Identity Mixer** library), mainly consist of a signature proof of knowledge of a CL signature $\sigma = (A, e, v)$ on a nonce n_1 as recalled below. Note that the protocols only consider proving knowledge of a valid credential and selective disclosure. For more advanced protocols, such as interval proofs and enumeration, we refer the reader to the **Identity Mixer** specification [17]

$$\begin{aligned}
 & SPK\{(e, \{m_i : i \in A_h\}, v) : \\
 & \quad \frac{Z}{\prod_{i \in A_r} R_i^{m_i}} \equiv \pm A^e S^v \prod_{j \in A_h} R_j^{m_j} \\
 & \quad \forall i \in A_h : m_i \in \{0, 1\}^{l_m + l_\phi + l_H + 2} \\
 & \quad e - 2^{l_e - 1} \in \{0, 1\}^{l'_e + l_\phi + l_H + 2} \\
 & \quad \}(n_1)
 \end{aligned} \tag{7}$$

with A_h and A_r the set of hidden, resp. revealed attributes, l_m, l_H, l_e the bit length of the attributes, the challenge and e respectively, l_ϕ a security parameter that governs the statistical zero-knowledgeness and l'_e the size of the interval the e values are taken from. m_0 is called the master secret and is never revealed.

This signature proof of knowledge can be converted into the following two protocols:

Construction of the proof. After receiving the nonce n_1 from the verifier, the prover builds the proof as follows:

1. Randomize CL-Signature $\sigma = (A, e, v)$:

$$\begin{aligned} r_A &\in_R \{0, 1\}^{l_n + l_\phi} \\ A' &= AS^{r_A} \pmod n \\ v' &= v - er_A, \\ e' &= e - 2^{l'_e - 1} \end{aligned}$$

2. Compute 1st round:

$$\begin{aligned} \tilde{e} &\in_R \pm\{0, 1\}^{l_{e'} + l_\phi + l_H}, \tilde{v}' \in_R \pm\{0, 1\}^{l_v + l_\phi + l_H}, \tilde{m}_i \in_R \{0, 1\}^{l_m + l_\phi + l_H} \quad \forall i \in A_h \\ \tilde{Z} &= (A')^{\tilde{e}} \left(\prod_{i \in A_h} R_i^{\tilde{m}_i} \right) (S^{\tilde{v}'}) \pmod n \end{aligned}$$

3. Compute challenge:

$$c = H(\text{context}, A', \tilde{Z}, n_1)$$

4. Compute 2nd round:

$$\begin{aligned} \hat{e} &= \tilde{e} + ce' \\ \hat{v}' &= \tilde{v}' + cv' \\ \hat{m}_i &= \tilde{m}_i + cm_i \quad \forall i \in A_h \end{aligned}$$

Let the proof $\pi = (c, A', \hat{e}, \hat{v}', \hat{m}_i, m_j \forall i \in A_h \text{ and } \forall j \in A_r)$

Verification of the proof by the verifier. See Sect. 3.

Decentralized Ciphertext-Policy Attribute-Based Encryption Scheme with Fast Decryption

Y. Sreenivasa Rao and Ratna Dutta

Department of Mathematics
Indian Institute of Technology Kharagpur
Kharagpur-721302, India
{ysrao, ratna}@maths.iitkgp.ernet.in

Abstract. In this paper, we propose an efficient multi-authority decentralized ciphertext-policy attribute-based encryption scheme dCP-ABE-MAS for monotone access structures (MAS). Our setup is without any central authority (CA) where all authorities function entirely independently and need not even be aware of each other. The scheme makes use of the minimal authorized sets representation of MAS to encrypt messages, and hence the size of ciphertext is linear in the number of minimal authorized sets in MAS and the number of bilinear pairings is *constant* during decryption. We describe several networks that can use dCP-ABE-MAS to control data access from unauthorized nodes. The proposed scheme resists collusion attacks and is secure against chosen plaintext attacks in the generic bilinear group model over prime order bilinear groups.

Keywords: attribute-based encryption, decentralized, multi-authority, monotone access structure.

1 Introduction

In Attribute-Based Encryption (ABE), each user is ascribed a set of descriptive attributes (or credentials), and secret key and ciphertext are associated with an access policy or a set of attributes. Decryption is then successful only when the attributes of ciphertext or secret key satisfy the access policy. ABE is classified as Key-Policy ABE (KP-ABE) [3] or Ciphertext-Policy ABE (CP-ABE) [4] according to whether the secret key or ciphertext is associated with an access policy, respectively. Since the invention of ABE [2], several improved ABE schemes [3–6] have been proposed. All the foregoing ABE schemes make use of a single trusted central authority (CA) to control the universe of attributes and issue secret keys to users that should not be compromised at all. Consequently, the CA can decrypt every ciphertext in the system encrypted under any access policy by calculating the required secret keys at any time, this is the *key escrow* problem of ABE. A solution to help mitigate the key escrow problem is distributing the functionality of the CA over many potentially untrusted authorities in such a way that as long as some of them are honest, the system would still be

secure. An ABE with this mechanism is the so-called *multi-authority* ABE. In this scenario, each authority controls a different domain of attributes and issues attribute-related secret keys to users.

Chase [10] devised the first multi-authority ABE as an affirmative solution to the open problem posed by Sahai and Waters [2] that consists of one fully trusted centralized authority (CA) and multiple (attribute) authorities. Every user is assigned a unique global identifier and the keys from different authorities are bound together by this identifier to counteract the *collusion attack*—multiple users can pool their secret keys obtained from different authorities to decrypt a ciphertext that they are not individually entitled to. As CA holds the system’s master secret, it can decrypt all the ciphertexts in the system, thereby cannot the key escrow resists. The first CA-free multi-authority ABE is proposed by Lin et al. [9] wherein Distributed Key Generation (DKG) protocol and Joint Zero Secret Sharing (JZSS) protocol are deployed to remove CA. All authorities must interact to execute DKG and JZSS protocols during system setup phase. However, the scheme is collusion-resistant up to collusion of m users, where m is a system wide parameter that should be fixed during setup, and the number of JZSS protocol executions, the computation and communication costs are all linear in m . Chase and Chow [11] proposed CA-free multi-authority ABE with user privacy that resolves the key escrow problem using distributed Pseudo Random Functions (PRF). In this setting, each pair of authorities will communicate with each other via a 2-party key exchange protocol to generate users’ secret keys during setup phase that incurs $\mathcal{O}(N^2)$ communication overhead on the system, where N is the fixed number of authorities. The foregoing constructions [10, 9, 11] can only handle a set of fixed number of authorities at system initialization which exploit AND-gate access policies in key-policy setting to prevent unauthorized data access.

Müller et al. [15] gave two multi-authority CP-ABE schemes which employ one CA and several authorities where the authorities work independently from each other. However, the CA can still decrypt all ciphertexts in the system. The first construction uses Disjunctive Normal Form (DNF) access policies to annotate ciphertexts, thereby achieves constant computation cost during decryption. The second scheme realizes any Linear Secret Sharing Scheme (LSSS) access policy and hence the computation cost for successful decryption is linear in minimum number of attributes required to compute the target vector, i.e., a vector that contains the secret as one of its components. Lewko and Waters [8] proposed a novel multi-authority CP-ABE scheme without CA that is decentralized, where all authorities function entirely independently and need not even be aware of each other. The concept of global identifier introduced by Chase [10] is used to “link” attribute-related secret keys together that are issued to the same user by different authorities, this in turn achieves collusion-resistant among any number of users. The same scheme works on both composite order and prime order bilinear groups. The security of the former is given in random oracle model and the security of latter one is analyzed in the generic group model. In both cases, the monotone access structures are realized by LSSS, the ciphertext size

Table 1. Comparison of [8] with Our (dCP-ABE-MAS) Scheme

Scheme	Key Generation		Encryption			Decryption		Access Policy
	$E_{\mathbb{G}}$	User Secret Key Size	$E_{\mathbb{G}}$	$E_{\mathbb{G}_T}$	Ciphertext Size	$E_{\mathbb{G}_T}$	P_e	
[8]	2γ	$\gamma B_{\mathbb{G}}$	3α	$2\alpha + 1$	$2\alpha B_{\mathbb{G}} + (\alpha + 1)B_{\mathbb{G}_T} + \tau$	$\mathcal{O}(\beta)$	$\mathcal{O}(\beta)$	LSSS
Our	2γ	$\gamma B_{\mathbb{G}}$	$2k$	k	$2kB_{\mathbb{G}} + kB_{\mathbb{G}_T} + \tau$	-	2	any MAS

$E_{\mathbb{G}}$ (or $E_{\mathbb{G}_T}$) = number of exponentiations in a group \mathbb{G} (or \mathbb{G}_T , resp.), P_e = number of pairing computations, $B_{\mathbb{G}}$ (or $B_{\mathbb{G}_T}$) = bit size of an element of \mathbb{G} (or \mathbb{G}_T , resp.), α = size of LSSS access structure, β = minimum number of attributes required for decryption, γ = number of attributes annotated to a user secret key, k = number of minimal sets in MAS, τ = size of an access structure.

is linear in the size of the LSSS, and the number of pairings is linear in the minimum number of attributes that satisfy the LSSS. Liu et al. [17] devised a LSSS-realizable multi-authority CP-ABE system which has multiple CAs and authorities. The scheme is adaptively secure without random oracles unlike [8].

In all the multi-authority KP/CP-ABE schemes except the one (CA based) in [15] discussed so far, the size of ciphertext is linear in the size of monotone span program or the number of attributes that are associated with ciphertexts and the number of bilinear pairing computations is linear in the minimum number of attributes required for successful decryption. Constant computation and low communication cost access control schemes are more practical where the computing resources have limited computing power and bandwidth is the primary concern. For these reasons, we provide a solution to help mitigate the problem of large ciphertext size and linear-size number of bilinear pairings in designing multi-authority ABE schemes.

Our Contribution. We propose dCP-ABE-MAS, which is a multi-authority CP-ABE in a decentralized setting for any monotone access structure (MAS). Every MAS, \mathbb{A} , can uniquely be represented by a set \mathbb{A}_0 of minimal authorized sets in \mathbb{A} (see Section 2.1). This scheme has the same functionality as the most robust and scalable multi-authority CP-ABE [8] to date. Even though the schemes [11, 9] exclude the requirement of the CA, they are not fully decentralized as the number of authorities is fixed ahead of time and all authorities are communicating each other during system setup unlike [8]. That is why we compare (in Table¹ 1) our dCP-ABE-MAS only with the decentralized scheme² of [8] in view of prime order bilinear group setting.

The ciphertext size in [8] is linear in the size, α , of LSSS, while the size of ciphertext in our construction grows linearly with k , the number of minimal authorized sets in the MAS. For (t, n) -threshold policy, where $1 < t < n$, the value of $k = n!/(n-t)! t!$ which will be larger than n , whereas there exist a

¹ The description of all the symbols in Table 1,3,4 is given at the bottom of Table 1.

² The scheme that works on prime order bilinear group and the security is analyzed in the generic group model.

LSSS with size $\alpha = n$ to realize the (t, n) -threshold policy. However, there are several classes of MAS for which the value of k is constant but the size of the monotone span program (or LSSS) computing the MAS is at least polynomial in the number of attributes in the access structure. As a trivial case, if one uses a single AND-gate with n attributes, the value of k will be 1, while the size of LSSS is equal to n , i.e., $\alpha = n$. We now consider some non-trivial cases from [18]. Let $\mathbb{A}_0 = \{B_1 = \{a_1, \dots, a_{\lceil n/2 \rceil}\}, B_2 = \{a_{\lceil n/2 \rceil + 1}, \dots, a_n\}\}$ be the set of minimal sets for a MAS, \mathbb{A} , over n attributes a_1, \dots, a_n . Then, $k = 2$ and the size, α , of LSSS computing \mathbb{A} is at least $\mathcal{O}(n)$. Similarly, if $\mathbb{A}_0 = \{B_1 = \{a_1, \dots, a_{\lceil n/3 \rceil}\}, B_2 = \{a_{\lceil n/3 \rceil + 1}, \dots, a_{\lceil 2n/3 \rceil}\}, B_3 = \{a_{\lceil 2n/3 \rceil + 1}, \dots, a_n\}\}$ is the set of minimal sets for a MAS, \mathbb{A} , then $k = 3$ but the size, α , of LSSS computing \mathbb{A} is at least $\mathcal{O}(n)$ (for more details see Section 2.1 in [18]). Thus, in such cases, our dCP-ABE-MAS scheme exhibits shorter ciphertext. Moreover, our approach requires only 2 pairing computations to decrypt any ciphertext. The user secret key size is linear in the number of attributes associated with the user.

An inherent drawback of [8] is that every authority can independently decrypt every ciphertext in the system, if the set of attributes controlled by the authority satisfies the LSSS access structure associated with the ciphertext. However, this can be avoided if each authorized set contains attributes from at least two different authorities. The same problem can be eliminated in our dCP-ABE-MAS if each minimal authorized set contains attributes from at least two different authorities. This fact follows from satisfiability condition given in Definition 2.

We discuss how our dCP-ABE-MAS can provide attractive solutions to fine-grained access control in various network scenarios and compare our work with the existing works in the area. Additionally, our multi-authority scheme provides a mechanism for packing multiple messages in a single ciphertext. This in turn reduces network traffic significantly. The proposed scheme is proven to be collusion-resistant and is secure against chosen plaintext attacks in the generic bilinear group model. To the best of our knowledge, our proposed multi-authority CP-ABE scheme is the only scheme in a decentralized framework where the decryption time is constant for general MAS.

2 Preliminaries

Definition 1. Let \mathbb{G} and \mathbb{G}_T be multiplicative cyclic groups of prime order p . Let g be a generator of \mathbb{G} . A mapping $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is said to be bilinear if $e(u^a, v^b) = e(u, v)^{ab}$, for all $u, v \in \mathbb{G}$ and $a, b \in \mathbb{Z}_p$ and non-degenerate if $e(g, g) \neq 1_T$ (where, 1_T is the unit element in \mathbb{G}_T). We say that \mathbb{G} is a bilinear group if the group operation in \mathbb{G} can be computed efficiently and there exists \mathbb{G}_T for which the bilinear map $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is efficiently computable.

2.1 Access Structure

In this section, we briefly review the concept of general access structures [7].

Let U be the universe of attributes and $|U| = n$. Let $\mathcal{P}(U)$ be the collection of all subsets of U . Every subset of $\mathcal{P}(U) \setminus \{\emptyset\}$ is called an *access structure*. An access structure \mathbb{A} is said to be *monotone access structure* (MAS) if

$$\{C \in \mathcal{P}(U) | C \supseteq B, \text{ for some } B \in \mathbb{A}\} \subseteq \mathbb{A}.$$

The sets in \mathbb{A} are called the authorized sets and the sets not in \mathbb{A} are called the unauthorized sets with respect to the monotone access structure \mathbb{A} . Then every superset of an authorized set is again authorized set in MAS.

A set B in a monotone access structure \mathbb{A} is a *minimal authorized set* in \mathbb{A} if there exists a set $D (\neq B)$ such that $D \subseteq B$, then $D \notin \mathbb{A}$. The set of all minimal authorized sets of \mathbb{A} , denoted by \mathbb{A}_0 , is called the *basis* of \mathbb{A} . Then we can generate \mathbb{A} from its basis \mathbb{A}_0 as follows:

$$\mathbb{A} = \{C \in \mathcal{P}(U) | C \supseteq B, \text{ for some } B \in \mathbb{A}_0\}. \quad (1)$$

Lemma 1. *The monotone access structure \mathbb{A} given in Eq. (1) is generated uniquely from its basis \mathbb{A}_0 .*

Proof. Suppose \mathbb{A}' is a monotone access structure generated from \mathbb{A}_0 . Then $\mathbb{A}' = \{C' \in \mathcal{P}(U) | C' \supseteq B', \text{ for some } B' \in \mathbb{A}_0\}$. We shall prove that $\mathbb{A} = \mathbb{A}'$. Let $C \in \mathbb{A}$. Then by Eq. (1), we have $C \supseteq B$, for some $B \in \mathbb{A}_0$ and hence $C \in \mathbb{A}'$. Therefore, $\mathbb{A} \subseteq \mathbb{A}'$. Similarly, we can have $\mathbb{A}' \subseteq \mathbb{A}$. Thus, $\mathbb{A} = \mathbb{A}'$. \square

In sum, every monotone access structure can be represented by its basis.

Definition 2. *Let \mathbb{A} be a monotone access structure and \mathbb{A}_0 be its basis. A set, L , of attributes satisfies \mathbb{A} , denoted as $L \models \mathbb{A}$ if and only if $L \supseteq B$, for some $B \in \mathbb{A}_0$, and otherwise L does not satisfy \mathbb{A} , denoted as $L \not\models \mathbb{A}$.*

3 Decentralized CP-ABE System

A decentralized CP-ABE system is composed mainly of a set \mathcal{A} of authorities, a trusted initializer and users. The *only* responsibility of trusted initializer is generation of system global public parameters, which are system wide public parameters available to every entity in the system, once during system initialization. Each authority $A_j \in \mathcal{A}$ controls a different set U^j of attributes and issues corresponding secret attribute keys to users. We note here that all authorities will work independently. As such, every authority is completely unaware of the existence of the other authorities in the system. Each user in the system is identified with a unique global identity $ID \in \{0, 1\}^*$ and is allowed to request secret attribute keys from the different authorities. At any point of time in the system, each user with identity ID possesses a set of secret attribute keys that reflects a set L_{ID} of attributes, which we call an attribute set of the user with identity ID .

Let $U = \bigcup_{A_j \in \mathcal{A}} U^j$, where $U^{j_1} \cap U^{j_2} = \emptyset$, for all $j_1 \neq j_2$, be the attribute universe of the system. Due to lack of global coordination between authorities, different authorities may hold the same attribute string. To overcome such scenario, we can treat each attribute as a tuple consisting of the attribute string and

the controlling authority identifier, for example (“supervisor”, j), where the attribute “supervisor” is held by the authority A_j . Consequently, the attributes (“supervisor”, j_1) and (“supervisor”, j_2) will be considered as distinct as long as $j_1 \neq j_2$.

The decentralized CP-ABE system consists of the following five algorithms. **System Initialization**(κ). At the initial system setup phase, a trusted initializer chooses global public parameters GP according to the security parameter κ . Any authority or any user in the system can make use of these parameters GP in order to perform their executions.

Authority Setup(GP, U^j). This algorithm is run by every authority $A_j \in \mathcal{A}$ *once* during initialization. It accepts as input the global public parameters GP and a set of attributes U^j for the authority A_j and outputs public key $\text{Pub}A_j$ and master secret key $\text{Mk}A_j$ of the authority A_j .

Authority KeyGen(GP, ID, a , $\text{Mk}A_j$). Every authority executes this algorithm upon receiving a secret attribute key request from the user. It will take as input global public parameters GP, a global identity ID of a user, an attribute a hold by some authority and the master secret key of the corresponding authority. It returns a secret attribute key $\text{SK}_{a,\text{ID}}$ for the identity ID.

Encrypt(GP, M , \mathbb{A} , $\{\text{Pub}A_j\}$). This algorithm is run by an encryptor and it takes as input the global public parameters GP, a message M to be encrypted, an access structure \mathbb{A} , and public keys of relevant authorities corresponding to all attributes appeared in \mathbb{A} . It then encrypts M under \mathbb{A} and returns the ciphertext CT, where \mathbb{A} is embedded into CT.

Decrypt(GP, CT, $\{\text{SK}_{a,\text{ID}}|a \in L_{\text{ID}}\}$). On receiving a ciphertext CT, a decryptor with identity ID runs this algorithm with the input the global public parameters GP, a ciphertext CT which is an encryption of M under \mathbb{A} , and $\{\text{SK}_{a,\text{ID}}|a \in L_{\text{ID}}\}$ is a set of secret attribute keys obtained for the same identity ID. Then it outputs the message M if the user attribute set L_{ID} satisfies the access structure \mathbb{A} ; otherwise, decryption fails.

3.1 Security Model

Following [8], we define a security model in terms of a game which is carried out between a challenger and an adversary, where the challenger plays the role of all authorities. The adversary can corrupt authorities statically, i.e., the adversary has to announce the list of corrupted authorities before obtaining the public keys of honest authorities, whereas key queries can be made adaptively.

Setup. First, the challenger obtains global public parameters GP. The adversary announces a set $\mathcal{A}' \subset \mathcal{A}$ of corrupt-authorities. Now, the challenger runs Authority Setup algorithm for each honest authority and gives all public keys to the adversary.

Key Query Phase 1. The adversary is allowed to make secret key queries for the attributes coupled with user global identities (a , ID), where the attributes a

are held by honest authorities. The challenger runs Authority KeyGen algorithm and returns the corresponding secret keys $SK_{a, \text{ID}}$ to the adversary.

Challenge. The adversary submits two equal length messages M_0, M_1 and an access structure \mathbb{A} . The access structure \mathbb{A} must obey the following constraint. Let F be a set of attributes belonging to the corrupt-authorities that are in \mathbb{A} . For each identity ID , let F_{ID} be the set of attributes in \mathbb{A} for which the adversary has queried (a, ID) . For each identity ID , the attribute set $F \cup F_{\text{ID}}$ must not satisfy the access structure \mathbb{A} , i.e., $(F \cup F_{\text{ID}}) \not\models \mathbb{A}$. The adversary needs to give the challenger the public keys of corrupt-authorities whose attributes are in \mathbb{A} . Now, The challenger flips a random coin $\mu \in \{0, 1\}$ and runs Encrypt algorithm in order to encrypt M_μ under \mathbb{A} . The resulting challenge ciphertext CT^* is given to the adversary.

Key Query Phase 2. The adversary can make additional secret key queries for (a, ID) with the same restriction on the challenge access structure stated in Challenge phase.

Guess. The adversary outputs a guess bit $\mu' \in \{0, 1\}$ for the challenger's secret coin μ and wins if $\mu' = \mu$.

The advantage of an adversary in this game is defined to be $|\Pr[\mu' = \mu] - \frac{1}{2}|$, where the probability is taken over all random coin tosses of both adversary and challenger.

Definition 3. *The decentralized CP-ABE system is said to be IND-CPA (ciphertext indistinguishability under chosen plaintext attacks) secure against static corruption of authorities if all polynomial time adversaries have at most a negligible advantage in the above security game.*

4 dCP-ABE-MAS

In this section, we present a decentralized CP-ABE scheme for monotone access structures, dCP-ABE-MAS. Note that every monotone access structure \mathbb{A} is represented by its basis \mathbb{A}_0 which is the set of minimal authorized sets in \mathbb{A} .

System Initialization(κ). During system initialization phase, a six tuple $\text{GP} = (p, \mathbb{G}, g, \mathbb{G}_T, e, \mathcal{H})$ is chosen as global public parameters, where p is a prime number greater than 2^κ , \mathbb{G}, \mathbb{G}_T are two multiplicative cyclic groups of same prime order p , g is a generator of \mathbb{G} , $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is a bilinear map and $\mathcal{H} : \{0, 1\}^* \rightarrow \mathbb{G}$ is a collision resistant hash function which will be modeled as a random oracle in our security proof.

Authority Setup(GP, U^j). Each authority $A_j \in \mathcal{A}$ possesses a set of attributes U^j . For each attribute $a \in U^j$, A_j selects two random exponents $t_a, t'_a \in \mathbb{Z}_p$, and computes $P_a = g^{t_a}, P'_a = e(g, g)^{t'_a}$. The public key of A_j is published as $\text{Pub}A_j = \{(P_a, P'_a) | a \in U^j\}$. The master secret key of the authority A_j is $\text{Mk}A_j = \{(t_a, t'_a) | a \in U^j\}$.

Authority KeyGen($\text{GP}, \text{ID}, a, \text{Mk}A_j$). When a user with unique global identity $\text{ID} \in \{0, 1\}^*$ requests for a secret key associated with an attribute a which is held by A_j , the authority A_j returns $SK_{a, \text{ID}} = g^{t'_a} \mathcal{H}(\text{ID})^{t_a}$ to the user.

Encrypt(GP, $M, \mathbb{A}_0, \{\text{PubA}_j\}$). Here \mathbb{A}_0 is the basis for a monotone access structure \mathbb{A} . Let $\mathbb{A}_0 = \{B_1, B_2, \dots, B_k\}$, where each $B_i \subset U$ is a minimal authorized set in \mathbb{A} . The set $\{\text{PubA}_j\}$ is a set of public keys of all authorities which are managing the attributes in \mathbb{A}_0 . In order to encrypt a message $M \in \mathbb{G}_T$, the encryptor chooses a random exponent $s_i \in \mathbb{Z}_p$, for each $i, 1 \leq i \leq k$, and computes

$$C_{i,1} = M \cdot \left(\prod_{a \in B_i} P'_a \right)^{s_i}, C_{i,2} = g^{s_i} \quad \text{and} \quad C_{i,3} = \left(\prod_{a \in B_i} P_a \right)^{s_i}. \quad (2)$$

The encryptor outputs the ciphertext $\text{CT} = \langle \mathbb{A}_0, \{C_{i,1}, C_{i,2}, C_{i,3} | 1 \leq i \leq k\} \rangle$.

Decrypt(GP, CT, $\{\text{SK}_{a,\text{ID}} | a \in L_{\text{ID}}\}$). When a user with global identity $\text{ID} \in \{0,1\}^*$ receives a ciphertext CT, it first computes $\mathcal{H}(\text{ID})$. Suppose the attribute set L_{ID} of this user satisfies the monotone access structure \mathbb{A} generated by $\mathbb{A}_0 = \{B_1, B_2, \dots, B_k\}$. Then $L_{\text{ID}} \supseteq B_i$, for some $B_i \in \mathbb{A}_0$. The receiver now aggregates the secret attribute keys associated with the attributes appeared in the minimal authorized set B_i and computes $K_i = \prod_{a \in B_i} (\text{SK}_{a,\text{ID}})$. The message can then be obtained by computing

$$C_{i,1} \cdot \frac{e(\mathcal{H}(\text{ID}), C_{i,3})}{e(K_i, C_{i,2})} = M \cdot e(g, g)^{s_i b'_i} \cdot \frac{e(\mathcal{H}(\text{ID}), g^{s_i b_i})}{e(g^{b'_i} \mathcal{H}(\text{ID})^{b_i}, g^{s_i})} = M,$$

where $b'_i = \sum_{b \in B_i} t'_a$ and $b_i = \sum_{b \in B_i} t_a$. We will use the notations b'_i and b_i in our security proof.

Remark 1. An encryptor can pack different messages, say $M_1, M_2, \dots, M_{k'}$, where k' is equal or smaller than the size of a basis of a monotone access structure, in a single ciphertext by using the following encryption algorithm.

multi.Encrypt(GP, $\{M_1, M_2, \dots, M_{k'}\}, \mathbb{A}_0, \{\text{PubA}_j\}$). Let \mathbb{A} be a monotone access structure generated by its basis $\mathbb{A}_0 = \{B_1, B_2, \dots, B_k\}$. For each $i, 1 \leq i \leq k$, the encryptor chooses a random exponent $s_i \in \mathbb{Z}_p$ and computes the ciphertext $\text{CT} = \langle \mathbb{A}_0, \{C_{i,1}, C_{i,2}, C_{i,3} | 1 \leq i \leq k\} \rangle$, where $C_{i,1} = M_i \cdot (\prod_{a \in B_i} P'_a)^{s_i}$, $C_{i,2} = g^{s_i}$ and $C_{i,3} = (\prod_{a \in B_i} P_a)^{s_i}$.

On receiving the ciphertext $\text{CT} = \langle \mathbb{A}_0, \{C_{i,1}, C_{i,2}, C_{i,3} | 1 \leq i \leq k\} \rangle$, the recipient can be recovered respective message M_i by executing the decryption algorithm **Decrypt**(CT, $\{\text{SK}_{a,\text{ID}} | a \in L_{\text{ID}}\}, \text{GP}$) of dCP-ABE-MAS. The deployment of this mechanism will be discussed in Section 5.

4.1 Security Analysis

In this section, we first argue our dCP-ABE-MAS is secure against collusion attacks. We then prove dCP-ABE-MAS is IND-CPA secure in the generic bilinear group model (we refer the reader to [4] for definition).

Security against collusion attacks. A scheme is said to be collusion-resistant if no two or more recipients can combine their secret keys in order to decrypt a message that they are not entitled to decrypt alone. We will show that if two

users with identities ID, ID' try to collude and combine their secret keys, they will fail in decryption process even though their attributes associated with secret keys satisfy the monotone access structure \mathbb{A} . Note that $\mathbb{A}_0 = \{B_1, B_2, \dots, B_k\}$ is a basis for \mathbb{A} .

The encryption algorithm blinds the message M with $e(g, g)^{s_i b'_i}$. Consequently, the decryptor needs to recover the blinding term $e(g, g)^{s_i b'_i}$ by coupling their secret keys for attribute and identity pairs (a, ID) with the respective ciphertext components. If the decryptor has a satisfying set of keys with the same identity ID , i.e., $\{SK_{a, ID} | a \in B_i\}$, for some i , then the decryptor can recover the blinding term from the following computation.

$$\frac{e(K_i, C_{i,2})}{e(\mathcal{H}(ID), C_{i,3})} = \frac{e(g, g)^{s_i b'_i} \cdot \prod_{a \in B_i} e(\mathcal{H}(ID), g)^{s_i t_a}}{\prod_{a \in B_i} e(\mathcal{H}(ID), g)^{s_i t_a}} = e(g, g)^{s_i b'_i}.$$

Suppose two users with different identities ID and ID' try to collude and combine their secret attribute keys such that $L_{ID} \not\supseteq B_i$ and $L_{ID'} \not\supseteq B_i$, for any $1 \leq i \leq k$ but $L_{ID} \cup L_{ID'} \supseteq B_i$, for some B_i . Then $K_i = \prod_{a \in B_{i, ID}} SK_{a, ID} \cdot \prod_{a \in B_{i, ID'}} SK_{a, ID'}$, where $B_{i, ID} = L_{ID} \cap B_i$ and $B_{i, ID'} = L_{ID'} \cap B_i$. Consequently, there will be some terms of the form $e(\mathcal{H}(ID), g)^{s_i t_a}$ in denominator and some terms of the form $e(\mathcal{H}(ID'), g)^{s_i t_a}$ in numerator which will not cancel with each other as \mathcal{H} is collision resistant, i.e., $\mathcal{H}(ID) \neq \mathcal{H}(ID')$, thereby preventing the recovery of the blinding term $e(g, g)^{s_i b'_i}$, so is the message M . This demonstrates that dCP-ABE-MAS scheme is collusion-resistant.

Theorem 1. *The dCP-ABE-MAS scheme is IND-CPA secure against static corruption of authorities in the generic group model.*

Proof. Let ADV_1 be an adversary who plays the original security game, say $GAME_1$, described in Section 3.1. According to $GAME_1$, the challenge ciphertext has a component $C_{i,1}$ which is either $M_0 \cdot e(g, g)^{s_i b'_i}$ or $M_1 \cdot e(g, g)^{s_i b'_i}$, and the adversary ADV_1 has to distinguish them. Consequently, we define a modified game, say $GAME_2$, as follows. Setup, Key Query Phase 1 and Key Query Phase 2 are similar to $GAME_1$, but the challenge ciphertext component $C_{i,1}$ in Challenge phase is computed as $C_{i,1} = e(g, g)^{s_i b'_i}$ if $\mu = 1$ and $C_{i,1} = e(g, g)^{\delta_i}$ if $\mu = 0$, where δ_i is selected uniformly at random from \mathbb{Z}_p , and other ciphertext components are computed in the same way analogous to Encrypt algorithm. Then we have the following claim. *Claim 1: If ADV_1 has advantage ϵ to win $GAME_1$, then there is an adversary who wins $GAME_2$ with advantage at least $\epsilon/2$.*

Proof of Claim 1: According to ADV_1 , we can construct an adversary ADV_2 as follows. In Setup, Key Query Phase 1 and Key Query Phase 2, ADV_2 forwards all messages it receives from ADV_1 to the challenger and all messages from the challenger to ADV_1 . In the Challenge phase, ADV_2 receives two messages M_0 and M_1 from ADV_1 and the challenge ciphertext CT^* from the challenger. Note that CT^* contains $C_{i,1}$ that is either $e(g, g)^{s_i b'_i}$ or $e(g, g)^{\delta_i}$. Now, ADV_2 flips a random coin $\nu \in \{0, 1\}$ and replaces $C_{i,1}$ by $C'_{i,1} = M_\nu \cdot C_{i,1}$ in CT^* to compute a modified ciphertext CT' and finally sends the resulting CT' to the adversary ADV_1 .

Guess: ADV_1 outputs his guess $\nu' \in \{0, 1\}$ on ν . If $\nu' = \nu$, ADV_2 outputs as its guess $\mu' = 1$; otherwise he outputs $\mu' = 0$.

- In the case where $\mu = 1$, CT' is a correct ciphertext of M_ν . Consequently, ADV_1 can output $\nu' = \nu$ with the advantage ϵ , i.e., $\Pr[\nu' = \nu | \mu = 1] = \frac{1}{2} + \epsilon$. Since ADV_2 guesses $\mu' = 1$ when $\nu' = \nu$, we get $\Pr[\mu' = \mu | \mu = 1] = \frac{1}{2} + \epsilon$.
- In the next case where $\mu = 0$, the challenge ciphertext CT^* is independent of the messages M_0 and M_1 , so ADV_1 cannot obtain any information about ν . Therefore, ADV_1 can output $\nu' \neq \nu$ with no advantage, i.e., $\Pr[\nu' \neq \nu | \mu = 0] = \frac{1}{2}$. Since ADV_2 guesses $\mu' = 0$ when $\nu' \neq \nu$, we get $\Pr[\mu' = \mu | \mu = 0] = \frac{1}{2}$.

Thus, advantage of $\text{ADV}_2 = |\Pr[\mu' = \mu] - \frac{1}{2}| \geq \frac{1}{2} \cdot (\frac{1}{2} + \epsilon) + \frac{1}{2} \cdot \frac{1}{2} - \frac{1}{2} = \frac{\epsilon}{2}$. This proves the claim 1.

This claim demonstrates that any adversary that has a non-negligible advantage in GAME_1 can have a non-negligible advantage in GAME_2 . We shall prove that no adversary can have non-negligible advantage in GAME_2 . From now on, we will discuss the advantage of the adversary in GAME_2 , wherein the adversary must distinguish between $e(g, g)^{s_i b'_i}$ and $e(g, g)^{\delta_i}$.

Simulation in GAME_2 : To simulate the modified security game GAME_2 , we use the generic bilinear group model given in [4]. Consider two injective random maps $\psi, \psi_T : \mathbb{Z}_p \rightarrow \{0, 1\}^{\lceil 3 \log(p) \rceil}$. In this model every element of \mathbb{G} and \mathbb{G}_T is encoded as an arbitrary random string from the adversary's point of view, i.e., $\mathbb{G} = \{\psi(x) | x \in \mathbb{Z}_p\}$ and $\mathbb{G}_T = \{\psi_T(x) | x \in \mathbb{Z}_p\}$. The adversary is given three oracles to compute group operations of \mathbb{G}, \mathbb{G}_T and to compute the bilinear pairing e . The input of all oracles are string representations of group elements. The adversary is allowed to perform group operations and pairing computations by interacting with the corresponding oracles only. It is assumed that the adversary can make queries to the group oracles on input strings that were previously been obtained from the simulator or were given from the oracles in response to the previous queries. This event occurs with high probability. Since $|\psi(\mathbb{Z}_p)| > p^3$ and $|\psi_T(\mathbb{Z}_p)| > p^3$, the probability of the adversary being able to guess an element (which it has not previously obtained) in the ranges of ψ, ψ_T is negligible.

The notations $g^x := \psi(x)$ and $e(g, g)^x := \psi_T(x)$ are used in the rest of the proof. With this notation, g and $e(g, g)$ can be represented as $\psi(1)$ and $\psi_T(1)$, respectively.

Setup: Note that \mathcal{A} is the set of all authorities in the system and U is the attribute universe. The simulator obtains the global public parameters GP from the trusted system initializer and gives $\psi(1)$ to the adversary. The adversary sends a corrupted authority list $\mathcal{A}' \subset \mathcal{A}$ to the simulator. For each attribute $a \in U$ controlled by honest authorities, the simulator chooses two new random values $t_a, t'_a \in \mathbb{Z}_p$, computes $g^{t_a}, e(g, g)^{t'_a}$ using respective group oracles and gives $P_a = \psi(t_a), P'_a = \psi_T(t'_a)$ to the adversary.

Query Phase 1: The adversary issues hash and secret key queries, and consequently the simulator responds as follows.

Hash queries: When the adversary requests $\mathcal{H}(\text{ID})$ for some user identity ID for the first time, the simulator chooses a new, unique random value $u_{\text{ID}} \in \mathbb{Z}_p$,

computes $g^{u_{\text{ID}}} = \psi(u_{\text{ID}})$ using group oracle and gives $\psi(u_{\text{ID}})$ to the adversary as $\mathcal{H}(\text{ID})$. The association between values u_{ID} and the user identities ID is stored in Hlist so that it can reply consistently for subsequent queries in the future.

Secret key queries: If the adversary requests for a secret key of an attribute a with identity ID , the simulator computes $g^{t_a} \mathcal{H}(\text{ID})^{t_a}$ using the group oracle and returns $\text{SK}_{a,\text{ID}} = \psi(t'_a + u_{\text{ID}} t_a)$ to the adversary. If $\mathcal{H}(\text{ID})$ has not been stored in Hlist , it is determined as above.

Challenge: In order to obtain a challenge ciphertext CT^* , the adversary specifies the basis $\mathbb{A}_0 = \{B_1, B_2, \dots, B_k\}$ of a monotone access structure \mathbb{A} along with the public keys $g^{t_a}, e(g, g)^{t_a}$ of attributes $a \in U$ which are controlled by corrupted authorities and appeared in \mathbb{A}_0 as members in several B_i . The simulator then checks the validity of these public keys by querying the group oracles. Now, the simulator chooses a random s_i for the i -th minimal set of \mathbb{A}_0 , for each i , $1 \leq i \leq k$ and computes $b_i = \sum_{a \in B_i} t_a$. The simulator then flips a random coin $\mu \in \{0, 1\}$ and if $\mu = 1$, he sets $\delta_i = s_i b'_i$, where $b'_i = \sum_{a \in B_i} t'_a$, otherwise δ_i is set to be a random value from \mathbb{Z}_p . The simulator finally computes the components of challenge ciphertext CT^* by using group oracles as follows.

$$C_{i,1} = \psi_T(\delta_i), C_{i,2} = \psi(s_i), C_{i,3} = \psi(s_i b_i) \text{ for all } i, 1 \leq i \leq k.$$

The ciphertext $\text{CT}^* = \langle \mathbb{A}_0, \{C_{i,1}, C_{i,2}, C_{i,3} | 1 \leq i \leq k\} \rangle$ is sent to the adversary.

Query Phase 2: The adversary issues more hash and secret key queries. The simulator responds as in Query Phase 1. We note that if the adversary requests for secret keys of a set of attributes that allow decryption in combination with secret keys obtained from corrupted authorities, then the simulator is aborted.

The adversary now can have in his hand, all values that consists of encodings of random values $\delta_i, 1, u_{\text{ID}}, t_a, t'_a, s_i$ and combination of these values given by the simulator (e.g., $\psi(t'_a + u_{\text{ID}} t_a)$) or results of queries on combination of these values to the oracles. In turn, we can think of each query of the adversary is a multivariate polynomial in the variables $\delta_i, 1, u_{\text{ID}}, t_a, t'_a, s_i$, where a ranges over the attributes controlled by honest authorities, i ranges over the minimal sets in the basis of monotonic access structure and ID ranges over the allowed user identities. We assume that any pair of the adversary's queries on two different polynomials result in two different answers. This assumption is false only when our choice of the random encodings of the variables ensures that the difference of two polynomial queries evaluates to zero. Following the security proof in [4], it can be claimed that the probability of any such collision is at most $\mathcal{O}(q^2/p)$, q being an upper bound on the number of oracle queries made by the adversary during the entire simulation. Therefore, the advantage of the adversary is at most $\mathcal{O}(q^2/p)$. We assume that no such random collisions occur while retain $1 - \mathcal{O}(q^2/p)$ probability mass.

Under this condition, we show that the view of the adversary in GAME_2 is identically distributed when $\delta_i = s_i b'_i$ if $\mu = 1$ and δ_i is random if $\mu = 0$, and hence the adversary cannot distinguish them in the generic bilinear group model. To prove this by contradiction, let us assume that the views are not identically distributed. The adversary's views can only differ when there exists two queries

Table 2. Possible adversary's query terms in \mathbb{G}_T (here, the variables a, a' are possible attributes, ID, ID' are authorized user identities and i, i' are indices of the minimal sets in the monotone access structure).

t_a	$t_a t_{a'}$	$u_{ID} u_{ID'}$	$b_i(t'_a + u_{ID} t_a)$	$s_i s_{i'}$
u_{ID}	$t_a u_{ID}$	$u_{ID'}(t'_a + u_{ID} t_a)$	$s_i(t'_a + u_{ID} t_a)$	$s_i s_{i'} b_{i'}$
$t'_a + u_{ID} t_a$	$t_{a'}(t'_a + u_{ID} t_a)$	$u_{ID} b_i$	$s_i b_i(t'_a + u_{ID} t_a)$	$s_i s_{i'} b_i b_{i'}$
b_i	$t_a b_i$	$u_{ID} s_i$	$b_i b_{i'}$	t'_a
s_i	$t_a s_i$	$u_{ID} s_i b_i$	$s_i b_{i'}$	b'_i
$s_i b_i$	$t_a s_i b_i$	$(t'_a + u_{ID} t_a)(t'_{a'} + u_{ID'} t_{a'})$	$s_i b_i b_{i'}$	

q_1 and q_2 in \mathbb{G}_T such that $q_1 \neq q_2$ with $q_1|_{(\delta_i=s_i b'_i)} = q_2|_{(\delta_i=s_i b'_i)}$, for at least one i . Fix one such i . Since δ_i only appears as $\psi_T(\delta_i)$ and elements of ψ_T cannot be used as input of this oracle takes elements of ψ as input, the adversary can only make queries of the following form involving δ_i : $q_1 = c_1 \delta_i + q'_1$ and $q_2 = c_2 \delta_i + q'_2$, for some q'_1 and q'_2 that do not contain δ_i , and for some constants c_1 and c_2 . Since $q_1|_{(\delta_i=s_i b'_i)} = q_2|_{(\delta_i=s_i b'_i)}$, we have $c_1 s_i b'_i + q'_1 = c_2 s_i b'_i + q'_2$ and it gives $q'_2 - q'_1 = (c_1 - c_2) s_i b'_i = c s_i b'_i$, for some constant $c \neq 0$. Therefore, the adversary can construct the query $\psi_T(c s_i b'_i)$, for some constant $c \neq 0$, yielding a contradiction to our claim 2 proved below. Hence the adversary's views in GAME_2 are identically distributed, i.e., the adversary has no non-negligible advantage in GAME_2 , so in the original game GAME_1 by claim 1.

Claim 2 : The adversary cannot make a query of the form $\psi_T(c s_i b'_i)$ for any non-zero constant c and any i .

Proof of Claim 2: To establish this claim, we examine the information given to the adversary during the entire simulation and perform case analysis based on that information.

In Table 2, we list all the possible adversary's query terms in \mathbb{G}_T by means of the bilinear map and group elements given to the adversary during the simulation. It can be seen that the adversary can query for an arbitrary linear combination of 1 (which is $\psi_T(1)$), δ_i and the terms given in Table 2. We will now show that no such linear combination can produce a term of the form $c s_i b'_i$ for any non-zero constant c and any i . Note that the adversary knows the values of t_a, t'_a for attributes a that are controlled by the corrupted authorities, so these can appear in a foregoing linear combinations as the coefficients of the terms given in Table 2.

We note that $s_i b'_i = \sum_{a \in B_i} s_i t'_a$. From Table 2 we see that the only way for an adversary to create a term containing $s_i t'_a$ is by pairing s_i with $t'_a + u_{ID} t_a$. Consequently, the adversary can create a query polynomial of the form

$$\sum_{a \in B} (c_{(i,a)} s_i t'_a + c_{(i,a,ID)} u_{ID} s_i t_a), \quad (3)$$

for some set of attributes B and non-zero constants $c_{(i,a)}, c_{(i,a,ID)}$. In order to get a query polynomial of the form $c s_i b'_i$ the adversary must add other terms to

cancel the extra terms $\sum_{a \in B} c_{(i,a, \text{ID})} u_{\text{ID}} s_i t_a$. For any terms $c_{(i,a, \text{ID})} u_{\text{ID}} s_i t_a$ where a is an attribute held by a corrupted authority, the value of t_a is revealed to the adversary, thereby the adversary can form the term $-c_{(i,a, \text{ID})} u_{\text{ID}} s_i t_a$ in order to cancel this from the polynomial given in Eq. (3). For terms $c_{(i,a, \text{ID})} u_{\text{ID}} s_i t_a$ where a is an attribute controlled by an uncorrupted authority, the adversary cannot construct terms to cancel these from the polynomial given in Eq. (3) since there is no term in Table 2 that enables the adversary to construct a term of the form $-c_{(i,a, \text{ID})} u_{\text{ID}} s_i t_a$. Consequently, the adversary's query polynomial cannot be of the form $c s_i b'_i$.

Suppose for some identity ID, a set B' of attributes in B belong to the corrupted authorities or the adversary has obtained secret keys $\{\text{SK}_{a, \text{ID}} | a \in B'\}$ such that $B' \supseteq B_i$, for some $i, 1 \leq i \leq k$. Then the adversary can construct a query polynomial of the form

$$\sum_{a \in B_i} (c s_i t'_a + c_{\text{ID}} u_{\text{ID}} s_i t_a), \quad (4)$$

for some non-zero constant c and c_{ID} . The query polynomial given in Eq. (4) is same as $c s_i \sum_{a \in B_i} t'_a + c_{\text{ID}} u_{\text{ID}} s_i \sum_{a \in B_i} t_a = c s_i b'_i + c_{\text{ID}} u_{\text{ID}} s_i b_i$. The extra term $c_{\text{ID}} u_{\text{ID}} s_i b_i$ here will be canceled by using the term $u_{\text{ID}} s_i b_i$ appeared in Table 2. In this case, even though the adversary becomes successful, the constraint mentioned in the Challenge phase of the security game is violated and simulator is aborted.

We have shown that the adversary cannot make a query polynomial of the form $c s_i b'_i$, for any constant $c \neq 0$ and any i , without violating the assumptions stated in the security game. This proves the claim 2 and hence the theorem. \square

5 Applications

In this section, we propose an access control scheme in various network scenarios that make use of our dCP-ABE-MAS and then compare our scheme with the existing schemes in the respective areas.

Vehicular Ad Hoc Network: Typically, a vehicular ad hoc network (VANET) mainly consists of three kinds of entities—trusted initializer (TI), road side units (RSUs) and vehicles which are equipped with wireless communication devices, called on-board units (OBUs). During registration phase, each vehicle is assigned by the TI a set of *persistent attributes* (e.g., year, model), which remains constant throughout the lifetime of a vehicle, and a set of different pseudonyms, which preserves location privacy of the vehicle. We assume that each vehicle is capable of changing pseudonyms from time to time. In addition, TI gives each vehicle a set of secret keys associated with the persistent attributes for each pseudonym of that vehicle. These attributes and keys are preloaded into vehicle's OBU.

There are several RSUs which are distributed across the network in a uniform fashion and each RSU provides infrastructure support for a specified region which we call communication range of that RSU. Each RSU controls a set of *dynamic attributes* (e.g., road name, vehicle speed). When a vehicle enters within

communication range of an RSU, the RSU gives it certain dynamic attributes along with corresponding secret attribute keys after receiving a certificate relating the current pseudonym of the vehicle. We assume that there are secure communication channels between vehicles and TI as well as vehicles and RSUs.

Note that the authorities in our dCP-ABE-MAS play the role of RSUs and the attribute universe is combination of all persistent and dynamic attributes involved in the network. Every persistent attribute is different from every dynamic attribute and the attributes controlled by two different RSUs are all different from each other. The pseudonym can be treated as vehicle's identity. The setup and key generation algorithms of TI are same as authorities' setup and key generation algorithms, respectively.

Vehicles can encrypt and decrypt messages. RSUs can also encrypt messages for a set of selected vehicles. When a vehicle wants to send a message M to other vehicles in the network regarding the road situation (e.g., a car accident is ahead), it decides firstly the intended vehicles (e.g., ambulance, police car, breakdown truck) and then formulates an associated MAS in terms of minimal authorized sets over some attributes (both persistent and dynamic), for example, $\mathbb{A}_0 = \{B_1, B_2, B_3\}$, where $B_1 = \{\text{ambulance, road1}\}$, $B_2 = \{\text{policecar, lane2}\}$ and $B_3 = \{\text{breakdowntruck, road2}\}$. The encryptor vehicle then uses the public keys of the attributes occurring in the access structure to encrypt a message and transmits the ciphertext. A recipient vehicle whose attribute set satisfies the access structure will only be able to decrypt the message.

Refer to the above example, consider a scenario where the encryptor vehicle needs to send a different message to each category of vehicles—ambulance, police car, breakdown truck. Consequently, it has to encrypt each message separately under respective access structure for each category. In turn, the number of encryptions will grow linearly with the number of categories. In such cases, the proposed multi.Encrypt algorithm (described in *Remark 1*) can pack multiple messages in a single ciphertext, thereby reduces network traffic significantly, in such a way that the respective message will only be decrypted by the intended category of vehicles. This helps in the widespread dissemination of messages and early decision making in such a highly dynamic network environments.

The comparison of proposed scheme, say Scheme 1 in the VANET scenario, with the existing scheme [12] is presented in Table 3, 4.

Distributed Cloud Network: The cloud storage system is composed of five entities: trusted initializer (TI), key generation authorities (KGAs), cloud, data owner (data provider) and users (data consumers). The only responsibility of TI is generation of global public parameters GP of the system and assignment of a unique global identity ID to each user in the system. Each key generation authority controls a different set of attributes and generates public and secret keys for all attributes that it holds. The KGAs are also responsible to distribute secret keys for users' attribute sets on request according to their role or identity. The KGAs could be scattered geographically far apart and execute assigned tasks independently. The authorities in our dCP-ABE-MAS act as KGAs.

Table 3. Comparison of Computation Costs

Scheme	Key Generation	Encryption			Decryption		
	E_G	E_G	E_{G_T}	P_e	E_G	E_{G_T}	P_e
[14]	$2\gamma + 2$	$4\alpha + 1$	1	-	-	$\mathcal{O}(\beta)$	$\mathcal{O}(\beta)$
[12, 13, 16]	2γ	3α	$2\alpha + 1$	1	-	$\mathcal{O}(\beta)$	$\mathcal{O}(\beta)$
Scheme 1,2	2γ	$2k$	k	-	-	-	2

Table 4. Comparison of Communication Overheads

Scheme	User Secret Key Size	Ciphertext Size	Access Policy	Requirement of CA
[14]	$(\gamma + 2)B_G$	$(3\alpha + 1)B_G + B_{G_T} + \tau$	LSSS	Yes
[12, 13, 16]	γB_G	$(2\alpha)B_G + (\alpha + 1)B_{G_T} + \tau$	LSSS	No
Scheme 1,2	γB_G	$2kB_G + kB_{G_T} + \tau$	any MAS	No

The cloud is an external storage server that allows the data owners to store their data in the cloud in order to share their data securely to intended users. The data owners enforce an access control policy in the form of a MAS into ciphertext in such a way that only intended users can recover the data and sign the message by employing an efficient attribute-based signature scheme. Finally, the ciphertext along with signature is sent to the cloud. The cloud first verifies the signature and stores the ciphertext if the signature is valid. Each user can obtain ciphertexts from the cloud on demand. However, the users can decrypt the ciphertext only if the set of attributes associated with their secret keys satisfy the access control policy embedded in the ciphertext.

Consider a health-care scenario where the patients can be data providers, and doctors, medical researchers and health insurance companies can be data consumers. For example, a patient wishes to store his medical history in the cloud for specific users as follows: brain scan records, M_1 , for any neurologist from hospital X, ECG (Electrocardiography) reports, M_2 , for any cardiologist and Ultrasound reports, M_3 , for any radiology researcher from any medical research center. In such setting, the multi.Encrypt algorithm (described in *Remark 1*) is well suited to pack all the three messages in a single ciphertext. To this end, the patient first formulates a MAS whose basis is $\mathbb{A}_0 = \{B_1, B_2, B_3\}$, where $B_1 = \{\text{neurologist, hospitalX}\}$, $B_2 = \{\text{cardiologist}\}$ and $B_3 = \{\text{radiologist, researcher}\}$. Once the policy is specified, multi.Encrypt algorithm is executed with the input the set of messages $\{M_1, M_2, M_3\}$, \mathbb{A}_0 and the respective public keys. Finally, the resulting ciphertext will be stored in the cloud. Refer to the decryption algorithm of dCP-ABE-MAS, only the intended users can decrypt the respective messages.

We compare our proposed construction, say Scheme 2 in the context of cloud storage, with the existing schemes [13, 14, 16] in Table 3, 4, where the ciphertext size is considered without signature to make consistent with other schemes.

Acknowledgement. The authors would like to thank the anonymous reviewers of this paper for their valuable comments and suggestions.

References

1. Shamir, A.: Identity-Based Cryptosystems and Signature Schemes. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
2. Sahai, A., Waters, B.: Fuzzy Identity-Based Encryption. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 457–473. Springer, Heidelberg (2005)
3. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute Based Encryption for Fine-Grained Access Control of Encrypted Data. In: ACM Conference on Computer and Communications Security, pp. 89–98 (2006)
4. Bethencourt, J., Sahai, A., Waters, B.: Ciphertext-Policy Attribute-Based Encryption. In: IEEE Symposium on Security and Privacy, pp. 321–334 (2007)
5. Waters, B.: Ciphertext-Policy Attribute-Based Encryption: An Expressive, Efficient, and Provably Secure Realization. In: Catalano, D., Fazio, N., Gennaro, R., Nicolosi, A. (eds.) PKC 2011. LNCS, vol. 6571, pp. 53–70. Springer, Heidelberg (2011)
6. Ibraimi, L., Tang, Q., Hartel, P., Jonker, W.: Efficient and Provable Secure Ciphertext-Policy Attribute-Based Encryption Schemes. In: Bao, F., Li, H., Wang, G. (eds.) ISPEC 2009. LNCS, vol. 5451, pp. 1–12. Springer, Heidelberg (2009)
7. Stinson, D.R.: Cryptography: Theory and Practice, 3rd edn. CRC Press (2006)
8. Lewko, A., Waters, B.: Decentralizing Attribute-Based Encryption. Cryptology ePrint Archive, Report 2010/351 (2010)
9. Lin, H., Cao, Z.-F., Liang, X., Shao, J.: Secure Threshold Multi Authority Attribute Based Encryption without a Central Authority. In: Chowdhury, D.R., Rijmen, V., Das, A. (eds.) INDOCRYPT 2008. LNCS, vol. 5365, pp. 426–436. Springer, Heidelberg (2008)
10. Chase, M.: Multi-authority Attribute Based Encryption. In: Vadhan, S.P. (ed.) TCC 2007. LNCS, vol. 4392, pp. 515–534. Springer, Heidelberg (2007)
11. Chase, M., Chow, S.S.M.: Improving Privacy and Security in Multi-Authority Attribute-Based Encryption. In: ACM Conference on Computer and Communications Security, pp. 121–130. ACM, New York (2009)
12. Ruj, S., Nayak, A., Stojmenovic, I.: Improved Access Control Mechanism in Vehicular Ad Hoc Networks. In: Frey, H., Li, X., Ruehrup, S. (eds.) ADHOC-NOW 2011. LNCS, vol. 6811, pp. 191–205. Springer, Heidelberg (2011)
13. Ruj, S., Stojmenovic, M., Nayak, A.: Privacy Preserving Access Control with Authentication for Securing Data in Clouds. In: 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 556–563 (2012)
14. Yang, K., Jia, X., Ren, K.: DAC-MACS: Effective Data Access Control for Multi-Authority Cloud Storage Systems. Cryptology ePrint Archive, Report 2012/419
15. Müller, S., Katzenbeisser, S., Eckert, C.: On Multi-Authority Ciphertext-Policy Attribute-Based Encryption. Bulletin of the Korean Mathematical Society 46(4), 803–817 (2009)
16. Ruj, S., Nayak, A., Stojmenovic, I.: DACC: Distributed access control in clouds. In: IEEE TrustCom 2011, pp. 91–98. IEEE (2011)
17. Liu, Z., Cao, Z., Huang, Q., Wong, D.S., Yuen, T.H.: Fully Secure Multi-authority Ciphertext-Policy Attribute-Based Encryption without Random Oracles. In: Atluri, V., Diaz, C. (eds.) ESORICS 2011. LNCS, vol. 6879, pp. 278–297. Springer, Heidelberg (2011)
18. Pandit, T., Barua, R.: Efficient Fully Secure Attribute-Based Encryption Schemes for General Access Structures. In: Takagi, T., Wang, G., Qin, Z., Jiang, S., Yu, Y. (eds.) ProvSec 2012. LNCS, vol. 7496, pp. 193–214. Springer, Heidelberg (2012)

Part II

Work in Progress

Security of Features Describing the Visual Appearance of Handwriting Samples Using the Bio-hash Algorithm of Vielhauer against an Evolutionary Algorithm Attack

Andreas Hasselberg¹, Rene Zimmermann¹, Christian Kraetzer¹,
Tobias Scheidat², Claus Vielhauer², and Karl Kümmel²

¹ Otto-von-Guericke-University Magdeburg, Germany

² Brandenburg University of Applied Sciences

Abstract. To improve the security and stability of biometric handwriting samples a Bio-Hash algorithm for handwriting was introduced in [1]. It utilizes features to describe how the sample was written, but the current set of features does not characterize the visual appearance of the sample itself. In this paper we present a set of new features derived from handwriting forensics and OCR algorithms to address this issue. Furthermore, here the security of the old and new sets of features is evaluated for their resilience against a new, fully automated attack trying to compute raw data matching a given hash vector.

The main contributions of this paper are: The introduction of new features with a potential to increase the attack resilience of the Bio-Hash algorithm, and, an improvement of the attack approach from [6] to produce more realistic looking synthetic handwriting signals.

Keywords: bio-hash, evolutionary algorithm, handwriting, biometrics.

1 Introduction

The most common approach to efficiently protect the biometric templates in a system (i.e. the reference data) is to transform the data using a one-way-function. An example for this kind of function is the biometric hash algorithm for dynamic handwriting (hereafter called Bio-Hash algorithm) from Vielhauer [1].

In [6] Kümmel et al. introduce an evolutionary algorithm based attack to the template space of the Bio-Hash algorithm. The main drawbacks to this rather successful attack are the fact that it relies on user interaction and the strongly artificial look of its output.

In this paper we make two contributions to the research in this field:

- A) We extend the feature space used by the Bio-Hash algorithm for dynamic handwriting based user authentication. Our new features aim at the description of the visual appearance of the sample itself, something that is amiss amongst the features used until now.

- B) We extend the attack from [6] to a completely automated (i.e. without the user interaction required in [6]) means to create synthetic handwritings which, on one hand, produces a Bio-Hash similar to the reference Bio-Hash and, on the other hand, look natural enough to be able to fool visual inspection of the input in an authentication scenario.

Our results imply for the first contribution that we introduce two (of the 13 new features) which show a strong potential to increase the attack resilience of the Bio-Hash algorithm, because they have been exceptionally hard to successfully attack in our evaluations. For the second main contribution, our improved attack generates much more realistic looking synthetic handwritings.

The paper is structured as follows: Section 2 summarizes some required basics. In section 3 the prerequisites, ideas and the algorithms to compute the 13 new features describing the visual appearance of online captured handwriting samples are presented. In section 4 the attack approach used for evaluation is described with its prerequisites and its implementation. Section 5 contains the evaluation results, while the final section 6 draws a conclusion and presents possible direction for future work.

2 Related Work

The template used in biometric authentication systems has to be protected for various security, legal and ethical reasons. For handwriting data, one very strong motivation for such protection is the fact that the template might contain enough information for re-engineering parts of the original biometric data, e.g. a handwritten signature. Amongst the various protection approaches one of the most prominent is the Bio-Hash algorithm of Vielhauer [1], which has been developed for handwriting, but could also be used for different biometrics (e.g. voice).

The Biometric Hash (short: Bio-Hash) algorithm was introduced by Vielhauer in [1] to generate individual, stable hash values from dynamic handwriting data as well as to perform biometric verification based on these hashes. Generally, the raw data of each dynamic handwriting sample derived from a handwriting digitizer device (e.g. Tablet PC) consists of a time-dependent sequence of physical values. Each sample point of such a sequence contains the five values horizontal pen tip position, vertical pen tip position, pen tip pressure and pen orientation angles altitude and azimuth.

During the enrollment process, n handwriting raw data samples are acquired per person. From each of $n-1$ raw data samples, one k -dimensional statistical feature vector containing static and dynamic features is calculated. The individual Interval Matrix (IM) consists of a vector containing a mapping interval length for each feature and an offset value vector. Both are calculated based on the analysis of intra-class variability of the person using the $n-1$ statistical feature vectors acquired during enrollment session. There are two parameters to influence the Bio-Hash generation by scaling the mapping intervals stored in the IM: the tolerance factor TF and tolerance

vector TV . TF is a global hash generation parameter, which is a scalar value. Using the TF , the mapping intervals of all features are scaled by the same global factor. In contrast, TV provides an individual scaling of the mapping interval of every single statistical feature. Based on the statistical feature vector derived from the remaining enrollment sample of the person and its individual IM , the interval mapping function determines the reference Bio-Hash vector b_{ref} of this user. Therefore, the feature dependent interval lengths and offsets provided by IM are used to map each of the corresponding statistical features to a hash value. Each further biometric hash is calculated in the same manner, whether it is used for biometric verification or hash generation. In case of verification, a Bio-Hash vector b , which is derived from the currently presented handwriting sample, is compared against the reference hash vector b_{ref} by using the Hamming distance and a predefined threshold. For more details of the particular calculation steps, the interested reader is referred to [1].

In order to build the feature vector, 131 features are extracted from each raw data sample. As described in [7] by Makrushin et al. the features can be classified by the signals required for calculation: 3 time based statistics features, 88 static spatial statistics features (time and order independent), 8 dynamic spatial statistics features, 22 pressure-based statistics features and 10 angles-based statistics features.

In [6] Kümmel et al. introduce an attack that uses a genetic algorithm for raw handwriting data reconstruction. In a Kerckhoffs' compliant setup it assumes that the attacker gains access to the Bio-Hash b_{ref} of a user and the corresponding IM and can therefore evaluate the fitness of artificially created handwritings by computing the Hamming distance to b_{ref} . Despite the fact that this attack works quite well, as shown in [6], it suffers from two rather severe drawbacks that are addressed in this paper: first, the attack in [6] is requiring user interaction for the implementation of selected features, and second, its output looks artificial to a human observer (see an comparison between a real handwriting and the outputs of the attack tool from [6] and our attack in figure 2 at the end of the document). Both drawbacks are addressed here, by introducing a new, fully automated, genetic algorithm based attack that aims at the generation of natural looking artificial handwritings.

3 Description of the New features

The feature extraction follows a pipeline which is reflected in the following four points summarised below: The necessary preprocessing, the determination of the baseline as well as the slope correction, the estimation of the reference lines and the approximation of the dominant slant of the written sample.

For the preprocessing, the majority of the following algorithms has its origin in OCR and is rather based on continuous handwritten samples than the discrete point sets returned by online systems for capturing of handwriting samples. Therefore, to adapt these algorithms the output generated by such systems is preprocessed in our approach by connecting the points using the Bresenham-algorithm [3]. The algorithm

capitalizes on the available online information from the sensor device, specifically the fact that the order of points and their coordinates are known. After this preprocessing, the sample from an online system looks much more like a real continuous writing and is usable for OCR approaches.

The line on which most of the lower ends of written letters align is called the baseline. It allows for a determination of the orientation of the writing direction. Our idea for the computation of the baseline is derived from the approach of Senior and Robinson [2]. In contrast to [2], our algorithm abstains from cutting descenders in the first step, because for our test samples no horizontal slope for the sample can be guaranteed. Instead our first step is to identify the point with the smallest y -coordinate in each column (of the x - y matrix representing the handwritten sample), i.e. for each distinct x -coordinate with points. Through the connection of points performed in the preprocessing, the need for an identification of unrequired points is kept to a minimum. In the worst case, every point of the sample might have been returned, as long as they were in different columns of the x - y matrix. The next step is to compute the local minima. All points except for the local minima will be discarded for the further baseline computation steps. This reduces the set of points to only the ones relevant for the baseline computation.

As starting point for further optimization a first baseline approximation is computed via linear regression on the remaining points. For these, their distance to the regression line is calculated using the Hesse normal form. Depending on whether the points lay above or below the regression line, their distances have either a positive or a negative sign. Given that only local minima were used in the computation of the first approximation of the baseline, it is safe to assume that most of these points lay close to the actual baseline and only a minority possesses a larger deviation. Therefore, we assume that with the first linear regression we already found a suitable first baseline approximation, i.e. we assume that the calculated distances have a zero mean Gaussian distribution. For this sample and the expected value it is possible to calculate the standard deviation so that all necessary parameters for the description of a Gaussian distribution are in place. Usually, the processing steps described above leave only a small number of points for further processing. Thus we approximate the Gaussian distribution with a Student's t -distribution. To compensate for outliers in distances before performing another linear regression, we cut off the outer 20 % of the distribution. For the remaining points we compute again a linear regression line, which is the final baseline returned by our approach. Here, it has to be mentioned that the baseline is already conceptually considered in [1] and used there for a preprocessing of the evaluation samples.

Finding the baseline allows for the extraction of the first seven features describing the appearance of the sample: (1) slope of the baseline, (2) angle between baseline and x -axis, (3) y -coordinate of interception of baseline and y -axis, (4) x - and (5) y -coordinate of the left interception point of baseline and bounding box, (6) x - and (7) y -coordinate of the right interception point of baseline and bounding box. After the extraction of these features, all points of the sample are rotated using the angle between baseline and x -axis to create a sample aligned parallel with the x -axis.

When learning to write in a language using the Latin alphabet in school, often special paper with four helper lines is used to help the pupils to keep the letter sizes relatively constant during the writing process. These four lines are called reference lines and even without them preprinted on a sheet of paper handwriting in languages using the Latin alphabet usually aligns itself on these (trained) lines. Our approach to compute the reference lines follows in parts the histogram based approach of Yanikoglu and Sandon [4]. The first step is to determine the histogram h_0 of the horizontal pixel density; one distinct x-coordinate is one distinct class in the histogram. Step two is to smooth the resulting histogram. This is done by accumulating the sum from the two previous classes, the class itself and the two subsequent classes.

This smoothing results in a far more stable computation, since it becomes more infrequent to observe single empty classes, i.e. with the value zeros in the histogram. In contrast to Yanikoglu and Sandon [4] we abstain from doing another smoothing step because the determination of the two inner reference lines is implemented here in a different way and the results using a second smoothing were too unstable in our experiments.

In cursive handwriting it is possible for the dot over an “i” to be at a higher position than the upper end of the word boundary or to even protrude in upper text lines. Therefore, starting from the center of the smoothed histogram s_0 , moving upwards and downwards, the first class not equaling null, is the upper respectively lower reference line. As we already determined the baseline in the previous pipeline step it is enough to simply rotate this line using the angle between baseline and x-axis to obtain the rotated baseline.

To approximate the core line, we use an approach similar to the approach for determining the baseline, but instead of identifying column minima we are looking for column maxima and their peak values. Unfortunately, the existence of ascenders in normal handwritten text complicates our approach. Under the assumption that letters without ascenders are more common than letters with, the majority of the maxima should be somewhere around the core line. That is why we use a sliding window of size of 15 % of the sample height (after correcting the slope) to identify the area with the majority of local maxima. For the points within this window we calculate the arithmetic mean, which is assumed to be the y-coordinate of the core line. With this all reference lines are in place and the following features can be extracted: (8) y-coordinate of the rotated baseline, (9) y-coordinate of the core line, (10) the height of ascenders, e.g. the distance between the upper reference line and the core line, (11) the height of the letter core, e.g. the distance between the core line and the baseline, (12) the height of descenders, e.g. the distance between baseline and the lower reference line.

Letters in words in cursive handwriting have the characteristic of displaying a certain slant. This slant usually lies somewhere between 20° and 130° [5], depending on whether the characters are more slanted to the left or the right. To determine the dominant slant, a histogram, describing the distribution of angles between a line through two consecutive points and the x-axis, is computed. The class width for this histogram

is defined by us to be 1 degree. All angles not contained in the interval $[20^\circ, 130^\circ]$ are neglected. For the remaining angles the arithmetic mean is calculated and returned as feature (13) - our approximation of the dominant slant of the written sample, which is also the final new feature considered in this paper.

4 Our Attack Approach

Like in [6], which is the basis for our attack considerations, the goal of the performed attack is to create raw data of handwriting which will produce the same Bio-Hash as a specific real handwriting sample. In addition to this potential authentication impact, the attack should be fully automated (in contrast to [6] where user interaction is required) and the created raw data should look like real handwriting – which does not mean to create a handwriting which looks like the original sample, because there is not enough information left in the Bio-Hash to re-create the original.

The targets for our attack are on one hand an extended version of Vielhauers Bio-Hash algorithm (see [7]), and on the other hand our version of this algorithm which is even further extended by the 13 new features described in section 3. Regarding the version that extracts 131 features out of the online writing sample, here our attack concentrates only on the features which are solely positional dependent, meaning features which can be calculated directly from the horizontal (x) and vertical (y) positions, plus three features which can be used to compute the bounding box of the writing. This reduces the number of features considered in this version of the Bio-Hash algorithm to 83 (out of 131).

4.1 Requirements of the Attack

The attack has certain requirements: In order to create raw data, which will produce the same Bio-Hash as a specific real handwriting, some information about the chosen real handwriting is necessary (see [6]). These required inputs to the attack are the Interval matrix (see section 2) and the corresponding reference Bio-Hash. This data is used to calculate a Bio-Hash vector representation of the actual synthetic raw data and also to determine the Hamming distance between this Bio-Hash and the reference Bio-Hash. That means the attacker will be able to evaluate how similar the Bio-Hash from the synthetic writing will be in regard to the reference Bio-Hash.

4.2 Algorithm Layout of the Attack

The attack presented here uses an evolutionary algorithm to compute improved synthetic handwritings. It extends the attack presented in [6] by removing the dependence on user input and creating much more natural looking raw handwriting samples.

The initial step for the approach used in this paper is to generate the synthetic handwriting required as input for the evolutionary algorithm. We choose to let a subject write a text and extract the single characters as raw sample data to compose a

dictionary that we call in the following the basic set. With such a basic set it is possible to create own words by simply concatenating individual characters drawn from the basic set. Several rules can be specified for the creation. Here, the following set of basic rules is applied:

1. Scale the character in a way that they have a similar height as the reference
2. Alignment of the characters to their base line with respect to their position on the y-axis
3. Use a fixed distance between the characters with respect to their position on the x-axis
4. Start a word with an upper-case character followed by lower case characters
5. Try to create a word with a bounding box as similar as possible to the bounding box derived from the reference Bio-Hash

The advantage of our approach is that it is possible to use the handwriting style of the basic set. Therefore, the newly created synthetic handwriting as input for the evolutionary algorithm is by design looking very natural. It has to be admitted that there are some deviations from this rule. For example the connections between letters can look unnatural since they are created by simply connecting the last point of a letter with a straight line to the first point of his successor.

In an evolutionary algorithm there are two general types of operations to change data: Mutation (1:1 operations) and recombination (n:n operations). A recombination tries to combine the positive features of different individuals together in one. The problem in this case is the computation of the features: Most of them are dependent on every single data point in the raw data, which means if two individuals are combined; the result would not have features partially of one and partially of another individual, but completely new ones. For this reason recombinations are excluded from our work and only mutations are used for our attack approach. Nevertheless, it has to be ensured that the applied mutations are capable of producing results across the whole sample space. This means especially:

- A data point can move by mutations to any location
- A sample can have any number of data points
- There can be any number of pen-ups in a handwriting and they can be everywhere

Additionally, mutations should not impose strong degradations to the naturalness of the original synthetic handwriting. The set of mutation operations used by us is: Point change (controlled by a radius), add or remove pen up, exchange a character (from the basic set) by another, dilation of a handwriting in width and/or height, rotation (of writing or character), changes of the slant (writing or character) and slight movements of a whole character. It has to be admitted that the sample space would only fully covered if the char-change-mutation would be able to create any number of data points, which is not the case in our work. Even so, the sample space is covered enough which is why there is no need for the implementation of add-character or delete-character mutations.

In the evolutionary algorithm, the fitness function for the rating of a generated synthetic writing is the hamming distance to the reference Bio-Hash and as selection method elitism is applied with parameterization of 50%. Thereby, only the best 80 individuals out of a generation of a total of 160 synthetic writing samples are carried over into the next generation.

5 Results

The goal of our evaluations is to determine, whether the new features introduced in section 3 improve the security in regards to the attack introduced in section 4. Due to the fact that the current Matlab implementation of the attack is not optimized for fast processing, the computation time for an attack can take over three days of time, depending on the number of sample points in the writings and the performance of the hardware. Since we only had a limited amount of time to conduct our evaluation; only attacks on seven sample signatures are performed here. Every attack is parameterized with 350 generations to compute the final synthetic writing.

For evaluation purposes we log the fitness of every individual per generation (i.e. the hamming distance to the reference Bio-Hash) and for every feature per generation the number of synthetic handwriting samples which implement the feature incorrectly, i.e. which do not achieve a feature value close enough to the reference Bio-Hash to successfully authenticate this feature.

A first approach to evaluate the usefulness of the new features against the attack presented in this paper is to survey if and how often a feature could be successfully created / implemented in the generated synthetic samples. This is of great interest because if a feature cannot be implemented in a synthetic writing it is very secure. In contrast, if most mutations lead to a correct implementation it would be very insecure. Using the number of incorrect implementations for every feature, a hit-score is computed in every generation. For the list of hit scores computed for a feature, a tendency indicator is created for every generation by subtracting the value for the current generation by its successor. A positive value for this tendency indicator signals an improvement for the new generation. A negative number implies degradation. At the end, out of the computed tendency indicators a global tendency for the feature is computed. Since only the correct implementations are of interest, only positive value tendency indicators are added up. The computed number is not the number of correct implementations of this feature, but gives nevertheless a value to compare the old and the new features in regards to their reachability. The hit-score average for the old features is 287.62 while the average for the new features is 349.40. These values lead to a first impression that the new features are far worse than the old ones. However there is a special case which can distort the basis data, which is when a feature was correctly implemented in every individual of the first generation and the implementation rate never degraded. In such a case the feature was correct in every instance of a synthetic writing, but never added something to the hit-score. To take a closer look onto this problem; these cases are counted for every feature in every of the seven attack simulations, class-divided for old and new features and normalized, i.e. divided

by the number of attack simulations and number of features. The resulting values are 1.08 for the old and 0.62 for the new features. This leads to the conclusion that, although the hit score is far worse for the new features, we cannot conclude that the new features are performing worse than the old ones.

The hit-score is an attempt to compare the two categories: old features and the new features introduced in this paper. However, one could take a single feature and compare it to the rest of the features. In this case two of the new features are very noticeable: While most of the features seem to be reproducible with average difficulty, the third (y-coordinate of interception of baseline and y-axis) and the fifth feature (y-coordinate of the left interception point of baseline and bounding box) create in a considerable proportion of the attack simulations the special case that not a single synthetic writing implemented them correctly. While in case of the fifth feature three out of seven attacked hashes resulted in not a single correct implementation; the third feature showed no correct implantation in five out of seven attacked hashes. Based on our evaluation data no other feature, neither in the old nor in the new set, created such a special case this often. This indicates that these two new features show a strong potential to increase the attack resilience of the Bio-Hash algorithm.

6 Summary and Future Work

Our evaluations show, that like the old feature set described in [7], every new feature can be implemented correctly in a synthetic writing created with our attack. However, we cannot say that the older features are more secure than the new ones in regards to the presented attack, or otherwise. Nevertheless, it was shown, that two features of the new ones seem to be exceptionally secure. Further tests should be done to support or reject this observation and to evaluate what makes these features more secure than the rest.

In regards to the outcomes of the attacks: The synthetic handwriting as output of the attack looks for the most part very naturally. An example for such a synthetic writing is shown in Figure 1. Nevertheless, it has to be admitted that in some cases unnatural characteristics, e.g. a straight linking line between two letters, are created.



Fig. 1. Examples for real (left) and artificially created handwriting - in the middle an output of the attack tool from [6] and right an output of the attack introduced here

Regarding open issues for future work, the first step should be the extension of the evaluation set used. The number of seven sample signatures used as attack target here are suitable for first indications on the performance of the new features and attack, but statistically significant evaluations would require larger variance.

Also, a wider variety in the basic set used as input for the evolutionary algorithm (and therefore a more diverse starting point for the attack) should be considered. This could be achieved by using handwriting samples from different persons in contrast to only one person used in this paper. A drawback might be that the output artificial writing might show a much stronger divergence regarding different instances of the same letter and would therefore lose some of its natural look.

Acknowledgements. This work is supported by the German Federal Ministry of Education and Research (BMBF), project "OptiBioHashEmbedded" under grant number 17N3109. The content of this document is under the sole responsibility of the authors. We also like to thank Prof. Jana Dittmann of the Otto-von-Guericke University Magdeburg and the StepOver GmbH for supporting the project "OptiBioHashEmbedded".

References

1. Vielhauer, C.: Biometric User Authentication for IT Security: From Fundamentals to Handwriting. Springer, New York (2006)
2. Senior, A.W., Robinson, A.J.: An Off-Line Cursive Handwriting Recognition System. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 309–321 (1998)
3. Bresenham, J.: Algorithm for computer control of a digital plotter. *IBM Systems Journal* 4 (1965)
4. Yanikoglu, B., Sanson, P.: Off-Line Cursive Handwriting Recognition Using Style Parameters, Dartmouth College, Computer Science, Version: 1993 (PCS-TR93-192) (1993), <http://www.cs.dartmouth.edu/reports/TR93-192.ps.Z>
5. Koppenhaver, K.: Forensic Document Examination: Principles and Practices. Humana Press (2007)
6. Kümmel, K., Vielhauer, C., Scheidat, T., Franke, D., Dittmann, J.: Handwriting Biometric Hash Attack: A Genetic Algorithm with User Interaction for Raw Data Reconstruction. In: De Decker, B., Schaumüller-Bichl, I. (eds.) CMS 2010. LNCS, vol. 6109, pp. 178–190. Springer, Heidelberg (2010)
7. Makrushin, A., Scheidat, T., Vielhauer, C.: Improving reliability of biometric hash generation through the selection of dynamic handwriting features. In: Shi, Y.Q., Katzenbeisser, S. (eds.) Transactions on DHMS VIII. LNCS, vol. 7228, pp. 19–41. Springer, Heidelberg (2012)

Video Watermarking Scheme with High Payload and Robustness against Geometric Distortion

Huajian Liu, Yiyao Li, and Martin Steinebach

Fraunhofer SIT, Rheinstr. 75, Darmstadt, Germany
{liu, steinebach}@sit.fraunhofer.de

Abstract. Besides copyright protection, digital video watermarking is also applied in non-security applications like second screen annotation, where high robustness against geometric distortions and high watermark payload are required. Robustness against geometric distortions, however, is still one of the major challenging issues in video watermarking, in particular for the schemes in compressed domain. In this paper, we propose a video watermarking scheme that can resist geometric attacks. The watermark embedding is performed in Fourier domain using patchwork method which is able to handle high embedding payload. Fast transform between block DCT and DFT enables the proposed scheme to be applicable directly in the compressed domain, significantly reducing the computation cost. Perceptual masking is applied in both DFT and DCT domains to ensure high visual quality. Experimental results demonstrate that the proposed scheme achieves satisfactory robustness against all kinds of attacks, including geometric distortions and frame dropping and swapping.

Keywords: video watermarking, annotation application, geometric distortion, temporal attack.

1 Introduction

Digital video watermarking is commonly used as an effective technique for copyright protection of video content [1]. With the increasing employment in practice, non-security applications based on watermarking technology also attract great interest. For instance, in the recent emerging second screen application, digital watermark can serve as an invisible annotation of the video content, which can be captured by a smartphone to retrieve related information of the ongoing scenes.

Depending on particular applications, video watermarking has to comply with corresponding requirements, such as transparency, payload, robustness and security. For copyright protection, security against removal attacks and robustness to common video processing are of high concern. In contrast, in non-security applications like second screen, the most significant requirements are high robustness against affine transforms and high payload while security becomes subordinate. This is because geometric distortions are inevitable during the capture of video content using a smartphone camera and a high payload of watermark information is needed to provide reasonable annotations.

In the past decades, many video watermarking algorithms have been proposed targeted for copyright protection applications [1-3]. While many video watermarking algorithms can survive common video processing, they lack robustness against geometric and temporal attacks, such as cropping, scaling, rotation, frame dropping, swapping, insertion, averaging and so forth [2-3]. For high efficiency, conventional video watermarking schemes tend to embed information in the compressed domain so as to avoid the complete decoding and encoding. Since almost all the video data are compressed as block DCT coefficients and motion vectors, such as MPEG-2, MPEG-4 and H.264, any geometric attack will regroup the blocks and entirely change the bit-stream structure, which leads to de-synchronization of the embedded watermark. In addition, temporal attacks can also compromise the watermark detection by disabling the watermark synchronization across frames. Therefore, how to tackle the geometric attacks and temporal attacks in compressed video domain is still a challenging topic.

The existing watermarking algorithms, which aim at achieving good robustness against geometric attacks, can be roughly categorized into three classes: template based scheme, feature based scheme and invariant domain based scheme [4].

Template based schemes embed a structured template in addition to the informative watermark to register the undergone geometric distortion and rectify the host data before performing watermark detection. Template based methods are reported to have good robustness against geometric attacks [4]. Its main drawback lies in the vulnerability to removal attacks. Because all copies share the same template, it can be discerned by collusion attack using multiple copies and be removed afterwards.

Feature based schemes use local or global features, which are invariant to affine transformation, to synchronize the watermark. In [5-6] Harris corners are used as local feature points in the watermarking. While the local feature points are fairly robust against scaling and rotation, they are vulnerable to cropping and require high computational complexity. In addition, some of local features suffer from most signal processing attacks like low-pass filtering. In contrast, in [7] the global feature of the histogram of a frame sequence is used for watermarking. Since the global feature is independent of the pixel position in a frame, it is robust to geometric attacks and slight cropping. However, histogram based schemes provide rather low watermark payload, which makes them of limited use in practical applications. The scheme in [7] embeds the watermark by modifying selected consecutive bins of the histogram of the average DC energy of frames. Averagely it can embed only 1 bit per 400 frames, which makes it not suitable for annotation applications like second screen. In addition, such histogram based scheme is not applicable in all types of videos. According to our experiments, when the video contains few motions between frames, the average DC energy of each frame will converge on a few bins in the histogram while the number of samples on other bins is close to zero. Thus, the embedding will fail due to the lack of sufficient bins to modify.

Invariant domain based schemes embed the watermark in a geometric transform invariant domain. The most popularly used invariant domain is Fourier-Mellin transform [8-9]. However, since log-polar mapping is involved, it is very difficult to implement watermarking in Fourier-Mellin domain without causing severe quality loss [8]. In [10] and [11], the full DCT is used as a scaling invariant domain for

watermarking, because the scaling of one frame has roughly equivalent effect to the truncation of high-frequency band in its full DCT domain. Thus, the watermark embedded in the low-frequency area will not be impacted by downscaling. Nevertheless, these schemes are not robust against other geometric distortions, like cropping and rotation. Therefore in [11] the authors have to propose another solution for cropping attack. Furthermore, embedding in low-frequency band may lead to more visual quality loss.

As previously mentioned temporal attacks like frame dropping and swapping may also destroy the watermark synchronization in video data [12]. The histogram based scheme in [7] is claimed to be robust against frame dropping, but according to our experiments, if the varying frames (scene changing) in a video clip are dropped, the distribution of the histogram of the average DC energy will be greatly changed so that the watermark cannot be detected. In [11], an additional strategy was proposed to resist frame dropping attack, which divides the video sequence into scenes and accordingly partitions the watermark into groups of bit strings. For each scene, the corresponding group of bit is redundantly embedded into all frames within the scene. However, this method can only ensure the watermark integrity in case of frame dropping inside a scene. If the whole scene is dropped or the video is truncated, the extracted watermark will become incomplete.

In this paper, we proposed a video watermarking scheme targeted for annotation applications like second screen. As security is not much concerned, the template based method is adopted to achieve the robustness against geometric distortions because of its good performance. Besides the informative watermark message, a structured template is embedded in the Fourier domain to register the undergone geometric transformations. The watermark embedding is done using an effective patchwork method in the same Fourier domain, which is able to handle a high watermark payload. As well known, high payload will cause more quality degradation. Hence, a perceptual masking is applied to reshape the embedded energy to ensure good visual quality. In addition, in order to tackle the problem of temporal attacks, each watermark segment in separate frames is indexed with a unique ID to be self-synchronizable. Furthermore, to lower the computation cost and achieve high efficiency, the embedding process is done directly in the compressed domain by applying a fast transform between block DCT and 2D-DFT. To demonstrate the performance of the proposed scheme, the experimental results are compared with other schemes which use feature based and invariant domain based methods resisting geometric distortions.

The rest of the paper is organized as follows. In section 2, we introduce the fast transform between block DCT and 2D-DFT. The proposed video watermark scheme is presented in Section 3. Experimental results are given in Section 4 and we conclude the paper in Section 5.

2 Fast Transform between Block 2D-DCT and 2D-DFT

Nearly all video data are stored and transmitted in compressed format and most commonly used video compression methods, such as MPEG-2, MPEG-4 and H.264, apply

block DCT for spatial redundancy reduction. Thus, if another transform domain, e.g. DFT and DWT, is applied in the watermark embedding, the video data has to be first decoded to its uncompressed format and then transformed to the watermarking domain. When the embedding is finished, the watermarked data must be again inversely transformed to the spatial domain and then transformed into block DCT. The whole process involves four mathematical transforms and is usually computationally complex and time consuming.

To avoid the complete decoding and encoding, a direct embedding using block DCT is highly desired. As both DCT and DFT are linear and invertible transforms, a linear relationship exists between block DCT coefficients and DFT coefficients of the full frame [13]. Therefore, we can directly attain a frame's DFT coefficient without first decoding it into its uncompressed format.

Given a frame Y of size $LN \times LM$, where L is the DCT block size, the relationship between its block 2D-DCT and 2D-DFT can be denoted by

$$F = P\hat{C}Q^T, \hat{C} = P^{-1}F(Q^{-1})^T \quad (1)$$

where F is the 2D-DFT of Y and \hat{C} is the stacked block DCT of Y . $P = A_c \hat{B}_{LN}^{-1}$ and $Q = A_r \hat{B}_{LM}^{-1}$, where A_c and A_r are the DFT transform matrices for column and row, \hat{B}_{LN} and \hat{B}_{LM} are the stacked block DCT transform matrices.

Note that the transform matrix P and Q are determined by M , N and L , i.e. the frame size and block size, being independent of the frame content. Since the frame size in a video is constant, these transform matrices can be calculated in advance. This will significantly reduce the computation cost at embedding.

3 Proposed Video Watermarking Scheme

As shown in Figure 1, the proposed video watermarking scheme consists of the following four parts: frame grouping, watermark encoding, template and watermark embedding, and watermark reshaping.

3.1 Frame Grouping

In contrast to image, video data normally contains a large number of frames. Effective utilization of this feature could help improve the watermarking performance in terms of payload, robustness and video quality.

In our scheme we apply a frame grouping strategy to separate the template and watermark embedding into different frames. The structured template can not only rectify the geometric distortion but also be used as a temporal synchronization signal.

Assuming a video clip that contains K frames, we divide all the frames into G groups, resulting in n frames in each group. Then the first t frames will be used for template embedding, denoted as template frame, and the rest $n-t$ frames for watermark embedding, denoted as watermark frame.

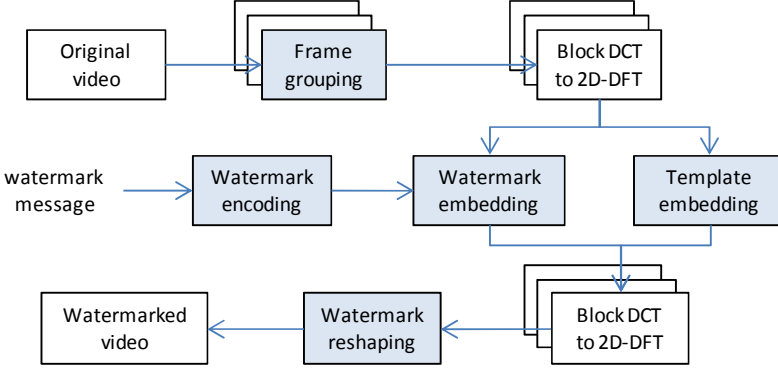


Fig. 1. Diagram of proposed watermarking scheme

3.2 Watermark Encoding

In the proposed scheme, we desire to embed a watermark payload which can meet the common requirement for reasonable annotation while preserving the video perceptual quality. In order to effectively utilize the watermark capacity of all frames, when necessary, the watermark message is segmented and spreads over the $n-t$ frames in each group. However, as mentioned in Section 1, watermarks embedded across frames will suffer from temporal attacks, which will disable the synchronization between the watermark segments in different frames. Hence, in order to be resynchronizable after distortions like frame missing, the watermark message is encoded as follows.

To ensure the integrity of extracted watermark, the original watermark message is first encoded using cyclic redundancy check (CRC) code. The encoded message m_i , $i=1,2,\dots,N_w$, is then divided into $n-t$ segments. For each segment, an indexing ID is attached to the watermark bits before embedding. Thus every single watermark segment becomes self-synchronizable. The bit length of the indexing ID is determined by the number of watermark segment. Furthermore, the resulting watermark bits in each segment are further encoded using BCH code, generating the final encoded watermark, w_k , $k=1,2,\dots,N_c$.

3.3 Watermark and Template Embedding

In every frame group, each watermark frame is used to embed one watermark segment w_k . The encoded watermark segment is embedded in the DFT domain of the corresponding frame using the patchwork algorithm in [14]. The DFT magnitude coefficients in the middle-frequency band are used for embedding in order to achieve better robustness against low-pass filtering attacks and lossy compression. In this work, the used frequency band ranges from 0.15 to 0.35 in the normalized frequency. The embedding payload in each frame is limited to N_c . The patchwork algorithm in [14] yields good robustness against a large number of attacks and the ability to resist cropping attack. Moreover, the patchwork technique enables blind detection which means that no original video data has to be supplied during the detection process.

In each template frame of every frame group, a structured template is embedded in the same DFT domain using the same approach as in [14], which can register the undergone geometric transformations. The template contains 14 points uniformly distributed along two lines in the first and second quadrants of the magnitude spectrum. The angles of the lines and the radii of points can be randomly chosen, controlled by a secret key. Nevertheless, they should avoid extremely low or high frequency band, because the energy embedded in high frequency can be easily reduced by lossy compression and the energy embedded in low frequency will cause noticeable artifacts. In this work, the template points reside in a normalized mid-frequency band of [0.25, 0.35]. In order to ensure a good fidelity on average, the same embedding strength is used for the template embedding in all frames.

In case that the watermarked video is geometrically distorted, the detected transformation parameters out of the template shall be used to rectify the video before carrying out watermark detection. Moreover, the embedded template information also offers a means of identifying frame groups or synchronizing video temporally in case that the video is truncated or some frames are dropped.

3.4 Watermark Reshaping

Although the embedding strength is adaptively controlled in the DFT domain based on the spectrum properties, like magnitude of the coefficients and their corresponding frequency bands, these properties doesn't cover all the characteristics of human visual system. For example, no local perceptual analysis of pixels can be done in Fourier domain, because Fourier transform only provide a global frequency analysis.

Since we obtain the DFT coefficients directly from block DCT coefficients out of compressed video stream, there is no access to the pixel values. After inversely transformed to block DCT, the embedded energy is distributed over the whole frame which will degrade the visual quality.

To tackle the perceptual error and conform the embedding to the visual content more precisely, we utilize Watson perceptual model in DCT domain [15] to reshape the embedded energy,

$$\Delta C(i, j, k)_m = \rho \cdot m'_{ijk} \cdot \Delta C(i, j, k) \quad (2)$$

where $\Delta C(i, j, k)$ is the original watermark energy for the DCT coefficient $C(i, j, k)$ in the k th block. m'_{ijk} is the normalized contrast masking threshold [15], ρ is a scaling factor to ensure the total embedding energy is conserved after reshaping.

$$\rho = \sqrt{\sum \Delta C(i, j, k)^2 / \sum (m'_{ijk} \cdot \Delta C(i, j, k))^2} \quad (3)$$

3.5 Watermark Detection

Since the watermarked video may have suffered from both geometric and temporal attacks, the number of frames may have changed and the position of frames may also

be different from the original video. Hence, in the detection process, we first scan the frames to detect the template information. If a valid template is found, the detected scaling and rotation parameters are used to rectify the whole video sequence. Since there might exist many frames with template information detected, we choose the one with the minimal MSE_T , which is the template detector response, indicating the accuracy of template detection, as global optimum.

Following geometric rectification, the watermark detection is performed. The detected template frames are considered as synchronization points of frame groups and the following frames are examined as a group for watermark detection. When the watermark message is embedded across frames, the detected watermark segments will be decoded and united together by their attached indexing ID. Since some frames may be dropped or swapped and the watermark embedded in some frame may be impaired by possible attacks, it may be necessary to combine detected watermark segments from more than one frame group to obtain a complete watermark message. Then the integrity of the detected message will be double-checked by the CRC code.

4 Experimental Results

In our experiments the size of frame group is set to 22, in which 2 frames are used for template embedding and 20 frames for watermark embedding. The indexing ID is 5 bit long and the watermark payload is 12 bits per frame. The net payload is 48 bits and the total payload reaches 240 bits per frame group after BCH and CRC encoding.

Three video sets with different resolution and frame rate are used to evaluate the proposed scheme as listed in Table 1. Every data set contains some or all of the following standard test video clips: Mobile Calendar, Park Run, Shields and Stockholm¹.

Table 1. Standard test data set information

Test Data	Set 1	Set 2	Set 3
Resolution	1920x1080	1280x720	720x576
Frames	252	504	252
Frame rate	25 fps	50 fps	25 fps
Color subsampling	4:2:0	4:2:0	4:2:0

To demonstrate the performance, we compare the results of our scheme with two other video watermarking schemes which uses feature based and invariant domain based methods to resist geometric distortions. One is the histogram based algorithm proposed in [7], the other is the full DCT based algorithm proposed in [10]. In the following tables, the three schemes are denoted as Hist, Dct and Our, respectively.

Table 2 shows the average PSNR of the four watermark videos. As only 2 bits are embedded by the histogram based algorithm, it yields the highest PSNR values. The full DCT based algorithm has embedded 12 bits. Regardless of the low watermark

¹ The used video test sets can be downloaded via

ftp://ftp.ldv.ei.tum.de/videolab/public/SVT_Test_Set/

Table 2. Average PSNR (dB) of watermarked videos

Schemes	Hist	Dct	Our
Mobile Calendar	63	34	53
Park Run	62	34.5	57
Schiolds	61	34.9	54
Stockholm	63	35	53

payload, the watermarked video is of worst quality. This is due to the embedding in low frequency band of DCT. In contrast, our scheme shows a good tradeoff between watermark payload and visual quality. The average PSNR is above 53dB, which means the embedded watermark is totally invisible.

In the evaluation of robustness, the following types of attacks are considered:

- Geometric attacks: rotation, scaling, cropping,
- Removal attack: lossy compression,
- Signal processing attack: blur, sharp,
- Temporal attack: temporal smoothing, frame dropping, frame swapping.

The test results regarding robustness against video processing and geometric attacks are shown in Table 3. All the values are the average testing results on all the three test sets in Table 1. A "-" means that the watermark detection fails. As can be observed, the histogram based scheme yields very high robustness on Mobile Calendar and Schields, but completely fails on Park Run and Stockholm. This result proves our discussion in the previous section. The video Park run and Stockholm don't have enough varieties between frames, so the histogram of the average DC energy contains only limited number of bins which makes robust embedding unfeasible.

The full DCT based algorithm shows overall low robustness. It is robust to scaling attack but fragile to rotation and cropping. Our scheme demonstrates in general high robustness against all the attacks. Moreover, it is independent of video types, so our scheme outperforms the other two schemes in terms of tradeoff between payload, robustness and adaptability.

The test results on frame dropping and frame swapping are listed in Table 4. Since in practical applications the frame dropping and swapping attacks are often combined with other attacks, the robustness against frame dropping and swapping is tested together with other seven attacks respectively, including geometric attacks and signal processing attacks. The listed values in Table 4 are the average testing values on all of the four videos in Test Data Set 3, i.e. with a resolution of 720x576.

As shown in Table 4, both our scheme and histogram based algorithm yield high robustness against any attack combination between frame dropping/swapping and other attacks. The bit error rate of the detected watermark using these two schemes is always kept to zero or very close to zero. The full DCT based algorithm loses the competition again with the much higher bit error rate values. These test results demonstrate that it is an effective mean to resist frame dropping/swapping by attaching an indexing ID to each watermark segment.

Table 3. Average detection BER after various attacks

Attacks	Mobile Calendar			Park Run			Schiolds			Stockholm		
Schemes	Hist	Dct	Our	Hist	Dct	Our	Hist	Dct	Our	Hist	Dct	Our
Rotation 1°	0	0.08	0.06	-	0.11	0	0	0.08	0.01	-	0.08	0
Rotation 2°	0	0.08	0	-	0.11	0	0	0.08	0	-	0.08	0
Rotation 3°	0	0.11	0.04	-	0.11	0	0	0.11	0.03	-	0.08	0
Scaling 0.5	0	0.08	0	-	0.05	0	0	0.05	0	-	0.08	0
Scaling 0.6	0	0.08	0	-	0.05	0	0	0.05	0	-	0.08	0
Scaling 0.7	0	0.08	0	-	0.05	0.03	0	0.05	0	-	0.08	0
Scaling 0.8	0	0.08	0	-	0.05	0	0	0.08	0	-	0.08	0
Cropping 12%	0	0.08	0	-	0.16	0	0	0.16	0	-	0.16	0
Cropping 22%	0	0.11	0	-	0.16	0	0	0.16	0	-	0.08	0
Cropping 32%	0	0.14	0	-	0.16	0	0	0.16	0	-	0.08	0
MPEG-2	0	0.11	0.04	-	0.11	0	0	0.11	0.01	-	0.08	0
MPEG-4	0	0.08	0	-	0.08	0	0	0.08	0	-	0.08	0
H.264	0	0.11	0.14	-	0.08	0	0	0.08	0	-	0.08	0
Blur	0	0.08	0	-	0.05	0	0	0.05	0	-	0.08	0
Sharpen	0	0.08	0	-	0.05	0	0	0.05	0	-	0.08	0
Temporal smoothing	0	0.08	0	-	0.05	0	0	0.05	0	-	0.08	0

Table 4. Average detection BER after frame dropping and swapping

Attacks	Drop 40 frames			Drop 120 frames			Frame swapping		
Schemes	Hist	Dct	Our	Hist	Dct	Our	Hist	Dct	Our
Rotation 2°	0	0.08	0	0	0.16	0	0	0.08	0
Scaling 0.8	0	0.08	0	0	0.08	0	0	0.08	0.04
Cropping 80x80	0	0.08	0.03	0	0.08	0.07	0	0.08	0
MPEG-4	0	0.08	0	0	0.16	0	0	0.08	0
Blur	0	0.08	0	0	0.08	0	0	0.08	0
Sharpen	0	0.08	0	0	0.08	0	0	0.08	0
Temporal smoothing	0	0.08	0	0	0.08	0	0	0.08	0

5 Conclusion

To be applicable in practical annotation applications, video watermarking has to be robust and efficient. Geometric and temporal attacks are more challenging than other attacks like lossy compression and transcoding, in particular for watermarking schemes in compressed domain. In this paper, a robust video watermarking scheme is proposed for annotation applications, which combines different strategies for fast embedding, high robustness and payload, and good visual quality. The watermark embedding is done in DFT domain using patchwork approach. Fast transform between block DCT and 2D-DFT enables the direct embedding in compressed domain. Structured template is embedded in selected frames to rectify possible geometric

distortions. Each watermark segment embedded in separate frames is individually indexed so that it is able to be resynchronized in case of temporal attacks.

In addition, although the embedding position of template can be randomized by a secret key, its security against removal attack needs further investigation and improvement in the future work for the applications with high security requirement.

References

1. Langelaar, G.C., Setyawan, I., Lagendijk, R.L.: Watermarking digital image and video data: A state-of-the-art overview. *IEEE Signal Processing Magazine* 17(5), 20–46 (2000)
2. Shojanazeri, H., Adnan, W.A.W., Ahmad, S.M.S., Saripan, M.I.: Analysis of watermarking techniques in video. In: *Proceedings of International Conference Hybrid Intelligent Systems, HIS*, pp. 486–492 (December 2011)
3. Jayamalar, T., Radha, V.: Survey on digital video watermarking techniques and attacks on watermarks. *International Journal of Engineering Science and Technology* 2(12), 6963–6967 (2010)
4. Zheng, D., Liu, Y., Zhao, J., Saddik, A.E.: A survey of RST invariant image watermarking algorithm. *ACM Computer Survey* 39(2), 1–91 (2007)
5. Liu, Y., Zhao, J.: A robust RST invariant image watermarking method based on locally detected features. In: *Proceedings of IEEE International Workshop on Haptic Audio Visual Environments and their Applications*, pp. 133–138 (October 2005)
6. Zhuang, C., Zhang, L., Tian, X., Xia, S.: A novel anti-geometric-attacking watermarking algorithm based on Harris feature points. In: *Proc. International Congress Image and Signal Processing, CISP*, vol. 2, pp. 1003–1007 (October 2011)
7. Chen, C., Ni, J., Huang, J.: Temporal statistic based video watermarking scheme robust against geometric attacks and frame dropping. In: Ho, A.T.S., Shi, Y.Q., Kim, H.J., Barni, M. (eds.) *IWDW 2009. LNCS*, vol. 5703, pp. 81–95. Springer, Heidelberg (2009)
8. O’Ruanaidh, J.J.K., Pun, T.: Rotation, scale and translation invariant digital image watermarking. In: *Proc. Int. Conf. Image Processing*, vol. 1, pp. 536–539 (October 1997)
9. Lin, C., Wu, M., Bloom, J., Cox, I., Miller, M., Lui, Y.: Rotation, scale, and translation resilient watermarking for images. *IEEE Transactions on Image Processing* 10(5), 767–782 (2001)
10. Ling, H., Wang, L., Zou, F., Chen, J.: Low-complexity video watermarking scheme resisting geometric distortions. In: *Proc. IEEE Int. Conf. Signal Proc., ICSP*, pp. 1861–1864 (2010)
11. Wang, Y., Pearmain, A.: Blind MPEG-2 video watermarking robust against geometric attacks: a set of approaches in DCT domain. *IEEE Trans. Image Processing* 15(6), 1536–1543 (2006)
12. Lin, E.T., Delp, E.J.: Temporal Synchronization in Video Watermarking. *IEEE Transactions on Signal Processing* 52(10), 3007–3022 (2004)
13. Davis, B.J., Nawab, S.H.: The relationship of transform coefficients for differing transforms and/or differing subblock sizes. *IEEE Trans. Signal Proc.* 52(5), 1458–1461 (2004)
14. Liu, H., Steinebach, M.: Improved Fourier domain template and patchwork embedding using spatial masking. In: *Proceedings of IS&T SPIE Electronic Imaging - Media Watermarking, Security, and Forensics*, no. 8303 (January 2012)
15. Watson, A.B.: Visually optimal DCT quantization matrices for individual images. In: *Data Compression Conference*, pp. 178–187 (1993)

Use of Linear Error-Correcting Subcodes in Flow Watermarking for Channels with Substitution and Deletion Errors

Boris Assanovich¹, William Puech², and Iuliia Tkachenko^{2,3}

¹ YK State University of Grodno, Belarus
bas@grsu.by

² LIRMM, UMR CNRS 5506
University of Montpellier 2, France
{William.Puech, Iuliia.Tkachenko}@lirmm.fr

³ Authentication Industries, Montpellier, France

Abstract. An invisible flow watermarking QIM scheme based on linear error-correcting subcodes for channels with substitution and deletion errors is proposed in this paper. The evaluation of scheme demonstrates similar to known scheme performance but with lower complexity as soon as its implementation is mainly based on linear decoding operations.

Keywords: Flow watermarking, inter-packet-delay, deletion and substitution errors, linear error-correcting codes, VT-codes, quantization index modulation.

1 Introduction

Recently, an active approach of traffic analysis called “flow watermarking” has been considered. This approach attempts to manipulate the statistical properties of packets flow to insert a watermark making it easier to detect the flow after passing through one or more relay hosts. To prevent an attacker to tolerate the packet delays and to eliminate embedded watermark, recent schemes have concentrated on making them “invisible”. This technique has been the subject of increased interest in the past decade, because it requires low computational and communication cost while providing high accuracy in linking traffic flows.

Flow watermarking is also classified as *interval*-based and *inter-packet-delay* (IPD)-based. The first type of watermarking technique is robust to packet losses but is vulnerable to the *multi-flow attack* [1]. The IPD-based flow watermarking, in which the watermarks are embedded into the time intervals between arrivals of packets, resists this attack [2]. The drawback of this scheme is that it can cause a lot of errors in decoding during the loss of packet synchronization in the watermark detection process.

A novel IPD-based flow watermarking scheme that can withstand packet losses has been done in [2]. In this scheme the watermark embedding has been done with the use of quantization index modulation (QIM) [3]. In this approach the embedded marks are invisible. To withstand packet losses authors develop a Hidden-Markov Model

(HMM) decoding scheme considering the communication channel with dependent deletion and substitution errors. However, the proposed watermark detector based on a maximum likelihood decoding algorithm paired with a forward-backward is of high complexity and requires a lot of computational resources.

In this paper we propose the alternative IPD-flow watermarking QIM scheme, based on the use of linear error-correcting codes and Varshamov–Tenengol'ts (VT) codes [4] to reduce the complexity of flow watermarking method [2].

2 Error Correcting Codes

2.1 Linear Error-Correcting Codes

Most practical error-correcting codes used today in watermarking are binary. Linearity is an important structural property of codes, allowing a concise representation of codes, the accompanying encoding and decoding rules as well as the determination of the errors are correctable/detectable. A very good survey of the theory of error-correcting codes is done in [5] and the only necessary definitions are used throughout a paper. The symbols of binary linear codes are the elements of a field $GF(2)=\{0;1\}$ which is a code alphabet. Generally, a binary code C is defined as a set of finite sequences (vectors) $\mathbf{x} = (x_1 \dots x_n)$, called codewords, encoded with the use of corresponding message vectors $\mathbf{b} = (b_1 \dots b_k)$ from code symbols $x_i, b_i \in GF(2)$.

Linear $[n,k,d]$ -code is defined by following parameters: Hamming distance between any binary codewords $d(\mathbf{x}_i; \mathbf{x}_j)$, weight of a codeword $wt(\mathbf{x}_i)$ and a code rate $R=k/n$. Any linear code C is completely defined by generator matrix \mathbf{G} and parity-check matrix \mathbf{H} whose columns and rows are respectively linearly independent. Every codeword of a linear block code C is a linear combination of the rows of a generator matrix \mathbf{G} . The error correction capacity t of linear error-correcting codes strictly depends on its minimum distance $d_{min}=\min\{d(\mathbf{x}_i; \mathbf{x}_j)\}$ and weight distribution of a code. To perform an error correction in codeword \mathbf{y} , corrupted by t or less errors, a rather simple method of bounded distance decoding with syndrome could be applied. It consists of following steps: the calculation of syndrome for a received word \mathbf{y}

$$\mathbf{S} = \mathbf{y} \cdot \mathbf{H}^T, \quad (1)$$

search for a most plausible error pattern \mathbf{e} , the estimation of transmitted codeword \mathbf{x}' . Decoder picks error pattern \mathbf{e} of smallest weight satisfying $\mathbf{e} \cdot \mathbf{H}^T = \mathbf{S}$. All procedures of decoding with syndrome are linear and only the second step requires a nonlinear operation that can be performed by look-up tables.

For example, the linear $[6,3,3]$ -code $C=\{(000000), (110100), (011010), (101110), (101001), (011101), (110011), (000111)\}$ is completely defined by its generator matrix \mathbf{G} [5, p.357-367]. To change its properties, a binary code can be easily modified by different techniques [5]. The number of its codewords can be increased or decreased. The process of deleting a codeword from the basis of C to obtain a new code C' , where the minimum weight of remains the same, is referred to as taking a subcode of C .

It is known that linear codes as the other error correcting are applied for channels with substitution errors when transmitted symbols are received as the other symbols. However, there are channels that suffer from synchronization errors, which are associated with not receiving transmitted symbols leading to deletion errors. Therefore, there is a compelling reason to consider codes that, not only correct substitution errors, but can also recover from deletion errors. Recently it has been proved [6] that linear codes and all cyclic codes can correct a single deletion or insertion error but not by both types of errors [6].

As opposed to [6] the application of binary linear codes for the correction of both types of errors by the same subcodes with the use of two different decoders will be proposed. In Section 2.2, we define the VT-codes and show how to get a subcode of a linear error-correcting code to combat with substitution and deletion errors.

2.2 VT-Codes for Deletion and Substitution Errors Correction

Given a parameter a , with $0 \leq a \leq n$, the Varshamov-Tenegol's (VT) code $VT_a(n)$ is the set of binary words $\mathbf{x}=(x_1 \dots x_n)$ of length n so that the equality satisfies [4]:

$$\sum_{i=1}^n ix_i \equiv a \pmod{(n+1)}. \quad (2)$$

These codes are single error correcting codes and optimal for $a=0$ as it was conjectured in [4, 7] and will be discussed in this paper.

For example, after calculation $\sum_{i=1}^n ix_i \equiv 0 \pmod{7}$ the code $VT_0(6)$ with block length $n = 6$ is $VT_0(6)=\{(000000),(001100),(010010),(011110),(100001),(101101),(110011),(110100),(111111)\}$. Any code $VT_0(n)$ can be used to communicate reliably over a channel that introduces at most one deletion in a block of length n . Levenshtein proposed a simple decoding algorithm [8] based on the deficiency in checksum and weight calculation for a VT code.

As an example, assume the code $VT_0(6)$ is used and $\mathbf{x}=(110100) \in VT_0(6)$ is transmitted over the channel. If the first bit in \mathbf{x} is deleted and $\mathbf{y}=(10100)$ is received, then the new checksum is 4, and the deficiency $D=7-4=3 > wt(\mathbf{y})=2$. The decoder inserts a 1 after $n-D=3$ 0's from the right to get (110100) . Thus a very simple algorithm of low complexity can be used to decode $VT_0(n)$ with deletion correction.

However, in general the $VT_0(n)$ codes are nonlinear and the dimension k is to get a linear $[n,k]$ is bounded by $k \leq \lfloor n/2 \rfloor$ [6]. We use this result and propose an approach to find a linear substitution and deletion correction code from VT-code. The algorithm contains of the following steps: organize the codewords of $VT_0(n)$ code in a lexicographically order; choose the k linear independent codewords of maximum weight preserving $d(\mathbf{x}_i; \mathbf{x}_j) \geq d_{min}$; produce \mathbf{G} and \mathbf{H} of C making the linear combinations of chosen VT-codewords. The use of this algorithm results in a subcode C' that has at least $k+1$ codewords of $VT_0(n)$ code as soon as the linear combination of any codeword with itself makes a codeword $(0\dots 0)$, which is also a codeword of $VT_0(n)$ code. Considering the use of the proposed above algorithm, the flowing generator and parity check matrixes for a modified $[6,3,3]$ -code C' have been made:

$$\mathbf{G}' = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}, \mathbf{H}' = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}. \tag{3}$$

The use of \mathbf{G}' and \mathbf{H}' from (3) results in a code set C' with an increased number of codewords that belong to $VT_0(6)$, compared with C . If we prune the codewords that are not the codewords of $VT_0(6)$ and a codeword with all zeros, we can make a subcode with the necessary properties $C^* = \{(110100), (011110), (101101), (110011)\}$ consisting of 4 codewords. C^* is a linear subcode with $d_{min}=3$, at the same time it is a $VT_0(6)$ code and it can be used for error correction of one substitution and one deletion error. The examination of C^* has shown, that its code rate is reduced to approximately by $1/2$ relatively to the code rate of C . The algorithm proposed below can be applied to an arbitrary code to find the error-correcting VT-code (EC-VTC) that is a subcode C^* of a linear code. For example, there is a EC-VTC, coinciding with linear error correcting code [8,2,5] [5, p.378], consisting of 4 codewords and subcoding $VT_0(8)$. It is easily seen that it corrects one deletion and two substitution errors. It is known that the size of any $VT_0(n)$ is about $2^n/n$ [6], then an additional limitation on its linear properties leads to a reduction in its rate of at least by $1/2$.

However, to perform the independent decoding of codewords from EC-VTC placed in a continuous bit stream the boundaries of codewords must be known. We implement the independent decoding of them by accurately making a set of codewords C^r from EC-VTC with possible reduction of a code rate R and inserting the periodic markers between the codewords as discussed in Section 3.

3 System Model

3.1 Scheme for Flow Watermarking

We use the described above linear codes in a flow watermarking [1], [2] considering the channel with deletion and substitution errors. The proposed watermark embedding scheme, depicted in Figure 1, has the same embedder and extractor based on QIM [2], but the different coding and decoding principles are used.

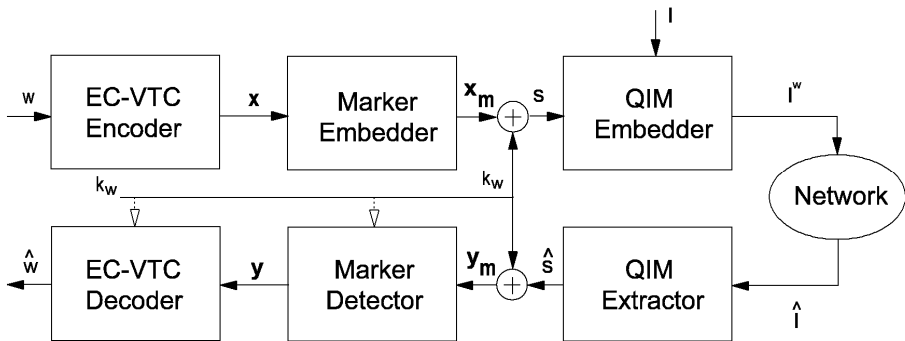


Fig. 1. System model

3.2 QIM Embedder and Extractor

For the watermark embedding the flow of IPDs is modified with the use of QIM watermarking (Fig.1). A quantization step size Δ , which is the distance between two quantizers, is used for QIM modulation:

$$I_i^w = \begin{cases} c\Delta, & \text{if } s_i = 0 \\ (c+0.5)\Delta, & \text{if } s_i = 1 \end{cases}. \quad (4)$$

As packets can only be delayed by QIM Embedder, we choose parameter c to be the smallest integer so that the change in I_i^w would delay the i -th packet. Then I^w is transmitted and after the transfer over the network it is received in the form of estimated sequence of IPDs \hat{I} and received by the QIM Extractor. For the flow \hat{I} processed by QIM Extractor, the following QIM demodulation function is used to recover the embedded bits \hat{s} :

$$\hat{s}_i = \begin{cases} \text{mod}(\lfloor 2\hat{I}_i/\Delta \rfloor, 2) & \text{if } 2\hat{I}_i/\Delta - \lfloor 2\hat{I}_i/\Delta \rfloor \leq 0.5 \\ \text{mod}(\lceil 2\hat{I}_i/\Delta \rceil, 2) & \text{if } 2\hat{I}_i/\Delta - \lfloor 2\hat{I}_i/\Delta \rfloor > 0.5 \end{cases}. \quad (5)$$

The embedding and extracting steps with possible IPDs distortion are presented in Figure 2.

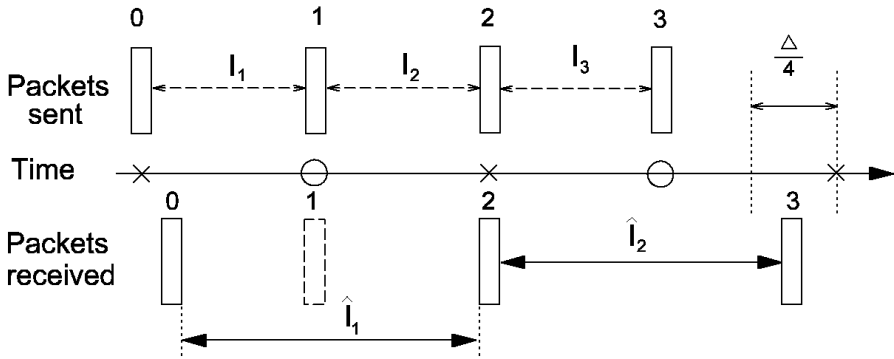


Fig. 2. An example of IPDs distortion caused by jitters

As it was discussed, the scheme in Figure 1 may be regarded as a communication channel with two types of errors: substitutions and deletions. The substitution error refers to bit flips due to network jitters or packet deletions that result in merger of two IPDs. Since during QIM demodulation we map each IPD to its closest quantizer, any jitter over $\Delta/4$ would possibly result in a substitution error (see Figure 2). The channel model developed in [2] handles the dependent substitution and deletion errors. However, to simplify the decoding we assume that the dependence exists only inside the received codeword, which is a reasonable limitation, as soon as any number of packet deletions results only in the presence or in the absence of a substitution error.

For example, in Figure 2 four packets 0, 1, 2, 3 are sent, three packets 0, 2, 3 are received and packet 1 is lost. The first two IPDs I_1 and I_2 are transformed into \hat{I}_1 and the size of last IPD I_2 is changed and evaluated as \hat{I}_2 . Hence the result of channel noise is the bit received before Packet 2 that is the merged of the two intervals $\hat{s}_1 = s_1 \oplus s_2$ and the bit flipped after receiving Packet 3 resulting in $\hat{s}_2 = \bar{s}_3$. In general $\hat{s}_i = \sum_{j=r+1}^i s_j$ and can take only 0 or 1 binary values, where r is the index of the last successfully received packet before i -th one.

Without loss of generality, we consider the packet deletion probability P_d and the packet substitution probability P_s to be identical for all packets and assume that Packet 0 is always synchronized. This assumption can be easily performed by the use of so-called frame synchronization based on special markers [6, 9] or on one or more codewords received without errors. This allows the scheme to be in the synchronized state before the decoding procedure and further evaluate the distance between w and of \hat{w} to decide whether the watermark is present. Considering that EC-VTC decoder is synchronized prior to decoding of a received sequence \hat{s} we describe its operation principles below.

3.3 EC-VTC Decoder

The original watermark $w=w_1w_2\dots w_N$ is a sequence of bits with each element from $GF(2)$. This sequence is divided into blocks $\mathbf{b}=(b_1\dots b_l)$ to produce the VT-codewords \mathbf{x} of length n by VTC Encoder. Then a codeword \mathbf{x} is concatenated with predefined marker pattern $\mathbf{z}=z_1z_2\dots z_m$ of length m making \mathbf{x}_m and xored with pseudo-random key sequence k_w , forming a sequence s , as depicted in Fig.1. The used key k_w is a sparse sequence containing a binary 1 only in one position of block with length $n+m$ and is applied for security and frame synchronization. Actually sequence s is made from concatenation of N codewords \mathbf{x}_m , has length $M=(n+m)N$ and is embedded in flow IPDs. I^w is transmitted and after transversing the network is received in the form of estimated sequence \hat{I} and demodulated. The result sequence \hat{s} is xored with key sequence k_w , separated into codewords y_m by marker detection [9] and further converted into VT-codewords \mathbf{y} containing possible substitution or/and deletion errors. The EC-VTC Decoder performs the error-correcting decoding using one of two algorithms, depending on the number of errors in \mathbf{y} occurred. The decision about the decoder type to be applied is based on the estimation of \mathbf{y} length. If the only one deletion is found, the Levenshtein's decoding algorithm [8] is used, and if the number of deletion errors is greater than one, the maximum likelihood decoding (MLD) maximizing $\Pr(\mathbf{x}^*=\mathbf{x}/\mathbf{y})$ is applied. The bounded distance decoding with the use of syndrome calculation is performed in case of absence of deletion errors or after their correction as well in process of frame synchronization.

For example, consider that a key generator outputs the sequence of digits $g=03\dots$ and a key sequence is made according to the expression $k_w=g \bmod (n+1)$. The received and extracted by demodulator sequence $\hat{s}=110100000.11110000$ is then xored with key sequence $k_w=000000000.001000000$ and decoded with the use of

mapping the subcode $C^r = \{(110100), (110011), (011110), (101101)\}$ to data blocks $\mathbf{b} = \{00, 01, 10, 11\}$. Hence, after marker detection, 2 codewords are obtained $\mathbf{y}_1 = 110100$, $\mathbf{y}_2 = 11010$. The calculation of (1) gives $S=0$ that can be used to indicate about the correct boundaries of received error-free codeword \mathbf{y}_1 , infer the presence of synchronization and allow to start decoding by applying the Levenshtein's algorithm for the deleted zero bit in the last position of \mathbf{y}_2 . However, if the second bit is also deleted and $\mathbf{y}_2 = 1101$, the evaluation of its length results in the use of MLD decoding. The use of VT-code as a subcode of any linear code preserves its minimum distance d_{min} and does not change code performance in channel with substitution errors which is well analyzed in [5]. However, the reduction in probability of codeword error when using MLD strictly depends on subcode used and on weight distribution $w_t(\mathbf{x}_i)$ of its codewords.

4 Performance Evaluation

In this section we evaluate the robustness to packet losses. The proposed watermarking scheme has been evaluated by simulation of packets, generated from independent Poisson process of rate 3 packets per second and length of about 4000 packets with shifted mean of 25 ms and standard deviation of 10 ms. Network jitters was simulated as Laplace distribution with zero mean and the same deviation. The pseudorandom bits of watermarks have been encoded by subcode C^r with added uniform marker $z=000$ and randomly embedded into 3600 flows with the use QIM modulation (4). The watermark parameters were taken similar to the values from [2] to get the approximately the same number of watermark bits as $N=50$, $n=9$, $M=450$. Note, that the block length of EC-VTC [6,3,3] with appended z marker bits results in $n=9$ and close to the sparsified version [2] of watermark. The watermark extraction was made with the use of QIM demodulation function (5) and the decoding was performed with the use of Levenshtein's, MLD and syndrome decoders.

The evaluation of the proposed scheme against packet deletions by considering the varying packet deletion probabilities $P_d = \{0.01, 0.02, 0.03, 0.1, 0.2\}$ has been done. The watermarks were randomly embedded into 3600 flows. Also the other 3660 unmarked flows were used to obtain the false positive rates. The detection threshold was chosen so that the false positive rate was kept below 1% for all deletion probabilities. True Positive (TP) detection rates for deletion ratios P_d gave corresponding values: 1% - 0.9999, 2% - 0.9998, 3% - 0.9998, 10% - 0.9951; 20% - 0.6655.

We see that the detector has rather high true positive rates, maintaining true positive rate (TP) up to 99%, even when less than 10% of packets were deleted. However the value of TP drops to 66% when packet deletion ratio is at 20%, which is rare in a network environment. Thus, in comparison with the other IPD-based watermarking schemes [1], [10] which suffer from desynchronization, the proposed scheme is robust against packet losses and network jitters presented as deletion and substitution errors. The use of pseudo-random key sequence on transmission side improves the security of overall scheme and provides the frame synchronization in watermarking system. To examine the visibility of proposed scheme, the Kolmogorov-Smirnov (K-S) has been performed on 4500 watermarked flows against

4500 unwatermarked flows and demonstrated the statistical invisibility of watermark according to the values of K-S distances that are below 0.03. Obviously, to defeat the multi-flow attack, as suggested in [1] the use of random function position of embedding positions can selected within the described above synchronization method.

5 Conclusion

An invisible flow watermarking scheme based on linear error correcting codes for channels with substitution and deletion errors, representing network jitter and packet drops, has been developed. The described scheme is based on relatively low-rate linear code, formed on the basis of proposed algorithm to create a linear error-correcting code that is a subcode of VT-code. Statistical and computational experiments demonstrate that proposed scheme is of similar to [2] performance, but has a much lower complexity, as soon as it uses a simpler implementation mainly based on linear decoding operations with much less space and time complexity and only perform MLD with the use of look-up tables when the packet loss increases significantly.

References

1. Kiyavash, N., Houmansadr, A., Borisov, N.: Multi-flow attacks against network flow watermarking schemes. In: USENIX Security Symposium, pp. 307–320 (2008)
2. Gong, X., Rodrigues, M., Kiyavash, N.: Invisible flow watermarks for channels with dependent substitution and deletion errors. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Proc., Kyoto, Japan, March 25-30, pp. 1773–1776 (2012)
3. Chen, B., Wornell, G.W.: Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inf. Th.* 47, 1423–1443 (2001)
4. Tenengol'ts, G.M.: Class of codes correcting bit loss and errors in the preceding bit. *Avtomat. Telemekh.* 37(5), 797–802 (1976)
5. Sklar, B.: *Digital Communications: Fundamentals and Applications*, 2nd edn. Prentice-Hall (2001, 2003)
6. Abdel-Ghaffar, K.A.S., Ferreira, H.C., Cheng, L.: Correcting deletions using linear and cyclic codes. *IEEE Trans. Inf. Th.* 56(10), 5223–5234 (2010)
7. Varshamov, R.P., Tenengol'ts, G.M.: Correction code for single asymmetric errors. *Avtomat. Telemekh.* 26(2), 286–290 (1965)
8. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklady* 10(8), 707–710 (1966)
9. Chen, J., Mitzenmacher, M., Ng, C., Varnica, N.: Concatenated codes for deletion channels. In: Proceedings of the 2003 IEEE International Symposium on Information Theory, Yokohama, Japan, June 29-July 4, p. 218 (2003)
10. Houmansadr, A., Kiyavash, N., Borisov, N.: RAINBOW: A Robust and Invisible Non-Blind watermark for network flows. In: Proceedings of the 16th Annual Network & Distributed System Security Symposium, San Diego, USA, February 8-11 (2009)

Detecting Resized Double JPEG Compressed Images – Using Support Vector Machine

Hieu Cuong Nguyen and Stefan Katzenbeisser

Computer Science Department, Darmstadt University of Technology, Germany
{cuong, katzenbeisser}@seceng.informatik.tu-darmstadt.de

Abstract. Since JPEG is the most widely used compression standard, detection of forgeries in JPEG images is necessary. In order to create a forged JPEG image, the image is usually loaded into a photo editing software, manipulated and then re-saved as JPEG. This yields to double JPEG compression artifacts, which can possibly reveal the forgery. Many techniques for the detection of double JPEG compressed images have been proposed. However, when the image is resized before the second compression step, the blocking artifacts of the first JPEG compression are destroyed. Therefore, most reported techniques for detecting double JPEG compression do not work for this case. In this paper, we propose a technique for detecting resized double JPEG compressed (called RD-JPEG) images. We first identify features that can discriminate RD-JPEG images from JPEG images and then use Support Vector Machines (SVM) as a classification tool. Experiments with many RD-JPEG images with different quality and scaling factors indicate that our technique works well.

Keywords: SVM, classification, image forensics, re-sampling.

1 Introduction

Due to the large number of available image processing tools, digital images can easily be altered without leaving visual evidence. Therefore, developing techniques for judging the authenticity of digital images became an urgent need. There are many types of image forgeries, which can be detected by different image forensic methods [1]. Since JPEG is the most popular image type and it is supported by many applications, it is worthwhile to develop forensic techniques for JPEG images.

Although there are many ways of making forgeries in a JPEG image, most share three main steps: 1) loading the JPEG image which is compressed by quality factor QF_1 to a photo editing software, 2) manipulating this image and 3) re-compressing it as a JPEG file with quality factor QF_2 . Consequently, the forged image is doubly JPEG compressed (called D-JPEG). Detecting artifacts of double JPEG compression is an important step to judge whether a JPEG image is authentic. To this end, several techniques have been developed [2–8]. The authors in [2, 3] found that when QF_1 is different from QF_2 , periodic artifacts are present in the histograms of the DCT coefficients of D-JPEG images. The periodicity can be recognized in the Fourier transform through peaks in the spectrum. Lin et al. [4] expanded the global approach of [3] by

locating the tampered regions in D-JPEG images. Bianchi et al. [5] proposed an enhanced version of [4], leading to an improvement of the accuracy of the algorithm. The authors in [6, 7] showed that the distribution of the most significant digit of the DCT coefficients in JPEG images follows the generalized Benford distribution. This distribution is very sensitive to double JPEG compression and this property can be applied to detect D-JPEG images. Chen et al. [8] proposed a set of image features, which have subsequently been evaluated by a SVM based classifier.

A limitation of these techniques is that they cannot detect D-JPEG images if the JPEG images are cropped before the second compression step is applied. The reason is that the corresponding blocking grids in the first compression and in the second compression are no longer aligned. In order to overcome this limitation, some other techniques have been proposed [9–11]. In [9] a blocking artifact characteristic matrix (BACM) is computed to measure the symmetric representation of the blocking artifacts introduced by JPEG compression. Since the symmetry of the BACM of a JPEG image is destroyed after the image is cropped, the BACM can be used as evidence for detecting cropped double JPEG compressed images. The authors of [10] model the linear dependency of the “within-block” pixels (pixels that are not on the border of segmented 8×8 image blocks), compute the probability of the pixel being linearly correlated to its neighbors and form the map of the probabilities of all pixels in the image. The map is converted to Fourier domain and several statistical features from the different peak energy distribution are extracted in order to discriminate cropped D-JPEG images from non-cropped D-JPEG images. A simple yet reliable technique to detect the presence of cropped double JPEG compression has been introduced in [11]. This technique is based on the observation that the DCT coefficients exhibit an integer periodicity when they are computed according to the grids of the primary JPEG compression.

Although [9–11] work well for detecting cropped double JPEG compressed images, they are defeated if the images are resized before the second compression. The reason is that due to the effect of re-sampling, the blocking artifacts will be broken. The authors of [12] demonstrated the influence of resizing on the detection results of [7, 8]. To the best of our knowledge, there are only a few techniques for detecting resized double JPEG compressed (RD-JPEG) images [13–15]. Kirchner and Gloe [13] apply a re-sampling detection technique (which was originally designed to work with uncompressed images) to JPEG images and analyze how the JPEG compression affects the detection output. A limitation of [13] is that the detection rates when applied to RD-JPEG images are very low if QF_1 is much larger than QF_2 . Besides, if the JPEG images are down-sampled before the second compression, the technique is mostly defeated. The technique [14] extracts neighboring joint density features and applies SVM to them. Although this technique works for both up-sampled images and down-sampled images by different interpolation methods, it is analyzed by the authors only for quality factors (both QF_1 and QF_2) of 75 and no information on false positives is given. Bianchi and Piva [15] proposed an algorithm, which can be summarized by some steps: 1) estimate the candidate resizing factor; 2) for each candidate factor, undo the image resizing operation and measure the NLDP (near lattice distribution property); 3) if the result is greater than a predefined threshold, label the image

as resized double JPEG compressed. Furthermore, the technique [15] can estimate both the resize factor and the quality factor of the first JPEG compression of the analyzed image. The experimental results in [15] show that it surpasses [13] in the same test condition, but similar to [13], it seems more difficult to detect when QF_1 is much larger than QF_2 .

In this paper, we propose a new technique to detect RD-JPEG images. The technique first reveals specific features of JPEG images by using a re-sampling detector. These features are subsequently fed to SVM-based classifiers in order to discriminate RD-JPEG images from JPEG images. In comparison to [13], our technique does not require to distinguish in detail the peaks caused by JPEG compression from the peaks caused by re-sampling. In comparison to [14], our approach does not need to extract complex image features for classification. The technique [15] consists of some intricate steps, which mostly use for the purpose of reverse engineering of resized double JPEG compressed images.

In Section 2, we briefly introduce state-of-the-art re-sampling detection techniques. Re-sampling detection is an important step of our technique and any of the mentioned re-sampling detectors can be used in our construction. The proposed detection algorithm for RD-JPEG images is explained in Section 3 and experimental results are shown in Section 4. Lastly, the paper is concluded in Section 5.

2 Techniques for Image Re-sampling Detection

To create a convincing forged image, the geometry of the image or some portions of it is often transformed. Once a geometric transformation (such as resizing or rotation) is applied to an image, a re-sampling process is involved. Interpolation is the central step of re-sampling in order to estimate the value of the image signal at intermediate positions to the original samples. Based on specific artifacts created by interpolation, there are several techniques to detect traces of re-sampling in digital images.

Gallagher [16] realized that low-order interpolated signals introduce periodicity in the variance function of their second derivatives. Based on this observation, the author proposed a technique to detect whether an image has been re-sampled. A limitation of this technique is that it works only in the case of image resizing. Using the Radon transform, Mahdian and Saic [17] improved [16] so that their technique can detect not only image resizing, but also image rotation. Popescu [18] noted that there are linear dependencies between neighboring pixels in re-sampled images. These correlations can be determined by using the Expectation/Maximization (EM) algorithm. The output of the algorithm is a matrix indicating the probability of every image sample being correlated to its neighbors (called p-map). The p-map of a re-sampled image usually contains periodic patterns, which are visible in the Fourier domain.

A drawback of [18] is that its computational complexity due to the use of the EM algorithm. Based on [18] some improved techniques have been introduced in [19] and [20]. The author in [19] showed that the p-map of a re-sampled image is periodic and this periodicity does not depend on the prediction weights that are used to compute the correlations of neighboring pixels. Therefore, he used a predefined set of

prediction weights to compute the correlation probability of every image sample and designed a fast re-sampling detection technique which bypasses the EM algorithm.

Although the values of prediction weights do not affect the periodicity of the p-map in theory, the authors in [20] found that the selected set of predefined weights can strongly affect the obtained results: using one predefined set of weights for detection, peaks can be recognized in the transformed p-map, but using another set, peaks are not evident (though the periodicity exists in theory). Therefore, they use a predefined set, which is chosen through experimentation and apply the Radon transformation to the probability map of the analyzed image in order to enhance the frequency peaks and consequently the robustness of the overall technique. An example for detection of (uncompressed) re-sampled images by using [20] is presented in Fig. 1.

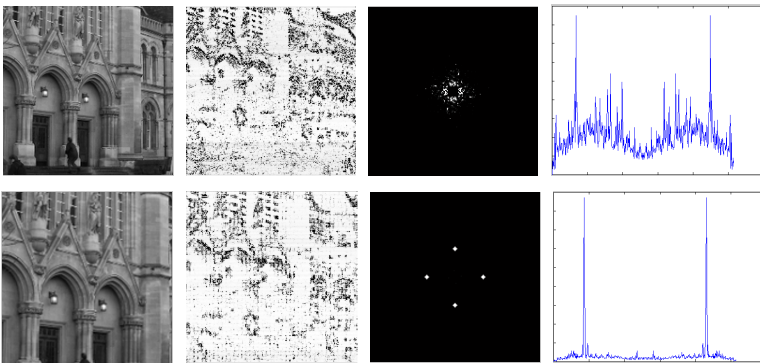


Fig. 1. Shown in the top row is the original image and shown in the bottom row is the re-sampled image by a factor of 1.2. Shown in the left most column are the original image and the re-sampled image. Shown in the middle columns are the p-maps and the magnitudes of the Fourier transforms of the p-maps. Shown in the right most column are the Fourier transformations of the Radon transforms of the p-maps.

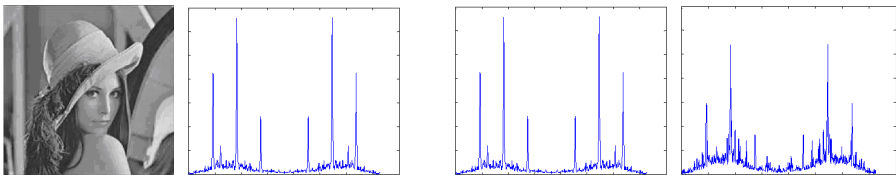


Fig. 2. Shown in the left is a JPEG image of Lena and in the right is the detection result by using a re-sampling detection technique

Fig. 3. Shown in the left is the detection result of the JPEG image of Lena and in the right the detection result of the RD-JPEG version of the same image

The techniques [16–20] work well for detecting traces of re-sampling in uncompressed images. However, they fail when applied to JPEG images. The reason is that JPEG compression has an effect similar to nearest neighbors interpolation and the re-sampling detector will get confused [16]. An example can be seen in Fig. 2, which

shows a JPEG image that has not been re-sampled, yet the spectrum when applying the re-sampling detector [17] contains strong peaks.

In the next section, we propose a technique which uses one of these re-sampling detectors as the first step for detecting RD-JPEG images. Although any mentioned re-sampling detection technique in this section can be used, we choose [17] because of its efficiency as well as its speed.

3 Proposed Technique for Detecting RD-JPEG Images

When using [17] to detect re-sampling in both JPEG images and RD-JPEG images, we empirically found that the detection results of RD-JPEG images seem to have more peaks than those of JPEG images. An example is shown in Fig. 3. This is because the detection result of a RD-JPEG image contains not only the peaks introduced by JPEG compression, but also the peaks due to re-sampling. Nevertheless, the difference is not always easy to recognize by human eyes. Besides, it is necessary to automatically classify RD-JPEG images from JPEG images. To this end, we first apply the technique [17] to JPEG images, and then extract the values of maximal peaks from the normalized Fourier spectrum. The extracted features are subsequently fed to SVM-based classifiers in order to discriminate RD-JPEG images from JPEG images. Since SVM is only a binary classifier, we use two different approaches to design SVM classifiers for detecting RD-JPEG images.

In the first approach, we design a single SVM classifier for directly distinguishing JPEG and RD-JPEG images, compressed by different quality factors. To this end, the features of a set of JPEG images and their re-sampled versions (the number of JPEG and re-sampled JPEG images are the same) are extracted for training a SVM classifier. This approach is simple and suitable for many situations in practice when we do not know the quality factors of the analyzed images. However, through experiments, reported in Section 4, we find that this technique works well mostly when QF_1 is lower than the QF_2 .

The second approach is based on the idea that while QF_1 of a double JPEG compressed image is usually not known to the analyst, QF_2 can reliably be computed from the bitstream of the JPEG image (see Appendix A). Thus, instead of using one single classifier for all quality factors, we design several different SVM classifiers, each of which distinguishes JPEG and RD-JPEG images for one specific value of QF_2 . Once the last quality factor of an analyzed JPEG image is known, the corresponding classifier will be applied to it. The method to design a classifier for a particular QF_2 is similar to the first approach: we first use a set of JPEG images and another set of RD-JPEG images (the numbers of images in both sets are the same and every image is compressed by QF_2) and then extract image features for training. In other words, the last quality factor of a tested image is first identified, and then the image will be analyzed by the corresponding classifier. In next section, we discuss experimental results for both approaches.

4 Experimental Results

First, we randomly choose 200 uncompressed images from the UCID image database [21]. We create 5 datasets of JPEG images by compressing the uncompressed images with the quality factors of 40, 50, 60, 70, and 80. The JPEG images are subsequently resized by a scaling factor of 1.2 and recompressed by different factors of 40, 50, 60, 70, and 80. As a result, we obtained 5 datasets of 1000 RD-JPEG images corresponding to each dataset of JPEG images.

To test the first approach, we create a single SVM classifier by using two groups of JPEG images and RD-JPEG images (with the scaling factor of 1.2) for training. After the training process (presented in Section 3) we apply the classifier to test RD-JPEG images. In training, we consider two cases of different quality factors: 1) 100 JPEG images compressed by a quality factor of 50 and 100 RD-JPEG images re-compressed by a quality factor of 70 ($QF_1=50$, $QF_2=70$ and scaling factor =1.2) and 2) 100 JPEG images compressed by a quality factor of 70 and 100 RD-JPEG images re-compressed by a quality factor of 80 ($QF_1=70$, $QF_2=80$ and scaling factor =1.2). Analyzing the detection results (see Table 1 and Table 2), we found that the technique works well for detecting RD-JPEG images where QF_1 is smaller than QF_2 . Otherwise, when QF_1 is larger than QF_2 , the detection rate is reduced. In our experiments, the false positive rates (computed by testing the classifier on datasets of JPEG images which have been compressed by different quality factors of 40, 50, 60, 70, and 80) are lower than 10% in the first case and lower than 8% in the second case.

In a more realistic scenario, we test the techniques on the RD-JPEG images, which have been resized with a different scaling factor than the factors are used in the training process. The datasets are created in the same way as above, except the scaling factor 1.1 is used instead of 1.2 (i.e. $QF_1=70$, $QF_2=80$ and scaling factor =1.1). Although the detection results (in Table 3) are clearly worse compare with Table 1 and Table 2, we found that the degradation is not significant; therefore, the technique can potentially work in case the scaling factor is unknown.

In the second approach, we consider 5 different cases corresponding to a QF_2 of 40, 50, 60, 70, and 80. The case of $QF_2=40$, we organize the training images into two groups: a group of 100 JPEG images (the quality factor of 40) and the other group of 100 RD-JPEG images ($QF_1=50$, $QF_2=40$ and scaling factor=1.2). The extracted features are used to train a SVM classifier that can be used to detect RD-JPEG images which compressed by the QF_2 of 40. We repeat this process for the other cases when QF_2 is 50, 60, 70, and 80. The detection results in testing RD-JPEG datasets are presented in Table 4. We noticed that following the second approach, the technique works well even if QF_1 is larger than QF_2 . The false positive rates are lower than 10% (9%, 8%, 5%, 6% and 3% when testing JPEG images compressed by the quality factors of 40, 50, 60, 70, and 80 respectively). Since JPEG compression with a lower factor produces stronger peaks in the Fourier spectrum, it obtains higher false positives.

Table 1. Detection results using a single SVM classifier (training JPEG images compressed by $QF=50$ and RD-JPEG images re-compressed by $QF_1=50, QF_2=70$) for RD-JPEG images by the scaling factor of 1.2 and by different quality factors (QF_1 in rows and QF_2 in columns)

	40	50	60	70	80
40	65.5%	91.0%	99.5%	99.5%	84.5%
50	52.5%	80.0%	97.0%	99.0%	87.0%
60	35.5%	77.5%	92.5%	98.5%	88.0%
70	19.5%	67.5%	87.0%	99.0%	84.0%
80	10.5%	45.0%	79.5%	91.5%	78.0%

Table 2. Detection results using a single SVM classifier (training JPEG images compressed by $QF=70$ and RD-JPEG images re-compressed by $QF_1=70, QF_2=80$) for RD-JPEG images by the scaling factor of 1.2 and by different quality factors (QF_1 in rows and QF_2 in columns)

	40	50	60	70	80
40	70.0%	94.0%	98.5%	99.0%	95.0%
50	62.0%	80.0%	92.5%	98.5%	98.0%
60	48.0%	76.0%	87.5%	96.5%	99.0%
70	33.5%	68.0%	83.0%	93.5%	99.0%
80	24.0%	57.0%	69.0%	81.0%	92.0%

Table 3. Detection results using a single SVM classifier (training JPEG images compressed by $QF=70$ and RD-JPEG images re-compressed by $QF_1=70, QF_2=80$) for RD-JPEG images by the scaling factor of 1.1 and by different quality factors (QF_1 in rows and QF_2 in columns)

	40	50	60	70	80
40	37.0%	57.0%	63.5%	78.0%	82.5%
50	37.0%	58.0%	63.5%	78.5%	83.0%
60	26.0%	48.0%	66.5%	77.5%	87.0%
70	13.5%	43.5%	68.0%	77.5%	73.0%
80	10.5%	39.0%	62.5%	77.0%	86.5%

Table 4. Detection results using dedicated SVM classifiers for RD-JPEG images (depending on the quality factor of the second compression) by the scaling factor of 1.2 and by different quality factors (QF_1 in rows and QF_2 in columns)

	40	50	60	70	80
40	95.0%	91.5%	89.5%	99.0%	98.0%
50	90.0%	90.0%	88.5%	98.5%	99.5%
60	89.5%	91.0%	97.5%	98.0%	100%
70	87.5%	85.0%	95.0%	99.5%	98.0%
80	85.0%	80.0%	96.0%	100%	99.0%

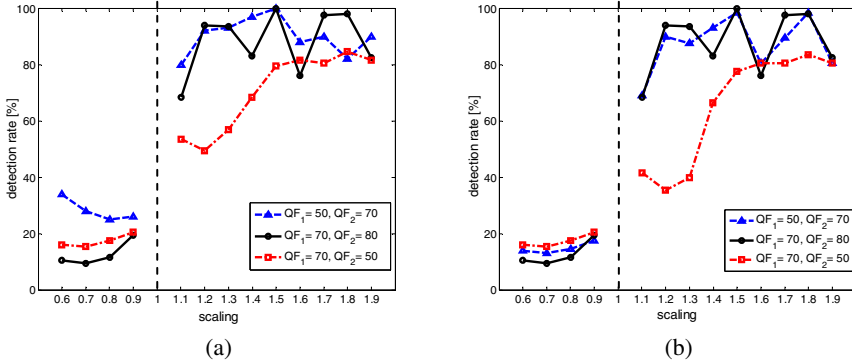


Fig. 4. Detection results for RD-JPEG images by different scaling factors: (a) the quality factors of the trained images and the test images are the same, (b) the quality factors of the trained images and the test images are different

In order to assess the influence of scaling factor, we test the proposed technique for detection of RD-JPEG images with various scaling factors. The RD-JPEG images are created by resizing JPEG images (firstly compressed by QF_1) of different scaling factors (from 0.6 to 1.9) and then they are recompressed (by a different quality factor QF_2). We consider three cases: 1) $QF_1=50$ and $QF_2=70$, 2) $QF_1=70$ and $QF_2=80$ and 3) $QF_1=70$ and $QF_2=50$. We create different datasets of JPEG images and RD-JPEG images and in each case, the training and testing processes of the classifiers are conducted as described before. The detection results in various scaling factors are shown in Fig. 4a. Due to missing information in the down-sampling process, the detection rates of the down-sampled images are very low. Detecting up-sampled images is possible with much higher rates. In some cases, the detection rates even reach about 100%. In this scenario, the test images are compressed with the same quality factors as the training images (but with different scaling factor). We found that scaling factors affect the detection results: typically the detection rates tend to increase.

Lastly, in a more realistic scenario, we apply the technique trained by one image type ($QF_1=70, QF_2=80$, scaling factor = 1.2) to images with different types ($QF_1=50$ and $QF_2=70, QF_1=70$ and $QF_2=50$, and scaling factor ranges from 0.6 to 1.9). The detection results are presented in Fig. 4b. Although the results deteriorate (compare with Fig. 4a), we found that the degradation is not significant; therefore, the technique can potentially work in a real condition.

5 Conclusion

In this paper, we designed a technique for detecting resized double JPEG compressed images. The technique is based on applying a re-sampling detector to JPEG images, and extracting features from strong peaks of the normalized Fourier transformation. Then the extracted features are fed into a SVM-based classifier in order to discriminate RD-JPEG images from JPEG images. We propose two methods to design SVM classifiers: one single global classifier and several classifiers depending on the quality

factor of the last compression. Although the first approach is simple and easy to use, the second approach achieves higher detection rates. In comparison with some existing techniques our technique has higher detection rates when the quality factor of the first compression is larger than the quality factor of the last compression and when detecting down-sampled images. We apply the technique to test RD-JPEG images resized with different scaling factors and found that the scaling factors can affect the detection results. In future, we will apply the technique for the detection of rotated double JPEG compressed images and use other re-sampling detectors in our technique so that we can compare their efficiency in the detector of RD-JPEG images.

References

1. Farid, H.: Image forgery detection. *IEEE Signal Processing Magazine* 26, 16–25 (2009)
2. Lukáš, J., Fridrich, J.: Estimation of Primary Quantization Matrix in Double Compressed JPEG Images. In: *Proc. Digital Forensic Research Workshop* (2003)
3. Popescu, A.: *Statistical Tools for Digital Image Forensics*. PhD Thesis (2004)
4. Lin, Z., He, J., Tang, X., Tang, C.-K.: Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis. *Pattern Recognition* 42, 2492–2501 (2009)
5. Bianchi, T., De Rosa, A., Piva, A.: Improved DCT coefficient analysis for forgery localization in JPEG images. In: *ICASSP 2011*, pp. 2444–2447 (2011)
6. Fu, D., Shi, Y.Q., Su, W.: A generalized Benford's law for JPEG coefficients and its applications in image forensics. In: *Proceedings of SPIE*, pp. 65051L–65051L-11 (2007)
7. Li, B., Shi, Y.Q., Huang, J.: Detecting doubly compressed JPEG images by using Mode Based First Digit Features. In: *2008 IEEE 10th Workshop on Multimedia Signal Processing*, pp. 730–735 (2008)
8. Chen, C., Shi, Y.Q., Su, W.: A machine learning based scheme for double JPEG compression detection. In: *2008 19th International Conference on Pattern Recognition*, pp. 1–4 (2008)
9. Luo, W., Qu, Z., Huang, J., Qiu, G.: A Novel Method for Detecting Cropped and Recompressed Image Block. In: *IEEE ICASSP 2007*, pp. 217–220 (2007)
10. Chen, Y., Hsu, C.: Image Tampering Detection by Blocking Periodicity Analysis in JPEG Compressed Images. In: *MMSP 2008*, pp. 803–808 (2008)
11. Bianchi, T., Piva, A.: Detection of Nonaligned Double JPEG Compression Based on Integer Periodicity Maps. *IEEE Trans. on Information Forensics and Security* 7, 842–848 (2012)
12. Sutthiwan, P., Shi, Y.Q.: Anti-Forensics of Double JPEG Compression Detection. In: Shi, Y.Q., Kim, H.-J., Perez-Gonzalez, F. (eds.) *IWDW 2011*. LNCS, vol. 7128, pp. 411–424. Springer, Heidelberg (2012)
13. Kirchner, M., Gloe, T.: On resampling detection in re-compressed images. In: *WIFS*, pp. 21–25 (2009)
14. Liu, Q., Sung, A.H.: A new approach for JPEG resize and image splicing detection. In: *Proceedings of the First ACM Workshop on Multimedia in Forensics, MiFor 2009*, p. 43 (2009)
15. Bianchi, T., Piva, A.: Reverse engineering of double JPEG compression in the presence of image resizing. In: *2012 IEEE International Workshop on Information Forensics and Security, WIFS*, pp. 127–132 (2012)
16. Gallagher, A.C.: Detection of Linear and Cubic Interpolation in JPEG Compressed Images. In: *The 2nd Canadian Conference on Computer and Robot Vision, CRV 2005*, pp. 65–72 (2005)
17. Mahdian, B., Saic, S.: Blind Authentication Using Periodic Properties of Interpolation. *IEEE Transactions on Information Forensics and Security* 3, 529–538 (2008)

18. Popescu, A.C., Farid, H.: Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on Signal Processing* 53, 758–767 (2005)
19. Kirchner, M.: Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue. In: *Proceedings of the 10th ACM Workshop on Multimedia and Security, MM&Sec 2008* (2008)
20. Nguyen, H.C., Katzenbeisser, S.: Robust Resampling Detection in Digital Images. In: De Decker, B., Chadwick, D.W. (eds.) *CMS 2012. LNCS*, vol. 7394, pp. 3–15. Springer, Heidelberg (2012)
21. Schaefer, G., Stich, M.: UCID: an uncompressed color image database. In: *Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia, San Jose, USA*, pp. 472–480 (2004)
22. Kornblum, J.: Using JPEG quantization tables to identify imagery processed by software. *Digital Investigation* 5, S21–S25 (2008)
23. Chandra, S., Ellis, C.S.: JPEG Compression Metric as a Quality Aware Image Transcoding. In: *Proceedings of USITS 1999* (1999)
24. Sallee, P.: *Matlab JPEG Toolbox*, <http://www.philsallee.com/jpegtbx/index.html>

Appendix A: Determining the Last Quality Factor of a JPEG Images

The compression ratios of JPEG images are controlled by the quantization tables which used in the compression process. In this paper, we focus on images stored in the JPEG Interchange File Format (JFIF). The JFIF is the most commonly used format for JPEG data [22]. The quantization table that was used to compress an image is stored in the JFIF header [23]. This table (called T_s) can be identified by using the JPEG Toolbox [24].

The most commonly used standard quantization tables are published by the International JPEG Group (IJG). Based on the standard table (T_b), and the quality factor (Q), the quantization table can be computed as follows:

$$S = \begin{cases} \frac{500}{Q} & \text{if } Q < 50 \\ 200 - 2Q & \text{otherwise} \end{cases}, \quad T_s[i] = \left\lfloor \frac{S * T_b[i] + 50}{100} \right\rfloor.$$

Conversely, when the tables T_b and T_s are known, the approximate value of the quality factor can be computed as follows [23]:

$$S' = \frac{T_s[i] * 100 - 50}{T_b[i]}, \quad Q' = \begin{cases} \left\lfloor \frac{200 - S'}{2} \right\rfloor & \text{if } S' \leq 100 \\ \left\lfloor \frac{5000}{S'} \right\rfloor & \text{otherwise} \end{cases}.$$

Note that the function to predict the quality factor involves integer computation on the quantization table (T_s) that introduce integer rounding errors, so the value of Q' is close to Q . Following a suggestion in [23], then the computed quality factor (Q') should be off by one or two.

Pit Stop for an Audio Steganography Algorithm

Andreas Westfeld¹, Jürgen Wurzer², Christian Fabian², and Ernst Piller²

¹ Dresden University of Applied Sciences, Germany

² St. Pölten University of Applied Sciences, Austria

Abstract. Steganography plays an important role in the field of secret communication. The security of such communication lies in the impossibility of proving that secret communication is taking place.

We evaluate the implementation of a previously published spread spectrum technique for steganography in auditive media. We have unveiled and solved several weaknesses that compromise undetectability.

The spread-spectrum approach of the technique under evaluation is rather unusual for steganography and makes the secret message fit to survive A/D and D/A conversions of analogue audio telephony, re-encoded speech channels of GSM/UMTS, or VoIP. Its impact to signal statistics, which is at least concealed by the lossy channel, is reduced. There is little published on robust audio steganography, its steganalysis, and evaluation, with the possible exception of audio watermarking, where undetectability is not as important.

Keywords: information hiding, steganalysis, spread spectrum BPSK, VoIP steganography.

1 Introduction

Steganography is the art and science of invisible communication. Its aim is the transmission of information embedded invisibly into cover data. Secure watermarking methods embed short messages protected against modifying attackers (robustness, watermarking security) while the existence of steganographically embedded information cannot be proven by a third party (indiscernibility, steganographic security).

In general, steganographic communication uses an error-free channel, hence messages are received unmodified. Digitised image or audio files reach the recipient virtually without errors when sent, e.g., as an e-mail attachment. The data link layer ensures a safe, i.e., mostly error-free, transmission. If every bit of the cover medium is received straight from the source, then the recipient can extract a possibly embedded message without any problem. However, analogue audio telephony with A/D and D/A conversions, re-encoded speech channels of GSM/UMTS, and VoIP telephony use lossy compression or even do without a data link layer. This is because emerging errors have little influence on the (auditive) quality and can therefore be tolerated.

Without error correction, distortions are acceptable only in irrelevant parts of the cover signal. However, typical steganographic methods prefer these locations

for hiding payload. The hidden message would experience the most interference in error-prone channels. Therefore, robust embedding functions have to add redundancy and change only locations that are carefully selected w.r.t. the proportion between unobtrusiveness and probability of error. This increases the risk of detection and permits only a small payload.

Information hiding techniques can be described in the classical triangle, i.e., a set of three characteristics: capacity, robustness, and undetectability. There are highly robust watermarking methods that offer small capacities and achieve perceptual transparency. Some watermarking methods are even robust against distortions in the time and frequency domains. Tachibana et al. introduced an algorithm that embeds a watermark by changing the power difference between the consecutive DFT frames [1]. It embeds 64 bits in a 30-second music sample. Compared to the proposed steganographic method this is a quarter of the payload in a host signal (cover) occupying 50 times the bandwidth. It is robust against radio transmission. However, it was not designed to be steganographically secure and the presence of a watermark is likely to be detected by calculating the statistics of the power difference without knowing the pseudo random pattern. Van der Veen et al. published an audio watermarking technology that survives air transmission on an acoustical path and numerous other robustness tests while being perceptually transparent [2]. The algorithm of Kirovski and Malvar [3] embeds about 1 bit per second (half as much as the one in [1]) and is even more robust (against the StirMark Benchmark [4]). Arnold et al. presented an adaptive spread phase modulation (ASPM) that embeds an inaudible watermark with good robustness [5]. Although watermarking algorithms are perceptually transparent, they are not intended to be steganographically secure.

Examples for robust steganography are rather rare and, in most cases, embed into images. Marvel et al. [6] developed a robust steganographic method for images based on spread spectrum modulation [7]. This technique enables the transmission of information below the noise or cover signal level (signal to noise ratio below 0 dB). Likewise it is difficult to jam, as long as transmitter and receiver are synchronised. Therefore, successful attacks de-synchronise the modulated signal [8]. Further examples robustly embed messages using DSSS in slow scan television signals [9] or in auditive media.

This paper evaluates a particular implementation of spread spectrum technique for steganography in auditive media, introduced by Nutzinger et al. in 2010 [10] and implemented by Nutzinger and Wurzer in 2011 [11]. This technique survived several robustness tests, such as noise addition, variable time delay, frequency shifting, GSM coding, air transmission, cropping, and resampling. It also did not show significant changes of perceived distortion level in hearing tests comparing original and modified signals. Finally, the phase spectrum and the time and frequency representation did not show significant changes [11].

What is the goal of this paper? As the title suggests, it is not a description of an implementation of an audio embedding method that is claimed to be secure, just an evaluation of a previously known method from the literature. It might well be a bit more secure than before, under particular assumptions. We can set some of

these assumptions as long as we want to play the attacker, but it would not say much about the security during a real application of the embedding method. It is probable that some of the attacks that we describe in this paper will be effective under certain conditions, and even successful through *other* steganographic (audio?) techniques. We unveil some weaknesses using rather simple methods that the implementors have not been aware of in their own validation of the embedding method. An evaluation at the given level of security—defined by the embedding method—does not require a universally working detector that is aware of all possible sources of stego signals, even if the steganographer could conceal some of the weaknesses using these sources. It is always advisable to identify the source of the weakness and revise the responsible part of the embedding method.

The paper is organised as follows. In the next section, the algorithm of the spread spectrum technique is described. Section 3 scrutinises the implementation of the spread spectrum technique. We found several weaknesses with proposed fixes in Sect. 4. Finally, Sect. 5 concludes the paper and gives an overview on our further work.

2 Spread Spectrum Algorithm

The steganographic algorithm of the StegIT-3 research project uses the audio signal of voice calls as its cover media. The voice call can be either a VoIP call or a mobile call over GSM or UMTS. The steganographic modulation for embedding the secret is applied at the decoded audio signal. The sample values of the uncompressed audio signal S_{float} are between the floating point values -1.0 and 1.0 . If the encoded audio signal uses the PCM16 codec, it will be converted as shown below:

$$S_{\text{float}} = S_{\text{PCM16}}/32768.0 \quad (1a)$$

$$S_{\text{PCM16}} = \lfloor S_{\text{float}} \cdot 32768.0 + 0.5 \rfloor \quad \text{if } S_{\text{float}} \geq 0 \quad (1b)$$

$$S_{\text{PCM16}} = \lceil S_{\text{float}} \cdot 32768.0 - 0.5 \rceil \quad \text{if } S_{\text{float}} < 0 \quad (1c)$$

The implementation of the StegIT-3 framework had a rounding bug. For more information see section 4.1.

For embedding, the original unchanged decoded voice signal (cover signal) $c(t)$ is used. By default, the sample rate f_s of a phone call is 8000 Hz, but the algorithm implementation would also work with any higher sample rate. At the sender, each secret bit is embedded as a chip sequence. One pseudo-noise chip sequence represents the bit value *false* while the other represents the bit value for *true*. A chip is represented by the value -1 or 1 ($V_{\text{chip}}(t)$). These sequences are generated by a linear feedback shift register (LFSR). Each chip of the chip sequence is embedded into the cover signal by the binary phase-shift keying (BPSK) modulation. The count of chips for one bit can be configured. It is a part of the stego key and also determines the transmission time for one embedded secret bit. The following equations show parameters for embedding a chip.

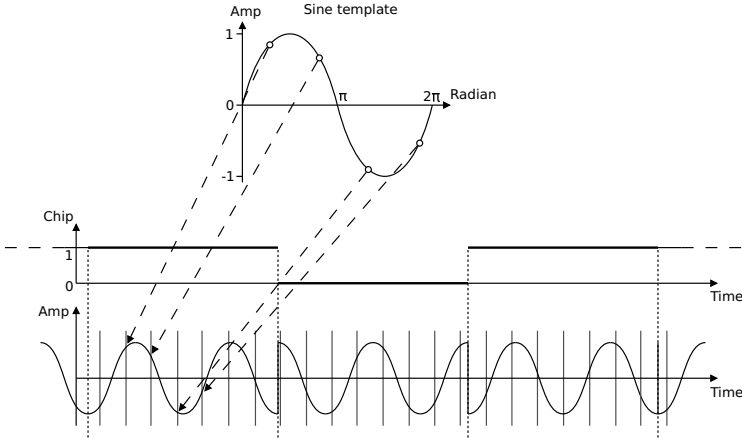


Fig. 1. Generation of the chips

$$500.0 \text{ Hz} \leq f \leq 3000.0 \text{ Hz} \quad \text{BPSK modulation freq.} \quad (2a)$$

$$T = 1/f \quad \text{period time} \quad (2b)$$

$$C_{\text{opc}} = 3 \dots 12 \quad \text{oscillations per chip} \quad (2c)$$

$$t_c = C_{\text{opc}} \cdot T \quad \text{chip period, chip time} \quad (2d)$$

$$V_{\text{chip}}(t) \quad \text{chip value } \{-1, 1\} \quad (2e)$$

$$t_{\text{start}} \quad \text{chip start offset} \quad (2f)$$

$$0 \leq \varphi \leq 2\pi \quad \text{phase for BSPK} \quad (2g)$$

For embedding the chip value $V_{\text{chip}}(t)$, the cover signal $c(t)$ is BPSK modulated according to Eq. 3, creating the modified (stego) audio signal

$$s(t) = c(t) + A_{\text{embed}}(t) \cdot V_{\text{chip}}(t) \cdot \cos(2\pi \cdot f \cdot t + \varphi). \quad (3)$$

$A_{\text{embed}}(t)$ and $V_{\text{chip}}(t)$ are constant for the embedding of one chip (chip time). The challenge of the algorithm is to find the perfect value for A_{embed} . The value of A_{embed} represents the amplitude for the BPSK modulation of one chip and affects the ability of the receiver to successfully extract the chip. A higher amplitude—while enhancing the quality of extraction—has negative impacts on the security of the steganographic algorithm. This algorithm uses a constant modulation frequency. An advanced version of this algorithm is described by Nutzinger [10]. Figure 1 shows the BPSK modulation of chips and Fig. 2 the cover and the stego signal with the added BPSK modulation chip signal.

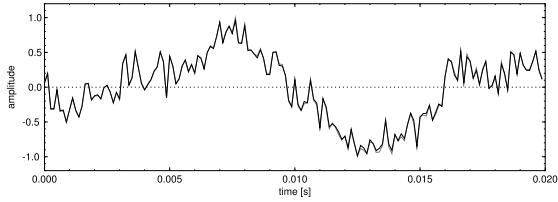


Fig. 2. Cover signal (dark line) and stego signal (light line)

The receiver has to extract the chips from the audio signal. The demodulated chip value at the receiver side is given by

$$V_{\text{chip,ext.}} = \begin{cases} 1 & \text{if } d \geq 0 \\ -1 & \text{if } d < 0 \end{cases}, \tag{4}$$

$$\text{with } d = \int_{t_{\text{start}}}^{t_{\text{start}}+t_c} s(t) \cdot \cos(2\pi \cdot f \cdot t + \varphi) dt. \tag{5}$$

3 Attacks

The goal of this project was to judge the security of the embedding algorithm. It is sensible to study all information before trying to detect traces of the embedding process in the output signal, however, we (involuntarily) played the attacker in two different setups. Due to an intellectual property issue, neither the C++ source code, nor the binary of the implementation was available in the first phase. We could only get a small number of WAV files (recorded phone calls), each file in several versions (without any message, with an embedded message in three different embedding configurations: “amp,” “bpsk,” and “phase,” in order to find out which one is the best). We also knew that the messages had been embedded with some kind of spread spectrum modulation.

3.1 Twin Peaks?

We started our research with the simplest attack we could think of, namely a histogram attack à la *Twin Peaks* [12], since the embedding method uses spread spectrum modulation (SS). Not knowing the exact implementation of the SS method, we assumed a simple direct sequence SS (DSSS) algorithm. We hoped that at least one of the three configurations (“amp,” “bpsk,” and “phase”) is close enough to our vague assumption. Figure 3 gives a concrete example of the assumed simple SS embedding¹ with a (zero mean) Gaussian cover signal. A PN sequence, consisting of random samples -1 and 1 only, is used to spread one symbol (e.g. a bit) over a longer time period. If this spreading sequence is added to the cover signal, the symbol can be decoded as the scalar product of the

¹ Which is indeed hard to match to the real implementation described in Sect. 2.

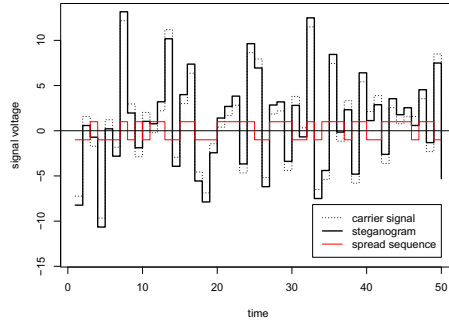


Fig. 3. A spreading PN sequence is added to the cover signal (dotted line), resulting in the stego signal (bold line)

spreading sequence and the stego signal. If the histogram of the cover signal has one peak, there might be two peaks in the histogram of the stego signal. (Hence the name of the attack.) Since the cover signal is Gaussian, the stego signal is the sum of two Gaussian distributions, with a mean distance determined by the spreading sequence ($1 - (-1) = 2$). If the variance of the cover signal is large (cf. Fig. 4, left), e.g. a louder part of a phone call, the resulting distribution is hard to distinguish from the original Gaussian. However, in more quiet passages of the phone call, the resulting distribution might show twin peaks (cf. Fig. 4, right), or a noticeable change of the distribution exploitable for the detection of the steganographic method. We expected at least a kind of automatic volume control of the spread sequence, which reduces the amplitude according to the cover signal. However, since the embedding method should be useful for phone calls, real-time properties are a concern. It may be that the control is delayed, in order not to delay the speech signal too much.

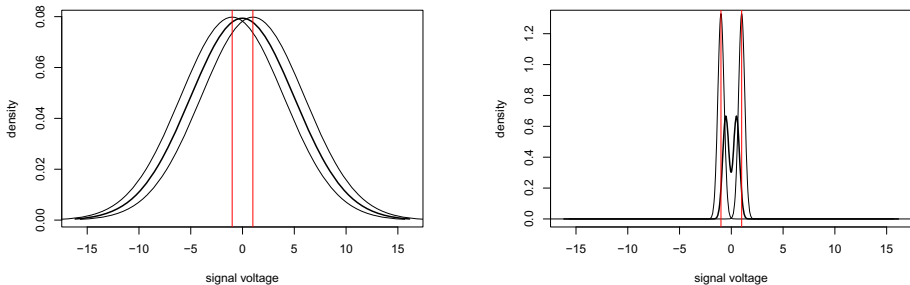


Fig. 4. If the cover signal is strong enough, the resulting composite distribution (bold line) still seems to be Gaussian (left), however, quiet passages of the stego signal might show twin peaks (right)

Surprisingly, our rather blind attack separated (even parts of) cover and stego signal in our test database perfectly. The detector worked in two steps. The first step selected quiet parts of the signal (a dispensable step as we will see later), the second created a histogram, which showed a single peak at zero for stego signals, but not for cover signals. While the cover signal contained about the same number of zeros as ones (maybe up to 30 % more), in stego signals we found twice as many zeros. Interestingly, this worked for the configurations “amp,” “bpsk,” and “phase” with the same threshold. If there are more than 1.5 as many zeros than ones, the signal is a detected stego signal.

To understand the reason, we had to wait until we finally received the source code (cf. Sect. 4.1).

3.2 “Steps”

A cover–stego attack is possible here, i.e., the synchronous confrontation of cover samples c_i and their corresponding stego samples s_i . The closer the samples to the diagonal $s_i = c_i$, the smaller the change caused by the embedding. Figure 5 opposes cover and stego samples, resulting in an overlay of diagonal stripes. Obviously, there is a mechanism that controls the embedding intensity in discrete steps. However, under more realistic conditions (stego only attack), an attacker has to estimate the cover from the stego signal. This is usually called “denoising.”

3.3 Saturn Sighted

We could model the cover sample from other, but temporally close stego samples:

$$c_i \sim s_{i-2} + s_{i-1} + s_{i+1} + s_{i+2}. \quad (6)$$

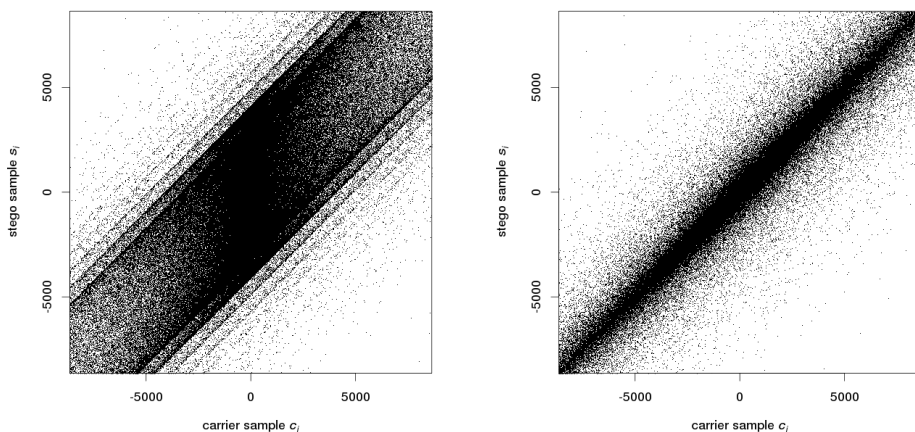


Fig. 5. Synchronous confrontation of cover and stego samples shows steps of a controller (left), and no steps after correction (right)

A similar approach was used during the BOWS-2 contest to estimate the unmarked magnitude of wavelet coefficients from its surrounding [13]. This has been successful, because the piece of watermark in s_i was independent of the watermark in the samples of the surrounding. Unfortunately, we cannot be sure or even expect this property in the case of the attacked audio signal here, since a chip time could be longer than a sample time. Nevertheless, we simply predicted the next cover sample by the current stego sample

$$\hat{c}_{i+1} = s_i \quad (7)$$

leading to the impressive, “astronomic” constellation in Fig. 6.

Although resembling the planet Saturn, it is technically a Lissajous curve that can be estimated using the following ellipse:

$$t^2 = \left(\frac{y - mx}{b} \right)^2 - \frac{x^2}{a^2}, \quad \text{with} \quad (8)$$

$$a = 3280,$$

$$b = 2850,$$

$$m = 0.496,$$

where t is a threshold parameter. Lissajous curves appear on an oscilloscope in X-Y mode, when the two inputs are sinusoidal signals. Here the two inputs come from the same, but time shifted signal, i.e. have the same frequency but different phase, resulting in an ellipse. The phase shift is determined by the time between two consecutive samples, i.e., depends on the sample rate. A detector can be constructed, which assumes a stego signal if a suspicious amount of samples occurs between the dashed ellipse ($t = 0.7$) and the dotted one ($t = 1.4$), compared to the total number of samples.

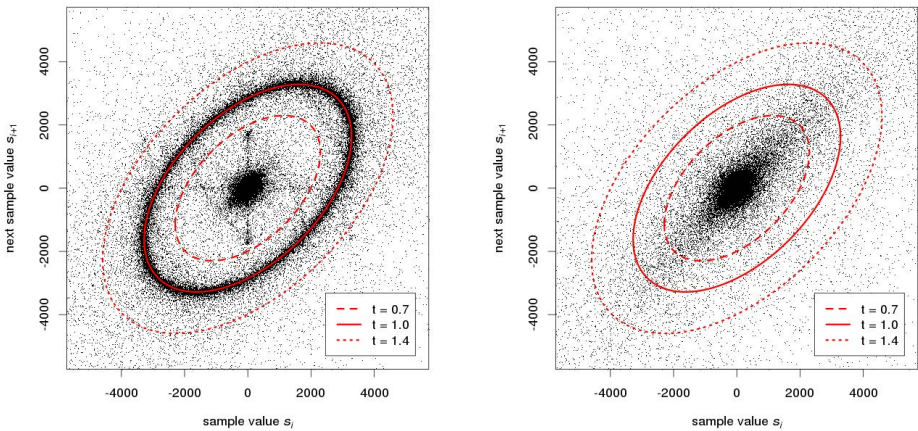



Fig. 6. Confrontation of consecutive stego samples (next sample value as an estimate of cover) shows a Lissajous curve (left), and after correction (right)

The strength of the effect might be surprising. It is *not* the usual parameterisation, but one for testing the algorithm’s behaviour on a GSM channel. In this test case it is of course audible. However, even with the “production” parameters, a small ring ($a \approx b \approx 30$: ;-) can be isolated in some quiet passages of the stego signal (and is covered by the dark centre of Fig. 6 otherwise). We admit that we could have missed this fact in the rest of the test cases if the accidental GSM test case would not have been included in our test database setup.

4 Countermeasures

4.1 Solitary Peak

The cover signal is read from an audio source, usually a telephone, sound card or WAV file (e.g., 8000 PCM samples per second, 16 bits per sample, 1 channel). The samples are signed integers. To support different rates and precisions, the embedding algorithm internally maps the raw data to a series of normalised `double` floating point samples in the range $-1 \dots 1$.

The conversion is implemented by a type cast from `double` to integer, followed by scaling down to the desired interval $[-1, 1]$. Finally, the `double` values are scaled up (and clipped, if necessary) to the original range, and casted back to integer. If the values are not changed in between by the embedding step, the final cast from `double` to integer is one-to-one, because the fractional part is zero.

However, if something is embedded, the final `double` values will also have non-zero fractional parts. The obvious, but careless, use of type cast takes revenge here. The cast operator (`y = (int)x;`) takes a numeric argument $x \in \mathbb{R}$ and returns an integer $y \in \mathbb{Z}$, formed by truncating the values in x toward 0 (cf. Fig. 7). Possible repair: `y = (int)(x + ((x < 0) ? -0.5 : 0.5));`

After the problem of spotty rounding around zero was solved, another problem was detected that occurs with some cover signals only, because of the asymmetry of the integer domain. There is an even number of 16 bit integers, one is neutral (0), 32767 are positive, 32768 are negative. If we negated -32768 (0x8000) there would be a sign overflow, resulting in the same (negative) value. The implemented mapping to $[-1.0, 1.0]$ divided all samples by 32767 and clipped -32768 to -1.0 . If the signal is sufficiently saturated, there will be peaks in the histogram of cover samples for the saturated values ($-32768, 32767$). The conversion routines of the mapping shifted the peak at -32768 to -32767 . In case of saturation this provides a rather safe feature for detection. The obvious repair is to change the divisor to 32768, and to clip positive values above 32767 when mapping back to integer.

4.2 Better Amplitude Adjustment Control for BPSK Modulation

At the sender’s side, the embedding algorithm determines whether the original cover signal is suitable for embedding the secret chip. In case decoding the original audio signal would result in the secret chip’s value, it is not necessary to

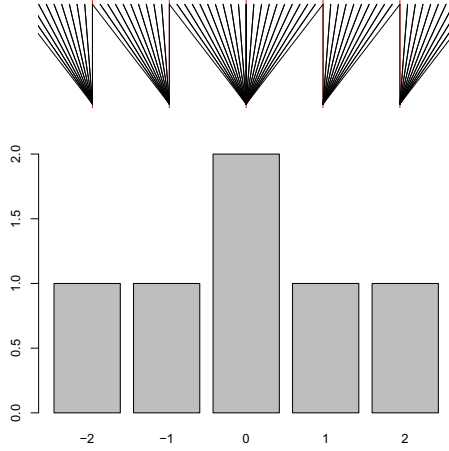


Fig. 7. Bin 0 collects from a double-width interval due to truncation towards zero, resulting in a histogram peak

modify the cover signal to embed the chip. This is shown in Eq. 9. The value of c_{embed} decides if a modification of the cover signal is necessary (cf. Eq. 10).

$$c_{\text{embed}} = \int_{t_{\text{start}}}^{t_{\text{start}}+t_c} c(t) \cdot V_{\text{chip}}(t) \cdot \cos(2\pi \cdot f \cdot t + \varphi) dt \quad (9)$$

$$\text{steganographic modification is } \begin{cases} \text{not necessary} & \text{if } c_{\text{embed}} > 0 \\ \text{necessary} & \text{if } c_{\text{embed}} \leq 0 \end{cases} \quad (10)$$

In case of embedding, the value of A_{embed} is increased step-by-step until a chip can be extracted correctly at the receiver’s side. However, this approach shows steps when plotting the cover against the stego signal (cf. Fig. 5). Also a fix minimum amplitude was added. The fix minimum amplitude increases the effect of “Saturn Rings” (cf. Sect. 3.3). In order to avoid “Saturn Rings” and “steps,” the amplitude A_{embed} is determined by a modified control mechanism (cf. Eq. 11 ... 15).

$$c_{\text{avg}} = \frac{1}{t_c} \int_{t_{\text{start}}}^{t_{\text{start}}+t_c} c(t) \cdot V_{\text{chip}}(t) \cdot \cos(2\pi \cdot f \cdot t + \varphi) dt \quad (11)$$

$$\text{steganographic modification is } \begin{cases} \text{not necessary} & \text{if } c_{\text{avg}} > 0 \\ \text{necessary} & \text{if } c_{\text{avg}} \leq 0 \end{cases} \quad (12)$$

$$\int_{t_{\text{start}}}^{t_{\text{start}}+t_c} c(t) - 2 \cdot c_{\text{avg}} \cdot V_{\text{chip}}(t) \cdot \cos(2\pi \cdot f \cdot t + \varphi) dt = 0 \quad (13)$$

$$s_{\text{ARV}} = \frac{1}{t_c} \int_{t_{\text{start}}}^{t_{\text{start}}+t_c} |c(t)| dt \quad (14)$$

$$A_{\text{embed}} = 2 \cdot |c_{\text{avg}}| + s_{\text{ARV}} \cdot A_{\text{add}} \quad (15)$$

The signal mean c_{avg} (cf. Eq. 11) determines whether embedding is necessary. This value is the basis in the new definition of the embedding amplitude A_{embed} (cf. Eq. 15). For a successful chip extraction at the receiver side, A_{embed} must be greater than the double of c_{avg} . The embedding amplitude A_{embed} is increased by an additional part A_{add} that is scaled with the averaged rectified values in the signal segment s_{ARV} to provide the receiver the necessary margin for extraction of the correct chip value. The scaling relieves the ‘‘Saturn’’ effect (cf. Sect. 3.3) compared to the initial choice of a constant margin (e.g., an unscaled $A_{\text{add}} = 0.1$).

5 Conclusions

We found several weaknesses in the implementation of a spread spectrum technique for steganography in auditive media. Some of them did not result from the embedding technique itself, but the mapping from the external cover representation to the internal working representation. It seems to be important to carefully check conversion and normalisation functions, their homogeneity around special values like 0, and their properties in case of saturation.

However, also the embedding algorithms itself showed weaknesses. It seems to be important to consider the difference between cover signal and stego signal during the design of the algorithm. Although an average attacker cannot access this difference signal, obtrusive properties might radiate through the cover’s shielding guard. It is also advisable to use pathologic signals, like rhythmic audio pulses, to test the integrity, for instance, of control mechanisms.

Be aware of correlations within the cover’s values. Such correlations will also occur in the stego signal. Often, such correlations can be used to ‘‘denoise’’ the signal or to ‘‘calibrate’’ statistics [14,15], even in audio streams.

Acknowledgments. This work was supported in the KIRAS programme for security research by the Austrian Federal Ministry for Transport, Innovation and Technology.

References

1. Tachibana, R., Shimizu, S., Nakamura, T., Kobayashi, S.: An audio watermarking method robust against time- and frequency-fluctuation. In: Delp III, E.J., Wong, P.W. (eds.) *Security, Steganography and Watermarking of Multimedia Contents III* (Proc. of SPIE), San Jose, CA, pp. 104–115 (2001)
2. van der Veen, M., Bruekers, F., Haitsma, J., Klaker, T., Lemma, A.N., Oomen, W.: Robust multi-functional and high-quality audio watermarking technology. In: 110th Audio Engineering Society Convention. Volume Convention Paper 5345 (2001)
3. Kirovski, D., Malvar, H.S.: Spread-spectrum watermarking of audio signals. *IEEE Trans. on Signal Processing* 51, 1020–1033 (2003)

4. Steinebach, M., Petitcolas, F., Raynal, F., Dittmann, J., Fontaine, C., Seibel, S., Fates, N., Ferri, L.: StirMark benchmark: audio watermarking attacks. In: International Conference on Information Technology: Coding and Computing, pp. 49–54 (2001)
5. Arnold, M., Baum, P.G., Voeßing, W.: A phase modulation audio watermarking technique. In: Katzenbeisser, S., Sadeghi, A.-R. (eds.) IH 2009. LNCS, vol. 5806, pp. 102–116. Springer, Heidelberg (2009)
6. Marvel, L.M., Boncelet, C.G., Retter, C.T.: Spread spectrum image steganography. *IEEE Transactions on Image Processing* 8, 1075–1083 (1999)
7. Pichholtz, R.L., Schilling, D.L., Milstein, L.B.: Theory of spread-spectrum communications—a tutorial. *IEEE Transactions on Communications* 30, 855–884 (1982)
8. Petitcolas, F.A.P., Anderson, R.J., Kuhn, M.G.: Attacks on copyright marking systems. In: Aucsmith, D. (ed.) IH 1998. LNCS, vol. 1525, pp. 218–238. Springer, Heidelberg (1998)
9. Westfeld, A.: Steganography for radio amateurs— A DSSS based approach for slow scan television. In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 201–215. Springer, Heidelberg (2007)
10. Nutzinger, M., Fabian, C., Marschalek, M.: Secure hybrid spread spectrum system for steganography in auditive media. In: 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IHH-MSP, pp. 78–81 (2010)
11. Nutzinger, M., Wurzer, J.: A novel phase coding technique for steganography in auditive media. In: 6th International Conference on Availability, Reliability and Security, ARES, pp. 91–98 (2011)
12. Maes, M.: Twin Peaks: The histogram attack to fixed depth image watermarks. In: Aucsmith, D. (ed.) IH 1998. LNCS, vol. 1525, pp. 290–305. Springer, Heidelberg (1998)
13. Bas, P., Westfeld, A.: Two key estimation techniques for the broken arrows watermarking scheme. In: Proc. of ACM Multimedia and Security Workshop, Princeton, NJ, USA, pp. 1–8 (2009)
14. Fridrich, J., Goljan, M., Hoge, D.: Steganalysis of JPEG images: Breaking the F5 algorithm. In: Petitcolas, F.A.P. (ed.) IH 2002. LNCS, vol. 2578, pp. 310–323. Springer, Heidelberg (2003)
15. Kodovský, J., Fridrich, J.: Calibration revisited. In: Proc. of ACM Multimedia and Security Workshop, Princeton, NJ, USA, pp. 63–73 (2009)

Robust Hash Algorithms for Text

Martin Steinebach^{1,2,3}, Peter Klöckner², Nils Reimers²,
Dominik Wienand², and Patrick Wolf³

¹ Fraunhofer SIT, Rheinstrasse 75, Darmstadt, Germany

² CASED, Mornewegstrasse 32, Darmstadt, Germany

³ CoSee GmbH, Rheinstrasse 75, Darmstadt, Germany

Abstract. We discuss and compare robust hash functions for natural text with respect to their performance regarding text modification and natural language watermark embedding. Our goal is to identify algorithms suitable for efficiently identifying watermarked copies of eBooks before watermark detection.

Keywords: Robust Hashing, Text Watermarking, Evaluation.

1 Introduction

While multimedia content and machine to machine data have seen a strong increase in recent years, written natural text still has an important role in information distribution, storing and distribution of knowledge as well as entertainment. Books, scientific papers, patents, news articles - it is easy to list many examples important in everyday life and work. Still, concepts for reliable authentication specifically designed for natural language text are rare. Most often they are based on cryptographic hash functions, which are secure and reliable, but often require a precision of reproduction that brings challenges to efficient data handling: While in contracts every single word may be of importance, in many other documents it is rather the meaning and the flow of ideas that counts. This is especially true if digital watermarking for natural language is applied as watermarking will change the wording but not the meaning of texts, e.g. by active/passive or enumeration modulation.

In this work we discuss and compare alternatives for natural language hashing. These hashes shall feature robustness comparable to robust image or audio hashes. As long as a human observer perceives copies of a work as the same, the hash should also be identical or at least similar. Our goal is to provide a system allowing the following work flow:

- Create a robust hash H of a text T
- Create n individually watermarked copies TM of T
- Use H to identify all n copies of TM

If no hash method robust against the embedding of a watermarking, for each TM a cryptographic hash needs to be computed and stored if we want to proof

TM to be a copy of T. At the same time, the only alternative to a hash is a comparison to the original copy of T. While this is acceptable with respect to computation speed and resilience to errors, here the big drawback is the need to distribute the original text. If the application is to scan the Internet for a secret document, the document often will or at least should not be available to the searching agent.

2 Motivation

As portable eBook reader become more and more common, the sale of eBooks grows. EBook revenues for 2012 are at \$1.3 billion, up 46% from 2011 [1] and forecasts for 2016 range from \$5 billion [2] to \$10 billion [3]. Copyright holders, i.e. publishers, will face the same challenges as the music or film industry with pirated versions of their intellectual content. The illegal distribution of eBooks is comparably simple, due to their small file size which is usually about 1 megabyte. It takes only a few minutes to find free versions of all books from the *Spiegel Bestsellerliste*. On the illegal channels one can also find the scanned version of printed books.

Watermarking on content level can help to determine the leakage if an eBook is found on an illegal channel. Watermarking works in the way that it modifies the content in non-noticeable way to include a unique ID. A publisher would then be able to check the channels if his eBooks are leaked and could then identify the source of the leaked copy.

This requires of course a method to verify that a given eBook found in one of the illegal channels belongs to the publisher. Taking the content of a found eBook and comparing it on text level with a list of all owned books would be quite inefficient. Also the publisher may want to outsource checking of these illegal channels to a third party but is not willing to hand out the content of its books.

Using a hashing algorithm like SHA-1 would fail as soon as there is a minimal modification on the content. This modification could be intentional in order not to be detected. But likely it's unintentional, due to format conversion, e.g. from ePub to PDF, due to an OCR error or maybe the eBook was split into parts. As mentioned before watermarking also introduces changes into the content of the book, hence each version would have a unique SHA-1 fingerprint.

A robust hash algorithm for text documents is therefore required. It should produce the same hash value for nearly identical contents. Obviously OCR errors, small modifications and watermarks should result in the same hash value. Still defining robustness requirements can be challenging: Should a substring, e.g. the first 10 or first 100 pages of a book, produce the same hash value?

3 Related Work

The goal to create a text authentication method robust against slight modifications is not new. As an example, plagiarism recognition faces this challenge on a regular base.

3.1 Cryptographic Hashing

Hash functions allow securely computing a short digest of a long message. Mathematically speaking, a hash function is a function that maps a variable length input message to an output message digest of fixed length. Cryptographic hash functions (such as SHA-1 or RIPEMD) are by design extremely sensitive to changes in the input data: even changing one bit in the data results in a totally different hash value.

3.2 Piecewise Hashing

Piecewise hashing, also called fuzzy hashing, combines cryptographic hashing and data segmentation. One of best know examples is Ssdeep that implements an algorithm called context triggered piecewise hashing (CTPH) presented by Kornblum in 2006 [4]. It divides a byte sequence into chunks and hashes each chunk separately using the Fowler algorithm. To represent the fingerprint of a chunk, CTPH encodes the least significant 6 bits of the FNV hash as a Base64 character. All these characters are concatenated to create the fingerprint.

3.3 Robust Hashing

Perceptual hashes usually extract features from an multimedia data which are relevant to human perception and base the digest on them; thus, by design the perceptual hash of two similar objects will be similar. Perceptual hashes are sometimes also called digital fingerprints as they allow to identify content through extracted features.

3.4 Text Hashing

In the following section we describe different hash functions for natural language text. There are many applications that use cryptographic hash functions and piecewise hashing for text hashing. These approaches have one common weakness: If the document is changed by natural language watermarking, these hash functions will fail. If piecewise hashing is applied, this depends on the relation between chunk size and distance between changes caused by watermark embedding.

4 Approaches

We implemented and evaluated three algorithms: WordToBit, Broder and SimHash. The first one is based on the simple idea of hashing each word in the input text to a single bit, while the latter two are borrowed from near duplicate detection methods used in web crawling.

4.1 Word to Bit

This algorithm creates a digest by hashing each word in a given text to a single bit. To do this, we first split the text into a list of word tokens. Then we convert each word to either 0 or 1. This conversion should map the space of world uniformly at random to the space $\{0, 1\}$. We use the least significant bit of the Java built in hashCode function of the word. Other, more efficient text hash algorithms could be used. The digest is the concatenation of all those bits and its length is thus equal to the number of words in the text.

For comparison a distance measure is required. The Hamming distance is not usable as a single deleted word causes the rest of the digest to be off. The Levenshtein distance is suitable comparably slow. If the hashes are close, computing it on small parts where the two hash values differ speeds it up.

We decided to use instead a sampling approach to measure the distance of two WordToBit hash values. One hash is designated the main hash and we then try to find sub-samples of the other hash in that main hash. The motivation behind this was to be able to detect parts of a text (e.g. a chapter) in digests. The hash is split into sub-samples of e.g. 128 bits size (equal to 128 words). We then compute the Hamming distance at each position of the main hash for all sub-samples. If the distance is below a given threshold (e.g. 1 / 4 of the sub-sample size), we assume the sub-sample to match at that position. If we can find a certain number of matches (one is usually enough) we consider the whole text to match.

4.2 Broder's Algorithm

Broder's algorithm, as described in [5], uses shingling to introduce a similarity measure for message digest. As mentioned before, Broder's as well as the Charikar's SimHash algorithm have been proven to be efficient to find near-duplicates in web crawling [7].

The algorithm uses m different Rabin fingerprint functions f_i , $0 \leq i < m$. The procedure starts with f_0 . Each subsequence of k tokens is fingerprinted with f_0 , which leads to $n - k + 1$ 64-bit values, called shingles, for f_0 . The smallest of these values is the first minvalue and the first value of the hash.

The algorithm proceeds with doing the same for $f_1, f_2 \dots f_{m-1}$. Thus the algorithm results in a hash consisting of m minvalues, which leads to a hash size of $64 \cdot m$ bits. We use $k = 8, m = 84$ in most test cases. Evaluation later showed that m can be reduced to 25, which further decreases the computation time. To estimate the similarity of two texts, we determine the number of equal minvalues in their hashes and call this B-Similarity. We consider two books to be a match if the B-Similarity is at least two.

4.3 SimHash

Charikar's random projection based approach [6] is used for finding near-duplicate web pages [7]. We adopted the proposed algorithm from Charikar and

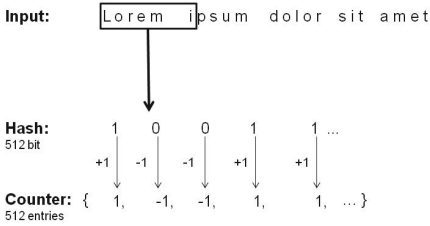


Fig. 1. SimHash mode of operation

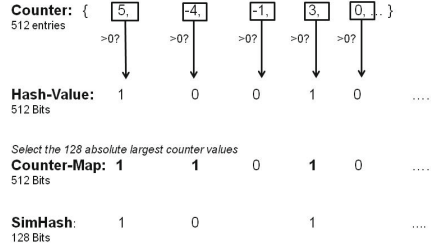


Fig. 2. SimHash: Final computation step

introduced slight changes, to yield a higher robustness. Charikar’s algorithm is a fingerprinting technique that enjoys the property that fingerprints of near-duplicates differ in a small number of bit positions. We will call in the following the proposed algorithm SimHash, following the name conversion from [7].

SimHash splits works on tokens of length n . Tokens can either single characters or words. We implemented both versions, for n -grams and word sequences of length n . We decided to use the n -grams version for the ease of use and also it seemed to have better robustness properties. Each n -gram is then randomly projected to the space $\{0, 1\}^k$. We use $n = 12$ and $k = 512$ and the random projection is computed by a reduced round SHA-512 implementation. As the random project does only need proper random properties, computing all rounds of SHA-512 is not necessary. Other, more efficient, pseudo-random projections could also been used.

For each n -gram we compute such a weakened SHA-512 hash value. At the same time we initialize a counter with 512 counter values. If the first bit of a SHA-512 hash equals one, we increment the first counter by one. Otherwise we decrement the counter. This is done for each bit of the hash value. Figure 1 depicts this. After hashing all n -grams and incrementing/decrementing the counter values, we convert the counter to a final hash value. In the original paper, each entry of the counter is converted to one if it is bigger than zero, else to zero. Given a random project, the expected value of each counter entry equals zero. A single changed n -gram may change the sign of a counter value, resulting to a flip in the final hash value.

To overcome this problem, we introduce a *compression step*. We select out of the 512 counter entries the 128 *most robust* entries, i.e. the entries with the largest absolute value. Changing one n -gram will not lead to a change of the sign for these counter entries. Figure 2 illustrates this final compression step.

For a given text document our SimHash algorithm returns a 512 bit large counter map and an either 128 or 512 bit SimHash value. The counter map contains a one if the counter entry belongs to the most robust entries, i.e. to one of the 128 with largest absolute value. The SimHash value can be either 128 bits, if we only convert the counters with the largest absolute values, or 512 if we simply convert all counter values.

There are several options to compute the similarity between two SimHash values. One option would be to simply compute the Hamming distance between the two compressed 128 SimHash values, ignoring the counter map. This can simply be done and the main benefit of this would be to create an easier indexing for the hash values.

For our evaluation we used a more complex comparison routine. One hash is declared as main hash. Its counter map is then used to extract from both inputs the 128 Bit SimHash value. Then the Hamming distance is computed. This introduces an asymmetric distance measure, but which could easily be made symmetric.

5 Evaluation

The evaluation of the algorithms was performed in two steps, which we call white box and black box test, respectively. White box tests should examine the properties of the present algorithms, i.e. we not only wanted to show the robustness of the algorithms but also quantify the robustness in some way. Black box tests abstract from the underlying algorithm. Here for given test scenario we were just interested in the false positive and false negative rates.

5.1 White Box Tests

To be a suitable hash algorithm, the algorithms should produce distinct hash values for distinct inputs. To test this, we extracted 1000 randomly selected articles from the German Wikipedia with a size between 9,700 and 335,000 bytes. The wiki markup was removed by WikiPrep [8] in order to gain the pure text content of the articles.

The distinction property for WordToBit is straight forward and was not further analyzed by us. As one can see in Figure 3, the mean Hamming distance of SimHash is 64, which is also the expected value for a random projection to a 128 bit space. For Broders algorithm, 998,780 out of the 999,000 possible article combinations produce a B-Similarity of zero. One pair produced a B-Similarity of 23, which is fairly high. This is due to the fact that one can find the same information or even the same paragraphs over different Wikipedia article, e.g. if a main article is split into several sub articles. In conclusion all algorithms provide sufficient distinction properties.

Revisions of Wikipedia offer a large corpus on manual modifications on text. Often a new revision changes only a typo, replaces a word or adds some information. We extract 9400 revision of various articles of the English Wikipedia to test our algorithms. We compared each revision with the previous revision of the article and computed the Levenshtein distance in order to get a measure for the performed modification. The result for SimHash and Broder can be seen in Figure 5 and Figure 6. In conclusion, both algorithm are robust to (small) human made modifications on text.

As mentioned in the introduction, in some cases a book is scanned before it is distributed. Either because only the printed version is available or also to

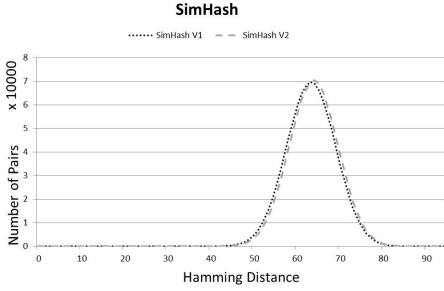


Fig. 3. Hamming distances of SimHash for 1000 randomly selected Wikipedia articles

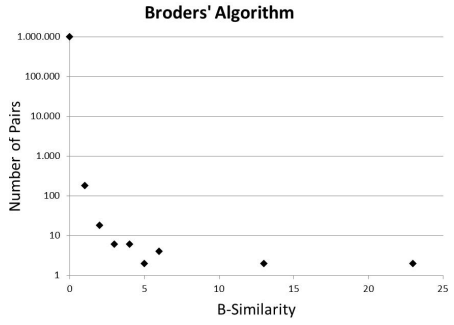


Fig. 4. Hamming distances of SimHash for 1000 randomly selected Wikipedia articles

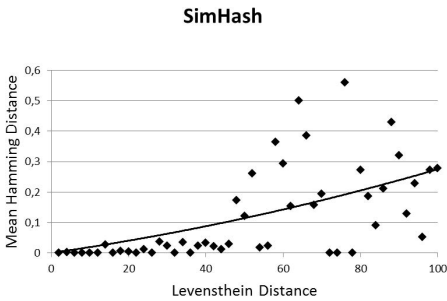


Fig. 5. SimHash revisions results

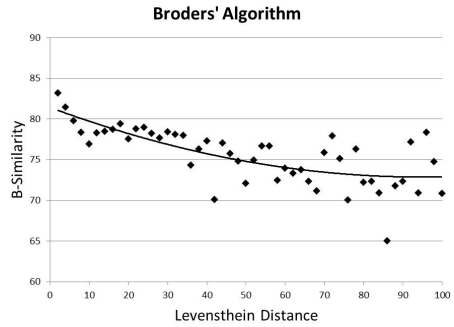


Fig. 6. Broder revisions results

remove watermarks or circumvent DRM. We simulated typical OCR errors: I's recognized as l's, l's recognized as I's, s' recognized as f's and rn's recognized as m's. We increased the error rate from 4, 8, 25 and 10 percent in ten steps to 40, 80, 25 and 100 percent. Each bar in the following figures represents the average distance over 10 books. The distance of the SimHashes grew from zero to 10, the B-Similarity of Broders' algorithm fell reciprocally from 84 to 10.

To evaluate the performance with respect to natural language watermarking, we used eight eBooks from the current top-selling charts as covers and created marked copies of it. The watermarked versions differ from the original cover in the order of some enumerations, for example "he was smart and cute" compared to "he was cute and smart". This is one of the few accepted concepts for natural language watermarking in German language.

It turned out that using Broder with $m = 84$ leads to 3.33% of the eBooks having a B-Similarity of 82 compared to their watermarked versions, 35% a B-Similarity of 83 and 61.66% a B-Similarity of 84. SimHash even resulted in each pair of hashes of the eBook and its watermarked version having a distance of

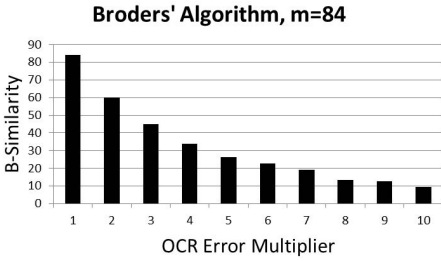


Fig. 7. Broder, OCR Simulation

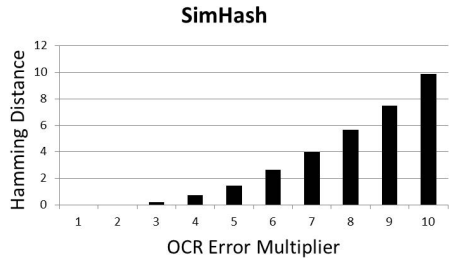


Fig. 8. SimHash, OCR Simulation

zero. This shows that the given watermarking algorithm has no impact on our hash methods' ability to detect matching eBooks.

5.2 Black Box Tests

To evaluate and compare the performance of all three algorithms, we implemented an automated runner for “test tracks” taken from our corpus of texts. The test tracks are sets of matching and not matching eBook pairs assembled to represent typical use case scenarios. The first test track “Publisher” simulates the case of a publisher scanning a collection of possible copies for a single work of his. The second test track, “FBI” simulates searching a list of suspicious files for a list of known “bad” works. There is also one internal test track that contains a mixture of comparisons from the white box tests. We applied each algorithm on the given test track and logged the runtime and the number of false positives/negatives.

We tuned all three algorithms to produce zero false positives/negatives on the two main tracks. However, as our corpus of real eBooks is not nearly large enough to yield significant results regarding false positives/negatives, these tests were mainly used to assess the runtimes of the algorithms. There is a 32 bit implementation of WordToBit (using ints) and a 64 bit implementation (using longs). Benchmarks were done on a 64 bit machine. On a 32 bit machine the results for WordToBit and WordToBit64 would roughly be reversed. SimHashV1 runs on word tokens, SimHashV2 on n-grams.

The Publisher Test Track has a 1:1 ratio of hash generations to comparisons with many real books. The FBI test track contains about 1000 works and comparisons are done on the cross product of those works. WordToBit performs extraordinarily bad because its expensive operation is the comparison as opposed to the other algorithms where hashing is expensive. The internal test track again features a 1:1 hash generation:comparison ratio but with a lower number of real books, which makes WordToBit perform better than on the publisher test track.

We used black box tests to tune the parameters for WordToBit and Broder. Test runs were done on an older version of the internal test track. We ordered the results by the sum of false positives and false negatives and then by runtime. Table 1 shows the top 10 results for the parameters of WordToBit. The best

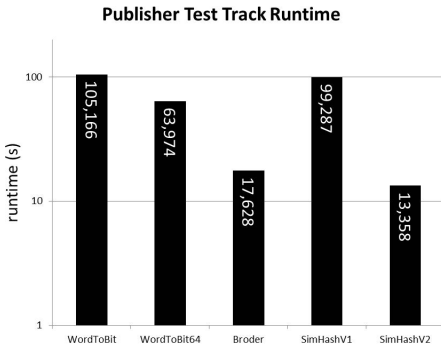


Fig. 9. Publisher Test Track Runtimes

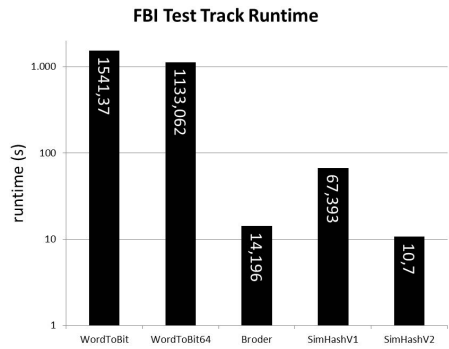


Fig. 10. FBI Test Track Runtimes

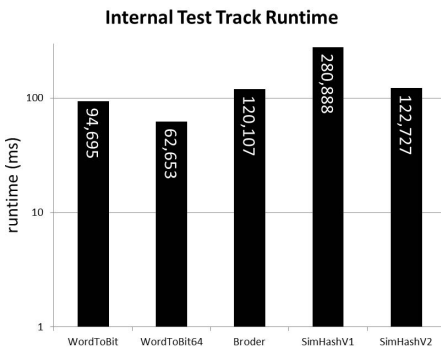


Fig. 11. Internal Test Track Runtimes

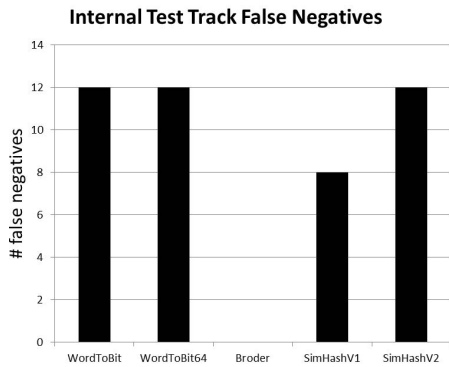


Fig. 12. Internal False Negatives

Table 1. Word to Bit Parameter Ranking

Rank	SubSampleSize	DistanceThreshold	Runtime	FalsePositives	FalseNegatives	False Sum
1	3	18	174904	0	3	3
2	3	15	177301	0	3	3
3	2	6	194477	0	4	4
4	2	8	196483	2	2	4
5	4	28	169801	0	5	5
6	3	12	174502	0	5	5
7	5	35	147376	0	6	6
8	4	24	168241	0	6	6
9	2	4	195907	0	6	6
10	5	30	147188	0	7	7

combined false rates can be achieved with subsample size 3 and the distance thresholds 15 and 18. For Broder in the top 10 the false sum was always 0, therefore runtime was the only discriminator. The best runtime of 40971 ms was achieved with $m=25$, $k=4$.

6 Conclusion

As a result of our evaluation, we come to the conclusion that all the algorithms are suited for specific tasks or can be applied to satisfy certain requirements. To find chapters or in general parts of a full text we recommend the use of WordToBit. To achieve low latency one should use SimHashV2 or Broder as both are faster than WordToBit. To achieve a high precision it is advisable to use Broder as here zero false negatives could be achieved. Still, all algorithms show that it is advisable to utilize robust text hashing in error- or noise-prone environments as their robustness is an important advantage compared to common hash methods.

We show that the combined use of hashing and watermarking is applicable. Together, these tools provide a promising way to individually mark written language and identify it later on without the need of huge data bases or the leakage of the cover text. One aspect that is common in hash protocols based on similarity search is that the search and comparison part of the hash detection part must not be neglected. Depending on the application, this can become more time-demanding than the actual hash calculation.

Acknowledgement. This work has been supported by the Federal Ministry of Education and Research via the project SIDIM (01IS10054A) in the funding initiative "KMU-innovativ" for the innovative SMEs target group.

References

1. Hoffelder, N.: AAP Reports US eBook Sales Up 46% in 2012, Now Well Over a Fifth of US Book Market
2. Wolf, M.: E-book market forecast to hit \$5.2B as the book industry burns
3. Wauters, R.: Total Mobile eBook Sales Forecast To Reach \$10B By 2016; Now Close To 1 Million Books In Kindle Store
4. Kornblum, J.: Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation* 3(S) (2006)
5. Broder, A., Glassman, S., Manasse, M., Zweig, G.: Syntactic Clustering of the Web. In: 6th International World Wide Web Conference, pp. 393–404 (April 1997)
6. Charikar, M.: Similarity estimation techniques from rounding algorithms. In: Proc. 34th Annual Symposium on Theory of Computing, STOC 2002, pp. 380–388 (2002)
7. Manku, G., Jain, A., Sarma, A.: Detecting near-duplicates for web crawling. In: Proceedings of the 16th International Conference on World Wide Web (2007)
8. Gabrilovich, E.: Wikipedia Preprocessor (WikiPrep),
<http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/>

Hardware Based Security Enhanced Direct Memory Access

Marcel Eckert², Igor Podebrad¹, and Bernd Klauer²

¹ Group Security, Threats Defense
Commerzbank AG, Frankfurt am Main
`igor.podebrad@commerzbank.com`

² Computer Engineering, Department of Electrical Engineering
Helmut Schmidt University, Hamburg
`{marcel.eckert,bernd.klauer}@hsu-hh.de`

Abstract. This paper presents an approach to prevent memory attacks enabled by DMA. DMA is a technique that is frequently used to release processors from simple memory transfers. DMA transfers are usually performed during idle times of the bus. A disadvantage of DMA transfers is that they are primarily unsupervised by anti malware agents. After the completion of a DMA activity the transferred data can be scanned for malicious codes. At this time the malicious structures are already in the memory and processor time is necessary to perform a malware scan. The approach presented in this paper enhances the DMA by a watchdog mechanisms that scans the data passing by and interrupts the processor after the detection of a malicious data or instruction sequence. Configurable hardware based on FPGAs is used to overcome the problem of frequently changing malware and malware signatures.

Keywords: Hardware Security, FPGA, Direct Memory Access, Malware.

1 Introduction

The security of modern computing systems is mainly based on software, such as anti malware agents or intrusion detection systems. Modern attack vectors consider hardware and software leaks for intrusion purposes. The main target today is software on all software abstraction levels in a computer. Hardware as a target is also moving into the focus of the "Dark Side" (configurable hardware is also infectable hardware) but this fact is not focus of this paper. Although traditional (software based) anti malware approaches are improving daily by enhanced snooping procedures and new malware signatures, Rutkowska has shown, that attacks are possible, which are undetectable by software [4].

Especially exploits taking advantage from the Direct Memory Access unit (DMA) showed the simplicity of malware injections directly into the main memory of computing systems, bypassing all software based security mechanisms. Rutkowskas attacks imposingly show, that the hardware engineering paradigm

"Software writers should provide security; Hardware should just be as fast as possible" [2] is definitively outdated. To design secure systems in future, security needs to become an issue to be addressed on all abstraction levels of computing systems.

A general introduction into the field of hardware based security is given in [6], with an excellent elaboration on DMA and related components like memory and interfaces in chapter 5. A comprehensive analysis of related work in the area of processor security is given in [1].

In the remainder of this paper we present our hardware based security enhancement for the DMA-functionality as an example for how to cut down selected hardware originated attack vectors. It's basic functionality is a tamper-proof hardware based, highly parallel executed snooping functionality on the data bus, that scans for signatures of malicious code structures interrupting if a signature is found.

2 Problem

DMA is a well known technique to release processors from time consuming workload caused by simple data transfers. The transfers are performed without supervision. Data and instruction are communicated between the memory as sink and source or between the memory and mass storage or interfaces. They reach their destination block-wise completely before being checked by anti malware agents. The DMA reports DMA completions by setting bits in status registers by interrupting the processor or by other status signals [5]. An anti malware agent can then check the result if the memory was the DMA target.

With the unsupervised DMA transfers stealth features can be implemented. Malicious code and data can be transferred. To complete success the attack pattern needs to launch the code before it has been checked by the anti malware agent.

Another opportunity to exploit the unsupervised DMAs is to infect the anti malware software directly.

3 Solution

3.1 Concept

The vulnerability of a system, based on the security leak imposed by the DMA-functionality will be solved by the introduction of a DMA-Watchdog (watchdog). The watchdog itself resides between the DMA-controller and the memory controller (of the main memory). The watchdog supervises the data part of the memory-bus with a number x of sensors (S_0 to S_x in figure 1).

The sensors provide a pattern matching functionality to identify malware. The detection algorithm of a single sensor can be complex in any order and is variable in principle. If one of the sensors is detecting his pattern it is signaling to the watchdog. The watchdog itself will now block the current DMA-transfer

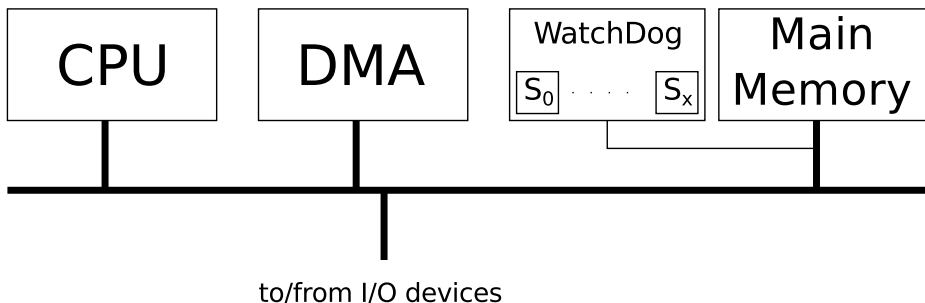


Fig. 1. watchdog residing between DMA-controller and main memory

and signals the processor a "bad" transfer where it can be handed over to the operating system and appropriate software (e.g. anti virus software).

The detection of different patterns is possible by the different sensors and is performed in parallel. For proving the effectiveness of our solution, the implemented proof of concept demonstrator is presented in the next section.

3.2 Proof of Concept

This section presents the proof of concept demonstrator for an hardware based DMA-Watchdog as described in the previous section. It is based on the *Partially Reconfigurable Heterogeneous System (PRHS)* framework as shown in subsection 3.2. The purpose of the proof on concept demonstrator is to prove the following theses:

1. A hardware based DMA-watchdog is able to detect malware infected DMA-to-Memory transfers.
2. There is no performance loss for the system processor if a hardware based DMA-watchdog is used.

PRHS Framework. Before starting to work on hardware based DMA-watchdogs, it is necessary to have a freely configurable framework for investigations. Already available frameworks, e.g. Xilinx Microblaze, suffer one big problem: Their hardware is either hardwired (hardcores) or the available soft-cores have closed sources. To gain full flexibility for future research and development, it was necessary to have control even over the configuration and architecture of the hardware. Therefore the *PRHS* framework has been developed. It is presented in the remainder of this section.

The framework has been designed for *Field Programmable Gate Array (FPGA)* usage only, with the intention to have a platform for research and education. Hence it consists of several modules allowing reuse and adaptivity for different *FPGAs* and Boards (it has already been used for Spartan3, Virtex5, Spartan6, Virtex6, Virtex7 *FPGAs* and evaluation boards hosting such devices (i.e. ML505,

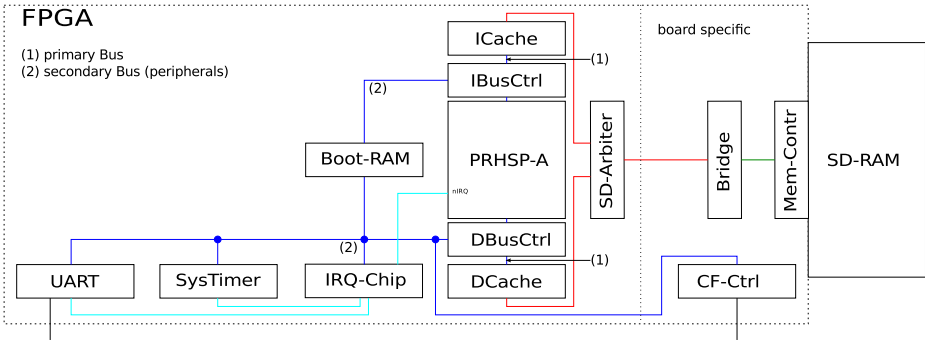


Fig. 2. schematic overview for base system of PRHS framework (hardware, detailed explanation of components is given in section 3.2)

ML605, VC707). All necessary hardware-sources (VHDL) have been implemented from scratch. Therefore the system is completely open. The software parts are adapted Open Source projects (Linux as OS, gcc and ucLibC for Cross Compiler toolchain) and therefore also fully accessible.

Hardware. For allowing flexibility, the *PRHS* framework comes with three fundamental systems:

- Small system: Combines the *PRHS Core - ARM Instruction Set (PRHSC-A)* with Block-Ram and an UART. This might serve as starting point for small embedded systems.
- Base system: Extends the small system to be able to run a customized Linux (*Linux for PRHS (L4PRHS)*) including an SD-RAM interface. Figure 2 gives a schematic overview.
- Reconfiguration system: Extends the base system with a partial reconfigurable area. (This feature is not used for the DMA-watchdog proof of concept demonstrator)

Description of the board independent components:

PRHSp-A. A self implemented processor, consisting of a core, which is instruction set compatible with the ARM8 instruction set, a system co-processor and a memory management unit.

Boot-Ram. On-chip Block Ram that contains the stage 1 boot-loader.

ICACHE/DCACHE. Instruction/Data Caches, different implementation variants exist, ranging from a simple bridge mechanism to a 32k Cache.

IBusCtrl/DBusCtrl. Bus controller to separate fast accesses to Caches/Memory (on primary Bus) and slow accesses to peripheral devices (on secondary Bus).

SD-Arbitrer. Bus Arbitrer to prevent mixing of SD-RAM memory access.

SysTimers. In-system programmable Timers.

UART. Provides input/output functionality over a RS232 Line.

Between the caches and the SD-Arbiters, a fast, board and *FPGA* independent protocol for SD-RAM access is implemented. It is mapped by board specific components to the appropriate SD-RAM protocol of a board.

Description of the board specific components:

Bridge. This device maps the *FPGA* independent SD-RAM protocol to the appropriate board specific SD-RAM protocol.

Mem-Contr. Board specific SD-RAM memory controller.

CF-Ctrl. Board specific (Compact) Flash controller. This device implements harddrive functionality.

Software. The *PRHS* framework also contains a software part including the following components:

L4PRHS An adapted Linux kernel (version 3.8), including all device driver modules to run properly on the hardware presented in the previous section.

cross compiler toolchain gcc based toolchain (including uClibc as Standard Library) to compile the adapted Linux kernel and develop software for the Framework on a high-level language base.

3.3 Proof of Concept Demonstrator

The general architecture for our proof of concept demonstrator, based on the base system of the *PRHS* framework is given in figure 3.

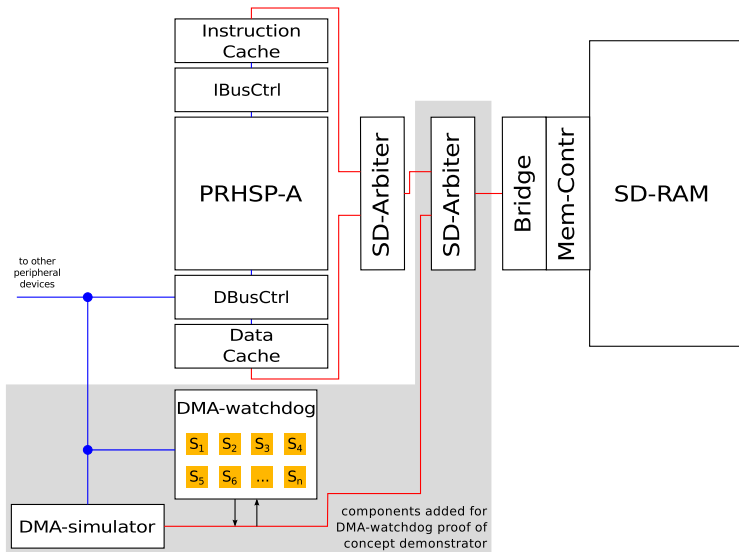


Fig. 3. General architecture for proof of concept demonstrator

To simulate and initiate DMA-functionality, a new device (DMA-simulator) has been added to the secondary data bus of the system. It is also connected to RAM via an SD-Arbitrer. Its task is to copy the content of internal Block-Ram (8kByte, contents programmable via software) to a given RAM address (programmable via software) on demand. A write request to RAM contains 256 Bits (32 Bytes) of Data. The 8 Kbyte content of the Block RAM is therefore transferred with 256 write requests to RAM.

A DMA-watchdog as described in the previous section has also been implemented. The sensors implement only a very simple pattern matching algorithm without any fuzziness. It simply compares its programmed (via software) pattern (256 bit wide) with the actual data signal (256 bits wide) of the DMA-to-memory transfer.

As the watchdog and the sensors itself are working independently from the system processor by design, there is no performance loss for the processor. Nonetheless, performance measurement has been done with dhrystone[7] and ramspeed[3] benchmarks and has proven: watchdog functionality isn't consuming processor time.

Detection functionality has been tested with a proof of concept demonstrator based system including 40 sensors. The two possibilities, DMA data contains (positive test) or doesn't contain "bad" data (negative test) were extensively tested. In both cases, DMA data and sensors patterns were generated randomly. For the negative test cases, one of the sensor patterns was randomly selected and inserted at a random position in the DMA data at each iteration. For both cases 1.000.000 iterations were carried out. In all iterations of both cases, the watchdog worked correctly.

4 Evaluation

Our proposal has shown, that a security mechanism can be implemented on a very low level, which doesn't cost any processor performance by design, because it doesn't use any computational resources of the processor. The design of the memory controller and the DMA-controller are also not affected by the introduction of a DMA-watchdog. Nevertheless, additional resources are necessary to implement the sensors and the watchdog, which can be seen as a very special kind of (co)-processor, consuming their own processing time.

The introduced approach is able to cut down only a certain part of the overall attack tree, namely the hardware related ones. This means that a hardware based solution is powerful but definitely not able to fulfill all requirements concerning detection at all levels. Nonetheless with our solution we are for the first time in the position to detect malicious code structures at a very low level in a tamper-proof way.

The proposed solution is "signature" based and therefore lacks the problem of actuality. Additionally, the number of detectable "signatures" is limited by the implemented number of sensors.

As mentioned in section 3 it would also be imaginable to implement a more complex behaviour based detection mechanism [8]. In this case, the sensors will

only signal at the "end" of a behavior. This results in data already transferred to RAM, but the entire transfer, or more exactly the "final" transfer will be prevented.

The current solution lacks the problem of a static detection algorithm. This is related to the hardware implementation of this algorithms. Some already available commercial systems like Convey HC-series or Xilinx Zynq and the academic *PRHS* framework (section 3.2) introduce the possibility to modify hardware components at system runtime via (re)configurable hardware based on *FPGA*. These technologies offers the possibility to exchange the hardware based detection algorithms for our proposed solution.

5 Conclusion and Future Work

In this paper we presented our solution for the problem of unsupervised DMA-transfers. The proposed solution introduced a hardware based DMA-watchdog. A proof of concept demonstrator, based on an ARM-system running Linux, has also been implemented to prove the feasibility of our solution and to underline one of the big advantages: this approach doesn't consume processor performance at all.

For the future we want to introduce the ability to also modify the hardware based detection algorithm by means of partial and dynamic reconfiguration. This might include also a behavior based detection algorithm.

References

1. Chhabra, S., Solihin, Y., Lal, R., Hoekstra, M.: An analysis of secure processor architectures. *Transactions on Computational Science* 7, 101–121 (2010)
2. Gueron, S., Stronqin, G., Seifert, J.-P., Chiou, D., Sendag, R., Yi, J.J.: Where does security stand? new vulnerabilities vs. trusted computing. *New Vulnerabilities vs. Trusted Computing* 27(6), 25–35 (2007)
3. Hollander, R.M., Bolotoff, P.V.: RAMspeed, a cache and memory benchmarking tool (2009), <http://alasilir.com/software/ramspeed/>
4. Rutkowska, J.: Beyond The CPU: Defeating Hardware Based RAM Acquisition (2009), <http://www.first.org/conference/2007/papers/rutkowska-joanna-slides.pdf>
5. Stewin, P., Bystrov, I.: Understanding dma malware. In: Flegel, U., Markatos, E., Robertson, W. (eds.) *DIMVA 2012*. LNCS, vol. 7591, pp. 21–41. Springer, Heidelberg (2013)
6. Wang, S., Ledley, R.S.: *Computer Architecture and Security: Fundamentals of Designing Secure Computer Systems*
7. Weicker, R.P.: Dhrystone: a synthetic systems programming benchmark. *Commun. ACM* 27(10), 1013–1030 (1984)
8. Ye, D., Moffie, M., Kaeli, D.: *A Benchmark Suite for BehaviorBased Security Mechanisms* (2005), <http://www.ece.neu.edu/groups/nucar/publications/SSATTM05.pdf>

Privacy Visor: Method for Preventing Face Image Detection by Using Differences in Human and Device Sensitivity

Takayuki Yamada¹, Seiichi Gohshi², and Isao Echizen¹

¹ National Institute of Informatics, Japan

² Kogakuin University, Japan

s5152331@yahoo.co.jp, gohshi@cc.kogakuin.ac.jp,

iechizen@nii.ac.jp

Abstract. A method is proposed for preventing unauthorized face image revelation through unintentional capture of facial images. Methods such as covering the face and painting particular patterns on the face effectively prevent detection of facial images but hinder face-to-face communication. The proposed method overcomes this problem through the use of a device worn on the face that transmits near-infrared signals that are picked up by camera image sensors, which makes faces in captured images undetectable. The device is similar in appearance to a pair of eyeglasses, and the signals cannot be seen by the human eye, so face-to-face communication is not hindered. Testing of a prototype "privacy visor" showed that captured facial images are sufficiently corrupted to prevent unauthorized face image revelation by face detection.

Keywords: Privacy, Unauthorized face image revelation, Face detection, Haar-like feature, Near-infrared LED.

1 Introduction

Due to the popularization of portable devices with built-in cameras and advances in social networking services and image search technologies, information such as when and where a photographed person was at the time the photograph was taken is revealed by the posting of photos online without the person's permission [1,2]. This has resulted in a greater need to protect the privacy of photographed individuals. A particularly serious problem is unauthorized face image revelation through the posting of images of people captured unintentionally and shared over the Internet. If, for example, your face or figure is unintentionally captured in an image taken by someone, and then that image is shared by posting it on a social networking site, information about where you were and when can be revealed through the face recognition process of an image retrieval service (e.g., Google Images) that can access the geographic location and shooting date and time information contained in the image's geotag, without your permission [3]. An experiment conducted at Carnegie Mellon University showed that the names of almost one-third of the people who participated could be determined by comparing the information in photographs taken of them with the information in

photographs posted on a social networking site. Furthermore, other information about some of the participants, including their interests and even their social security number, was found [4].

In this paper, we describe a method we have developed for preventing unauthorized face image revelation through facial recognition from images captured with a digital camera that does not hinder face-to-face communication. It is based on a method for preventing video recording in movie theaters using near-infrared (IR) signals [5]. No new functions need to be added to existing cameras or networking services because IR signals are used to add noise to the facial portions of captured images. We have developed a prototype wearable device (a "privacy visor") that implements this method. The device is worn on the face like a pair of eyeglasses, so the user does not have a feeling of strangeness. Near-IR light emitting diodes (LEDs) on the device are located near the eyes and nose. Prototype testing demonstrated that our method effectively prevents unauthorized face image revelation through image capture.

The next section describes previous methods for preventing unauthorized face image revelation. Section 3 describes various methods developed for detecting faces. Our proposed method for making faces in captured images undetectable is presented in Section 4, and our prototype wearable device implementing this method is described in Section 5. In Section 6, we describe our evaluation of the prototype implementation, present the results, and discuss them. We close in Section 7 with a summary of the key points.

2 Previous Methods

Methods proposed for preventing unauthorized face image revelation include hiding one's face with an unfolded shell [6] and painting particular patterns on one's face [7]. The first method physically protects the user's privacy by using material in the shape of a shell (a "Wearable Privacy Shell") that can be folded and unfolded. When folded, it functions as a fashion accessory; when unfolded, it functions as a face shield, preventing unintentional capture of the wearer's facial image. The second method prevents identification of the person by using particular coloring of the hair and special paint patterns on the face that cause facial recognition methods to fail. However, such methods interfere with face-to-face communication because they hide a large portion of the face and/or distract the attention of the person to whom the wearer is communicating.

The method we have developed for preventing unauthorized face image revelation through the unintentional capture of facial images does not hinder face-to-face communication. It is implemented in a wearable device (a privacy visor) that makes face detection impossible by irradiating near-IR signals, which do not affect human vision. They affect only the imaging devices used in cameras.

3 Face Detection

Face detection is the key to facial image processing [8] as it is the first step in facial recognition. The method most commonly used for face detection is the one reported

by Viola and Jones in 2004 [9]. The "Viola-Jones method" is based on a multi-scale detection algorithm that uses cascade composition of the Haar-like features, image integration, and a cascade architecture with strong classifiers. It achieves highly accurate and high-speed detection.

3.1 Haar-Like Features

Haar-like features are rectangular image features used for object recognition. Figure 1 shows the basic patterns of Haar-like features. The black areas represent dark features (negative areas), and the white ones represent bright features (positive areas). As shown in Figure 2, Haar-like features are superposed on a detection area, and their values are calculated by subtracting the average of the pixel values in the black areas $s(r_2)$ from that of those in the white areas $s(r_1)$ for each detection area. That is, the value of the Haar-like features $h(r_1, r_2)$ is given by

$$h(r_1, r_2) = s(r_1) - s(r_2). \quad (1)$$

Changing the position and size of the basic patterns of the Haar-like features in a detection area makes various Haar-like features applicable to the detection area.

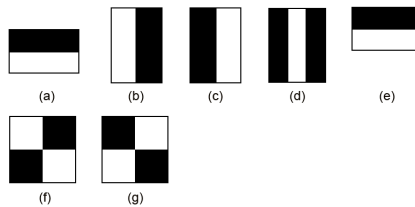


Fig. 1. Basic patterns of Haar-like features

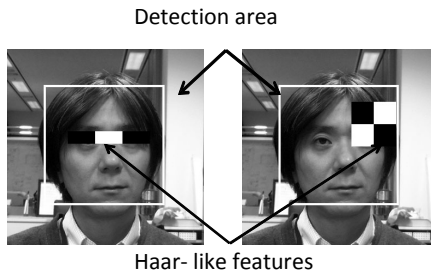


Fig. 2. Example of Haar-like features in detection area

3.2 Cascade Architecture

A weak classifier is composed of several different Haar-like features. It calculates the value of a given feature, and the value is compared with a threshold value. A strong classifier comprises various weak classifiers. Strong classifiers are arranged in a cascade architecture in order of complexity, as shown in Figure 3. The composition of

the weak classifiers and the connection order of the strong classifiers are determined in advance by supervised learning using positive (facial) and negative (non-facial) images. Haar-like features effective for face detection are chosen by supervised learning. As shown in Figure 3, a strong classifier determines "1: True" or "0: False" for each detection area. In the case of "False," the process is terminated and then restarted for the next detection area. In the case of "True" for the Nth strong classifier, the detection area is identified as a face candidate.

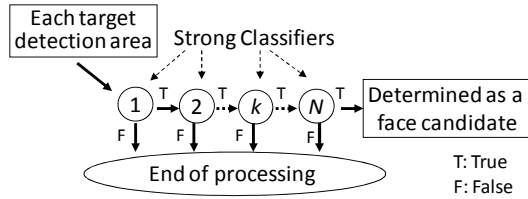


Fig. 3. Cascade architecture

4 Proposed Method

The proposed method for preventing unauthorized face image revelation through unintentional capture of facial images does not interfere with face-to-face communication in physical space. The near-IR signals used to add noise to the facial portion of a captured image cannot be seen by the human eye and hence do not hinder face-to-face communication. Moreover, no new functions need to be added to existing cameras and/or networking services.

4.1 Principle

Our proposed method is based on the difference between human sensory perception and recording device characteristics. It uses near-IR signals to corrupt images captured by a charge-coupled device (CCD) or CMOS device. According to the International Commission on Illumination (CIE), the wavelength of visible light ranges from 380 to 780 nm [10]. In contrast, the wavelengths that can be picked up by such image sensor devices as the CCDs and CMOS devices used in digital cameras and camcorders range from 200 to 1100 nm. This ability to pick up wavelengths outside the visible range gives digital camcorders the high level of luminous sensitivity needed for shooting in the dark [11].

Our proposed method adds a noise signal corresponding to the near-IR signals between 800 and 1000 nm, which people cannot see but to which sensor devices react. This noise signal is generated by LEDs located near the eyes and nose so that it prevents face detection, the first step in the face recognition process. In particular, the noise signal distorts the Haar-like features [9] around the eyes and nose, which are used in the face detection process. This means that new functions do not need to be added to cameras, social networking services, or image retrieval services. Our purpose is to establish a method that prevents identification of a person without causing

physical discomfort to the user. We do this by irradiating near-IR signals from near a person's eyes and nose that react with only the imaging device in a camera, thereby adding noise to captured facial images. These signals do not affect the person's vision but do cause facial detection misjudgment. The near-IR irradiation from near the eyes and nose can be prototyped by implementing a near-IR light source in a pair of glasses or goggles, which is something that is commonly worn, as a noise source. The means we propose for achieving our purpose is a wearable device (a privacy visor) in the shape of goggles that incorporates near-IR LEDs. The transmitted near-IR signals act as a noise source, which makes the face in captured images undetectable.

4.2 Arrangement of Near-IR LEDs

The near-IR LEDs must be effectively arranged to prevent the strong and weak classifiers from classifying the input image as an object. To interfere with the weak classifiers, it is necessary to change the difference in luminance between the positive and negative areas of each feature. We analyzed the Haar-like features effective in face detection by using supervised learning and determined into which portion of the face a noise light source should be arranged.

To determine the composition of the Haar-like features to be trained by supervised learning, we used an Open CV [12] example cascade that had been trained in advance by using 5000 facial and 3000 non-facial images [13]. By setting the pixel value of the positive area r_1 to +1 and setting the pixel value of the negative area r_2 to -1, we identified the partial regions that have a large absolute value and determined the part where the effect of face detection using the change in the luminosity value is large. The superposition of the Haar-like features as determined by using the first strong classifier ($k=1$) and by using the 10-th strong classifier ($k=10$) is shown in Figure 4. A positive area representing bright features is concentrated on the circumference of the nose, and a negative area representing dark features is concentrated on the circumference of the eyes and nose. To make face detection fail, we have to make the negative area bright or make the positive area dark so that features are obscured. Near-IR LEDs can make a dark area bright. We therefore focus on the negative area.

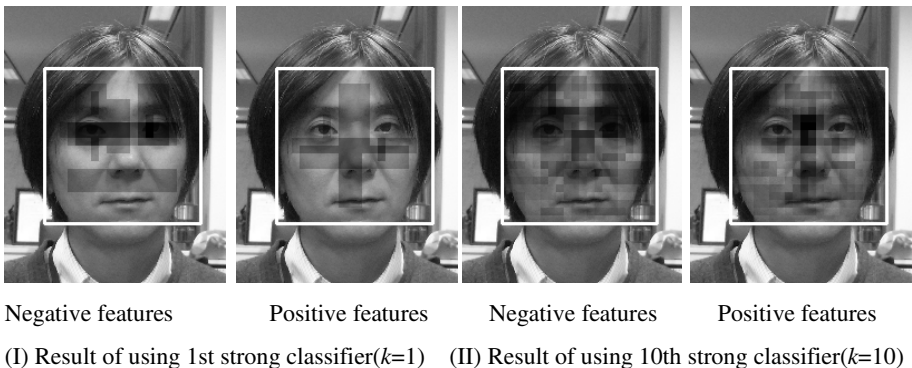


Fig. 4. Superposition of Haar-like features in detection area

By considering the combination of the negative area with the area where a device with a goggle form could be worn, we determined that the near-IR LEDs should be arranged around the eyes and along the periphery of the nose bridge.

5 Prototype

5.1 Description

An overview of the prototype privacy visor is shown in Figure 5, and the specifications are listed in Table 1. The prototype is a pair of commercial goggles to which near-IR LEDs have been attached. The LEDs have a peak wavelength of 870 nm and are positioned so as to maximize the distortion of the Haar-like features.

Table 1. Specifications of prototype privacy visor

Near-IR LEDs	Type: Chip-type with lens; Number: 11; Peak wavelength: 870 nm; Radiation intensity: 600 mW/sr; Radiation angle: $\pm 15^\circ$; Rated current: 1 A; Rated power consumption: 2.1 W
Goggles	Materials: Plastic frame; Polycarbonate lenses
Power	Lithium-ion battery (3.7 V \times 3) 2000 mA/h

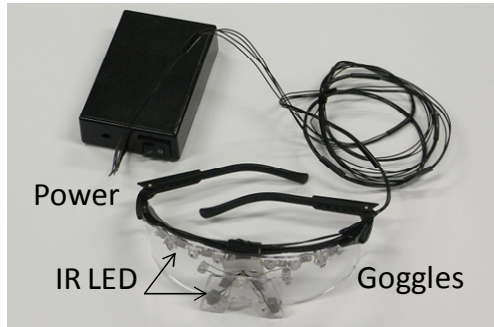


Fig. 5. Overview of prototype privacy visor

5.2 Configuration

Near-IR LEDs

We used chip-type-with-lens near-infrared LEDs based on the sensitivity of human eye and imaging sensor device characteristics by considering the basic types of LEDs (bullet, reflective, and chip with lens), the peak wavelength, the spectral width, etc.

Goggles

We used goggles with a plastic frame and polycarbonate lenses because goggles facilitate the "attachment" and "detachment" of near-IR LEDs to the human body. To effectively interfere with face recognition and maximize the noise effect in captured images, the noise light source must be carefully placed on the face. Because face detection uses the Haar-like features of several areas on the face, face detection cannot be prevented by simply wearing sunglasses. Therefore, to distort the large Haar-like features around the eyes and along the bridge of the nose, we attached 11 near-IR LEDs to commercial goggles on the basis of the analysis results described in Section 4.2. They were positioned around the eyes (3 above each eye; 1 on the inside of each eye) and around the nose (1 on each side of the nose; 1 on the glabella). Because unintentional image capture can also occur from a slant as well as from the front, image capture from a slant must also be prevented. Therefore, the six LEDs above the eyes were arranged in the normal direction of the curved surface of the lenses.

An example of the privacy visor in use is shown in Figure 6. When the noise light source is turned on (right-side images in Figure 6 (i) and (ii)), the near-IR signals are picked up by the image sensor device of a camera as noise. Detection of the face is thus impossible because this added noise greatly changes the Haar-like feature. When the noise light source is turned off, the goggles revert to a form common in the physical world (left-side images in Figure 6 (i) and (ii)), and thus do not interfere with face-to-face communication in physical space.

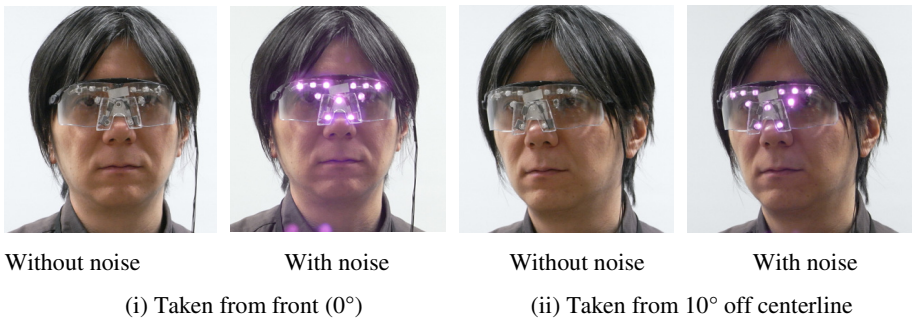


Fig. 6. Example of privacy visor in use

6 Evaluation

We evaluated our prototype privacy visor with the assistance of ten participants (age: 20–40) using a commercial digital camera (1/2.3-inch CCD; ~10 million effective pixels). The room illumination had a light intensity of 67.5 lux. The camera was set to "Camera focus: spot AF," "Photometry: multi-aperture," "Iris: f/3.3 (automatic setup)," "Exposure time: 1/10 s (automatic setup)." We took images of each participant under three conditions: (a) not wearing the privacy visor, (b) wearing it

without IR emission, and (c) wearing it with IR emission. They were taken from three directions (0° , 10° off centerline, and 20° off centerline) and at distances of 1–20 m.

6.1 Method

The images were evaluated using an Open CV face detection API and the same example [7] used to consider the arrangement of the near-IR LEDs. The images had a resolution of 3264×2448 pixels, and the cascade was used to set the detection area to various sizes (from 20×20 pixels to the maximum number of pixels so that it fit into the image, magnified by a scale factor of 1.1) and to various positions (a loop stride of program is at least two pixels, determined by the scale) from corner to corner.

A detection area classified as an object by all strong classifiers was considered to be a face candidate. After all face candidates were detected, a single candidate was focused on, and the number of neighbor candidates M , that had a size different from and were included in the focused on candidate was counted. If M was two or more, it was determined that there was a face in the detection area, and the face was detected. If M was less than two, especially if the scales of the included candidate were not continuous, it was determined that there was not a face in the detection area, and a face was not detected.

6.2 Results

The face detection results for the images taken from the closest distance (1 m) are shown in Figure 7. Each rectangle indicates a candidate classified as an object by all the classifiers in the cascade. The different colors represent different scales. The number of people detected is plotted in Figure 8. The number of people detected increased in the order of (c), (b), and (a). The plots show that a person not wearing the privacy visor or wearing it without noise could be detected for certain directions and distances while a person wearing the privacy visor with noise could not be detected for any direction or distance. This means that the near-IR signals had a greater noise effect than the visor itself. Details of the evaluation results for each angle are given below.

Images Taken from Front

As shown in Figure 7 (i), for conditions (a) and (b), a face was detected because other candidates were included in the outer candidate for more than two continuous scales. For (c), a face was not detected because there was no candidate on the actual face. As shown in Figure 8 (i), for conditions (a) and (b), all the faces were detected when the distance was less than 16 m. When it was 16 m or more, the number of faces detected decreased moderately. For (c), no face was detected at any distance.

Images Taken from 20° off Centerline

As shown in Figure 7 (ii), for conditions (a) and (b), a face was detected because other candidates were included in the outer candidate for more than two continuous scales. For (c), a face was not detected because there was no other candidate in the

outer candidate. As shown in Figure 8 (ii), for conditions (a) and (b), all the faces were detected when the distance was less than 11 m. When it was 11 m or more, the number of faces detected decreased rapidly. For (c), no face was detected at any distance, the same as for (i). In this evaluation, the maximum slant was up to 20° off the centerline. If it is 20° or more, face detection becomes difficult.

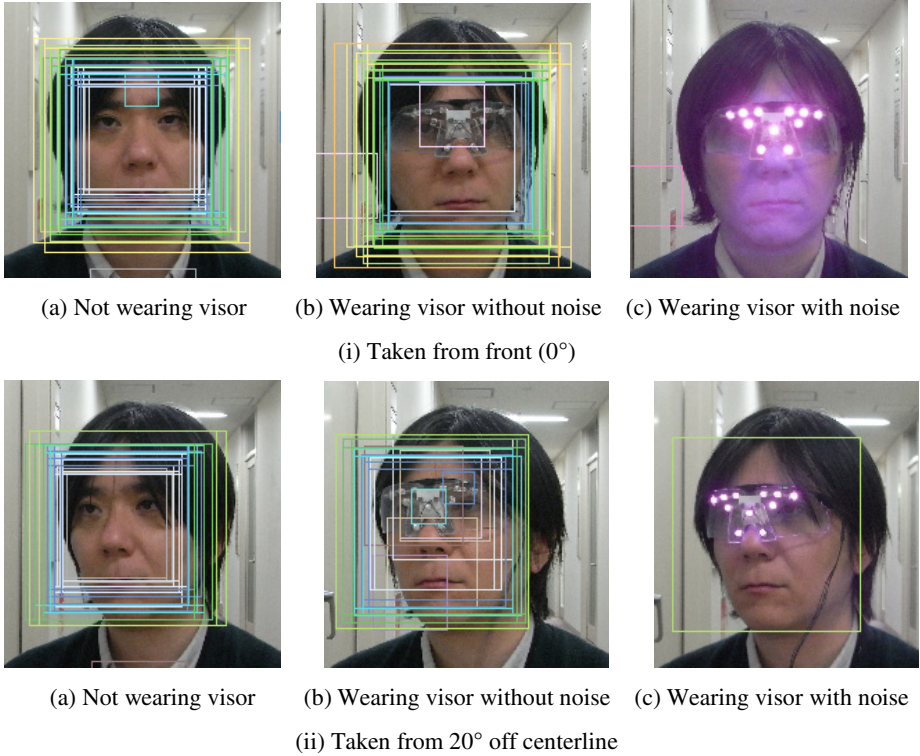


Fig. 7. Face detection results for pictures taken from 1 m

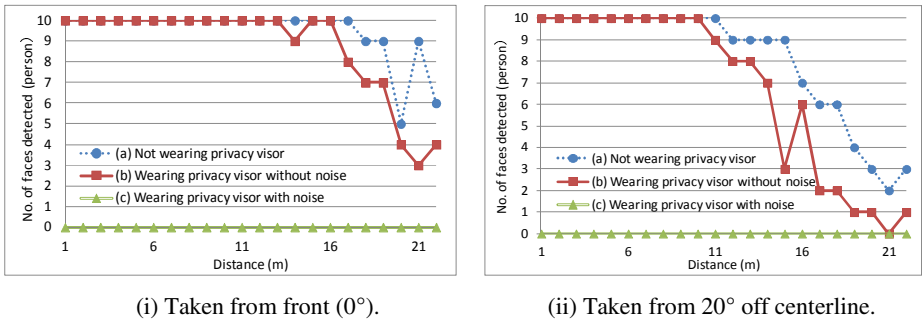


Fig. 8. Number of faces detected

7 Conclusion

The method we proposed in this paper prevents unauthorized face image revelation through unintentional capture of facial images. It adds invisible noise signals to images captured with an image sensor, thereby preventing the revelation of sensitive information via the face recognition process of image retrieval services. Specifically, our method prevents the recognition of people's faces by adding noise to imaged facial images by irradiating from near a person's eyes and nose near-IR signals that react with only the imaging device on a camera and do not affect the user's vision. These noise signals cause facial detection to fail, and facial detection is required for facial recognition. Testing of a prototype privacy visor implementing this method demonstrated that it can effectively prevent unauthorized face image revelation by interfering with the facial images, thus validating the feasibility of our proposed method.

We are now working on a method that uses absorption/reflective material and the reflection properties of light so that a power supply is not needed.

References

1. Cutillo, L., Molva, R.: Safebook: A privacy-preserving online social network leveraging on real-life trust. *IEEE Communications Magazine* 47(12), 94–101 (2009)
2. Debatin, B., Lovejoy, J., Horn, A.: Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences. *Journal of Computer-Mediated Communication* 15(1), 83–108 (2009)
3. Blackman, J.: Omniveillance, Google, Privacy in Public, and the Right to Your Digital Identity: A Tort for Recording and Disseminating an Individual's Image Over the Internet. *49 Santa Clara Law Review* 313, 341–392 (2009)
4. Face Recognition Study FAQ, <http://www.heinz.cmu.edu/~acquisti/face-recognition-study-FAQ/>
5. Yamada, T., Gohshi, S., Echizen, I.: Preventing re-recording based on difference between sensory perceptions of humans and devices. In: *Proc. of the 17th International Conference on Image Processing, ICIP 2010*, pp. 993–996 (2010)
6. GAIA, VEASYBLE, <http://www.veasyble.com/index.html>
7. Harvey, A.: CV Dazzle, <http://ahprojects.com/projects/cv-dazzle>
8. Feris, R.S., de Campos, T.E., Cesar Jr., R.M.: Detection and Tracking of Facial Features in Video Sequences. In: Cairó, O., Cantú, F.J. (eds.) *MICAI 2000*. LNCS, vol. 1793, pp. 127–135. Springer, Heidelberg (2000)
9. Viola, P., Jones, M.: Robust Real-Time Face Detection. *International Journal of Computer Vision (IJCV)* 57(2), 134–157 (2004)
10. Schanda, J. (ed.): *Colorimetry: Understanding the CIE System*. Wiley-Interscience (2007)
11. Holst, G., Lomheim, T.: *CMOS/CCD Sensors and Camera Systems*. SPIE-International Society for Optical Engine (2007)
12. Bradski, G., Kaehler, A.: *Learning Open CV Computer Vision with the Open CV Library*. O'Reilly Media (2008)
13. Lienhart, R., Kuranov, A., Pisarevsky, V.: Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. In: Michaelis, B., Krell, G. (eds.) *DAGM 2003*. LNCS, vol. 2781, pp. 297–304. Springer, Heidelberg (2003)

E-Learning of IT Security Threats: A Game Prototype for Children

Jana Fruth, Carsten Schulze, Marleen Rohde, and Jana Dittmann

Otto-von-Guericke-University of Magdeburg
P.O. Box 4120, 39016 Magdeburg, Germany
{fruth,dittmann}@ovgu.de,
{carsten.schulze,marleen.rohde}@st.ovgu.de

Abstract. In this paper an e-learning game prototype for primary school children (aged between 7 and 9 years) is introduced. The game teaches children about IT security threats, which they encounter using the Internet. The game is separated into three mini games: virus infection of the computer, inviting somebody in social networks, chatting with strangers. The game design used metaphors and based on standard guidelines of infantile learning environments (e.g. paradigm of simplicity, multidimensional stimuli, characters). Furthermore, the results of a user study of 36 primary school children are presented. In the future, the prototype would be extended by additional metaphors.

Keywords: e-learning, IT security threats, internet, game, children.

1 Introduction and Motivation

Nowadays more and more children are using personal computers and the Internet. More than half of the primary-school pupils (6 to 10 years old) use these technologies regularly [1]. Surfing the Internet may expose children to many threats and risks, such as malicious codes infecting their personal computers, or their personal data being spied on [1, 2]. There are various concepts to raise children's awareness to IT security threats. Most parents set rules for their children's Internet usage (e.g. time limits). At school children are taught about the proper usage of personal computers by their teachers in computer science or IT classes. Furthermore, many different initiatives and websites offer information for both the parents and their children about IT security threats (see Section 2). However, according to research results and personal experiences those concepts are not reaching the children [1, 2]. To overcome this problem this article introduces an educational game prototype for children based on subjects of IT security. This method for raising children's awareness can also be adopted for safety (physical integrity) related subjects. The educational game prototype can be extended to cover subjects referring to safety related aspects of mobile toys like robots. The article is structured as follows: in Section 2 a short overview of the state of the art of methods for raising the awareness to safety and IT security related topics are illustrated. In Section 3 the educational game prototype "InSiKids" (engl.

“Internet Security for Kids”) mentioned above is introduced. In Section 4 the results of the usability test that was performed with the prototype “InSiKids” is presented and discussed. The success of intermediation of IT security threats metaphors to primary school children and the confirmation of assumptions is checked. Section 5 concludes the paper and shows future prospects.

2 State of the Art

In this section the state of the art of children’s psychology of learning, playing games, using the Internet are described. Furthermore, current techniques of computer game development and current IT security awareness raising methods for children are illustrated.

2.1 Children’s Psychology

How children learn: Remo H. Largo, the Swiss paediatrician, says that children learn in different ways [3]: *Social learning:* The child imitates the behaviour of role models, like parents and other children. *Learning by experiences with the objective environment:* The child becomes acquainted with its environment by occupying itself with objectives via its motion activities and senses. So the child develops for example a comprehension for the dimension, shape or color of different objects. *Learning by education:* According to Largo the learning opportunities should be adapted to children’s development-specific interests. Ideally, teachers creates the learning environment in such a way, that a child is able to gain experience and new comprehension on its own. This will be successful if a child has the comprehension for this specific learning assignment.

Why children play games: The terms ‘play’ and ‘game’ [4, 5] need to be differentiated. Both, play and games are guided by rules, while rules of play (e.g. fantasy play) are flexible, games (e.g. basketball) are governed by explicit rules, which are not negotiable [6]. Various researchers [4, 7] claim, that play is being essential for children’s healthy development. It’s a cornerstone of children’s development of cognitive, social skills, and a fundament to learn higher complex concepts if children are elder.

How they play games: *Conventional games:* Children in the primary school age prefer various games. Examples are sports (e.g. cycling), and traditional games (e.g. hide and seek). Favourite games of girls include verbal games, role playing, play with dolls [8]. Boys often play construction games and games [8] involving physically activities, like ball games [6]. *Computer and online games:* Among conventional games, computer games are an inherent part of primary school pupils leisure time. About 13% of children between 6 and 9 years play computer games every day, about 40% of them play regularly (once and/or several times a week) [9]. The majority of children in this age play not longer than one hour. The most used device are portable games consoles, such as Nintendo DS. Online games are not so common in this age. Only 15% plays online games once a week. Offline games are more common. Nearly half of the children in this

age play them. The favourite games differ between boys and girls. Boys often play ‘FIFA’, ‘Mario Kart’, and ‘Pokémon’. The favourite games for girls are ‘The Sims’, ‘Singstar’, and ‘Wii Sports’.

How and why children use the Internet: Aloud a survey throughout the EU [2] 60% of children aged between 6 and 9 years use the Internet. Their favourite Internet activities during their leisure time are surfing the Internet, viewing websites for children, watching films and videos online. It’s not so common that peers in this age communicate via social networks. Girls are more active in social networks than boys. Only 5% of the 6 to 7 years old and 13% of the 8 to 9 years old are a member of a social network community [9]. Furthermore more often, primary school children are invited by their teachers, to use the Internet to do their homework.

2.2 Development of Computer Games for Children

Modern game development is an iterative process [10]: firstly game ideas are generated and formalised, afterwards the game is tested and test results are evaluated. The iterative process has to be repeated, if the evaluation identifies some problems with the game design. A modern game design should have six core elements [11]: challenge, goals, rewards, rules, interactivity, and decision making. Game designers usually distinguish between demographic groups, differ in age and gender [12]. Amongst other factors, specific skills of a demographic group define a user specific game design. The group of ‘kids’, children aged between 7 and 9 years, are very interested in computer game playing, usually have reading skills, and start logical thought. The challenge for a game designer for those kids is to avoid overwhelming them with too much information (see passage “Children as Users”). For the development of computer games programmer could choose various game engines¹. A common game engine is the XNA Game Studio 4.0 [13]. It’s a programming environment provided by Microsoft, which includes the XNA Framework. It allows an easy game development of small game projects, with a comparative small implementation effort in comparison to standard game implementations. Amongst other functions XNA provides the window management, the display of 2D and 3D graphics, the handling of user inputs (keyboard, game controller) and the output of sound. The game described in Chapter 3 is developed using the XNA framework because of the easy way of 2D game implementation. In our opinion the use of 2D games with a simple visual design are adequate to allow learning without distracting children.

Children as Users: The way children think differs a lot from adults. Children in general are used to thinking in a world of fantasy and dream of magic [14]. Furthermore, young children have problems reading long and especially complicated texts [15]. Therefore, texts should be short, easy to understand and the information has to be limited. If the children are overwhelmed with too much information, they will easily feel frustrated, lose their concentration

¹ <http://www.indiedb.com/engines>, last access: 14. June 2013.

on the task at hand [16]. To support the learning process of children the use of multimedia is appropriate [14]. The use of metaphors ensures that the children will be able to understand the complex information. Those metaphors should wrap the complex information (IT security) into something the children know from their daily life [15]. The prototype “InSiKids” realises this metaphorical approach for a target audience of primary school pupils aged between 7 and 9 years Section 3. The game prototype was evaluated with test methods adapted to children (“thinking aloud technique”, “active intervention method” and “retrospection” method) [17]. But gender particularities must be observed. Research results of the developmental psychology validate developmental differences in cognitive skills between girls and boys [16] Section 4. Girls in comparison to boys tend to have better verbal skills (e.g. spelling, writing, linguistic understanding), while boys tend to have an affinity for technics resulting in comparatively more interest in the functionality of technical devices [16].

2.3 Awareness Methods to IT Security Threats for Children

To raise the children’s awareness to problems and questions of IT security while using the Internet various procurement methods exist:

Parents and school: IT security is a complex subject, which children mainly learn about by asking their parents and friends (see peers) or while taking computer science / IT classes at school [1, 2]. Most parents arrange specific rules with their children, e.g. time limits for using the Internet [18]. At school the children are often instructed only on how to use computers, but rarely the risks and threats they can encounter [1]. **Peers:** Children learn while interacting with their friends (peers). The interaction with qualified others is essential in learning new things [19]. **Initiatives:** Many initiatives have been formed to convey the crucial knowledge about IT security to children as well as their parents. Initiatives like the website ‘klickSafe’² focus on increasing the awareness of Internet users in general to possible threats and conveying the appropriate behaviour in such critical situations. The website ‘fragFinn.de’³ is a *web search engine* especially developed for children, which provides a safe way searching the internet. Another approach to convey this subject to children as the target audience is the use of websites containing *games, comics and quizzes*. This approach is relatively wide spread since it takes advantage of things children like, such as ‘Sheeplive’⁴. Even though these subjects have been widely discussed before the stated projects seem to fail to convey the crucial information to their target audience [1, 2]. Therefore, new approaches have to be taken. The e-learning game prototype, introduced in Section 3, is a new approach to teach children IT security threats.

² www.klicksafe.de, last access: 14. June 2013.

³ www.fragfinn.de, last access: 14. June 2013.

⁴ at.sheeplive.eu, last access: 14. June 2013.

3 Game Prototype: E-Learning of IT Security for Children

Based on standard guidelines of infantile learning environments (e.g. paradigm of simplicity, multidimensional stimuli, characters, simple descriptions) the game prototype was developed [14, 15]. Key feature of this prototype are the utilized metaphors for the security threats, which should relate the abstract threats to known everyday situations for the children.

User specific game design: In the game the children become part of the company of hero characters. Every character stands for a mini-game about a particular security threat (chatting, publishing of personal information, malicious codes) (see Table 1). A mascot provides advice and explanation through the game. The text was presented in speech bubbles as in comic or manga with a big font size. The mini-games were keep short to prevent exhaustion.



Fig. 1. Main menu



Fig. 2. Social network game

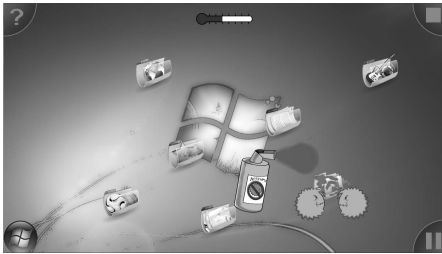


Fig. 3. Virus game

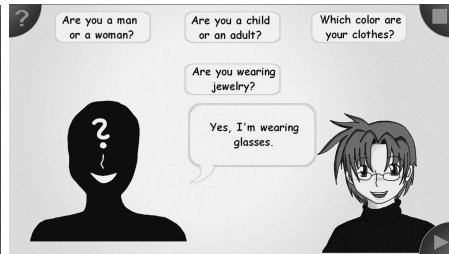


Fig. 4. Chat room game

Technical description of game prototype: The game is implemented with hand drawn two dimension (2D) sprites⁵ and backgrounds. In the field of game development sprites are an established technique, which allows the easy creation and handling of geometries in comparison to other techniques. Only point and click user interaction method was implemented to keep the interaction simple.

⁵ In the field of game development a “sprite” is defined as, an 2D image or an animation, which is included into a larger game scene [11].

The user interface is designed using the well-known icons from music players. If the infantile player successfully finishes a mini game, he earns a medal, which is placed on the character icon button in the main menu. Furthermore, children can easily navigate between the mini games through the corresponding icon of the character or through scrolling with the arrow buttons Figure 1 on page 166. In the following two sections the creation and realisation of the metaphors is further described. **Design of IT Security Threat Metaphors:** Metaphors are essential for teaching children about abstract security threats and appropriate precautions. Children cannot associate a computer virus, as they do not touch or see (problem of non-suitable warning messages of virus scanner for children) the threat to their security and safety. Therefore, the goal is to develop suitable metaphors to form proper environment for the children to raise their awareness of security threats. In addition, the metaphor must be chosen age-based for maximum learning success but not every metaphor can be realized in an educational game. Using the Internet children are threatened by the same threats as adults but children are more susceptible to threats because of their stage of development. Thus the threats in Table 1 are gathered without limitation. Afterwards, the threats are rated on the basis of media use of children [9, 20] and the following three are chosen to be implemented in the educational game: Social network: publishing of personal data, Chat rooms: strangers in the Internet, and Malicious codes: viruses. **Realisation of Metaphors in the Game Prototype:** To generate meta-phors for threats, the appropriate precautions are collected and set into the age-based context. Around the threat and precaution a metaphor is developed to provide the children with situations they could relate to. At the beginning of the mini games the mascot gives a short introduction to the situation of the character and provides the player with a task. The children are provided with additional information about the consequences of their choices, e.g. if you misplace your party invitations, the party could be over crowded with strangers. The *social network game* is designed as visual novel. The character and her conscience embodied by little devil and little angel propose different solutions to place the birthday party invitations in school Figure 2 on page 166 a. In the *virus game* the children should react to a virus infection which was displayed as desktop-icon eating monsters Figure 3 on page 166. They could try to use a spray to distroy the viruses, but these keep spawning until all data, which is symbolised by the desktop icons being destroyed. The solution to win is to shut down the system. The *chat room game* was designed as a memory game. Five characters are presented to the children, so they could memorize them. Afterwards they had to interview a randomly chosen out of the five and then guess which the chosen character was Figure 4 on page 166.

4 User Study

The educational game prototype “InSiKids” is designed for primary-school pupils aged between 7 and 9. To evaluate the knowledge transfer via the games and the knowledge, which is present prior to the test, an online questionnaire developed

by using the LimeSurvey Framework⁶ has been used. By reason of absence of a pre test with a time interval to the main test, only knowledge not learning effects could be measured. **Test environment:** The usability test of “InSiKids” took place at the trilingual international primary school in Magdeburg. The children are familiar with their school environment so the stress during the test are decreased [15]. The children could use their own personal laptops⁷ and had been gaining experience with computers for over a year. The usability test was performed within two third grade classes. The number of pupils was 17 respectively 19 in each class. The results of the test are not representative, because of the small size of the test group (36 pupils). Therefore, only tendencies for knowledge transfer effects for the game prototype could be derived from the test results.

Test Realisation: Testing the game prototype was done in different phases: **Preparation** was done by installing the game to the children’s laptops prior to the test. **Introduction:** The team and the prototype were introduced to the children. The operating concept and the test procedure in general were introduced and explained to them. The duration of this phase should not exceed 10 minutes, because children tend to lose their concentration rather quickly. **Game testing:** The children could start playing the games. They were free to start with which game they wanted to play the most. The children were only admonished to play every game at least once. This phase lasts for approximately 25 minutes. In the process of testing the team split in groups which were distributed to the two classes. Two team members stayed in each class to answer the children’s questions and take notes on their behavior while playing. **Break:** With a duration of 15 minutes for recreation. **Questionnaire:** With a duration of approximately 20 minutes the children evaluated the games via an online questionnaire. **Certificates and Conclusion:** Certificates were handed out to the children, on which they could have their names signed on and thus could join the company of heroes. Mentioning the certificates to the children resulted in a higher motivation to play the games and to learn about the presented IT security subjects.

Test Results: The test results of the evaluation of the educational game prototype using descriptive statistics [21] are presented. The evaluation results are collected by a non-standardized questionnaire, which was self-developed. The questions presented are categorized into four categories: *sociodemographic characteristics* (age, gender) and questions considering *previous knowledge* (use of different technical devices and frequency of use of the Internet), questions about the *three mini games* (virus, social network, chat rooms), and the *personal consternation* of the test persons concerning IT security threats. The test results were separated into gender groups, which can be reasoned by cognitive differences between girls and boys based on the findings of developmental psychology [16]. 34 children participated in the test. Four data sets were invalid, so 30 valid data sets (20 female, 10 male) were evaluated. To provide comparable results for

⁶ www.limesurvey.org, last access: 14. June 2013.

⁷ www.intel.com/content/www/us/en/intel-learning-series/classmatepc-convertible.html, last access: 14. June 2013.

groups differing in size the collected data was normalized. **Sociodemographic characteristics:** the girls average age was 8,4 years and the boys average age was 8,8 years. **Previous knowledge:** At home boys (60%) are predominantly use laptops while girls predominantly use personal computers (50%) instead of other technical devices. The boys (50%, 4-6 times a week) use the Internet more often than the girls (50%, 2-3 times a week). 15% of the girls do not use the Internet while all interviewed boys use the Internet between one and six times a week. In the following the results of the comprehension questions to the **three mini games** (virus, social network, chat) are presented. The questions are analysed by using a rating system. Correct answers are coded with a score of 2 (more relevant) or 1 (relevant), and incorrect answers with a score of 0. In case a child gets a score of zero in one group of questions, he could not achieve a higher score. This is based on the assumption, that the metaphors and the communication of the knowledge about IT security threats fail. It should be measured if the teaching of metaphors of IT security threats were successful. Because of the missing pre test, only knowledge effects not learning effects could be measures. Besides the illustration of test results via block diagrams, the data is also statistically analysed [21]. The *independent two-sample t-test* verifies the relationship of the means of two population on basis on the means of two independent samples. In our case, the t-test is used to identify differences between the test results of girls and boys. *Cohen's d effect size* determines the practical relevance of significant results for small samples and is defined as: <0,3 minor, 0,3-0,5 middle, >0,5 major, >0,7 strong. In the mini game **“virus”** the children could achieve an overall score of seven Figure 5 on page 170. The polynomial trend lines in Figures 5 - 7 symbolise the distribution of the sample data. In comparison to the boys, the answers of the girls are more variant. The t-test results and the minor effect size of Cohen's d (0,26) show no indication for a difference between the results of girls and boys. **“social network”** game an overall score of seven could be achieved Figure 6 on page 170. The scores of the girls as well as the boys are accumulated in the middle and higher range. In comparison to the boys in average girls achieve higher scores. The t-test results are not significant. But the middle effect size of Cohen's d (0,73) show an indication for a difference between the results of girls and boys. In the **“chat”** game an overall score of four could be achieved. In this game the distribution of the score is relatively homogeneous for both groups Figure 7 on page 170. T-test and Cohen's effect size (0,21) are not significant.

Discussion: In this section the test results are discussed and our findings are represented. The test goal were to measure if the teaching of metaphors of IT security threats to the assessed children were successful. Due to the missing of a pre test only knowledge effects could be measured instead of learning effects. The results of the virus and the chat game are no significant. Only the results for the social network game indicate a higher knowledge of girls in comparison to the boys. It can be explained by the well designed metaphor of the social network threat, which seem to address the girls more than the boys. Otherwise, in comparison to the boys, girls use more often social networks. This fact can

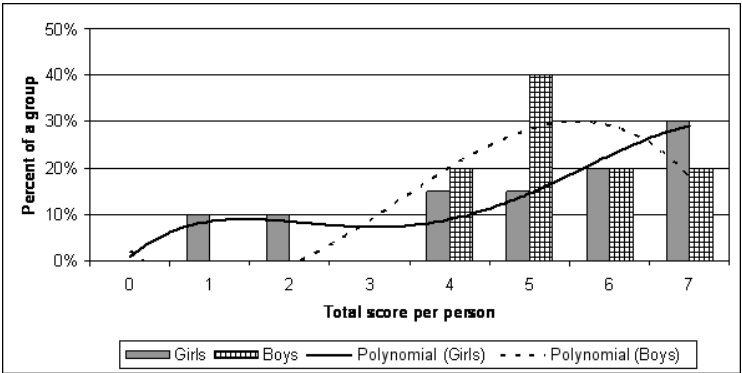


Fig. 5. Test results: computer virus game

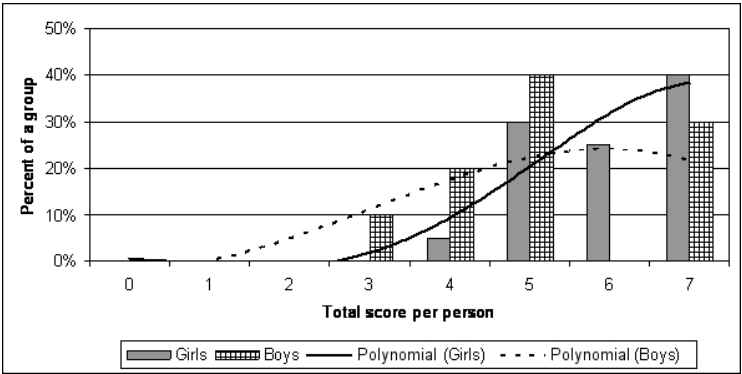


Fig. 6. Test results: social network game

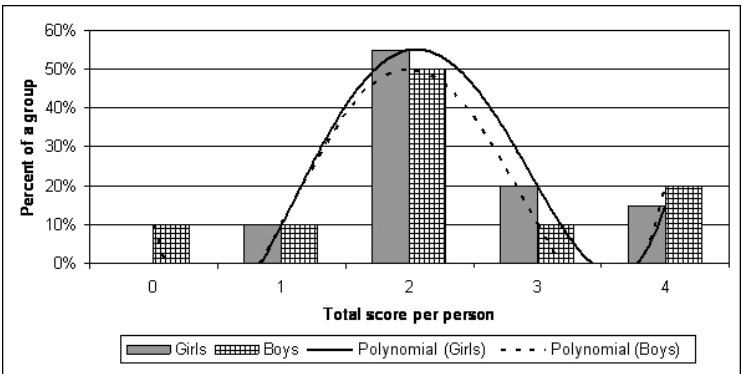


Fig. 7. Test results: chat game

be a reason for girls better results. **Lessons learnt:** *User study:* It needs a lot of planning to realise a user study. Enough time has to be planned for preparation, realisation of the test and analysis of the test results. *Test preparing:* include the search of a suitable test environment (e.g. a school), test persons (e.g. pupils), and assisting personal (e.g. class teachers). Amongst others, the information leaflets (e.g. parent's letter of agreement), the questionnaire, and the technical systems have to be prepared. *Test realisation:* From our experience, it is beneficial to have the help of persons, who are familiar with the children. They could facilitate a smooth test process. It's helpful to have sufficient personal for different tasks during the test. In our test only two people had to monitor children's behaviour and to assist pupils by playing the game. *Analysis:* Amongst other things, it has to be determined while preparing the test, what kind of property (e.g. knowledge) should be measured. For example, learning effects could be measured if a certain time period has to be elapsed between a pre test and the main test. The consultation of experts of empirical analysis, such as psychologists, could help to plan an scientific evaluation with representative results.

5 Conclusion and Future Work

The user study with 36 primary school children shows, the metaphors for exemplary IT security threats (virus infection of the computer, inviting somebody in social networks, chatting with strangers) partially support children's way of playful learning. In the future, the e-learning game prototype 'InSiKids' is to be extended and improved in *design, content and techniques*, respectively. The latter can be done in synchronizing and adding sound effects to the games to enhance the learning effect. Clues to help the children solve the tasks can be extended and a reference page to look up terms and definitions could be implemented. Due to the prototype's mini game structure more games should be easily added. The medals which can be won in the games could be extended to a bronze, silver and gold medal according to how well the children did solving the task. Additionally, new metaphors for existing and other IT security threats are to be designed and realised in the prototype. The concept of teaching of IT security threats via metaphors is to be extended and assigned for other user groups, e.g. teenagers. Furthermore, the *test setup and test realisation* is to be improved. Adapted (e.g. improved questionnaire) and improved test methods are to be used in future user studies. To consider a learning effect future studies and usability tests should be designed for long-term studies.

Acknowledgments. We want to thank the "Dreisprachige Internationale Grundschule Magdeburg", her schoolmaster Irina Horstmann, the teachers, parents and children of the third classes of the year 2012/2013, Sebastian Stellmacher, Dennis Hartmann, Michael Knuth, Volkmar Hinz, and the Acagamics e.V. Jana Fruth is funded by the German Ministry of Education and Science (BMBF), project 01IM10002A. The presented work is part of the ViERforES project [22].

References

- [1] Livingstone, S., Haddon, L., Görzig, A., Ólafsson, K.: EU Kids Online II - Final Report (2011)
- [2] Livingstone, S., Haddon, L.: EU Kids Online - Final Report (2009)
- [3] Largo, R.H., Beglinger, M.: Schülerjahre: Wie Kinder besser lernen (2010)
- [4] Piaget, J.: Play, dreams and imitation in childhood (1962)
- [5] Rubin, K.H., Fein, G.G., Vandenberg, B.: Play. Handbook of Child Psychology (1983)
- [6] Pellegrini, A.D.: The role of play in human development (2009)
- [7] Singer, D.G., Golinkoff, R.M., Hirsh-Pasek, K.: Play=learning: How play motivates and enhances children's cognitive and social-emotional growth (2006)
- [8] Herrmann Laux: Was kinder heute spielen: Anknüpfungspunkte für die schule (2009)
- [9] Behrens, P., Schmid, T., König, T., Rathgeb, T.: KIM-Studie 2010. Medienpädagogischer Forschungverbund Südwest, Stuttgart (2010)
- [10] Fullerton, T., Swain, C., Hoffman, S.: Game design workshop: A playcentric approach to creating innovative games (2008)
- [11] Perry, D., DeMaria, R.: David perry on game design: A brainstorming toolbox (2009)
- [12] Schell, J.: The art of game design: A book of lenses (2008)
- [13] Microsoft: Xna game studio 4.0 (2012)
- [14] Menzel, W., et al.: Design and evaluation of security multimedia warnings for children's smartphones (2012)
- [15] Nielsen, J.: Children's websites: Usability issues in designing for kids (2010)
- [16] Berk, L.E.: Child Development, 9 edn. Pearson (March 2012)
- [17] Kesteren, I.E., et al.: Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects, pp. 41–49. ACM Press (2003)
- [18] Kuhlmann, S., Hoppe, T., Fruth, J., Dittmann, J.: Voruntersuchungen und erste Ergebnisse zur Webseitengestaltung für die Situationsbewusste Unterstützung von Kindern in IT-Sicherheitsfragen. In: Informatik 2012, 42. Jahrestagung der Gesellschaft für Informatik, Braunschweig (2012)
- [19] Fuhrer, U.: Cultivating Minds: Identity as Meaning-Making Practice. Routledge Chapman & Hall (2004)
- [20] Behrens, P., Rathgeb, T.: JIM-Studie 2011. Medienpädagogischer Forschungverbund Südwest, Stuttgart (2011)
- [21] Howell, D.C.: Statistical Methods for Psychology, 8 edn. Wadsworth Inc. Fulfillment (January 2012)
- [22] (2013) (June 18, 2013)

Hiding Information in Social Networks from De-anonymization Attacks by Using Identity Separation

Gábor György Gulyás and Sándor Imre

Department of Networked Systems and Services,
Budapest University of Technology and Economics,
Magyar tudósok krt. 2., H-1117 Budapest, Hungary
{gulyasg, imre}@hit.bme.hu

Abstract. Social networks allow their users to mark their profile attributes, relationships as private in order to guarantee privacy, although private information get occasionally published within sanitized datasets offered to third parties, such as business partners. Today, powerful de-anonymization attacks exist that enable the finding of corresponding nodes within these datasets and public network data (e.g., crawls of other networks) solely by considering structural information. In this paper, we propose an identity management technique, namely identity separation, as a tool for hiding information from attacks aiming to achieve large-scale re-identification. By simulation experiments we compare the protective strength of identity management to the state-of-the-art attack. We show that while a large fraction of participating users are required to repel the attack, with the proper settings it is possible to effectively hide information, even for a handful of users. In addition, we propose a user-controllable method based on decoy nodes, which turn out to be successful for information hiding as at most 3.33% of hidden nodes are revealed in our experiments.

Keywords: social networks, privacy, de-anonymization, identity separation.

1 Introduction

The basic concept of online social networks is to provide an interface for managing social relationships. However, social networks are not the only services that have an underlying graph structure, and recently several network alignment attacks have been published in which attackers aimed to breach the privacy of nodes within anonymized networks (e.g., obtained for business or research purposes) by using data from other (social) networks [1–4,11]. Basically, such attacks can have two goals, i.e., to achieve node or edge privacy breach (or both). In case of the first, the attacker learns the identity of a node, or some otherwise hidden profile information, and in the second case the attacker realizes the existence of a hidden relationship.

The first attack of its kind was introduced by Narayanan and Shmatikov in 2009 [1], who proposed a structural re-identification algorithm being able to de-anonymize users at large-scale, by using data from another social network. In their main experiment they de-anonymized 30.8% of nodes being mutually present in a Twitter and a Flickr crawl. Recently it has been shown that location information can also be re-identified with similar methods [4]. As there are many services based on the graph structure (or implicitly having one), it is likely that more similar attacks will be discovered.

Attacks capable of achieving large-scale re-identification consist of two sequential phases, the global and local re-identification phase [9]. In the first phase the algorithm seeks for globally outstanding nodes (called the seeds), e.g., by their degree, and then the second phase extends the seed set in an iterative way, locally comparing nodes being connected to the seed set.

For instance, an attacker may obtain datasets as depicted on figure 1, wishing to know an otherwise inaccessible private attribute: who prefers tea or coffee (dashed or thick bordered nodes). She initializes the seed set by re-identifying (or mapping) $v_{Alice} \leftrightarrow v_7$ and $v_{Bob} \leftrightarrow v_3$ as they have globally the highest degree in both networks (global re-identification phase). Next, she looks for nodes with locally unique degree values connecting to both seeds, and picks $deg(v_{Harry}) = 3$. By looking for nodes within the common neighbors of v_3, v_7 with the same degree, she maps $v_{Harry} \leftrightarrow v_4$. Then, the process continues with additional iterations.

In this paper we propose a privacy-enhancing method related to the identity partitioning technique [8], called identity separation, to tackle these attacks. Identity separation allows a user to have multiple unlinkable profiles in the same network, which results in multiple unlinkable nodes in the sanitized graph (e.g., as the service provider is also unaware of the link between the identities).

We present effect of identity separation by the example of Fred on Fig. 1, who created two unlinkable profiles, v_8 for pretending being a coffee fan towards his closer friends (Alice, Ed, Greg), but also created v_{12} for maintaining relationships with tea lovers (Harry, Jennie). By applying the attack algorithm, it can be seen that the hidden drink preference of Fred will not be discovered by third parties.

Our main contributions are as follows. By simulation measurements we characterize how resistant the attack in [1] is against different features of identity

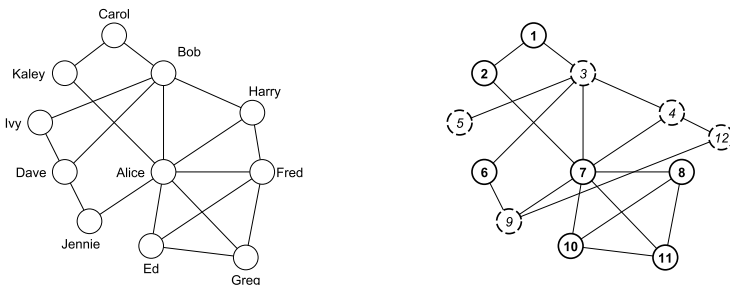


Fig. 1. Example providing insights on de-anonymization and identity separation (left: auxiliary network G_{src} ; right: sanitized network G_{tar})

separation, e.g., splitting nodes or deleting edges. In experiments we show that by using these features the quantity of information revealed by the attacker can be reduced as low as 3.21% and even lower, while identity separation cannot efficiently repel the attack on the network level. We additionally propose a method using decoys that can effectively protect privacy in a user-controllable way, for which we measure the quantity of revealed information well under 4%, even when only a few users adopt this technique.

2 Related Work

The first attack proposed by Narayanan and Shmatikov in [1] (to which we later refer as Nar09) used seeding of 4-cliques, and its local re-identification phase works similarly as described in the example of Section 1, being based on a propagation step which is iterated on the neighbors of the seed set until new nodes can be identified (already identified nodes are revisited). In each iteration, candidates for the currently inspected source node are selected from target graph nodes, sharing at least a common mapped neighbor with it. At this point the algorithm calculates a score based on cosine similarity for each candidate. If there is an outstanding candidate, a reverse match checking is executed to verify the proposed mapping from a reversed point of view. If the result of reverse checking equals the source node, a new mapping is registered.

However, since then several attacks appeared, here we include the most relevant works. Narayanan et al. in 2011 presented a specialized version of their attack [2], which was capable of achieving a higher recall rate, but was specialized for the task of working on two snapshots of the same network. More recently, Wei et al. also proposed another algorithm in [3] challenging [1]; however, we argue that their work goes beyond [1], for at least two reasons. First, in their paper there is no evaluation of their algorithm against the perturbation strategy proposed in [1], although it is definitely more realistic than what is used in [3]. As the perturbation strategy of [1] deletes edges (only), this is remarkable deficiency. Our second remark is also related to their experiments, which were performed on quite small graphs having fewer than a thousand nodes; there are no experimental results that their algorithm perform also better on networks having tens of thousands of nodes (or larger). Finally, there are some other works developing the original idea further in specific directions, such as in the case for de-anonymizing location traces by Srivatsa and Hicks [4]; however, as to the best of our knowledge no work provides better results than [1] in general, we chose this attack as the state-of-the-art, and work with it in our experiments.

For preventing de-anonymization, we consider a user centered privacy protection mechanism (instead of graph sanitization applied by the service provider), one that can be applied to existing services – otherwise one might consider using revised service models, such as distributed social networks [6]. In our previous work we analytically showed that identity separation is an effective tool against 4-clique based global re-identification [5] (as described later, we use structural identity separation models from this work).

Recently, Beato et al. proposed the friend-in-the-middle model [7], where a proxy-like nodes serve as mediators to hide connections, and presented the viability of their model (successfully) on the Slashdot network [10]. In contrast to their work, we focus also on information hiding working even for a few nodes only, and in addition, identity separation is a rather powerful method allowing a fine-grained management of information [8], e.g., it allows hiding profile information beside relationships. Lastly, we note that as network structure has a notable bias on results, we carry out experiments on multiple datasets.

3 Datasets, Modeling and Simulation Settings

We partially base our notation on the one used in [1]. Given a graph G_{tar} to be de-anonymized by using an auxiliary data source G_{src} , let $\tilde{V}_{src} \subseteq V_{src}, \tilde{V}_{tar} \subseteq V_{tar}$ denote the set of nodes mutually existing in both. Due to the presence of nodes using identity separation, ground truth information is represented by two mappings, $\mu_G : \tilde{V}_{src} \rightarrow \tilde{V}_{tar}$ denote mapping between nodes that are intact, and $\lambda_G : \tilde{V}_{src} \rightrightarrows \tilde{V}_{tar}$ denote mappings between nodes in G_{src} and the sets of their separated identities in G_{tar} . Running a deterministic re-identification attack on (G_{src}, G_{tar}) results in a re-identification mapping denoted as $\mu : V_{src} \rightarrow V_{tar}$.

3.1 Social Network Data and Modeling Identity Separation

During our experiments we used multiple datasets with different characteristics in order to avoid related biases. In addition, we used large networks, as brute-force attacks can be mounted against smaller ones. We obtained two datasets from the SNAP collection [10], namely the Slashdot network crawled in 2009 (82,168 nodes, 504,230 edges) and the Epinions network crawled in 2002 (75,879 nodes, 405,740 edges). The third dataset is a subgraph exported from the LiveJournal network crawled in 2010 (at our dept.; consisting of 66,752 nodes, 619,512 edges).

For modeling identity separation, it would be desirable to analyze real-world data on user behavior, but to the best of our knowledge, such datasets are unavailable and there are no trivial ways of crawling one (yet). Fortunately, data on a functionality similar to identity separation is available: structural information of social circles extracted from Google+, Twitter and Facebook [10]. We found in this data that the number of circles has a power-law distribution, for instance in the Twitter dataset we measured $\alpha = 2.31$ (933 ego networks, $x_{min} = 2, x_{max} = 18$). Many users did not duplicate their connections (44.6%), and only a fragment of them had more than twice as many connections in their circles compared to the number of their unique acquaintances (6.07%). While it is not possible to draw strong conclusions from these observations, we believe they indicate the real nature of identity separation (the usability of this dataset is limited by the absence of hidden connections).

Thus, due to the lack of data, we used the probability based models we introduced in [5], which describe identity separation from a structural point of view,

and allow deriving test data from real-world datasets. These models capture identity separation as splitting a node, and assigning previously existing edges to the new nodes. The number of new identities is modeled with a random variable Y (with unspecified distribution), which we either set to a fixed value, or model it with a random variable having a power-law-like distribution. For edge sorting, there are four models in [5] regarding whether it is allowed to delete edges (i.e., an edge becomes private), or to duplicate edges, from which we used three in our experiments. While the basic model is simple and easy to work with (no edge deletion or duplication allowed), we used the realistic model to capture real-life behavior, too (both operations are allowed). We additionally used the best model describing a privacy oriented user behavior (no edge duplication, but deletion allowed), and omitted the worst model (edge duplication only).

3.2 Data Preparation

During the test data creation process first we derived a pair of source and target graphs (G_{src}, G_{tar}) having desired overlap of nodes and edges, and then modeled identity separation on a subset of nodes in the target graph. We used the perturbation strategy proposed by Narayanan and Shmatikov [1]. Their algorithm considers the initial graph as the ground truth of real connections, from which graphs G_{src}, G_{tar} are extracted with the desired fraction of overlapping nodes (α_v), and then edges are deleted independently to achieve edge overlap α_e .

We found $\alpha_v = 0.5$, $\alpha_e = 0.75$ to be a good trade-off at which a significant level of uncertainty is present in the data (capturing the essence of a life-like scenario), but the Nar09 attack is still capable of identifying a large ratio of the co-existing nodes¹. Identity separation is then modeled on the target graph by uniformly sampling a given percent of nodes with at least $deg(v) = 2$ (this ratio is maintained for the ground truth nodes), and then nodes are split and their edges are sorted according to the settings of the currently used model.

3.3 Calibrating Attack Parameters and Measuring Success Rate

By comparing the directed and undirected versions of Nar09, we found little difference in results, therefore, due to this reason and for sake of simplicity, in our experiments we used undirected networks. Next, we run several measurements to find the optimal parameters of the attack. We found choosing randomly a 1,000 from the top 25% (by node degree) of mutually existing nodes to be a redundant choice modeling a strong attacker (as 750 seeds were enough for reaching the high-end of large scale propagation).

Seed location sensitivity of the algorithm is known for small networks [9]. In contrast, we found that seed location matters less for large networks, likely

¹ Without adding perturbation Nar09 could correctly identify 52.55% of coexisting nodes in the Epinions graph, 68.34% in the Slashdot graph, and 88.55% in the LiveJournal graph; identification rates were consequently proportional to the ratio of one-degree nodes.

because the greater redundancy in topology against perturbation, and larger ground truth sizes. Therefore, in each experiment we created two random perturbations, and run simulations twice on both with a different seed set. We observed only minor deviations in results, usually less than a percent.

The Nar09 algorithm has another important parameter (Θ) for controlling the ratio of true and false positives. The attack produced fairly low error rates even for small values of Θ , hence we choose to work with $\Theta = 0.01$. The error rate stayed well under 3% in most of experiments, with only a few exceptions when it went above slightly this value.

We use two measures for evaluating simulation results. The *recall rate* reflects the extent of re-identification (this itself can be used due to constantly negligible error rates), describing success from an attacker point of view. It is calculated by dividing the number of correct identifications with the number of mutually existing nodes (seeds are excluded from the results).

The *disclosure rate* quantifies information the attacker learned from users who applied identity separation, describing an overall protection efficiency from a user point of view. As current identity separation models are bound to structural information, we use a measure reflecting the average percent of edges that the attacker successfully revealed (this can be extended for other types of information in future experiments, e.g., sensitive profile attributes).

4 Characterizing Weaknesses of the Nar09 Algorithm

In the first part of our experiments, in order to discover the strongest privacy-enhancing identity separation mechanisms, we investigated the efficiency of features in different models against the Nar09 algorithm.

4.1 Measuring Sensitivity to the Number of Identities

Foremost, we tested the Nar09 algorithm against the *basic model with uniform edge sorting probability*. Simulations of the attack were executed for all networks having a ratio of users applying identity separation of $R \in [0.0, 0.9]$ (with stepping 0.1). For the selected users a fixed number of new identities were created ($Y \in [2, 5]$). We summarized results on Fig. 2; however, we omitted results for cases of $Y = 3, Y = 4$, as these can be easily inferred from the rest.

Opposing our initial expectations, the basic model with $Y = 2$ and uniform edge sorting probability is not effective in stopping the attack. For the Epinions and Slashdot networks the recall rate mildly decreased until the ratio of privacy-protecting users reached circa $R = 0.5$. For the LiveJournal graph the recall rate shows relevant fault tolerance of the attack (likely because of network structure, as this is also the most dense test network), e.g., Nar09 still correctly identified 15.12% of users even for $R = 0.7$. When participating users had five new identities, results were more promising, as recall rates dropped below 10% at $R = 0.5$ for all networks.

We also tested what if edges are sorted according to power-law distribution having $Y = 5$. These experiments resulted in a slightly higher true positive

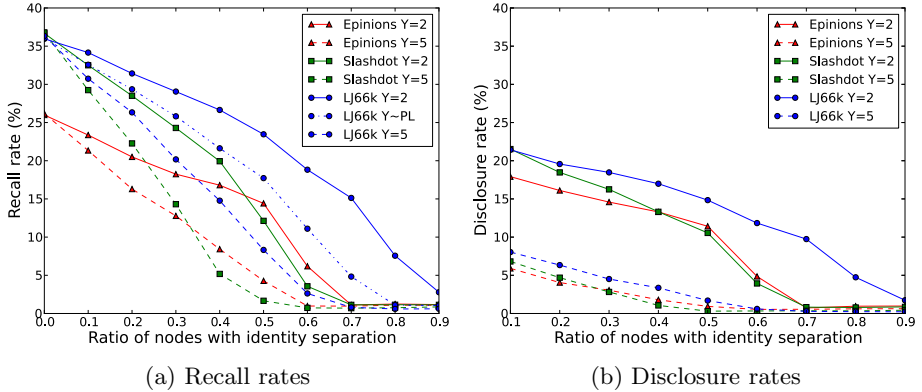


Fig. 2. Experimental results using the basic identity separation model

rate, which is not very surprising: if edges are not uniformly distributed it is more likely for an identity to appear that has more of the original edges than the others (with higher chances to be re-identified). In another comparative experiment we modeled a variable number of new identities with power-law-like distribution with $Y \in [2, 5]$ and uniform edge sorting probability. Results were properly centered between cases $Y = 2$ and $Y = 5$ as the LiveJournal example shows on Fig. 2a.

Even though by looking at the recall rates the basic model seems ineffective in impeding the attack, the disclosure rates imply better results. As shown on Fig. 2b, disclosure rates are significantly lower compared to recall rates². From this point of view using the basic model with $Y = 5$ and uniform edge sorting probability provides strong protection for even a small ratio of applying users: the disclosure rate is at most 8.03% when $R = 0.1$. By comparing the results of the two measures, we conclude that by using the basic model it is not feasible to repel the attack, however, by using a higher number of identities the access of the attacker to information can be effectively limited.

While conducting the analysis, we found that the recall rate was notably higher for users of identity separation ($\forall v \in \text{dom}(\lambda_G)$) than for others ($\forall v' \in \tilde{V}_{src}$). For low values of R this difference in the recall was almost constant and disappeared for high values³. This turned out to be a bias caused by the seeding strategy: after changing to mixed seeding with an equal ratio of seeds selected from $\text{dom}(\mu_G)$ and $\text{dom}(\lambda_G)$, while the overall recall rate remained equivalent the difference disappeared for the LiveJournal and Slashdot networks, and significantly decreased for the Epinions. Most importantly (from the user perspective), the disclosure rates stayed equivalently low regardless of the used seeding strategy.

² As the disclosure rate is measured for $\forall v \in \text{dom}(\lambda_G)$, results start from $R = 0.1$.

³ We omit plotting this on a figure due to space limitations, however, the difference was as follows: $\text{avg}(\Delta_{\text{Slashdot}}) = 2.34\% \forall R \in [0.1, 0.4]$; $\text{avg}(\Delta_{\text{Epinions}}) = 6.13\% \forall R \in [0.1, 0.5]$; $\text{avg}(\Delta_{\text{LiveJournal}}) = 4.05\% \forall R \in [0.1, 0.7]$

This finding has an interesting impact for the attacker on choosing the proper seeding strategy. Using a simple seeding mechanism seems to be a natural choice, but adding fault tolerance against identity separation is not trivial: the analysis provided by in [5] shows that the seeding method discussed in [1] is not very resistant to identity separation. Thus, using a simpler choice of seed identification, the attacker will also have a higher rate of correct identification for nodes protecting their privacy.

4.2 Measuring Sensitivity to Deletion of Edges

We used the realistic and best models to test the Nar09 against additional edge perturbation [5], as edge deletion is allowed during the edge sorting phase within these models. Since details are not explicitly defined in [5], we used three different settings in our experiments. For all of them edge sorting probabilities are calculated according to multivariate normal distribution as $P(X_1 = x_1, \dots, X_y = x_y) \sim \mathcal{N}_y(\boldsymbol{\eta}, \boldsymbol{\Sigma})$, where y denotes the current number of identities. We set each value of $\boldsymbol{\eta}$ to $(y)^{-1}$ and configure $\boldsymbol{\Sigma}$ in a way to have higher probabilities for events when the sum of the new edges are relatively close to original node degree (in the best model when the sum was higher than the original degree, the distribution was simply recalculated).

The first setting is the *realistic model with minimal deletion*, in which each edge is assigned to each identity, and if there is still ample space left, random edges are assigned to those identities. In this setting edges are not deleted if it is not necessary. In the setting of the *realistic model with random deletion* new identities take a portion of edges proportional to (x_1, \dots, x_y) . This setting is expected to delete unassigned edges proportionally to $\prod(1 - x_i)$. We also included a setting with the best model for comparison, namely the *best model with random deletion*⁴.

We ran simulations for all models in all the test networks with $Y = 2$, and found that recall rates strongly correlate with results of the basic model (although being slightly better); thus, these models are also incapable of repelling the attack. Fortunately, disclosure rates show significant progress from the basic model; as an example, results for the Epinions network are depicted on Fig. 3a. We conclude that while these models are also incapable of stopping large-scale propagation, they yet perform better in privacy protection.

4.3 Simulating Multiple Model Settings in Parallel

In previous subsections we described experiments in which settings of different models were used homogeneously. Naturally, the question arises whether the observed differences remain if multiple settings are allowed in the same network in a mixed way? Hence in another experiment we allowed three settings in parallel: basic model with uniform edge sorting probability (34% of R), realistic model

⁴ We note, however, that none of these settings capture aggressive edge deletion, but it might be interesting to investigate such settings in the future.

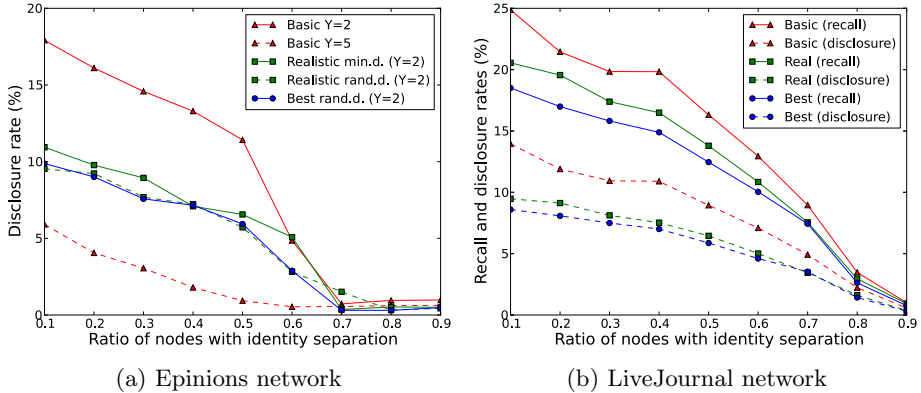


Fig. 3. (a) Disclosure rates for different models in the Epinions network; (b) multiple models present in parallel in the LiveJournal network.

with random deletion (33% of R), best model with random deletion (33% of R). We found that for the users of each setting was proportional to results measured in previous experiments, for instance, users of the best model achieved the lowest recall and disclosure rates. Simulation results in the LiveJournal graph are plotted on Fig. 3b for demonstration (results were measured for homogeneous groups consisting of nodes having the same setting).

5 Searching for Strongest User Protection Mechanisms

5.1 Measuring the Best Trivial Strategies

Previously we characterized weaknesses of the Nar09 attack, and it turned out that while none of the previously analyzed defense strategies can effectively stop it, some forms of identity separation can reduce the amount of accessible information. It also turned out that increasing the number of new identities has a powerful impact on the disclosure rate, while edge perturbation has a less, but yet remarkable effect. Thus, the best model with a high number of identities seems to be the most effective setting.

We run the best model with $Y = 5$ (using the same distribution as described in Section 4) on all test networks. Results revealed even this method cannot prevent large-scale re-identification when a relatively low ratio of users apply the technique. Instead, for all networks the re-identification rate converged to a hypothetical linear line monotonically decreasing as R increase (see Fig. 4a). Fortunately, the setting had more convincing results for disclosure rates: even for $R = 0.1$ the disclosure rate topped at 2.22%. Disclosure values continued to fall as the ratio of defending users increased.

In addition, we also examined disclosure rates for cases when participation were very low such as 1% of V_{tar} , meaning only a few tens or hundreds of users

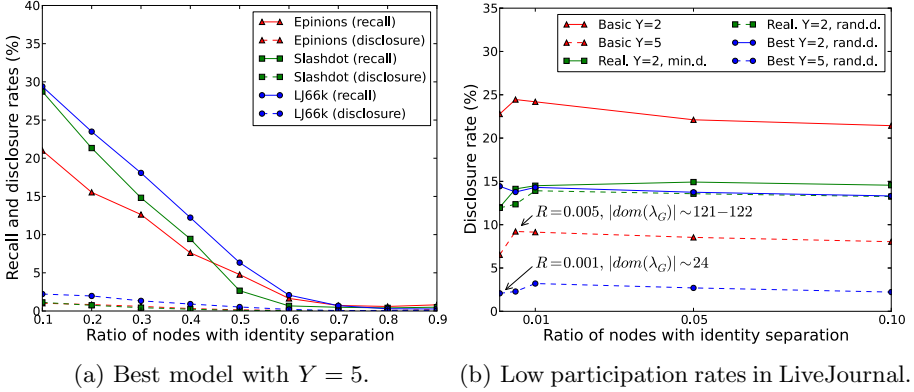


Fig. 4. Analysis of the most effective privacy-enhancing strategies

using identity separation from \tilde{V}_{tar} . As seen on Fig. 4b, our experiments resulted in approximately constant disclosure rates for all models (variability for $R < 0.01$ is likely to be due to the small sample sizes). Therefore we conclude that even if only a few users use the best model with $Y = 5$, their privacy is protected as the attacker can reveal only a few percent of sensitive information.

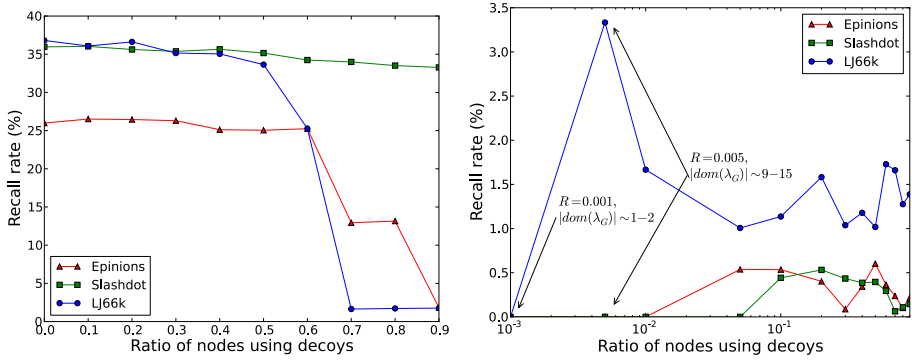
5.2 Placing the User in the Position of Decision

Strategy proposed previously lacks user control, i.e., the user cannot influence on what information she wishes to hide from the attacker. Here, we introduce a simple model that puts the user into decisive position by utilizing decoy identities. Nevertheless, we note that this model is a simple example, and ones used in real-life situations can be adapted to other hypothetical attacker strategies and to the type of information for hiding (e.g., one may consider using structural steganography for hiding nodes [11]).

We applied the following strategy on nodes $v_i \in \tilde{V}_{tar}$ that have at least 30 neighbors⁵. First, we create a decoy node v_i^{decoy} representing non-sensitive connections with the goal of capturing the attention of attacker algorithm (this may be a public profile as well). Node v_i^{decoy} is assigned 90% of the acquaintances of v_i . Next, a hidden node v_i^{hidden} is created having the rest 10% of neighbors (i.e., sensitive relationships), and an additional 10% that overlaps with the neighbors of v_i^{decoy} (i.e., modeling overlapping relationships).

This model showed promising results after being applied to our test data sets. While from the attacker point of view the algorithm was successful, as being able to produce high recall rates (until large number of decoys appeared – see details on Fig. 5a), privacy-protecting nodes achieved of revealing little sensitive information as shown on Fig. 5b. Recall rates were typically small for hidden

⁵ Resulting in a significantly smaller set of applicable nodes, e.g., in LiveJournal, even for $R = 0.9$ only $|dom(\lambda_G)| \approx 11.2\%$ of \tilde{V}_{tar} .



(a) Recall sensitivity to the use of decoys. (b) Results for nodes with decoys.

Fig. 5. Recall rates for whole networks and nodes using decoy nodes

nodes, less or equal 0.6% within the Slashdot and the Epinion networks, and with one exception less or equal 1.66% within the LiveJournal network. Misguidance was also successful when only a few users used it⁶.

6 Future Work and Conclusions

In this paper, we analyzed different models of identity separation to evaluate their effect in repelling structural de-anonymization attacks and in information hiding. By our experiments we found that if identity separation is used in a non-cooperative way, it is not possible to avoid large-scale re-identification regardless of the used strategy, unless a large fraction of users is involved. This finding sets a direction for future work: is there a way for cooperating users to tackle these attacks more effectively?

We also used another measure in our experiments, reflecting the quantity of information a successful attack reveals. This metric showed more promising results: experiments confirmed that using multiple identities and allowing to hide some connections favors user privacy. Moreover, in our experiments using five identities with a moderate preference for hiding or duplicating edges proved to be eligible to achieve a high rate of information hiding. Numerically, in the LiveJournal network we measured an information disclosure of 2.08% when only circa 24 users applied the proposed strategy, and yet results were well under 4% in other cases as well.

However, using five identities is not realistic for most users, and in this case it is not possible to control what the attacker may reveal. Therefore, we proposed a method of using decoys, which seemingly did not affect the success of the attack (unless used by more than half of the users); however, as the attacker almost

⁶ Appearing variability is due to small sample sizes, and almost negligible. For instance, the recall of 3.33% means one node being identified correctly within 29 cases.

discovered decoy nodes only, a minority of hidden nodes were found: less or equal 0.6% within the Slashdot and the Epinion networks, and with one exception less or equal 1.66% within the LiveJournal network. This method also produced suitable results when applied by a few nodes only (i.e., numerically 1-2).

Therefore, we have provided guidelines (in the form of two models) for effectively realizing information hiding in social networks, that can be applied to existing social networks, even without the consent of the service provider. As our closing word, we designate an interesting direction as future work: what strategies should a user follow, if identity separation can be applied in both networks of G_{src} and G_{tar} ?

References

1. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: 30th IEEE Symposium on Security and Privacy, pp. 173–187. IEEE Press, New York (2009)
2. Narayanan, A., Shi, E., Rubinstein, B.I.P.: Link prediction by de-anonymization: How we won the kaggle social network challenge. In: The 2011 International Joint Conference on Neural Networks, pp. 1825–1834. IEEE Press, New York (2011)
3. Peng, W., Li, F., Zou, X., Wu, J.: Seed and Grow: An attack against anonymized social networks. In: 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, pp. 587–595. IEEE Press, New York (2012)
4. Srivatsa, M., Hicks, M.: Deanonymizing mobility traces: Using social network as a side-channel. In: 2012 ACM Conference on Computer and Communications Security, pp. 628–637. ACM Press, New York (2012)
5. Gulyás, G.G., Imre, S.: Analysis of Identity Separation Against a Passive Clique-Based De-anonymization Attack. *Infocommunications Journal* III(4), 11–20 (2011)
6. Cuttillo, L.A., Molva, R., Strufe, T.: Safebook: a Privacy Preserving Online Social Network Leveraging on Real-Life Trust. *IEEE Communications Magazine* 47(12), 94–101 (2009)
7. Beato, F., Conti, M., Preneel, B.: Friend in the Middle (FiM): Tackling De-Anonymization in Social Networks. In: 5th IEEE International Workshop on Security and Social Networking (2013)
8. Clauß, S., Kesdogan, D., Kölsch, T.: Privacy enhancing identity management: protection against re-identification and profiling. In: 2005 Workshop on Digital Identity Management, pp. 83–94. ACM Press, New York (2005)
9. Gulyás, G.G., Imre, S.: Measuring Local Topological Anonymity in Social Networks. In: 12th International Conference on Data Mining Workshops, pp. 563–570. IEEE Press, New York (2012)
10. Stanford Large Network Dataset Collection, <http://snap.stanford.edu/data/index.html>
11. Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In: 16th International Conference on World Wide Web, pp. 181–190. ACM Press, New York (2007)

An Equivalent Access Based Approach for Building Collaboration Model between Distinct Access Control Models

Xiaofeng Xia

Heidelberg Institute for Theoretical Studies (HITS),
D-69118 Heidelberg, Germany
Xiaofeng.Xia@h-its.org

Abstract. Organizations collaborate with each other for resource sharing and task performing. To protect their resources from unauthorized access the organization domains adopt own access control models. The collaboration thus faces a problem that how a secure collaboration is built between the domains with distinct access control models. Currently there are approaches focusing on role based access control model (RBAC), where role mapping is considered to be the main technique. It assumes that all organizations adopt RBAC model, then builds a global access control policy on role mappings. However if the organization domains, also including collaboration domain, use distinct access control models, role mapping and global policy can not be built on these models. In this paper we propose an equivalent access based approach and introduce a mediator involved collaboration pattern, where access control model entities have corresponding mapping and linking sets on which the equivalent accesses are built. Collaboration also introduces the "Inter Domain Role Mapping" (IDRM) problem and we thus propose new algorithms for IDRM problem based on flat and hierarchical role structures, in addition we also introduce the necessary constraints transforming between organization and collaboration domains. Finally we analyze our algorithms and present the testing and comparison results with existed approaches.

Keywords: collaboration model, distinct access control models, equivalent access, mapping set, linking set.

1 Introduction

When several organizations want to make a collaboration, they could share resources among each other such that some common tasks can be completed. The collaboration pattern discussed in this paper refers to that for the resources shared from participating organization domains, the collaboration domain can have its own access control model. Practical security policy configurations tell us that the security models or policies are not all-purpose. To protect their resources from unauthorized access the organization domains adopt different access control models, e.g. RBAC [10], mandatory access control (MAC)[12], and discretionary access control (DAC)[6]. These models have different model entities related to permissions, which we call core model semantics. For example RBAC model constructs roles, MAC model has security labels. Currently there

are approaches, e.g. in [3] and [4], focusing on RBAC, which assume that all organizations adopt RBAC model, then build a global access control policy on role mappings. A global policy can be generated, because all domains have the same core model semantics, however if the domains use distinct access control models, role mapping and global policy can not be built on these models.

Organizational collaboration also introduces the IDRM problem in [3], which means to find out the minimal role set covering requested permissions from collaboration domain. This problem can then be generalized to distinct models and be defined as finding out an "appropriate" set of core model semantics covering requested permission set.

The third problem for organization collaboration is constraints transforming. As the model entities are mapped between domains, from the perspective of participators there are some constraints that must also be held in collaboration domain, e.g. for RBAC model, the separation of duty constraint(SSD)[9].

Therefore in this paper our contributions are (1)building a collaboration model between distinct access control models; (2)the necessary algorithms of figuring out an appropriate set of core model semantics to requested permission set; (3)constraints transforming between distinct models. The rest sections of this paper are organized as following: section 2 describes related work, in section 3 we present a new collaboration model based on equivalent access and section 4 illustrates the supporting algorithms and methods on building the collaboration model. Our testing and comparison results to algorithms is presented in section 5. Finally we have the conclusion of this paper in section 6.

2 Related Work

The RBAC[10] [11] model provides role-permission management, role hierarchy, and separation of duty constraints. For Lattice Based Access Control(LBAC) or MAC model [8], the information flow is restricted by the constraints on security labels and clearances. DAC[6] model emphasizes owning relationships of resources and permission delegation to be the way of authorization. In the past years RBAC is the most concerned model due to the of its conforming to organization structure.

An context-dependent RBAC model[7] is proposed to enforce access control in web-based collaboration environments. Organization based access control(OrBAC)[1] is constructed from a RBAC model as concrete level, and OrBAC then refers to common organizational contextual entities as abstract level. Based on OrBAC, PolyOrBAC[2] is proposed to implement the collaboration between organizations having OrBAC model in their domains. It takes advantage of abstract organizational entities and Web Services mechanisms, e.g. UDDI, XML, SOAP, to enforce a global framework of collaboration for engaging organization domains.

Role mapping[4] helps one domain obtaining accesses to resources from other domains by role-inheritances across domains. A global access control policy is specified to merge the engaging organization's local policies. This approach also assumes that all domains adopt RBAC model. Due to these contributions on RBAC and collaboration, we start to focus on organization domains with distinct access control models.

The other work on collaboration, or inter-domain operation, refers to IDRM problem. In [3], the proposed greedy-search based algorithm is an approximate solution to

IDRM problem, however simple greedy-search has local-maxima problem, therefore a probability based greedy-search algorithm in this paper is used to avoid local-maxima and get better approximate solution. In section 5 we will discuss the problems of these approaches comparing with ours. To improve the algorithms, [5] presents another idea on greedy-search, they note that the assumption of IDRM problem should be more complex and practical. IDRM problem can be reduced to a weighted-set cover problem instead of minimal set cover problem in [3]. However the algorithm by [5] can not avoid local-maxima either.

3 Equivalent Access and Collaboration Model

3.1 Preliminary Definitions

An organization domain or collaboration domain \mathcal{D} should contain part of the following entity sets and relations:

- *User, Resource, and Action*: the sets of system users, resources, and operations on resources;
- T : the set of Tag objects, e.g. roles, security labels;
As we constructed DAC model with a role based way [6], we view the model semantics as Tag objects, i.e. both role and label object can be instantiated by a Tag class which has at least two attributes: $\langle type, name \rangle$.
- $Permission \subseteq Resource \times Action$, a set of permissions;

And some predicates and functions:

- $Reslabel(Resource, T)$: the assignment relation between a resource and a security label;
- $mayAccess(User, Resource, Action)$: a common predicate indicating an access request from some user to make an operation on some resource.
- $Usetag : User \times T$, is to indicate the relation that certain Tag objects (roles or labels) are assigned to some user.
- $PS : T \rightarrow Permission$: the permissions assigned or held by a Tag object;
- $PR : Permission \rightarrow T$: the tags holding current permission;

3.2 Collaboration Model Based on Equivalent Access

Any access request of a user to some resource can be enforced by different access control models. We introduce "equivalent access", which is related to two domain's access control policies. Since in organizational collaborations the preliminary goal is to find appropriate resources, equivalent access refers to that a user's access to some resource under collaboration domain policy has equivalent evaluation results as that under participating domain policy.

Equivalent access should be the preliminary goal of organization collaborations, i.e. the constructing process of collaboration is to find equivalent accesses for the required resources in participating domains.

The collaboration scenario we discussed here refers to a collaboration domain, denoted as \mathcal{D}_c , and a series of original domains, i.e. $(\mathcal{D}_c; \mathcal{D}_1, \dots, \mathcal{D}_n), n \geq 2$. Each domain applies own access control model and policy. For a collaboration group $(\mathcal{D}_c; \mathcal{D}_1, \dots, \mathcal{D}_n)$, there exists two sorts of entity relations between collaboration domain (\mathcal{D}_c) and other participating domains ($\mathcal{D}_i, i \in [1, n]$); one is the entity mapping set, the other is the entity linking set. We denote the former as \mathcal{Q} , which maps the entities of \mathcal{D}_i onto those of \mathcal{D}_c simply. The mapping means that for any resource $e_0 \in \mathcal{D}_i$, it has a corresponding virtual resource $e'_0 \in \mathcal{D}_c$. The mappings are classified into “user”, “resource” and “action”, i.e. $\{\zeta^u, \zeta^e, \zeta^a\}$.

Another relation is entity linking set, denoted as \mathcal{L} , which need to be computed and will be introduced in following parts.

Definition 1. For a collaboration group $(\mathcal{D}_c; \mathcal{D}_1, \dots, \mathcal{D}_n)$, considering any participating domain $\mathcal{D}_i, i \in [1, n]$ and its mapping set \mathcal{Q} , there are $u \in user_c$ and $e \in resource_c$, as well as $u' \in user_i$ and $e' \in resource_i$, such that $\langle u, u' \rangle \in \zeta_1^u$ and $\langle e, e' \rangle \in \zeta_{\langle \mathcal{D}_i, \mathcal{D}_c \rangle}^e$; we say that the access by u to e is equivalent to that by u' to e' under two policies \mathcal{P}_c and \mathcal{P}_i , if for the substitutions $\theta_{\mathcal{D}_c} = \{U_x/u, E_x/e, A_x/read\}$ and $\theta_{\mathcal{D}_i} = \{U_x/u', E_x/e', A_x/read'\}$ and:

$$\mathcal{P}_c \models mayAccess(U_x, E_x, A_x)[\theta_{\mathcal{D}_c}] \wedge \mathcal{P}_i \models mayAccess(U_x, E_x, A_x)[\theta_{\mathcal{D}_i}] \quad (1)$$

Then the **equivalent access** is denoted as:

$$mayAccess(U_x, E_x, A_x)[\theta_{\mathcal{D}_c}, \theta_{\mathcal{D}_i}] \mid \langle \mathcal{P}_c, \mathcal{P}_i \rangle \quad (2)$$

Definition 2. The elements of entity linking set indicate the pairs of related “Tag” objects respectively from collaboration (\mathcal{D}_c) and original (\mathcal{D}_i) domains. When two substitutions towards their own policies \mathcal{P}_c and \mathcal{P}_i have equivalent access, a set $S_{\mathcal{D}_c}$ indicates the “Tag” objects which satisfy the request by $\theta_{\mathcal{D}_c}$ and a set $S_{\mathcal{D}_i}$ indicates those by $\theta_{\mathcal{D}_i}$, then the **entity linking set** $\mathcal{L}_{\langle \mathcal{D}_i, \mathcal{D}_c \rangle}$ is defined as following rule:

$$\mathcal{L}_{\langle \mathcal{D}_i, \mathcal{D}_c \rangle} = \{ \langle r, l \rangle \mid \langle r, l \rangle \in S_{\mathcal{D}_c} \times S_{\mathcal{D}_i} \}. \quad (3)$$

Definition 3. For a collaboration group $(\mathcal{D}_c; \mathcal{D}_1, \dots, \mathcal{D}_n)$, where all domain’s model has the form of $\{\mathcal{D}_R, \mathcal{D}_M, \mathcal{D}_S\}$, considering any original domain $\mathcal{D}_i, i \in [1, n]$ and its mapping set $\mathcal{Q}_{\langle \mathcal{D}_c, \mathcal{D}_i \rangle}$ with \mathcal{D}_c , the **collaboration model** Γ of the group will be defined by the above definitions of organization domain as a union of pairs:

$$\Gamma = \bigcup_{i=1}^n \langle \mathcal{Q}_{\langle \mathcal{D}_c, \mathcal{D}_i \rangle}, \mathcal{L}_{\langle \mathcal{D}_c, \mathcal{D}_i \rangle} \rangle \quad (4)$$

4 Building Collaboration Model between Distinct Access Control Models

In this section we will analyze the problems of building a collaboration model, then introduce the algorithms we use to build the model, as well as the methods to transfer constraints into collaboration domain. According to the definition of our new collaboration

model, there are basically 3 steps to enforce:(1)finding out equivalent accesses;(2)try to minimize the scale of disclosure of the organization's policy information involved into collaboration;(2)domain constraints should be transferred into collaboration domain by configuring them on the policy entities in collaboration domain.

4.1 RBAC as Participator's Model

4.1.1 Minimal Role Set Covering Requested Permissions

A greedy-search based algorithm (GSA) is proposed to get a solution to IDRМ problem (NP-complete) in [3]. Basically the algorithm handles each candidate role with taking all its permissions that can cover as much as possible target permissions, then put this role into solution set. [3] also provides another probabilistic-greedy-search algorithm (IGSA-PROB) which executes candidate role handling with probability p (near 1). Greedy-search based algorithm however does not guarantee to find the optimal solution R' . It is an H_n -approximation algorithm for IDRМ problem. The IDRМ approaches proposed in [3] hence has the following problems:(1)the GSA algorithm is non-terminating and will probably not find any solution;(2)the GSA algorithm has local-maxima problem ;(3)the IGSA-PROB algorithm searches with probability p , while the local-maxima problem cannot be effectively avoided;(4)the inheritance hierarchy of roles can be applied to the IDRМ problem.

The GSA and IGSA-PROB algorithms select only the roles which have permissions as a subset of required permission set to be candidates. Thus it makes the algorithms non-terminating. We build collaboration model by entity mapping and linking sets. The entity mapping set ensures that only the requests involving mapped entities will be allowed, which means that even if a role r is linked into, but only the mapped permissions will be allowed. This enables our algorithm to terminate. Towards solving IDRМ problem we propose three algorithms, the input of them includes RQ , requested permission set; R , set of all roles; P , set of all permissions; R_S , set of initially selected roles; in turn the output has TS , set of candidate roles. They are specified formally in appendix.

I. Improved GSA Algorithm (IGSA)

- (1) finding out all the roles from R , which have intersected permissions with requested RQ , and put them into R_S .
- (2) for a role r in set R_S , if r 's permission set covers larger part of RQ than any other roles in R_S , then put r into candidate set TS , and remove r from R_S as well as remove the covered permissions of r from RQ .
- (3) if RQ is not empty, then go to step (2).

II. Improved Algorithm for Local-Maxima (IGSAL)

- (1) for each permission finding out those which are assigned to a single role r .
- (2) for the other roles in R , remove the permissions assigned to them, but also assigned to the role r .
- (3) comparing each role r' with all of the other roles, if one of the permissions of r' belongs to another role r^* and r^* has more permissions than r' , then remove all of the overlapped permissions from r' .

- (4) if the permissions of r' are all removed, then r' should also be removed from R .
- (5) performing the steps of algorithm I to compute candidate set TS .

III. Algorithm for Hierarchical Roles (HCHY)

- (1) initially put the roles which have no parent roles, into set S_1 , remove them from their child roles' parents list, then make a new set S_2 .
- (2) for each role r in R , if it has no parent roles and it dose not belong to S_1 and S_2 , and if the convergent class set $Convergent_Classes$ is empty, then make a new convergent class set and add r into it; if $Convergent_Classes$ is not empty, then check every convergent class set C in it, if the current role r belongs to the child role set of any role in C , add r into C .
- (3) remove r from the parent role set of each child role of r , add r into S_2 .
- (4) make a new set S_3 ; for S_3 and each permission p of P , make another new set S_4 , thus for each role r' which holds p , if there is a convergent class set C containing r' , add r' to S_4 .
- (5) after checking all of the roles having p , add S_4 into S_3 ; make new sets S_5 and S_6 .
- (6) by a recursive process "recurse", compute the combinations of sets in S_3 and return the minimal combination results.

4.1.2 Constraints of Participating Domain

Figuring out the minimal set roles covering requested permissions is the first step to enable the collaboration process, in addition, we must see that some RBAC constraints should also be held in collaboration domain. Here we focus on the static separation of duty constraint(SSD), which is defined as the following statements where "assigned_user(r)" indicates the set of users holding the role " r ", and "assigned_tag(u)" indicates the set of roles being assigned to user " u "[9].

- $SSD \subseteq (2^R \times N)$, R is the set of roles and N is the set of natural numbers.
- $\forall \langle rs, n \rangle \in SSD \forall t \subseteq rs. |t| \geq n \rightarrow \bigcap_{r \in t} assigned_user(r) = \phi$

Now we know that rs is a set related to SSD , the possible " n -tuple" sets from rs is $C_{|rs|}^n$, which means the possibilities of picking n elements from $|rs|$ elements. For each possibility we define the set s_k of all involved permissions, thus $C_{|rs|}^n$ sets are defined as the following statements, where PS_{SSD} indicates the permission sets for each of the SSD constraint elements in participating domain:

$$\forall r_1, r_2, \dots, r_n \in rs. s_k = \bigcup_{i=1}^n PS(r_i)$$

$$PS_{SSD} = \{s_1, s_2, \dots, s_k\}, k = C_{|rs|}^n.$$

When the participating domain adopts RBAC model, the collaboration domain has also RBAC or DAC model (our DAC model is built by a "role" based way), it is necessary to note that there are 3 new constraints setting for collaboration domain's policy. They refer to in collaboration domain: (1) none of the "Tag" objects can have the whole permissions related to anyone of the SSD elements; (2) no user's permissions can cover

the whole permissions related to one SSD element; (3)if the collaboration domain has RBAC model, then configuring new SSD constraints from the role sets which have the requested permissions. The 3 constraints are formally defined as the following statements. When the collaboration domain has MAC model, then only the constraint <1> should be held, since in MAC model each user holds one security label. Each member of P_{SSD} will be mapped to corresponding permission sets $s'_i, i \in [1, k]$ in collaboration domain and the permission sets accordingly to P'_{SSD} in collaboration domain.

$$\begin{aligned}
 <1> \quad \forall s'_i \in P'_{SSD} \forall t \in T_{\mathcal{D}_c}. s'_i \not\subseteq PS(t). \\
 <2> \quad \forall s'_i \in P'_{SSD} \forall t \in T_{\mathcal{D}_c} \forall u \in U_{\mathcal{D}_c} \forall l' \in assigned_tag(u). s'_i \not\subseteq \bigcup PS(l') \\
 \quad \text{where } g = |s'_i| \geq 1, s'_i = \{p_j | j \in [1, g]\} \\
 <3> \quad \forall \langle rs_c, m \rangle \in SSD_{s'_i} \forall t' \in rs_c. |t'| \geq m \rightarrow \bigcap_{l \in t'} assigned_user(l) = \phi \\
 \quad \text{where } rs_c \subseteq T_{\mathcal{D}_c} \wedge o_d \subseteq rs_c \wedge m = |o_d| \wedge SSD_{s'_i} = \{o_d | o_d = \{r_s^1, r_s^2, \dots, r_s^g\}\}
 \end{aligned}$$

4.2 MAC as Participant's Model

If the participating domain adopts a mandatory access control model, then a resource has exactly one label. When the requested resources and operations are confirmed, these resources can be simply mapped onto different security labels to which they are assigned in participating domain. In this section we discuss on the Bell Lapadula model (BL)[12][6] in collaboration, and the other Biba model is about integrity, which is dual to BL model.

The MAC model assigns for each object exactly one security label and for each user or subject only one security clearance. Comparing with the scenario where RBAC as participant's model, we only need to find out the labels of resources lying in the requested permissions, then these labels can provide equivalent accesses. To prevent disallowed information flow in collaboration domain, additional constraints must be added to collaboration domain policies. Since finding out the labels of resources is trivial, we provide only the definition of newly created constraint in collaboration domain. Assuming that a collaboration model Γ and one of the participating domains \mathcal{D}_i and the collaboration domain \mathcal{D}_c are defined as in section 3.

Single Label Constraint

$$\begin{aligned}
 <1> \quad P'_r = \{ \langle e, a \rangle | \forall e' \in Resource_{\mathcal{D}_c} \exists e \in Resource_{\mathcal{D}_i}. r \in T_{\mathcal{D}_c} \wedge \langle e', a \rangle \in PS(r) \wedge \langle e', e \rangle \in \zeta^e \}. \\
 \quad P'_r \subseteq RQ \wedge \left| \bigcup \{ l | \forall \langle e, a \rangle \in P'_r \wedge Reslabel(e, l) \} \right| = 1. \\
 <2> \quad \forall u \in U_{\mathcal{D}_c} \forall l, r \in T_{\mathcal{D}_c}. Usertag(u, l) \wedge Usertag(u, r) \rightarrow (l = r) \\
 <3> \quad T' = \{ l | \forall u \in U_{\mathcal{D}_c}. Usertag(u, l) \} \\
 \quad \forall l \in T' \exists t \in T_{\mathcal{D}_i}. P'_l \subseteq RQ \wedge \left| \{ t | \forall \langle e, a \rangle \in P'_l \wedge Reslabel(e, t) \} \right| = \{ t \}
 \end{aligned}$$

In the collaboration domain, the information flow policy of participating domain should be held. Single label constraint will make restrictions on the labels of the resources which are shared in collaboration domain. Each "Tag" object can be assigned with the permissions, whose mapping entities in participating domain have the same security

label. Each user or subject in collaboration domain can have either only one "Tag" object or multiple "Tag" objects which are assigned with the permissions related to same security label. Therefore the above constraint is expressed with the following formula: $\langle 1 \rangle \wedge (\langle 2 \rangle \vee \langle 3 \rangle)$.

4.3 DAC as Participator's Model

In a collaboration process, if the required permissions are provided from a participating domain with DAC model, the delegation of these permissions will not be considered in collaboration domain, since only the access permissions are necessary, while not the delegation permissions.

In our DAC model definitions, resource and different operations construct permissions for which different roles are created. Each resource has an owner, who is assigned "owner role" of the resource. The "owner role" inherits all of the permissions from other relevant roles.

Participating domain only needs to provide the basic roles which are related to the requested permissions. Although our DAC model adopts a "role" based way, in DAC model, there is no high level roles which hold large number of permissions related to different resources. Thus the previous algorithm of finding minimized role set for requested permissions will not be applied in DAC model. In participating domain with DAC model, there are no special constraints to be ensured in collaboration domain.

5 Analysis on Algorithm Properties and Testing Results

We present algorithms IGSA, IGSAL, and HCHY for handling minimal role set problem in section 5. Our collaboration model Γ verifies the entity mapping and linking sets, by which it is helpful to introduce non-required permissions. Only the collaboration relevant permissions, that is, the resources and operations are kept as entity mappings in collaboration model Γ , can be allowed for access.

As discussed in [3], the GSA has local-maxima problem and can be solved by GSA-PROB (probability based greedy search algorithm). By analyzing the problem we found that the permission assignment relationship, i.e. one permission assigned to multiple roles, causes local-maxima problem. Our IGSAL algorithm tries to remove this "multi-inheritance" from the role-permission relation, then the greedy search can be applied to resulted roles and permissions. To describe the complexity characteristics of these 3 algorithms, we assume that the size of requested permissions is N . Comparing with IGSA and GSA-PROB algorithms, IGSAL spends computation on preprocessing the role-permission relations, then starts a greedy search to obtain solution. However on efficiency of algorithm, IGSAL has a nested loop for checking all of the requested permissions, which makes a $\mathcal{O}(N^2)$ complexity. Since the complexity of greedy search referring to IGSA and GSA-PROB is $\mathcal{O}(\ln N)$ [3] and the second step of IGSAL is also greedy search, the final complexity of IGSAL is still $\mathcal{O}(N^2)$. By randomly generating permissions and the assignment relationships, a testing for handling 100 roles and 43000 50000 permissions and the size of requested permission ranges from 1000 to 15000. Table 1 shows that IGSAL is less efficient than IGSA, but more precise.

Table 1. Comparison of IGSA and IGSAL on efficiency

Role size	Perm size	Requested perms	Time consuming(IGSA/IGSAL)	Solution size(IGSA/IGSAL)
100	41613	10^3	71 / 5334	80 / 78
100	45807	2×10^3	79 / 14549	90 / 87
100	46055	3×10^3	90 / 23011	91 / 91
100	43696	4×10^3	104 / 31864	93 / 89
100	45252	5×10^3	113 / 43066	96 / 95
100	44701	6×10^3	121 / 54115	98 / 96
100	48191	7×10^3	193 / 81417	99 / 97
100	44323	8×10^3	143 / 84534	99 / 99
100	45879	9×10^3	221 / 109845	98 / 97
100	43841	10^4	164 / 110684	97 / 95
100	47209	11×10^3	243 / 161712	98 / 98
100	45088	12×10^3	266 / 161768	99 / 98
100	46269	13×10^3	269 / 188546	100 / 98
100	44134	14×10^3	300 / 197264	98 / 97
100	44036	15×10^3	299 / 217346	99 / 97

It is mentioned that the role hierarchy can be used to provide minimal role set for requested permissions. The collaboration model can ensure that only mapped and linked entities related permissions can be allowed to access, even if there is a high level role is involved and has more permission than requested. Therefore from one or multiple role hierarchies in an organization domain one can find out the powerful roles to cover as much as possible requested permissions. The hierarchies discussed in section 4 is called convergent classes. The algorithm HCHY computes firstly the convergent classes of roles contained in an access control model, which will make a time consuming with complexity $\mathcal{O}(C_1)$. C_1 indicates that a constant time consuming on convergent classes, since the roles and role hierarchies in a domain has already been determined in advance. It is only necessary to compute it once. The second step of HCHY algorithm is to input the requested permissions, which takes time complexity $\mathcal{O}(N)$. Finally we need to figure out by a recursive process the minimal set of roles covering requested permissions, which is only related to the size of roles, hence the complexity of this process varies by the number of involved role hierarchies, assuming C_2 . The total time complexity of HCHY on requested permissions is $\mathcal{O}(N) + C_1 + C_2$. By Table 2 we can see that HCHY is faster than IGSA.

In an organization domain with RBAC model, it adopts flat role structure or hierarchical role structure. our algorithms IGSA, IGSAL, and HCHY can handle and make use of both of these role structures.

Table 2. Performance testing of HCHY

Role size	Perm size	Requested perms	Convergent Classes	Time consuming	Solution
91	66610	10^3	60	29	3
91	66610	2×10^3	60	57	5
91	66610	3×10^3	60	65	7
91	66610	4×10^3	60	70	8
91	66610	5×10^3	60	75	7
91	66610	6×10^3	60	96	8
91	66610	7×10^3	60	86	10
91	66610	8×10^3	60	94	12
91	66610	9×10^3	60	106	15
91	66610	10^4	60	103	14
91	66610	11×10^3	60	119	16
91	66610	12×10^3	60	128	15
91	66610	13×10^3	60	149	21
91	66610	14×10^3	60	156	21
91	66610	15×10^3	60	157	22

6 Conclusion

In this paper we handle 3 problems in organizational collaboration: (1) a secure collaboration is built between the domains with the distinct access control models (2) finding out an "appropriate" set of core model semantics covering requested permission set (3) constraints transforming between organization and collaboration domains. We present an equivalent access based approach and introduce a mediator involved collaboration pattern for the first problem. New algorithms are in turn proposed for IDRM problem based on flat and hierarchical role structures. Then some new constraints are presented for the third problem. Finally we analyze our algorithms and present the testing and comparison results with existed approaches.

The collaboration pattern with "mediator" works for both situations that there is or there is no domain access control model in collaboration. The access control policies of participating domains are respected. In our future work, we will implement the mediator role, the collaboration model, and transformed constraints in XACML.

References

1. Kalam, A., Benferhat, S., Mieke, A., et al.: Organization based access control. In: Proceedings of the 4th Workshop on Policies for Distributed Systems and Networks, p. 120 (2003)
2. Kalam, A., Deswarte, Y., Bima, A., et al.: Access control for collaborative system: a web services based approach. In: Proceedings of International Conference on Web Services (2007)
3. Du, S., Joshi, J.: Supporting authorization query and inter-domain role mapping in presence of hybrid role hierarchy. In: Proceedings of the Eleventh ACM Symposium on Access Control Models and Technologies, pp. 228–236 (2006)
4. Joshi, J., Shafiq, B., Bertino, E.: Secure Interoperation in a Multidomain Environment Employing RBAC Policies. *IEEE Trans. on Knowl. and Data Eng.* 17(11), 1557–1577 (2005)
5. Chen, L., Crampton, J.: Inter-domain role mapping and least privilege. In: Proceedings of ACM Symposium on Access Control Models and Technologies (2007)
6. Osborn, S., Sandhu, R., Munawar, Q.: Configuring Role-Based Access Control to Enforce Mandatory and Discretionary Access Control Policies. *ACM Transactions on Information and System Security* 3(2), 85–106 (2000)
7. Wolf, R., Schneider, M.: Context-Dependent Access Control for Web-Based Collaboration Environments with Role-Based Approach. In: Gorodetsky, V., Popyack, L.J., Skormin, V.A. (eds.) *MMM-ACNS 2003*. LNCS, vol. 2776, pp. 267–278. Springer, Heidelberg (2003)
8. Sandhu, R.: Lattice based access control. *Journal of Computer* 26(11), 9–19 (1993)
9. Incits: ANSI INCITS 359-2004 for information technology role based access control (2004)
10. Sandhu, R.S., Coyne, E.J., Feinstein, H.L., Youman, C.E.: Role-based access control models. *IEEE Computer* 29(2) (1996)
11. Sandhu, R.S., Bhamidipati, V., Munawar, Q.: The ARBAC97 Model for Role-Based Administration of Roles. *ACM Transactions on Information and System Security*. Special Issue on Role-Based Access Control 2(1) (1999)
12. Bell, D.E., LaPadula, L.J.: Secure computer systems: Mathematical foundations and model. MITRE technical report 2547, vol. I (1973)
13. Nilson, U., Maluszynski, J.: Logic, programming and Prolog. pp. 14–16. John Wiley & Sons Ltd. (1995)

Part III

Extended Abstracts

Authentication with Time Features for Keystroke Dynamics on Touchscreens

Matthias Trojahn¹, Florian Arndt¹, and Frank Ortmeier²

¹ Volkswagen AG, Wolfsburg, Germany

² Otto-von-Guericke University of Magdeburg, Computer Systems in Engineering,
Magdeburg, Germany
{matthias.trojahn,florian.arndt1}@volkswagen.de,
frank.ortmeier@ovgu.de

Abstract. Keystroke authentication is a well known method to secure the mobile devices. Especially, the increasing amount of personal and sensitive data stored on these devices makes a secure authentication system necessary. Traditional security techniques like the four-digit PIN-input are insufficient and do not correspond to the present password standards. A keystroke behavior based authentication system could increase the security. Different researches have been published based on keystroke authentication systems with traditional PC keypads. But the keystroke behavior on touchscreens, as they are nowadays used on smartphones, are not analysed before.

Keywords: keystroke authentication, mobile devices, capacitive display.

1 Introduction

Today, smartphones are not only used like normal telephones to phone or write SMS's. This changed with the introduction of the iPhone in the year 2007. With this or other smartphones the number of security relevant data and information which are stored on the smartphone (or provided through applications) are increased.

Different studies showed already an improvement for the authentication if keystroke dynamics are used [1,2]. But the existing publications are mainly dealing with computer keyboards or 12-key hardware keyboard of mobile phones.

In this paper, we will discuss the standard features of keystroke dynamics on touchscreen devices. The goal is to see that an authentication with a touchscreen keyboard can be done. On the following questions our research is focused. If an authentication is done with a touchscreen keyboard using time features, the same error rates can be achieved compared with the existing keystroke dynamics studies.

2 Keystroke Authentication Background

Keystroke behaviour can be described as a biometric characteristic of a person. In particular how this person is typing on a keyboard [3]. It is used like other

biometric methods to verify a person. Furthermore, the rhythm how a person is typing can be calculated by different points but at least the time differences are used [4].

In general, two basic types of events can be recorded: The duration time which describes how long a key is pressed (time between pressing and releasing a key) is the first type. The second one describes the time period between n keystrokes, defined by the n press events. This is called n -graph [4]. Several variants of the time period exist. The most used is the digraph where $n=2$. In addition, some publications are using the combination of three key presses. This is called the trigraph [5]. Basically, each value over $n > 1$ is possible in order to determine the time differences. But with a higher value the information decreases which can be extracted by the input. The reason with a higher value is that an average over n events is calculated.

3 Experimental Design and First Results

In our experiment, the subjects were asked to enter a predefined, 17-digit pass phrase on a smartphone (ten times in a row). For the experiment we used a Samsung Galaxy Nexus. To record the information of the keyboard we implemented a soft keyboard, in addition, to an application where the subject had to type the pass phrase.

As a first evaluation we calculated the standard features (duration time of the keystroke, the digraph and trigraph). Figure 1 shows the extracted data for five randomly selected subjects (like [6,7]). The left figure shows the data of the digraph and while the right represents the data of the duration time.

On the left, it can be seen that the average duration is in most cases more constant over the time. However, there are differences between the people. Even the general speed or the time between single digraphs of one subject differs. The differences between one subject can be explained on the basis of experience. E.g. the fifth person has a lot of experience while subject number four has no experience. This is the reason why the fourth person has a higher value for each

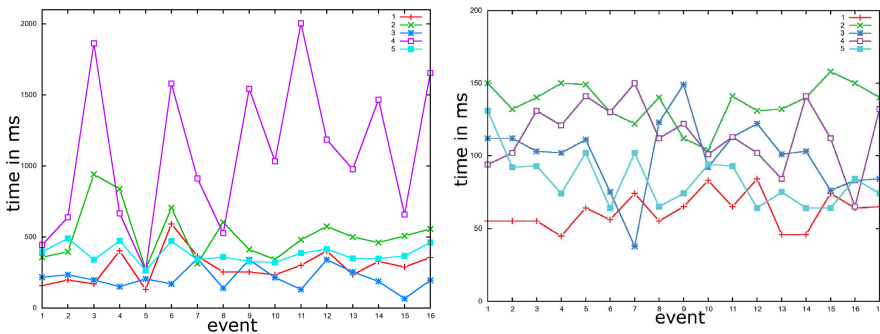


Fig. 1. (left) Digraph for five different users, (right) duration time of the same subjects

digraph than the others. The small value for the digraph at fifth time event for all subjects can be explained by a double letter in the pass phrase. No subject has to search the next letter in this situation.

The rhythm of the duration time (Figure 1 on the right), is less constant between individuals and, in addition, between different attempts by one person. Furthermore, the duration time tends to be less than the digraph. A person needs more time to press the next key than to hold a key.

4 Conclusion

The first result of this experiment shows that there are inter-differences between subjects for the time features and intra-similarities between different attempts of one user. This has to be evaluated more in a bigger experiment. On touchscreen keyboards, which are now installed in nearly every smartphone, besides the well-known features, other possibilities for typing behavior can be recorded. Examples for this are the pressure or the size of the fingertip during typing. These can be used in combination with the time values for authentication [8].

References

1. Joyce, R., Gupta, G.: Identity authentication based on keystroke latencies. *Commun. ACM* 33, 168–176 (1990)
2. Ord, T., Furnell, S.: User authentication for keypad-based devices using keystroke analysis. In: *Proc. 2nd Int'l Network Conf. (INC 2000)*, pp. 263–272 (2000)
3. Monrose, F., Rubin, A.D.: Authentication via keystroke dynamics. In: *Proceedings of the 4th ACM Conf. on Computer and Communications Security, CCS 1997*, pp. 48–56. ACM, New York (1997)
4. Moskovitch, R., Feher, C., Messerman, A., Kirschnick, N., Mustafic, T., Camtepe, A., Löhlein, B., Heister, U., Möller, S., Rokach, L., Elovici, Y.: Identity theft, computers and behavioral biometrics. In: *Proceedings of the 2009 IEEE Intl. Conf. on Intelligence and Security Informatics, ISI 2009*, pp. 155–160. IEEE Press, Piscataway (2009)
5. Choraś, M., Mroczkowski, P.: Keystroke dynamics for biometrics identification. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) *ICANNGA 2007, Part II. LNCS*, vol. 4432, pp. 424–431. Springer, Heidelberg (2007)
6. Lau, E., Liu, X., Xiao, C., Yu, X.: Enhanced user authentication through keystroke biometrics. In: *Computer and Network Security (2004)*
7. Clarke, N.L., Furnell, S.M.: Authenticating mobile phone users using keystroke analysis. *Int. J. Inf. Sec.*, 1–14 (2006)
8. Ross, A., Jain, A.K.: Information fusion in biometrics. *Pattern Recognition Letters* 24, 2115–2125 (2003)

Visibility Assessment of Latent Fingerprints on Challenging Substrates in Spectroscopic Scans

Mario Hildebrandt, Andrey Makrushin, Kun Qian, and Jana Dittmann

Research Group on Multimedia and Security, Otto-von-Guericke University of Magdeburg,
Universitaetsplatz 2, 39106 Magdeburg, Germany
{hildebrandt,makrushin,kqian,dittmann}@iti.cs.uni-magdeburg.de

Abstract. Our objectives for crime scene forensics are to find the substrates on which finger traces are visible in limited ranges of the electromagnetic spectrum using UV-VIS reflection spectroscopy and to determine the optimal ranges in the interval from 163 to 844 nm. We subjectively assess the visibility of fingerprints within detailed scans with a resolution of 500 ppi and compare the results with those of an automatic visibility assessment based on the streakiness score. Ten different substrates are evaluated, each with three fingerprints from different donors. Streakiness score is confirmed to be a suitable fingerprint visibility indicator on non-structured substrates. We identify two substrates, namely metallic paint and blued metal, on which ridge lines become visible exclusively in UV range from 200 to 400 nm and from 210 to 300 nm correspondingly.

Keywords: Forensics, latent fingerprints, UV-VIS spectroscopy, fingerprint visibility assessment.

1 Extended Abstract

As the most common form of trace evidence left at crime scenes, latent fingerprints pose the most challenges for forensic experts [1]. Since they are present rather often yet invisible unless brought out with chemical and/or physical processes designed to enhance the visibility of fingerprint residues [2] (e.g. fuming, powdering, or deposition of chemical reagent), it is important to identify substrate dependent technique or reagent and the appropriate processing sequence for successful enhancement.

In today's crime scene forensics three kinds of substrate properties are known [2][3]: 1) porous substrates (e.g. paper) are absorbent and provide good adhesion for latent fingerprints; 2) non-porous substrates (e.g. glass) do not absorb, so latent fingerprints on such substrates are more susceptible to damage; 3) semi-porous substrates (e.g. glossy cardboard) both resist and absorb fingerprint residue, so it may or may not soak in depending on its viscosity as well as on the absorbent properties of the substrates. Since substrates have different characteristics regarding the absorption of fingerprint residues and light reflection, different fingerprint development techniques are required. On challenging substrates these techniques have an invasive nature as they involve applying alien substances on the fingerprint residue, thus causing

the potential risk of compromising the evidence. Hence, various non-invasive approaches are developed to visualize latent fingerprints, e.g. by using various sensors and light sources [4], or Chromatic White Light (CWL) sensors [3][5][6]. It is common that substrates reflect specular light differently depending on the wavelength of the light source. This effect can be used to render fingerprint residues visible and to acquire them in detailed scans. However, inherent properties of substrates (e.g. translucency or structure) might interfere with the scanning, making the identification of the substrate of crucial importance.

We apply the spectroscopic acquisition with a lateral resolution of 500 ppi in the range from 163 to 844 nm on latent fingerprints deposited on ten substrates and then assess the fingerprint visibility in those detailed scans subjectively and by calculating the streakiness scores [5].

In the subjective visibility evaluation of the fingerprint a label corresponding to its suitability for forensic investigations is assigned. In particular, we distinguish between four classes of fingerprint visibility: 1) invisible, i.e. fingerprint is not perceptible for the naked eye; 2) weak pattern, i.e. fingerprint is partially visible but not usable for forensic investigations; 3) fair pattern, i.e. fingerprint is visible but covered by the substrate's texture or structure, and this usually requires longer investigation time; 4) perfect pattern, i.e. fingerprint can be easily segmented using binarization.

The streakiness score S [5] is used as the objective visibility measure. However, in contrast to [5] we do not use substrate specific filters therefore achieve fully automatic processing. The basic idea is assigning S to determine the visibility of the fingerprint at each particular wavelength. Since S is the relative number of pixels in ridge lines in a binarized fingerprint pattern, its value rests in the interval $[0, 1]$, where 0 denotes a completely invisible fingerprint pattern and 1 a perfectly visible one. It relies on the assumption that a fingerprint forms a striated pattern consisting of local ridge-valley patterns with constant frequencies and directions. A preprocessing in the Fourier domain is applied to enhance the ridge-valley pattern, afterwards an enhancement based on the approach from [7] is applied. The first step consists of multiplying the Fourier representation with a ring filter around its origin (see [8]) followed by a zero-mean normalization. The second step consists of estimating local orientations and a corresponding reliability map, both based on gradients. Furthermore, local frequencies are determined for reliable orientations. This is the foundation for locally applying corresponding Gabor filters for the enhancement of the ridge pattern. At last, the filtered image is binarized to calculate the streakiness score.

As a benchmark, we adapt the concept of differential images from [6] to eliminate possible interference introduced by the substrate and compare the visibility of fingerprints on original and differential images. On cooperative substrates the fingerprint visibility on original and differential images is expected to be equal. On challenging substrates with a constant reflection the originally invisible fingerprints become visible in differential scans. On challenging substrates with a random reflection the visibility is not expected to increase. The original approach subtracts the data from the same area of the substrate without a fingerprint. In our proposed approach, the image with the fingerprint is divided by the values from the images without the fingerprint: The intensity image $I(\lambda_w)$ at one particular wavelength λ_w in the interval $[163, 844]$ nm

with $w \in [0, 2047]$ can be determined by $I(\lambda_w) = R(\lambda_w) \cdot S(\lambda_w) \cdot Sens(\lambda_w)$, where $R(\lambda_w)$ is the reflection of the measured material, $S(\lambda_w)$ is the spectral distribution of the light source, and $Sens(\lambda_w)$ is the sensitivity of the spectrometer. Then $R(\lambda_w)$ can be divided into the reflection of the substrate $R_s(\lambda_w)$ and the specific absorption and fluorescence of the fingerprint residue $R_f(\lambda_w)$. By performing a division of $I(\lambda_w)$ by the spectral data from a reference scan $I_{ref}(\lambda_w)$, $S(\lambda_w)$ and $Sens(\lambda_w)$, as well as $R_s(\lambda_w)$ are eliminated, leading to an enhanced contrast.

In total we evaluate ten substrates that are common at crime scenes as summarized in Table 1, each with three fingerprints from different donors. The substrates are divided into the classes cooperative, moderate and challenging based on observations for the CWL sensor [6]. On cooperative non-structured and non-textured substrates we can assume the following ranges streakiness scores S for the subjective visibility classes: $S < 0.1$ indicates an invisible fingerprint, $0.1 < S < 0.3$ usually corresponds with a weak pattern, $0.3 < S < 0.5$ indicates a fair pattern which probably contains smeared areas, $S > 0.5$ indicates a good to perfect fingerprint visibility. However, since we do not apply substrate specific filters those ranges do not apply on moderate and challenging substrates because higher values of streakiness scores are likely to be caused by surface properties.

Table 1. Substrates, integration times and subjective evaluation results of original and differential images (***) - perfect, ** - good, * - weak, and o - invisible fingerprint pattern)

Substrate	Integration time (ms)	Original image		Differential image		Comparison to CWL [9]	Substrate property
		Wavelength range (nm)	Quality	Wavelength range (nm)	Quality		
White furniture surface	100	163-844	***	163-844	***	identical	non-porous, cooperative
Metallic paint	100	200-385	***	195-385	***	better	non-porous, challenging
Glass	50	210-710	***	200-844	***	identical	non-porous cooperative
Copy paper	250	220-250	*	220-250	**	identical/better	porous, challenging
5 Euro cent coin	100	220-520	***	220-844	***	identical/better	non-porous cooperative
Blued metal	500	210-300	*	210-300	**	better	semi-porous, challenging
Non-metallic paint	250	163-844	O	200-220	*/**	identical/worse	non-porous, challenging
Brushed stainless steel	100	163-844	*/**	210-270	**	worse	non-porous, moderate
Beech veneer	100	163-844	**	163-844	***	identical	non-porous, moderate
Golden oak veneer	250	163-844	o/*	163-844	o/*	worse	non-porous, moderate

On cooperative substrates (white furniture surface, glass and 5 Euro cent coin) the UV-VIS spectroscopy images yield satisfactory visibility throughout the entire spectrum with S exceeding 0.5. On copy paper the images are equally unusable as those captured using a CWL sensor. On beech veneer the visibility is identical with that on CWL images. On non-metallic paint, brushed stainless steel and golden oak veneer the visibility is slightly worse than that of a CWL sensor. The calculation of S on structured surfaces (beech veneer, golden oak veneer, and brushed stainless steel) makes no sense since the streakiness of a surface dominates that of a fingerprint. Most notably, the spectrometer acquires a perfect fingerprint image within the UV band on metallic paint, whereas the transparent layer causes non-deterministic noise while using a CWL sensor [6]. Here, S yields 0.7 between 200 and 400nm and 0.05 for the spectrum of visible light. This can be explained by the fact that the covering transparent paint is non-transparent for UV radiation. On blued metal, fresh fingerprints are slightly visible in the UV band, which is an improvement compared to a CWL sensor.

The important conclusion is that the streakiness score generally reflects judgments of human experts about the fingerprint visibility. However, in order to extract meaningful streakiness scores a substrate specific preprocessing is necessary.

Acknowledgement. The authors wish to thank Michael Ulrich for the list of common substrates from crime scenes. The work in this paper has been funded in part by the German Federal Ministry of Education and Science (BMBF) through the Research Program under Contract No. FKZ: 13N10818.

References

1. Champod, C., et al.: Fingerprints and other ridge skin impressions. CRC Press (2004)
2. Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) et al.: The fingerprint sourcebook, U.S. Department of Justice (2011)
3. Makrushin, A., et al.: Advanced techniques for latent fingerprint detection and validation using a CWL device. In: Proc. SPIE, vol. 8436, p. 84360Y (2012)
4. Bleay, S.M., Sears, V.G., et al.: Fingerprint Source Book, Home Office (2012), <https://www.gov.uk/government/publications/fingerprint-source-book>
5. Makrushin, A., et al.: Visibility enhancement and validation of segmented latent fingerprints in crime scene forensics. In: Proc. SPIE, vol. 8665, p. 866508 (2013)
6. Hildebrandt, M., et al.: Benchmarking contact-less surface measurement devices for fingerprint acquisition in forensic investigations: results for a differential scan approach with a chromatic white light sensor. In: Proc. DSP 2011, pp. 1–6 (2011)
7. Hong, L., Wan, Y., Jain, A.: Fingerprint Image Enhancement: Algorithm and Performance Evaluation. IEEE Trans. on PAMI 30, 777–789 (1998)
8. Wu, C., Tulyakov, S., Govindaraju, V.: Image quality measures for fingerprint image enhancement. In: Gunsel, B., Jain, A.K., Tekalp, A.M., Sankur, B. (eds.) MRCS 2006. LNCS, vol. 4105, pp. 215–222. Springer, Heidelberg (2006)

Creation of a Public Corpus of Contact-Less Acquired Latent Fingerprints without Privacy Implications

Mario Hildebrandt¹, Jennifer Sturm¹, Jana Dittmann¹, and Claus Vielhauer²

¹ Research Group on Multimedia and Security, Otto-von-Guericke University of Magdeburg,
Universitaetsplatz 2, 39106 Magdeburg, Germany

{hildebrandt, sturm, dittmann}@iti.cs.uni-magdeburg.de

² Brandenburg University of Applied Sciences, Magdeburger Str. 50,
14770 Brandenburg an der Havel, Germany

vielhauer@fh-brandenburg.de

Abstract. Data sets of biometric or forensic samples are an important basis for evaluations and research. Especially biometric data is considered as personal data, which is protected by privacy regulations. Since the data cannot be altered or revoked, at least in some countries, this poses a challenge because rights must be granted to the data's subject. In particular in Germany and probably in the entire European Union after its reformation of the data protection legislation it is challenging to use such data. Furthermore, with respect to latent fingerprints only very few public data sets exist nowadays. We propose the creation of a public data set without privacy implications consisting of latent fingerprints from artificial fingerprint patterns. On the foundation of a first set of 50 fingerprints on a compact disk surface we report challenges that need to be solved in order to create realistic samples.

Keywords: Forensics, latent fingerprints, public data set, privacy aspects.

1 Extended Abstract

In biometrics and forensics evaluations of new and existing techniques are very important in order to determine error rates especially when such methods are used e.g. for authentication systems or for the analysis of evidence. In forensics the Daubert challenge [1] comprises several factors that can be assessed by a judge prior to admitting evidence in court. Furthermore, researchers need to show that their proposed methods pose an advance in science and technology. Here, public data sets help comparing the evaluation results with each other without any bias. However, especially biometric data need to be considered in most cases as personal data [2]. Thus, it is covered by privacy regulations and data protection acts. Moreover, the possibility of replicating biometric traits or traces can increase the risk for identity theft [5]. Hence, in some legislations, such as in Germany [3], the public usage of biometric data sets is challenging since certain rights must be granted to the person concerned. Furthermore, it is expected that the reform of the European data protection legislation will strengthen individual rights as well [4], leading to higher barriers in creating and

using public biometric data sets. In contrast to similar fields such as biometric systems where biometric data and templates can be transferred and stored in an encrypted or obfuscated manner (e.g. in fingerprint authentication [9]) to ensure privacy, an access to the original raw fingerprint data is necessary for research purposes and various evaluations. Moreover, in forensics an access to the fingerprint image is necessary because the examiner is responsible for a final decision.

We address this challenge by proposing a data set of contact-less acquired printed artificially created latent fingerprints for the evaluation of forensic techniques. In doing so, we want to generate latent fingerprints as realistic as possible while retaining the ability to detect them as motivated by Kiltz et al. [5]. To avoid any privacy implications artificial fingerprint patterns are generated using SFinGe [6] and afterwards printed using a Canon Pixma iP4950 ink-jet printer with the technique of Schwarz [7]. The intention of SFinGe is the creation of fingerprints for the evaluation of biometric systems. For that, it is supported to add noise, distortions and sensor influences. However, in forensics latent fingerprints can be found on various substrates that potentially require the acquisition with different sensors. The printing process ensures that the fingerprint pattern can be applied to a broad variety of such substrates and thus creating realistic conditions for forensic investigations. In our first experiments, the samples are digitized using a Keyence VK-x110 series confocal laser scanning microscope which captures topography and intensity data using a laser and color data by using a CCD camera. Furthermore, a color-intensity image is generated by combining the color data and the laser data. Since this measurement device stores the digitized trace within a proprietary format, we convert the image data for the public database into simple binary objects consisting of an 8 bit header with the width and the height of the data field followed by either a field of 32 bit little endian floating point values for topography and laser intensity data or 32 bit ARGB values for color and color-intensity data. This allows for analyzing the data with various tools. The four binary objects are accompanied with a meta-data file with the sensor and its parameters which are necessary for the interpretation of the data.

In our poster, we show and discuss the challenges of the creation of the data set on the foundation of 50 samples. In particular, the reproducibility of the printing results and the overall realism of the acquired fingerprints need to be further enhanced in order to replace real biometric traces from the human. The reproducibility of the printing results is primarily affected by the reliability of the printing process. Nozzles or the entire print head tend to be clotted [5] since the artificial sweat has different properties than the manufacturer ink. The realism of the printed latent fingerprints trace is also negatively affected by this effect leading to a visible pattern of amino acid dots instead of continuous ridge line impressions. Hence, in order to address those challenges, we propose a work around to enhance the ridge clarity within the digitized data which connects neighboring dots to create continuous ridges. The method applies a hit or miss operator as a first processing step leading to a binarized image of multiple dots of amino acid. In the second step a triangulation is used to connect the dots. A threshold for the maximum distance between dots is applied to avoid connecting dots between different ridges. The result is a binarized fingerprint pattern similar to the original sample. However, this approach would likely not work

on non-smooth or textured surfaces due to the surface noise. Furthermore, such a processing is unsuitable to achieve realistic traces.

For quality measurement, we adapt a correlation based measure from [8] to determine whether the digitized trace and the original pattern are sufficiently similar. It is based on the Pearson product-moment correlation coefficient of the images. However, this requires an exact alignment and scaling of the data.

In future work the reliability of the printing process and the realism of the printed patterns need to be increased in order to create usable data sets for forensic sciences.

Acknowledgement. The work in this paper has been funded in part by the German Federal Ministry of Education and Science (BMBF) through the Research Program under Contract No. FKZ: 13N10818 and FKZ: 13N10816.

References

1. Dixon, L., Gill, B.: Changes in the Standards for Admitting Expert Evidence in Federal Civil Cases Since the Daubert Decision. RAND Institute for Civil Justice (2001) ISBN: 0-8330-3088-4
2. Article 29 Data Protection Working Party. Opinion 3/2012 on developments in biometric technologies (2012), http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp193_en.pdf
3. Federal data protection act (BDSG) (2010), http://www.bfdi.bund.de/EN/DataProtectionActs/Artikel/BDSG_idFv01092009.pdf?__blob=publicationFile
4. European Commission. How does the data protection reform strengthen citizens' rights? (2012), http://ec.europa.eu/justice/data-protection/document/review2012/factsheets/2_en.pdf
5. Kiltz, S., Hildebrandt, M., Dittmann, J., Vielhauer, C., Kraetzer, C.: Printed fingerprints: a framework and first results towards detection of artificially printed latent fingerprints for forensics. In: Image Quality and System Performance VIII, Proceedings of SPIE, vol. 7867 (2011)
6. Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: Handbook of Fingerprint Recognition, 2nd edn. Springer, London (2009)
7. Schwarz, L.: An amino acid model for latent fingerprints on porous surfaces. Journal of Forensic Sciences 54(6), 1323–1326 (2009)
8. Sturm, J., Hildebrandt, M., Dittmann, J., Vielhauer, C.: High quality training materials to detect printed fingerprints: Benchmarking three different acquisition sensors producing printing templates. In: International Workshop on Biometrics and Forensics, IWBF 2013, Lisbon, Portugal (2013)
9. Barni, M., Bianchi, T., Catalano, D., Di Raimondo, M., Labati, R.D., Failla, P., Fiore, D., Lazzeretti, R., Piuri, V., Scotti, F., Piva, A.: Privacy-preserving fingercode authentication. In: Proceedings of the 12th ACM workshop on Multimedia and Security, MM&Sec 2010, pp. 231–240 (2010)

SocACL: An ASP-Based Access Control Language for Online Social Networks

Edward Caprin and Yan Zhang

Artificial Intelligence Research Group
School of Computing, Engineering and Mathematics
University of Western Sydney, Kingswood, Australia
{e.caprin,y.zhang}@uws.edu.au

Abstract. *Online Social Networks* (OSNs), such as Facebook, encourage their users to disclose significant amounts of personal information to facilitate connecting and sharing content with other users. This has resulted in some OSNs holding vast amounts of information about their users; all of which is readily available via their profile page. As such, OSNs are particularly vulnerable to privacy breach attacks. With the impact these breaches varying from simply embarrassing the user, to negatively influencing the decision of a potential employer, identity theft and even physical harm it is important that these breaches are addressed. In this research we approach privacy management in OSNs as an access control problem, proposing a fine-grained, formal *Attribute-Based Access Control* (ABAC) language; *SocACL* (Social Access Control Language). SocACL is based on *Answer Set Programming* (ASP) and allows for policy specification using the most abundant sources of information available in OSNs; user attributes and relationships.

Keywords: Answer Set Programming, Online Social Networks, privacy, access control, Attribute-Based Access Control.

1 Introduction

Online Social Networks (OSNs), such as Facebook and LinkedIn, encourage their users to disclose significant amounts of personal information to facilitate connecting and sharing content with other users. This has resulted in some OSNs holding vast amounts of information about their users; all of which is readily available via their profile page. As such, OSNs are particularly vulnerable to privacy breach attacks [3]. With the impact these breaches varying from simply embarrassing the user, to negatively influencing the decision of a potential employer, identity theft and even physical harm it is important that these breaches are addressed. OSN operators have responded to privacy concerns by providing user customisable privacy settings. However, these have proven ineffective, often resulting in settings that do not reflect the intentions of the user [5]. This is in part due to the coarse-grained nature of the information on which these settings are based. In this research we approach privacy management in OSNs

as an access control problem, proposing a fine-grained, formal *Attribute-Based Access Control* (ABAC) language; *SocACL* (Social Access Control Language). SocACL is based on *Answer Set Programming* (ASP) and allows for policy specification using the most abundant sources of information available in OSNs; user attributes and relationships.

2 Answer Set Programming (ASP)

The semantics of SocACL is defined as a translation to ASP. ASP is a form of declarative programming well suited to representing domain specific knowledge [1], making it ideal for capturing the wide range of features found in OSNs. An ASP program is a finite set of rules that describes some set of knowledge and are used with inference engines, such as DLV [4], to generate sets of conclusions that can be inferred from the program called *answer sets*, on which we base SocACL's policy evaluation system.

3 SocACL EBNF

```

Query = NAME 'asks' NAME · ACT · OBJ · PURPOSE '?'
Policy = {NAME 'says' (Rule | Definition) ';' }
Rule = Head ['if' Body]
Definition = Def-Obli | Def-RelC | Def-Desc
Head = Auth | Attr ':' SF · PIF | Rel-Dir ':' SF | Dele
Body = ( ['not'] [Prin 'says'] BTerm | Aggr | Cons ){',' Body}
BTerm = Attr | Desc | Rel-Dir | Rel-Sind | Rel-Rind
Auth = ('allow' | 'deny') · Prin · ACT · OBJ · PURPOSE · OBLI-NAME
Attr = Prin · ATTR-NAME [ {·Val} ]
Def-Obli = 'define' · 'obligation' · OBLI-NAME · ACT · Prin · NUM
Def-RelC = 'define' · 'relchain' · RELCHAIN-NAME · ('Body')
Def-Desc = 'define' · 'description' · DESC-NAME · VAR · ('Body')
Aggr = VAR '=' Aggr-Op · VAR · ('Body') | Aggr-Op · VAR · ('Body') · Aggr-Cmp
Aggr-Cmp = ('exactly' | 'atleast' | 'atmost') · Val | 'between' · Val · Val
Aggr-Op = 'count' | 'sum' | 'min' | 'max'
Desc = SUB · 'description' · DESC-NAME
Rel-Dir = SUB · 'relationship' · REL-TYPE · SUB
Rel-Sind = SUB · 'sindRelationship' · RELCHAIN-NAME · SUB
Rel-Rind = SUB · 'rindRelationship' · NUM · SUB
Cons = Val ('<' | '>' | '≤' | '≥' | '=' | '≠') Val
Prin = SUB | OBJ
Val = NAME | VAR | NUM

```

NAMES start with a lowercase letter and can contain letters, numbers and underscores, while VARs start with a uppercase letter. SUB and OBJ is a NAME or VAR that identifies a subject or an object. SF and PIF are the *sensitivity flag* and *primary instance flag*, respectively. These are used during the SocACL negotiation process, which is not covered in this paper.

4 SocACL Example

Suppose we have some hypothetical OSN with a member Alice that has a coworker Bob (eq. (1)) and considers this relationship non-sensitive. She is envious of people enrolled at the prestigious University of Learning (eq. (2)), treating this attribute as non-sensitive and a non-primary instance.

alice **says Me · relationship** · coworker · bob : **ns**; (1)

alice **says Me · envious · Other : ns · np if**
Other · enrolled · “UoL”; (2)

Below we find the ASP translation of eq. (1) and (2) respectively. For eq. (3) and (4) arity 1 denotes the principal providing this attribute. In eq. (4) arity 1, 4, and 5 of “enrolled” are underscores, the anonymous variable of DLV used as a placeholder for values that do not matter for this rule. Meaning that for “enrolled” it does not matter who provides this attribute (arity 1). Arities 4 and 5 are the SF and PIF respectively, since these are used only by the SocACL negotiation process it does not matter what their values are when used as decision criteria.

relationship(alice, alice, bob, coworker, **ns**). (3)

envious(alice, alice, **Other**, **ns**, **np**) ← enrolled(_, **Other**, “UoL”, _, _). (4)

5 Related Work and Conclusion

With relationships an integral part of any OSN there have been various access control framework proposals based on them. *Relationship-Based Access Control* (ReBAC) [2] and its supporting language specifies policies in terms of the accessors relationship with the owner. ReBAC’s modelling of relationships differs from that of SocACL. ReBAC relationships can be composed from “smaller” relationships, e.g. “grandparent” can be composed from “parent parent”, which can also be inverted. With SocACL allowing for distance based relationships these compositions pose a problem; is “grandparent” a 1st- or 2nd-degree relationship? Instead, SocACL allows for indirect relationships to be expressed as a sequence of direct relationships at each “hop”. Furthermore, SocACL relationships can be used in conjunction with attributes and the aggregate operations count, sum, min and max. This allows for rules such as *Allow access to “friends”, with red hair, which have 5 “friends” in common with me*; something not possible with ReBAC.

In SocACL we have an access control language with features tailored to OSNs. SocACL utilises the two most abundant sources of information in OSNs as decision criteria; information about the user and their relationships with others.

References

1. Baral, C.: Knowledge Representation, Reasoning and Declarative Problem Solving, 1st edn. Cambridge University Press (2010)
2. Fong, P.W.L.: Relationship-based access control: protection model and policy language. In: Proc. of the 1st ACM Conf. on Data and Application Security and Privacy, CODASPY 2011, pp. 191–202. ACM, New York (2011)
3. Gao, H., Hu, J., Huang, T., Wang, J., Chen, Y.: Security Issues in Online Social Networks. *IEEE Internet Computing* 15(4), 56–63 (2011)
4. Leone, N., Pfeifer, G., Faber, W., Eiter, T., Gottlob, G., Perri, S., Scarcello, F.: The dlv system for knowledge representation and reasoning. *ACM Trans. Comput. Logic* 7(3), 499–562 (2006)
5. Madejski, M., Johnson, M., Bellovin, S.M.: A Study of Privacy Settings Errors in an Online Social Network. In: Proc. of 2012 IEEE Int. Conf. on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 340–345 (March 2012)

Watermark Resynchronization: An Efficient Approach Based on Eulerian Tours around a Robust Skeleton

Konstantinos Raftopoulos¹, Klimis Ntalianis², Paraskevi Tzouveli³,
Nicolas Tsapatsoulis⁴, Aleatha Parker-Wood^{1,5}, and Marin Ferecatu¹

¹ École SITI, Conservatoire National des Arts et Métiers,
292 Rue St Martin FR-75141 Paris Cedex 03, France

{konstantinos.raftopoulos,marin.ferecatu}@cnam.fr

² Technological Educational Institute of Athens, Department of Marketing –
Online Computing Group, Egaleo 12242, Athens, Greece

kda175@gmail.com

³ National Technical University of Athens, Electrical and Computer Engineering Department,
Zografou 15773, Athens, Greece

tpar@image.ntua.gr

⁴ Cyprus University of Technology, Department of Communication and Internet Studies
Limassol, CYPRUS

nicolas.tsapatsoulis@cut.ac.cy

⁵ University of California Santa Cruz, Storage Systems Research Center,
Santa Cruz, California USA

aleatha@soe.ucsc.edu

Abstract. The existing block-based approaches, albeit successful in resisting frequency domain attacks, are sensitive to geometric distortions due to the lack of reference in repositioning the block grid. In this paper an RST-invariant block-based approach is proposed, aiming at protecting image objects. The term “image object” refers to *semantically contiguous* parts of images that have a specific contour boundary. The proposed approach is based on shape information, since the watermark is embedded in image blocks, the location and orientation of which are defined by *Eulerian* tours that are appropriately arranged in layers, around an object’s *robust skeleton*. Images from the Polymathic project are used to illustrate the technique.

Keywords: Object-based Watermarking, Resynchronization, Skeleton Transform, Global-Local Transformation, Eulerian Tours.

1 Introduction

Block-based watermarking approaches using spread spectrum modulation of pseudo-random signals, even though successfully improving watermark resistance to various frequency domain attacks, cannot withstand rotation, because the rectangular grid arrangement is changed and thus synchronization is lost. A reference point for placing the rectangular block-grid for watermark recovery is not easy to obtain after image/object rotation, if *a priori* registration information is not available. The situation

is illustrated in [1] where elaborate methods of high complexity are proposed to alleviate this problem. The main contribution of this paper is that it incorporates a certain noise resistant method of extracting shape information[2] during the block-based watermark embedding process in such a way that the watermarked block's location is readily identifiable at the retrieval phase, even after RST transformation, cropping or Gaussian boundary noise. The proposed system is tested in its ability to provide geometrically resistant copyright protection of semantic objects having an explicit boundary, most suited for protecting certain objects in an image, or explicit creations of artists that have to be distributed and reused in different contexts. Another application is inferring image provenance, allowing artists to find source images quickly by examining watermarks in derivative files, similar to activity inference [3].

2 The Method

A robust skeleton of each object is initially extracted in a vector form. To introduce enhanced noise resistance to the extracted skeleton, we use the Global-Local transformation [2] to map the noise contour back to a smoothed version. Starting then from the skeleton endpoints (marked as small circles in the images below), the pseudorandom watermark sequence is embedded in the DCT domain of the blocks along the skeleton's Eulerian tour. When the first tour is completed the Eulerian tour is extended outwards, in consecutive layers towards the object's boundary until the watermark sequence is spread to the whole object. During watermark detection, initially the skeleton of the candidate object is extracted and the potentially watermarked blocks that are located along the extended Eulerian tour are matched against the respective blocks around the initial skeleton. Only tours starting from the skeleton endpoints are examined. The success of the method stems from the fact that based on the Global-Local transformation, the robust skeleton is extracted from a smooth version of the original contour that retains certain metric information.

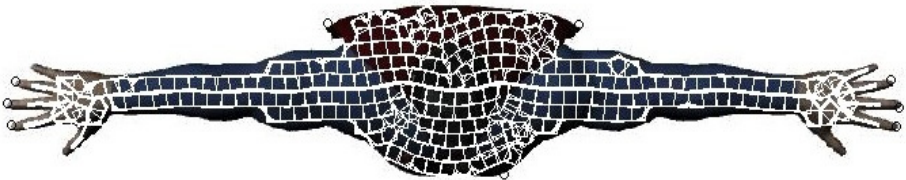


Fig. 1. Watermarked blocks arranged across the Eulerian tours around the skeleton of the Superman top view

3 Experiments



Fig. 2. The location of the watermarked blocks arranged in layers around the skeleton is illustrated for the Superman figure.

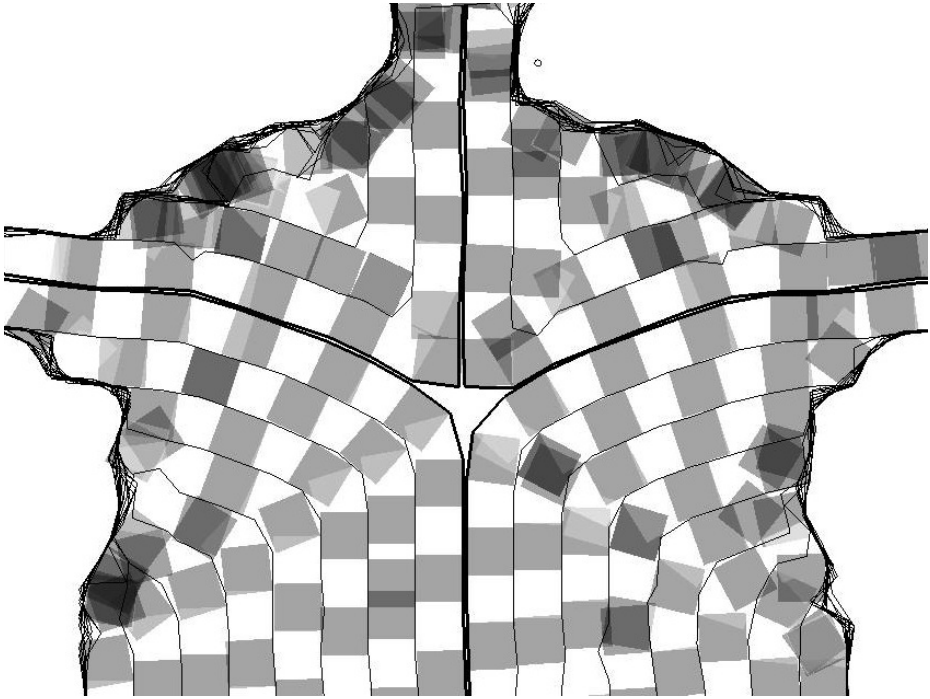


Fig. 3. A closer view of the Eulerian tours at the back view of the Superman image. Some overlap has been allowed for better illustration

Acknowledgements. This research has been supported by the Project FUI Polymathic, financed by the French government.

References

1. Lin, C., Wu, M., Bloom, J., Cox, I., Miller, M., Lui, Y.: Rotation, scale, and translation resilient watermarking for images. *IEEE Transaction on Image Processing* 10(5), 767–782 (2001)
2. Raftopoulos, K.A., Kollias, S.D.: The Global–Local transformation for noise resistant shape representation. *Computer Vision and Image Understanding* 115(8), 1170–1186 (2011)
3. Gaugaz, J., Costache, S., Chirita, P.A., Firan, C.S., Nejd, W.: Activity based links as a ranking factor in semantic desktop search. In: *Latin American Web Conference, IEEE LA-WEB 2008*, pp. 49–57 (2008)

Part IV
Keynotes

Face Recognition from Degraded Images – Super Resolution Approach by Non-adaptive Image-Independent Compressive Sensing Dictionaries

Sabah A. Jassim

Department of Applied Computing, University of Buckingham, Buckingham, UK
sabah.jassim@buckingham.ac.uk

Abstract. In recent years, the emergence of the new paradigm of compressive sensing (CS) has led to the development of innovative image/signal processing and analysis tools that can be exploited to efficiently deal with serious challenges in pattern recognitions. This paper is concerned with the use of CS tools and dictionaries for face recognition, and in particular when dealing with uncontrolled conditions, e.g. faces captured at a distance in surveillance scenarios or in post-rioting forensic, whereby the images are severely degraded/blurred and of low-resolution. We present the results of our recent investigations¹ into the construction of over-complete dictionaries that recover super-resolved face images from any input low-resolution degraded face image. These results demonstrate that non-adaptive image-independent implicitly designed dictionaries that guarantee the recovery of sparse signals achieve face recognition accuracy levels and yield significant recognition rates that are as good as if not better than those achieved by a recently proposed image-based learnt dictionaries. We shall also show that a variety of random dictionaries known to satisfy the Restricted Isometry Property (RIP), achieve similar accuracy rates, and thereby removing the need for training images. The high quality of the super-resolved images provides great potential for forensics and crime/terrorism fighting.

Keywords: Compressive Sensing, Super-resolution, RIP dictionaries, Face recognition, CS classifiers.

1 Introduction

Automatic face recognition in uncontrolled conditions and in particular when dealing with surveillance tasks is far from reliable due to the degraded nature of captured images. Image resolution enhancement is deemed necessary for face recognition, where the camera is at a distant from imaged face yielding small low resolution, blurred and low-quality image for matching. Image degradation results from a variety of recording conditions: subject on the move, unstable sensors, out of focus optical system, or abnormal weather and atmospheric conditions such as thermal waves.

¹ Conducted as part of the PhD research project of Nadia Al-Hassan supervised jointly with Harin Sellahewa.

Recognising faces when matching low-resolution (LR) degraded small images against a gallery of high-resolution good size face images, is traditionally dealt with by preprocessing procedures primarily using the so called super-resolution methods which aims to reconstruct a higher resolution version of the LR image. Hennings-Yeomans *et al* [1] proposed to perform super-resolution and recognition simultaneously. The performance of this method depends on the training database. He and Zhang in [2] have developed an SR technique that constructs a high-resolution face image, from a sequence of low-resolution images, to be processed by Gabor feature based recognition.

The emergence of compressive sensing (CS) theory and sparse representation has led to a plethora of image/signal processing and analysis tools that can be exploited to efficiently deal with serious challenges in pattern recognitions including the face recognition in uncontrolled conditioned. In particular, the development of efficient l_1 -minimization procedures to obtain sparse solutions of certain underdetermined linear systems has led to the emergence of new SR schemes that aim recover high quality super-resolved images from low resolution degraded images ([3], [4], [5], [6], & [7]). Such approaches are motivated by the fact that images in general, and more so degraded ones, can be well-approximated by a sparse expansion in terms of suitable bases such as wavelets. Yang et al in [4], [5] proposed a method to reconstruct super-resolved image from a single low-resolution image using a pair of overcomplete dictionaries D_H and D_L whose columns are constructed, through a learning process, from a number of randomly selected patches of high and low resolution training datasets of face images. This pair of image-trained dictionaries is referred to as the LD system.

In this paper, we briefly discuss CS dictionary construction for various purposes in pattern recognition, but our main focus will on CS based image SR approach for face recognition in uncontrolled conditions. We shall demonstrate that non-adaptive dictionaries, implicitly constructed without using images, perform as well as the LD dictionary, if not better. We introduce an implicit approach to CS dictionary construction, an example of which was developed by the author's team at Buckingham, and investigate its performance in comparison to that of the LD scheme as well as a number of different random dictionaries in terms of the quality of their super-resolved images, face recognition accuracy and CS relevant statistical parameters. For completion, we also present the performance of a non-CS based iterative SR method [8] and of matching in low-resolution.

The rest of the paper is organized as follows. Sections 2 and 3 provide a brief review of Super resolution and Compressive Sensing respectively. In section 4, we shall discuss a recently designed CS approach to image RS using different types of dictionaries, and discuss the properties of these dictionaries that are relevant to the recovery of a sparse signal from a down-sampled degraded version of images. In section 5, we shall conduct experiments to compare the performance of a known face recognition scheme when applied to super-resolved images using the different types of dictionaries as well as to the original LR images. In the conclusion, section 6, we shall briefly describe the contribution of the paper and also highlight benefits of using certain types of implicitly constructed CS dictionaries conclusions.

2 Super Resolution

Super-resolution (SR) is an inverse problem used as a pre-processing technique to recover a high-resolution (HR) image from one or more low-resolution (LR) images. Generally, SR techniques to obtain a HR image from an observed LR input image \mathbf{y} may be modelled as a solution \mathbf{x} of the matrix equation:

$$\mathbf{y} = \mathbf{S}\mathbf{B}\mathbf{x} + \boldsymbol{\eta} \quad (1)$$

where \mathbf{B} is a point-spread function with a blurring effect, \mathbf{S} is a down sampling function, and $\boldsymbol{\eta}$ is additive noise. Various traditional non-CS based super-resolution techniques have been developed, and the most common of these are variants of the Iterative Back Projection (IBP) SR scheme which can super-resolve a single or multiple input LR image(s). The standard single LR image IBP scheme generates the initial HR image \mathbf{x}_0 simply by decimating the pixels of the LR image \mathbf{y} and using Bicubic interpolation. At the n th iteration, $n > 0$, an error image \mathbf{x}_e of the size of the $\mathbf{x}_{(n-1)}$ image is calculated by: (1) convoluting the $\mathbf{x}_{(n-1)}$ image with an appropriate degradation function, (2) down sample the resulting image to obtain $\mathbf{y}_{(n)}$, and (3) \mathbf{x}_e is obtained from $(\mathbf{y} - \mathbf{y}_{(n)})$ by up-sampling. The n th iteration output the n th version of the HR image simply by calculating $\mathbf{x}_{(n)} = (\mathbf{x}_{(n-1)} + \mathbf{x}_e)$, which represents the back projection of the difference $(\mathbf{y} - \mathbf{y}_{(n)})$ onto $\mathbf{x}_{(n-1)}$. The iteration procedure terminates either when the energy of the error term $(\mathbf{y} - \mathbf{y}_{(n)})$ is reduced below a certain threshold or the number of iterations reached a fixed maximum number. In [9], interesting variants of the IBP have been proposed that simply pack-project additional terms in each iteration, representing high frequency information in \mathbf{x}_0 . These variants include the use of Canny edge information and the Gabor filter to preserve edges in different directions.

The main challenge in recovering \mathbf{x} is the modelling of the unknown blurring function. Gaussian functions with different blurring effect have been considered as a suitable model for use in SR procedures, but they do not reflect severe degradation conditions seen in surveillance scenarios. A suitable model can be based on the use of atmospheric turbulence functions of different strengths (i.e. degradation functions that model environmental conditions caused by variation in temperature, wind speed and exposure time) which extends the effect of the Gaussian functions. In the frequency domain such functions are of the form:

$$H(u, v) = e^{-k(u^2+v^2)^{5/6}} \quad (2)$$

where k is a constant that reflects severity of blurring. We label degradation as severe if $k \in]0.045, 0.09]$; mild (similar to most Gaussian blurring functions) if $k \in]0.02, 0.04[$; and low if $k \in]0, 0.02[$, Figure 4 illustrate the effect for different values of k . In what follows, we shall adopt this model of degradation for a number of k values in these ranges to test performance of face recognition from LR images.

3 Compressive Sensing

Compressive sensing, also known as sparse recovery, is a novel paradigm of signal sampling that greatly relaxes the stringent limitations of the conventional

Shannon-Nyquist Sampling Theorem, for signals that can be approximated by a sparse expansion in terms of a suitable basis. Image compression tools (e.g. JPEG and JPEG2000) use a DCT or wavelet transforms to obtain different approximately sparse representation of any input image. The concatenation of 2 bases one constructed from wavelet functions and the other from sinusoid functions is expected to be of benefits for image processing/analysis tasks, [10]. Each of these bases provides. The underlying principle of CS is that the number of linear measurements needed to reconstruct a compressed signal should be proportional to the compressed size of the signal, not the uncompressed size. The central challenge for CS is the construction of preferably non-adaptive relatively small number of linear measurements that can guarantee the reconstruction of a sparse or approximately sparse signal. Such a set of linear measurements are represented by rows of an over complete dictionary, [11], i.e. an $m \times n$ matrix whose columns form a spanning set of m -dimensional vectors to be used to decompose the signal. Dictionaries generalize vector space basis, and are represented by overcomplete $m \times n$ matrices, ($m \ll n$), whose columns are expected to form a pool of \mathbb{R}^m bases. In this case, any vector in \mathbb{R}^m can have multiple representations in terms of the different bases each capturing different features perhaps at different scales. A main premise of this work is that good CS dictionaries can be constructed implicitly from certain pools of bases by concatenation.

Once a suitable underdetermined dictionary $\mathbf{D} = \{d_1, d_2, \dots, d_n\} \in \mathbb{R}^{m \times n}$ is created, the main step in CS based tools is then the recovery of the sparsest solution of the equation: $\mathbf{y} = \mathbf{D}\mathbf{x}$, where \mathbf{y} is the observed vector, i.e. finding $\hat{\mathbf{x}} \in \mathbb{R}^n$ such that:

$$\hat{\mathbf{x}} = \min_{\mathbf{y}} \|\mathbf{x}\|_0 \text{ subject to } \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \quad (3)$$

This l_0 -minimization problem, known as the (P_0) problem, is computationally NP-hard. If \mathbf{x} is sparse and \mathbf{D} is suitably selected, then we can find a unique solution of the l_1 - minimisation (P_1) problem:

$$\hat{\mathbf{x}} = \min_{\mathbf{y}} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 . \quad (4)$$

This is a convex optimisation problem which is amenable to linear programming. Note that, if \mathbf{x} is small, then the Least Square (LS) method can be used to solve the corresponding the l_2 - minimisation (P_2) problem:

$$\hat{\mathbf{x}} = (\mathbf{arg\,min}_{\mathbf{x}: \mathbf{D}\mathbf{x}=\mathbf{b}} \|\mathbf{x}\|_2) = \mathbf{D}^*(\mathbf{D}\mathbf{D}^*)^{-1}\mathbf{b}.$$

However, the LS solution is not desired in many applications such as when \mathbf{x} is spiky. Therefore, the use of the l_1 - minimisation to recover the solution of (P_0) problem have been the subject of intense investigations. Bruckstein et al, [10], discuss two basic questions about (P_0) : (1) Under what conditions, does it have a unique solution? and (2) Given a feasible solution, is there a simple test to verify that is a global minimizer? We now discuss the dependence of these uniqueness requirements on certain parameters and properties of the matrix \mathbf{D} .

The *sparke* an $m \times n$ matrix D , denoted by $sp(D)$ is the minimum number of linearly dependent columns of D . It is clear that $sp(D) \leq m+1$. Equality occurs when D has a full row rank, and then D is said to be of **full sparke**.

Theorem 1: (see [10]) If every $(sp(D)-1)$ columns of D are linearly independent then every $(sp(D)/2)$ -sparse x can be recovered uniquely from Dx . //

This theorem provides an efficient strategy, that we adopt here, for the implicit construction of suitable CS-dictionaries by concatenating certain sets of \mathbb{R}^m bases.

An $m \times n$ dictionary D satisfies the Null Space Property (NSP) of order k if for each size k set $\Omega \subset \{1, \dots, n\}$ and nonzero vector $\underline{z} \in \text{Ker}(D)$,

$$\|\underline{z}_\Omega\|_1 \leq \|\underline{z}_{\Omega^c}\|_1,$$

where \underline{z}_A is obtained from \underline{z} by making 0 all coordinates not indexed by $A \subset \{1, \dots, n\}$.

Theorem 2: ([12]) An $m \times n$ dictionary D satisfies NSP of order k iff every k -sparse solution x can be recovered by ℓ_1 -minimization. //

It is not difficult to show that if D satisfies NSP of order k then every k columns of D are linearly independent. Consequently, NSP of order $2k$ guarantee uniqueness by Theorem 1 while Theorem 2 then showing the way of recovering the sparsest solution.

Candes and Tao, [13], introduced Restricted Isometry Property (RIP) as sufficient for ℓ_1 - recovery: an $m \times n$ dictionary D , $m \ll n$, satisfy the RIP of order k if there is a constant $0 < \delta_k < 1$, such that for any k -sparse signal $x \in \mathbb{R}^n$:

$$(1 - \delta_k)\|x\|_2^2 \leq \|Dx\|_2^2 \leq (1 + \delta_k)\|x\|_2^2 \tag{5}$$

The smallest δ_k is called the restricted Isometry constant (RIC) of order k , and if D satisfies RIP of order k , then any $2k$ -columns sub-matrix of D must be well- conditioned, [14]. The condition number of a matrix is the ratio of its maximum to the minimum singular values. Checking this property for all $2k$ - columns submatrices is computationally infeasible as it requires exhaustive check of all $\binom{n}{2k}$ submatrices. The statistical version of the restricted Isometry property (STRIP) provides a computationally easier to check version of the RIP property. It requires computation of condition numbers of sufficiently large uniformly randomly selected such submatrices. Gan *et al* [15], developed a STRIP performance bound in terms of the mutual coherence μ of the dictionary which is an indicator of the dependence between columns of the matrix. The coherence of a matrix provides information about the likelihood of guaranteed recovery of the sparse solution, and is defined as the largest absolute normalized inner product of pairs of columns a_i and a_j , i.e.

$$\mu(A) = \max_{1 \leq i < j \leq n} \frac{|(a_i, a_j)|}{\|a_i\|_2 \|a_j\|_2} \quad (6)$$

It is not difficult to show that if D is a dictionary with unit column vector and coherence μ then D satisfies RIP of order k with $\delta_k \leq (k-1)\mu$, (see [15]). When $n \gg m$, the coherence value has been shown, [15], to be bounded below by $\frac{1}{\sqrt{m}}$, which is reasonably tighter than the Welch bound $\sqrt{\frac{n-m}{m(n-1)}}$.

Due to unfeasibility of exhaustive search, we shall follow a statistical sampling approach when estimating the strength of RIP of the various dictionaries in terms of the condition number and coherence, or when testing for linear independence of the $2k$ columns submatrices of the dictionary. There are a number of efficient sparse recovery algorithms that have been developed including the Homotopy method (LARS) and the Iteratively Reweighted Least Square method (IRLA), [12].

4 CS-Based Superresolution

Here, we briefly describe a recently developed CS-based method to super-resolve low resolution degraded images which uses underdetermined dictionaries that are assumed to satisfy RIP. We list a number of approaches to dictionary construction including the LD approach, random constructions, and a new construction strategy that is independent of training images but designed to implicitly be of full sparke. We shall test the strength of RIP, as described above, for the constructed dictionaries and use the corresponding pairs of dictionaries to reconstruct super resolved images from low resolution degraded face images.

4.1 CS Approach for Image SR

It requires the use of 2 CS dictionaries: a Low resolution matrix D_L of size 100×512 and a High resolution matrix D_H of size 25×512 . The input to this scheme is a degraded low row resolution small image L_r , and the output is super-resolved to double the size image that is meant to be of "high quality". The L_r image is resized by decimating its pixels and Bi-cubic interpolation to obtain double the size image LR which is still degraded. Three more versions of the LR image are created by applying 3 spatial filters to highlights edges in different directions. These four images are subdivided into blocks of size 5, and the pixel values in the 4 versions are turned into a column vector of $100=4 \times 25$ by concatenation. In order to avoid the appearance of blocking artefacts, the LR image will be subdivided into overlapping blocks.

Initialise a HR image of the same size of the LR image for the super-resolved image. The 5×5 blocks are then processed iteratively as follows:

1. Let \mathbf{y} be corresponding 100-dimensional vector.
2. Find the sparse solution \mathbf{z} of the underdetermined equation $\mathbf{y} = \mathbf{D}_L \mathbf{z}$.
3. Calculate the 25-dimensional HR patch \mathbf{x} using the matrix multiplication $\mathbf{x} = \mathbf{D}_H \mathbf{z}$,
4. Back-projection the 2-dimensional 5x5 version of \mathbf{x} onto the existing HR.

In the rest of this section we describe the pairs of Dictionaries \mathbf{D}_L and \mathbf{D}_H for the various dictionary construction strategies adopted in this paper.

4.2 The Image-Based Learnt Dictionary

This construction was proposed by Yang et al, (see [4], [5], & [6]) and used for super-resolution based face recognition. It is based on learning dictionaries using patches from a large training set of high resolution images of good quality that exhibit similar statistical characteristics of the pattern recognition task under investigation. We shall refer to this construction as LD dictionary. The \mathbf{D}_H and \mathbf{D}_L dictionaries are created as follows:

1. A sufficiently large number of high resolution (HR) images (here Face images) are selected and each divided into patches of 5x5 pixels. Patches overlap.
2. Randomly sampling raw patches from a training HR images, and each selected patch is transformed into a normalised vector that is added as a column to the \mathbf{D}_H .
3. Generate a set LR of blurred versions of the HR images, and create 3 other filtered versions, and the columns of \mathbf{D}_L are constructed in a similar way as in above, but by concatenating the patches from the LR images and their 3 filtered version. Again the columns are to be normalised.

4.3 Random Dictionaries

CS randomly constructed measurement matrices that satisfy the Restricted Isometry Condition include Gaussian, Toeplitz and Circular random Matrices. For Gaussian Random Matrix (GRM), the entries $x_{i,j}$ of the CS matrix of size $m \times n$ are independently sampled from a normal distribution $x_{i,j} \sim N(0, 1/m)$, the l_2 -norm was used to normalize each columns in the dictionary. In order to recover super resolved image from a single LR image for face recognition via sparse representation, two overcomplete dictionaries \mathbf{D}_H , \mathbf{D}_L of size 25×512 and 100×512 respectively have been generated from a zero mean Gaussian distribution with variance $1/25$.

Toeplitz-Circular Random measurement matrix (TCRM) are another class of RIP dictionaries that have been widely used. Bajwa et al. [16], have shown that Toeplitz-structured matrices are sufficient to recover undersampled sparse signals. Toeplitz and Circular matrices of the size $k \times n$ are respectively of the form:

$$T = \begin{bmatrix} t_n & t_{n-1} & \dots & t_1 \\ t_{n+1} & t_n & \dots & t_2 \\ \vdots & \vdots & \ddots & \vdots \\ t_{n+k-1} & t_{n+k-2} & \dots & t_k \end{bmatrix}, \text{ and } C = \begin{bmatrix} t_n & t_{n-1} & \dots & t_1 \\ t_1 & t_n & \dots & t_2 \\ \vdots & \vdots & \ddots & \vdots \\ t_{n-1} & t_{n-2} & \dots & t_k \end{bmatrix}$$

For image reconstruction, D_H and D_L are generated as TCRM matrices, where the first row consists of standard Gaussian random variables, and the rest of the rows are permuted versions of the first row as shown above.

4.4 Iteratively Constructed Full Spark Dictionaries

Full-sparke dictionaries is class of full row rank overcomplete $m \times n$ dictionaries, where $m \ll n$, so that each m -columns sub-matrix is a basis of \mathbb{R}^m . Here we describe an example on how to construct such matrices by starting with an invertible $m \times m$ matrix and iteratively appending a set of image independent linearly independent m -column vectors in \mathbb{R}^m while maintaining the full sparke property after every addition. One way to maintain the full sparke is to insist that every new column can only be generated by the full columns of the previous inserted submatrices. In this, paper we present a simple example of such a dictionary, but in the future we shall investigate algebraic construction method using group finite actions on \mathbb{R}^m .

Our example of full sparke dictionaries, referred to as LID, is of the form:

$$D = [A_{p_1}, A_{p_2}, \dots, A_{p_k}, C(A_{p_{k+1}})].$$

For $i=1, \dots, k+1$, the p_i 's, are distinct real numbers >1 , and

$$A_{p_i} = \begin{pmatrix} 1 & \frac{1}{p_i} & \frac{1}{p_i^2} \dots & \frac{1}{p_i^{m-1}} \\ \frac{1}{p_i} & 1 & \frac{1}{p_i} \dots & \frac{1}{p_i^{m-2}} \\ & & \vdots & \\ \frac{1}{p_i^{m-1}} & \frac{1}{p_i^{m-2}} & \frac{1}{p_i^{m-3}} \dots & 1 \end{pmatrix}.$$

Note that $k = \lfloor n/m \rfloor$ and the last sub-matrix of D is simply the first $(n-km)$ columns. Then, the $m \times n$ LID dictionary is obtained from the following matrix after normalising its columns using the l_2 -norm.

For our experimental purposes we the LID high-dictionary D_H is generated from using integers $p_i > 1$. For simplicity, the low-dictionary D_L was created from a Standard Gaussian Random Matrix (GRM).

4.5 Comparison of RIP Parameters for Different D_H Dictionaries

Here we present some comparisons of the “strength” of the RIP for the LD and LID dictionaries. Whenever exhaustive search is infeasible we conducted a statistical testing by taking random sample of 100 cases. To test for *full sparke* property, we

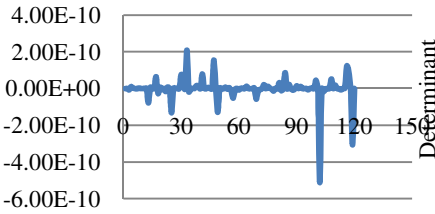


Fig. 1. Determinant of submatrices from the High-resolution LD Dictionary

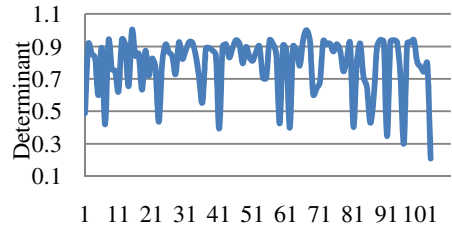


Fig. 2. Determinant of submatrices from the High-resolution LID dictionary

evaluated the determinants, as indicator of linear independence, for more than a hundred randomly selected sample of 25×25 submatrices of the corresponding D_H dictionaries. Although in theory, the LD dictionary may statistically satisfy NSP of order 12, Figure 1 shows that the determinant of most 25×25 submatrices is so small (almost zero) and hence the full sparke property is not satisfied. In contrast, figure 2 confirm that the LID is indeed fully sparke.

The next experiment to calculate another RIP indicator, namely the condition number of 25×25 submatrices of 4 of the DH dictionaries for LD, LID, GRM and theTCRM. These condition numbers are expected to be bounded by RIC of order $2k$, with $k=12$. Table 1, below, displays the mean and standard deviation of the condition numbers for 100 randomly selected submatrices and the condition number of the full size 25×512 matrix.

Table 1. Mean and Standard deviation for CN for a hundred random sub-matrices of different sizes

submatrices	Dictionaries							
	LID		LD		GRM		TCRM	
	mean	std	mean	std	mean	std	mean	std
25x25	3.08	3.14	3.34E+16	1.79 E+17	279.36	597.79	85.19	155.68
Full matrix	1.977		1.00E+15		1.43		1.453	

These results again demonstrate that the overcomplete LID dictionary is well-conditioned in comparison to all others for the various submatrices but for the full matrix GRM and TCRM have similar condition numbers that are better than the LID. Moreover, the condition number of the LD is extremely large for all cases, which make these dictionaries very ill conditioned.

Another test relates to calculating the row-rank and coherence values for the various dictionaries. It is well known that the highest sparsity recovered signal for any dictionary $= (1 + \text{row rank})/2$, and coherence μ must satisfy $0.2 = 1/\sqrt{m} \leq \mu \leq 1$. Again, results in Table 2 highlight the superiority of the LID dictionary.

Table 2. Row-rank and Coherence

Dictionary	LID	LD	GRM	TCRM
Row Rank	25	24	25	25
Coherence	0.9958	< 0.2	0.7438	0.7318

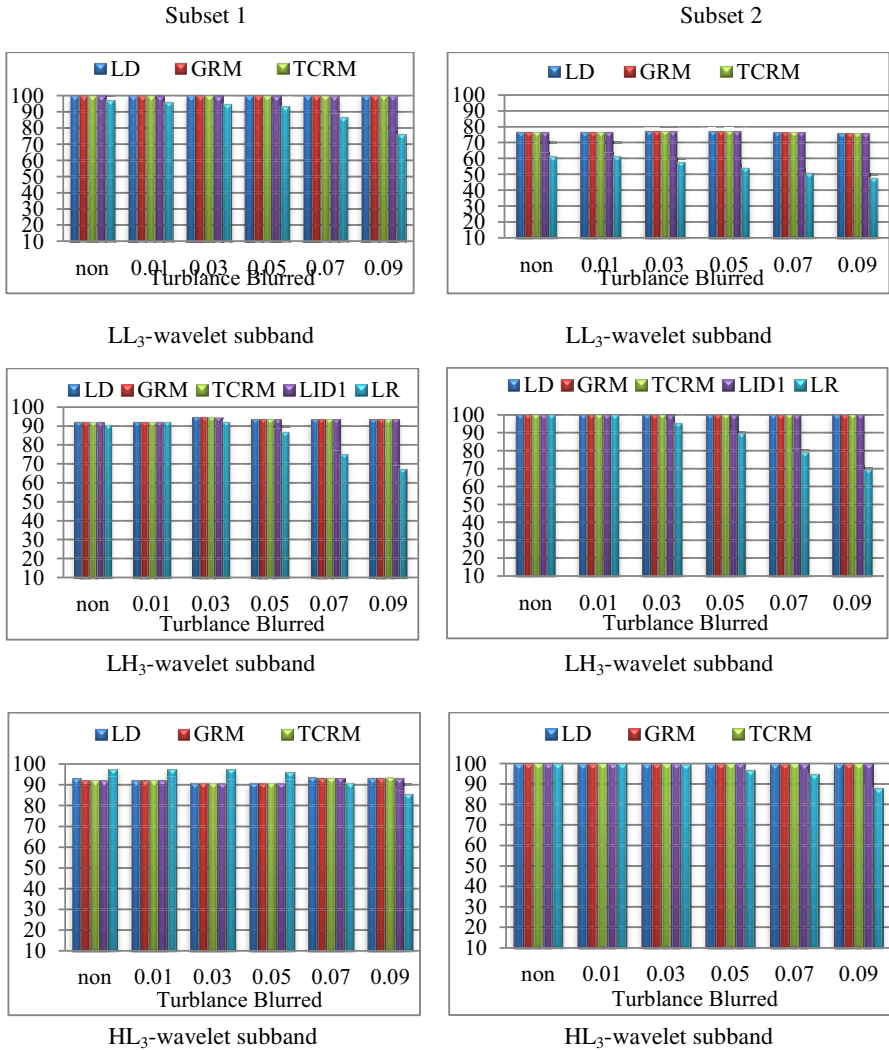


Fig. 3. Recognition accuracy rates using different dictionaries and in comparisons with matching in low-resolution

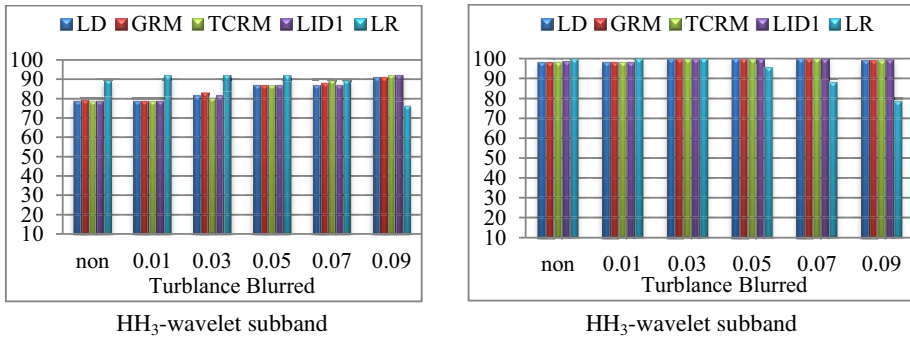


Fig. 3. (Continued)

Finally, in order to test the viability of the CS-based SR scheme, we compared the visual quality of large set of SR enhanced images obtained from the application of the various discussed dictionaries, as well as the IISR scheme and interpolation methods. The results of the visual inspection reveal that there is little difference in the quality of the recovered HR image using the various dictionaries, but a noticeable improvement that can be noticed in SR methods, including IISR, over the low-resolution images and interpolation methods at every level of degradation. Figure 3, below, show one example but this pattern was repeated over all the images. PSNR values calculated between the output SR image and the original images confirm the same pattern, but we omit these results. Unsurprisingly and regardless of the method used in the SR procedure, the quality of SR images decreases as the level of blurring increases. With increased level of blurring there is no difference in image quality obtained by different dictionary methods. But the dictionary methods produced slight improvement on the IISR method, and superiority over the interpolation methods at every level of blurring.

5 Face Recognition – Experimental Results

In this section, we shall compare the performance of face recognition using the corresponding super-resolved images. We test the performance of different dictionary methods as well as a state-of-the-art methods to reconstruct super resolved face image from a single LR image with different magnification blur. The image sets were sampled from a publically available face database and, comparisons recognition rates presented with matching in low-resolution domain. We use a simple but efficient wavelet-based face recognition scheme, whereby the training as well the matching image are wavelet decomposed to level 3 and each of the each of the subbands at level 3 (i.e. LL3, HL3, LH3 and HH3) is used as a face feature vector and Euclidian distance is used for matching against the face feature vectors of the templates.

To test the performance of the wavelet face recognition schemes post the CS-based SR preprocessing schemes, we used face images from the Extended Yale B database. This database consists of 2,414 frontal-face images of 38 individuals. The cropped

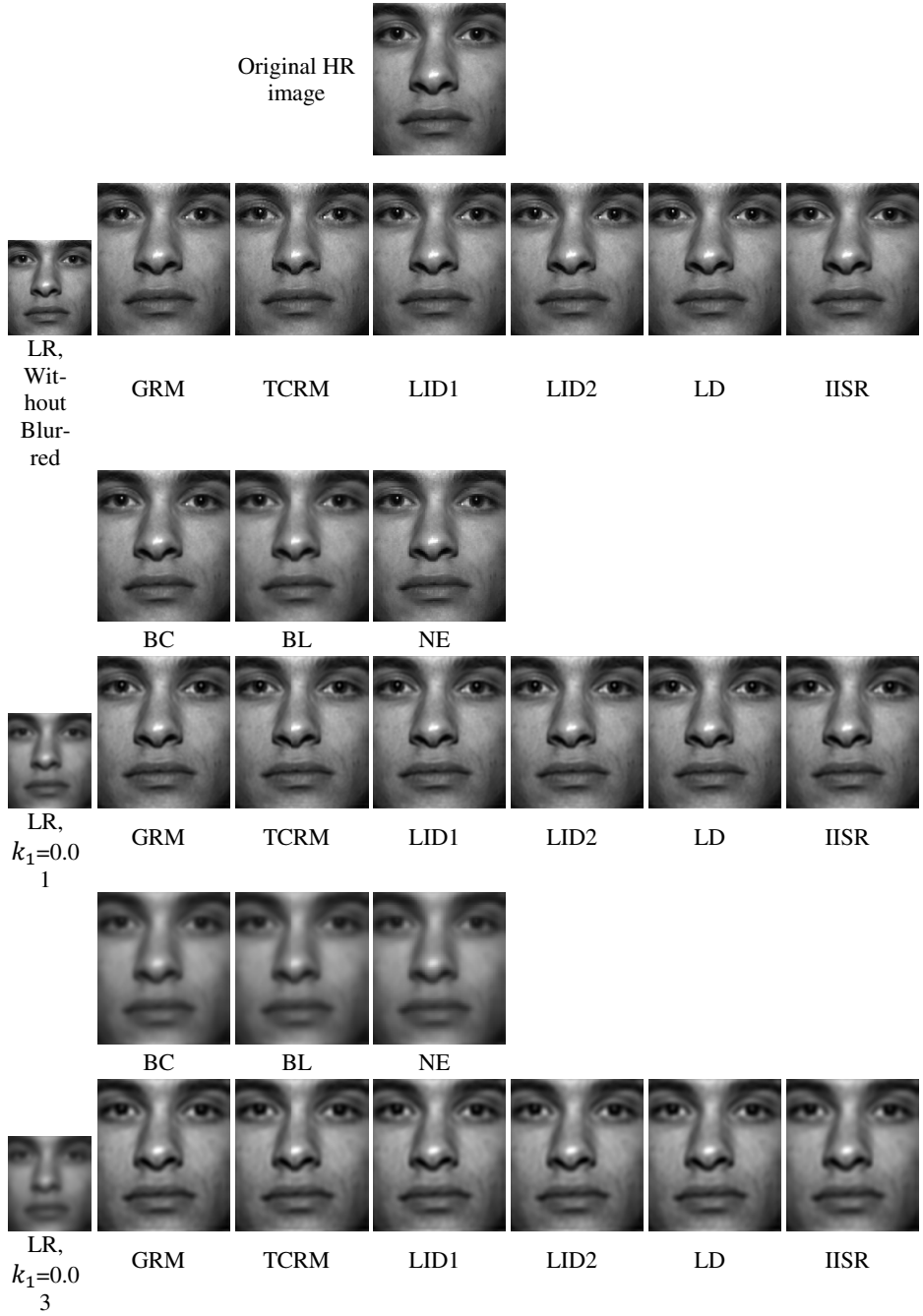


Fig. 4. Super Resolved image by different SR methods and well non-interpolation methods

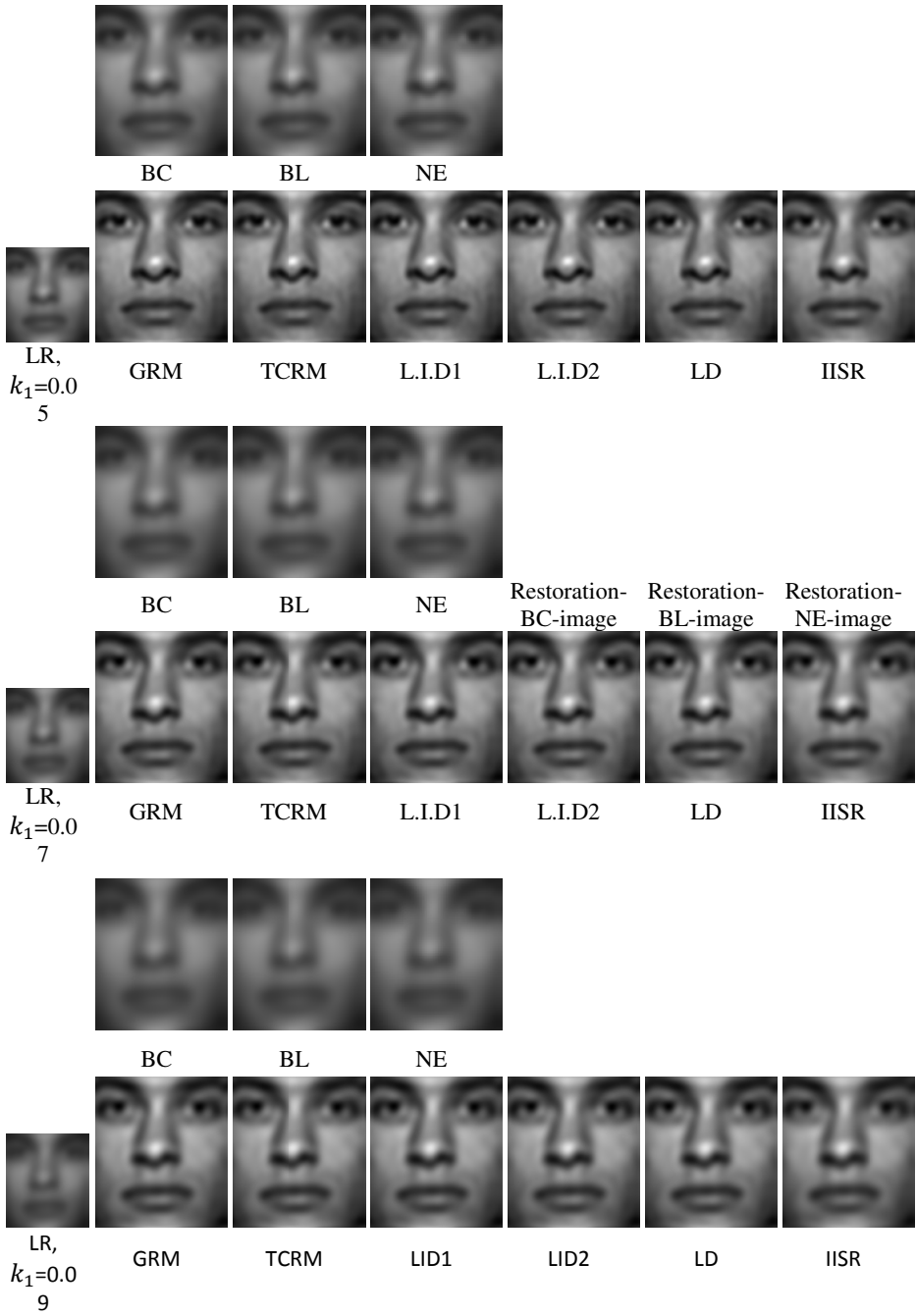


Fig. 4. (Continued)

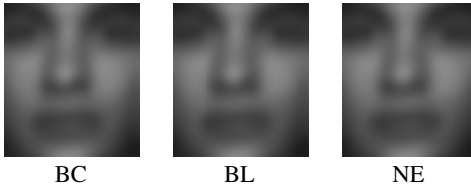


Fig. 4. (Continued)

and normalized 192×168 face images were captured under various laboratory-controlled lighting conditions. For each subject, we selected the P00A+00E+00 image for the gallery set and the other images for testing. To construct LD that depends on images, three images for each subject were selected from the well-lit face images in set 1 that were not included in the gallery/test images.

To simulate face recognition under the recording condition of low resolution degraded quality images, we first apply the degradation function defined by equation 2 for different values of k on the high resolution images in the reasonably lit sets (1 and 2) of the database. The low resolution degraded Lr images are finally obtained by down-sampling the degraded images by a factor of 2.

Each test Lr face image is first super-resolved using each of the built dictionaries, the IISR scheme or simply by up-sampling and interpolation.

For feature extraction, on the resulting HR templates as well as the SR-resolved test images, we use the Z-score normalized coefficients of the the subbands of the Haar wavelet decomposed face images at level three. Matching will be based on the City Block distance between a probe image and a gallery image.

The experimental results are shown in Figure 3 for the different subbands and differently illuminated sets of images. The various charts display the accuracy rate at each level of degradation function. As can be seen, there is no significant difference in identification accuracy rates, between the different dictionaries methods. Moreover, the accuracy rates seem to be maintained at the same level for different degradation level. In comparisons to the method of matching the LR images with down-sampled gallery images, the performance of the dictionary based methods are far more superior and much more apparent as the image quality deteriorates from mild to severe degradation. However, the picture is surprisingly different for mild image degradation (i.e. $k_1 < 0.7$) when we use the HL3 or HH3 subband as feature vectors for set 1 images which are recorded in a slightly better illumination condition. Note that the full set of experiments, part of which are not included here, indicate similar patterns of performance, i.e. a much better accuracy are achieved by all the dictionaries than that achieved by the various interpolation methods such as nearest, bilinear and Bi-cubic and the difference in performance increases as the of degradation get more severe..

The observed pattern of variation in the performance of face recognition, using different sub-bands is consistent with known results for wavelet-based face recognition without degradation reported in the literature, (see [17]). For this work we also conducted face recognition for exactly the same conditions on resolution and blurring but using the Principle Component Analysis technique (PCA). The results exhibit similar patterns and are omitted here.

6 Conclusion

We have studied the RIP property for random, and not so random, constructions of CS related overcomplete dictionaries as well as an existing dictionary that trained on a set of high-resolution face images. These dictionaries were used to generate super resolved image with the aim of using for face recognition in uncontrolled conditions where the input is degraded blurred LR image with a wide range of degradation. This results effectively support the use of SR based techniques that employ CS dictionaries for recovering super-resolved images that are suitable for face recognition. More importantly, that there is no need for using image sets for training dictionaries, because non-adaptive dictionaries perform equally well if not better in some cases. In an attempt to find possible explanation, we conducted a number of tests of numerical matrix parameters relevant to the RIP condition. We note that the learning image-based dictionary is highly ill conditioned and far from satisfying the RIP related conditions discussed in the literature. Perhaps the use of image patches with the same statistical parameters of general face image patches compensate for the lack of RIP properties.

Further studies are needed to test other implicit construction of RIP dictionaries. Indeed, we have developed a new method which aims to implicitly satisfy the known bounds on singular values. Such approaches can be exploited to use RIP dictionaries for revocable face biometric instead of the traditional random projection.

References

- [1] Hennings-Yeomans, P.H., Baker, S., Kumar, B.V.: Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
- [2] He, J., Zhang, D.: Face super-resolution reconstruction and recognition from low-resolution image sequences. In: 2nd International Conference on Computer Engineering and Technology (ICCET), vol. 2, p. V2-620 (2010)
- [3] Zeyde, R., Elad, M., Protter, M.: On Single Image Scale-Up Using Sparse-Representations. In: Boissonnat, J.-D., Chenin, P., Cohen, A., Gout, C., Lyche, T., Mazure, M.-L., Schumaker, L. (eds.) Curves and Surfaces 2011. LNCS, vol. 6920, pp. 711–730. Springer, Heidelberg (2012)
- [4] Yang, J., et al.: Image super-resolution as sparse representation of raw image patches. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
- [5] Yang, J., et al.: Image super-resolution via sparse representation. *IEEE Transactions on Image Processing* 11(19), 2861–2873 (2010)
- [6] Yang, J., et al.: Face hallucination via sparse coding. In: 15th IEEE International Conference on Image Processing, pp. 1264–1267 (2008)
- [7] Wang, S., Zhang, L., Liang, Y., Pan, Q.: Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis, pp. 2216–2223 (2012)
- [8] AL-Hassan, N., Jassim, S.A., Sellahewa, H.: Enhancing face recognition at a distance using super resolution. In: *MMSec*, pp. 123–132. ACM (2012)

- [9] Makwana, R.R., Mehta, N.D.: Single Image Super-Resolution VIA Iterative Back Projection Based Canny Edge Detection and a Gabor Filter Prior. *International Journal of Soft Computing and Engineering (IJSCE)* 3(1), 2231–2307 (2013)
- [10] Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review* 51(1), 34–81 (2009)
- [11] Rubinstein, R., Bruckstein, A.M., Elad, M.: Dictionaries for sparse representation modeling. *Proceedings of the IEEE* 6(98), 1045–1057 (2010)
- [12] Fornasier, M., Rauhut, H.: Compressive Sensing. In: Scherzer, O. (ed.) *Handbook of Mathematical Methods in Imaging*, pp. 187–228. Springer (2011)
- [13] Candes, E.J., Tao, T.: Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies? *IEEE Transaction on Information Theory* 12(52), 5406–5425 (2006)
- [14] Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. *Constructive Approximation* 3(28), 253–263 (2008)
- [15] Gan, L., Ling, C., Do, T.T., Tran, T.D.: Analysis of the statistical restricted isometry property for deterministic sensing matrices using Stein’s method. Citeseer (2009)
- [16] Bajwa, W.U., Haupt, J.D., Raz, G.M.: Toeplitz-Structured Compressed Sensing Matrices. In: *IEEE 14th Workshop on SSP 2007*, pp. 294–298 (2007)
- [17] Al-Obaydy, W., Sellahewa, H.: On using high-definition body worn cameras for face recognition from a distance. In: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N.C., Fairhurst, M.C. (eds.) *BioID 2011. LNCS*, vol. 6583, pp. 193–204. Springer, Heidelberg (2011)

Trustworthy Software Development

Sachar Paulus¹, Nazila Gol Mohammadi², and Thorsten Weyer²

¹Department of Economics, Brandenburg University of Applied Sciences,
14770 Brandenburg an der Havel, Germany

paulus@fh-brandenburg.de

²paluno – The Ruhr Institute for Software Technology,
University of Duisburg-Essen, 45127 Essen, Germany

{nazila.golmohammadi, thorsten.weyer}@paluno.uni-due.de

Abstract. This paper presents an overview on how existing development methodologies and practices support the creation of trustworthy software. Trustworthy software is key for a successful and trusted usage of software, specifically in the Cloud. To better understand what trustworthy software applications actually mean, the concepts of trustworthiness and trust are defined and put in contrast to each other. Furthermore, we identify attributes of software applications that support trustworthiness. Based on this groundwork, some well-known software development methodologies and best practices are analyzed with respect on how they support the systematic engineering of trustworthy software. Finally, the state of the art is discussed in a qualitative way, and an outlook on necessary research efforts and technological innovations is given.

Keywords: Software development, Trustworthiness, Trust, Trustworthy software, Trustworthy development practices.

1 Introduction

In the last years, many attempts have been made to overcome the issue of insecure and untrusted software. A number of terms have been used to catch the expectation on how “solid” a piece of software should be: secure, safe, dependable and trusted. Only in recent years literature related to (secure) software developments has seen the introduction of socio-technical systems (STS) (for more details, see [1]). This concept allows to distinguish between the actual trust that users of software put into the functioning / delivery of the software in questions on the one side, and the trustworthiness of the software, i.e. properties (we call them attributes) that justify the trust that users put “into” the software. Whereas trust should primarily be the subject of the “maintenance” of the relationship between the user and the software in use (“in operations”), trustworthiness is primarily acquired during the development process of the software and can mostly only be “lost” later on.

The software creation process, neither, has been addressed adequately both in theory and practice until recently regarding topics like trust, trustworthiness or similar, except either purely theoretical approaches (such as formal proofs or other forms

of verification (e.g. [2]) or on a functional level only (using e.g. security patterns [3]). As such, an analysis of existing software development practices / methodologies with a specific view on trustworthiness is new to the field. This research has been carried out as part of the OPTET project, and the results will be presented in this paper in adequate detail. As an overview publication, it summarizes results of other very recent publications [1].

This paper is structured as follows: in a first section, we define the notions of trust and trustworthiness and introduce the concept of trustworthiness attributes. The next section presents the analysis of the different development methodologies and practices in light of trustworthiness, followed by an analysis section on the state-of-the-art to summarize what is available today, and where there is more research needed to achieve the goal of trustworthy software. A last section summarizes the research carried out and shortly indicates the future work planned in the OPTET project.

2 Fundamentals

In this section we introduce the two basic concepts “trust” and “trustworthiness” in order to be able to analyze how trustworthiness is addressed by different software development disciplines. Both concepts focus on the outcome of the STS but are different in the view of the trustor and trustee(s) perspective. In general, trust is the trustor's prior estimation that an STS will provide an appropriate outcome, while trustworthiness is the probability that the same STS will successfully meet all of the trustors' requirements. The balance between trust and trustworthiness is a core issue for software development because any imbalance (over-cautiousness or misplaced trust) could lead to serious negative impact, e.g. concerning the acceptance of the software by its (potential) users.

2.1 The Notion “Trust”

We define trust in a system as a property of each individual trustor, expressed in terms of probabilities and reflecting the strength of their belief that engaging in the system for some purpose will produce an acceptable outcome. Thus, trust characterizes a state where the outcome is still unknown, based on each trustor's subjective perceptions and requirements. A stakeholder would decide to place trust on an STS if his trust criterion was successfully met; in other words, their perceptions exceed or meet its requirements. A trustor having engaged in a system for multiple transactions can (or will) update the current trust level of that STS by observing past outcomes.

A presence of subjective factors in trust decisions means that two different trustors may have different levels of trust for the same STS to provide the same outcome in the future, even if they both have observed exactly the same system outcomes in the past. More specifically, subjective perceptions can depend on trustor attributes, which capture social factors such as age, gender, cultural background, level of experience with Internet-based applications, and view on laws. Subjective requirements, on the other hand, are represented by so-called trust attributes that quantify the anticipated

utility gains or losses with respect to each anticipated outcome. Thus, relatively high levels of trust alone may not be adequate to determine a positive decision (e.g., if the minimum thresholds from requirements are even higher). Similarly, it is possible to engage in a system even if one's trust for an acceptable outcome is low (e.g., if the utility gains from this outcome are sufficiently high).

2.2 The Notion “Trustworthiness”

We regard trustworthiness as an objective property of the STS, based on the existence (or nonexistence) of appropriate properties and countermeasures that reduce the likelihood of unacceptable outcomes. A stakeholder (e.g., the system designer, a party performing certification) shall decide to what extent a system is trustworthy based on trustworthiness criteria. These criteria are logical expressions in terms of systems attributes, referred to as quality attributes. For example, trustworthiness may be evaluated with respect to the confidentiality of sensitive information, the integrity of valuable information, the availability of critical data, the response time or accuracy of outputs. Such quality attributes shall be quantified by measuring systems' (or individual components') properties and/or behavior. Objectivity in assessing trustworthiness for a particular attribute is based on meeting certain predefined metrics for this attribute or based on compliance of the design process for this attribute to our predefined system specifications.

Thus, the trustworthiness of an STS may be evaluated compared to a target performance level, or the target may be its ability to prevent a threat from becoming active. Such issues are defined by the trustworthiness attributes that have a dual interpretation. Until recently, trustworthiness was primarily investigated from a security or loyalty perspective while assuming that single properties (certification, certain technologies or methodologies) of services lead to trustworthiness and even to trust in it by users. Compared to this approach, we reasonably assume that such a one-dimensional approach is insufficient to capture all the factors that contribute to an STS's trustworthiness and instead we consider a multitude of attributes.

In this paper, our definition for trustworthiness attributes reflects the design-time aspects. A trustworthiness attribute in this sense is a property of the system that indicates its capability to prevent potential threats to cause an unexpected and undesired outcome, e.g., a resilience assurance that it will not produce an unacceptable outcome.

2.3 Trustworthiness of a Software Application

In order to prove to be trustworthy, software applications could promise to cover a set of various quality attributes [1],[4] depending on their domain and target users. Trustworthiness should promise a wide spectrum including reliability, security, performance, and user experience. But trustworthiness is domain- and application-dependent, and a relative attribute that means that if a system is trustworthy with respect to some Quality of Service (QoS) like performance, it would not necessarily be successful in being secure. Consequently, trustworthiness and trust should not be regarded as a single construct with a single effect, they are rather strongly context

systematic realization of trustworthiness to a system under development. In next section, the result and evaluation of these studies is presented.

3 Review of Development Models and Practices

Recently, a number of development practices have been proposed, both from a theoretical as well as from a practical point of view, to address security of the software to-be-developed. As described above, security is an important component of trustworthy software, but neither is it the only one, nor will it be sufficient to look solely at preserving / creating a good level of security to attain trustworthiness. For example, transparency plays an important role for the creation of trust, and therefore for the trustworthiness of software.

In this section, we will look into the major software engineering processes or process enhancements that target security to build a „secure“ software system and identify corresponding innovation potential, specifically towards extending security to trustworthiness. A more exhaustive overview of development methodologies can for instance be found in Jayaswal and Patton's "Design for Trustworthy Software" [20], though it does not specify how these methodologies contribute to the trustworthiness of the product. This reference documents their generic characteristics and an overview of the historical evolution of different development strategies and lifecycle models.

We will briefly describe which elements of the development approaches will actually increase or inhibit trust, and how the approaches could be used for modeling trustworthiness.

3.1 Plan-Driven

In a plan-driven process [17] one typically plans and schedules all of the process activities before the work can start. The *Waterfall model* is a well-known example of plan-driven development that typically includes the following phases:

- Requirements analysis
- System design
- Implementation
- Testing (unit, integration and system testing)
- Deployment
- Operation and maintenance

Many of the simplistic software manufacturing projects follow a plan-driven model. This approach has been followed by industrial software development for a long time. It is relatively easy to assure non-functional requirements throughout the rest of the process, but the key issue is that they need to be identified completely in the first phase. Plan-driven processes such as the Waterfall model originate from aerospace and other manufacturing industries, where robustness and correctness is usually an important concern, but are often considered being too rigorous, inflexible and a bit old-fashioned for many software development projects. There are examples of

Waterfall trustworthy software development processes in the literature, e.g. COCOMO. Therefore, there should be means to assure trustworthiness and enhance the process. There can be more formal variants of this process, for instance the B method [21], where a mathematical model of the specification is created and then automatically transferred into code. For the general plan-driven process we consider the following trustworthiness characteristics to be valid:

Trustworthiness gains:

- Formal system variants are well suited to the development of systems that have stringent safety, reliability or security – and thus potentially also trustworthiness – requirements.

Trustworthiness losses:

- Vulnerable with vague, missing or incorrect security and trustworthiness requirements in the first place.
- Does not offer significant cost-benefits over other approaches, which on a tight budget can lead to less focus on trustworthiness.
- Little flexibility if new attacks or types of vulnerabilities are discovered late in the development process.
- Usability for modeling trustworthiness

In a plan-driven process one can apply structured testing on units as well as on a system as a whole. In addition, it is relatively easy to keep track of the implementation of safety, reliability or security and potentially also trustworthiness requirements. As such, the plan-driven approach supports modeling in general, but not specifically for trustworthiness.

3.2 Incremental

Incremental development (cf. [19]) represents a broad range of related methodologies where initial implementations are presented to the user at regular intervals until the software satisfies the user expectations (or the money runs out). A fundamental principle is that not all requirements can be known completely prior to development. Thus, they are evolving as the software is being developed. Incremental development covers most of the agile approaches and prototype development, although it could be enhanced by other approaches to become more formal in terms of trustworthiness.

Trustworthiness gains:

- New and evolving requirements for trust may be incorporated as part of an iterative process.
- The customer will have a good sense of ownership and understanding of the product after participating in the development process.

Trustworthiness losses:

- Mismatch between organizational procedures/policies and a more informal or agile process.
- Little documentation, increasing complexity and long-lifetime systems may result in security flaws. Especially, documentation on non-functional aspects that are crosscutting among different software features implementation could not be well documented.
- Security and trustworthiness can be difficult to test and evaluate, specifically by the user, and may therefore lose focus on the development.

Incremental development allows new and evolving requirements for trustworthiness to be incorporated as part of an iterative process. Iterative processes allow for modeling of properties, but changes to the model that reflect changed or more detailed customer expectations, will in turn require changing the design and code, eventually in another iteration. Additionally, there are no specific trustworthiness modeling capabilities.

3.3 Reuse-Oriented

Very few systems today are created completely from scratch; in most cases there is some sort of reuse of design or code from other sources within or outside the organization (cf. [19]). Existing code can typically be used as-is, modified as needed or wrapped with an interface. Reuse is of particular relevance for service-oriented systems where services are mixed and matched in order to create larger systems. Reuse-oriented methodologies can be very ad-hoc, and often there are no other means to assure trustworthiness.

Trustworthiness gains:

- The system can be based on existing parts that are known to be trustworthy. This does not, however, mean that the composition is just as trustworthy as the sum of its parts.
- An existing, trustworthy part may increase trust (e.g. a known, trusted authentication).

Trustworthiness losses:

- Use of components that are "not-invented-here" leads to uncertainty.
- Increased complexity due to heterogeneous component assembly.
- The use of existing components in a different context than originally targeted may under certain circumstances (e.g. unmonitored re-use of in-house developed components) jeopardize an existing security / trustworthiness property.

This approach has both pros and cons regarding trustworthiness modeling. On the positive side, already existing, trustworthy and trusted components may lead to easier,

trustworthiness modeling for the overall solution; adequate software assurance, e.g. a security certification, or source code availability may help in improving trustworthiness of re-used “foreign” components. The drawback is that there is a risk that the trustworthiness of the combined system may decrease due to the combination with less trustworthy components.

3.4 Model-Driven

Model-driven engineering (MDE) [22] (encompassing the OMG term Model-driven Architecture (MDA) and others) refers to the process of creating domain models to represent application structure, behavior and requirements within particular domains, and the use of transformations that can analyze certain aspects of these models and then create artifacts such as code and simulators. A lot of the development effort is put into the application design, and the reuse of patterns and best practices is central during the modeling.

Trustworthiness gains:

- Coding practices that are deemed insecure or unreliable can be eliminated through the use of formal reasoning.
- Coding policies related to trustworthiness, reliability and security could be systematically added to the generated code.
- Problems that lead to trustworthiness concerns can, at least theoretically, be detected early during model analysis and simulation.
- Separation of concerns allows trust issues to be independent of platform, and also less complicated models and a better combination of different expertise.

Trustworthiness losses:

- Systems developed with such methods tends to be expensive to maintain, and may therefore suffer from lack of updates.
- Requires significant training and tool support, which might become outdated.
- A structured, model-driven approach does not prevent the forgetting of security and trustworthiness requirements.
- Later changes during development need to review and potentially change the model.
- The (time and space) complexity of the formal verification of especially non-functional properties may lead to omitting certain necessary computations when the project is under time and resource pressure.

With a model-driven approach it is possible to eliminate deemed insecure or unreliable design and coding practices. An early model analysis and simulation with regards to trustworthiness concerns is possible and of high value. In addition, model-driven security tests could improve the trustworthiness. However, in general, there are no specific trustworthiness related modeling properties, it is just model-driven.

The major drawback (and risk) is that the computational complexity for verifying non-functional properties is very high.

3.5 Test-Driven

Test-driven development is considered to be part of agile development practices. In test-driven development, developers first implement test code that is able to test corresponding requirements, and only after that the actual code of a module, a function, a class etc. The main purpose for test-driven development is to increase the test coverage, thereby allowing for a higher quality assurance and thus requirements coverage, specifically related to non-functional aspects. The drawback of test-driven approaches consists in the fact that due to the necessary micro-iterations the design of the software is subject to on-going changes. This makes e.g. the combination of model-driven and test-driven approaches rather impossible.

Trustworthiness gains:

- The high degree of test coverage (that could be up to 100%) assures the implementation of trustworthiness related requirements.

Trustworthiness losses:

- The programming technique cannot be combined with (formal) assurance methodologies, e.g. using model-driven approaches, Common Criteria, or formal verification.

Test-driven development is well suited for assuring the presence of well-described trustworthiness requirements. Moreover, this approach can be successfully used to address changes of the threat landscape. A major drawback, though, is that it cannot easily be combined with modeling techniques that are used for formal assurance methodologies.

3.6 Common Criteria ISO 15408

The Common Criteria (CC) is a standardized approach [24] to evaluate security properties of (information) systems. A “Target of Evaluation” is tested against so-called “Security Targets” that are composed of given Functional Security Requirements and Security Assurance Requirements (both addressing development and operations) and are selected based on a protection requirement evaluation. Furthermore, the evaluation can be performed at different strengths called “Evaluation Assurance Level”.

On the downside, there are some disadvantages: the development model is quite stiff, and does not easily allow for an adjustment to specific environments. Furthermore, Common Criteria is an „all-or-nothing“ approach, one can limit the Target of Evaluation or the Evaluation Assurance Level, but it is rather difficult to then express the overall security / trustworthiness of a system with metrics related to CC.

Trustworthiness gains:

- Evaluations related to security and assurance indicates to what level the target application can be trusted.
- CC evaluations are performed by (trusted) third parties.
- There are security profiles for various types of application domains.

Trustworthiness losses:

- Protection profiles are not tailored for Cloud services.
- A CC certification can be misunderstood to prove the security / trustworthiness of a system, but it actually does only provide evidence for a very specific property of a small portion of the system.

The Common Criteria approach is unrelated to modeling in general, although the higher evaluation assurance levels would benefit from modeling. The functional security requirements may well serve as input for a (security-related) trustworthiness modeling, whereas the security assurance requirements, as the properties of the development process itself, shall be used for a modeling of the developing organization. Note that these constitute two different modeling approaches.

3.7 ISO 21827 Systems Security Engineering - Capability Maturity Model

Systems Security Engineering - Capability Maturity Model (SSE-CMM) is a specific application of the more generic Capability Maturity Model of the Software Engineering Institute at Carnegie Mellon University. Originally, in 1996 SSE-CCM was an initiative of the NSA, but was given over later to the International Systems Security Engineering Association, that published it as ISO 21827 in 2003. In contrast to the previous examples, SSE-CMM targets the developing organization and not the product / service to be developed. There are a number of so-called "base practices" (11 security base practices and 11 project and organizational base practices) that can be fulfilled to different levels of maturity. The maturity levels are identical to CMM.

Trustworthiness gains:

- The developing organization gains more and more experience in developing secure and more generically good quality software.
- The use of a quality-related maturity model infers that user-centric non-functional requirements, such as security and trustworthiness, will be taken into account.

Trustworthiness losses:

- This is an organizational approach rather than a system-centric approach; hence there is not really any guarantee about the trustworthiness of the developed application (which could e.g. be put to use in another way than it was intended for).

This approach focuses on the development of trustworthiness for the developing organization, instead on the to-be developed software, service or system. The security base practices may serve as input for modeling trustworthiness requirements when modeling the development process.

3.8 Building Security in Maturity Model / OpenSAMM

The Building Security In Maturity Model (BSIMM) [23] initiative has recognized the caveat of ISO 21827 being oriented towards the developing organization, and has proposed a maturity model that is centralized around the software to be developed. It defines activities in four groups (Governance, Intelligence, SSDL Touch points, Deployment) that are rated in their maturity according to three levels. OpenSAMM is a very similar approach that has the same origin, but developed slightly differently and is now an OWASP project.

This standard presents an ideal starting point for developing trustworthiness activities within an organization, since it allows tracking the maturity of the development process in terms of addressing security requirements – this could also be used for trustworthiness.

Trustworthiness gains:

- The maturity-oriented approach requires the identification of security (and potentially) trustworthiness properties and assures their existence according to different levels of assurance.
- The probability of producing a secure (and trustworthy) system is high.

Trustworthiness losses:

- There is no evidence that the system actually is trustworthy or secure.

This approach means to develop trustworthiness for the developing organization, instead of the to-be developed software, service, or system. The security base practices may serve as input for modeling trustworthiness requirements when modeling the development process.

3.9 Microsoft SDL

In 2001, Microsoft has started the security-oriented software engineering process that has probably had the largest impact across the whole software industry. Yet, the „process“ was more a collection of individual activities along the software development lifecycle than a real structured approach. The focus point of the Microsoft SDL - that has been adopted by a large number of organizations in different variants - is that every single measure was optimized over time to either have a positive ROI or it was dropped again. This results in a number of industry-proven best practices for enhancing the security of software. Since there is no standardized list of activities, there is no benchmark to map activities against.

Trustworthiness gains:

- The world's largest software manufacturer does use this approach.
- The identified measures have proven to be usable and effective over the course of more than a decade.

Trustworthiness losses:

- There is no evidence that the system actually is trustworthy or even secure.

Microsoft SDL is a development-related threat modeling and was Microsoft's major investment to increase the trustworthiness of its products („Trustworthy Computing Initiative“). The comparability is only given if more detailed parameters are specified. For the modeling of trustworthiness, this method is only of limited help.

3.10 Methodologies Not Covered in This Paper

During the analysis process, a significant number of other methodologies and approaches have been investigated, among others, ISO 27002, OWASP Clasp or TOGAF. We dropped these here since they either replicate some of the capabilities already mentioned above or because their contribution to trustworthiness showed to be rather small.

4 Conclusions from the State of the Art Analysis

After having analyzed the different methodologies and best practices, we can make two major observations. The first observation is related to the nature of the methodologies and best practices. There are two major types of approaches:

- *Evidence-based* approaches that concentrate on evidences, i.e. some sort of qualitative “proof” that a certain level of security, safety etc. is actually met, and
- *Improvement-based* approaches that concentrate on improving the overall situation within the software developing organization with regards to more or less specific requirements.

Evidence-based approaches are typically relatively rigid and therefore often not used in practice, except there is an explicit need, e.g. for a certification in a specific market context. The origin of evidence-based approaches is either research or a strongly regulated market, such as e.g. the defense sector. In contrast to those, improvement-based approaches allow for customization and are therefore much better suited for the application in various industries, but lack in general the possibility to create any kind of evidence that the software developed actually fulfills some even fundamental trustworthiness expectations.

Assuming that evidence-based and improvement-based approaches are – graphically speaking – at the opposite ends of a continuous one-dimensional space, a way to improve trustworthiness of software applications might be to identify approaches that

are “sitting in between” these two types (for example, by picking and choosing elements of different approaches, augmented with some additional capabilities). One option might be to release the burden of qualitative evidence creation by switching to / encompassing evidences based on quantitative aspects. We propose to investigate how metrics for the trustworthiness attributes presented in Section 2 can be used to create evidences by applying selected elements of the improvement-based approaches.

A second major observation relates to the scope of the activities described in the methodologies and best practices. There are three types of “scope”:

- *Product-centric* approaches emphasize the creation and/or verification of attributes of the to-be-developed software,
- *Process-centric* approaches concentrate on process steps that need to be adhered to enable the fulfillment of the expected goal and
- *Organization-centric* approaches focus on the capabilities of the developing organization, looking at a longer-term enablement to sustainably develop trustworthy software.

Some approaches combine the scope, e.g. Common Criteria both mandates verifying product-related and process-related requirements, whereas others, such as SSE-CMM [25] concentrate on only one scope. Current scientific discussions targeting trustworthiness related attributes are mainly focusing on product-centric approaches which is very understandable given the fact that this is the only approach that focuses on evidences on the software itself, whereas practices used in industry often tend towards a more process- or even organization-centric approach (SSE-CMM, CMM, ISO 9001). We therefore propose to investigate how to evolve the above-mentioned evidence-based activities around metrics towards covering process- and organization-centric approaches.

5 Conclusion and Future Work

In this paper we presented an overview on how existing development methods and practices support the development of trustworthy software. To this aim, we first elaborated on the notion of trust and trustworthiness and presented a general taxonomy for trustworthiness attributes of software. Then we analyzed some well-known general software development methodologies and practices with respect on how they support the development of trustworthy software.

As we have shown in the paper, existing software design methodologies have some capacities in ensuring security. But, the treatment of other trustworthiness attributes and requirements in software development is not yet well studied. Trustworthiness attributes that have major impact on acceptance of STS, must be taken to account, analyzed, and documented as thoroughly as possible. In this way the transparency of the decisions under taken during the development will remove potentially the uncertainty of stakeholders of respective software.

The main ideas and findings of our work will be further investigated. It is important to understand how the trustworthiness attributes and the corresponding system properties can be addressed in the system to be in a systematic way. As a next step, we will investigate trustworthiness evaluation techniques for enabling and providing effective measurements and metrics to assess trustworthiness of systems under development. Furthermore, we will develop an Eclipse Process Framework (EPF) based plug-in that will support the process of establishing trustworthiness attributes into a system and guiding the developer through the development activities. Using this plug-in during the development process, the corresponding project team will be supported by guidelines, architectural patterns, and process chunks for developing trustworthy software and later on to analyze the results and evaluate the trustworthiness of the developed software.

Acknowledgements. This research was carried out with the help of the European Commission's 7th framework program, notably the project "OPTET". We specifically would like to thank all participants of Work Package 3 for contributing to the analysis of the methodologies and best practices.

References

1. Gol Mohammadi, N., Paulus, S., Bishr, M., Metzger, A., Koennecke, H., Hartenstein S., Pohl, K.: An Analysis of Software Quality Attributes and Their Contribution to Trustworthiness. In: 3rd International conference on Cloud Computing and Service Science (CLOSER), Special Session on Security Governance and SLAs in Cloud Computing – CloudSecGov, available in SCITEPRESS Digital Library, to appear in Springer-Verlag, SSRI, Aachen (2013)
2. Leveson, N., Stolzy, J.: Safety analysis using Petri nets. *IEEE Transactions on Software Engineering* 13(3), 386–397 (1987)
3. Schumacher, M., Fernandez-Buglioni, E., Hybertson, D., Buschmann, F., Sommerlad, P.: *Security Patterns: Integrating Security and Systems Engineering*. Wiley Series in Software Design. Wiley (2005)
4. Mei, H., Huang G., Xie, T.: Internetware: A software paradigm for internet computing, pp. 26–31. *IEEE Computer Society* (2012)
5. Araújo Neto, A., Vieira, M.: Untrustworthiness: A Trust-Based Security Metric. In: 4th International Conference on Risks and Security of Internet and Systems (CRiSIS), France, pp. 123–126 (2009)
6. San-Martín, S., Camarero, C.: A Cross-National Study on Online Consumer Perceptions, Trust, and Loyalty. *Journal of Organizational Computing and Electronic Commerce* 22, 64–86 (2012)
7. Chen, C., Wang, K., Liao, S., Zhang, Q., Dai, Y.: A Novel Server-based Application Execution Architecture. In: International Conference on Computational Science and Engineering, CSE 2009, vol. 2, pp. 678–683 (2009)
8. Harris, L.C., Goode, M.M.: The four levels of loyalty and the pivotal role of trust: a study of online service dynamics. *Journal of Retailing* 80(2), 139–158 (2004)
9. S-Cube: Quality Reference Model for SBA. S-Cube - European Network of Excellence (2008), http://www.s-cube-network.eu/results/deliverables/wp-jra-1.3/Reference_Model_for_SBA.pdf/view

10. ISO/IEC 9126-1: Software Engineering – Product quality – Part: Quality Model, International Organization of Standardization, Geneva, Switzerland (2001)
11. Gómez, M., Carbó, J., Benac-Earle, C.: An Anticipatory Trust Model for Open Distributed Systems. In: Butz, M.V., Sigaud, O., Pezzulo, G., Baldassarre, G. (eds.) ABIALS 2006. LNCS (LNAI), vol. 4520, pp. 307–324. Springer, Heidelberg (2007)
12. Yolum, P., Singh, M.P.: Engineering self-organizing referral networks for trustworthy service selection. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 35(3), 396–407 (2005)
13. Yan, Z., Goel, G.: An adaptive trust control model for a trustworthy component software platform. In: Xiao, B., Yang, L.T., Ma, J., Muller-Schloer, C., Hua, Y. (eds.) ATC 2007. LNCS, vol. 4610, pp. 226–238. Springer, Heidelberg (2007)
14. Boehm, B.W., Brown, J.R., Lipow, M.: Quantitative Evaluation of Software Quality. In: *Proceedings of the 2nd International Conference on Software Engineering (ICSE)*, pp. 592–605. IEEE Computer Society Press, Los Alamitos (1976)
15. Adrion, W., Branstad, M., Cherniavsky, J.: Validation, Verification, and Testing of Computer Software. *ACM Computing Surveys* 14, 159–192 (1982)
16. McCall, J.A., Richards, P.K., Walters, G.F.: Factors in Software Quality. Volume I. Concepts and Definitions of Software Quality. US Department of Commerce, National Technical Information Service (NTIS), Final technical rept. (1977)
17. Royce, W.W.: Managing the Development of Large Software Systems: Concepts and Techniques. In: *IEEE WESTCON*, Los Angeles CA, pp. 1–9 (1970)
18. Boehm, B.: A Spiral Model of Software Development and Enhancement. *IEEE Computer* 21(5), 61–72 (1988)
19. Sommerville, I.: *Software Engineering*, 9th edn. Pearson, Boston (2011)
20. Jayaswal, B.K., Patton, P.C.: *Design for Trustworthy Software: Tools, Techniques and Methodology for Developing Robust Software*. Prentice Hall (2011)
21. Wordworth, J.: *Software Engineering with B*. Addison Wesley Longman (1996)
22. Schmidt, D.C.: Model-Driven Engineering. *IEEE Computer* 39(2), 25–31 (2006)
23. McGraw, G., Chess, B.: A Software Security Framework: Working Towards a Realistic Maturity Model. *InformIT* (October 2008)
24. ISO/IEC 15408: Information technology - Security techniques - Evaluation criteria for IT security - Part 1: Introduction and general model, Geneva, Switzerland (2009)
25. ISO/IEC 21827:2002: Information technology – Systems Security Engineering – Capability Maturity Model (SSE-CMM) Geneva, Switzerland (2002)

Author Index

- Arndt, Florian 197
Assanovich, Boris 105
- Caprin, Edward 207
- De Decker, Bart 34
De Strycker, Lieven 50
Dittmann, Jana 162, 200, 204
Driessen, Benedikt 18
Dürmuth, Markus 18
Dutta, Ratna 66
- Echizen, Isao 152
Eckert, Marcel 145
- Fabian, Christian 123
Ferecatu, Marin 211
Fruth, Jana 162
- Gohshi, Seiichi 152
Gulyás, Gábor György 173
- Hämmerle-Uhl, Jutta 3
Hasselberg, Andreas 85
Hildebrandt, Mario 200, 204
- Imre, Sándor 173
- Jassim, Sabah A. 217
- Katzenbeisser, Stefan 113
Klauer, Bernd 145
Klößner, Peter 135
Kraetzer, Christian 85
Kümmel, Karl 85
- Lapon, Jorn 34, 50
Li, Yiyao 95
Liu, Huajian 95
- Makrushin, Andrey 200
Mohammadi, Nazila Gol 233
- Naessens, Vincent 34, 50
Nguyen, Hieu Cuong 113
Ntalianis, Klimis 211
- Ortmeier, Frank 197
Ottoy, Geoffrey 50
- Parker-Wood, Aleatha 211
Paulus, Sachar 233
Piller, Ernst 123
Pober, Michael 3
Podebrad, Igor 145
Preneel, Bart 50
Puech, William 105
- Qian, Kun 200
- Raftopoulos, Konstantinos 211
Rao, Y. Sreenivasa 66
Reimers, Nils 135
Rohde, Marleen 162
- Scheidat, Tobias 85
Schulze, Carsten 162
Steinebach, Martin 95, 135
Sturm, Jennifer 204
- Tkachenko, Iuliia 105
Trojahn, Matthias 197
Tsapatsoulis, Nicolas 211
Tzouveli, Paraskevi 211
- Uhl, Andreas 3
- Vielhauer, Claus 85, 204
Vossaert, Jan 34
- Westfeld, Andreas 123
Weyer, Thorsten 233
Wienand, Dominik 135
Wolf, Patrick 135
Wurzer, Jürgen 123
- Xia, Xiaofeng 185
- Yamada, Takayuki 152
- Zhang, Yan 207
Zimmermann, Rene 85