Micro-structure based physical object identification on mobile platforms

DIEPHUIS, Maurits

Abstract

Physical object protection includes all techniques to identify or authenticate objects to determine their origin. In this thesis we envision an approach based on the microscopic surface structure of an object. These so called micro-structures are both unique to the object and currently non-cloneable and thus serve as natural identifiers. Moreover, micro-structure based security schemes have relatively cheap enrollment, and are noninvasive, leaving the original object untouched. This thesis contains a number of methods to allow micro-structure identification using a hand-held mobile phone, without any modification to the object or the device. Primarily, a novel descriptor was designed to robustly identify distorted micro-structures: SketchPrint. It requires no training, is relatively stable yet information rich, and needs but a small number of enrolled descriptor vectors per sample. As it both captures geometrical and micro-structure information from its region of interest, it does not require any exhaustive geometrical re-ranking or aggregation.

Reference


URN : urn:nbn:ch:unige-962499
DOI : 10.13097/archive-ouverte/unige:96249
Micro-structure based physical object identification on mobile platforms

THÈSE

présentée à la Faculté des sciences de l’Université de Genève
pour obtenir le grade de Docteur ‘es sciences, mention informatique

par

Maurits DIEPHUIS

de

Seria (Brunei)

Thèse N° 5091

GENÈVE
Repro-Mail - Université de Genève
2017
Micro-structure based physical object identification on mobile platforms
DOCTORAT Œ SCIENCES, MENTION INFORMATIQUE

Thèse de Monsieur Maurits DIEPHUIS

intitulée :

«Micro-structure Based Physical Object Identification on Mobile Platforms»

La Faculté des sciences, sur le préavis de

Monsieur S. VOLOSHYNOVSKIY, professeur et directeur de thèse
Département d'informatique

Monsieur T. PUN, professeur ordinaire
Département d'informatique

Monsieur T. HOLOTYAK, docteur
Département d'informatique

Monsieur P. BAS, docteur
Laboratoire d'automatique, Génie informatique et Signal, Ecole Centrale de Lille,
Villeneuve d'ascq, France

Monsieur K. EGIAZARIAN, professeur
Institute of Signal Processing, Tampere University of Technology, Finland

autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 12 juin 2017

Thèse - 5091 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".
First and foremost, I would like to thank Prof. Sviatoslav Voloshynovskiy who gave me the opportunity to work and live in Switzerland and whose relentless drive and elegant understanding of how things work, helped me grow. Prof. Thiery Pun, whose natural presence has created a more than pleasant academic environment. Taras Holotyak for his everlasting optimism and technical support.

Thanks to Guillaume Chanel and Mohammad Soleymani, for being my friends and office mates, you have helped me through many a hard moment. Thanks to Edward Akerboom and Uli Beisel, André du Croix, Margo Khokhlova, Ciska Kürpershoek and Annemijn Diephuis, on whose support I have depended endlessly.

Thanks to Fokko Beekhof, without whom I wouldn’t have come to Switzerland, Farzad Farhadzadeh for many a cheerful discussion on math and academic life and Jason Ramapuram for the daily discussions on deep learning.

Thanks to Germaine Gusthiot who ran the faculty, Nicolas Mayencourt and Daniel Agulleiro, who kept my servers running smoothly.

Thanks to my parents, Gerhard en Stieneke Diephuis.

This leaves me with the endless list. The list of people who helped me to get there, or where just always there when needed. In complete random order: Andreas Wombacher, Ferdi van der Heijden, Vroni Wiesgickl, het Schaap, Sander Bockting, Stéphane Marchand-Maillet, Pablo Strasser, Joep Kierkels, Sohrab Ferdowsi, Dince Kostadinov, Stephan Kruisman, Lionel Blondé, Conor Fitzpatrick, Wen-Ru Sheu, Caterina Doglioni, Anne, Martin, Per and Eileen Wingaard Sjøqvist, Paco Roman-Rangel, Ke Sun, Michal Muszynski, Alexandros Kalousis, Séverine Cloix, Wolfgang Domzig, Theodoros Kostoulas, Magda Gregorova, Amina Mollaysa, Géraldine Coppin, Maike Kuno, Vanessa Wormer, Wesley Huisman, Oliver O’Hanlon and Helen Gmuer
Résumé

La protection physique d’un objet inclut toutes les techniques qui peuvent identifier ou bien authentifier des objets afin de tracer et vérifier leurs origines. Les techniques les plus connues ont tendance à altérer les objets d’une manière invasive, en ajoutant par exemple des marques sur des colis ou des hologrammes sur des billets de banques.

Dans cette thèse, nous concevons une approche différente de la protection physique d’un objet, basée sur la structure microscopique de sa surface. Ces micro-structures sont uniques et non-reproductibles et peuvent donc servir comme identifiants naturels. De plus, les systèmes de sécurité basés sur les microstructures sont peu onéreux et non-invasifs, laissant les objets complètement intacts. Enfin, la vérification des résultats peut être effectuée facilement par les utilisateurs eux-mêmes.

Cette thèse accroît de manière importante les connaissances dans ce domaine et ce de plusieurs manières. Un certain nombre de méthodes permettent aux microstructures d’être analysées aisément par des smartphones ou tout autres supports d’acquisition sans aucune modification à apporter au support ou à l’éclairage. Deuxièmement, les algorithmes développés requièrent uniquement que les objets se trouvent dans le champs d’acquisition du support : les algorithmes n’ont pas besoin de marques ou de logiciels spécialisés, aussi bien pour l’acquisition que pour le stockage de ces images de micro-structures.

Le procédé rendant cela possible n’est en revanche pas aisé. Les photographies des microstructures sont visuellement médiocres et manquent de bords et de régions proéminentes. De plus, un certain degré de dégradation est observé sur les photographies prises à l’aide d’un téléphone mobile due, par exemple, à l’inclinaison par rapport à la surface ou un éclairage non-idéal. Enfin, même si le nombre d’objets peut être important, le nombre d’échantillons dont les algorithmes doivent apprendre est lui limité.

Cette thèse apporte des solutions pour la reconnaissance des microstructures, ainsi qu’un cadre pour les effectuer. Dans ce cadre, des paramètres et facteurs de l’environnement

Premièrement, cette thèse démontre comment les microstructures peuvent être associées en se basant sur des particularités dites robustes et la géométrie. Plus précisément, un algorithme basé sur les affinités est capable d’analyser des jeux de données de taille limitées, et qui contiennent plus de 50 de valeurs aberrantes.

Deuxièmement, une technique robuste et innovante de discrimination a été développé, appelée « Sketchprint ». Elle a été élaborée spécifiquement pour l’identification des microstructures déformées. Cette technique ne requiert aucun entraînement, elle est relativement stable et riche en information, et ne requiert que l’enregistrement de quelques vecteurs descriptifs pour chaque échantillon. Comme cette procédure collecte la microstructure et la géométrie de la zone d’intérêt, il n’est pas nécessaire d’utiliser des techniques de reclassement et/ou d’assemblment géométrique.

Finalement, un modèle statistique dit « Bag of Word » a été développé et contient le mode d’identification. Il collectionne toute information pertinente, type et quantité de descriptifs utilisés, la robustesse désirée vis-à-vis du bruit optique, et des choix quant aux assemblages et performances requises. Ce modèle a été vérifié empiriquement.
Abstract

Physical object protection includes all techniques to identify or authenticate objects to determine their origin. High quality counterfeited products are ever more prevalent and widespread nowadays. Luxurious items have always been popular targets, but the last two decades, driven by ever cheaper and sophisticated manufacturing technology, have also seen fake medication, fake industrial and aerospace parts.

Popular countermeasure techniques tend to invasively alter an object, for example by adding markings, holograms, chips or using expensive printing techniques. These techniques hinge on the assumption that they are either to hard or to expensive to replicate.

In this thesis we envision a different approach to physical object protection based on the microscopic surface structure of the object’s surface. This micro-structure is both unique to the object and currently non-cloneable and thus serves as natural identifier. Moreover, micro-structure based security schemes have relatively cheap enrollment, are non-invasive leaving the original object untouched. Lastly, verification can be done by ordinary consumers without any particular expertise.

This thesis will extend the state-of-the-art in several important ways. It shows a number of methods to allow micro-structure based object protection on hand-held mobile platforms, both for enrollment and verification without any kind of modification of the acquisition device or lighting. Secondly, the developed algorithms only require the object to be in the field of view and do not need any aid, such as a priori known printed mark on the package, to acquire and extract the sought micro-structure.

The process that enables this is not trivial. Optically acquired micro-structures are visually poor, lack edges and salient regions. Further serious degradation is caused by the mobile phone, the angle under which it is held, and under what lighting. Lastly, while one can expect the number of enrolled objects to be huge, there are only but a few acquisitions available per unique sample, to learn from.

This thesis proposes a number of solutions for micro-structure verification and frameworks in which they can be applied, explicitly modeling user selectable parameters and
environmental factors that influence the design and performance of micro-structure based authentication and identification frameworks.

Firstly, this thesis demonstrates how micro-structures may be matched based on so called robust features and geometry. Specifically, it proposes an affinity based algorithm that can match small sets of points, of which over 50% are outliers.

Secondly, a novel robust descriptor was developed and is patent pending: Sketchprint. Specifically designed to robustly identify distorted micro-structures, it requires no training, is relatively stable and information rich, and requires but a small number of enrolled descriptor vectors per sample. As it both captures geometrical and micro-structure information from its region of interest, it doesn’t require any exhaustive geometrical re-ranking or aggregation.

Tangent to Sketchprint, to address aggregation and compact descriptor representations, a statistical model of a Bag-of-Word content identification model has been built and theoretically analyzed. It captures all relevant parameters, from the type and quantity of the used descriptors, the desired robustness to noise, architectural choices such as the deployed type of pooling, and ties those into the predicted final system performance. This model has also been empirically verified.
Acronyms

**MSER** Maximally Stable Extremal Region

**FAMOS** Forensic Authentication Micro-structure Optical Set

**BOF** Bag of Features

**BOW** Bag of Words

**PUF** Physical Uncloneable Function

**LUT** Look Up Table

**SP** SketchPrint

**ACF** Auto Correlation Function

**DLT** Direct Linear Transform

**ROC** Radar Operator Characteristic

**EDM** Euclidean Distance Matrix

**DoG** Difference of Gaussian

**ORB** Oriented FAST and Rotated BRIEF

**RP** Random Projection

**RANSAC** RANdom SAmple Consensus

**VQ** Vector Quantization

**CBIR** Content Based Information Retrieval

**PMF** Probability Mass Function

**ANMS** Adaptive Non Maximum Suppression

**ROI** Region of Interest

**DL** Deep Learning

**CNN** Convolutional Neural Network

**SAE** Sparse Auto Encoder

**SC** Sparse Coding

**ICA** Independent Component Analysis
**Symbols**

RICA  Reconstruction Independent Component Analysis  
PCA   Principal Component Analysis  
KLD   Kullback–Leibler Divergence  
MI    Mutual Information  
BSC   Binary Symmetric Channel  
MP    Mega Pixel  
AWGN  Additive White Gaussian Noise  
DOS   Distance Order Statistics  
SVM   Support Vector Machine  
ZCA   Zero Components Analysis  
RFID  Radio-Frequency IDentification
### Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>Scalar random variable</td>
</tr>
<tr>
<td>$\mathbf{X}$</td>
<td>Vector random variable</td>
</tr>
<tr>
<td>$x$</td>
<td>Scalar, realization of a random variable $X$</td>
</tr>
<tr>
<td>$\mathbf{x}$</td>
<td>Vector, realization of a random variable $\mathbf{X}$</td>
</tr>
<tr>
<td>$\mathcal{X}$</td>
<td>Alphabet, or the set of values a realization $x$ of a random variable $X$ may take on.</td>
</tr>
<tr>
<td>$\mathbf{X} \sim p_{\mathbf{X}}(\mathbf{x})$</td>
<td>Random variable $\mathbf{X}$ is distributed according to $p_{\mathbf{X}}(\mathbf{x})$</td>
</tr>
<tr>
<td>$\mathbf{X} \sim p(\mathbf{x})$</td>
<td></td>
</tr>
<tr>
<td>$\mathcal{D}(\cdot</td>
<td></td>
</tr>
<tr>
<td>$I(\cdot, \cdot)$</td>
<td>Mutual Information</td>
</tr>
<tr>
<td>$H(\cdot)$</td>
<td>Entropy</td>
</tr>
<tr>
<td>$\mathbb{E}{\cdot}$</td>
<td>Expectation</td>
</tr>
<tr>
<td>$Q(\cdot)$</td>
<td>Q-function</td>
</tr>
<tr>
<td>$x[i, j]$</td>
<td>Element with index $i, j$ from a 2-dimensional vector $\mathbf{x}$</td>
</tr>
<tr>
<td>$\hat{x}$</td>
<td>Scalar that is being estimated or measured</td>
</tr>
<tr>
<td>$x^N$ or $\mathbf{x}$</td>
<td>Equivalent to ${x[1]...,x[N]}$</td>
</tr>
<tr>
<td>$\mapsto$</td>
<td>Mapping</td>
</tr>
<tr>
<td>$\mathcal{N}(\mu, \sigma^2)$</td>
<td>Gaussian probability density function with mean $\mu$ and variance $\sigma^2_X$</td>
</tr>
<tr>
<td>$\mathcal{B}(L, P_b)$</td>
<td>Binomial distribution with length $L$ and probability $P_b$</td>
</tr>
<tr>
<td>$\cdot$</td>
<td>Multiplication</td>
</tr>
<tr>
<td>$\cdot$</td>
<td>Element wise multiplication</td>
</tr>
<tr>
<td>*</td>
<td>Convolution</td>
</tr>
<tr>
<td>*</td>
<td>Cross correlation</td>
</tr>
</tbody>
</table>
\( F \)  
Fourier Transform

\( \triangleq \)  
Equal to by definition

\( \tilde{x}^k(m) \)  
The projected descriptor \( k \) from image \( m \), its individual elements are \( \tilde{x}^k(m)[i] \)

\( b_{\tilde{x}^k(m)} \)  
The quantized projected descriptor \( k \) from image \( m \)
# Contents

## Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>viii</td>
</tr>
</tbody>
</table>

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>xi</td>
</tr>
</tbody>
</table>

## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>xv</td>
</tr>
</tbody>
</table>

## Chapter 1 Introduction

### 1.1 Object Protection

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.1 Physical alterations</td>
<td>3</td>
</tr>
<tr>
<td>1.1.2 Cryptographic hashing</td>
<td>5</td>
</tr>
<tr>
<td>1.1.3 Watermarking</td>
<td>5</td>
</tr>
<tr>
<td>1.1.4 Physical uncloneable functions</td>
<td>6</td>
</tr>
<tr>
<td>1.1.5 Micro-structures</td>
<td>7</td>
</tr>
<tr>
<td>1.1.6 Identification and Authentication Architectures</td>
<td>8</td>
</tr>
<tr>
<td>1.1.7 Links to Biometrics</td>
<td>8</td>
</tr>
</tbody>
</table>

### 1.2 FAMOS Datasets

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2.1 FAMOS-W</td>
<td>9</td>
</tr>
<tr>
<td>1.2.2 FAMOS-L and FAMOS-S</td>
<td>10</td>
</tr>
</tbody>
</table>

### 1.3 Scope

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

### 1.4 Thesis outline

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
</tr>
</tbody>
</table>

### 1.5 Contributions

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
</tr>
</tbody>
</table>

## Chapter 2 Geometry based micro-structure verification

### 2.1 Introduction

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
</tr>
</tbody>
</table>

### 2.2 Problem formulation

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.1 Requirements and Methodology</td>
<td>22</td>
</tr>
</tbody>
</table>

### 2.3 Micro-structure verification with alignment

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1 Alignment tests</td>
<td>23</td>
</tr>
<tr>
<td>2.3.2 Results</td>
<td>26</td>
</tr>
</tbody>
</table>

### 2.4 Micro-structure verification without alignment

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
</tr>
</tbody>
</table>

### 2.5 Feature geometry based authentication

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
</tr>
</tbody>
</table>

### 2.6 Affinity based Authentication

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6.1 Affinity algorithm</td>
<td>31</td>
</tr>
<tr>
<td>2.6.2 Hybrid identification</td>
<td>32</td>
</tr>
</tbody>
</table>

### 2.7 Feature geometry based authentication with Quantization

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
</tr>
</tbody>
</table>

### 2.8 Results

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
</tr>
</tbody>
</table>

### 2.9 Conclusion

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
</tr>
</tbody>
</table>
3 The SketchPrint descriptor
   3.1 Introduction ........................................ 65
   3.2 Related work and challenges ......................... 66
      3.2.1 Existing identification architectures ...... 67
      3.2.2 Existing hand-crafted descriptors ....... 68
      3.2.3 Machine learning methods ................... 69
      3.2.4 Learned feature representations ........... 69
      3.2.5 Supervised Learning ......................... 76
      3.2.6 Learning key-point reliability ............. 76
   3.3 SketchPrint Objectives .............................. 78
   3.4 The SketchPrint Algorithm .......................... 78
      3.4.1 Detector ....................................... 79
      3.4.2 Descriptor ..................................... 79
   3.5 Design elements ..................................... 80
      3.5.1 Key-point stability ............................ 80
      3.5.2 Stability testing protocol .................... 87
      3.5.3 Results ........................................ 88
      3.5.4 SketchPrint descriptor filtering ........... 88
   3.6 Experimental results ................................ 92
      3.6.1 Descriptor distinguishability: ground truth 92
      3.6.2 Descriptor performance evaluation: geometrically blind matching 94
      3.6.3 Identification performance evaluation: Bag of Words (BOW) .... 96
   3.7 The Distance Order Statistics framework for micro-structure images 102
   3.8 Experimental validation .............................. 104
      3.8.1 Quantization ................................... 105
   3.9 Conclusions .......................................... 113
   3.10 Future work ......................................... 113

4 Bag-of-Word based Content Identification .................. 115
   4.1 Introduction ......................................... 115
   4.1.1 Approach ......................................... 118
   4.2 BOW-based content identification: Problem statement 120
   4.3 Statistical model of BOW content identification .... 122
   4.4 Performance analysis: Uncompressed features ....... 127
   4.5 Performance analysis: Compressed features ......... 131
   4.6 Simulation ............................................ 133
   4.7 Application: Privacy preserving content identification 141
      4.7.1 Existing privacy preserving approaches ....... 142
      4.7.2 Architecture .................................... 143
      4.7.3 Performance, complexity and privacy protection 145

5 Conclusions and Future work ................................ 147

A The FAMOS-W dataset .................................. 151
   A.1 Introduction ...................................... 151
   A.2 FAMOS-W ........................................... 151
A.3 Alignment ................................................................. 152
A.4 ACF-Template based Alignment ................................. 153
  A.4.1 Alignment algorithm ........................................... 153
    A.4.1.1 Design considerations ................................ 153
A.5 Basic patch based identification ................................. 156
  A.5.1 Noise modeling ................................................. 156
    A.5.1.1 Transformed Domain ................................... 158
    A.5.1.2 Experimental Results .................................. 159
A.6 Conclusion ............................................................ 162

B Quantization of SIFT descriptors 165
  B.1 Introduction ....................................................... 165
  B.2 Random Projection and Binarization ......................... 166
  B.3 Performance Analysis .......................................... 167
    B.3.1 Descriptor statistics ...................................... 167
    B.3.2 Channel Distortions ...................................... 168

C The Pharmaceutical package dataset 175
  C.1 Introduction ....................................................... 175

Bibliography 177
# List of Figures

1.1 An example showing the front and back of a genuine and fake Chanel perfume package [1]. Even though they are visually different when presented side-by-side, the counterfeits easily pass as real when a genuine example isn’t present for comparison. ................................................. 2

1.2 An example of a real Canon 580 EX2 flash and a counterfeit one [2]. For all practical purposes, they are indistinguishable to an end user. .......... 3

1.3 Examples of optically acquired micro-structures from different materials such as aluminum (1.3a, 1.3b), paper (1.3c, 1.3d), a proprietary printing process used for a passport (1.3e), rubber (1.3f) and steel (1.3g, 1.3h). The paper sample 1.3e was acquired with a handheld mobile phone, whereas 1.3d was obtained with an industrial camera (Appendix A, Figure A.1). All others were acquired with the Dino hand-held USB micro-scope (Figure 1.4). Histogram equalization was used for visualization purposes. .... 4

1.4 Micro-structure acquisition from paper and plastic with a mobile phone and a hand-held Dino micro-scope. ....................................................... 4

1.5 Overview of the object protection domain. ................................. 5

1.6 The authentication and identification architectures as deployed throughout this work. ................................................................. 10

1.7 Two FAMOS-W acquisitions of a single sample from the RAF camera (1.7a and the NIK (1.7b). ................................................................. 11

1.8 Multiple FAMOS-W acquisitions of a single extracted 128 × 128 micro-structure sample from the RAF (1.8a, 1.8b), and NIK (1.8c, 1.8d) camera. It clearly shows how acquisition circumstances, in particular the light, hugely influence the resulting capture. Histogram equalization and unscharp masking were used for visualization purposes. ......................... 11

1.9 Two acquisitions made by a mobile phone from the FAMOS-L set, with the alignment mark included. ......................................................... 12

1.10 Multiple FAMOS-L acquisitions of a single extracted 600×600 pixel micro-structure from two different physical objects (Figures 1.10a-1.10b and 1.10c-1.10d). Histogram equalization and unscharp masking were used for visualization purposes. ................................. 12

1.11 Multiple acquisitions of two extracted 150 × 100 pixel micro-structure samples (1.11a-1.11b and 1.11c-1.11d) taken with a mobile phone from the FAMOS-S set. It clearly shows how acquisition circumstances, most notably the light and scale, hugely influence the resulting capture. Histogram equalization and unscharp masking were used for visualization purposes. ................................. 12

1.12 Schematic breakdown of all methods in this work and their role in Authentication and Identification architectures. ................................. 14
2.1 Schematic overview of all presented methods in this Chapter for geometry based micro-structure verification, with and without sample alignment and their application to authentication and identification. ............................. 22
2.2 Basic Authentication and Identification setups. ................................. 24
2.3 Schematic overview of the authentication and identification definitions, next to their performance evaluation. ....................................................... 27
2.4 Geometrical alignment of query acquisitions in authentication and identification scenarios. For authentication, a query may be aligned against an identical enrolled copy, whereas for identification alignment can only be done using a common template or mark present on all enrolled samples. 28
2.5 The used masks for geometric alignment when performing identification for FAMOS-W RAF (2.5a) and FAMOS-L (2.5b). ................................. 34
2.6 Results for congealing and funneling based registration using SIFTFlow, for identification (2.6a) and authentication (2.6b) on the FAMOS-W RAF set. ................................................................. 35
2.7 SIFTflow based registration results for identification (2.7a) and authentication (2.7b) using the FAMOS-L set. ....................................................... 35
2.8 Viola based registration results for identification (2.8a) and authentication (2.8b) using the FAMOS-W RAF set. ....................................................... 35
2.9 Viola based registration results for identification (2.9a) and authentication (2.9b) using the FAMOS-L set. ....................................................... 36
2.10 An example of ACF-Template based registration results for identification (2.10a) and authentication (2.10b) using the FAMOS-L set. .................. 36
2.11 Alignment results for FAMOS-L using the ACF-Template method, as shown in the example in Figure 2.10, might seem impressive at first. Therefore, the entire FAMOS-L dataset was aligned with this method and image patches were extracted as identification fingerprint. The ROC curves for this identification system are shown, where the enrolled and query samples were 640 × 800 and 150 × 100. Clearly the performance for both is unacceptable. ....................................................... 37
2.12 Examples of feature based registration for identification (2.12a-2.12b) and authentication (2.12c-2.12d) using the FAMOS-W RAF set. For the identification, the methods matches based on the printed mark only, whereas authentication may match two acquisitions from an identical sample. Only the latter is possible, as results clearly show. For visualization purposes, Figure 2.12d only shows 1’000 matches. ....................................................... 37
2.13 Feature based registration for identification (2.13a-2.13b) and authentication (2.13c-2.13d) using the FAMOS-L set. For the identification, the methods matches based on the printed mark only, whereas authentication may match two acquisitions from an identical sample. Again, only the latter is possible, as results clearly show. ....................................................... 38
2.14 Example of Delauney based local geometry matching [3] using SIFT points from the two largest scales (2.14a, 2.14b) and the first 50 features ordered by scale (2.14c, 2.14d), on two FAMOS-L images from an identical sample encoded against 500 k-means basis vectors. Both image pairs have no matching signatures. ....................................................... 40
2.15 Quad based geometric descriptors. ..................................................... 41
2.16 Schematics of the major components of the feature based identification algorithm. ....................................................... 42
2.17 Illustration of the false matches and outliers between SIFT features from two FAMOS-L acquisitions from an identical sample. Figure 2.17a shows a result when SIFT descriptors are matched using Lowe’s non-ambigu matching rule. Figure 2.17b shows the results from applying rootSIFT [4]. The latter performs better, but the difference is to small to be a sufficient stand alone solution for this particular dataset. 45

2.18 In- and outliers results from the matching of features from an enrolled 100 × 150 patch against a high resolution image from the FAMOS-L set using traditional SIFT features and rootSIFT [4]. The correct sub-patch is marked in blue. Clearly, neither method achieves a satisfactory result. 45

2.19 Illustration of the false matches and outliers between SIFT features from two FAMOS-W RAF acquisitions from an identical sample. Figure 2.19a shows a result when SIFT descriptors are matched using Lowe’s non-ambigu matching rule. Visually it is obvious that most notable in the bottom left corners, a number of inliers has been falsely rejected. Figure 2.19b shows the results from applying DSP-SIFT [5, 6]. For big image samples from an industrial camera, such as these from the FAMOS-W RAF set, domain-size pooling SIFT gives a 9.5% improvement. 46

2.20 In- and outliers results from the matching of features from an enrolled patch against a high resolution image from the FAMOS-L set using traditional SIFT features and DSP-SIFT [5, 6]. The correct sub-patch is marked in blue. Both methods achieve two correct matches, although the SIFT based matching wrongly marks an inlying match as false. 46

2.21 Schematics of the major components of the affinity based identification algorithm. All individual building blocks are seen schematically in Figure 2.22. 47

2.22 Diagram of the affinity based geometrical matching algorithm for feature points. 48

2.23 Example results from the RAF 2592 × 1944 versus RAF 150 × 100 setup for geometry (Figures 2.23a-2.23b) and affinity (Figures 2.23c-2.23d) based matching. Affinity falsely accepts (2.23d) when the number of matches drops to the minimum and all points are outliers. Geometry also exhibits false accepts (2.23d) the moment it may fit 3 out of 4 matches. 49

2.24 Example results from the NIK 1601 × 1201 versus NIK 150 × 100 setup for geometry (Figures 2.24a-2.24c) and affinity (Figures 2.24d-2.24f) based matching. Both algorithms suffer from false accepts if a sufficient number outliers is consistent in their error, effectively making them a consistent inlier set. 50

2.25 Example results from the RAF 2592 × 1944 versus NIK 150 × 100 setup for geometry (Figures 2.25a-2.25c) and affinity (Figures 2.25d-2.25f) based matching. Both the geometry (2.25a) and the affinity algorithm (2.25d) give a true accept, but erroneously include 2 respectively 1 outlier in the final set. False accepts occur the moment erroneous matches are consistent with certain geometrical transform. 51
2.26 Example results from the NIK 1601 × 1201 versus RAF 150 × 100 setup for geometry (Figures 2.26a-2.26c) and affinity (Figures 2.26d-2.26f) based matching. A trade off must be made for geometry based matching on the minimum size of a subset that is coherent in its rigid transformation. Smaller allowable subsets increase the number of false accepts (Figure 2.26b), whereas increasing this number will result in obvious misses (Figure 2.26c). ................................................................. 52

2.27 Example results from the FAMOS S 150 × 100 versus 150 × 100 setup for geometry (Figure 2.27a) and affinity (Figures 2.27b-2.27c) based matching. 53

2.28 Example results from the FAMOS-L 150 × 100 versus 3264 × 2448 setup for geometry (Figure 2.28a) and affinity (Figure 2.28b) based matching. 53

2.29 Two examples showing a feature based authentication attempt between a large query image and a small enrolled micro-structure patch. Figure 2.29a shows the failure of the feature based algorithm (Section 2.5) due to the large number of outliers. Figure 2.29b shows the results for the modified hybrid approach, as detailed in Section 2.6.2. .................. 54

2.30 Estimated intra and inter hamming distance probability mass functions for projected binarized SIFT descriptors from the FAMOS-L set, using the full resolution image of 3264 × 2448 to extract query features (Figures 2.30a-2.30c) or a small 150 × 100 (Figures 2.30d-2.30f) patch where the dimensionality was retained (Figures 2.30a, 2.30d) or reduced to 64 (Figures 2.30b, 2.30e) and 32 (Figures 2.30c-2.30f). These Probability Mass Function (PMF)’s have been estimated using a ground truth, showing that any (geometrical) matching procedure must be able to deal with the fact that there are many erroneous, or inter matches between descriptor that need to be filtered out. In practice, this is not possible .................. 57

2.31 Estimated intra and inter hamming distance probability mass functions for projected binarized SIFT descriptors from the FAMOS-W RAF-RAF and FAMOS-W NIK-NIK set where either the full resolution image was taken to enroll descriptors (Figures 2.31d-2.31f and 2.31j-2.31l) or just the descriptors from a small patch of 150 × 100 (Figures 2.31a-2.31c and 2.31g-2.31i). The dimensionality was retained at the original 128, or reduced to 64 and 32 respectively. These PMF’s have been estimated using a ground truth, showing that any (geometrical) matching procedure must be able to deal with the fact that there are many erroneous, or inter matches between descriptor that need to be filtered out. In practice, this is not possible. .................. 58

2.32 Examples of a true (Figure 2.32a) and false accept (Figure 2.32b results from the FAMOS-W RAF 2592 × 1944 versus RAF 2592 × 1944 geometry based matching where the original descriptors have been quantized to 32 dimensions. Both the true and false accepts have matching descriptors that adhere to a similarity transform. .................. 59

2.33 Examples of true (Figures 2.33a-2.33c) and false accept (Figures 2.33d-2.33f) results from the FAMOS-W RAF 2592 × 1944 versus RAF 150 × 100 geometry based matching where the original descriptors have been quantized to 32, 64 and 128 dimensions. Both the true and false accepts have a small number of matching descriptors that adhere to a similarity transform. .................. 59
2.34 Examples of true (Figures 2.34a, 2.34c) and false (Figures 2.34b, 2.34d) accepts for the FAMOS-L 3264 × 2448 dataset, using the original 128 and reduced 64 quantized feature vectors. ................................................................. 60

2.35 Examples of true (Figure 2.35a) and false (Figures 2.35b) accepts for the FAMOS-L dataset where samples of 150 × 100 where used to enroll descriptors from, quantized to the 128 dimensions. ......................................................... 61

3.1 An example of package authentication and identification using a mobile phone. A package can be recognized based on its printed design, and it can be identified using its micro-structure. ................................................................. 66

3.2 Identification architectures based on (3.2a) direct geometric matching using RANdom SAmple Consensus (RANSAC), (3.2b) BOW with geometric re-ranking and (3.2c) SketchPrint descriptors without geometric re-ranking. ................................................................. 67

3.3 Sparse Autoencoder en RICA basis functions for patches sampled from the aligned FAMOS-W RAF set. These algorithms may discover lower dimensional structure in (image) data, but micro-structures are much more noise-like than natural scene images. ................................................................. 71

3.4 Random Projection based fingerprinting. ................................................................. 72

3.5 Architecture of a single layer convolutional and max pooling network used for feature extraction from aligned FAMOS-W RAF images, using filters obtained from Sparse Auto Encoder (SAE), Reconstruction Independent Component Analysis (RICA) and Random Projection (RP). ................................................................. 72

3.6 Authentication results using a convolutional network feature basis on SAE, RICA and RP basis-functions for the aligned FAMOS-W RAF set, with two different sized max pooling regions of 5 × 5 and 19 × 19. It is notable that there is little difference between results from different sized pooling layers. Lastly, using independently generated RP features gives similar performance to those that are trained. ................................................................. 75

3.7 Authentication results using Random Projections, projecting aligned 128 × 128 FAMOS-W RAF patches directly down to a 64 dimensional feature vector. These results are superior to those of the used convolutional network in Figure 3.6. ................................................................. 75

3.8 Illustration of the extraction of sub image patches around stable and unstable SIFT feature points in the FAMOS-L dataset using an annotated ground truth over a known set of distortions. ................................................................. 77

3.9 Basic block diagram of the SketchPrint algorithm, consisting of two main stages. The detector part spatially clusters found key-points to find a small but robust subset of key-points. The descriptor part traces over the image grid saving fine-grained data, while filtering out signals that are non informative. ................................................................. 81

3.10 Examples of SketchPrint descriptors extracted from the original natural, text and micro-structure images (3.10a), (3.10d), (3.10g) and distorted copies (3.10b), (3.10e), (3.10h) and their comparison (3.10c), (3.10f), (3.10i). 82

3.11 An example showing SketchPrint spatial clustering. Detected points are shown in blue. Found clusters are indicated with a green bounding box, next to single point clusters indicated with green spheres. The latter are removed completely. ................................................................. 83
3.12 The main stages of the SketchPrint descriptor algorithm for the original image and distorted image: (3.12a), (3.12d) original and distorted images with the ORB key-points detection, (3.12b), (3.12e) key-points filtering, (3.12c), (3.12f) SketchPrints extraction and (3.12g), (3.12i) SketchPrints filtering and an example of a match. 84

3.13 The definition of the probabilities of error. In this example, \( y^i_g \) is correctly matched to \( x^i_g \) and \( y^j_g \) to \( x^j_g \), whereas \( x^k_g \) is a false accept and \( x^i_g \) is a miss. 88

3.14 Example of the used error measures, demonstrated on a single image. The ground truth is ascertained by a priori knowledge of the induced distortions and exhaustive geometric matching with RANSAC. The red sphere indicates the radius within which matches are accepted. Matching vertices have blue edges. 89

3.15 An example highlighting that sketching over an edge results in highly similar descriptors, even if the originating images are not a like. Such candidate SketchPrints are rejected using order statistics by stipulating that the local variances must be of the same order, e.g., the variance of the local variance may not be dominated by a single factor. 93

3.16 Comparison of ROCs for SIFT, ORB and SketchPrint descriptors, using a priori geometrical information to exclude the impact of the key-point detectors under: A projective transform, AWGN with noise standard deviation 10 and JPEG compression QF=80 (3.16a) and a projective transform and gamma correction (3.16b). 94

3.17 Schematic of how ordered distances between descriptor sets originating from identical and from different images are collected and graphed. 96

3.18 Examples of the used text documents, packages and random micro-structure images. 97

3.19 100 enrolled SketchPrints (3.19a), (3.19d), (3.19g), 100 enrolled SIFTs (3.19b), (3.19e), (3.19h) and 1'000 enrolled SIFTs (3.19c), (3.19f), (3.19i) performance on text documents (3.19a-3.19c), packages (3.19d-3.19f) and random micro-structure images (3.19g-3.19i) where the query and the enrolled item were identical (self-assessment). 99

3.20 100 enrolled SketchPrints (3.20a), (3.20d), (3.20g), 100 enrolled SIFTs (3.20b), (3.20e), (3.20h) and 1'000 enrolled SIFTs (3.20c), (3.20f), (3.20i) performance on text documents (3.20a)-3.20g), packages (3.20d-3.20f) and random micro-structure images (3.20g-3.20i) on the original and a distorted copy. The distortions were done using a projective transformation, additive Gaussian noise and JPEG compression. 100

3.21 Identification performance for SketchPrint (error-less and therefor lacking a curve), SIFT and ORB on the UCID database using simple BOW aggregation without any (vector) quantization, meaning that all original descriptors are stored and used for the enrollment and queries. 101

3.22 Identification architecture based on sets of individual descriptors and ordered distance statistics. 103

3.23 Order statistics example for two acquisitions of two micro-structures from FAMOS-L using SketchPrint and SIFT: (3.23a-3.23b a sample 1 and (3.23c-3.23d) a sample 2 both acquired 2 times, (3.23e) the DOS for 100 enrolled SketchPrint, (3.23f) the DOS for 1’000 SIFT and (3.23g) 10’000 SIFT. The latter can not distinguish these micro-structures due to the overlap of inter and intra statistics. 106
3.24 Boxplot for the first 10 distance order statistics for the FAMOS-L set with the intra distances in red and inter in blue: (3.24a) SketchPrint, (3.24b) 1'000 enrolled SIFT, (3.24c) 10'000 enrolled SIFT and (3.24d) 500 enrolled ORB based identification. ........................................... 107
3.25 Boxplot for the first 10 distance order statistics for SketchPrint and the FAMOS-W set with the intra distances in red and inter in blue for: (3.25a) RAF camera, (3.25b) OMRON camera. ................................. 107
3.26 Boxplot for the first 10 distance order statistics for ORB and the FAMOS-W RAF dataset with the intra distances in red and inter in blue for: (3.26a) 250 and (3.26b) 500 enrolled descriptors. ......................... 108
3.27 Boxplot for the first 10 distance order statistics for ORB and the FAMOS-W NIK dataset with the intra distances in red and inter in blue for: (3.27a) 250 and (3.27b) 500 enrolled descriptors. ............................... 108
3.28 Boxplot for the first 10 distance order statistics for SIFT and the FAMOS-W RAF dataset with the intra distances in red and inter in blue for: (3.28a) 1'000 and (3.28b) 10'000 enrolled descriptors. ......................... 109
3.29 Boxplot for the first 10 distance order statistics for SIFT and the FAMOS-W NIK dataset with the intra distances in red and inter in blue for: (3.29a) 1’000 and (3.29b) 10’000 enrolled descriptors. ......................... 109
3.30 Boxplot for the first 10 distance order statistics for SketchPrint (3.30a), SIFT (3.30b) and ORB (3.30c) for FAMOS-W dataset, matching the RAF camera against the NIK, with the intra distances in red and inter in blue. 110
3.31 Boxplot for the first 10 distance order statistics for the FAMOS-L set with the intra distances in red and inter in blue after projection and quantization for (3.31a) SketchPrint, (3.31b) 1’000 enrolled SIFT. ............ 112

4.1 Block diagram of a BOW identification system. ........................................ 116
4.2 Basic building blocks of a Bag of Content based identification system. ....... 121
4.3 Encoding and pooling for BOW based identification ................................. 123
4.4 Fundamental building blocks of a BOW based identification system. ........... 126
4.5 Descriptor matching performance ......................................................... 137
4.6 ORB similarity score distributions ....................................................... 138
4.7 Impact of the number of deployed descriptors in BOW based identification 139
4.8 Optimal BOW thresholds ................................................................. 140
4.9 The diagram of privacy preserving image search in a cloud system. ............ 142
4.10 Secret key sharing using geometrical information. .................................. 144
4.11 Complete schematic of the proposed architecture for private image identifi-
cation in cloud systems. .............................................................. 144
4.12 Identification performance with obfuscation of the query descriptors for dif-
f erent sizes of $J_x$ and $J_y$ and a single $synchro$ case where geometry information is used. ......................................................... 146

A.1 The industrial acquisition device, including the feeder, belt, camera, light-
ing and integrated computer screen which were used for the FAMOS-W acquisi-
tion. ..................................................................... 152
A.2 The basic autocorrelation based alignment algorithm, as used for the FAMOS-W dataset. ................................................................. 154
A.3 The FAMOS-W template used for alignment and its corresponding Au-
tocorrelation function. ......................................................... 154
A.4 Illustration of Auto Correlation Function (ACF) peak sensitivity to fluctuations in the segmented mark. [7] ................................................. 155
A.5 Radar Operator Characteristic (ROC) curve for all camera combinations from the FAMOS-W set using $\rho_{xy}$ in the real domain. .......................... 157
A.6 Authentication performance on synthetic data. ................................. 160
A.7 Spectra of the noise showing noise correlation of the FAMOS-W dataset. . 161
A.8 Authentication results on the FAMOS-W dataset in the direct domain. . 161
A.9 Authentication results on the FAMOS-W dataset in the DCT domain. . 162
A.10 Examples of synthetic micro-structures [8]. From left to right, the 2D Ising model (A.10a), The Moving Gaussian Average (A.10b), Perlin ($1/f$) noise (A.10c) and Brownian Fractal noise (A.10d). .......................... 163

B.1 Binarized SIFT generating framework. ............................................. 166
B.2 The average correlation per image with in descriptors before and after projection and binarization for the Caltech Airplane set, FAMOS-W and FAMOS-L. The average correlation between original SIFT descriptors per image is shown via the normalized Auto Correlation Function in Figures B.2a-B.2c. The results after projection and binarization can be seen in B.2d-B.2f. ................................................................. 169
B.3 Normalized histogram of the pair-wise entropy $H(B_x[i], B_x[j]), i \neq j$, for random projected and quantized SIFT descriptors from the Caltech Airplane set (Figure B.3a), FAMOS-W (Figure B.3b) and FAMOS-L (Figure B.3c). ................................................................. 169
B.4 Empirical distance distributions, and a fitted Gaussian distribution, between descriptor vectors in the original domain (Figures B.4a-B.4c) and projected quantized domain (Figures B.4d-B.4f) for the Caltech Airplane set, FAMOS-W and FAMOS-L. ................................................................. 170
B.5 The probability of bit error, $P_b$ for for the tested distortions distortions for a set of projected SIFT descriptors. ............................................. 172
B.5 Channel statistics for various distortions for a set of original and projected SIFT descriptors. $x$ and $y$ denote the codebook of original and the distorted sft descriptors. The projected counterparts are denoted by: $b_x$ and $b_y$. Descriptors from identical indices are matching descriptors. . . . . 174
C.1 Examples of acquisitions from the Pharmaceutical package dataset. ....... 176
Chapter 1

Introduction

We consider a scenario in which customs or an end user wishes to authenticate a shipment of medicine. Counterfeit drugs, after all, have been a regular phenomena the last decade [9]. In the absence of any other evidence pointing to a possible counterfeit shipment, actually authenticating a (medical) package is not a trivial task. Furthermore, the amount and type of physical products that are counterfeited is large and varied, including documents, consumer products but also counterfeit chips [10], aviation parts or military hardware [11].

Traditional anti-counterfeiting methods usually rely on modifying the product with an industrial process that is either very hard to replicate, or so costly that counterfeiting becomes economically nonviable. Examples include sophisticated printing patterns, holograms, Radio-Frequency IDentification (RFID), special inks etc. However, ever cheaper and sophisticated manufacturing technology have enabled counterfeiters to replicate and falsify a much broader range of physical products then was previously possible.

There have been some attempts, by brand owners at educating their consumers, explicitly highlighting what details to look for in genuine and counterfeit products [1]. Examples can be seen in Figure 1.1. These differences, although clear when the genuine and fake product are presented side by side, are tough to spot in practice and place the burden on the end consumer. Worse, there are many examples of counterfeits that are near indistinguishable. Figure 1.2 from [2] shows two Canon flashes. One is fake. Close investigation reveals that the fake flash has a deviating glue pattern below a rubber protection band. This type of high quality counterfeit is ever more prevalent and one can simply not expect a consumer to spot the difference.

We consider a consumer, who generally lacks both the time and knowledge to scrutiny a product. It is therefore that we envision another approach, one in which a customs
Chapter 1. Introduction

Figure 1.1 – An example showing the front and back of a genuine and fake Chanel perfume package [1]. Even though they are visually different when presented side-by-side, the counterfeits easily pass as real when a genuine example isn’t present for comparison.

officer or end consumer takes an ordinary mobile phone, and makes a (macro) picture of the products surface structure, or micro-structure. This picture is send to a server, which compares it with the enrolled micro-structure pictures of authentic products, taken at fabrication. A match indicates that the presented product is indeed authentic.

The essence of this scenario is the so-called identification and authentication problem. We aim to identify a photographed sample, potentially corrupted by noise, compression and geometric (acquisition) distortions, against a growing database of original samples, in an efficient way.

The key driver is the actual optically acquired micro-structures, which serves as Physical Uncloneable Function (PUF), as they are both unique and currently non-cloneable [12]. Relatively cheap enrollment, the non-invasive nature of the protection and the ability to be easily verified efficiently by consumers make micro-structure based protection schemes very attractive.

Key components of these architectures share many similarities with biometric, watermarking and digital fingerprinting schemes. They include the selection of robust or invariant features, the ability to recover from major geometrical and lighting distortions, dimensionality reduction and quantization and aggregation techniques.
Chapter 1. Introduction

1.1 Object Protection

Object protection compasses all methods that aim to identify or authenticate objects such that their origin may be verified and tracked.

State-of-the-art methods for object protection may broadly be divided into protection methods for physical and digital objects, methods that are invasive, i.e., they add something to the object, or non-invasive such as fingerprinting or using forensics, and lastly, methods that rely on randomness or structured information. An overview can be seen in Figure 1.5, and we will briefly review them, next to highlighting a number of differences with biometrics based security.
Figure 1.3 – Examples of optically acquired micro-structures from different materials such as aluminum (1.3a, 1.3b), paper (1.3c, 1.3d), a proprietary printing process used for a passport (1.3e), rubber (1.3f) and steel (1.3g, 1.3h). The paper sample 1.3e was acquired with a handheld mobile phone, whereas 1.3d was obtained with an industrial camera (Appendix A, Figure A.1). All others were acquired with the Dino hand-held USB micro-scope (Figure 1.4). Histogram equalization was used for visualization purposes.

Figure 1.4 – Micro-structure acquisition from paper and plastic with a mobile phone and a hand-held Dino micro-scope.
1.1.1 Physical alterations

The most well known active techniques, for the protection of physical objects, invasively mark and alter an object, such as the usage of special inks, printing, and holograms on bank notes. Alterations may also be done using synthetic randomness. These are random alterations caused by some type physical interaction, such as a laser burn, which may also serve as PUF (Section 1.1.4).

This type of protection hinges on the assumption that the used manufacturing is either impossible or to costly to replicate. Even if this assumption would be universally valid, there are a number of downsides. Products need to be altered, not all techniques are capable of uniquely marking each product making them unsuitable for identification, it also potentially makes the product manufacturer dependent on a third party. Lastly, the end users needs to know the details of the used protection in order to validate its authenticity itself.

1.1.2 Cryptographic hashing

Modern information security depends critically on asymmetric, or one-way functions. These are operations that are very easy to do in one direction, but extremely hard to reverse. The most famous example of this is the multiplication of two very large prime
numbers. Multiplication is straightforward, but currently no algorithm is known that can find the factors in polynomial time, and it is suspected to be in \textit{NP}. For our application, the protection of physical objects such as consumer goods, the main drawback is more practical in nature. Using cryptographic primitives requires embedding using a specialized device such as a smart card or authentication token. This economically unfeasible for ordinary goods [13, 14].

1.1.3 Watermarking

Other traditional methods for digital object protection include a wide range of watermarking methods [15] in which information is added to either be resilient against malicious object modifications, carry limited information or function as an indicator for any kind of tampering. Watermark can be designed such that stakeholders can satisfy conflicting requirements such as the robustness against attacks, the visibility and the payload. Physical objects that may be printed can all be potentially watermarked, but generating a unique (key dependent) watermark per object is not trivial for mass production when cost-efficient off-set printing is used.

Fingerprinting passively extracts a (binary) signature from an object that may serve as a unique indicator. A wide variety of algorithms exists with cryptographic hashes on one side of the spectrum and robust features which sacrifice discriminative power for resilience on the other.

Active content fingerprinting [16] is a more recent development and is a hybrid form between fingerprinting and watermarking. It minimally alters an object to ensure that an extracted fingerprint is of superior quality in terms of robustness and information payload for a given vector length.

1.1.4 Physical uncloneable functions

Physical uncloneable functions are random like structures and phenomena that are difficult to statistically model, currently impossible to clone, but relatively easy to measure, store and verify.

Natural or inherently present randomness occurs in the (surface) structure of physical objects. Examples include micro-structures and human biometrics.

Natural randomness may also be induced by some kind of fabrication process with uncontrollable parameters or naturally occurring. Examples of the first include random interaction patterns of ink and paper, burn patterns from laser engravings or printing
press imperfections [17]. Examples of optically acquired natural micro-structures are seen in Figure 1.3.

Synthetic randomness includes all deterministic methods that actively modify physical objects with a mark [18]. The marks have an information payload and are hard to clone as their physical embedding process gives rise to unique random interactions between the ink and the object. Examples include graphical codes [19, 20] or the result from a laser burn.

1.1.5 Micro-structures

Micro-structures are physical uncloneable functions (PUF), and the central focus of this work. They are measurements taken from the inherently random surface structure of physical objects. The usage of micro-structures for object identification and authentication has been researched using different modalities to acquire the measurements.

So called speckle patterns [21, 13, 14], which occur when a light source is reflected of a surface and the scattered result is projected on a screen, have been used successfully for identifying paper, both using a laser [12] and LED lighting [22]. Usage of coherent light, in different wave-bands is also prevalent [23, 24, 25].

This work will focus on optical acquired micro-structures, using an ordinary hand-held consumer mobile phone without any modification or lighting aid (incoherent light) of any kind, photographing objects that have no marks or other indicators to where the specific Region of Interest (ROI) is residing on the object, other than it is somewhere within the field-of-view.

Key element for all micro-structure based systems is that the random structures serve as object identifiers that are both unique and currently non-cloneable [12]. They can relatively easily be passively enrolled without any object modifications, and form the basis for a variety of authentication and identification architectures.

Secondly, the processing chain for optical micro-structures shares similar components and technological challenges with traditional biometric, watermarking and fingerprinting schemes. These include the selection of robust or invariant features, the ability to recover from major geometrical and lighting distortions, dimensionality reduction and quantization and aggregation techniques.
1.1.6 Identification and Authentication Architectures

Object identification and authentication architectures are both based on the premises one can ascribe a unique identifier to said object. In this work, this is the optically acquired micro-structure or any fingerprint that was subsequently extracted from this image.

The enrollment phase is similar for both architectures, but the verification stage for identification and authentication differ in a simple, but fundamental way. Identification returns an estimated identification (number) based on the received data, which is matched against the enrollment database, or issues a reject. Contrarily, authentication takes both the object’s data and the claimed identifier and decides if the two correspond based on the received data, and the stored enrolled data that pertains to the passed query identifier. An overview can be seen in Figure 1.6.

Generic requirements for any object verification system include accuracy, speed, needed memory, storage, resilience, scalability and flexibility [26]. Specifically in this work, using micro-structures from mediocre mobile phone acquisitions, there are three important factors in play:

- Due to the very nature of the system, the protection of (consumer) objects and physical goods, the number of stored items, and the number of items that gets enrolled during the lifetime of the system is potentially huge.
- The original micro-structures images need to be stored, to make future systems backwards compatible with any modifications to the used processing techniques, e.g., the fingerprinting or robust descriptor algorithms.
- Deployed algorithms must not be data-dependent. Due to the constant enrollment of new items, which may or may not exhibit similar statistical properties, algorithms that critically depend on processing the entire dataset are of limited use. This includes well known methods such as Principal Component Analysis (PCA) and the SAE.

1.1.7 Links to Biometrics

Similar to the usage of PUFs for object security, biometrics compasses all methods and features used to uniquely identify a human. They have a certain robustness against errors, whilst being as hard as possible to clone or falsify. The architectures for enrollment and verification are similar for both fields, sharing tools and building blocks, such as an acquisition phase, the extraction of features and a decision making process governing both identification and authentication.
There are, however, a number of significant differences, which we will briefly touch. For an in depth review, see [27].

Contrarily to biometrics, protected physical objects, such as consumer goods, are part of the public domain and thus in full disposal to an attacker. For PUF based physical object protection, this is not a problem. The PUFs generally exhibit high entropy, reproducing a clone from original PUF features or the actual physical sample is generally not technically feasible or economically non-viable. In contrast, biometrics, such as fingerprints, exhibit low entropy and may be reproduced with relative ease. As such, the field has developed a multitude of methods to protect and obfuscate stored biometric templates and augmented verification frameworks, to limit the information an attacker has, if a database with enrolled biometric primitives is compromised [28, 29, 30]. This also means that for verification in the field, the biometrics’ hard-ware and protocol requirements are more stringent to guard against information leaks.

PUF based physical object security may be enhanced more easily as objects can potentially be modified by a manufacturing process that leaves a random mark, such as a burn from a laser. Acquisitions can also relatively easy be done with more than one modality, e.g., using different wavelengths. Obviously, humans may not be modified to enhance extracted biometrics. Augmentations in biometric based security is usually based on executing more than a single verification round, using more primitives or for example, testing for liveliness to see if the presented sample is actually a human and not a fabricated physical copy.

Because of the near noise-like nature of, for example, optically acquired PUF’s such as the micro-structures used in this work, equipment for enrollment must ideally be of (very) high quality. Biometrics, such as fingerprints, with inherent salient detail and structure, are much easier to acquire.

Lastly, PUF’s for object security should ideally be enrolled at the source, or manufacturing before release into the public domain. Human biometrics may be enrolled at any time in the life-cycle. The latter also suffers less from degradation due to aging than physical objects due to wear and tear from everyday usage.

1.2 FAMOS Datasets

The Forensic Authentication Micro-structure Optical Set (FAMOS) datasets contain optically acquired micro-structures from standard consumer cardboard packages. There are 3 different sets. FAMOS-W contains images acquired with industrial cameras in a
controlled setting. In contrast, FAMOS-L and FAMOS-S images were acquired inside an ordinary office with a hand-held mobile phone.

1.2.1 FAMOS-W

FAMOS-W [31, 32], is a freely available forensic dataset, produced by the University of Geneva, comprised of micro-structures taken from 5000 unique packages with two different industrial cameras, designated as RAF and NIK. It contains three acquisitions per sample from two cameras resulting in 30000 images in total. A printed mark was used to ensure that the specific designated region with the micro-structure was captured. The cameras were fixed above the targets and used ring lighting to ensure high quality acquisitions. A brief overview of the FAMOS-W dataset and the results is presented in Appendix A.

1.2.2 FAMOS-L and FAMOS-S

FAMOS-L and FAMOS-S are two datasets that were acquired with a consumer mobile phone, as demonstrated in Figures 1.9 and 1.11. The deployed mobile device is a Samsung Galaxy S3 phone, which has an on board camera producing $3264 \times 2448$ (8MP) images from the Sony manufactured IMX145 sensor.

No special light or any other modifications were used. Further more, the samples have been acquired in different rooms to verify the impact of variations in the light conditions between the enrollment and identification phases.
Figure 1.7 – Two FAMOS-W acquisitions of a single sample from the RAF camera (1.7a and the NIK (1.7b).

Figure 1.8 – Multiple FAMOS-W acquisitions of a single extracted 128 × 128 micro-structure sample from the RAF (1.8a, 1.8b), and NIK (1.8c, 1.8d) camera. It clearly shows how acquisition circumstances, in particular the light, hugely influence the resulting capture. Histogram equalization and unscharp masking were used for visualization purposes.

FAMOS-L consists of 312 samples of 1024 × 768 pixels, acquired twice. The samples contain a printed mark, which potentially may aid the extraction of the predefined micro-structure image patch. Sample acquisitions can be seen in Figure 1.9.

The second mobile set, FAMOS-S consists of 19719 unique samples, that were acquired twice giving 39438 images in total. Samples are 150 × 100 pixels and do not contain any printed mark. Examples can be seen in Figure 1.11.

1.3 Scope

The objective of this thesis is the identification of micro-structure images taken with an ordinary hand-held mobile phone without any a priori knowledge of the sample or its (geometrical) distortions.
Figure 1.9 – Two acquisitions made by a mobile phone from the FAMOS-L set, with the alignment mark included.

Figure 1.10 – Multiple FAMOS-L acquisitions of a single extracted $600 \times 600$ pixel micro-structure from two different physical objects (Figures 1.10a-1.10b and 1.10c-1.10d). Histogram equalization and unscharp masking were used for visualization purposes.

Figure 1.11 – Multiple acquisitions of two extracted $150 \times 100$ pixel micro-structure samples (1.11a-1.11b and 1.11c-1.11d) taken with a mobile phone from the FAMOS-S set. It clearly shows how acquisition circumstances, most notably the light and scale, hugely influence the resulting capture. Histogram equalization and unscharp masking were used for visualization purposes.
As the geometrical distortions are a natural consequence of the acquisition method, and micro-structures lack the intrinsic structure to compensate this, all methods and processing chains that have limited or no resilience against these deviations are out of scope. This includes most fingerprinting methods, such as used in iris recognition [33], dimensionality reduction methods and basis pursuit algorithms [34].

This thesis uses two primary micro-structure datasets. The first one, FAMOS-W, is comprised of micro-structures, acquired with industrial cameras with ring lighting fixed above the conveyor with physical objects. It demonstrated the maximum attainable performance when acquisition parameter are controllable.

The second primary dataset, FAMOS-L, is completely comprised of acquisitions taken with an unmodified hand-held consumer mobile phone. Only ambient indoor light was used. These images suffer from a multitude of distortions. There are serious geometic distortions as the phone is hand-held, non-linear lens distortions, the lighting isn’t uniform for a sample, or similar between acquisitions.

Both datasets contain a limited number of acquisitions, 2 or 3 per camera, per unique object to distill information from. Secondly, micro-structures are visually poor lacking the salient detail that is common in natural scene images.

To overcome the lack of training data, the near noise like nature of the images and the geometrical and lighting distortions caused by the acquisitions, two feature based algorithms have been developed.

The first allows to geometrically match a query sample against those in a dataset using a small set of features riddled with outliers.

The second concerns the development of a new robust descriptor, Sketchprint, designed specifically to be stable and informative, yet requiring a factor 10 to 20 less enrolled vectors than more common methods.

Although current state-of-the-art deep learning methods for image recognition surpass those based on the older BOW paradigm, they require a massive number of samples per class or per label to train on. Furthermore, classes should be separable in some feature space. The mobile images from FAMOS-L are visually poor, yet orders of magnitudes larger (10 Mega Pixel (MP)), and with but 3 to 6 acquisitions per sample, have no meaningful training data. This means that BOW architectures to aggregate robust features into a single fixed dimensional representation are a viable solution.

To research the trade-off between the different requirements, parameters, and BOW architectural design choices, a comprehensive statistical model was developed that is
Figure 1.12 – Schematic breakdown of all methods in this work and their role in Authentication and Identification architectures.

able to predict the final performance. This model has been empirically validated and illustrated with a use-case of designing a privacy-preserving content search framework.

1.4 Thesis outline

This thesis will both review and introduce new methods to identify objects using optically acquired micro-structures. Table 1.1 shows a brief overview of the used micro-structure databases and the primary developed and tested methods. Figure 1.12 shows a similar overview of all used methods, demonstrating their usage in either identification or authentication architectures.
Chapter 1. Introduction

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Result summarization</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment and fingerprinting</td>
<td>FAMOS W</td>
<td>Yes</td>
<td>Appendix A</td>
</tr>
<tr>
<td></td>
<td>FAMOS L</td>
<td>Not satisfactory</td>
<td>Chapter 2</td>
</tr>
<tr>
<td></td>
<td>FAMOS S</td>
<td>Not applicable as mark is absent</td>
<td>Chapter 2</td>
</tr>
<tr>
<td>Exhaustive geometric matching</td>
<td>FAMOS W</td>
<td>Yes</td>
<td>Appendix A</td>
</tr>
<tr>
<td>based on robust features</td>
<td>FAMOS L</td>
<td>Yes</td>
<td>Chapter 2, Appendix B</td>
</tr>
<tr>
<td></td>
<td>FAMOS S</td>
<td>No</td>
<td>Chapter 2</td>
</tr>
<tr>
<td>Affinity based matching</td>
<td>FAMOS W</td>
<td>Yes, but not necessary</td>
<td>Appendix A</td>
</tr>
<tr>
<td></td>
<td>FAMOS L</td>
<td>Yes, but not necessary</td>
<td>Chapter 2</td>
</tr>
<tr>
<td></td>
<td>FAMOS S</td>
<td>Yes</td>
<td>Chapter 2</td>
</tr>
<tr>
<td>Robust descriptors and Look</td>
<td>FAMOS W</td>
<td>Yes, but computational infeasible</td>
<td>Chapter 3</td>
</tr>
<tr>
<td>Up Table (LUT)</td>
<td>FAMOS L</td>
<td>Yes, but computational infeasible</td>
<td>Chapter 3, Appendix B</td>
</tr>
<tr>
<td></td>
<td>FAMOS S</td>
<td>No, due to lack of robust descriptors</td>
<td>Chapter 3</td>
</tr>
<tr>
<td>Bag of Features (BOF) based approaches</td>
<td>FAMOS W</td>
<td>Yes</td>
<td>Chapter 4</td>
</tr>
<tr>
<td></td>
<td>FAMOS L</td>
<td>Yes</td>
<td>Chapter 4</td>
</tr>
<tr>
<td></td>
<td>FAMOS S</td>
<td>No</td>
<td>Chapter 4</td>
</tr>
<tr>
<td>SketchPrint (SP)</td>
<td>FAMOS W</td>
<td>Yes</td>
<td>Chapter 3</td>
</tr>
<tr>
<td></td>
<td>FAMOS L</td>
<td>Yes</td>
<td>Chapter 3, Appendix B</td>
</tr>
<tr>
<td></td>
<td>FAMOS S</td>
<td>Yes</td>
<td>Chapter 3</td>
</tr>
</tbody>
</table>

Table 1.1 – Brief overview of the micro-structure datasets, identification and authentication methods and results.

Geometry based micro-structure identification

Chapter 2 formally introduces the authentication and identification framework, and the used micro-structure datasets, FAMOS-W, FAMOS-L and FAMOS-S. The last two datasets are micro-structure sets that have been acquired with a hand-held unmodified mobile phone.

It demonstrates that it is possible to authenticated FAMOS-W, FAMOS-L and FAMOS-S samples without the use of a printed mark or other a priori information using robust features. In particular, it introduces affinity based matching, designed for FAMOS-L and FAMOS-S, to be able to geometrically match robust features from a very small set that contains over 50% outliers.

The SketchPrint descriptor

The main contribution of Chapter 3 is the introduction of a new robust discriminative descriptor: SketchPrint. It is designed specifically to robustly identify micro-structure images which contain very little salient detail, be extremely stable against non-linear geometrical distortions, information rich and thus discriminative, and, to require at least an order of magnitude less descriptors to be enrolled per image than current state-of-the-art. The latter will also contain an overview of the so-called Deep Learning (DL) methods.
and demonstrate their limitations on datasets such as FAMOS-L, thus necessitating the design of the new handcrafted feature, Sketchprint.

In this Chapter, we will also consider each design criteria in depth and compare every component against state-of-art descriptors. Most significantly, we will cover the criteria currently used to a priori measure if a key-point will be stable against distortions, next to introducing a new measure based on spatial clustering. Robustness and discriminative power is tested, not only on micro-structures, but on natural images, sparse product pictures and text, both on an individual basis and as part of a BOW architecture. Lastly, it introduces the Ordered Statistics Framework, which allows to model and test system performance for a varying number of enrolled descriptors and quantization parameter.

**Bag-of-Word based Content Identification**

A formal statistical model for BOW based content identification is introduced in Chapter 4. This model is amongst the first that theoretically predict the identification performance taking all relevant parameters into account. These range from the type and quantity of the underlying descriptors, the quantization and size of the codebook, the desired performance, to architectural choices such as the used type of pooling. An empirical simulation on real data is used to validate the theoretical results. Lastly, Chapter 4 shows how a practical application, a privacy preserving search, can be build using the results of the theoretical model.

**Micro-structure alignment**

Appendix A briefly introduces the FAMOS-W dataset, that contains a printed mark which was used to align samples as to undo any geometrical distortion and extract the sought micro-structure patch. It covers the used alignment algorithm, the basic matching results based on cross-correlation and a more advanced correlated noise model. It serves as a performance indicator for what performance is feasible when more information is available.

**Quantization of SIFT descriptors**

Appendix B details the quantization of SIFT descriptors based on random projections. Ideally, for any binarized feature vector one would want bits to be equally likely, bits within feature vectors to be independent. For any identification or authentication architecture that relies on binary features as identifier, one would like feature vectors originating from different objects to not only be independent but also maximally different to
be able to deal with noise. Appendix B covers all these requirements, empirically testing when needed on both real life image data as micro-structure data from FAMOS-L.

1.5 Contributions

The main contributions of this Thesis are:

- The design of an authentication scheme for non aligned micro-structure images based on binarized robust features.
- The development of an authentication algorithm to geometrically match a small subset features originating from a query and database image, when the number of feature outliers exceeds 50 % percent.
- The development of a new feature detector algorithm that increases key-point robustness by spatially clustering their geometric locations, such that the found centroids form new feature points.
- A novel (universal) feature descriptor was designed, Sketchprint, that specifically also works on very smooth edge-less (micro-structure) images that have little to no salient regions. It simultaneously captures local geometrical information next to a detailed descriptor vector of its region of interest.
- The formulation and definition of a statistical model for BOW based feature identification.
- The design of an architecture that allows for privacy preserving semantical image search, in which the privacy and identification capabilities are explicitly defined and modeled.
Chapter 2

Geometry based micro-structure verification

In this chapter, we address physical object identification on mobile phones. It is well known that consumer goods are counterfeited on a massive scale in certain regions of the world, illustrating how existing counter measures fall short or don’t exist at all, as can be seen in the local absence of laws pertaining to brand protection. This chapter introduces a technological tool that allows the consumer to quickly identify a product or package with a mobile device using physical non-cloneable features in the form of a surface micro-structure image. This natural occurring identifier allows a producer to track and trace all its products and gives the consumer a powerful instrument to confirm the authenticity of an offered product. Work in this chapter is based on the following publications: [35, 36, 37, 31].

2.1 Introduction

Physical uncloneable features (PUF’s) such as optically acquired micro-structures (Figures 1.3 and 1.4) are an attractive emerging tool in the field of physical object protection [21, 13, 14]. Currently, micro-structures can not be counterfeited and as such can serve as a cheap non-evasive security and authentication identifier for packages, passports, or other physical goods next to being a natural occurring serial number for tracking and tracing applications.

Micro-structures and their process-chain share many aspects with the biometrics field. Examples include the robust extraction of the relevant structure, the selection and deployment of robust features, dimensionality reduction and quantization, and naturally,
a need for fast and near error-less identification and authentication on an as large as possible database thus approaching the information-theoretic limits of identification [38]. Naturally, identification capacity, memory-complexity and robustness are conflicting requirements which should be addressed in the design of any identification system.

The success of physical object identification critically depends on a proper selection of:

- The informative features extracted from images.
- Some sort of (alignment) mechanism to ensure that the designated image patch is extracted.
- A set of decision rules matched with the statistics of the features and distortions.

In previous studies, we introduced FAMOS-W [31, 32], a freely available forensic dataset comprised of micro-structures taken from 5000 unique packages with two different cameras and three acquisition rounds resulting in 30000 images in total. A printed mark was used to ensure that the specific designated region with the micro-structure was captured. Its basic statistical properties were examined and base identification performance for original, dimensionality reduced and quantized data was established. A brief overview of the FAMOS-W dataset, basic deployed algorithms and results are presented in Appendix A.

This chapter extends the current state-of-art two fold. Firstly, by demonstrating that FAMOS-W micro-structure samples can be identified from raw query images from the entire package without alignment based on the printed mark, by deploying robust features and a heuristic matching procedure [36]. The latter approach thus greatly enhances robustness by being partly invariant to geometrical distortions in sacrifice of earlier memory-efficient and fast fingerprinting methods.

Secondly, we will depart from the industrial and purposefully build hand-held cameras in favor of a basic consumer mobile phone camera without any special lighting or adaptation. This dataset is designated FAMOS-L (Section 1.2.2) and its hand-held acquisition can be seen in Figures 1.9 next to the resulting cropped out micro-structure patches in Figures 1.11 and 1.10.

Mobile hand-held acquired images differ severely from their industrial counterparts:

- They are acquired under varying light conditions, pending where the user took the picture.
- The resolution differs between mobile devices.
- The images have non-linear geographical distortions, both from the lens and the manner in which the mobile device was held.
• There is limited control over the digital processing pipeline of the device, such as compression or white-balance correction.

Micro-structure verification can be performed in two distinct modes, identification and authentication. The (technical) requirements differ significantly between those two applications, and we will consider both in depth in this chapter.

Identification

Micro-structure identification, by definition, necessitates the ability to quickly compare a query against a database with enrolled (labeled) samples. Although it is not impossible to use the raw image data, speed and storage requirements dictate need for the extraction of a low dimensional fingerprint. For micro-structure images, geometrical distortions caused by the optical acquisition next to determining the exact patch extraction coordinates are the most critical step. This broadly leaves three options:

• The framework processes images prior to fingerprint extraction with an alignment or synchronization algorithm using a printed mark as a template. Depending on the specific mark, this may be done using robust features [7]. For the FAMOS-W dataset (Section A.3) alignment was achieved using the ACF based algorithm, which is covered in detail in Appendix A.

• If alignment is not possible, nor the extraction of a completely geometrical invariant feature descriptor, identification may still be achieved if exhaustively matching a query against the enrolled set, can be done sufficiently fast.

• The used fingerprinting algorithm must be completely robust against geometrical distortions and extract all image data that is within the field of view. This approach is taken in Chapter 3, which introduces the SketchPrint descriptor.

Authentication

Micro-structure authentication is technically less challenging than identification as it allows to compare a labeled query image against its claimed labeled enrolled sample, thus providing the decision making algorithm with much more information. Next to all algorithms used for identification, the following approaches are also potentially viable:

• Alignment may still be done, but not against a predetermined universal mark, but between the presented query image and the claimed corresponding enrolled image. A multitude of methods for this alignment scenario where images of an identical subject or object need to be matched exists, including robust feature
Figure 2.1 – Schematic overview of all presented methods in this Chapter for geometry based micro-structure verification, with and without sample alignment and their application to authentication and identification.

• Forgoing alignment due to the lack of a printed mark, or a mark that does carry sufficient information, a match may also be evaluated based on the number and quality of matching robust features in either image. This approach is covered in Sections 2.4, 2.6 and 2.7.

2.1.1 Organization

An overview of the used approaches, methods, and applicability to either package authentication of identification within the scope of this Chapter can be seen in Figure 2.1.

The first mayor research question is whether or not micro-structure samples can be **aligned**, i.e., can geometrically distortions from both enrollment and verification acquisition be corrected. Geometrical distortions, caused for example by a hand-held acquisition device are by far the largest and most significant source of errors. These distortions not only influence the micro-structure data itself, but also make it more difficult to extract micro-structure data from an exact pre-defined region. Due to their random nature, small ROI shifts will lead to large signal deviations.
Alignment methods are covered in Section A.3, and test four principal state-of-art methods: congealing, convex methods, ACF-Template matching using the ACF, and geometric invariant features. All methods are tested in two modalities, namely for the authentication and identification setup. Specifically when aligning query image samples, pending on targeting either two, the amount of information an alignment algorithm may leverage, varies. Sections 2.1 and 2.1 will cover this in detail.

Section 2.4 covers the scenarios when an alignment phase is foregone or simply unfeasible. State-of-the-art methods either based on forms of robust hashing or geometric matching are reviewed and tested empirically. They include geometrically matching robust features (Section 2.5), Delauney triangulation, spectral graph matching and Quadrilateral (Quad) features. The Robust feature matching algorithm is extended in two ways. The Affinity algorithm covered in Section 2.6 was designed to be able to handle a sparse set of features with a significant amount of outliers. Section 2.7 shows how robust feature matching can be made more computationally efficient by deploying quantization and compression, next to the incurred authentication performance penalty.

Finally, Section 2.6.2 demonstrates how performance may be enhanced by firstly roughly attempting to align a sample from FAMOS-L, prior to deploying methods designed for non-aligned samples.

Summarizing, this chapter extends the state-of-the-art primarily by negating the use of a printed mark for alignment and patch extraction whilst using a hand-held mobile phone as an acquisition device. To enable this, three principle authentication methods of our own were developed and tested:

- An image-based similarity score using design-based alignment followed by traditional fingerprinting.
- Two variants of feature based geometric matching.
- A hybrid method that deploys rough alignment and accept or rejects features based on geometric consistency according to a predefined family of models.

## 2.2 Problem formulation

### Identification

In the identification problem, we assume that the enrolled database contains images acquired from $M$ objects. We suppose that each image of size $N_1 \times N_2$ is lexicographically ordered into a sequence $x(m) \in \mathbb{R}^N$, where $N = N_1 \times N_2$ with $1 \leq m \leq M$. The object
that is to be tested is represented by its own vector $y$ which might originate from the observation of some $x(m)$ contained in the database through the degradation channel $p(y|x(m))$ or any randomly generated vector $x'$ which is not linked with data stored in the database.

Identification can be considered as a composite $(M + 1)$ hypothesis testing problem [39], [38]:

\[
\begin{cases}
H_0 : p(y|H_0) = p(y|x'), \\
H_m : p(y|H_m) = p(y|x(m)), 1 \leq m \leq M,
\end{cases}
\]  

(2.1)

where hypothesis $H_0$ corresponds to the case when some object $x'$, unrelated to the database with enrolled samples, is presented to the system, and hypothesis $H_m$ denotes a valid case where an object under consideration corresponds to an enrolled data item $x(m), 1 \leq m \leq M$.

To validate the system’s performance, the following measures can be used. The \textit{probability of successful attack} $P_{sa}$, which denotes the case when an unrelated (counterfeited) object is accepted as one of the enrolled items. Secondly, the \textit{probability of incorrect identification} $P_{ic}$ denotes the situation in which an enrolled object with index $m$ is wrongly decoded, as an object with some index $m' \neq m$.

For each enrolled object, the identification produces an estimation of an index $m$ by applying a binary test $\phi : Y^N \times X^N \mapsto \{0,1\}$, where 0 stands for rejection and 1 for acceptance for each object.

In particular the acceptance or rejection decision of a query is performed using a distance metric and a threshold $\gamma$ as $\phi(y,x(m)) \leq \gamma N$, i.e., if $\phi(y,x(m)) \leq \gamma N$, the decision is 1 and 0, otherwise. Such a system may produce a list of indices and it is implicitly
assumed that properly selecting threshold $\gamma$ ensures that only the correct index $m$ will be in the final list.

In this case, the probability $P_{sa}$ can be defined as:

$$P_{sa} = \Pr[\bigcup_{m=1}^{M} \phi(Y, x(m)) \leq \gamma N | \mathcal{H}_0].$$  \hspace{1cm} (2.2)

The probability $P_{ic}$ consists of the probabilities of two possible events under hypothesis $\mathcal{H}_m$, i.e., the probability that the object with index $m$ is not found in the database, consisting of $M$ objects, denoted as probability of miss $P_{m}$ and the probability that the object with index $m$ is falsely accepted or identified with index $m'$, denoted as the probability of false acceptance $P_{f}$. These two probabilities are defined as:

$$P_{m} = \Pr[\phi(Y, x(m)) \geq \gamma N | \mathcal{H}_m],$$  \hspace{1cm} (2.3)

$$P_{f} = \Pr[\bigcup_{m \neq m'} \phi(Y, x(m)) \leq \gamma N | \mathcal{H}_m],$$  \hspace{1cm} (2.4)

which provides the upper bound to $P_{ic} \leq P_{m} + P_{f}$ according to the union bound [38].

Given an optimal design of $\phi$ matched with the observation model $p(y|x(m))$ and selection of $N$, one can demonstrate that the $P_{f}$ can be made negligibly small, independent of the number of items $M$ in the database. Contrarily, $P_{sa}$ and $P_{m}$ depend on the identification capacity $C_{id} = I(X; Y)$, where $I(\cdot; \cdot)$ denotes the mutual information [40]. Both may be made negligibly small, if the number of database items satisfies $\frac{1}{N} \log_2 M \leq C_{id}$. Moreover, these probabilities are equal, if all items in the database are statistically independent and an attacker submits queries that are generated independently from enrolled data [38]. Work in [41] investigates both terms following a scenario in which attackers have prior statistical knowledge on the enrolled database items.

The principal challenge in designing an identification system based on micro-structure images is overcoming the lack of accurate statistical models $p(y|x(m))$ and $p(y|x')$, which determine system performance under hypothesis $\mathcal{H}_0$ and $\mathcal{H}_m$.

If and only if explicit geometric alignment is assumed between items, and that the residual distortions are additive and Gaussian in nature [31, 32], the hypothesis from (2.1) may be reformulated as:

$$
\begin{align*}
\mathcal{H}_m' & : y = x' + z, \\
\mathcal{H}_m & : y = x(m) + z.
\end{align*}
$$  \hspace{1cm} (2.5)
In this case, the similarity measure $\phi$ is reduced to the Euclidean distance. The lower performance bounds for such systems, using the high quality samples from FAMOS-W and near perfect alignment, have been tested and are summarized in Appendix A.

**Authentication**

In the second part of this Chapter, we will deal primarily with micro-structure datasets for which geometric alignment can not be attained (accurately). Therefore, in following, we will evaluate $P_m$ and $P_f = P_{sa}$ for an authentication system based on semi-robust local invariant features for which a geometrical relation can be established between a query and an enrolled sample.

Authentication may be considered as a binary hypothesis testing problem:

\[
\begin{align*}
H_0 & : p(y|H_0) = p(y|x(m')), m' \neq m \\
H_m & : p(y|H_m) = p(y|x(m)),
\end{align*}
\]  

(2.6)

for which the corresponding $P_m$ and $P_f$ are:

\[
P_m = \Pr[Y \text{ is rejected } | H_m],
\]  

(2.7)

\[
P_f = \Pr[\bigcup_{m \neq m'} Y \text{ is accepted as } m'|H_m].
\]  

(2.8)

**Performance evaluation**

Summarizing, in this work, both authentication and identification will be evaluated in terms of their probability of miss $P_m$ and probability of false alarm $P_f$. A schematic overview is seen in Figure 2.3.

**2.2.1 Requirements and Methodology**

For micro-structure verification, there are two major algorithmic approaches and two application design goals. The approaches are whether or not the algorithm will attempt alignment to the samples to undo any geometrical distortions prior to processing, or use the images as is. The second major distinction is whether the developed verification framework will use identification or authentication. A complete overview of all methods is shown in Figure 2.1.
Alignment algorithms need information to be able to ascertain the incurred geometrical distortions. In an authentication scenario, this can be done by comparing a presented query sample against an enrolled copy as the query also passes a sample identifier. In an identification scenario, only a query sample is presented to the user. This means that alignment needs to be done on the basis of a priori knowledge for all samples, such as the presence of a common mark or logo. A schematic overview of the alignment procedure for both authentication and identification is shown in Figure 2.4.

When an alignment phase is skipped, or is simply technically not feasible, the algorithms must work with the images in their original acquired form. Especially for images taken with a hand-held mobile phone, these distortions can be significant. This means that whatever method is used, it must exhibit sufficient invariance against these geometrical distortions. In an authentication scenario, one has a possibility of matching invariant features from a query and an enrolled copy on a one by one basis. Identification, however, requires the comparison of a query against the entire enrolled dataset. This scenario ideally requires the usage of geometrically invariant hashes or a fast matching procedure.
2.3 Micro-structure verification with alignment

This section will review and empirically test four micro-structure alignment methods, both in an identification and authentication scenario. It contains three well established image registration methods, next to introducing an algorithm of our own. These are:

- A funneling and congealing method.
- An information-theoretic approach combined with a convex optimizer as is commonly seen in medical image registration.
- ACF-Template based matching.
- Feature based registration, applied widely in stitching and camera calibration.

Specifically, these methods will be tested on micro-structure images in both an identification scenario, where only a query image is present, and an authentication scenario, which would allow the registration of two images against each other as the query also contains a label or identifier. Schematically, this is shown in Figure 2.4.
Funneling and Congealing methods

This family of methods traces their origin to image recognition, where images exhibit huge amounts of intra-class variability. Facial images, for example, will differ in lighting, background and pose. Considerable effort has been put in reducing this intra class variability as doing so enhances the performance of recognition and identification systems. Congealing is the process of iteratively transforming a stack of images such that they become more similar, for example by extracting and clustering SIFT descriptors as intermediate features and using entropy as a fitness measure [42]. Work from [43, 44] uses deep learning to extract intermediate features.

SIFTflow [45] uses a variant of the optical flow objective function, next to densely sampled SIFT descriptors as basic input instead of raw pixel values. Optical flow [46] is usually understood as the distribution of the apparent velocities of selected objects in different image frames, for example from a video. It has shown excellent results for both scene alignment, stitching semantically similar yet visually very different images together next to being able to stitch relative sparse images, such as those taken from the Marsian surface.

Convex registration methods

Image registration is widely used and researched in the medical domain. It involves fusing and aligning images from possibly different modalities as a pre-cursor to many surgical visualization, planning and navigation steps [47, 48].

It may use features that are invasively introduced or be based on the images alone. The later, known as intrinsic methods my use hand placed markers to aid the registration process or derive information from naturally present image structures such as texture or edges by automatically detecting robust features or simply use all raw image data as is.

Registration methods all broadly share the following components [49]:

- A deformation model, that defines the expected geometrical distortions between images, be it rigid, affine or any other deformation. Common approaches also include describing the transform as a linear combination of basis functions, such from the Fourier or Wavelet transform. Others includes using splines and non-parametric models [50].

The number of free parameters that need be optimized obviously influence the computational burden and the nature of the solution space tremendously.
• The used feature. In the most simple case, the feature might be the raw image data, or matching landmark points that are selected by a human operator in the different images. Algorithms may also use features derived from the image, such as a derivative, or use feature detectors such as SIFT.

• The similarity measure or objective function. The used criteria that needs to be maximized when seeking the best registration. In its most simple form, although insufficient, this may be the L2 distance between raw images or between all landmark points.

• The optimization technique. The (numerical) optimization methods that search for global extrema of the given objective function by manipulating the set of supplied model parameters.

Information-theoretic approaches were pioneered by Viola [51] and Maes [52] who both advocated the use of Mutual Information (MI) between two images to determine the quality of the registration. An overview can be found in [53]. As a base comparison this work will test the Viola algorithm combined with simplex optimization [54].

The primary drawback of this family of registration methods is two-fold. Firstly, transforming the image and determining the similarity measure for every algorithm iteration is very computational expensive. Secondly, there is no guarantee that the obtained extrema in parameter-space is global. In other words, the found mapping need not correspond to the sought solution for a particular user application.

**ACF-Template based matching**

ACF-Template based matching was successfully deployed on the FAMOS-W dataset [32, 31]. At the heart of this method is the fact that it uses an a priori known mark on the physical object to undo geometrical distortions and extract a pre-designated region with a micro-structure. If successful, this method enables the application of dimensionality reduction, fingerprint and quantization schemes, which in turn allows for computationally fast and storage efficient identification architecture.

In short, this algorithm takes the two dimensional ACF of the segmented printed mark and detects its local maxima. The geometrical coordinates of these points $y_g$ are matched against those of an other image or template $x_g$: \{ $y_g \leftrightarrow x_g$ \} resulting in a set of linear equations \{ $y_g = Ax_g$ \} from which scaling and rotation may be derived. After correcting these, translation requires a final pass using traditional cross correlation between the segmented marks. In particular, the usage of the ACF, makes ACF-Template based
matching robust against small printing fluctuations and any noise caused by the segmentation step. It is also parameter free. All the Algorithm details and test are shown in Appendix A.

**Feature based registration**

Computer vision features such as SIFT [55] or SURF [56] have long been deployed in applications such as the stitching of pictures to form a panorama, estimate pose, calibrate a camera, or for structure from motion algorithms [57, 58]. After extraction, these features need to be matched, which is generally done in two stages.

Contemporary feature descriptors $x = (x_g, x_d)$ have two distinct components, an $n$ dimensional location in (scale) space $x_g$ and a descriptor vector that encodes a ROI around the feature location $x_d$. In the first stage, an initial matching $\{y_d \leftrightarrow x_d\}$ is attained by exhaustively matching the descriptor vectors $y_d$ and $x_d$ against each other.

Secondly all algorithms assume a distortion model between the features from the images. If these distortions are modeled as affine, the relationship between matching feature pairs $\{y_g \leftrightarrow x_g\}$ can be expressed as $y_g = Ax_g$. This model not only defines the transformation that can potentially be undone to align images, but also allows the stitching algorithms to filter out outliers from the initial set of matching feature points, for example, using RANSAC.

The basic feature based registration algorithm that will be tested on FAMOS-W and FAMOS-L was earlier used successfully on other datasets [7] and is covered in depth in Section 2.5.

### 2.3.1 Alignment tests

All above mentioned alignment methods were tested on both FAMOS-W and FAMOS-L for two distinct scenario’s, identification and authentication (Figure 2.4).

For identification the algorithm must be able to align a presented sample based on that sample alone, which means that the only a priori information that can be leveraged is the printed mark.

In the authentication framework, the system is presented with a query image and its supposed label. The alignment algorithm may therefore use the original enrolled item with that particular label. This means that next to the printed mark, the micro-structures can potentially be used to help alignment.


2.3.2 Results

Four alignment algorithms were tested, on FAMOS-W and FAMOS-L in both the identification and authentication scenario.

Funneling and Congealing methods: SIFTflow

The result for the SIFTflow algorithm applied to FAMOS-W and FAMOS-L are shown in Figure 2.6 and 2.7.

The SIFTflow algorithm is capable of leveraging micro-structure surface information for its registration. Although not flawless, it achieves reasonable alignment results for authentication for both datasets (Figures 2.6b and 2.7b). However, this method still leaves more post alignment residual errors than feature based registration.

Identification results for FAMOS-W and FAMOS-L (Figures 2.6a and 2.7a) are better than those of the Viola Maes based algorithm, thought hardly sufficient for our application.

Convex registration methods: Viola Maes

Results for algorithm based on the work of [59, 52] which use simplex in combination with mutual information can be seen in Figure 2.8 and 2.9 for FAMOS-W and FAMOS-L.

Regardless of the identification or authentication setup, results for FAMOS-W (Figures 2.8a-2.8b) show the algorithm terminating in a non optimal maximum. Results for FAMOS-L (Figures 2.9a-2.9b) clearly show residual scaling errors.

In our specific application context, the fast verification of physical objects using micro-structure images, iterative registration methods are already disadvantageous due to their computational complexity. The fact that they can not register images that are visually sparse, makes them non applicable as a whole.

ACF-Template based matching

ACF-Template based alignment is by far the simplest, in terms of parameters, and fastest method of those tested. For FAMOS-W, its alignment is precise enough to enable further dimensionality reduction and fingerprinting applications for fast identification. These results are covered in detail in Appendix A.
A visualization of an identification and authentication result for FAMOS-L can be seen in Figure 2.10. Although some residual error is left, especially the identification (Figure 2.10a) result is superior to all other alignment methods.

This method was therefore tested against the entire FAMOS-L dataset. The key requirement for this method is the quality of the mark and the type of occurring distortions. Mobile acquisitions, of a printed mark, of which dataset FAMOS-L is made up, suffer in that respect. Secondly, the distortions both from the lens and the fact that the mobile device is hand-held are non-linear. This information can not be sufficiently deduced from the deformations in the printed marks of FAMOS-L and is the source of residual geometric errors after alignment.

Although significantly more precise than other alignment methods, this unsurprisingly means that ACF-Template based matching gives unacceptable results for an identification system on the FAMOS-L dataset, as is shown in the ROC curve in Figure 2.11.

**Feature based registration**

Figure 2.12 shows an example of feature based registration results for FAMOS-W for identification and authentication (Figure 2.12c). The identification result (Figures 2.13a-2.13b) clearly shows some residual scaling between the query image and the template, meaning that for the extraction of a micro-structure patch, feature based registration is insufficient. Authentication (Figures 2.12c-2.12d) is clearly viable for FAMOS-W based on features, if the entire image is enrolled and stored in the database.

The results for FAMOS-L are shown in Figure 2.13. Clearly the four crosses that form the FAMOS-L mark are not sufficient to register an image in the identification context (Figures 2.13a-2.13b). The micro-structure is however, sufficiently rich to generate a multitude relatively stable features, such that images takes from an identical physical sample, may be registered against each other (Figures 2.13c-2.13d).

**Conclusion**

In the identification scenario, where queries can only be aligned against an a priori known mark, the best results are given by ACF-Template based matching, and only for the FAMOS-W dataset (Appendix A). This set has a high quality printed mark that is sufficiently information rich. The mark is in the center of the acquisition, the sample is laying flat on a conveyor and is photographed with an industrial camera and light.
Chapter 2. *Micro-structures*

Table 2.1 – Result summary for the four basic image registration methods on the FAMOS-W RAF and FAMOS-L datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Application</th>
<th>Features</th>
<th>ACF-Template</th>
<th>Convex</th>
<th>Congealing</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAMOS-W</td>
<td>Identification</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>Authentication</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>FAMOS-L</td>
<td>Identification</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>Authentication</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

Contrarily, the mark in the FAMOS-L mobile dataset is printed by an ordinary laser printer, which gives considerably more variation throughout the set. Furthermore, the mark is on the edges of the camera view, where the (mobile) lens is most prone to distortions and flaws. Finally, these samples were acquired hand-held, producing an ever wider rage of distortions that need to be corrected. No algorithm was able to do this in an identification scenario.

In an authentication scenario, where a query image may be aligned against an identical enrolled copy, deploying features and robust geometric matching gives superior alignment results for authentication, on both FAMOS-W and L datasets. There are two mayor drawbacks to this type of approach. Firstly, geometric matching is computational intensive, and secondly, it requires storing the original images of the dataset, next to the extracted features. Both issues will be address in Section 2.4.
Figure 2.6 – Results for congealing and funneling based registration using SIFT-Flow, for identification (2.6a) and authentication (2.6b) on the FAMOS-W RAF set.

Figure 2.7 – SIFTflow based registration results for identification (2.7a) and authentication (2.7b) using the FAMOS-L set.

Figure 2.8 – Viola based registration results for identification (2.8a) and authentication (2.8b) using the FAMOS-W RAF set.
Figure 2.9 – Viola based registration results for identification (2.9a) and authentication (2.9b) using the FAMOS-L set.

Figure 2.10 – An example of ACF-Template based registration results for identification (2.10a) and authentication (2.10b) using the FAMOS-L set.
Figure 2.11 – Alignment results for FAMOS-L using the ACF-Template method, as shown in the example in Figure 2.10, might seem impressive at first. Therefore, the entire FAMOS-L dataset was aligned with this method and image patches were extracted as identification fingerprint. The ROC curves for this identification system are shown, where the enrolled and query samples were 640 × 800 and 150 × 100. Clearly the performance for both is unacceptable.

Figure 2.12 – Examples of feature based registration for identification (2.12a-2.12b) and authentication (2.12c-2.12d) using the FAMOS-W RAF set. For the identification, the methods matches based on the printed mark only, whereas authentication may match two acquisitions from an identical sample. Only the latter is possible, as results clearly show. For visualization purposes, Figure 2.12d only shows 1’000 matches.
Figure 2.13 – Feature based registration for identification (2.13a-2.13b) and authentication (2.13c-2.13d) using the FAMOS-L set. For the identification, the methods matches based on the printed mark only, whereas authentication may match two acquisitions from an identical sample. Again, only the latter is possible, as results clearly show.
Chapter 2. Micro-structures

2.4 Micro-structure verification without alignment

This section focuses on micro-structure verification without prior alignment for both authentication and identification. Broadly speaking, this means that either one must be able to extract a feature vector or fingerprint that is completely invariant to geometrical distortions caused by the acquisition, or should be able to form a match between a query and an enrolled sample directly.

This section will review a number of state-of-the-art methods for geometric invariant hashing and their applicability. Section 2.4 will introduce a robust but exhaustive matching procedure. This algorithm will be extended for robustness in Section 2.6. Finally Section 2.7 will aim to speed up the matching using quantization.

Robust features can also form the basis for a geometrically invariant hash by grouping them following some criteria and encoding the result. In this Section, we will review three popular approaches, namely one based on Delauney triangulation, Quad descriptors and Spectral graph matching, followed by two authentication algorithms of our own, in Sections 2.5 and 2.6.

Delauney Triangulation

Delauney triangulation is commonly used to select the features (vertices of the found triangles) [60, 3, 61] which will be jointly encoded or to select the image patches (surface covered by the triangles) for further processing [62]. Delauney triangulation is sensitive to the appearance and disappearance of features, as all triangles this vertex belonged to, will change. Work in [3] tries to overcome this by using a multi-scale feature (SIFT) and building triangulations for each scale-space level. Per triangle, the vertices are encoded by taking the SIFT descriptor belonging to that particular vertex and matching it against a codebook of $k$-means basis functions. The indices of the three closest ones are written out, giving a 9 integer code, or signature, per triangle.

When applied to a micro-structure image from FAMOS-L and a copy that is synthetically rotated by 15 degrees, this methods results in about 10% matching signatures. However, examples of this algorithm applied to actual multiple FAMOS-L acquisitions, as seen in Figure 2.14, result in no matches at all. The used number of $k$-means basis, between 50 and 5'000, nor the scale-selection method have any influence on this result.

There are a number of causes for this failure. Features found in micro-structure images are inherently weak as the underlying image is quite devoid of surface structure or salient regions. The two direct consequences of this are that subtle changes in acquisition angle
Chapter 2. Micro-structures

Figure 2.14 – Example of Delauney based local geometry matching [3] using SIFT points from the two largest scales (2.14a, 2.14b) and the first 50 features ordered by scale (2.14c, 2.14d), on two FAMOS-L images from an identical sample encoded against 500 k-means basis vectors. Both image pairs have no matching signatures.

and light make a significant number of these features appear and disappear, and that descriptor vectors will not be discriminative.

Lastly, the originating SIFT scale-space, nor a fixed or quantified one, is not a suitable metric for robustness, a point that will be extensively covered in Chapter 3.

Spectral Graph matching

The graph isomorphism problem is in NP (with subgraph isomorphism belonging to NP-C [63]), that appears in many sub domains of computer science such as pose estimation, shape analysis, wide baseline stereo image matching and (medical) image registration. Spectral methods operate by taking the spectra, or eigenvalues, of the graph Laplacian matrices [64, 65]. All methods need a (learned) model or criteria to be able to formulate an objective function that allows to iteratively improve a matching [66, 67, 68].

Micro-structures are semi-random structures and individual samples don’t adhere to any (known) visual model, such as a shape outline. There is no natural coherence between
features from a single micro-structure image, which means that in order to build up graph edges, a rule must be applied, such as simply connecting neighboring vertices or connecting features whose originating descriptors quantize to the same (k-means) basis-vector. Connecting spatial nearest neighbors nodes is similar to the first step of basic Delauney triangulation (Section 2.4) and does not produce stable results. Neither does forming a graph by making an edge between nodes when the SIFT descriptors for these nodes match an identical trained basis-vector.

Although a simple deformation model, such as an affine transformation, can definitely be assumed between matching features from two acquisitions taken from an identical physical sample, we have found that this family of methods offers no advantage over more simple traditional computer vision methods, as detailed in Section 2.5 and 2.6.

Quad descriptors

Quad descriptors have their origin in matching star images [69, 70] against known constellations. They use two feature points as anchors to form a new scale and rotation normalized axis system in which all other nearby points are encoded and written out in pairs of two, with the anchors, to form 4-dimensional feature vectors, as shown in Figure 2.15. The main challenge, in absence of a reliable method to select anchor points in the FAMOS-L set, is the combinatorial explosion of possible 4D feature vectors. Secondly, the geometrically renormalized feature constellations, as they emerge from the FAMOS-L set, contain an overwhelming number of identical configurations that are not distinctive for a unique source image.

Figure 2.15 – Quad based geometric descriptors.
2.5 Feature geometry based authentication

It is possible to build a geometric feature based decision heuristic that can be used for identification while negating the need for explicit alignment or the presence of a printed mark. However, detecting features in micro-structures using methods designed for computer vision tasks and natural images, is not completely straightforward.

Micro-structure images show a glaring absence of prominent edges and corners where key-points or salient features are traditionally detected and the descriptors are computed. Contrarily, surface micro-structures represent slowly varying random fields that can potentially be modeled using autoregressive or Markov fields [8].

This means that micro-structure features, detected by methods such as SIFT, tend to be sparse and located in the lower scale-spaces. The latter makes them less robust, unreliable, and small feature sets leave less room for redundancy.

This means that even in a scenario where acquisition conditions, such as the lighting, can be accurately controlled, one can expect a significant number of features that can not, or only poorly, be matched between micro-structure acquisitions.

The geometric feature based authentication algorithm takes the following steps:

- Images are acquired and stored without any post processing. SIFT features are acquired from each image surface. The SIFT algorithm is modified to not reject weak features. The database thus contains an image, the SIFT feature vectors and a unique identifier for each enrolled package.
- To identify a package, its image is taken and SIFT feature points and descriptor vectors are attained.
- SIFT descriptors from the query and the database image are matched using the so-called non ambigu matching rule [55], which only accepts a match when the second best match for that particular point is significantly worse. This factor is usually between 1.4 and 1.6. Most importantly, it does not allow more than a
single match per feature point pair, discards outliers and is very effective in finding a stable subset of features that more or less are geometrically consistent.

- The algorithm coarsely filters matches using Hough pose-space clustering [55, 71] followed by RANSAC [58] and then attempts to attain a relation $y_g = Ax_g$ using least squares, where $A$ is a non rigid similarity transformation.
- An image for which this geometrical relation can be established between the query and the database image is deemed authentic and the index of the corresponding matched features is declared as $\hat{m}$.

The algorithm schematics are shown in Figure 2.16. Results for this algorithm can be seen in Table 2.2 and Section 2.8. Example matching from all used datasets are illustrated in Figures 2.23-2.28.

## 2.6 Affinity Based Authentication

To be able to deal with matching feature sets that are small or have an exceedingly large number of outliers, we firstly worked on achieving a better initial match using two state of the art methods, namely rootSIFT [4] and domain-size pooling SIFT [5, 6].

The rootSIFT algorithm [4] uses a Hellinger kernel to convert SIFT descriptors to a new feature map prior to matching. SIFT descriptors are histograms, which means that the $L_2$ distance might not be an optical comparison metric. Using the Hellinger kernel has been shown to give significant improvement in retrieval tasks and is trivially implemented.

Examples of matching original SIFT and rootSIFT descriptor matching are shown in Figure 2.17 for FAMOS-L when using large image patches, whereas Figure 2.18 shows a result when only a $150 \times 100$ patch has been enrolled. For high resolution images rootSIFT indeed gives an advantage, but crucially, not for small image patches, which are the focus of this Section.

Domain-sized pooling SIFT, or DSP-SIFT [5, 6] is a modification of the original SIFT algorithm where gradient orientations are gathered and pooled over all scales simultaneously in the spatial pyramid. These are thus re-scaled to such that they all have identical size. DSP-SIFT has shown substantial improvement over traditional SIFT when used in retrieval architectures.

An example of matching two samples from the FAMOS-W RAF set using traditional SIFT and DSP-SIFT is shown in Figure 2.19. It is clear that in this example, contrarily to conventional matching, DSP-SIFT does not falsely mark correct matches as outliers.
DSP-SIFT based matching was tested on full resolution images from the entire FAMOS-W RAF set and FAMOS-L giving on average an improvement of 9.5% for FAMOS-W RAF and 7.2% for FAMOS-L. This percentage indicates the average increase in found inlier matches, where the ground truth is established by RANSAC.

However, when small patches are used for enrollment, as shown in Figure 2.20, where a small 150 × 100 patch was matched against a original sized query image, the competitive advantage is insufficient.

Therefore, work was done to modify the feature based matching and outlier detection algorithm. Specifically, the envisioned solution should not rely on Hough pose-space clustering or RANSAC, again to be able to deal with a small amount of matches with a relatively large (over 50 percent) amount of outliers. Secondly, as detailed in Section 2.7, when quantization and compression is used on the descriptors, the matching becomes more prone to errors. The stipulation that a matching subset must adhere to a ridged similarity transform is than in itself not sufficient, as such as subset can nearly always be found.

The general philosophy behind affinity is to quantify the relationship between data points, mostly based on other points within a certain distance sphere, where near similar points have larger affinity. It has been applied successfully to clustering [72] and dimensionality reduction [73].

The used affinity algorithm builds a so-called Euclidean Distance Matrix (EDM) for features both in the query image and for those in each enrolled image [74, 75]. These are compared to each other to remove outliers. This algorithm is an effective simplification reminiscent of [76, 77]. An example of this type of matching can be seen in Figure 2.29b.
(a) The residual 29 false matches and outliers from traditional SIFT matching. Clearly, the left bottom matches should in reality be labeled as inliers.

(b) The residual 16 false matches and outliers from rootSIFT matching.

Figure 2.17 – Illustration of the false matches and outliers between SIFT features from two FAMOS-L acquisitions from an identical sample. Figure 2.17a shows a result when SIFT descriptors are matched using Lowe’s non-ambigu matching rule. Figure 2.17b shows the results from applying rootSIFT [4]. The latter performs better, but the difference is too small to be a sufficient stand alone solution for this particular dataset.

(a) Inliers from traditional SIFT matching.
(b) Outliers from traditional SIFT matching.
(c) Inliers from rootSIFT matching.
(d) Outliers from rootSIFT matching.

Figure 2.18 – In- and outliers results from the matching of features from an enrolled 100 × 150 patch against a high resolution image from the FAMOS-L set using traditional SIFT features and rootSIFT [4]. The correct sub-patch is marked in blue. Clearly, neither method achieves a satisfactory result.
Figure 2.19 – Illustration of the false matches and outliers between SIFT features from two FAMOS-W RAF acquisitions from an identical sample. Figure 2.19a shows a result when SIFT descriptors are matched using Lowe’s non-ambiguous matching rule. Visually it is obvious that most notable in the bottom left corners, a number of inliers has been falsely rejected. Figure 2.19b shows the results from applying DSP-SIFT [5, 6]. For big image samples from an industrial camera, such as these from the FAMOS-W RAF set, domain-size pooling SIFT gives a 9.5% improvement.

Figure 2.20 – In- and outliers results from the matching of features from an enrolled patch against a high resolution image from the FAMOS-L set using traditional SIFT features and DSP-SIFT [5, 6]. The correct sub-patch is marked in blue. Both methods achieve two correct matches, although the SIFT based matching wrongly marks an inlying match as false.
2.6.1 Affinity algorithm

The algorithm, as schematically seen in Figure 2.21, takes the following steps:

- SIFT features and descriptors are determined and stored.

- SIFT features from a query and database image are matched with the non ambigu matching rule [55] and a relatively loose threshold of 1.2. All points without valid matches are discarded. The crucial detail is that the query and enrolled feature set with potential matches will now have the same size, as points are uniquely matched, or not at all.

- For all points in the query image, the geometric distance to all other points is determined. The result is an EDM, as if the points form a fully connected graph, of which the upper triangle is stored. The same is done for the matching points in the database image.

- Points in the query image and in the enrolled image are geometrically normalized. The normalization function translates each set of 2D homogeneous points so that their centroid is at the origin and their mean distance from the origin is $\sqrt{2}$ [71].

- The two normalized EDM matrices are divided by each other to form a final feature matrix.

- Matching points pairs are accepted to the final set, only if their distances to neighbors in the query image are consistent with a certain geometrical transformation to those in the database image. In practice, this means rejecting points whose differences exceed the standard deviation of the feature matrix.

- A query image is deemed authentic, if is has three matching point-pairs in the final set between which a relation $y_g = Ax_g$ can be established.

Figure 2.21 – Schematics of the major components of the affinity based identification algorithm. All individual building blocks are seen schematically in Figure 2.22.
Figure 2.22 – Diagram of the affinity based geometrical matching algorithm for feature points.
Chapter 2. Micro-structures

(a) True accept, geometry based matching
(b) False accept, geometry based matching
(c) True accept, affinity based matching
(d) False accept, affinity based matching

Figure 2.23 – Example results from the RAF 2592 × 1944 versus RAF 150 × 100 setup for geometry (Figures 2.23a-2.23b) and affinity (Figures 2.23c-2.23d) based matching. Affinity falsely accepts (2.23d) when the number of matches drops to the minimum and all points are outliers. Geometry also exhibits false accepts (2.23d) the moment it may fit 3 out of 4 matches.

2.6.2 Hybrid identification

As evident from the ROC curve in Figure 2.11, ACF-Template based matching (Section 2.3) is not sufficiently precise to perform identification on the FAMOS-L dataset using an aligned patch as fingerprint.

It can, however, be used as a pre-processing step prior to feature based authentication. Roughly aligning samples prior to authentication improves performance significantly, as shown in Table 2.3, especially when using small enrolled image patches for FAMOS-L. Notably, the probability of false alarm is ten times smaller whilst to probability of miss, goes to zero.
Figure 2.24 – Example results from the NIK 1601 × 1201 versus NIK 150 × 100 setup for geometry (Figures 2.24a-2.24c) and affinity (Figures 2.24d-2.24f) based matching. Both algorithms suffer from false accepts if a sufficient number outliers is consistent in their error, effectively making them a consistent inlier set.
Figure 2.25 – Example results from the RAF $2592 \times 1944$ versus NIK $150 \times 100$ setup for geometry (Figures 2.25a-2.25c) and affinity (Figures 2.25d-2.25f) based matching. Both the geometry (2.25a) and the affinity algorithm (2.25d) give a true accept, but erroneously include 2 respectively 1 outlier in the final set. False accepts occur the moment erroneous matches are consistent with certain geometrical transform.

Table 2.2 – The probabilities of false alarm, $P^F_M$, $P^A_M$, and probabilities of miss $P^F_P$, $P^A_P$ for feature-geometry and affinity based authentication.
Figure 2.26 – Example results from the NIK 1601 × 1201 versus RAF 150 × 100 setup for geometry (Figures 2.26a-2.26c) and affinity (Figures 2.26d-2.26f) based matching. A trade off must be made for geometry based matching on the minimum size of a subset that is coherent in its rigid transformation. Smaller allowable subsets increase the number of false accepts (Figure 2.26b), whereas increasing this number will result in obvious misses (Figure 2.26c).

Table 2.3 – The probability of false alarm, $P^H_F$ and probability of miss $P^H_M$ for hybrid based authentication on the FAMOS-L set. Roughly aligning samples in itself does not allow for fingerprint based identification, but as a precursor to feature based authentication, it gives a dramatic performance increase.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Camera</th>
<th>Enrolled</th>
<th>Query</th>
<th>$P^H_F$</th>
<th>$P^H_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAMOS-L</td>
<td>Mobile</td>
<td>3264 × 2448</td>
<td>3264 × 2448</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3264 × 2448</td>
<td>100 × 150</td>
<td>0.02</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 2.27 – Example results from the FAMOS S 150 × 100 versus 150 × 100 setup for geometry (Figure 2.27a) and affinity (Figures 2.27b-2.27c) based matching.

Figure 2.28 – Example results from the FAMOS-L 150 × 100 versus 3264 × 2448 setup for geometry (Figure 2.28a) and affinity (Figure 2.28b) based matching.
Figure 2.29 – Two examples showing a feature based authentication attempt between a large query image and a small enrolled micro-structure patch. Figure 2.29a shows the failure of the feature based algorithm (Section 2.5) due to the large number of outliers. Figure 2.29b shows the results for the modified hybrid approach, as detailed in Section 2.6.2.
2.7 Feature geometry based authentication with Quantization

In order to ease the computational and storage burden, SIFT descriptors may be reduced in dimensionality and quantized to shorter binary fingerprints. Especially, when working with natural images, this may be applied with little performance loss [37]. In this section we will empirically test the influence of dimensionality reduction and quantization using random projections on the geometrical matching in an authentication scenario.

Appendix B contains in depth coverage of the statistical behavior of binarized random projected SIFT descriptors, including independence tests and the probabilities of error $p_b$ between fingerprints originating from an image and a distorted copy.

Specifically, we will follow the basic feature matching procedure from Section 2.5 (Figure 2.16), but with compressed binarized descriptors. The initial matching remains based on the non-ambigu matching rule [55], using the Hamming distance.

**Ground truth**

To illustrate the problems that must be overcome in this specific authentication scenario, a ground truth was established in the following manner:

- A ground truth was established by matching the original descriptors using the non-ambigu matching rule [55]. As this step filters out outliers very effectively, this ground truth represents a best case scenario. Manual annotation is not feasible due the size of FAMOS-W and FAMOS-L and the fact that the distortions are not known. This gives a matching for each pair of image acquisitions, per unique physical sample $\{x_g \leftrightarrow y_g\}$.

- All descriptors are reduced in dimensionality and quantized to 128, 64, and 32 dimensions.

- Intra class distances are made up from distances between matching descriptors, as indicated by the ground truth. Inter class distances are all other distances, meaning distances between non matching descriptors from an image pair, and those from all other images in the dataset.

The resulting estimated probability mass function for intra and inter class distances demonstrate how much further processing is needed using further geometric matching. Figure 2.30 shows the results for FAMOS-L and all FAMOS-W results are shown in Figure 2.31.
Chapter 2. Micro-structures

The long overlapping tails of the estimated inter class distributions indicate that even in authentication scenario’s and after filtering out outliers using the matching-rule, there will still be descriptors whose best match is false. Obviously, the overlap between intra and inter class distances becomes bigger the more the descriptor vectors are compressed.

There are a couple of noteworthy phenomena. For industrial cameras, Figure 2.31, compression, especially down to 32 dimensions hurts performance more than using a significant smaller image patch for the descriptor enrollment. For the mobile dataset, Figure 2.30, this is not the case. Enrolling small image patches results in completely overlapping distance classes, for any quantization level (Figures 2.30d-2.30f).

In conclusion, any geometric matching procedure must be able to overcome these hurdles.

**Empirical tests**

Quantization tests were run on three datasets, namely:

- The FAMOS-W, RAF versus RAF set, where the full resolution image of $2592 \times 1944$ was to extract features from, both for enrollment and the queries (Figure 2.32).
- The FAMOS-W, RAF versus RAF set, using a small $150 \times 100$ image patch for feature enrollment (Figure 2.33).
- The full resolution FAMOS-L set, using the original $3264 \times 2448$ images (Figure 2.34).
- The FAMOS-L set where descriptors were enrolled from $150 \times 100$ image patches (Figure 2.35).

Using full resolution images from FAMOS-W for the enrollment and query descriptors, quantization has no performance impact when matching images from identical samples. Figure 2.32a shows such an example using 32 dimensions. The problem, is that in the current geometric matching formulation, it is universally possible, to find a small subset of geometrical matching quantized descriptors, for any pair of non-identical images, i.e. the probability of an authentication false accept, $p_f$, goes to 1. Such a false accept is illustrated in Figure 2.32b. For this specific set, this can be mitigated by adding another classification rule stipulating that the number of matches must exceed $n$ prior to acceptance. However, next to being a new user set parameter, this solution is not universal for our micro-structure datasets.

Figure 2.33 illustrates true and false accepts from FAMOS-W, RAF, when only a small patch of $150 \times 100$ is used for enrollment, for all quantization levels. Although it is visually quite easy to distinguish a matching set from a true and from a false accept, all
Figure 2.30 – Estimated intra and inter hamming distance probability mass functions for projected binarized SIFT descriptors from the FAMOS-L set, using the full resolution image of 3264 × 2448 to extract query features (Figures 2.30a-2.30c) or a small 150 × 100 (Figures 2.30d-2.30f) patch where the dimensionality was retained (Figures 2.30a, 2.30d) or reduced to 64 (Figures 2.30b, 2.30e) and 32 (Figures 2.30c-2.30f). These PMF’s have been estimated using a ground truth, showing that any (geometrical) matching procedure must be able to deal with the fact that there are many erroneous, or inter matches between descriptor that need to be filtered out. In practice, this is not possible.

These matches correspond to a valid non rigid transformation. Section 2.6 introduced an algorithm that is capable of doing this.

Results from the mobile set, FAMOS-L for full resolutions images are seen in Figure 2.34 and for small enrolled patches in Figure 2.35. Quantization is only possible when the original dimensionality is retained and full resolution images are used. Quantized features, even of the original dimension, from the small enrolled patches will always attain a match that is consistent with a certain rigid transformation against (any) query images. A match by the basic geometrical algorithm is uninformative in this setup.

2.8 Results

Results for the feature based identification on the FAMOS-W dataset can be seen in Table 2.2. Although these are obviously very promising, one should make two critical
Figure 2.31 – Estimated intra and inter hamming distance probability mass functions for projected binarized SIFT descriptors from the FAMOS-W RAF-RAF and FAMOS-W NIK-NIK set where either the full resolution image was taken to enroll descriptors (Figures 2.31d-2.31f and 2.31j-2.31l) or just the descriptors from a small patch of 150 × 100 (Figures 2.31a-2.31c and 2.31g-2.31i). The dimensionality was retained at the original 128, or reduced to 64 and 32 respectively. These PMF’s have been estimated using a ground truth, showing that any (geometrical) matching procedure must be able to deal with the fact that there are many erroneous, or inter matches between descriptor that need to be filtered out. In practice, this is not possible.
Figure 2.32 – Examples of a true (Figure 2.32a) and false accept (Figure 2.32b) results from the FAMOS-W RAF 2592 × 1944 versus RAF 2592 × 1944 geometry based matching where the original descriptors have been quantized to 32 dimensions. Both the true and false accepts have matching descriptors that adhere to a similarity transform.

Figure 2.33 – Examples of true (Figures 2.33a-2.33c) and false accept (Figures 2.33d-2.33f) results from the FAMOS-W RAF 2592 × 1944 versus RAF 150 × 100 geometry based matching where the original descriptors have been quantized to 32, 64 and 128 dimensions. Both the true and false accepts have a small number of matching descriptors that adhere to a similarity transform.
Figure 2.34 – Examples of true (Figures 2.34a, 2.34c) and false (Figures 2.34b, 2.34d) accepts for the FAMOS-L $3264 \times 2448$ dataset, using the original 128 and reduced 64 quantized feature vectors.
Figure 2.35 – Examples of true (Figure 2.35a) and false (Figures 2.35b) accepts for the FAMOS-L dataset where samples of $150 \times 100$ were used to enroll descriptors from, quantized to the 128 dimensions.
notes. Firstly, the main disadvantage is that this algorithm is computationally expensive. Secondly, as earlier stated, the FAMOS-W dataset is acquired using an industrial setup and although the photographed samples are misaligned, the lighting and magnification are stable.

In general, one would like to enroll as small a sample as possible. Working with enrolled acquisitions that, in the case of FAMOS-L, are $150 \times 100$ instead of $3264 \times 2448$ eases storage and computational requirements tremendously. The affinity algorithm makes this a reasonable viable solution, keeping the probability of miss low or near zero at the cost of an increased probability of false alarm, when small resolution samples are enrolled.

There are still a number of noteworthy issues. Figures 2.23 shows two false accepts for both algorithms, using FAMOS-W, RAF-RAF, where a small patch of $150 \times 100$ was used for enrollment. As the number of matches drops, the affinity algorithm (Figure 2.23d) suffers from the fact that is does not have sufficient information to decide which point, if any, is the outlier. Small sets made up solely from outliers can thus pass. Geometry is also hurt the moment the number of final matches is four or smaller, as it is always possible to select a subset of three matches to fit a ridged transformation on a plane.

Similar issues arise for FAMOS-W, NIK-NIK, as seen in Figure 2.24. Affinity and geometry algorithms suffers from false accepts the moment outliers are consistent in their deviation from the set, falsely making them the inliers. This happens for all datasets, the moment the enrolled sample is small, and the system only has a couple of descriptor matches to base its authentication decision on.

Tempting as it may to introduce an extra rule that stipulates a minimum number of matches to limit the number of false accepts, this can not be done without introducing more misses, as illustrated by Figures 2.26b and 2.26c. The only viable solution to this issue is to increase the size of the image patch from which features are enrolled.

Results could also be improved for the industrial set by leveraging a priori knowledge. Its cameras are placed above the objects surface and the introduced deformations are therefore limited. Already implicitly done by the affinity algorithm, the geometry variant could be improved by limiting the allowable rigid transformation. A miss, such as illustrated in Figure 2.25f could be avoided in this manor.

Roughly aligning a sample, even though by itself not sufficient to allow the deployment of fingerprinting methods, does improve feature based authentication when used as a second step. However, a printed mark must be present to aid the first alignment step.
Feature quantization prior to geometric matching can most likely be applied successfully when the images the features are extracted from are taken with a high quality industrial camera, with minor adaptations to the matching procedure. However, for mobile images, especially the small ones, from the FAMOS-L dataset, this is not possible.

Very small mobile hand-held samples, without any mark, as FAMOS-S is build up from, form the toughest challenge and predictably give the worst performance. Even with a feature detector, tuned to not reject weak descriptors, FAMOS-S samples rarely have more than 10 robust features of which 3 to 4 will be distinct enough to form a match against an enrolled sample. In general, the affinity matching is more forgiving and is able to process borderline matches of 3 or 4 points.

2.9 Conclusion

In this chapter we have shown that it is feasible to uniquely identify a set of packages based on micro-structures that were acquired using an unmodified hand-held consumer mobile phone without the aided use of a printed mark and with relatively small sample sizes.

To be able to cope with the large variability in image quality due to, for example, lighting fluctuations or the distance and angle the phone was held from the package, our authentication algorithm sacrifices speed in favor of robustness.

Most difficulties arise when the number of matching features drops to five or less. By definition any three points in a plane can be selected to conform to a geometrical transform. Increasing the minimum number results in misses, the reverse in false accepts. For industrial image sets, the camera is fixed over the conveyor, and thus the geometrical distortions are limited to a subset. This a priori information could be used to improve results, but are not, in part because they are not applicable to mobile image sets. Affinity, in general, is able to find small clusters of inliers, but also struggles the moment the outliers make up 70% or more of a dataset. It too could be improved by leveraging a priori knowledge on the allowable deformations, or simply restrict itself to smaller set of geometrical transformations. However, in practice, the only real viable solution is to increase the image size from which enrolled descriptors are taken.

Quantization and compression may be applied, but its success depends more critically on the image quality and size. Especially for (small) images taken with a mobile phone, compression makes authentication infeasible.
Improvement to the affinity algorithm has been difficult mostly because local geometric methods proved to be too sensitive to appearance and disappearance of geometric keypoints.

Any direction that promises to reduce the number of parameters, is naturally worthy of investigation. The current affinity algorithm uses a simple threshold based on the variance of the values in the Euclidean distance matrix, which is not able to distinguish in- and outliers in all situations. A situation that could be improved with a superior statistical model of the Euclidean Distance Matrices.

In conclusion, it is possible to build a viable authentication system, using images solely acquired by mobile phone. But the images from which descriptors are extracted, must preferable be 2 MP and up.

Summarizing, this chapter has shown that micro-structure based identification is possible, if the printed mark is of high enough quality to undo a rigid transformation. Authentication is possible, without the mark, and a mobile phone using robust features.

In the next Chapter, we will introduce a new descriptor that aims to combine the best of both worlds. Its detector works on raw mark-less micro-structures and is robust, whereas its descriptor behaves very much like fingerprint, thus allowing identification.
Chapter 3

The SketchPrint descriptor

In this chapter, a new framework for local content identification is introduced based on the SketchPrint descriptor. It extends the properties of local descriptors to a more informative and discriminative, yet geometrically invariant content representation. In particular, it allows images to be compactly represented by less than a hundred SketchPrint descriptors, does not require any training or a final geometrical re-ranking stage.

Although specifically designed for visually poor and distorted micro-structure images, Sketchprint has been proven to also work on natural images, text and packages.

This chapter is primarily based on the ideas and results presented in [78, 79, 80, 81].

3.1 Introduction

Mobile object identification based on visual features has many applications in physical object security [31]. Discriminative and robust content representation plays a central role in object and content identification [36, 35].

Most visual descriptors originate from the computer vision domain and trade-off specificity against robustness to signal processing and geometrical distortions. The latter not only means a certain geometrical invariance, but such computer vision features also serve as the base input for semantic recognition. Unfortunately, performance is hampered when images are visually poor, lack distinct edges and corners or exhibit many repeated similar structures, as is for example the case with text or micro-structures. Further more, computer vision feature based system easily require over 1’000 descriptors per image and lack a quantitative method to order or select a subset. In practice, this means that an exhaustive final geometric re-ranking procedure is needed.
As discussed in Chapter 2, this means that for our application, relatively large micro-structure images must be retained to extract features from, and crucially, the geometry must be kept. Secondly, due to the exhaustive nature of geometric feature matching, they can only be used for authentication frameworks or in a final geometric re-ranking stage.

In this chapter, SketchPrint, a new descriptor for local content identification is introduced. It is specifically tailored to identify physical objects from images taken with a normal mobile phone. Its descriptor payload is therefore more descriptive than traditional computer vision features based on aggregation or histograms, requires neither training nor any exhaustive geometrical re-ranking procedure.

Importantly, SketchPrint can work on multiple scale levels. It is able to handle printed text and logo’s and works on micro-structures. As such, it can be used to recognize an object such as a package, by first classifying it by its design, and then identifying it, using its micro-structure. An example with hand-held mobile phone images is shown in Figure 3.1.

### 3.2 Related work and challenges

Formally, the goal of identification consists of the estimation of index $\hat{m}$, assigned to the object at enrollment stage, based on a probe or query image $y$.

Our goal is broadly to develop a framework that can efficiently identify an object visually using a hand held mobile device without any physical modifications to this object or special lighting. This means that the images will contain a broad range of distortions that need to be addressed on a device with limited computational power and bandwidth. Furthermore, we wish that our framework can identify objects based on their printed design, next to their micro-structure. Lastly, our framework should be able to work without any exhaustive geometrical re-ranking of a query against the enrolled dataset.
3.2.1 Existing identification architectures

It is commonly known that local descriptors accompanied by spatial information on their originating geometrical positions are a powerful tool for image recognition and alignment [82]. Disregarding the complexity, one can achieve excellent identification results using RANSAC based matching as shown in Figure 3.2a. The drawback of this approach is the need to perform exhaustive matching over all features over all images in the repository.

Partly to ease the computational burden, there exists a family of methods based on the bag-of-features (BOW) model applied to local image descriptors [83] or local geometric configurations of descriptors, also known as geometric hashing [84] (Figure 3.2b). The main idea behind this approach is to reduce the search space of size $M$ for the RANSAC based geometric matching by producing a short list $L(y)$ of images of a size not exceeding a 1’000 containing similar descriptors to the query. The main advantage of BOW based encoding using a histogram of descriptors is its relatively high accuracy for retrieval and acceptable complexity and memory storage for moderately sized databases. Further improvements were achieved via pyramid based BOW frameworks [85], advanced aggregation strategies such as VLAD [86] and methods such as Fisher vectors [87]. The latter extends the BOW frameworks by integrating higher order statistics near attained BOW centroids. Another direction of research aims at reducing the memory footprint and computational complexity by compressing local descriptors to very short binary codes along with their geometrical coordinates within the originating images [88, 89]. Identification accuracy, however, suffers from such optimizations.

Finally, our architecture, shown in Figure 3.2c, does not contain any geometric re-ranking stage and is based solely on robust and discriminative descriptors, i.e., SketchPrint. Since content descriptors form a basis of the considered approaches, we will consider the main shortcomings of existing descriptors.
3.2.2 Existing hand-crafted descriptors

Most computer vision features have two distinct parts, a detector that determines the spatial location of a feature, and a descriptor that captures the properties of the local neighborhood. Detected key-points are usually located in salient image regions such as corners [90] or local extrema in scale-space [91, 92].

Descriptors robustly encode a region of interest around the point. BRIEF, BRISK, ORB and FAST [93, 89, 94] record the difference between selected image pixel pairs. Many methods, such as SIFT [55], SURF [56], AKAZE [95] and CHoG [88] build a histogram over the local gradients. The latter uses Vector Quantization (VQ) for further compression. It should be noted that despite its age, SIFT remains a top tier feature algorithm with solid all round performance.

The main shortcomings of existing detectors and descriptors for physical object detection are:

**Detectors:**

- Very weak robustness to scaling and changing of lightening conditions which results in high birth and death probabilities for key-points.
- Key-points exhibit low location precision.
- There is an absence of measures for the selection or quantitative ordering of reliable key-points.
- Images that lack visual distinctive features such as sharp corners result in unstable performance or lack key-points all together.

**Descriptors:**

- Weak discriminative power results from descriptors being computed around near similar regions, such as corners or text-character edges. More information is also lost when aggregated histograms are used.
- The robustness to geometric transformations remains limited.

Most systems overcome existing weaknesses such as the low-discriminative power and the inability to select reliable features by extracting a multitude of features, in the order of 2'000-3'000 per image indiscriminately. In some cases even by bypassing the detector and simple sampling descriptors from a fixed grid. This necessitates deploying advanced descriptor compression, aggregation and compression strategies.

Even so, the geometrical positions or relationships between these local descriptors still needs to be encoded and stored to ensure (approximate) geometric re-ranking at the final stage to prune the identification results (Figure 3.2b).
3.2.3 Machine learning methods

Hand-crafted features are still prevalent in domains as computer vision, audio and natural language processing. They need expert domain knowledge, their design is labor intensive and time consuming, and in most cases, they do not generalize to other domains or new problems. In contrast, deep learning has made significant progress in recent years, replacing and outperforming handcrafted features using massive training datasets and a host of algorithms for unsupervised and semi-supervised feature learning and architectures for hierarchical feature extraction.

This section will therefore review and test a number of machine learning methods for their suitability for our datasets and application scenario.

3.2.4 Learned feature representations

Ideally, one would like to have an algorithm learn and build feature representations. In recent years, especially when it comes to semantic image recognition and object segmentation tasks, unsupervised learning algorithms that are able to learn feature representations from unlabeled data, such as those based on (variational) auto-encoders [96, 97], SAE networks [98, 99] or Sparse Coding (SC) [100, 101] have shown state-of-the-art performance on datasets such as MNIST [102], CIFAR-100 [103] and Imagenet [104]. Specifically for textures, work from [105, 106] used sparse features for data modeling and classification.

In general these approaches have two stages, the first in which a representation is learned from unlabeled data, followed by a second stage where the learned features are applied to the labeled data to learn a classifier. Critically, the used unlabeled data may be reasonably different from the actual test set, as shown in [101] where random natural images were used to ascertain a feature representation, which was then used to specifically classify elephant and rhino images. Obviously, learned representations can be applied to many other (semantic) image classification tasks.

So called Transfer Learning [107] takes this even further by showing that entire deep networks, that have been trained on a large dataset such as Imagenet, may be used for other image classification tasks with but a few final fine-tuning training iterations using the new dataset. Alternatively, these pre-trained networks may be used as feature extractors by simply removing the last layer such that they output not a class label probability but a feature vector. The latter can then be used as input data to train a new (more simple) architecture on.
To investigate the viability of these approaches, a single convolutional and max pooling network tested on 5000 $128 \times 128$ aligned images from the FAMOS-W RAF set, using filters build up from an SAE and RICA and RP. Secondly, these architectures are also compared to classical Random Projection based fingerprinting [32] (Section B.2).

**Sparse auto-encoder**

A (sparse) auto-encoder is a neural network that is trained unsupervised using back-propagation, where the target labels are the original data. Given $N$ unlabeled training data vectors $\{x^1, x^2, \ldots, x^N\}$, it sets $y^i = x^i$ for $i \in \{1, \ldots, N\}$. It then learns a function $h_{W,b}(x) \approx y$. By placing additional constraints on the solution, such as the number of hidden units in the network, lower dimensional structure may be revealed from the data. Obviously, if $x$ is an i.i.d. realization of a Gaussian random vector $X$, compression will not be possible, nor will the found basis functions reveal any underlying structure. In general, this type of network may find solutions close to those of PCA. Additional structure may be revealed by allowing the number of hidden units to be potentially large, whilst enforcing sparsity of their activations as a constrained solution. Sparsity in this context, means that the average activation $\rho$ of each hidden neuron should be small. This is achieved by adding an extra term to the objective function that penalizes the actual activation $\hat{\rho}_j$ of a hidden neuron $j$, from diverging from the user set sparsity parameter $\rho$ [108]. Here this term is the sum of the Kullback–Leibler Divergence (KLD) over the activation of all hidden nodes $1, 2, \ldots, J$, $\sum_{j=1}^J KL(\rho\|\hat{\rho}_j)$, where $KL(\rho\|\hat{\rho}_j) = \rho \log \rho + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j}$. The final objective function $J(W, b)$ that can be minimized using back-propagation is:

$$J(W, b) = \frac{1}{2} \|h_{W,b}(x) - y\|_2^2 + \lambda \|W\|_2 + \beta \sum_{j=1}^J KL(\rho\|\hat{\rho}_j). \quad (3.1)$$

Alternatively, this criteria may be expressed as:

$$\min_W \lambda \|\sigma(Wx)\|_2 + \frac{1}{2} \|\sigma(W^T \sigma(Wx)) - x\|_2^2, \quad (3.2)$$

where $\sigma$ is the activation function.

**RICA**

*Reconstruction* Independent Component Analysis (ICA) or RICA [109] aims to overcome the orthogonality constraint of traditional ICA by replacing it with a soft reconstruction...
Figure 3.3 – Sparse Autoencoder en RICA basis functions for patches sampled from the aligned FAMOS-W RAF set. These algorithms may discover lower dimensional structure in (image) data, but micro-structures are much more noise-like than natural scene images.

penalty. ICA finds sparse basis functions from whitened data as follows:

\[
\min \| Wx \|_1 \\
\text{s.t. } W^T W = I,
\]

(3.3)

whereas RICA uses the following criteria:

\[
\min_W \lambda \| Wx \|_1 + \frac{1}{2} \| W^T W x - x \|_2^2.
\]

(3.4)

This formulation allows for sparse over complete features. Intuitively, one can see that this formulation is quite identical to that of the SAE in (3.2), but without the activation function and with an \( L_1 \) constraint. Much like auto-encoders, it too may be optimized with back-propagation.
Random Projection based Fingerprinting

Finally, a completely separate RP based fingerprinting technique [110] was used to contrast the performance of the convolutional networks. It is detailed in Sections B.2-C.1 and schematically shown in Figure 3.4.

Given a micro-structure image patch $x(m) \in \mathcal{R}^{N_1 \times N_2}$, the projection is done as follows:

$$\tilde{x}(m) = W_{L \times N}^{L \times N} x(m),$$  \hspace{1cm} (3.5)

where $L$ is the number of dimensions $W_{L \times N}^{L \times N}$ will map to, $N = N_1 \times N_2$ is the length of the input column vector, which for the used micro-structures is $128 \times 128$. Random matrix $W_{L \times N}^{L \times N} = (W_1, W_2, \ldots, W_N)^T$ consists of a set of approximately orthonormal basis vectors, where all elements are generated as $W_i[j] \sim \mathcal{N}(0, \frac{1}{N})$, $1 \leq i \leq N, 1 \leq j \leq L$.

Figure 3.4 – Random Projection based fingerprinting.

Figure 3.5 – Architecture of a single layer convolutional and max pooling network used for feature extraction from aligned FAMOS-W RAF images, using filters obtained from SAE, RICA and RP.
Networks

To deploy the SAE and RICA features, two distinct networks are needed. First, both algorithms are trained unsupervised on a sampled subset of the data to learn features or basis functions. Secondly, a simple convolutional network was used to extract the final feature vectors, that function as identification fingerprints.

The SAE and RICA algorithms can not be used directly on FAMOS-W data. Instead both take 10'000 uniformly sampled $8 \times 8$ gray-scale image patches from the aligned FAMOS-W RAF set. Using larger image patches is computationally infeasible and gives unstable results. Both networks are then used to learn two filter banks, one set of 64: $8 \times 8$ features and a second set of 25: $8 \times 8$ features.

Specifically for the SAE algorithm the sampled patches were normalized and contrast stretched slightly to $[0.1, 0.9]$ to accommodate the sigmoid activation function. For the RICA algorithm, samples were decorrelated using Zero Components Analysis (ZCA) whitening [103]. Finally, the RP features can be generated data independently. The SAE and RICA found basis functions are visualized in Figures 3.3a-3.3d.

The second convolutional network, shown in Figure 3.5, uses the earlier learned features as filters, followed by a max-pooling layer without any further training. The filter bank, either formed with SAE, RICA or RP features contains 64 or 25 filters, sized $8 \times 8$, resulting in 64 or 25 feature maps, each of $(128 - 8 + 1) \times (128 - 8 + 1)$, followed by max pooling over regions of $19 \times 19$ and $5 \times 5$. The last layer simply concatenates the results.

Test setup

Learned or generated features from the SAE, RICA and RP were used as filter-banks for the convolutional network (Figure 3.5), either with 64: $8 \times 8$ features or with a set of 25: $8 \times 8$ features. Max-pooling was performed with a region of $19 \times 19$ and $5 \times 5$.

The FAMOS-W RAF set contains 3 acquisitions per unique sample. The first set was used to train basis functions on. The final test dataset is formed by $2 \times 5000$ aligned $128 \times 128$ patches from the other two acquisition rounds.

The convolutional network is used as feature extractor, from which intra and inter class statistics are derived.

Finally, Random Projection based fingerprinting simply uses a single RP matrix to project the $128 \times 128$ FAMOS-W RAF patches down to a $1 \times 64$ dimensional fingerprint.
Results

The ROC results for the two convolutional networks, when their encoded features are used for identification are shown in Figure 3.6. Although they might seem promising, there are a couple of critical issues that need be addressed.

Firstly, with only 3 samples per unique physical object (or class), of which there are 5000, only unsupervised techniques that encode feature vectors are applicable. Data augmentation is commonly used to enhance performance of deep learning networks and to prevent over-fitting, but will never be applicable in this case. Firstly, because all our samples from all objects are most probably generated by the same (unknown) distribution. Secondly, tests on FAMOS-L have shown that artificially induced distortions, such as a contrast shift or a geometrical transformation, are a bad substitute for the real life distortions that occur when both the object and the acquisition device are hand-held by someone.

Secondly, the data used for finding basis-functions and encoding was already aligned perfectly, using template based alignment. There are too little training samples to allow for cropping and extraction deviations.

Thirdly, both generating the basis functions, and the encoding by the convolutional network is very computational intensive, and subject to many hand picked hyper parameter choices, that critically influence the result. Figure 3.5 shows the RICA results for max pooling with domain sizes of 5 × 5 and 19 × 19, where the latter outperforms the former. Network architecture design and efficient hyper parameter selection remain open and active fields of research.

Fourthly, the final encoded fingerprint has a dimension of 3136, which is extremely large for this particular dataset considering other fingerprinting methods.

Fifthly, the original images from both FAMOS-W and FAMOS-L are large, between 2 and 8 megapixel. Only a fraction, in the form of a tiny patch of 8 × 8 is currently used, as the basis pursuit algorithms and (convolutional) networks simply can not process larger input.

Sixly, the used pooling regions are large. Max pooling with a size of 19×19, outperformed the much more commonly used smaller region of 5 × 5. This suggests that very little discriminative features get extracted.

Finally, there is no clear difference between results obtained using learned sparse or over-complete basis functions, when compared against those generated by RP.
Figure 3.6 – Authentication results using a convolutional network feature basis on SAE, RICA and RP basis-functions for the aligned FAMOS-W RAF set, with two different sized max pooling regions of $5 \times 5$ and $19 \times 19$. It is notable that there is little difference between results from different sized pooling layers. Lastly, using independently generated RP features gives similar performance to those that are trained.

Figure 3.7 – Authentication results using Random Projections, projecting aligned $128 \times 128$ FAMOS-W RAF patches directly down to a 64 dimensional feature vector. These results are superior to those of the used convolutional network in Figure 3.6.

In conclusion, contrast all the above points with Figure 3.7. It shows the results for the same dataset, where a $L = 64$ dimensional fingerprint vector was extracted, simply using random projection based fingerprinting. For this particular dataset, it achieves error-less performance, be it with a tiny margin. Never the less, for such a relatively simple form of projection it does highlight that the above tested unsupervised techniques offer no real advantage over simple fingerprinting or even using the raw image data itself.
3.2.5 Supervised Learning

Supervised learning was used for a completely different subproblem, in our domain, namely the stability of detected key-points. A number of experiments was done so investigate if a classifier would be able to separate reliable from unreliable key-points, using a labeled ground truth.

3.2.6 Learning key-point reliability

There have been data driven supervised learning approaches to filter out reliable key-points. For example by distorting a copy of an image and logging which features survived. This information, either from the original image patch, or the pay-load descriptor may then be used to train a classifier [111, 112]. Specifically, for our datasets, tests have shown that artificially induced distortions, which are needed to be able to automatically establish a ground truth, are not similar to the distortions caused by real life acquisition, both for the industrial FAMOS-W and mobile FAMOS-L set. The latter tends to hurt key-point performance significantly more. The main reason for this, is that differences in lighting allow for minute detail to appear and disappear in the acquisitions.

We used three fairly standard proof of concept architectures and classifiers to briefly empirically test this approach:

- A SIFT based BOW architecture.
- A Convolutional Neural Network (CNN) (DeCAF) combined with transfer learning and logistic regression.
- A pre-trained CNN (DeCAF) as feature extractor followed by logistic regression.

Training set

Two training sets were build. In all cases the training set was build up by following the same procedure detailed in Section 3.5.2 and shown in Figure 3.8. SIFT features were detected in an original image and a distorted copy. Because the distortions are a priori known, a ground truth can be established for the detected features in the distorted copy. Surviving and disappearing key-points are labeled, whilst newly formed key-points detected in the distorted copy are ignored. This is a simplification as it does not predict the appearance of new points in the second distorted image.

SIFT descriptors can written out to a two class dataset, with stable and unstable descriptors.
Chapter 3. SketchPrint

Secondly one may also use the underlying image. The SIFT descriptor is formed from a scale dependent ROI around the detected feature point. This similar sized region is used to extract the underlying sub image patches. The resized and rectified patches form the new two-class dataset, from which a classifier is trained to determine if there is difference between regions resulting in stable and unstable key-points.

**BOW**

The BOW architecture used the original labeled SIFT descriptors, followed by *k*-means clustering and feature mapping using a homogeneous kernel map [113, 114] to be able to efficiently use a linear Support Vector Machine (SVM). Prior to the emergence of deep learning this was a standard simple architecture for (semantic) retrieval tasks. Here it was used to separate the descriptors in to two classes, to quickly ascertain its viability as a potential solution.

**CNN**

The CNN networks were both *DeCAF* networks [115], both of which were pre-trained with weights from *Imagenet* [104]. Using pre-trained networks is the current norm, in absence of sufficient (labeled) training data.

Pre-trained networks may be trivially adapted by swapping out layers, and doing a final number of fine-tuning training epochs on the target training data. The first used CNN
was adapted by replacing the last output layer for a two-class logistic regression output layer. The network was then fine-tuned on the image patches from stable and unstable key-points.

Secondly, if one again removes the final output layer of a pre-trained DeCAF network, such that the output layer becomes the last dense connected layer with a 1'000 neurons, the network effectively becomes a pre-trained feature extractor [107]. These extracted features may then once again be used as input to a new classifier or network, in our case, again a single logistic regression output layer.

Results

None of these architectures achieved over 64% accuracy in predicting if a key-point would survive the distortions as outlined in Section 3.5.2. In other scenario’s, such as image stitching, this would certainly be beneficial. However, our specific goal is to use these key-points as a robust feature, for which this percentage does not suffice.

3.3 SketchPrint Objectives

SketchPrint aims at resolving the main shortcomings of existing methods by:

a) Having a key-point detector that is more stable against distortions.

b) Capture geometrical informations as intrinsic part of the descriptor.

c) Using a local descriptive content description that is highly discriminative such that it may be used as a fingerprint.

d) Be applicable to images that are visually poor, e.g, little or no edges and corners.

e) Negating any geometric re-ranking as suggested by the architecture shown in Figure 3.2c.

f) Does not require a training dataset.

3.4 The SketchPrint Algorithm

In this section, we consider the SketchPrint algorithm at the heart of the an identification system with the architecture presented in Figure 3.2c. The name SketchPrint comes from the sketching as a way of extracting features and content fingerprinting. We consider a sketch to be a read out of a signal between any reference system defined by two key-points. This signal should be properly processed to ensure its invariance to the signal
processing distortions and geometric transformations. An example is visualized in Figure 3.10 showing the SketchPrint descriptor on images, micro-structures and text.

The SketchPrint consists of four main stages:

- Key-point detection
- Key-point selection
- SketchPrint extraction
- SketchPrint filtering

### 3.4.1 Detector

SketchPrint can work with any key-point detector in principle. However, since its detector sketches between pairs of key-points, it needs very few but very stable key-points. The first, because the number of potential sketches that could be extracted and need to be examined is a combinatorial combination of the latter. Similar, because a sketch descriptor is always extracted between a pair of two detected key-points, it is more susceptible to key-point death and birth as a key-point may contribute to more than one extracted SketchPrint.

It iterates through the following procedure to select a limited set of reliable points.

1. Features are obtained using any detector. In this chapter, this is ORB.
2. All points are spatially clustered by forming fully connected graphs between vertices that are within a distance $\lambda$ of each other.
3. Fully connected components are replaced by a single new vertex, whose parameters are the average of the component it came from.
4. Connected components with less members than a certain threshold $\theta$ are removed.
5. Parameters $\lambda$ and $\theta$ are set in practice to attain approximately 32 final new feature points.

An example of spatial clustering can be seen in Figure 3.11. A schematic overview can be seen in Figure 3.9 while Section 3.5.1 and Tables 3.1 and 3.2 cover key-point reliability in detail.

### 3.4.2 Descriptor

The descriptors aim at finding the most discriminative sketches for an image and creating a robust representation.
1. Candidate point-pairs are initially selected based on the spatial distance between them. It may not exceed twice the number of interpollants used in the final descriptor representation nor be less than half.

2. A $3 \times 3$ region of interest is taken around each point.

3. A total of 9 lines are traced over the original image between those 9 point-pairs.

4. All gathered traces are averaged to a single signal.

5. The signal is re-interpolated to a fixed number, re-scaled and re-normalized.

6. The variance of the local variances is determined. If this value falls below a certain threshold, the candidate trace is rejected.

7. The order statistics of the local variances are determined. Candidate traces which do not attain 80 percent of the total value at the halfway pivot, are rejected.

8. Finally, SketchPrints may be ordered, written out, or processed with a further quantization step.

A schematic visualization on different media is seen in Figure 3.10, where a single SketchPrint is seen on a natural image, a micro-stricture, and a logo for both the original and a noisy geometrically distorted copy.

The basic algorithm schematics can be seen in Figure 3.9 and a work flow example in Figure 3.12. Section 3.5 details the major design choices and considerations.

### 3.5 Design elements

In this section we will cover two major design elements of SketchPrint, namely the selection of reliable key-points and the filtering of informative descriptors. Where relevant, the Sketchprint algorithm design choices will be tested against state-of-the-art competitive methods.

#### 3.5.1 Key-point stability

SketchPrint relies on little but very robust key-points. The need to able to select a small subset of reliable key-points is two fold. As a trace is by definition between a pair of key-points, it is more vulnerable to key-point births and deaths. Secondly, the number of potential SketchPrints for $n$ key-points is $\binom{n}{2}$. The basis algorithm also excludes the two common computer vision strategies to deal with spurious or non-informative feature points, BOW type aggregation and exhaustive geometric matching.

Currently, there exists a number of strategies for selecting reliable feature key-points:
Figure 3.9 – Basic block diagram of the SketchPrint algorithm, consisting of two main stages. The detector part spatially clusters found key-points to find a small but robust subset of key-points. The descriptor part traces over the image grid saving fine-grained data, while filtering out signals that are non-informative.
Figure 3.10 – Examples of SketchPrint descriptors extracted from the original natural, text and micro-structure images (3.10a), (3.10d), (3.10g) and distorted copies (3.10b), (3.10e), (3.10h) and their comparison (3.10c), (3.10f), (3.10i).
Figure 3.11 – An example showing SketchPrint spatial clustering. Detected points are shown in blue. Found clusters are indicated with a green bounding box, next to single point clusters indicated with green spheres. The latter are removed completely.
Figure 3.12 – The main stages of the SketchPrint descriptor algorithm for the original image and distorted image: (3.12a), (3.12d) original and distorted images with the ORB key-points detection, (3.12b), (3.12e) key-points filtering, (3.12c), (3.12f) SketchPrints extraction and (3.12g), (3.12i) SketchPrints filtering and an example of a match.
• Selecting large scale key-points for algorithms that use scale-space.
• Selecting key-points that have large gradient magnitudes.
• Selecting key-points that have multiple distinct gradients.
• Selecting key-points whose location is in a Maximally Stable Extremal Region (MSER) [116].
• Using Adaptive Non Maximum Suppression (ANMS) [117, 118].
• Tuning algorithm specific parameters to reject weak key-points.
• Selecting key-points using supervised learning [111].

In this Section, all these methods will be empirically tested. An overview of all tested methods and relevant parameters can be seen in Table 3.1.

Scale-space

The scale-space of an image [91] refers to a stack of smoothed image copies, usually build with a Gaussian kernel. The basic idea is the observation that humans perceive objects differently depending on the scale or distance from that object, for example seeing a tree from a distance, and individual branches and leaves once close. By detecting features in each scale-space layer, or in each progressively blurred, re-sampled image, one can detect both individual parts next to objects as a whole. Naturally, points detected in the largest scale space, are detected in an image with only low frequency components. These points are thus stable against all noise which will have been filtered out by repeatedly applying the Gaussian filter, which acts as a low pass filter.

Gradient magnitude

All key-point detectors define a measure to ascertain if a candidate point lies on an edge or a corner, and if so, what the quality of the found corner is. The oldest, used by the Moravec corner detector [119] considers a corner to be a point with low self-similarity. The Harris [120] algorithm builds up the so-called $2 \times 2$ Hessian matrix, for each pixel. Two large eigenvalues are indicative of a corner. However, the authors show that evaluating the determinant and the trace is also sufficient, which is computationally less intensive. SIFT also uses this measure, in a post processing step to filter out weak candidate key-points. Regardless of the specific measure, in all cases the magnitude of the key-point gradient is indicative for the distinguishability of the found point and its region, and thus its robustness [121, 122].
Multiple gradient magnitude

SIFT determines the major gradient by determining the (smoothed) histogram of gradient orientations in a ROI typically 1.5 times larger than the scale of the key-point [55, 114]. When this histogram has multiple extrema within 80% of the maximum, the key-point is written out multiple times with different major gradients, leading to a feature with multiple distinct gradients.

Key-points within an MSER

MSER is a so called blob detector [123, 124, 125]. It finds stable regions by iteratively thresholding an image and examining how connected components behave for each threshold level. Stable regions have low variations between levels. MSER has proven to be a reliable region detector which is robust against viewpoint and light changes. This has led to the hypothesis that detected key-points that also lie inside an MSER should be more stable than those that are not.

Adaptive Non Maximum Suppression

ANMS [117, 118] attempts to reduce the number of key-points while retaining a selection that is both spatially (Poisson or Uniformly) distributed over the image and robust. This is achieved by suppressing neighboring key-points within a certain radius of a selected key-point where the radius is relatively dependent on the gradient strength of the closest significantly stronger neighboring key-point.

Algorithm specific parameters

Many feature detectors have algorithm specific parameters. The most known basic parameter for the Harris detector [120] is the scale-selection. For the SIFT implementation by [114] the most important ones are the peak threshold and the edge threshold. The first determines the absolute value threshold below which peaks in the Difference of Gaussian (DoG) pyramid are removed, the second removes peaks whose curvature is too small. Points originating from strong circular peaks should be more reliable.
3.5.2 Stability testing protocol

For testing, the images from the INRIA Holiday set [126] were used. It contains 1491 amateur holiday images from a wide variety of scenes. From this set, three distorted copies were made distorted by:

- A projective transformation.
- A projective transformation followed by Additive White Gaussian noise and JPEG compression.
- A projective transformation followed by a Gamma transform.

Query points $y_g$ and enrolled points $x_g$ are matched using a bounded distance decoder with radius $\gamma$. Points may not be assigned to multiple other points. If there are more candidates only the best is assigned as matched, all others are either re-assigned within the radius $\gamma$ or left without a match. This is illustrated in Figures 3.13 and 3.14.

The Probability of Miss, $P_M$, the probability of False Alarm, $P_F$, are then defined as follows:

- $p_M$: After matching all points from set $y_g$ to set $x_g$, $\{y_g \rightarrow x_g\}$, all non-assigned enrolled points $x^i_g$ are counted as a miss, i.e:

$$P_M = Pr[d(x^i_g, Y^i_g) > \gamma]. \quad (3.6)$$

- $p_{FA}$: After matching all points from set $y_g$ to set $x_g$, $\{y_g \rightarrow x_g\}$, all non assigned points $x^i_g$ that are within the radius $\gamma$ of a point $y^j_g$ where $\{y^j_g \not\rightarrow x^i_g\}$ are counted as false alarm.

$$P_{FA} = Pr[d(x^i_g, Y^j_g) \leq \gamma], \quad i \neq j. \quad (3.7)$$

These metrics were selected to explicitly reflect the two following situations which are both unfavorable. Many detectors can be set to detect hundreds if not thousandths of key-points to the point of almost uniformly covering an image. A large percentage of these feature points will match those in the distorted image, not due the feature robustness, but simply due to abundance of features. The inverse situation in which only two or three features are selected and matched, must also be avoided.

Specifically for SketchPrint, we seek a robustness criteria that will help selected a smallish subset of approximately 20 key-points that are extremely stable.
3.5.3 Results

There is no single method currently that is decisively able to predict or select stable key-points for the given distortions. Table 3.2 shows conclusively that common heuristics found in literature such as scale based selection, selecting features with significant or with double gradients all offer no real advantage.

Good results can be obtained the moment the number of detected key-points is very high. Reliability is then not so much a function of the key-point strength per se, but caused by the fact that all points have a good uniform coverage over the entire image.

In regimes with but a few selected features, only spatial clustering results in relatively low $P_F$ values.

3.5.4 SketchPrint descriptor filtering

To produce a controllable and robust number of descriptors, we apply two types of filtering to extracted cross-sections between two key-points.

To maximize the information content of a descriptor, as expressed by its entropy, the local variance is measured. It is assumed that the normalized cross-section samples follow an i.i.d. Gaussian distribution with $\mathcal{N}(0, \sigma_X^2)$ for which the differential entropy is $h(X) = \frac{L}{2} \ln(2\pi e\sigma_X^2)$, thus cross-sections with a large variance maximize the discriminative nature of the descriptor, where $L$ is the length of a cross-section.
Figure 3.14 – Example of the used error measures, demonstrated on a single image. The ground truth is ascertained by a priori knowledge of the induced distortions and exhaustive geometric matching with RANSAC. The red sphere indicates the radius within which matches are accepted. Matching vertices have blue edges.
<table>
<thead>
<tr>
<th>Name</th>
<th>Feature selection method</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT H</td>
<td>Peakthreshold and Edgethreshold set to zero to get the maximum number of features. Peakthreshold removes peaks in DoG scale space that are too small in absolute value. Edgethreshold does the same, but uses the curvature as a metric.</td>
</tr>
<tr>
<td>SIFT L</td>
<td>Peakthreshold set to 20.</td>
</tr>
<tr>
<td>SIFT SC</td>
<td>Order feature points by scale and collect the 32 highest, at most.</td>
</tr>
<tr>
<td>SIFT DG</td>
<td>Only select feature points with double gradients.</td>
</tr>
<tr>
<td>SIFT SCDG</td>
<td>Only select feature points with double gradients and order by scale. Collect the 32 highest, at most.</td>
</tr>
<tr>
<td>SIFT MSER</td>
<td>Only retain feature points that are within MSER regions.</td>
</tr>
<tr>
<td>ORB H</td>
<td>Collect maximum amount of features.</td>
</tr>
<tr>
<td>ORB M</td>
<td>Select 100 features, ordered by scale.</td>
</tr>
<tr>
<td>ORB L</td>
<td>Select 32 features, ordered by scale.</td>
</tr>
<tr>
<td>Harris H</td>
<td>Collect maximum amount of features.</td>
</tr>
<tr>
<td>Harris GL</td>
<td>Order features by gradient strength and collect 32 at most.</td>
</tr>
<tr>
<td>Harris UD</td>
<td>Up and down sample image by 1.5 and 2/3 prior to feature detection.</td>
</tr>
<tr>
<td>Harris UDM</td>
<td>Up and down sample image by 3 and 1/3 prior to feature detection.</td>
</tr>
<tr>
<td>Harris ANMS</td>
<td>Harris algorithm using ANMS.</td>
</tr>
<tr>
<td>CL 1</td>
<td>Select 64 enrolled and 128 query ORB points maximally via scale and cluster.</td>
</tr>
<tr>
<td>CL 2</td>
<td>Select 64 enrolled and 128 query ORB points maximally via scale and cluster. Reject clusters with less than 2 members.</td>
</tr>
<tr>
<td>CL 3</td>
<td>Select 32 enrolled and 64 query ORB points maximally via scale and cluster.</td>
</tr>
<tr>
<td>CL 4</td>
<td>Select 32 enrolled and 64 query ORB points maximally via scale and cluster. Reject clusters with less than 2 members.</td>
</tr>
<tr>
<td>CL 5</td>
<td>Select 64 enrolled and 128 query ORB points maximally via scale and cluster.</td>
</tr>
<tr>
<td>CL 6</td>
<td>Select 64 enrolled and 128 query ORB points maximally via scale and cluster. Reject clusters with less than 2 members.</td>
</tr>
</tbody>
</table>

Table 3.1 – Tested key-point stability selection methods.
<table>
<thead>
<tr>
<th>Detector</th>
<th>Projective</th>
<th>Projective, AWGN, JPEG</th>
<th>Projective and Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_F$</td>
<td>$P_M$</td>
<td>$#x$</td>
</tr>
<tr>
<td>SIFT H</td>
<td>0.04</td>
<td>0.19</td>
<td>551.00</td>
</tr>
<tr>
<td>SIFT L</td>
<td>0.04</td>
<td>0.30</td>
<td>7.50</td>
</tr>
<tr>
<td>SIFT SC</td>
<td>0.03</td>
<td>0.33</td>
<td>30.50</td>
</tr>
<tr>
<td>SIFT DG</td>
<td>0.00</td>
<td>0.30</td>
<td>68.50</td>
</tr>
<tr>
<td>SIFT SCDG</td>
<td>0.10</td>
<td>0.61</td>
<td>15.50</td>
</tr>
<tr>
<td>SIFT MSER</td>
<td>0.00</td>
<td>0.71</td>
<td>50.50</td>
</tr>
<tr>
<td>ORB H</td>
<td>0.66</td>
<td>0.25</td>
<td>445.00</td>
</tr>
<tr>
<td>ORB M</td>
<td>0.56</td>
<td>0.27</td>
<td>88.50</td>
</tr>
<tr>
<td>ORB L</td>
<td>0.27</td>
<td>0.27</td>
<td>29.50</td>
</tr>
<tr>
<td>Harris H</td>
<td>0.05</td>
<td>0.25</td>
<td>134.50</td>
</tr>
<tr>
<td>Harris GL</td>
<td>0.09</td>
<td>0.31</td>
<td>32.00</td>
</tr>
<tr>
<td>Harris UD</td>
<td>0.06</td>
<td>0.21</td>
<td>109.00</td>
</tr>
<tr>
<td>Harris UDM</td>
<td>0.02</td>
<td>0.23</td>
<td>44.50</td>
</tr>
<tr>
<td>Harris ANMS</td>
<td>0.61</td>
<td>1.30</td>
<td>136.00</td>
</tr>
<tr>
<td>CL 1</td>
<td>0.05</td>
<td>0.23</td>
<td>17.50</td>
</tr>
<tr>
<td>CL 2</td>
<td>0.04</td>
<td>0.36</td>
<td>9.50</td>
</tr>
<tr>
<td>CL 3</td>
<td>0.02</td>
<td>0.29</td>
<td>12.00</td>
</tr>
<tr>
<td>CL 4</td>
<td>0.08</td>
<td>0.39</td>
<td>7.50</td>
</tr>
<tr>
<td>CL 5</td>
<td>0.08</td>
<td>0.23</td>
<td>17.50</td>
</tr>
<tr>
<td>CL 6</td>
<td>0.00</td>
<td>0.23</td>
<td>9.50</td>
</tr>
</tbody>
</table>

Table 3.2 – Key-point stability expressed via the probability of miss, $p_M$ and false alarm, $p_F$ for all detector regimes and three types of distortions. $\#x$ and $\#y$ are the average number of detected features at enrollment and query.
Chapter 3. SketchPrint

Given the normalized cross-section $x(m, j)$ for an image $m$ with $1 \leq j \leq K_m$ cross-sections, we compute the local variance in a sliding window of size $W = 5$ along the cross-section using the estimation $\sigma^2_{X_i}(m, j)$, $1 \leq i \leq L$. The variance of these local variances is subject to thresholding. Secondly, the resulting estimates are sorted to produce order statistics $\sigma^{2,OS}_{X_i}(m, j)$, excluding 10 samples at the beginning and end of the cross-section. Begin and end points are defined by the feature points, which usually rest on edges or corners. This type of variability is present around all feature points and therefore discarded as it is not discriminative.

The algorithm requires the order statistics $\sigma^{2,OS}_{X_i}(m, j) \geq T$, where $T = 15$ to comply with the informativeness condition. Order statistics not only help select robust discriminative cross-sections, they are instrumental in removing traces that exhibit a step function, as can be seen in Figure 3.15. These can be highly similar, even if the underlaying images are not.

Finally, the algorithm outputs $J_m$ out of $K_m$ descriptors per image $m$. In the case a fixed number of descriptors $J$ is needed for all images, the descriptors possessing the smallest local variances are filtered out.

3.6 Experimental results

3.6.1 Descriptor distinguishability: ground truth

As a base-line initial comparison, the descriptors from SketchPrint, SIFT and ORB were tested for distinguishability and robustness on the UCID database containing 1'338 natural images [127]. From these images a second copy was distorted with a projective transformation, followed by AWGN and JPEG compression and gamma correction. These are distortions that are quite typical for images acquired with a hand-held mobile phone.

As all distortions are known, a ground truth can be established to where features should be optimally detected. This means that features are detected in the original images, and in the distorted copy the detector is bypassed, forcing feature points to their theoretical location, given the known distortions. Although one can expect that the scale-space extrema will be different in the distorted image copy, fixing these parameter nevertheless gives an idea on the descriptors’ robustness.

Figure 3.16 shows the ROC curves for the three descriptors. The SketchPrint descriptor was tested in two versions: with intensity normalization and using the gradient along the cross-section (Figure 3.16). As expected, SketchPrint outperforms other descriptors on natural images as it retains by far the most original image information. SketchPrint
Figure 3.15 – An example highlighting that sketching over an edge results in highly similar descriptors, even if the originating images are not a like. Such candidate SketchPrints are rejected using order statistics by stipulating that the local variances must be of the same order, e.g., the variance of the local variance may not be dominated by a single factor.

descriptors based on normalized intensity also outperform the gradient based version. Finally, it is worth mentioning that SketchPrint is very distinctive as proven by the very low probability of false accept $P_F$. Practically, this means that only several descriptors suffice for a unique image representation thus greatly reducing the memory and storage burden without negatively influencing performance. SketchPrint robustness expressed by the probability of miss, $P_M$, is also considerably lower than is the case for both SIFT and ORB.

The distinguishability results for SketchPrint are a direct consequence of its descriptor design that retains more information. This design choice was taken in order to allow this descriptor to function as a robust hash, for which but a few need to be enrolled in order
Figure 3.16 – Comparison of ROCs for SIFT, ORB and SketchPrint descriptors, using a priori geometrical information to exclude the impact of the key-point detectors under: A projective transform, AWGN with noise standard deviation $10^{-4}$ and JPEG compression QF=80 (3.16a) and a projective transform and gamma correction (3.16b).

to authenticate or identify an image. Naturally, applying a histogram or quantization function to the region around a found key-point, such as is done by SIFT and ORB, makes them more robust against minor noise fluctuations. SketchPrint solves this, in part, by selecting a subset of robust key-points prior to invoking the detector.

3.6.2 Descriptor performance evaluation: geometrically blind matching

The above test from Section 3.6.1 is artificial in the sense that it assumes a known geometrical correspondence between descriptors, but serves as a useful bound on the maximum attainable distinguishability and robustness for a descriptor. In practice this is not the case. In the most basic identification setup, descriptors originating from a query image are matched blindly against the list of enrolled descriptors, initially without using geometrical information.

To illustrate the core performance an initial test was done using images from text, packages and micro-structures. From all of these a second distorted copy was made, which serves as the noisy query image. The used distortion was a combination of a projective transform, additive Gaussian noise and JPEG compression. Example images are shown in Figure 3.18.

In this first test, we have enrolled a maximum of a 100 SketchPrints, 100 and 1’000 SIFT descriptors selected based on the maximum gradient magnitude in each point.
Table 3.3 – Summarization of the distinguishability results when descriptors are geometrically blind matched, as seen in Figure 3.20. These percentages indicate the average number of matched descriptors, per enrolled image. For example, a 90% score indicates that on average 90% of the descriptors extracted from an individual image, matched correctly between the enrolled and query set.

The query contained an unconstrained number of extracted descriptors. Discarding all geometrical information, the statistics of the ordered distances for descriptors from the query and the enrolled dataset were captured. Schematically, this procedure can be seen in Figure 3.17.

As a test for base performance and correlation amongst descriptors from different images of the same category, the order statistics of distances between the enrolled 100 SketchPrints, 100 SIFTs and 1’000 SIFTs are shown in Figure 3.19. It clearly shows that enrolled SketchPrint descriptors are unique, i.e., the smallest distance between two different images is higher than the distance with the order 100 for SketchPrint for the same image (with the exception of image (d) where less descriptors have been enrolled). For the SIFT descriptor this condition is not always satisfied, especially in the case where a 1’000 descriptors have been enrolled.

The final results where a distorted query is matched against enrolled examples can be seen in Figure 3.20. If the query corresponds to the correct image, the order statistics of distances should slowly grow from the smallest Euclidean distance which corresponds to the best attained match. In contrast, if the query is taken from the wrong image, the distances are very large. If the descriptors are unique and robust and the query’s descriptors do not contain a lot of redundant descriptors, the flat slope of the matched distance order statistics should extend up to the number of enrolled descriptors in the ideal case. After this point, the order statistics should increase rapidly to the level of the non-matched ones. Secondly, the largest ‘matched’ distance in the list of order statistics should not exceed the minimum for the non-matched images. This condition guarantees that the images are clearly distinguishable.

The results for SketchPrint clearly indicate that one can achieve a performance that is competitive to 1’000 SIFT and superior to 100 SIFT performance. A summarization of the results, is shown in Table 3.3, which shows the percentage of query matching descriptors, averaged over all query images.
3.6.3 Identification performance evaluation: BOW

Finally, to validate SketchPrint in full identification mode with aggregation, we used the UCID database of real 1’338 images [127] which have been distorted with a projective transformation, followed by Additive White Gaussian Noise (AWGN) and JPEG compression. In an extreme case, this database of extracted descriptors might be stored on a mobile device. Therefore, we stipulated that each individual image would be represented just by 32 SketchPrint, SIFT and ORB descriptors per image.

A simple BOW aggregation scheme was used to build a single vector per image, prior to testing. Here the number of basis vectors that samples are encoded against, is simply formed by the original descriptor vectors of all enrolled images. This is not efficient, but allows us to ascertain the best possible performance excluding quantization effects.

A statistical model of architecture is the focus of Chapter 4. After enrollment, all descriptors from a query are simply exhaustively matched against the codebook with all enrolled vectors.

The final identification test between descriptors from a query and the enrolled set is based on the earlier mentioned ordered statistics where the decision rule was based on the 3rd ordered distance.

Although the database is relatively small, it was our intention to study system performance when no BOW compression is used. The SketchPrints have been of length 128 and quantized to 8 bits per sample to be directly comparable to the SIFT descriptors.
Figure 3.18 – Examples of the used text documents, packages and random microstructure images.
while the binary ORB descriptors are 256 bits. The systems based on SketchPrint identified all images error-less with different distortions. The ROC plot for SIFT and ORB is shown in Figure 3.21, which shows that SIFT slightly outperforms ORB, and are both outperformed by SketchPrint.
Figure 3.19 – 100 enrolled SketchPrints (3.19a), (3.19d), (3.19g), 100 enrolled SIFTs (3.19b), (3.19e), (3.19h) and 1’000 enrolled SIFTs (3.19c), (3.19f), (3.19i) performance on text documents (3.19a-3.19c), packages (3.19d-3.19f) and random micro-structure images (3.19g-3.19i) where the query and the enrolled item were identical (self-assessment).
Figure 3.20 – 100 enrolled SketchPrints (3.20a), (3.20d), (3.20g), 100 enrolled SIFTs (3.20b), (3.20e), (3.20h) and 1’000 enrolled SIFTs (3.20c), (3.20f), (3.20i) performance on text documents (3.20a)-3.20g), packages (3.20d-3.20f) and random micro-structure images (3.20g-3.20i) on the original and a distorted copy. The distortions were done using a projective transformation, additive Gaussian noise and JPEG compression.
Figure 3.21 – Identification performance for SketchPrint (error-less and therefore lacking a curve), SIFT and ORB on the UCID database using simple BOW aggregation without any (vector) quantization, meaning that all original descriptors are stored and used for the enrollment and queries.
3.7 The Distance Order Statistics framework for micro-structure images

In this section, an identification architecture will be analyzed based on distance order statistics. This architecture does not require geometrical re-ranking.

Let a database $\mathcal{D}$ of enrolled feature vectors contains a collection of $M$ items, represented by their descriptors $\{x^k(w)\}$, $1 \leq w \leq M$ and $1 \leq k \leq J_x(w)$, where $J_x(w)$ denotes the number of descriptors for an item $w$.

The identification problem is to find the best match between a query, represented by a set of descriptors $\{y^j\}$, $1 \leq j \leq J_y$, and those in the database $\mathcal{D}$. The system should produce an index estimate $\hat{w}$ for the best match, or an empty set $\emptyset$ if the query is not related to the database. Alternatively, the system may produce a list of indices with best matches.

The proposed architecture is shown in Figure 3.22. Query descriptors $\{y^j\}$, extracted from image $y$, $1 \leq j \leq J_y$, are matched against the database $\mathcal{D}$ by computing the distance $d_{kj}(w) = ||x^k(w) - y^j||^2$. It is important to point out that any fast approximate distance computation, like those based on product vector quantization [86] may be used. However, our goal is to demonstrate the feasibility of identifying random micro-structures based solely on descriptors without any further geometric processing and re-ranking.

Given the distances $d_{kj}(w)$, one can consider several system designs such that for each descriptor $y^j$ the system returns:

- The whole set of distances for all descriptors stored in the database $\mathcal{D}$.
- Only the list of descriptors and their indices that are within some $\epsilon N$ from $y^j$, i.e., $\epsilon$-$NN$ list $\mathcal{L}(y^j) = \{w : d_{kj}(w) \leq \epsilon N\}$, or the closest $\ell$-$NN$ descriptors and their indices.
- The best match occurs when $\ell = 1$ or equivalently $\hat{w} = \arg\min_{w,k} d_{kj}(w)$.

To investigate the theoretical performance limits, we will primarily consider the first case, to ascertain the minimum probability of miss, leaving memory and complexity issues aside.

Independent of the approach used, all distances computed above for all $J_y$ descriptors are combined to a common stack and the $L \leq J_y$-smallest distances $\tilde{d}_r(w)$, $1 \leq r \leq L$, such that $\tilde{d}_1(w) \leq \tilde{d}_2(w) \leq \cdots \leq \tilde{d}_L(w)$, are produced based on the analysis of Distance Order Statistics (DOS) as shown in Figure 3.22.
The identification decision rule is based on the order statistic detector, which can produce both a soft and hard decision. The latter is defined as:

\[
D_r(w) = \begin{cases} 
1, & \text{if } \hat{d}_r(w) \leq T_r, \\
0, & \text{otherwise}, 
\end{cases}
\] \hspace{1cm} (3.8)

using a set of trained thresholds \(\{T_r\}, 1 \leq r \leq L\) for each order statistic \(r\). These thresholds are set specifically for our application domain, such that the probability of miss for the \(r\)-th DOS is bounded as \(P_{Mr} \leq \epsilon'\), with \(\epsilon'\) to be a small non-negative constant, since SketchPrints are discriminative in nature and thus naturally exhibit small values of probability of false acceptance \(P_{FAr}\). It is important that descriptors are not missed at the decision stage. The decision outputs \(D_r(w, k)\) may also be soft values, computed proportionally to the statistics of correct and incorrect DOSs. This option will be considered in future research.

Following the general BOW with RANSAC strategy the framework should produce a list \(\mathcal{L}(y)\) of the most likely candidates which can then be geometrically re-ranked. The corresponding list decoder is:

\[
\mathcal{L}(y) = \{w : S(w) \geq T\},
\] \hspace{1cm} (3.9)
where

\[ S(w) = \sum_{r=1}^{L} D_r(w). \]  

(3.10)

Note that for simplicity this models the individual order statistics as independent whereas in reality they are dependent. System performance may then be evaluated by the probability of missing a correct item \( w \):

\[ P_M = \Pr[S(w) \leq T \mid w], \]  

(3.11)

for a chosen threshold \( T \) and the probability of falsely accepting a non-related item as matching to an enrolled item in the database:

\[ P_{FA} = \Pr[S(w) > T \mid w'] \quad w' \neq w. \]  

(3.12)

The resulting average list size of retrieved indices is then:

\[ \mathbb{E}\{|\mathcal{L}(y)|\} \simeq MP_{FA}, \]  

(3.13)

assuming the probability of miss is selected as \( P_M \leq \epsilon \).

### 3.8 Experimental validation

For the empirical tests, both FAMOS-W (Appendix A) and FAMOS-L (Chapter 2) were used.

Every image was enrolled once:

(a) For FAMOS-L, with a maximum of 100 SketchPrint descriptors. In Practice for FAMOS-L, this results in an average of 17 descriptors per image sample. For FAMOS-W, with a maximum of 32 SketchPrint descriptors.

(b) With a 1'000 SIFT selected via gradient magnitude.

(c) With an unlimited regime with on average 10'000 SIFT descriptors per image.

[114].

This enrolled set was tested against the other acquired images, that functioned as query. At query side, the number of descriptors was not constrained resulting in on average 50 SketchPrint descriptors for FAMOS-L and 110 for FAMOS-W. SIFT consistently enrolled 10'000 descriptor for both datasets.
A working example is shown in Figure 3.23. The intra class is formed from the ordered statistic distances from descriptors originating from identical images and the inter class distances for those between non-identical images.

An overview of all used descriptor types and datasets can be seen in Table 3.4. It shows the average retrieved descriptor list size $|\mathcal{L}|$ when the decoding rule (3.10) uses 1 to 10 individual order statistics and critically, $P_M = 0$.

When cumulative descriptor statistics are used, the original image may be trivially identified by looking at the intersection between the origins of matched descriptors in the cumulative set. This is shown in Table 3.5 for FAMOS L.

The results for the whole dataset are shown in Figures 3.24-3.30. It should be clear, as the intra and inter first order distances for SketchPrint do not overlap for the FAMOS-L set, that for this set, error-less identification is possible even without additional decoding rules.

For FAMOS-W, SIFT remains an excellent descriptor choice, if and only if, the large number of enrollments is acceptable.

SIFT does struggle more with FAMOS-L, where error-less identification ($|\mathcal{L}| = 1$) is not possible in this particular framework, which can be explained by the fact that neither its feature-points are not stable enough, nor its descriptors discriminative enough for the degraded mobile micro-structure images.

### 3.8.1 Quantization

Both the SketchPrint and SIFT descriptors from the FAMOS-L dataset may be projected and quantized to ease storage and computational burdens. The specific method, using random projections, is covered in depth in Appendix B. Within the DOS framework, this can be done with very little loss of performance as seen in Table 3.4 and Figure 3.31.
Figure 3.23 – Order statistics example for two acquisitions of two micro-structures from FAMOS-L using SketchPrint and SIFT: (3.23a-3.23b) a sample 1 and (3.23c-3.23d) a sample 2 both acquired 2 times, (3.23e) the DOS for 100 enrolled SketchPrint, (3.23f) the DOS for 1’000 SIFT and (3.23g) 10’000 SIFT. The latter can not distinguish these micro-structures due to the overlap of inter and intra statistics.
Figure 3.24 – Boxplot for the first 10 distance order statistics for the FAMOS-L set with the intra distances in red and inter in blue: (3.24a) SketchPrint, (3.24b) 1’000 enrolled SIFT, (3.24c) 10’000 enrolled SIFT and (3.24d) 500 enrolled ORB based identification.

Figure 3.25 – Boxplot for the first 10 distance order statistics for SketchPrint and the FAMOS-W set with the intra distances in red and inter in blue for: (3.25a) RAF camera, (3.25b) OMRON camera.
Figure 3.26 – Boxplot for the first 10 distance order statistics for ORB and the FAMOS-W RAF dataset with the intra distances in red and inter in blue for:
(a) ORB 250, FAMOS-W, RAF camera
(b) ORB 500, FAMOS-W, RAF camera

Figure 3.27 – Boxplot for the first 10 distance order statistics for ORB and the FAMOS-W NIK dataset with the intra distances in red and inter in blue for:
(a) ORB 250, FAMOS-W, NIK camera
(b) ORB 500, FAMOS-W, NIK camera

(3.26a) 250 and (3.26b) 500 enrolled descriptors.

(3.27a) 250 and (3.27b) 500 enrolled descriptors.
Figure 3.28 – Boxplot for the first 10 distance order statistics for SIFT and the FAMOS-W RAF dataset with the intra distances in red and inter in blue for: (3.28a) 1’000 and (3.28b) 10’000 enrolled descriptors.

Figure 3.29 – Boxplot for the first 10 distance order statistics for SIFT and the FAMOS-W NIK dataset with the intra distances in red and inter in blue for: (3.29a) 1’000 and (3.29b) 10’000 enrolled descriptors.
Figure 3.30 – Boxplot for the first 10 distance order statistics for SketchPrint (3.30a), SIFT (3.30b) and ORB (3.30c) for FAMOS-W dataset, matching the RAF camera against the NIK, with the intra distances in red and inter in blue.
Chapter 3. SketchPrint

Table 3.4 – The average retrieved descriptor list size |ℒ| for $P_M = 0$, per used order statistic (3.10), for SketchPrint, SIFT and ORB using the FAMOS-W and FAMOS-L database.

<table>
<thead>
<tr>
<th>Order Statistic</th>
<th>SketchPrint</th>
<th>ORB 250</th>
<th>ORB 500</th>
<th>SIFT 1’000</th>
<th>SIFT 10’000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SketchPrint</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ORB 250</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ORB 500</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SIFT 1’000</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>SIFT 10’000</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order Statistic</th>
<th>SketchPrint</th>
<th>ORB 250</th>
<th>ORB 500</th>
<th>SIFT 1’000</th>
<th>SIFT 10’000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SketchPrint</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ORB 250</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ORB 500</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>SIFT 1’000</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SIFT 10’000</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order Statistic</th>
<th>SketchPrint</th>
<th>ORB 250</th>
<th>ORB 500</th>
<th>SIFT 1’000</th>
<th>SIFT 10’000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SketchPrint</td>
<td>26</td>
<td>13</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>ORB 250</td>
<td>4925</td>
<td>4973</td>
<td>4947</td>
<td>4980</td>
<td>4967</td>
</tr>
<tr>
<td>ORB 500</td>
<td>4978</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
</tr>
<tr>
<td>SIFT 1’000</td>
<td>4978</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
</tr>
<tr>
<td>SIFT 10’000</td>
<td>4978</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
</tr>
</tbody>
</table>

Table 3.5 – The average retrieved final list size |ℒ| per cumulatively used order statistic for $P_M = 0$ (3.10), for SketchPrint and SIFT for FAMOS-L.

<table>
<thead>
<tr>
<th>Order Statistic</th>
<th>SketchPrint</th>
<th>ORB 500</th>
<th>SIFT 1’000</th>
<th>SIFT 10’000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SketchPrint</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ORB 500</td>
<td>234</td>
<td>105</td>
<td>115</td>
<td>102</td>
</tr>
<tr>
<td>SIFT 1’000</td>
<td>19</td>
<td>19</td>
<td>23</td>
<td>27</td>
</tr>
<tr>
<td>SIFT 10’000</td>
<td>17</td>
<td>21</td>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order Statistic</th>
<th>SketchPrint</th>
<th>ORB 500</th>
<th>SIFT 1’000</th>
<th>SIFT 10’000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SketchPrint</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SIFT 1’000</td>
<td>39</td>
<td>40</td>
<td>32</td>
<td>25</td>
</tr>
<tr>
<td>SIFT 10’000</td>
<td>39</td>
<td>40</td>
<td>32</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 3.4 – The average retrieved descriptor list size |ℒ| for $P_M = 0$, per used order statistic (3.10), for SketchPrint, SIFT and ORB using the FAMOS-W and FAMOS-L database.

<table>
<thead>
<tr>
<th>Order Statistic</th>
<th>SketchPrint</th>
<th>ORB 500</th>
<th>SIFT 1’000</th>
<th>SIFT 10’000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SketchPrint</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SIFT 1’000</td>
<td>18</td>
<td>14</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>SIFT 10’000</td>
<td>16</td>
<td>13</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>
Figure 3.31 – Boxplot for the first 10 distance order statistics for the FAMOS-L set with the intra distances in red and inter in blue after projection and quantization for (3.31a) SketchPrint, (3.31b) 1’000 enrolled SIFT.
3.9 Conclusions

In this chapter, we have detailed the SketchPrint algorithm and shown the feasibility of using this framework for identifying objects via their micro-structure using a mobile phone. SketchPrint is able to handle severely degraded images with little or no structure, captures implicit geometrical information, as it uses two key-points, and requires a significantly smaller number of enrollments per sample than other methods.

3.10 Future work

There are a number of interesting directions to further develop SketchPrint in. We will discuss the possibilities for the detector and the descriptor part of the algorithm separately.

Detector improvement

Any improvement in feature stability, or metric to predict feature stability and robustness would not only help SketchPrint, but any feature detection algorithm.

Deep learning based algorithms have consistently, if not dramatically improved the performance in image retrieval and recognition tasks the last couple of years. Somewhat surprisingly, all current state of the art feature detectors are still hand-crafted algorithms. However, using deep learning based approaches for feature detection is an emerging field.

CNN architectures, although traditionally associated with classification problems, may also be used for (object) localization by using regression. In this case the error-function becomes the $L_2$ distance between the coordinates of the box enclosing the sought object in the training set and the network output [128, 129]. Alternatively, one may use guided back-propagation or the deconvolution approach where back-propagation is done up to the input data, thus visually showing which part of the image was used by the network to classify the object in the image. These data-gradients can be used for automatic object selection and cropping [130, 131, 132, 133, 134].

In similar fashion to having a training set with bounded boxes and minimizing the $L_2$ distance between the network output and the training coordinates, a network can also be made to detect key-points. In this case, the training set is comprised of either hand-placed markers or earlier detected key-points with a conventional feature detector. The network takes raw images as input, and the final output layer converts the input from the multi dimensional tensor into a fixed sized feature vector, which is build up from
coordinates of the learned key-points. One can even train different network architectures to detect a specific single type of key-point.

This approach has been successful in detecting landmark features on frontal facial images from the Toronto face dataset [135]. The number of sought key-points is limited and more or less fixed and they are situated on different distinct facial points such as the center of an eye and nose, or a mouth corner.

Although it is possible to train a CNN based in our case on the VGG [128] architecture, to detect simple salient features such as prominent local extrema, or an end point of a scratch, the current networks put much more emphasis on the location of the training feature than the actual surroundings. Meaning, that in their current state, they learn filter weights for an expected key-point position, but can not use these weights to find a similar point, elsewhere in the image.

Micro-structures images are invariably more difficult than images displaying ordinary scenes. Their training key-points can not be be selected by hand, nor is there a correct fixed number per sample.

The challenges notwithstanding, using deep-networks as trainable feature detectors is a highly promising direction.

Descriptor improvement

In terms of the specific design of the SketchPrint descriptors, there are a number of important directions. Although we have found it to be unhelpful to read out SketchPrint descriptors in scale-space, the number of preselected interpolation points is by definition tied in to the image and final SketchPrint size and scale. Automatically choosing this number on an image by image basis would be preferable.

Secondly, because SketchPrint depends on a key-point pair, it is more sensitive to the birth and death of feature points. A SketchPrint variant, which would be used in parallel with the original, that is based on the ROI around a single point, similar to state-of-the-art descriptors, while retaining similar distinguishability would be advantageous.

Thirdly, SketchPrint is used as a single identifying descriptor or in a list-decoding setup. Advanced aggregation and quantization techniques could still be applied over a set of SketchPrints.

Lastly, Sketchprint is potentially a good candidate to deploy ACF [16, 136] on top of. This method makes an invisible modification to the digital image, much like watermarking, to enhance the final extracted feature vectors.
Chapter 4

Bag-of-Word based Content Identification

Bag-of-Word frameworks are used wide-spread for (semantic) image identification, yet their performance is often only assessed experimentally. This chapter presents a statistical framework to analyse the performance of Bag-of-Word based content identification systems. It aims at modeling all individual building blocks such as descriptor robustness and compression, assignment accuracy, the pooling layers and the final decision making progress. The impact of geometrical side information, most notably used in re-ranking for retrieval applications, is also covered. Finally, the developed theoretical model is empirically validated.

This chapter is based on [137, 138, 139].

4.1 Introduction

The BOW framework has been widely used in content search systems, biometric recognition applications based on face or gate and more recently in multimedia security applications including copy detection, black list tracking, content blocking and commercial content ranking systems. The recent systems based on BOW can easily handle large-scale search or recognition problems, even on mobile phones. The BOW approach is based on the construction of a visual alphabet or dictionary build with the clustering of low-level features such as discriminative and robust descriptors. It has proved to be very effective for both identification and (semantic) retrieval applications and allows for an efficient implementation based on inverted indices.
Traditionally, VQ is used for clustering to find basis vectors or centroids, based for example on $k$-means or an other unsupervised learning algorithm [140, 88]. The resulting code-book, or visual alphabet is a hierarchical structure composed of centroids, which is used to encode and aggregate the descriptors extracted from each image into a single fixed dimensional feature vector.

In its most elementary form for retrieval and classification applications, a training set of images is used to generate a set of features which are then clustered. The found clusters are known as visual words, and allow to describe an image with a fixed length histogram where each bucket pertains to a (trained) visual word. Votes are gathered by matching image descriptors to this fixed set of visual words.

A query image is converted into the BOW space after which the search system returns those database entries that are statistically similar to those of the query. In this case, a global similarity measure between the images is approximated by a set of local similarity measures in the space of the learned visual alphabet. Finally, the system may return a list of images using $\epsilon$-NN or $k$-NN search. These images are most often also re-ranked using the geometrical positions of the original descriptors. In most cases this amounts to grouping and matching points via exhaustive descriptor matching followed by RANSAC to find point groups that adhere to a similar (affine) transformation (Figure 4.1).

Therefore, the crucial elements of BOW systems include:

- Learning or gathering robust, discriminative and invariant local or global descriptors.
- The construction of a visual alphabet.
- Encoding of the image descriptors against the learned visual alphabet.
- Pooling the results to fixed-length vectors.
- The final decision making process.
- The list size of the result, if the system returns multiple candidate matches.
In the context of this work, BOW frameworks can play an important role. Firstly, our micro-structure datasets lack sufficient training data for contemporary (deep) learning methods. Secondly, they form an elegant way to aggregate multiple descriptors into a single fixed dimensional vector. Even though the SketchPrint descriptor (Chapter 3) was designed to work without aggregation, other state-of-the-art descriptors are not, and for all it can still be an attractive and viable option to reduce the memory and storage burden.

The state-of-the-art in the BOW research covers several main directions:

Descriptors

The growth of mobile platforms has spurred the development of very short (approximately 32 bits) and binary descriptors such as CHOG [88] and ORB [89] as a replacement for their non-binary, long and computationally heavier counterparts such as SIFT [55] and SURF [56].

Encoding & assignment

The design of new encoding and assignment strategies of descriptors against the visual word codebook is a crucial step. It has been shown repeatedly to be more critical to performance than descriptor or basis-function choice. [141, 142]. Approximations and encodings of the descriptors in the visual word space can be roughly classified in three main groups:

- VQ and hard-assignment [143] where the $i$th descriptor $x^i$ is quantized by a coarse VQ $Q_c(.)$ to its nearest visual word resulting in the approximate $\hat{x}^i = Q_c(x^i)$.
- Source coding with refinement [140] quantizes the descriptor $x^i$ using the VQ $Q_c(.)$ to its nearest visual word while simultaneously storing the quantized refinement information of the descriptor with respect to its quantized version given by the fine quantization $Q_f(.)$ of the residual vector: $\hat{x}^i = Q_c(x^i) + Q_f(x^i - Q_c(x^i))$.
- Soft-assignment [144], [145] approximates the descriptor $x^i$ in the visual word space $c^j$, $1 \leq j \leq J$, with a linear approximation: $\hat{x}^i = \sum_{j=1}^{J} w_j c^j$. Strategies such as sparse coding and locally-constrained linear coding (LLC) [146] include different selection strategies of their weight coefficients $\{w_j\}$.
- Using advanced pooling methods [147] that are applicable outside the BOW context as they do not consider feature dimensions to be independent, contrarily to normal max pooling.
• Triangulation embedding methods [148] that aim to improve both the embedding and aggregation stages of the BOW framework.

Pooling

These aggregation methods in BOW frameworks allow the system to deal with the fact that the number of features in the enrolled image and those of the queries are different. Secondly, they help deal systems with the fact that feature images will lack geometrical consistency. The two most used forms are the so called average and max pooling methods. [144]

Security

Recent work has also shown that it is possible to visually reconstruct images from information deduced from the visual codebooks and the descriptors [149, 150]. Section 4.7 will explicitly model the trade-off between identification performance and privacy preserving search and show-case a practical application. Especially for physical object security, where the actual objects are accessible to the public or when a sample may be reconstructed from its features, this trade-off is important.

4.1.1 Approach

Currently, the design of most existing BOW based systems focuses heavily on memory and complexity aspects in view of the large-scale nature of the search problem. A process that still involves a significant amount of heuristics, best practices and engineering. An example of such a contemporary heuristic is the usage of pooling layers and specifically the use of max pooling in favor of sum or average pooling. This has repeatedly been shown empirically to improve performance [144, 141].

Performance of BOW systems is mostly evaluated empirically. To the best of our knowledge, there is no published work that rigorously links all design elements of BOW-systems to its final performance. This includes the robustness and general properties of the descriptors, feature representation and compression, the encoding and decoding operations, pooling and all system hyper-parameters. Finally, the security of BOW-based systems is not completely explored and lacks systematic study.

Therefore, in this chapter, we are not interested in following some particular design of BOW-based system but will rather consider and explicitly model, the fundamental underlying assumptions, indicate the shortcomings and weaknesses of existing solutions
next to proposing solutions to these problems. In particular, decision theory will be used to evaluate the performance and security of BOW-based systems.

Currently, most BOW systems are used for Content Based Information Retrieval (CBIR), object recognition and copy detection. We will consider content identification where $M$ items are enrolled and given some query, the system should determine the corresponding matching item or issue a rejection. If it is not feasible to return the index of a single matching item, the system should retrieve a list with indices, obviously ensuring that the sought item is on that list.

The content identification problem is traditionally analyzed using the content fingerprinting formulation. Existing theoretical works [151, 152, 153, 38] mostly consider content identification based on a single enrolled fingerprint, sufficiently long to represent the content robustly. In most theoretical works, it is assumed that the enrolled and query fingerprints are perfectly aligned or synchronized, with [151] as notable exception, where fingerprint de-synchronization was modeled by a random shift parameter. In practice it is not feasible to design a single super fingerprint or descriptor that is invariant to a multitude of distortion types and severities, which is the driving force behind the usage of hundreds if not thousands of short fixed-length local descriptors. Individually they all have a reasonable robustness against signal processing distortions and aggregated collectively, against geometrical ones.

However, in this case the length of the deployed descriptors does not satisfy the asymptotic assumptions underlying the theoretical work in [151, 152, 153, 38] which makes the analysis of practical BOW-systems intractable.

Work from [141, 142] showed empirically that for semantic image recognition tasks, the algorithm used to attain the basis-functions (such as k-means in the most simple case) is of relatively little importance in comparison to the chosen encoding method, i.e., the specific way in which raw descriptors are aggregated into a fixed dimensional feature vector, using the found basis-functions.

Since there is little work on the theoretical analysis of BOW-systems’ performance besides [154] and none on BOW based content identification, the goal of this chapter is therefore to provide a simple and tractable model that allows to analyze, optimize and guide the design of BOW systems.

This chapter will consider two cases of compressed and non-compressed features to reveal the theoretical limits of BOW based identification systems, analyze the impact of descriptor compression and encoding next to highlighting the impact of geometrical consistency between the descriptors on overall system performance. Such a formulation was not considered in earlier studies.
Specifically, we will model the establishment of the final retrieved list with possible matching candidates as a function of all the BOW design elements, parameter choices, and distortions. This is done for two distinct scenarios. The first in which the features are not synchronized, meaning that the number of features in each enrolled image and query image is different. Secondly, the upper performance limits are ascertained by assuming that the features are synchronized, a scenario which corresponds to the best achievable geometrical matching between enrolled and query descriptors.

4.2 BOW-based content identification: Problem statement

A database with content is defined as a collection of $M$ items $x(m)$ represented by their descriptors $x(m) = \{x^1(m), \cdots, x^{J_x(m)}(m)\}$, $1 \leq m \leq M$, with each descriptor $x^i(m) \in \mathcal{X}_L$, $1 \leq i \leq J_x(m)$ and $L$ is the dimensionality of the used descriptor type.

The problem is to decide whether a query $y = \{y^1, \cdots, y^{J_y}\}$ is related to some enrolled elements or not. In general, the number of descriptors pertaining to a query is not equal to the number of enrolled descriptors per image, $J_y \neq J_x(m)$. The system should produce a list of indices $\mathcal{L}(y)$ whilst ensuring that the correct index $m$ is always on this list or an empty set, if the query $y$ is not related to any item in the database. The cardinality of this list or the number of retrieved indices on the list is denoted as $|\mathcal{L}(y)|$.

System performance is evaluated by the probability of missing a correct item $m$ on the list $\mathcal{L}(y)$ and the probability of falsely accepting an unrelated item $m'$ as related to some item $m$ in the database. This leads to the average list of accepted items $\mathbb{E}|\mathcal{L}(y)|$. The analysis will focus on the performance of the identification system for a given database size $M$, the parameters of the used descriptors as well as their numbers $J_x(m)$ and $J_y$.

The core idea behind the BOW systems consists in finding a single fixed length representation for each individual enrolled image. This feature vector is build up from coefficients that provide a robust and accurate approximation of the descriptors in terms of (trained) basis vectors or so-called visual words. Often, the original descriptors are compressed or approximated first $x^i \rightarrow \hat{x}^i$. There is a multitude of methods for encoding descriptors in the visual word space, including VQ (hard)-assignment [143], source coding with refinement [140] or soft-assignment [144, 145]. All these representations allow for fast $\epsilon$-NN or $k$-NN search.

However, to reveal the theoretical limits of the identification systems based on BOW, we will assume that the descriptors are uncompressed or very accurately approximated [155].
For these reasons, we will consider the equivalent model shown in Figure 4.2 consisting of enrollment and identification via the equivalent codebook \( C_x = (x^1, \cdots, x^J)^T \in \mathbb{R}^{J \times L} \), where \( J \) is the number of codewords in the visual codebook. This equivalent codebook contains all unique descriptors, the composition of which gives a particular image \( x(m) \) representation. It is noteworthy that the representation of each image in terms of the equivalent codebook with an appropriate indexing structure in the form of inverted files makes it possible to obtain an efficient search [140].

In this analysis, we will assume that each image \( x(m), 1 \leq m \leq M \), is represented by \( J_x(m) \) descriptors \( x^i(m), 1 \leq i \leq J_x(m) \) as shown in Figure 4.2. The visual codebook \( C_x \) contains all these descriptors labeled as \( x^j, 1 \leq j \leq J \). In this case, the size of visual codebook \( J = |C_x| = \sum_{m=1}^{M} \sum_{i=1}^{J_x(m)} \). Note that in this part we explicitly assume that the architecture does not deploy any clustering to build up the visual codebook nor does it use descriptor compression, as to reveal the theoretical performance limit.

At the end of the enrollment stage (Figure 4.2), the encoder produces a database that is organized in the form of an inverted index file, i.e., each codeword with the index \( j \in \{1, \cdots, J\} \) contains a list of images \( m \in \{1, \cdots, M\} \) containing this visual word. At the identification stage shown in Figure 4.2, the query image \( y \), represented by its \( J_y \) descriptors, is presented to the identification system. The list decoder seeks for all \( \epsilon \)-NN
or \( k \)-NN codewords \( x^j, j \in \{1, \cdots, J\} \) in the codebook \( C_x \) thus producing \( J_y \) lists \( \mathcal{L}(y^k) \) for all query descriptors. The joint decoder observes these lists and tests the database for the corresponding image indices containing the identified features and makes a final decision in favor of an index \( \hat{m} \) possessing the largest number of matched codewords.

### 4.3 Statistical model of BOW content identification

The statistical model of BOW content identification includes the definition of:

(a) The statistics of descriptors \( x^i \).

(b) The statistical observation model \( p(y^k|x^i) \).

(c) A model for the encoding.

(d) A decision model to deal with the absence or presence of descriptors and their geometric consistency verification.

(e) A final model for the global identification decision.

A schematic of the entire model can be seen in Figures 4.2 and 4.4.

**Database of descriptors**

In this chapter, we will assume that the descriptors \( x^i \in X^L \) are i.i.d. much like Oriented FAST and Rotated BRIEF (ORB), [89] following some distribution \( X^i \sim p(x^i) = \prod_{n=1}^{L} p(x_n^i) \). Obviously, one can consider different descriptors: local or global, sparse or dense and binary. The i.i.d. assumption is not valid for SIFT descriptors which manifests a high correlation between elements [37]. See Appendix B for details.

**Statistical observation model**

The statistical observation model for the entire image is expressed in terms of the statistical model for the local descriptors:

\[
p(y|x(m)) \equiv \prod_{k=1}^{J_y} \prod_{i=1}^{J_x(m)} p(y^k|x^i(m)), \tag{4.1}
\]

which reduces to \( p(y|x(m)) \equiv \prod_{k=1}^{J'} p(y^k|x^k(m)) \) in the synchronised case, i.e., the exact correspondence between the descriptors is known with \( J' = \min\{J_x(m), J_y\} \).
The above probabilistic model can be also mapped into some metric space via
\[
p(y^k|x^i(m)) \propto e^{-\frac{d(y^k, x^i(m))}{2}}
\]
assuming an exponential family of distortions, where \(d(y^k, x^i(m))\) represents
a distance between two descriptors and 2 is a scaling factor.

**Model of encoding & assignment**

We will consider hard assignment to investigate system performance under the minimum
requested memory storage requirements \([140, 156]\) \(^1\).

The encoding matrix can be generally constructed as \(C_x(m) = (c^1_x(m), \ldots, c^{J_x}(m)) \in \mathbb{R}^{J_x \times J_x(m)}\), where each column \(c^i_x(m)\) stands for the code representing the encoding of the
descriptor \(x^i(m)\), \(1 \leq i \leq J_x(m)\) with respect to the visual codebook \(C_x\). In the case of
hard assignment, \(C_x(m) \in \{0, 1\}^{J_x \times J_x(m)}\) with the elements \(c^j_x(m) = 1\) for \(j : x^j = x^i(m)\)
or zero-distance, i.e., \(j \in L(x^i(m))\) with the list:

\[
L(x^i(m)) = \{j \in \{1, \ldots, J\} : d(x^j, x^i(m)) = 0\}. 
\]  (4.2)

The encoding process based on the pooling is shown in Figure 4.3.

\[
C_x(m)=
\begin{bmatrix}
1 & i & J_x(m) \\
1 & 1 & J_x(m) \\
1 & 1 & 1 \\
J & 1 & 1 \\
\end{bmatrix}
\rightarrow \text{Pooling} \rightarrow \text{Enrolled BOF vector}
\]

\[
C_y =
\begin{bmatrix}
1 & k & J_y \\
1 & 1 & J_y \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
\end{bmatrix}
\rightarrow \text{Pooling} \rightarrow \text{Query BOF vector}
\]

Figure 4.3 – The block diagram of encoding or pooling at the enrollment and
identification to produce a fixed-length representation of length \(J\).

Given the case that the descriptors are matched without geometrical consistency, i.e.,
they are desynchronised, and there are generally a different number of descriptors in

\(^1\) The hard or soft assignment represents a trade-off between memory storage and decoding complexity.
the enrolled image $J_x(m)$ and query $J_y$, pooling is used. To address this, there are two common types of pooling: average- and max- pooling. In the case of hard assignment at the enrollment stage they are equivalent. The enrolled fixed-length sparse code for the image $m$ is $d_x(m) = (d^1_x(m), \ldots, d^J_x(m))^T$ which is obtained as:

$$d^av_x(m) = \sum_{i=1}^{J_x(m)} c^i_{x_j}(m), \quad (4.3)$$

in average-pooling and:

$$d^{max}_x(m) = \max_{1 \leq i \leq J_x(m)} c^i_{x_j}(m), \quad (4.4)$$

in max pooling.

**Descriptor presence model and geometric consistency verification**

Given a query $y$ consisting of $J_y$ descriptors, the encoding matrix for the query is defined as $C_y = (c^1_y, \ldots, c^{J_y}_y) \in \{0,1\}^{J \times J_y}$, with $c^j_y = 1$ for $j \in \mathcal{L}(y^k)$ with the list $\mathcal{L}(y^k) = \{ j \in \{1, \ldots, J\} : d(x^j, y^k) \leq \epsilon_L \}$. This decoder corresponds to the BDD or $\epsilon$-NN decoder which seeks all \{x^j\} NNs in the radius $\epsilon L$ from the query descriptor $y^k$.

The performance of the descriptors is measured in terms of their ROCs defined by the probabilities of miss $P^D_M = \Pr\{d(x^i, Y^k) \geq \epsilon L \}$ and probability of false acceptance $P^D_F = \Pr\{d(x^i, Y^k) < \epsilon L \}$, where $\epsilon$ is the threshold.

The corresponding pooled fixed-length vectors are defined for average pooling as:

$$d^{av}_y = \sum_{k=1}^{J_y} c^k_{y_j}, \quad (4.5)$$

and for max pooling as:

$$d^{max}_y = \max_{1 \leq k \leq J_y} c^k_{y_j}. \quad (4.6)$$

In following, we will only consider max pooling due to its reported superior performance [144].
The statistics of matrix $C_y$ are completely defined by the probabilities of descriptor miss $P_M$ and false acceptance $P_F$ as defined above.

**Final decoding model**

The final decision is based on a list decoder that produces a list of possible candidates:

$$\mathcal{L}(y) = \{m \in \{1, \cdots, M\} : t(m) \geq \tau J\}, \quad (4.7)$$

where $t(m) = d_x^T(m)d_y$ stands for the similarity score between two vectors which, for example, can be the cosine distance, that is often used in BOW systems, if the vectors are normalized by their norms $||d_x(m)||$ and $||d_y||$.

**Remark:** In the case when the correspondence between the descriptors from two images is established, one can estimate the upper bound of the system’s performance by evaluating the similarity between two matrices as $t(m) = C_x(m) \odot C_y$, where $\odot$ denotes the Frobenius inner product.\(^2\)

---

\(^2\) In the synchronized case, the matrices are of the same size.
Detect some local feature set

Extract feature descriptor vectors

\[ x_1^{(1)} \]
\[ \ldots \]
\[ x_J^{(1)} \]
\[ \vdots \]
\[ x_1^{(2)} \]
\[ \ldots \]
\[ x_J^{(2)} \]

Assignment followed by pooling per centroid, per set of descriptors from a single image

\[ d_x^{(1)} = (d_{x_1}^{(1)}(1), \ldots, d_{x_J}^{(1)}(1)) \]
\[ d_x^{(2)} = (d_{x_1}^{(2)}(1), \ldots, d_{x_J}^{(2)}(2)) \]

Assignment against enrollment centroids

\[ y_1 \]
\[ \ldots \]
\[ y_J \]
\[ \vdots \]

Assignment followed by Pooling per centroid, per set of descriptors from a single image

\[ d_y^{(1)} = (d_{y_1}^{(1)}(1), \ldots, d_{y_J}^{(1)}(1)) \]

Final Query BOW vector

\[ d_y^{(1)} \]
\[ \ldots \]
\[ d_y^{(2)} \]
\[ \vdots \]

Final Query BOW vector

\[ d_y^{(1)} \]
\[ \ldots \]
\[ d_y^{(2)} \]

Assignment against enrollment centroids

\[ y_1 \]
\[ \ldots \]
\[ y_J \]
\[ \vdots \]

Detection/Local Feature Set

\[ x(1) \]
\[ x(2) \]

Figure 4.4 – Fundamental building blocks of a BOW based identification system.
4.4 Performance analysis: Uncompressed features

The overall system performance is evaluated by the probability of miss \( P_M \), i.e., the correct \( m \) does not appear on the decoder’s list under the hypothesis \( \mathcal{H}_m \), \( P_M = \text{Pr}\{T(m) \leq \tau J_e|\mathcal{H}_m}\) and by the probability of false acceptance \( P_F \), i.e., an incorrect \( m' \) appears on the decoder’s list under the hypothesis \( \mathcal{H}_{m'} \), \( P_F = \text{Pr}\{T(m) > \tau J_e|\mathcal{H}_{m'}\} \), where \( \tau \) is the threshold and \( J_e \) stands for the equivalent length under different pooling strategies. The average list size can be estimated as \( \mathbb{E}\{|L(y)|\} = MP_F \). In the case of unique identification, the list size is 1.

Without loss of generality, we will assume that the same number of descriptors is enrolled for all images, i.e., \( J_x(m) = J_x \), which is a reasonable assumption for most of the identification systems where the enrollment is under the control.

The sufficient statistic in the case of max pooling and perfect synchronisation is:

\[
T(m) \sim \begin{cases} 
\mathcal{B}(J_e, \theta(m)), & \text{for } \mathcal{H}_m, \\
\mathcal{B}(J_e, \theta(m')), & \text{for } \mathcal{H}_{m'}, 
\end{cases}
\]  

where \( J_e = \min\{J_x, J_y\} \) and for max pooling: \( \theta(m) = 1 - (1 - P_D^D)(1 - P_F^D)^{J_y-1} \) and \( \theta(m') = 1 - (1 - P_F^D)^{J_y} \) and for the perfectly synchronised case: \( \theta(m) = P_D^D \) and \( \theta(m') = P_F^D \).

Proof. In this proof we derive the distribution of parameter \( T(m) \) under the hypotheses \( \mathcal{H}(m) \) and \( \mathcal{H}(m') \) for the max pooling. The equation (4.8) can be rewritten as:

\[
t(m) = d_x^T(m)d_y = \sum_{j=1}^{J_x} d_{x_j}(m)d_{y_j} = \sum_{j'=1}^{J_y} d_{y_{j'}},
\]

where \( J_e = \min\{J_x, J_y\} \). The parameter \( T(m) \) represents a sum of \( J_e \) i.i.d. Bernoulli random variables \( D_{y_{j'}} \) with the probability of success \( \text{Pr}\{D_y = 1|\mathcal{H}(m')\} \) under the hypothesis \( \mathcal{H}(m') \). Max pooling makes a positive decision on the presence of the descriptor, if at least one positive decision out of \( J_y \) is observed, which is defined as:
\[
\theta(m') = \Pr\{D_y = 1 | \mathcal{H}(m')\} = \Pr\{\bigcup_{k=1}^{J_y} E_F\} \\
= 1 - \Pr\{\bigcap_{k=1}^{J_y} \bar{E}_F\} \\
= 1 - \prod_{k=1}^{J_y} (1 - P_D) \\
= 1 - (1 - P_D)^{J_y}, \tag{4.10}
\]

where \( E_F \) denotes an event consisting in a false acceptance and \( \bar{E}_F \) denotes non-\( E_F \), i.e., the correct rejection.

Similarly, under the hypothesis \( \mathcal{H}(m) \), max pooling will make a positive decision, if there is at least one match originating from the true descriptor or from \((J_y - 1)\) falsely matched descriptors:

\[
\theta(m) = \Pr\{D_y = 1 | \mathcal{H}(m)\} = \Pr\{E_c \cup \bigcup_{k=1}^{J_y-1} E_F\} \\
= 1 - \Pr\{E_c \cap \bigcap_{k=1}^{J_y-1} \bar{E}_F\} \\
= 1 - (1 - P_D^D)(1 - P_D^D)^{J_y-1}, \tag{4.11}
\]

where \( E_c \) and \( \bar{E}_c \) denote the correct match with probability \( P_D^D \) and miss (complement to the correct match) with probability \((1 - P_D^D)\), respectively.

The summation of \( J_e \) independent Bernoulli random variables in (4.9) results in a Binomial random variable with the distributions \( T(m) \sim \mathcal{B}(J_e, \theta(m')) \) under \( \mathcal{H}_{m'} \) and \( T(m) \sim \mathcal{B}(J_e, \theta(m)) \) under \( \mathcal{H}_m \).

In a similar fashion, one can obtain the distortion of parameter \( T(m) \) for the synchronised case. The only difference consists in the fact that there is no summation over all \( J_y \). This results in \( \theta(m') = P_D^D \) and \( \theta(m) = P_D^D \).

\[\square\]

The performance of the content identification system is estimated based on a list decoder which is characterized by the probability of miss:
\[ P_M = \Pr\{T(m) \leq \tau J_e | H_m \} \]
\[ = \sum_{d=0}^{\tau J_e} \binom{J_e}{d} \theta^d(m)(1 - \theta(m))^{J_e - d} \]
\[ \leq 2^{-J_e D(\tau || \theta(m))}, \quad (4.12) \]

where \( D(\tau || \theta(m)) \) denotes the divergence and the probability of false acceptance is:

\[ P_F = \Pr\{T(m) > \tau J_e | H_m' \} \]
\[ = \sum_{d=\tau J_e}^{J_e} \binom{J_e}{d} \theta^d(m')(1 - \theta(m'))^{J_e - d} \]
\[ \leq 2^{-J_e D(\tau || \theta(m'))}, \quad (4.13) \]

which results into the average list of candidates \( \mathbb{E}\{|L(Y)|\} = MP_F \). The threshold should satisfy \( 0 \leq \theta(m') < \tau < \theta(m) \leq 1 \).

Using the notion of the identification rate as
\[ R = \frac{1}{J_e} \log_2 M \]
defined for large \( J_e \), one can target the condition \( R \leq D(\tau || \theta(m')) \) to keep the list of retrieved candidates small.

In the case of average-pooling, the resulting vectors consist of the sum of binomial random vectors with different lengths and probabilities. To present the results in a tractable form, we use a Gaussian approximation which results in:

\[ T(m) \sim \begin{cases} 
N(\mu(m), \sigma^2(m)), & \text{for } H_m, \\
N(\mu(m'), \sigma^2(m')), & \text{for } H_{m'}, 
\end{cases} \quad (4.14) \]

where \( \mu(m) = J_e(P^D_F + (J_y-1)P^D_P) \) and \( \mu(m') = J_eJ_yP^D_P \), \( \sigma^2(m) = J_e(P^D_D(1 - P^D_F) + (J_y-1)P^D_P(1 - P^D_P)) \) and \( \sigma^2(m') = J_eJ_yP^D_P(1 - P^D_P) \) with \( J_e = \min\{J_x, J_y\} \) and a new threshold \( J_t = J_eJ_y \).

**Proof.** In this proof, we will consider the statistics of \( T(m) \) under average pooling. The main difference with max pooling consists in the averaging of all positive matches in the variable \( d_{y,j'} \) in (4.9) rather than just indicating the positive success in any of the \( J_y \) outcomes.

Under the hypothesis \( H_{m'} \), when only false matches can lead to the success event, there is a sum of \( J_e \) Bernoulli random variables with success probability \( P^D_P \). Therefore, the summation of \( J_e \) independent Binomial random variables \( D_{y,j'} \sim B(J_y, P^D_P) \) in (4.9)
Chapter 4. BOW based Content Identification

results in a Binomial random variable with the distribution \( T(m) \sim B(J_e J_y, P_D^D) \) under \( \mathcal{H}_{m'} \).

Under the hypothesis \( \mathcal{H}_m \), the random variable \( D_yj' \) in (4.9) contains one correct match event with probability \( P_D^D \) and \( (J_y - 1) \) matches with probability \( P_D^F \). Unfortunately, it is not tractable to present the resulting distribution containing a sum of Binomial random variables with different dimensionalities and probabilities. Therefore, we will use the approximation of the sum of independent random variables via the Gaussian distribution according to the central limit theorem resulting in \( T(m) \sim N(\mu(m), \sigma^2(m)) \), where \( \mu(m) = J_e(P_D^D + (J_y - 1)P_D^F) \) and \( \sigma^2(m) = J_e(P_D^D(1 - P_D^D) + (J_y - 1)P_D^F(1 - P_D^F)) \).

Similarly, one can approximate the Binomial distribution under the distribution \( \mathcal{H}_{m'} \) as \( T(m) \sim N(\mu(m'), \sigma^2(m')) \) where \( \mu(m') = J_e J_y P_D^D \) and \( \sigma^2(m') = J_e J_y P_D^D(1 - P_D^D) \).

Using a Gaussian approximation of Hamming distances, assuming that the distances contribute to non-negative values only, the corresponding performance under average-pooling is:

\[
P_M = \Pr\{T(m) \leq \tau J_l | \mathcal{H}_m\} = \int_0^{\tau J_l} \frac{1}{\sqrt{2\pi \sigma^2(m)}} e^{-\frac{(\mu(m) - \tau J_l)^2}{2\sigma^2(m)}} \, dt = Q \left( \frac{\mu(m) - \tau J_l}{\sigma(m)} \right), \tag{4.15}
\]

\[
P_F = \Pr\{T(m) > \tau J_l | \mathcal{H}_{m'}\} = \int_{\tau J_l}^{+\infty} \frac{1}{\sqrt{2\pi \sigma^2(m')}} e^{-\frac{(\mu(m') - \tau J_l)^2}{2\sigma^2(m')}} \, dt = Q \left( \frac{\tau J_l - \mu(m')}{\sigma(m')} \right), \tag{4.16}
\]

with the average list of candidates \( \mathbb{E}\{|\mathcal{L}(Y)|\} = MP_F \). In some applications, it is interesting to keep both probabilities of errors small. In this case, one can follow the strategy to minimise the maximum probability of error under optimal \( \tau \) and \( \epsilon \) defined as \( (\hat{\tau}, \hat{\epsilon}) = \arg\min_{\tau, \epsilon} \max\{P_M(\tau, \epsilon), P_F(\tau, \epsilon)\} \). We will investigate this problem numerically.
Chapter 4. *BOW based Content Identification* 131

For technical reasons we will first fix $\epsilon$ and estimate $\tau$. In the case of max pooling and perfect synchronisation, the above maximisation is achieved when $P_M(\tau, \epsilon) = P_F(\tau, \epsilon)$. The equality of (4.12) and (4.13) leads to the equality $D(\tau\|\theta(m)) = D(\tau\|\theta(m'))$ that yields:

$$\hat{\tau} = \frac{\log \frac{1-\theta(m')}{1-\theta(m)}}{\log \frac{\theta(m')(1-\theta(m'))}{\theta(m)(1-\theta(m))}}.$$  (4.17)

From the equality of the Q-function arguments in the equations (4.15) and (4.16), one can obtain for average-pooling:

$$\hat{\tau} = \frac{1}{J_t} \frac{\sigma(m')\mu(m) + \sigma(m)\mu(m')}{\sigma(m') + \sigma(m)}.  \quad (4.18)$$

### 4.5 Performance analysis: Compressed features

Visual codebook compression has three main objectives:

(a) To enable efficient search via indexing based on inverted files.
(b) To minimize satisfaction of memory storage requirements.
(c) Enable visual codebook learning based on the available training data.$^3$

The different BOW architectures can be generalised as most share a common hierarchical design of their visual codebook where $\epsilon$-NN descriptors are clustered and represented by a common centroid. Then the query is first tested against all centroids and the nearest centroids are retrieved. The descriptors belonging to the found clusters represented by these centroids are explored to find the $\epsilon$-NN or $k$-NN descriptors.

**Impact of compression**

Disregarding a particular assignment technique, i.e., either hard- or soft-assignment, or source coding with refinement, the compression introduces an average descriptor distortion $D_c = \mathbb{E}\{d(X^j, \hat{X}^j)\}$, where $\hat{X}^j$ stands for the compressed version of $j$th descriptor $X^j$ and $d(,,)$ denotes the Euclidean or Hamming distance.

---

3. The last concerns mostly CBIR systems. In content identification applications, the codebook design can be optimised for all available enrolled samples.
The statistical model of compression is considered as a mapping \( p(\hat{x}^j|x^j) \). In a high-rate compression regime, i.e., regime of small distortions, the additive model can be used \( x^j = \hat{x}^j + z^j \), where \( z^j \) is the compression noise independent of \( \hat{x}^j \) [157]. In the case of binary descriptors such as ORB, the distortion \( D \) can be also interpreted as 
\[
\Pr\{X^j_i \neq \hat{X}^j_i\} \leq D
\]
where \( X^j_i \in \{0, 1\} \). If the binary descriptor is considered as a Bernoulli source with \( \Pr\{X^j_i = 1\} = p \), then from rate-distortion theory it is known that 
\[
D(R_D) = H_2^{-1}(H_2(p) - R_D(D)), \quad 0 \leq D \leq \min\{p, 1-p\}
\]
where \( R_D \) is the rate of the descriptor compression, and \( H_2(.) \) denotes the binary entropy [157]. For example, if the 256-bit-ORB descriptor with \( p = 0.5 \) is compressed to 64 bits, i.e., with the rate \( R_D = 0.25 \) bit/sample, the distortion is \( D(R_D) = 0.2145 \). Practically, the compression of the descriptor leads to extra degradation in the chain \( Y^j - X^j - \hat{X}^j \), when the compressed descriptors are stored instead of the original ones. It results in the degradation of the ROC for the descriptors. Mostly it concerns the degradation for the geometrically consistent descriptors with 
\[
\Pr\{Y^j_i \neq \hat{X}^j_i\} = D \ast P_b, \quad \text{where } P_b \text{ is the probability of bit error, and for the geometrically inconsistent descriptors } \Pr\{Y^k_i \neq \hat{X}^j_i\} = D \ast 0.5 = 0.5
\]
where \( D \ast P_b = D(1 - P_b) + P_b(1 - D) \). It means that the probability mass function for the geometrically consistent descriptors is closer to one of geometrically inconsistent ones and it has higher variance. Altogether, it leads to an increase of \( P_D^M \) and \( P_D^F \).

The same arguments are valid for the SIFT descriptors. We will consider a particular method of their compression based on the product-of-vector quantizer [140] and investigate the impact of compression on the ROC curves in Section 4.6.

**Impact on indexing structure**

The indexing structure of modern BOW systems is mostly based on hierarchical clustering. The generalisation of different indexing structures addressing the memory-complexity trade-off with the rate \( R \) close to the identification capacity \( I(X^j; Y^j) \) is given in [156]. This study covered the case of unique decoding or identification where each object is represented by a single descriptor in the codebook. In contrast, in BOW systems the rate \( R > I(X^j; Y^j) \) and consequently the system is not able to produce a unique decision. Instead, as considered in the previous Section, the list of the most likely candidates is retrieved based on list decoding. Moreover, the descriptors are additionally compressed in the BOW systems in contrast to the identification setup considered in [156] which reduces the identification rate \( I(\hat{X}^j; Y^j) \leq I(X^j; Y^j) \). Therefore, this analysis is not directly applicable.

Summarizing different methods of clustering based search, we can categories them depending on:
(a) The method of generation of the centroids, i.e., $p(c^j|x^j)$, $p(c^j|y^j)$ or $p(c^j|x^j,y^j)$.
(b) The decoding method, i.e., unique or list decoding
(c) The covering principle.

According to [156], all these methods achieve identification capacity, but generally the centroid generation based on $p(c^j|x^j,y^j)$ with list encoding shows the best memory-complexity trade-off. Note that the main memory burden comes from the storage of uncompressed descriptors. Therefore, the optimal compression of descriptors is crucial.

In this chapter, we only considered hard assignment with list decoding. Therefore, in our model the centroids are assumed to be generated as $p(c^j|x^j)$ without covering and list decoding similarly to [140]. The descriptors can be stored in their original form or compressed. This is reflected by the corresponding $P^D_M$ and $P^D_F$.

In conclusion, besides the obvious advantage of more efficient memory storage, the compression of descriptors has several important negative implications. First, the compression of descriptors leads to the increase of $P^D_M$ and $P^D_F$ which degrades overall system performance in terms of $P_M$ and $P_F$. Secondly, the compression of descriptors extends the search region for the $\epsilon$-NN centroid thus leading to an increased search complexity. Finally, it reduces the entropy of the descriptors and thus increases the potential for attacks.

### 4.6 Simulation

Since the overall system performance is determined by the statistics of $T(m)$ which is in turn defined by $P^D_M$ and $P^D_F$, we first investigated the typical ROC curves for SIFT and ORB descriptors shown in Figure 4.5 for the *copydays* database [126]. We have tested about 100’000 descriptors. As expected, SIFT produces better results, it is, however, relatively slow.

The experimental distributions of parameter $T(m)$ for the matched and non-matched pairs of descriptors are shown in Figures 4.5a and 4.5b for the ORB and SIFT descriptors, respectively. In addition, to investigate the impact of descriptor compression, we tested the product-of-vector quantizer proposed in [140] with a block size of 8 and 256 centroids in each block giving 64 bits in total per descriptor. The inter- and intra-class pdfs for the symmetric case, i.e., both the query and enrolled descriptors are quantized, is referred to as *quantized SIFT*, and the asymmetric case, i.e., the query descriptor is soft while the enrolled descriptor is quantized, is referred to as *soft SIFT*, and is shown in Figure 4.5b in comparison with the original SIFT pdfs. The ROCs for the ORB and SIFT descriptors are shown in Figure 4.5c. The results are obtained for 100’000
matched pairs of descriptors. It should be remarked that the original SIFT descriptors demonstrate significantly better performance in comparison to ORB: at least in 2 orders of magnitude in terms of the $P_{DF}$. However, ORB was about from 5 to 8 times faster in our experiments. The product-of-vector quantizer has demonstrated remarkable performance. In particular, very good results were obtained for the quantized SIFT with 64 bits which outperforms 256 bit ORB. However, as it was mentioned, the computational burden of SIFT is an essential bottleneck in practical on-line applications. Therefore, there is a need for binary, fast and performant features which generate descriptors directly without any compression.

To validate the accuracy of the developed mathematical model, we experimentally tested the distribution of $T(m)$ in (4.8) and (4.14), for the average pooling, max pooling and the synchronised case using the $P_{DM}$ and $P_{DF}$ for the ORB descriptors. It should noted that it follows expectation that SIFT provides better results. However, our objective was to validate and exemplify the performance of practical systems that can be used in on-line applications such as on mobile phones. Therefore, we focused on the ORB descriptors leaving the comparison of different descriptors out of scope. That is why our objective was to, given the ROC curve of any descriptor, predict the performance of the BOW identification system.

Accordingly, the main hypotheses we wanted to test here were:

1. To confirm that max pooling is superior to average pooling as is commonly believed.
2. To investigate the gap between pooling strategies that operate under the geometrical ambiguity with those of the perfectly synchronized case.

For purely demonstrative purposes, we have chosen the operational point $P_{DM} = 0.2954$ and $P_{DF} = 0.0101$ on the ORB ROC curve in Figure 4.5c and investigate the sufficient statistic $T(m)$ under $\mathcal{H}_m$ and $\mathcal{H}_m'$ as shown in Figure 4.6. The theoretical counterparts follow the experimental results remarkably well. The results were achieved in 10’000 experiments for each type of pooling. The number of descriptors in the enrolled images and query was identical and equals 500.

To obtain the objective performance of BOW systems, we investigated the ROC curves for the above operational point. In addition, it is interesting to highlight the impact of the different relationship between $J_x$ and $J_y$ on the overall system performance, which is why we considered four cases:

(a) $J_x = J_y = 50$ to simulate some practical situations.
(b) $J_x = J_y = 500$ to investigate the impact of an increased number of descriptors.
(c) $J_x = 500$ and $J_y = 50$ to study the impact of cropping in the query.
(d) $J_x = 50$ and $J_y = 500$ to highlight the impact of collage or query acquisition with some background interfering objects or features.

The results of this study are shown in Figure 4.7 and leads to the following conclusions.

First, the overall performance of max pooling and average pooling for this operational point is comparative (Figure 4.7a).

Second, as expected the synchronised case represents the lower bound in all tests and there is a considerable gap with average and max pooling. This signifies the importance of geometrical synchronisation. We believe that the BOW methods should strongly benefit from local geometrical consistency verification at the early stages of descriptor matching in contrast to the existing architecture presented in Figure 4.1, where geometrical verification is performed in the last stage.

Third, the increase of descriptors from $J_x = J_y = 50$ to $J_x = J_y = 500$ drastically increases the performances of the synchronised case (Figure 4.7b). This is in accordance with detection and information theory. The increase of the number of descriptors in the synchronised case is equivalent to the creation of one super descriptor, or vector, with a total length of 500 × 256 bits in the ORB case. In contrast, the performance of average pooling and max pooling drops. This is due to the fact that the BOW system generates a lot of false matches for the non-synchronised descriptors and the increase of their number leads to the masking of and interference with, correct matches. In addition, the drop in performance for max pooling is higher. This is due to the fact that average pooling generates a sort of soft information about the number of matches in the considered model while max pooling represents a binary decision rule that just declares the presence or absence of at least one match.

Forth, the asymmetric case of $J_x = 500$ and $J_y = 50$, that represents a sort of cropping, does not show any difference with the case of $J_x = J_y = 50$ (Figure 4.7c).

Fifth, the asymmetric case of $J_x = 50$ and $J_y = 500$ represents a case with many false descriptors in the query image (Figure 4.7d). The synchronised case shows good immunity to that. However, average pooling and max pooling manifest a drastic drop in their performance. This is due to the overwhelming amount of false matching. That is why the sole matching of descriptors is not sufficient without additional information about their orientations or geometrical positions.

Finally, the overall performance of BOW content identification system is summarised in terms of the min max $\{P_M(\tau, \epsilon), P_F(\tau, \epsilon)\}$ as a function of $\epsilon$ and $\tau$ in Figure 4.8. It is interesting to note that the system has a global minimum for the optimal pair of thresholds $\tau, \epsilon$. 

We investigated the optimal pair of $\tau, \epsilon$ leading to the minimisation of the maxmin error (Figure 4.8a). It should be noted that the experimental thresholds are very accurately predicted by the derived formulas (4.17) and (4.18). Therefore, the optimisation of BOW systems is straightforward. It is also remarkable that in our model max pooling did not demonstrate any advantage over average pooling, contrarily to common belief. Similar performance to max pooling was observed under the proper $\epsilon$ and $\tau$. Interestingly, both methods have the same global minimum for the same $\epsilon=0.25$ (Figure 4.8b). However, the optimal $\tau$s are different for max pooling and average pooling (Figure 4.8a). Therefore, if one specifies the descriptor, our model suggests a set of optimal parameters under average and max pooling to optimize overall system performance.
Figure 4.5 – Typical performance of descriptors for matched and non-matched descriptors under scaling 0.5, rotation $10^0$ and JPEG 75: (4.5a) histograms for ORB (uncompressed-uncompressed) descriptors, (4.5b) histograms for SIFT (uncompressed-uncompressed, uncompressed-compressed, compressed-compressed) and (4.5c) ROCs for ORB and SIFT descriptors.
Figure 4.6 – Distributions of similarity score for the ORB descriptor computed experimentally with the corresponding theoretical predictions: (4.6a) average pooling, (4.6b) max pooling and (4.6c) synchronised case.
Figure 4.7 – Impact of the number of descriptors $J_x$ and $J_y$ on the ROC for the ORD descriptor for the descriptor operational point determined by $P_M^D = 0.2954$ and $P_F^D = 0.0101$ for average-, max pooling and synchronised case: (4.7a) $J_x = J_y = 50$, (4.7b) $J_x = J_y = 500$, (4.7c) $J_x = 500$, $J_y = 50$ and (4.7d) $J_x = 50$, $J_y = 500$. 
Figure 4.8 – The optimal set of thresholds $\epsilon$ and $\tau$ for the ORB descriptor computed according to $(\hat{\tau}, \hat{\epsilon}) = \arg\min_{\tau, \epsilon} \max\{P_M(\tau, \epsilon), P_F(\tau, \epsilon)\}$ for the average pooling, max pooling and synchronised cases: (4.8a) the relationship between the thresholds and (4.8b) the achievable error in the function of $\epsilon$ for the optimal $\tau$. Remarkably, the average pooling and max pooling achieve the same error for optimised parameters.
4.7 Application: Privacy preserving content identification

This section will show case a practical application, that allows for privacy preserving search based on the introduced statistical model for BOW systems. Privacy preserving content identification and nearest neighbor search are active fields of research in numerous medical applications, privacy-sensitive multimedia collections and biometrics. Furthermore, there is a huge trend in outsourcing storage and services to cloud-based systems such as Amazon AWS. Sensitive data may obviously be encrypted but this step significantly complicates basic services such as the ability to search. Although possible in principle, cryptographic primitives, such as homomorphic methods, that enable basic search or fuzzy similarity matching, are complicated, computational heavy for all parties and scale poorly on large databases [158].

As an example, a practical scenario involving these issues is a (medical) researcher who wishes to find similar images from external sources where neither parties want to disclose all data to each other or to the server.

Our system model, seen in Fig. 4.9, includes three principle parties. The data owners that provide the original content and render services from an external provider or simply, the server. The data users wish to search through the collection. Following biometric primitives [159], neither original data is stored in the cloud, nor does the user send (biometric) templates as query.

We model the main privacy threats in our content identification system as follows:

(a) The reconstruction of the original query image from the derived query features which are actually sent.

(b) The reconstruction of database entries from the stored features on the server.

(c) The ability of the server to learn what a particular user is searching for.

This particular application is based on recent results [160, 161] demonstrating that local descriptors along with their positions in the originating images can be used successfully to reconstruct visually similar and recognizable images. The original geometric information is however vital for this attack to succeed.

The basic principle of this framework is to split the raw local descriptors into two parts where the payload of the descriptor $x_d$ can be shared in the public domain and the originating geometrical positions $x_g$ are kept private and shared only between authorized parties as shown in Fig. 4.10. This means that the statical BOW model, introduced in this chapter can be used to deduce all relevant aspects of this particular applications.

Specifically, the following aspects will be analyzed:
Chapter 4. BOW based Content Identification

Data owner

<table>
<thead>
<tr>
<th>Images</th>
<th>Feature extraction</th>
</tr>
</thead>
</table>

Data owner

<table>
<thead>
<tr>
<th>Images</th>
<th>Feature extraction</th>
</tr>
</thead>
</table>

Data owner

<table>
<thead>
<tr>
<th>Images</th>
<th>Feature extraction</th>
</tr>
</thead>
</table>

Data user

<table>
<thead>
<tr>
<th>Image query</th>
<th>Feature extraction</th>
</tr>
</thead>
</table>

| Meta data |

Figure 4.9 – The diagram of privacy preserving image search in a cloud system.

- What is the server ambiguity, i.e., cardinality $|\mathcal{L}(y_d)|$?

- How does geometrical information $y_g$ enhance accuracy and reduce the ambiguity of the data user, i.e., cardinality $|\mathcal{L}(y_d, y_g)|$?

- Is it possible to efficiently ambiguize a query in terms of an enlargement of $|\mathcal{L}(y_{dr})|$ but without significant loss of performance and an increase of complexity for the data user?

4.7.1 Existing privacy preserving approaches

One can distinguish several main approaches that address the privacy-preservation problem:

- Identification based on homomorphic encryption [162, 163].

- Secure embedding or fingerprinting [164, 165, 166, 150, 149].

- Attribute based encryption [167, 168].

Homomorphic methods allow computations between vectors in the encrypted domain, ensuring that the server neither sees the original data nor the query. Such methods are however computational intensive, the search for similar items is exhaustive and cipher texts are larger than the original data vectors causing a significant bandwidth and storage burden [162, 163].

Secure embedding is based on the mapping and encoding of images or features into some space. The basis vectors for this space can be learned or generated at random. The image features are represented in this space and then assigned or quantized. The trade-off between the accuracy of the computation in this space, the amount of information that has to be shared with the server and its ability to reconstruct the query determines the privacy protection.
Algorithms based on random basis vectors produce a quantized version, which may be over complete. The design of privacy preserving quantizers is considered in [166]. The quantization of the $L$ most reliable components is proposed in [165] with simultaneous randomization of the weak components. Methods for general randomization with low-search complexity, also known as bounded distance decoding, are addressed in [164].

Methods based on learned basis systems or codebooks in the BOW context deploy privacy preserving methods by limiting or protecting the information needed to reconstruct the used codebook [150, 149].

The main advantage of embedding or fingerprinting methods is that they allow for efficient searching routines based on Hamming distances. The inherent information loss harms performance but simultaneously ensures that full recovery of the query or database is not feasible. Formally, in the identification case, when the number of images $M \sim 2^{NC}$, where $N$ is the dimensionality of the used feature in Euclidean space and $C$ stands for the identification capacity [159], i.e. the ability to error-less identify a sought item, the condition of the Johnson-Lindenstrauss lemma, when $\mathbb{R}^N \to \mathbb{R}^L$, with $L > \frac{8 \ln M}{\epsilon^2}$, $0 < \epsilon < 1$, is not satisfied due to the exponential behaviour of $M$ with $N$. This loss will be significant in many practical situation. Therefore, the above methods can only be deployed for privacy preserving measures, when the embedding codebook is small or when list decoding is used to retrieve a list with possible matching candidates.

Finally, attribute coding applies encryption to stored content based on a secret or attribute which is extracted from that same content. When the query is in proximity to the database entry, the attribute is also close and the decryption, potentially with errors, might be performed. Encrypting and processing meta-data obviously lightens the computational burden, but still requires the client to attain and process the complete database attribute list [167].

### 4.7.2 Architecture

The basic operating principle is to split the original local descriptors from each image into two parts where the payload of the descriptor $x_d$ can be shared in the public domain and the originating geometrical positions $x_g$ are kept private and shared only between authorized parties as shown in Fig. 4.10.

A user only sends a query consisting of local descriptors $y_d$ to the server without any geometric information. The server uses a BOW retrieval framework to return a list of candidates based on descriptor similarity together with encrypted geometrical information. The user then has to perform the final re-ranking locally. It is assumed that the
data user will contact the data owner directly bypassing the cloud server to obtain the authorization to access the geometrical data for the content of interest.

It is important to point out that a pair \((x_g, y_g)\) is considered to be a shared secret between the data owner and data user, which is not available to the server.

Summarizing, the identification system is considered as a search tool that returns a list of database entry indices that are close to the query \(y\) given some similarity metric. The size of this list determines the ambiguity of the server and the residual workload the data user has to execute to finalize the query.

The cardinality of the list can be further increased via ambiquization. Here the user intentionally adds false descriptors to the query, results from which he later filters out. The proposed framework is schematically presented in Fig. 4.11.
4.7.3 Performance, complexity and privacy protection

Impact of geometrical information

Assuming that only authorized users are privy to geometrical information, whereas the server only has access to the descriptor payloads, the upper performance limits may be obtained for the optimized parameters $\epsilon$ and $\tau$. The resulting performance is shown in Figure 4.8. The gap in performance between authorized users and the server is drastic. Practically, it means that if there are $M = 1$ Mio entries in the database, the server list size is $|\mathcal{L}(y_d)| \cong 10^3$, while an authorized user may prune this list to $|\mathcal{L}(y_d, y_g)| \cong 1$ with vanishingly small probability of error. In this example, the additional work load for the client is comprised of $10^3$ geometrical matchings, which is feasible for thin clients.

Finally, the communication load between the server and data users can be estimated, taking the compression and encoding of descriptors and geometrical coordinates into account.

One can estimate the amount of bytes to be communicated for $J_x = 500$ descriptors per images and with the retrieved list size $|\mathcal{L}(y_d)| = 10^3$:

- ORB descriptor: $1000$ candidates $\times$ 2KB (geometry) $\times$ 12KB (descriptors) = 23 MB,
- compressed SIFT descriptor: $1000$ candidates $\times$ 2KB (geometry) $\times$ 3.9KB (descriptors) = 7.6MB and
- CHOG descriptor with the compressed geometrical information: $1000$ candidates $\times$ 4KB (geometry and descriptor) = 3.9MB.

These estimates are promising and reasonably modest in comparison to the homomorphic encryption load.

Impact of ambiguization

The query identity can be protected further via ambiguization. Contrary to existing randomization solutions that degrade the query by adding noise or applying dimensionality reduction, the proposed method adds controlled randomness that can be filtered out by an authorized party.

For demonstration purposes, we assume that $J_x = 50$ and the data user sends his query with $J_y = 50$ correct and 100, 200, 300, 400 and 500 randomly added descriptors degrading the performance of the server as shown in Figure 4.12. The list size $\mathbb{E}\{|\mathcal{L}(y_{rd})|\} \simeq MP_F$ increases with $P_F$ and $J_y$ causing corresponding growth of the
communication rate. The informed data user, however, filters out the indices of those images obtained from the randomly added descriptors and achieves the performance identical to that of an informed synchronized system.
Chapter 5

Conclusions and Future work

In this thesis, we have presented an approach to physical object protection based on their microscopic surface structure. Optically acquired, these so-called micro-structures are both unique and currently non-cloneable and thus serve as a natural identifier. The current state-of-the-art was extended in a number of ways.

First and foremost, in this thesis, we build algorithms and frameworks to allow a common end user to use a hand held mobile device to acquire and verify objects using their micro-structure. The physical objects are completely unmodified and the hand held mobile device contains no special features or lighting of any kind. Several contributions have made this possible.

Micro-structures, by their vary nature, are noise like images, mostly devoid of any edges or distinct salient regions. Secondly, although the amount of physical objects is potentially very large, the number of acquisitions per unique sample is very low, typically not more than 3. Acquisition via a hand held mobile phone then adds a multitude of (geometrical) distortions. Any algorithm in the identification and authentication chain thus has very little to no a priori knowledge to work upon, nor can much information be derived from the samples. These primary challenges have been overcome in a number of ways.

Firstly, we have shown how micro-structures may be matched based on so called robust features and geometry. Specifically, it prosed an affinity based algorithm that can match small sets of points, of which over 50% are outliers.

Secondly, a novel robust descriptor was developed: SketchPrint. Specifically designed to robustly identify distorted micro-structures, it requires no training, is relatively stable and information rich, and requires relatively little enrolled descriptor vectors per sample. As it both captures geometrical and micro-structure information from its region
of interest, it does not require any exhaustive geometrical re-ranking, or aggregation. SketchPrint has been tested against the state-of-art, both on an individual descriptor level and within an aggregation framework. It shows superior performance over state-of-the-art methods, when micro-structure images are more degraded and achieves better or similar performance for micro-structures that have been acquired with industrial camera’s, whilst enrolling up to an order less of descriptors per sample. Lastly, it can be applied to a broad range of visually sparse images such as consumer package designs and text.

Lastly, a statistical model of a Bag of Word content identification architecture has been build. It captures all relevant parameters, from the type and quantity of the used descriptors, the desired robustness to noise, architectural choices such as the deployed type of pooling, and ties those in to the predicted end performance. This model has been empirically verified and used to design an application that enables privacy preserving content search.

Summarizing, this work presents the first steps to practical object identification based on their micro-structures using a mobile phone.

In this respect, we envision a number of emerging future directions and extensions. A number of those will be briefly touched upon.

Micro-structures are currently modeled as Gaussian, with correlated noise. Although theoretical identification and authentication frameworks build upon this assumption come sufficiently close to those using real data, this model is incomplete. A better synthetic micro-structure model, for example based on a fractal distribution, would aid future tools and models.

The used affinity algorithm in Chapter 2 is exhaustive, exchanging computational efficiency for robustness. Improvement has been difficult in part because local geometric methods proved to be too sensitive to the birth and death of geometric key-points, whereas the descriptors where not distinctive enough to allow quantization and dimensionality reduction. It thus remains an open issue. Secondly, the current affinity algorithm uses a simple threshold based on the variance of the values in the Euclidean distance matrix, which is not able to distinguish in- and outliers in all situations. A situation that could be improved with a superior statistical model of the Euclidean Distance Matrices.

The SketchPrint algorithm has a number of interesting future directions. All modifications that lessen the number of algorithm parameters and decouple existing ones from the size and scale of the image would be an improvement.
All measures that increase key-point stability would not only benefit SketchPrint, but all contemporary feature detectors. Currently, state-of-the-art detectors are still all handcrafted algorithms. A promising venue to explore are convolutional networks which can also be trained to detect stable features, especially for micro-structures.

The SketchPrint descriptor can be enhanced in two distinct directions. Firstly, the incorporation of techniques such as active content fingerprinting to achieve more robust and distinguishable SketchPrint descriptors. Secondly, SketchPrint is currently used stand-alone, but could possibly also benefit from further encoding and aggregation techniques.

The current statistical BOW model can be extended into a number of major directions. First of all, the current model assumes that descriptor vectors are independently distributed. Secondly, it lacks a theoretical analysis of more advanced aggregation and quantization techniques. Lastly, it is obvious that when an image suffers from degradations, this will influence feature descriptors that are based on this image. At the (hard) encoding stage, this might ultimately cause a query descriptor to be encoded against a different matching basis-vector centroid then at enrollment. A model that can accurately predict this type of miss for complex descriptors is currently lacking.
Appendix A

The FAMOS-W dataset

A.1 Introduction

The field of physical object security based on surface micro-structures lacks common shared data for the development, testing and benchmarking of authentication and identification technologies. Often datasets are proprietary or lack sufficient size for statistically significant results. In this Appendix, two publicly available datasets, the Forensic Authentication Micro-structure Optical Set (FAMOS) are described. They are available via http://sip.unige.ch/famos.

This Appendix is based primarily on [36, 31, 110].

A.2 FAMOS-W

To acquire a massive set of micro-structure images under relatively stable conditions, an experimental industrial system was developed. Two color cameras, designated RAF and NIK respectively, are deployed above a conveyor belt that feeds the paper samples through the system. The system can process up to 20000 samples in a single run. Lighting is identical and consists of a white led ring light together with an angled one approximately 90mm above the surface. The RAF camera has a resolution of 2592×1944 (5Mp) with a sensor size of 5.7×4.4mm and a pixel size of 2.2μm. It has an optical magnification of 1 : 0.9. The NIK camera has a resolution of 1601×1201 (2Mp), a sensor size of 7×5.2mm, a pixel size of 4.4μm and no optical magnification. The entire setup is shown in Figure A.1.

No preprocessing is applied, which ensures that future users are free to design their own preprocessing and feature extraction methods.
Appendix A. FAMOS

Figure A.1 – The industrial acquisition device, including the feeder, belt, camera, lighting and integrated computer screen which were used for the FAMOS-W acquisition.

The FAMOS-W dataset contains 5000 unique samples, acquired with the two different cameras, three times each giving a total of 30000 images. Raw acquisition examples for both cameras of a single sample are shown in Figures 1.7a-1.7b and extracted patches in Figures 1.8a-1.8a.

A.3 Alignment

Alignment or registration refers to the whole body of procedures that takes an acquired image from a physical object and undoes any kind of geometrical distortion prior to extracting a predetermined image patch with the sought micro-structure. Although in principle related to the fields of image stitching [169, 57, 58, 170] and image registration [46, 52, 171] methods from these domains fail to align micro-structures satisfactory. Feature based methods are generally not precise enough leaving residual geometrical fluctuations, mostly because micro-structures have little features that also lack in robustness. Image registration methods mostly use convex methods to optimize a geometrical transformation model using a metric such as mutual information between a template and a distorted image to assess the current fit. Micro-structure images, even when marked, are visually poor, lacking edges and corners ensuring that image registration methods, including those based on (learned) congealing [44] and funneling [45, 43, 42] consistently fail to find any meaningful solution. Furthermore, they are very computational intensive, pending the space of allowable distortions that need be accounted for.
Specifically for FAMOS-W, so called template based alignment was used, but it is not universally applicable. An overview of all methods for other datasets, such as those from a mobile phone in FAMOS-L is seen in Chapter 2.

A.4 ACF-Template based Alignment

This form of alignment uses a priori knowledge about the true shape of a printed mark on the object to assess any distortions that have occurred during acquisition. Its performance critically hinges on the quality of the mark and the surface it is manufactured on. For high quality goods, where the mark for example is a machined part of a logo from metal, this method attains the best alignment performance for both identification and authentication scenarios.

A.4.1 Alignment algorithm

The Alignment algorithm (Figure A.2) [7], as applied to FAMOS-W, broadly executes the following steps:

- The image is segmented into a binary image with the printed mark as new foreground.
- The ACF is determined of the mark.
- The peaks in the ACF are detected.
- Peaks in the ACF are matched against those of a predefined template using the Hungarian Algorithm [172].
- Rotation and scaling can be ascertained by solving $y_g = Ax_g$ for corresponding peak locations $\{y_g \leftrightarrow x_g\}$ and transformation matrix $A$, using the Direct Linear Transform (DLT) [58].
- Translation is finally obtained using normalized cross-correlation between the template and the target image.

A.4.1.1 Design considerations

The print quality and segmentation accuracy of the printed mark is of vital importance for the alignment procedure. Figure A.4 shows how fluctuations in the extracted shape influence the peaks in the ACF. This is the primary reason the printed mark used in FAMOS W contains a periodic repeating pattern of dots (Figure A.3). The resulting redundancy creates stable peaks in the ACF even if the segmentation results fluctuate.
Figure A.2 – The basic autocorrelation based alignment algorithm, as used for the FAMOS-W dataset.

Figure A.3 – The FAMOS-W template used for alignment and its corresponding Autocorrelation function.
Figure A.4 – Illustration of ACF peak sensitivity to fluctuations in the segmented mark. [7]
For this technique to work, both the segmentation result and the manufacturing quality must be consistent for all enrolled objects.
A.5 Basic patch based identification

The most basic identification architecture, previously seen in Section 2.2 and Figure 1.6, consists of alignment of both enrolled and presented queries followed by comparison of the extracted image patches in the real domain.

Obviously, one of the main challenges of developing an optimal identification system, based on micro-structures, is the fact that the distributions \( p(y|x') \) and \( p(y|x(m)) \) are hard to obtain [173].

However, once all micro-structures are aligned and rid of geometrical disturbances, one can assume that the residual distortions are additive and Gaussian in nature [36]. The hypothesis from (2.1) can be then reformulated as:

\[
\begin{align*}
H_{m'} : y &= x' + z, \\
H_m : y &= x(m) + z.
\end{align*}
\]

(A.1)

where \( H_{m'} \) corresponds to a mis identification and \( H_m \) denotes a valid case where the identifier is indeed \( m \). Following the assumption that the noise is Gaussian, i.e. \( Z \sim \mathcal{N}(0, \sigma^2_N I_N) \), the cross correlation coefficient \( \rho_{xy} \) is a sufficient statistic, between vectors \( x \) and \( y \).

The base line results for this setup can be seen in Figure A.5. For an information theoretic analysis on fingerprinting techniques using the FAMOS-W set, we refer to [32].

A.5.1 Noise modeling

Work in the previous Section hinges on two important assumptions. Firstly, that all geometrical distortions have been corrected and secondly, that residual distortions are additive and Gaussian in nature. This allows the cross correlation coefficient \( \rho_{xy} \) to be a sufficient statistic. Performance of this metric will obviously drop, the moment these assumptions are not met. Chapter 2 covers methods when geometrical distortions can not be corrected, whereas this Section will focus on methods and possibilities when the second criteria is not met.

Major difficulties will arise the moment the noise is not Gaussian or when the variance of the noise is signal dependent and therefore varies not only between micro-structure images, but also within individual samples. Classical approaches include the application of variance-stabilizing transforms to the original (noisy) patches, such that they
closely follow a Gaussian distribution in the transformed domain. Examples include the Anscombe transform [174, 175, 176], random projections [31], or the homomorphic approach mapping multiplicative noise to additive noise with stationary variance [177]. An important drawback of such (non-linear) homomorphic methods is that the mapping may not exist at all, or if applied, that this transformation will also distort noise free samples. [173]

In this section, we will assume that the additive noise is correlated with some covariance matrix, which can be empirically measured. An enhanced noise observation model is attained, and corresponding rules for the matching of observations and fingerprints are derived. This is done for both synthetic data and micro-structures from the FAMOS-W dataset in the direct and transformed domain.

In the previous section [178, 31, 32], it was assumed that both the source $X$ and the noise were $Z$ are i.i.d. Gaussian, from which it follows that the correlation is a sufficient statistic in the direct domain.

If one relaxes these assumptions and allow for correlated noise and data, i.e., we assume that $X \sim \mathcal{N}(0, K_{xx})$ and $Z \sim \mathcal{N}(0, K_{zz})$. We make no assumptions about the counterfeit items $X'$ other than that they are generated from the same distribution as authentic items. Consequently, the hypothesis test can be reformulated in terms of the distributions:
\[
\{ 
\begin{align*}
H_0 & : \ Y \sim N(0, K_{xx} + K_{zz}), \\
H_m & : \ Y \sim N(x(m), K_{zz}).
\end{align*}
\] (A.2)

Following the Neyman-Pearson framework, we can then formulate the decision rule for a chosen threshold \(\gamma\) which can be developed using Bayes’ rule:

\[
\frac{Pr[H_m | y]}{Pr[H_0 | y]} > \gamma \iff \frac{f(y | H_m)p(H_m)}{f(y | H_0)p(H_0)} > \gamma. \tag{A.3}
\]

Assuming \(p(H_0) = p(H_m) = \frac{1}{2}\) yields:

\[
\frac{f(y | H_m)}{f(y | H_0)} > \gamma \tag{A.4}
\]

\[
\frac{1}{\sqrt{2\pi |K_{zz}|}} \exp\left(-\frac{1}{2}(y - x(m))^T K_{zz}^{-1}(y - x(m))\right) > \gamma, \tag{A.5}
\]

where \(|.|\) denotes the determinant of the matrix, and which can be further developed to obtain a sufficient statistic \cite{179}:

\[
t(y) = y^T(K_{xx} + K_{zz})^{-1}y - (y - x(m))^T K_{zz}^{-1}(y - x(m)), \tag{A.6}
\]

\[
t(y) = y^T(K_{xx} + K_{zz})^{-1}y - y^T K_{zz}^{-1}y + 2y^T K_{zz}^{-1}x(m) - x^T(m)K_{zz}^{-1}x(m). \tag{A.7}
\]

A simplified sufficient statistic \(s(y)\) can be formulated by assuming that the terms \(y^T(K_{xx} + K_{zz})^{-1}y\), \(y^T K_{zz}^{-1}y\) and \(x^T(m)K_{zz}^{-1}x(m)\) are all constant for the expected observations:

\[
s(y) = y^T K_{zz}^{-1}x(m). \tag{A.8}
\]

A.5.1.1 Transformed Domain

The calculation of a fingerprint involves a transform, which we assume to be linear and orthogonal. An example of such a transform is the DCT, but random projections, which are approximately orthogonal \cite{38, 180, 32} can be considered as well. Let

\[
\tilde{X} = WX, \quad \tilde{Y} = WY, \tag{A.9}
\]
then
\[ \tilde{Y} = WY = W(X + Z) = WX + WZ, \]  
(A.10)

where \( \tilde{X} \sim N(0, K_{\tilde{x}\tilde{x}}) \), \( \tilde{Z} \sim N(0, K_{\tilde{z}\tilde{z}}) \), and \( K_{\tilde{x}\tilde{x}} = WK_{\tilde{x}\tilde{x}}W^T \) and \( K_{\tilde{z}\tilde{z}} = WK_{\tilde{z}\tilde{z}}W^T \).

The hypothesis test in the transformed domain is then:
\[
\begin{align*}
\mathcal{H}_0 &: \tilde{Y} = \tilde{X}' + \tilde{Z}, \\
\mathcal{H}_m &: \tilde{Y} = \tilde{x}(m) + \tilde{Z},
\end{align*}
\]  
(A.11)

which can be reformulated as in the direct domain:
\[
\begin{align*}
\mathcal{H}_0 &: \tilde{Y} \sim N(0, K_{\tilde{x}\tilde{x}} + K_{\tilde{z}\tilde{z}}), \\
\mathcal{H}_m &: \tilde{Y} \sim N(\tilde{x}(m), K_{\tilde{z}\tilde{z}}),
\end{align*}
\]  
(A.12)

leading to a sufficient statistic in the transformed domain:
\[
s(\tilde{y}) = \tilde{y}^T K_{\tilde{z}\tilde{z}}^{-1} \tilde{x}(m).
\]  
(A.13)

The challenge for systems designers is to choose a transform that leads to the most informative and robust fingerprints, which can be evaluated in information-theoretical terms [32].

### A.5.1.2 Experimental Results

We have opted to use the DCT as transform because of its energy-compacting properties which offers a good future perspective for dimensionality reduction.

The benchmark metric that is used for comparison is the sum inner product, both in the direct and in the DCT domain [178]:
\[
\begin{align*}
  r(y) &= y^T x(m), \quad \text{(A.14)} \\
  r(\tilde{y}) &= \tilde{y}^T \tilde{x}(m). \quad \text{(A.15)}
\end{align*}
\]
Let the probability of miss and false alarm for a chosen threshold $t$ be defined as:

$$
p_{m}^{R} = \Pr [R < t \mid \mathcal{H}_{m}] \quad p_{f}^{R} = \Pr [R \geq t \mid \mathcal{H}_{0}], \quad (A.16)
$$

$$
p_{m}^{\tilde{R}} = \Pr [\tilde{R} < t \mid \mathcal{H}_{m}] \quad p_{f}^{\tilde{R}} = \Pr [\tilde{R} \geq t \mid \mathcal{H}_{0}], \quad (A.17)
$$

$$
p_{m}^{S} = \Pr [S < t \mid \mathcal{H}_{m}] \quad p_{f}^{S} = \Pr [S \geq t \mid \mathcal{H}_{0}], \quad (A.18)
$$

$$
p_{m}^{\tilde{S}} = \Pr [\tilde{S} < t \mid \mathcal{H}_{m}] \quad p_{f}^{\tilde{S}} = \Pr [\tilde{S} \geq t \mid \mathcal{H}_{0}], \quad (A.19)
$$

where $r(y) \sim \tilde{R}$, $r(\tilde{y}) \sim \tilde{R}$ and $s(y) \sim \tilde{S}$, $s(\tilde{y}) \sim \tilde{S}$. The authentication performance will be analysed in terms of a Receiver Operating Characteristic (ROC) curve.

The performance has first been evaluated for synthetic data, where both $X \sim \mathcal{N}(0, I_{N})$ and $Z$ were generated by stationary Gauss-Markov processes with $\rho_{Z} = 0.85$, i.e. $Z \sim \mathcal{N}(0, K_{zz})$ resulting in two sets of 4096 vectors with a dimensionality of 128. The results are visible in Figure A.6, indicating in all cases that the use of the proposed measures leads to greater precision than the use of regular correlations. It also demonstrated that there is no difference between the direct domain and the linear transformed domain.

A justification for the assumption of correlated noise can be seen in Figure A.7, where the spectra of the differences between acquisition images of an identical sample are shown. The non-uniformity of the spectra confirms the dominance of low frequencies, indicating a correlation between the elements of the noise.

Figure A.8 shows the results for both metrics $t(y)$ and $s(y)$ for the FAMOS-W set in the direct domain. As is evident, the sufficient statistical metric $s(y)$ attains error-less authentication performance when the enrollment and verification camera are identical.
Figure A.7 – Spectra of the noise showing noise correlation of the FAMOS-W dataset.

Figure A.8 – Authentication results on the FAMOS-W dataset in the direct domain.

The worst performance is observed when these cameras are different, as seen in Figure A.8c, but the fact remains that deploying $s(y)$ leads to a significant performance improvement.

The results for the FAMOS-W set in the DCT domain can be seen in Figure A.9. In the situation where the enrollment and authentication camera are identical and the RAF camera is deployed, as seen in Figure A.9a, the performance is near identical to that of the direct domain as is evident from $r(\tilde{y})$, and $s(\tilde{y})$, which again attains error less performance. Small discrepancies between the direct and the DCT domain can be seen in Figure A.9b where the NIK camera has been deployed. There we see that the performance of $r(\tilde{y})$, and $s(\tilde{y})$ is much closer to each other, and the latter no longer leads to error less performance. Finally, as expected, the worst performance is observed when the enrollment and authentication cameras are not identical, as seen in Figure A.8c. Most notable here is the fact that $s(\tilde{y})$ deviates from $s(y)$ in the direct domain.
A.6 Conclusion

In conclusion, we have shown an alignment algorithm that allows to extract a patch from a previous defined ROI and undo geometrical distortion sufficiently, that it allows to apply fingerprinting methods.

Furthermore, we have shown the elementary statistical properties of the database and modeled the performance of an authentication framework on the FAMOS-W dataset. Secondly we have shown that a sufficient statistical metric based on a model with correlated noise leads to significantly better performance on the real world FAMOS-W set in comparison to metrics that assume the noise to be additive white Gaussian.

There are a significant number of directions for future research. Primarily, this will focus on finding a more accurate statistical image model to describe the optical micro-structures. Although the Gaussian model is a reasonable choice, visually one can see that micro-structures also exhibit small scratches and areas with different local variance and local mean, which are currently not modeled. A number of more involved synthetic examples can be seen in Figure A.10 which will be investigated in the future.

Future research will also include dimensionality reduction methods that optimize robustness against distortions. Secondly, the development and testing of improved matching techniques, specifically soft decoding, which uses implicit information about the magnitude of $X$. Further more we aim to develop a thorough information-theoretical analysis based on the framework introduced in [32].
Figure A.10 – Examples of synthetic micro-structures [8]. From left to right, the 2D Ising model (A.10a), The Moving Gaussian Average (A.10b), Perlin (1/f) noise (A.10c) and Brownian Fractal noise (A.10d).
Appendix B

Quantization of SIFT descriptors

B.1 Introduction

In this Appendix we shall show how random projections and simple sign based quantization may be used to reduce the dimensionality of SIFT descriptors without a significant loss in performance. It is primarily based on work in [37].

SIFT descriptors are used widespread across multiple domains, ranging from computer vision applications such as the matching and stitching of images to providing an intrinsic part of semantic image recognition algorithms. Dimensionality reduction and quantization of SIFT descriptors is attractive because it eases the storage burden and determining Hamming distances is computationally efficient [181, 126, 182].

Research in [183, 184] proposes to use coarsely quantized random projections to build a descriptor hash. An information-theoretic analysis of content identification based on similarly constructed hashes is performed in [38]. Work from [185, 186, 187] shows that using the median per dimension as a hard threshold is an effective strategy, when not reducing dimensionality.

Requirements

In general, one would like quantized SIFT descriptors, or any binarized feature vector, to exhibit the following properties:

1. Bits should be equally likely, i.e., \( Pr[B_{x(j)}(m) = 0] = Pr[B_{x(j)}(m) = 1] = 0.5 \) and \( H(B_{x(j)}(m)) = 1 \) for, \( \forall m, \forall k, \) and all \( j \in \{1, ..., L\} \) where \( L = 128 \) for SIFT descriptors.
2. Bits within a feature vector should be independent, i.e. $H(B_{x^k(m)}, B_{x^k(j)}) = H(B_{x^k(m)}) + H(B_{x^k(j)}) = 2$, $\forall m, \forall k$, and all $i, j \in \{1, ..., L\}$, $j \neq i$.

For completeness, any identification or authentication architecture that relies on binary feature vectors, should also insure the following:

3. Feature vectors $B_{x^k(m)}, B_{x^k(n)}$ originating from different objects $m$ and $n$ should be independent of each other, i.e. $H(B_{x^k(m)}, B_{x^k(n)}) = H(B_{x^k(m)}) + H(B_{x^k(n)}), \forall k, m \neq n$.

4. Feature vectors $B_{x^k(m)}, B_{x^k(m)}$ originating from the same object, should be correlated as modeled by a Binary Symmetric Channel (BSC) with some probability of bit error, $p_b$, i.e. $H(B_{x^{k'}(m)} | B_{x^k(m)}) = H(p)$, for $\forall m, \forall k$.

In Section B.2, dimension reduction and quantization of SIFT descriptor vectors will be realized using random projections and sign based quantization. Statistical methods to test the four above mentioned requirements for binarized feature vectors are found in Section B.3.1.

B.2 Random Projection and Binarization

The binary fingerprints are obtained via a two staged process: a dimensionality reduction $W^{L \times N}$ and binarization $Q(.)$ [188] shown in Figure B.1.

The dimensionality reduction of the $k$-th SIFT descriptor from image $m$, $x^k(m)$, is done as follows:

$$\tilde{x}^k(m) = W^{L \times N} x^k(m), \quad (B.1)$$

where $W^{L \times N} \in \mathbb{R}^{L \times N}$. $L$ is the number of dimensions $W^{L \times N}$ will map to, $N$ is the length of the input column vector, which for SIFT vectors is 128. Random matrix $W^{L \times N} = (W_1, W_2, \ldots, W_N)^T$ consists of a set of approximately orthonormal basis vectors.
Appendix B. Quantization

vectors, where all elements are generated as $W_i[j] \sim \mathcal{N}(0, \frac{1}{N})$, $1 \leq i \leq N, 1 \leq j \leq L$, and as such behaves as an approximate orthoprojector. It guarantees that the projected vectors $\tilde{x}^k(m)$ could be considered as realisations of a Gaussian source with a covariance matrix that converges to diagonal in probability [189].

Also, as stated in the introduction, one can reasonably expect raw SIFT descriptors to be correlated amongst each other. It was shown that random projections de-correlate data vectors amongst each other under certain assumptions [38]. Furthermore, such a dimensionality reduction is extremely fast and can be applied over time to an ever changing dataset.

Finally, binarization is simply done by extracting and storing the sign of all individual elements of all projected SIFT vectors:

$$b_{x^k(m)} = \{\text{sign}(\tilde{x}^k(m)[1]), \text{sign}(\tilde{x}^k(m)[2]), \ldots, \text{sign}(\tilde{x}^k(m)[L])\},$$

(B.2)

where for $i \in \{1 \ldots L\}$, $b^k_{x(m)}[i] \in \{0, 1\}$ and $\forall a, \text{sign}(a) = 1$ if $a \geq 0$ and 0 otherwise. The goal of using such a threshold is justified by maximizing the entropy of the extracted binary data which should converge to 1 bit per sample. The binarized projected descriptors are then stored in the database in the bin corresponding to image index $m$.

B.3 Performance Analysis

B.3.1 Descriptor statistics

In order to justify the discriminative power of both the original and binarized projected descriptors, we used the joint empirical entropy [181]. The maximum amount of uniquely distinguishable typical sequences generated from a given binary stationary source is limited by $2^{H(B_x)}$, where $H(B_x)$ is the joint entropy of $L$ random variables $B_{x_i}$, $i \in \{1, 2, \ldots, L\}$. One needs to accurately estimate $H(B_x)$ to evaluate this bound. Estimating such a function of a joint distribution is extremely complex in a high dimensional space. The other case follows the assumption that the bits are jointly independent, i.e. $H(B_x) = \sum_{i=1}^{L} H(B_{x_i}) = LH(B_x)$. This situation only requires the estimation of the marginal distributions. This approach while lacking a certain accuracy, provides an upper bound on the sought discriminative power, according to the chain rule for entropy [181].

Three datasets were used for empirical testing. Firstly, the Airplane subset from the Caltech 101 imageset [190] from which 36’000 raw SIFT were extracted. Secondly, the
micro-structure datasets FAMOS-W and FAMOS-L were used, both from which 50’000 uniformly sampled SIFT descriptors were extracted.

To empirically test the discriminative power, or requirements (1) and (2) in Section C.1, the sampled SIFT descriptors from all datasets were processed following the above advocated approach. Transformation of the original SIFT descriptors into the binary domain according to B.1 and B.2 converges them to i.i.d. binary strings, which, with high probability, exhibit $p(1) = p(0) = 0.5$.

To demonstrate bit independence within feature vectors, two test were performed. Firstly, the auto-correlation function from each of the individual binarized projected fingerprints was calculated. These were then averaged. Figure B.2 shows that there is asymptotically no correlation between bits of individual binarized projected SIFT descriptors, contrarily to the original domain.

Secondly, an information-theoretic driven approach used in biometrics [191] was used, consisting in the estimation of $H(B_x)$ under the pair-wise independence assumption. Figure B.3 shows that pair-wise entropy for projected quantized SIFT descriptors, from the Caltech set is 2 bits, with a probability of 0.7. The FAMOS-W and L dataset perform less in this respect with probabilities of 0.45 and 0.3.

However, our experiments show that there is a certain amount of correlation between binarized projected descriptors. The distances between all descriptor vectors in the original and quantized projected domain are shown in Figure B.4. The measured Hamming distances between all binarized projected descriptors from the Caltech 101 Airplane dataset can be modeled with Bernoulli distribution $B(L, 0.36)$, FAMOS-W with $B(L, 0.44)$ and those from FAMOS-L with $B(L, 0.46)$.

### B.3.2 Channel Distortions

To examine how binarized projected descriptors behave under various image distortions and how their discriminative power is influenced, a digital communications approach has been deployed. Assuming that the influence of various distortions in the original raw SIFT domain will be reflected by bit flips in the binarized projected domain, the probability of bit error $P_b$ [181] will be used to measure performance. Then, the resulting discriminative power of the binarized projected SIFT descriptors for a particular channel distortion can be evaluated as $2^{N(1 - H_2(P_b))} = 2^{128(1 - H_2(P_b))}$, where $H_2(P_b) = -P_b \log_2 P_b - (1 - P_b) \log_2(1 - P_b)$ denotes the binary entropy [181].

Images in the Airplane dataset were exposed to the following distortions (Table B.1). They were rotated with parameter $\theta$ in degrees and scaled with parameters $s = s_x = s_y$. 
Figure B.2 – The average correlation per image with in descriptors before and after projection and binarization for the Caltech Airplane set, FAMOS-W and FAMOS-L. The average correlation between original SIFT descriptors per image is shown via the normalized Auto Correlation Function in Figures B.2a-B.2c. The results after projection and binarization can be seen in B.2d-B.2f.

Figure B.3 – Normalized histogram of the pair-wise entropy $H(B_x[i], B_x[j])$, $i \neq j$, for random projected and quantized SIFT descriptors from the Caltech Airplane set (Figure B.3a), FAMOS-W (Figure B.3b) and FAMOS-L (Figure B.3c).
Appendix B. Quantization

Figure B.4 – Empirical distance distributions, and a fitted Gaussian distribution, between descriptor vectors in the original domain (Figures B.4a-B.4c) and projected quantized domain (Figures B.4d-B.4f) for the Caltech Airplane set, FAMOS-W and FAMOS-L.

$s_y$. The Similarity transformation was deployed with parameters $s$, $\theta$ and $t_x, t_y = 1$. Additive White Gaussian Noise was added to images with variance $\sigma^2$. And finally, images were subjected to JPEG compression with quality factor $q$.

To accurately model channel distortions for real and projected descriptors, one needs to annotate the ground truth, i.e., one must assure that the matching of descriptors between two identical but distorted images is correct. We distinguish the following situation: if a match between point correspondences $\{x^k(m) \leftrightarrow y^k(m)\}$ is correct, the distance between the matching descriptors $x^k(m)$ and $y^k(m)$ is kept in the set. As no matching metric is flawless, and a manual annotation is not feasible, the following procedure is followed. For all points, we check, using a weak geometry based rule, if the matching point $y^k(m)$ is in the expected region. If not, this distance between $\{x^k(m) \leftrightarrow y^k(m)\}$ is removed from the set. The same rule is also used to reassign matches. This method of assessment is possible as we know the channel distortions that will be applied in the tests. Note that this is a different procedure then the matching rule, as proposed by [55].
Similarity Transform, parameters $s, \theta$ (degrees).

<table>
<thead>
<tr>
<th>Distortion parameters</th>
<th>Relative match</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.2, 10)</td>
<td>33%</td>
</tr>
<tr>
<td>(0.4, 20)</td>
<td>85%</td>
</tr>
<tr>
<td>(0.6, 30)</td>
<td>86%</td>
</tr>
<tr>
<td>(0.8, 40)</td>
<td>90%</td>
</tr>
<tr>
<td>(1, 10)</td>
<td>100%</td>
</tr>
<tr>
<td>(1.2, 10)</td>
<td>93%</td>
</tr>
<tr>
<td>(1.4, 20)</td>
<td>92%</td>
</tr>
<tr>
<td>(1.6, 30)</td>
<td>93%</td>
</tr>
<tr>
<td>(1.8, 40)</td>
<td>93%</td>
</tr>
</tbody>
</table>

Scale transform, parameter $s$.

<table>
<thead>
<tr>
<th>Distortion parameters</th>
<th>Relative match</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>92%</td>
</tr>
<tr>
<td>0.5</td>
<td>96%</td>
</tr>
<tr>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1.25</td>
</tr>
<tr>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

JPEG, parameter is quality factor $q$.

<table>
<thead>
<tr>
<th>Distortion parameters</th>
<th>Relative match</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>95%</td>
</tr>
<tr>
<td>20</td>
<td>96%</td>
</tr>
<tr>
<td>30</td>
<td>97%</td>
</tr>
<tr>
<td>30</td>
<td>98%</td>
</tr>
<tr>
<td>50</td>
<td>99%</td>
</tr>
<tr>
<td>60</td>
<td>99%</td>
</tr>
<tr>
<td>70</td>
<td>99%</td>
</tr>
<tr>
<td>80</td>
<td>100%</td>
</tr>
<tr>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

Rotation transform, parameter $\theta$ (degrees).

<table>
<thead>
<tr>
<th>Distortion parameters</th>
<th>Relative match</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>89%</td>
</tr>
<tr>
<td>70</td>
<td>90%</td>
</tr>
<tr>
<td>60</td>
<td>92%</td>
</tr>
<tr>
<td>50</td>
<td>90%</td>
</tr>
<tr>
<td>40</td>
<td>92%</td>
</tr>
<tr>
<td>30</td>
<td>92%</td>
</tr>
<tr>
<td>20</td>
<td>93%</td>
</tr>
<tr>
<td>10</td>
<td>94%</td>
</tr>
<tr>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>

AWGN, parameter $\sigma$.

<table>
<thead>
<tr>
<th>Distortion parameters</th>
<th>Relative match</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>90%</td>
</tr>
<tr>
<td>0.08</td>
<td>93%</td>
</tr>
<tr>
<td>0.06</td>
<td>92%</td>
</tr>
<tr>
<td>0.04</td>
<td>92%</td>
</tr>
<tr>
<td>0.02</td>
<td>92%</td>
</tr>
<tr>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table B.1 – Impact of signal processing and geometrical distortions on fingerprint matching. This Table shows the used channel distortions and their parameters. Furthermore it presents the percentage of attained matches between binarized projected descriptors that are identical to the matches attained by the original SIFT descriptors for various distortions. The latter are treated as the baseline and set to a 100 percent.

Two class distances were measured, the distance between matching descriptors, or intra-class distances, and the distance between non matching descriptors, or inter-class distances.

Our experimental results (Figure B.5) demonstrate that both original and binarized projected SIFT descriptors have a reasonable resilience against these distortions. The maximal $P_b = 0.08$ was induced by the similarity transform with $s = 0.2, \theta = 10$. Therefore, discriminative power in this identification protocol is bounded by $2^{128 \cdot 0.60}$ images. Note, that this bound is obtained under the independence assumption and can therefore, not be considered tight due to the correlation between binarized projected SIFT descriptors.

The second and third column in Figure B.5 show the attained intra- and inter-class distances for these channel distortions and parameter that resulted in the largest probability of $P_b$. As the class distances overlap, it shows that perfect identification is not possible, neither in the original domain, nor in the binarized projected domain for these distortions.

The FAMOS-W and L datasets both have multiple real life acquisitions negating the need for artificially introduced distortions. The behavior of SIFT vectors for these sets is covered extensively in Chapters 2 and 3.
Figure B.5 – The probability of bit error, $P_b$ for for the tested distortions distortions for a set of projected SIFT descriptors.
Appendix B. Quantization

(a) Euclidean distance between $x$ and $y$ for $\theta = 50$ degrees rotation.

(b) Hamming distance between $b_x$ and $b_y$ for $\theta = 50$ degrees rotation.

(c) Euclidean distance between $x$ and $y$ for scale distortion of $s = 0.2$.

(d) Hamming distance between $b_x$ and $b_y$ for scale distortion of $s = 0.2$.

(e) Euclidean distance between $x$ and $y$ for the similarity transform distortion with $(s = 0.2, \theta = 10)$.

(f) Hamming distance between $b_x$ and $b_y$ for the similarity transform distortion with $(s = 0.2, \theta = 10)$. 
Appendix B. Quantization

Figure B.5 – Channel statistics for various distortions for a set of original and projected SIFT descriptors. \( \mathbf{x} \) and \( \mathbf{y} \) denote the codebook of original and the distorted sift descriptors. The projected counterparts are denoted by: \( \mathbf{b}_\mathbf{x} \) and \( \mathbf{b}_\mathbf{y} \). Descriptors from identical indices are matching descriptors.

(g) Euclidean distance between \( \mathbf{x} \) and \( \mathbf{y} \) for AWGN with \( \sigma^2 = 0.1 \).

(h) Hamming distance between \( \mathbf{b}_\mathbf{x} \) and \( \mathbf{b}_\mathbf{y} \) for AWGN with \( \sigma^2 = 0.1 \).

(i) Euclidean distance between \( \mathbf{x} \) and \( \mathbf{y} \) for JPEG compression distortion with \( q = 10 \).

(j) Hamming distance between \( \mathbf{b}_\mathbf{x} \) and \( \mathbf{b}_\mathbf{y} \) for JPEG compression distortion \( q = 10 \).
Appendix C

The Pharmaceutical package dataset

C.1 Introduction

The pharmaceutical dataset is currently a smallish but realistically set up dataset with pictures from pharmaceutical packages. Its goal is to serve as a testing set for package identification algorithms, for which very limited training data is available, and to contain images as would be shot by a casual mobile phone user.

It is comprised of 450 unique packages with Roman and Cyrillic printing. These packages have been in circulation and have normal wear and tear. Each package was acquired six times. Once with a flatbed scanner and three separate times with a hand-held mobile phone.

The mobile phone acquisitions were done without any special lighting or equipment and with little regard for the quality of the picture, other than that it must contain the box proper. A single sample is always acquired with the flash. The used phone was a Samsung Galaxy SIII. Examples results may be seen in Figure C.1.
Figure C.1 – Examples of acquisitions from the Pharmaceutical package dataset.
Bibliography


[139] ——, “Privacy preserving multimedia content identification for cloud based bag-of-feature architectures,” in Workshop on Cloud Based Media in conjunction with
the 2015 IEEE International Conference on Multimedia and Expo (ICME 2015),
Torino, Italy, July, 3 2015.


