FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Insensitive Image Comparison in the Absence of Training Data

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der technischen Wissenschaften

by

Sebastian Zambanini
Registration Number 9925897

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Priv. Doz. Dr. Martin Kampel

The dissertation has been reviewed by:

(Priv.-Doz. Dr. Martin Kampel) (Prof. Dr. Dieter W. Fellner)

Wien, Aug. 19th, 2014

(Sebastian Zambanini)

Erklärung zur Verfassung der Arbeit

Sebastian Zambanini
Kreuzgasse 49/2/25, 1180 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasser)

Danksagung

An dieser Stelle möchte ich mich bei all jenen Personen bedanken, die mich bei der Erstellung dieser Dissertation unterstützt haben.

Ein besonderer Dank gilt selbstverständlich Martin Kampel für die Betreuung dieser Dissertation. Bei ihm und Robert Sablatnig möchte ich mich auch dafür bedanken, mir die Möglichkeit zur wissenschaftlichen Betätigung gegeben zu haben, ohne die diese Dissertation nicht entstanden wäre. Bei Klaus Vondrovec möchte ich mich für die Unterstützung in numismatischen Fragen und für die gute Projektzusammenarbeit bedanken. Dieter W. Fellner danke ich für die Begutachtung dieser Arbeit.

Ein Doktoratsstudium ist ein langjähriges Unterfangen, welches durch eine nette und inspirierende Kollegenschaft sowie gegenseitige Unterstützung stark gefördert wird. Aus diesem Grund möchte ich mich bei all meinen Arbeitskollegen bedanken, die mich über die Jahre begleitet haben, insbesondere bei Michael Hödlmoser, Andreas Zweng, Albert Kavelar, Markus Diem, Stefan Fiel, Florian Kleber, Rainer Planinc and Fabian Hollaus.

Nicht zuletzt möchte ich mich bei meiner Familie und bei meinen Eltern im Besonderen für die immerwährende und bedingungslose Unterstützung auf meinem Lebensweg bedanken.

Den größten Dank möchte ich aber meiner geliebten Julia und unserem Sohn Anton aussprechen, denen ich diese Arbeit widme. Julia, Danke für Deine uneingeschränkte moralische Unterstützung und Geduld mit mir. Anton, Deine Geburt und Dein tägliches Lächeln hat mir mehr als alles andere den Ansporn gegeben, diese Arbeit fertig zu stellen.

The research presented in this dissertation thesis has been partly supported by the Austrian Science Fund (FWF) under the grant TRP140-N23-2010 (ILAC).

Abstract

This thesis deals with the problem of automatically estimating the visual similarity of two objects shown in an image pair. Visual image comparison is a challenging task in the presence of appearance variations between objects, as the similarity estimation has to be made insensitive to the variations without losing the essential information necessary for differentiation. A common and effective methodology to handle appearance variations is to exploit machine learning techniques where the intra-class variations are learned by means of representative example images. However, this methodology relies on large amounts of a-priori available image data which might be infeasible in practice. Therefore, the work presented in this thesis aims at the robust classification with the aid of an insensitive image-to-image similarity estimation. Consequently, an exemplar-based classification pipeline is presented whose individual steps treat different aspects of appearance variability. The task of recognizing ancient coins is used as motivating example and main application area of the presented methods due to the challenging nature of ancient coins in terms of illumination effects, non-rigid spatial deformations, image clutter and inter-class similarity.

In the first part of the pipeline the segmentation of roughly circular objects like ancient coins is treated in order to make the visual comparison insensitive to object location and scale as well as background clutter. The second part deals with the illumination-insensitive extraction of image features, with a special focus on textureless objects like coins. Textureless objects exhibit more complex appearance variations under illumination changes than textured objects, which have been the main objects of interest in computer vision research on illumination insensitivity so far. Thus, an exhaustive evaluation of low-level image representations for recognizing textureless objects under illumination changes is presented. The findings of this study are utilized to construct a local image descriptor that outperforms state-of-the-art descriptors under illumination changes. Finally, in the last part the insensitivity against non-rigid local deformations is addressed, as this type of appearance variations typically occurs within instances of the same coin class. It is shown that by imposing both appearance-based and geometric constraints on the optimization framework for correspondence search one can use the matching costs for exemplar-based coin classification in a coarse-to-fine manner. However, the classification performance of this methodology suffers from the computational demands of using only weak geometric constraints. Appearance-driven feature matching followed by an evaluation of the geometric plausibility of the detected correspondences allows to use stronger geometric constraints and consequently leads to a faster and more reliable similarity estimation.

Kurzfassung

Diese Dissertation beschäftigt sich mit dem Problem, die visuelle Ähnlichkeit von in Bildern dargestellten Objekten zu bestimmen. Die besondere Schwierigkeit einer solchen Ähnlichkeitsbestimmung liegt darin, das Ähnlichkeitsmaß insensitiv gegenüber Veränderungen im Aussehen der Objekte zu machen, ohne die nützliche Information zur Unterscheidung von Objekten zu verlieren. Ein gebräuchlicher und effektiver Weg dafür ist der Einsatz von Techniken des maschinellen Lernens, bei denen die Variationen innerhalb einer Objektklasse anhand von repräsentativen Beispielbildern automatisch gelernt werden. Für einen effektiven Einsatz müssen diese Bilder aber in einer großen Zahl vorhanden sein, was in der Praxis nicht immer möglich ist. Aus diesem Grund verfolgt die vorgestellte Arbeit das Ziel, eine robuste Klassifizierung mithilfe eines Bildähnlichkeitsmaßes zu erreichen. Es wird ein exemplar-basierter Klassifizierungsprozess vorgestellt, dessen Einzelschritte unterschiedliche Aspekte von Objektvariabilität behandeln. Dieser Prozess ist von dem Problem der automatischen Klassifizierung antiker Münzen motiviert, da für diese Objekte eine Vielzahl von Variabilitäten wie Beleuchtungseffekte, räumliche Verformungen oder unvollständige Bilddaten berücksichtigt werden müssen.

Im ersten Teil der Arbeit wird eine Bildsegmentierung von annähernd runden Objekten vorgestellt, die es ermöglicht, den Bildvergleich unabhängig vom Hintergrund und der Größe des Objektes im Bild durchzuführen. Der zweite Teil untersucht die Berechnung von beleuchtungsinsensitiven Bildmerkmalen mit dem Hauptaugenmerk auf untexturierte Objekte. Untexturierte Objekte wie beispielsweise antike Münzen bestehen aus nur einer einheitlichen Farbe und sind daher unter Beleuchtungsunterschieden schwieriger zu erkennen als texturierte Objekte, weshalb sie in der Vergangenheit im Bereich der Computer Vision Forschung großteils vernachlässigt wurden. Aus diesem Grund werden in der Arbeit in einer umfangreichen Studie einfache pixelbasierte Merkmale auf ihre Insensitivität zu Beleuchtungsveränderungen untersucht. Die Erkenntnisse dieser Studie werden in weiterer Folge dazu genutzt, einen lokalen Bilddeskriptor zu entwickeln, der unter Beleuchtungsunterschieden leistungsfähiger als bestehende Deskriptoren ist. Der letzte Teil der Arbeit wird der Insensitivität gegenüber räumlichen Verformungen gewidmet, wie sie beispielsweise innerhalb von Objekten eines Münztyps vorkommen. Es wird gezeigt, dass das Ergebnis einer Suche nach zusammengehörenden Bildpunkten, die sowohl aussehensbasierte als auch geometrische Kriterien berücksichtigt, dazu genutzt werden kann, eine schrittweise Klassifizierung von Münzen zu erreichen. Jedoch erlaubt dieser rechenintensive Prozess lediglich die Berücksichtigung von einfachen geometrischen Bedingungen. Aus diesem Grund wird eine verbesserte Methode vorgestellt, die die zuverlässigsten Korrespondenzen von Bildpunkten dazu verwendet, um die Ähnlichkeit aus deren geometrischer Plausibilität abzuleiten, was zu einem schneller berechenbaren und leistungsfähigerem Ähnlichkeitsmaß führt.

Contents

1 Introduction **1**
- 1.1 Motivation and Scope of Work . . . 4
- 1.2 Summary of Contributions . . . 6
- 1.3 Thesis Structure . . . 8

2 Background and Related Work **11**
- 2.1 Invariance in Visual Object Comparison . . . 11
  - 2.1.1 Translation Invariance . . . 11
  - 2.1.2 Scale Invariance . . . 12
  - 2.1.3 Illumination Invariance . . . 14
  - 2.1.4 Invariance to Non-Rigid Deformations . . . 19
  - 2.1.5 Invariance to Incomplete Image Data . . . 20
- 2.2 Image Segmentation . . . 20
- 2.3 Local Image Descriptors . . . 23
  - 2.3.1 Low-Level Per-Pixel Feature Extraction . . . 24
  - 2.3.2 Spatial Encoding . . . 24
  - 2.3.3 Feature Vector Postprocessing . . . 26
  - 2.3.4 Illumination Insensitivity of Local Descriptors . . . 27
- 2.4 Correspondence-Based Image Similarity . . . 27
  - 2.4.1 Without Geometric Constraints . . . 29
  - 2.4.2 With Geometric Constraints . . . 30
- 2.5 Image-Based Coin Analysis . . . 32
  - 2.5.1 Introduction to Ancient Coinage . . . 32
  - 2.5.2 The Challenges of Image-Based Ancient Coin Analysis . . . 34
  - 2.5.3 Coin Image Segmentation . . . 35
  - 2.5.4 Image-Based Coin Classification . . . 36
- 2.6 Summary and Innovative Aspects of the Thesis . . . 38

3 Shape-Controlled Object Segmentation **41**
- 3.1 Methodology . . . 41
  - 3.1.1 Saliency Extraction . . . 42
  - 3.1.2 Shape-Controlled Thresholding . . . 43
- 3.2 Experiments . . . 46
  - 3.2.1 Comparison with Manual Coin Segmentations . . . 47

3.2.2 Evaluation for Shape-Based Coin Identification . . . 49
3.3 Summary . . . 51

4 Illumination-Insensitive Feature Extraction **53**
4.1 Evaluation of Low-Level Image Representations . . . 53
4.1.1 Low-Level Image Representations . . . 54
4.1.2 Experiments . . . 60
4.2 LIDRIC: A Local Image Descriptor Robust to Illumination Changes . . . 70
4.2.1 LIDRIC Descriptor Construction . . . 71
4.2.2 Experiments . . . 73
4.3 Summary . . . 80

5 Correspondence-Based Image Similarity **83**
5.1 Dense Feature Matching for Hierarchical Coarse-to-Fine Exemplar-Based Classification . . . 84
5.1.1 Image Similarity from SIFT Flow . . . 84
5.1.2 Insensitivity to Coin Rotations . . . 87
5.1.3 Hierarchical Coarse-To-Fine Classification . . . 88
5.2 Improved Similarity from Feature Correspondences by Evaluating Geometric Plausibility (GP) . . . 89
5.2.1 Feature Extraction and First-Order Matching . . . 89
5.2.2 Similarity Estimation from First-Order Correspondences . . . 91
5.3 Experiments . . . 92
5.3.1 Roman Republican Coin Datasets . . . 93
5.3.2 Comparison of Coin Matching with and without Geometric Constraints 95
5.3.3 Hierarchical Classification Performance and Runtime Analysis of SIFT Flow Method . . . 98
5.3.4 Analysis of Coin Rotation Insensitivity of SIFT Flow Method . . . 99
5.3.5 Comparison of Coin Classification Methods on Multi-Source Dataset . 101
5.3.6 Classification Performance on Large-Scale Single-Source Dataset . . . 104
5.4 Summary . . . 105

6 Conclusions **107**
6.1 Limitations . . . 108
6.2 Future Work and Implications . . . 110

Bibliography **113**

List of Acronyms **133**

CHAPTER 1

Introduction

In computer vision a central and recurrent problem is to determine the similarity of images or parts of images. For instance, in face recognition [Li and Jain, 2011] a probe image needs to be compared to gallery face images in order to find the most similar one and to consequently identify a person. Object tracking in a video stream [Yilmaz et al., 2006] also requires a notion of image similarity in order to detect the image part which looks most similar to the object appearance defined in the previous frame. Another example is Content-Based Image Retrieval (CBIR) [Datta et al., 2008] where an user might want to search for images with a color layout similar to a query image. Image retrieval can be also based on higher-level semantics [Liu et al., 2007] where image similarity is defined in terms of learned image categories like car, plane, motorbike etc. These examples show that there is a broad spectrum of the meaning of the term *image similarity* as well as of technical approaches needed to compute a both robust and discriminative similarity measure for a given application domain.

Essential to the use of image similarity measures is the question which information has to be extracted from the images and which information needs to be ignored, as the image similarity measure should be based only on the necessary image information and not be disturbed by other data inherent in the image. For instance, if image comparison is rather focused on an object shown somewhere in the image than on the complete image, image parts belonging to the background should not be considered, either implicitly within the method or explicitly by rejecting background image parts in a preprocessing step. Ignoring unnecessary, disturbing image information is also related to the concepts of *invariance* and *insensitivity*. Invariance in computer vision is the property of a feature or outcome of an operation to remain unaltered by a certain, defined set of image variations [Fisher et al., 2014]. Invariance can be proofed theoretically, whereas insensitivity is a looser constraint needed in domains where true invariants do not exist (e.g. illumination, see Section 2.1.3). Thus, in contrast to invariance, insensitivity can be defined only in a relative but not in an absolute manner: it can be examined only if a method is more or less insensitive to a given type of image transformation than another method and if a method is or is not invariant to a given type of image transformation.

In the context of object-based image similarity the issues raised above mean that we want

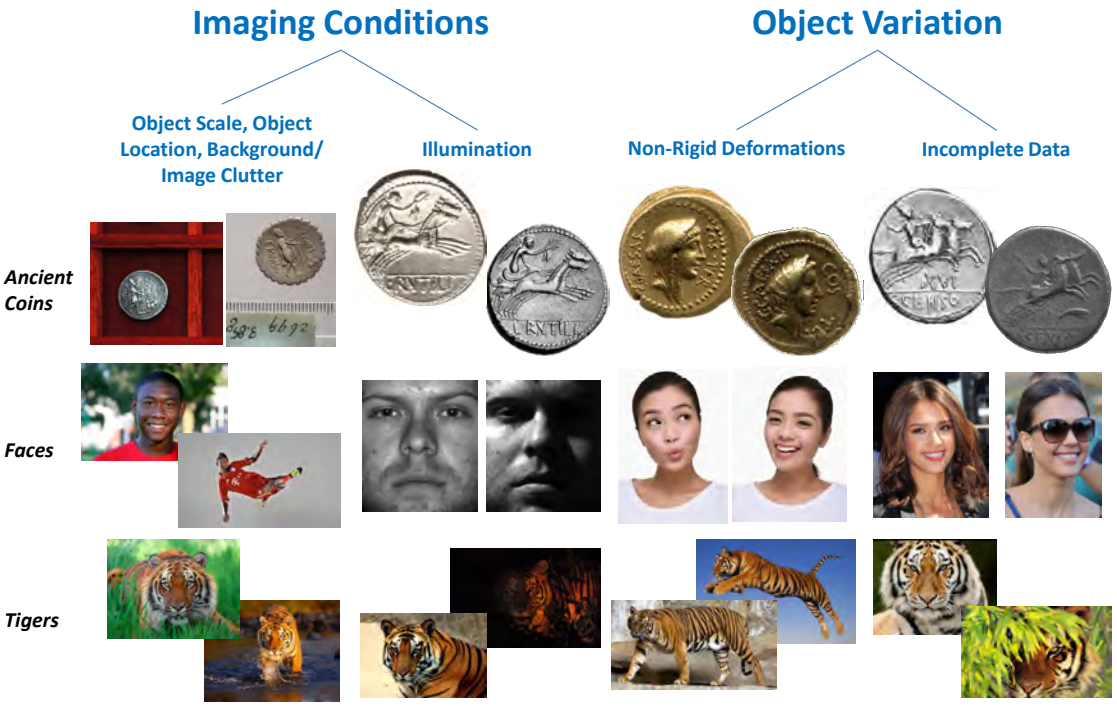


Figure 1.1: Causes of appearance variability, illustrated by example images for the application scenarios of *classifying ancient coins*, *identifying human faces* and *recognizing tigers*[1].

to compare two images of objects in such a way that, on the one hand, the resulting similarity measure is not affected by various appearance variations and, on the other hand, the similarity measure is discriminative enough to be suitable for the given task. As these two properties stand in a trade-off relationship with each other, we seek for methods which optimally balance them [Varma and Ray, 2007]. Consequently, it is beneficial to identify the types of possible image variations for a given application area, since each kind of invariance or insensitivity might unnecessarily decrease the discriminative power of the image similarity measure. In general, two main types of variations can be identified which are also illustrated in Figure 1.1 for the object types of ancient coins, human faces and tigers:

- **Variations due to imaging conditions**

  Object scale and location, background clutter: Whenever imaging conditions cannot be controlled, computer vision methods need to be made robust against the resulting image appearance variations. Arbitrary camera viewpoints lead to an unknown, arbitrary location and scale of the object to be recognized. The imaged background is considered as image clutter which potentially disturbs the recognition process. As a solution, the object

[1] Please note that the figure does not show a complete categorization of image variations (see [Frisby and Stone, 2010] for a comprehensive taxonomy).

has to be detected in a preprocessing step or feature extraction needs to be invariant to scale and location.

Illumination: Feature extraction is also required to ignore effects of illumination change like object brightness (e.g. day/night shot of tiger), highlights (e.g. metallic surfaces like coins) or shading (e.g. different illumination directions during face image acquisition).

- **Variations due to alterations of the imaged object**:

  Non-rigid deformations: The object to be recognized may also have varying physical shapes. For instance, the individual specimens of an ancient coin class can be non-rigidly deformed due to the hand-made dies used in ancient times for coin minting. However, non-rigid deformation does not only occur between different instances of the same class, but also for the same "physical object" acquired at different times, like the expressions on a human face. Also a tiger has a non-rigid shape which means that the relative position of body parts may change from one image to another. Consequently, the underlying recognition model needs a certain spatial flexibility to be insensitive to deformations of the object.

  Incomplete Data: Finally, as the imaged object might not be completely visible, feature extraction and recognition has to be designed in such a way that missing or disturbed data does not corrupt the recognition process. For example, on ancient coins abrasions and fragmentation can possibly lead to missing data. Missing data is also often related to occlusions which can be caused by uncontrolled imaging conditions as well, e.g. a bush occluding parts of the tiger.

The research presented in this thesis aims at making visual object comparison insensitive to the different types of appearance variabilities described above. In contrast to the prevalent methodology of learning image variability by training a classifier with image samples [Duda et al., 2012], a direct way of comparing images without the need for an offline training stage is followed. Consequently, individual methodologies are presented and evaluated, each one treating certain aspects of appearance variability and insensitivity:

- Object segmentation is proposed to deal with varying object locations and scales on different backgrounds.
- Illumination variations are intensively examined by means of synthetic and real image data leading to a new illumination-insensitive image descriptor.
- Image similarity is estimated based on a local correspondence model which shows high discriminative power while being robust against non-rigid deformations and missing image data.

The methods can be used in conjunction to obtain a holistic image similarity framework in the end. Besides the insensitivity concern, reducing the runtime of the image search is treated as well.

1.1 Motivation and Scope of Work

Although coping with image variabilities is a major concern in computer vision research, there are issues that have not been examined accordingly. The common methodology for treating image and object variation is using machine learning techniques [Duda et al., 2012]. The idea is to represent the overall set of variabilities in the training data set which is then used to learn models able to capture the variability of a given class [Varma and Zisserman, 2005, Russell et al., 2008, Wright et al., 2009, Xiang et al., 2014]. The amount of training data needed depends on the difficulty and degree of variability between images, but state-of-the-art image classification results are based on training set sizes in the order of millions of images [Torralba et al., 2008, Deng et al., 2009][2]. The digital era and easy access to online images makes the collection of training data for broad categories like tigers possible [Chen et al., 2013], but this is a condition which cannot be fulfilled in certain domains like ancient coin recognition or identification of (non-famous) persons.

Due to this training data problem, in this thesis the insensitive comparison of images without using a-priori knowledge in the form of learned recognition models is pursued. For the problem of image classification this means that no discriminative or generative models are learned in an offline training phase but a query image is compared to class-reference images in a dataset and classified to the most similar class image. This exemplar-based or nearest neighbor classification scheme is in line with recent works that show superior results for scenarios with limited number of training samples per class [Dreuw et al., 2009, Hua and Akbarzadeh, 2009, Pishchulin et al., 2012, Yao and Fei-Fei, 2012]. The benefit of such a classification scheme is that the variability can be treated during classification by means of insensitive image comparison, without relying on a training dataset that includes the overall appearance variability of a class. Moreover, there are additional advantages of exemplar-based classification compared to learning-based classification [Boiman et al., 2008]: it can naturally handle a higher a number of classes, it is non-parametric and therefore not subjected to over-fitting, and it needs no time-consuming offline learning phase which is particularly beneficial for dynamic databases as changing classes/training sets is instantaneous in exemplar-based classifiers.

The complexity of exemplar-based classification does not lie in the model used, but rather in the image similarity metric which has to consider appearance variabilities as shown in Figure 1.1. In the presence of image clutter and non-rigid deformations, establishing and assessing corresponding feature points between images has shown to be a prominent approach for this task [Berg et al., 2005, Duchenne et al., 2011b, Jorstad et al., 2011, Liu et al., 2011, Kim et al., 2013]. In this thesis, this line of research is further investigated in order to obtain a deeper understanding how local correspondences can be exploited to estimate image similarity. In this context, local image descriptors play an essential role as insensitive descriptions of the local image structures are needed to establish reliable correspondences. However, although a vast amount of descriptors have been proposed in the past (see Section 2.3 for an overview), insensitivity to illumination conditions has been only marginally treated, dealing exclusively with global brightness changes as common on textured surfaces. Nevertheless, as can be seen in Figure 1.1, textureless surfaces like coins or parts of the human face exhibit a higher degree of appearance variability

[2] see for instance the results of the *Large Scale Visual Recognition Challenge 2013* at http://www.image-net.org/challenges/LSVRC/2013/index (accessed on June 8th, 2014).

under illumination changes. Therefore, in this thesis illumination insensitivity on textureless surfaces is further examined, aiming for a more robust local image descriptor in the presence of illumination changes.

Within a correspondence-based recognition framework, location and scale differences of objects are commonly treated by detecting salient keypoints with a canonical scale [Lowe, 2004, Hassner et al., 2012]. However, due to the discriminative power-invariance trade-off and non-perfect repeatability of keypoint detection, knowing the image region that contains the object of interest before classification is conducive to a higher rate of correct correspondences and thus for a more reliable image similarity metric. Therefore, object segmentation is proposed as a first step for correspondence-based image classification in this thesis. For objects like ancient coins keypoint detection is more error-prone than segmenting the coin which can be achieved in a robust manner because of the known roughly circular shape of the coins.

Throughout this thesis, the task of classifying ancient coins is used as motivating example and main application area of the presented methods since the aforementioned challenging problems are inherently present in this type of data, as can also be seen in Figure 1.1:

- ancient coins can be imaged at different image locations, scales and on various backgrounds.

- ancient coins are textureless objects with a 3D relief, and thus the appearance of the coin surface in a 2D image is strongly influenced by the illumination conditions.

- ancient coins possibly show a high level of variability within a class due to their non-industrial manufacturing and abrasions over the centuries.

- ancient coin classes are numerous and have different levels of rarity: for instance, in the *Museum of Fine Arts* in *Vienna* around 3900 coins of the Roman Republican age are available, but for only 237 of the 1900 classes (including the subclasses) defined in [Crawford, 1974] more than three coins are available [Zambanini and Kampel, 2012]. Hence, the number of training samples is limited which heavily impedes the successful learning of appearance variabilities by machine learning methods.

The manual classification of ancient coins is in general a time-consuming and difficult task, even for trained experts [Grierson, 1975]. From a numismatic perspective, the motivation for the research presented in this thesis is that an image-based analysis has the potential to disburden the daily work of numismatists, but also to support more efficient research procedures in the future, e.g. the automatic clustering of coin hoards [Zaharieva et al., 2008]. Therefore, the methods presented in this thesis allow to automatically assess the visual similarity of coins, which in turn enables exemplar-based classification.

However, although the experiments in this thesis focus on ancient coin classification, the proposed methods are not restricted to this specific task. Instead, they generally treat objects with the appearance variabilities illustrated in Figure 1.1 and thus have the potential to be used and adapted for a wider class of problems (see Chapter 6 for a more detailed discussion of this issue).

1.2 Summary of Contributions

The general topic of this thesis is to automatically estimate the visual similarity of two objects shown in an image pair, and particularly insensitivity to appearance variabilities as depicted in Figure 1.1 is examined. Insensitivity to the various types of image variations are treated as individual subproblems and thus contributions to different aspects of insensitive image comparison are made:

1. **Insensitivity to Object Scale, Location and Background**

   - A shape-controlled approach for object segmentation is proposed [Zambanini and Kampel, 2009] which empirically proves to be robust and fast, in contrast to shape-unaware approaches. The method exploits known a-priori information about the shape of the object to be segmented, e.g. the roughly circular shape of an ancient coin. With the known location of the object, the image can be cut and normalized to a specified standard size to achieve insensitivity to object location and scale. This avoids using scale-invariant feature detection [Lowe, 2004] where only a sparse set of partially non-corresponding keypoints can be used for image comparison. Instead, dense matching can be performed which uses the overall visual data for comparison.
   - The circular object segmentation method [Zambanini and Kampel, 2009] is successfully used as a preprocessing step in various works on ancient coin recognition [Zambanini and Kampel, 2011, Zambanini and Kampel, 2012, Zambanini et al., 2013, Zambanini et al., 2014, Zambanini and Kampel, 2014, Kavelar et al., 2014]. Additionally, the resulting shape of the coin border is used as a feature for coin identification [Huber-Mörk et al., 2011]. Due to wearing the shape of the border is a characteristic feature of a coin specimen. In [Huber-Mörk et al., 2011] this property is exploited to build a system for automatic coin identification, motivated by the problem of illegal online coin trade which was combatted by the EU-funded COINS project [Zaharieva et al., 2007c] by means of an automatic retrieval of images of stolen coins in the internet.

2. **Insensitivity to Illumination Conditions**

   - A comprehensive evaluation of pixel-wise low-level features proposed in literature is conducted [Zambanini and Kampel, 2013a]. Unlike previous studies [Chen et al., 2000, Osadchy et al., 2007, Moreels and Perona, 2007, Van De Sande et al., 2010] the influence of material specularity, object texturedness and amount of illumination direction change is investigated. The experiments reveal that jets of oriented even Gabor filter responses are the features of choice for capturing object characteristics in an illumination-insensitive way, and that the single-scale representation can be extended towards multiple scales for improved performance.
   - The controlled evaluation is enabled by a new synthetic image dataset built from 3D historical coin models. The dataset makes it possible to directly compare the performance of the features under different conditions without introducing a bias

due to different objects used between datasets. As a contribution to other researchers in this field the dataset is made publicly available[3].

- The findings of the study are used to develop a new illumination-insensitive local image descriptor [Zambanini and Kampel, 2013b] which empirically shows to outperform state-of-the-art descriptors such as SIFT [Lowe, 2004], SURF [Bay et al., 2008], FREAK [Alahi et al., 2012], DAISY [Tola et al., 2010] or MROGH [Fan et al., 2012] under illumination changes. The source code of the descriptor is also made publicly available[4].

3. **Insensitivity to Non-Rigid Deformations and Incomplete Data**

   - In order to cope with the training data problem, an exemplar-based classification method is proposed [Zambanini and Kampel, 2011]. It uses a dense grid of local image features which are optimally matched by means of data-driven and geometric constraints to infer about image similarity. Due to the local appearance description and geometric regularization the method allows for the flexible matching needed in presence of non-rigid deformations. Because of the locality of the similarity estimation the method is also not vulnerable to occlusions or incomplete data.
   - A coarse-to-fine hierarchical classification scheme [Zambanini and Kampel, 2012] is introduced to decrease runtime. The main drawback of exemplar-based classification is the runtime which is theoretically linear to the number of classes in the dataset. It is shown that the runtime can be reduced to approximately one seventh without decreasing the classification rate.
   - It is further shown that the similarity estimation can be improved by evaluating data-driven matchings for their geometric plausibility [Zambanini et al., 2014], instead of regularizing the matching process by geometric constraints. As a consequence, geometric constraints have to be evaluated only once in the similarity estimation process, hence more complex constraints can be used which improves both the runtime and the classification performance.

On the application side, the individual methods allow to build a complete pipeline for classification of ancient coins, as depicted in Figure 1.2. This procedure is used in [Zambanini et al., 2014] (described in Chapter 5) and shows to outperform existing ancient coin classification approaches [Kampel and Zaharieva, 2008, Arandjelović, 2010]. The proposed methodology can be integrated into a numismatic coin classification tool whose impact on the numismatic research community would be versatile. First of all, it would help to save considerable amounts of time for everyone dealing with historical coins, as classification needs no longer to be based on printed reference books such as [Crawford, 1974] which might also be expensive and hardly available. The tool could be also filled with different sets of data and be used for various purposes such as die analysis [Howgego, 2005] or pre-classification of hoard finds.

[3] Synthetic Image Dataset for Illumination Robustness Evaluation (SIDIRE), http://www.caa.tuwien.ac.at/cvl/people/zamba/sidire/

[4] Local Image Descriptor Robust to Illumination Changes (LIDRIC), http://www.caa.tuwien.ac.at/cvl/people/zamba/lidric/LIDRIC_v1.02.rar

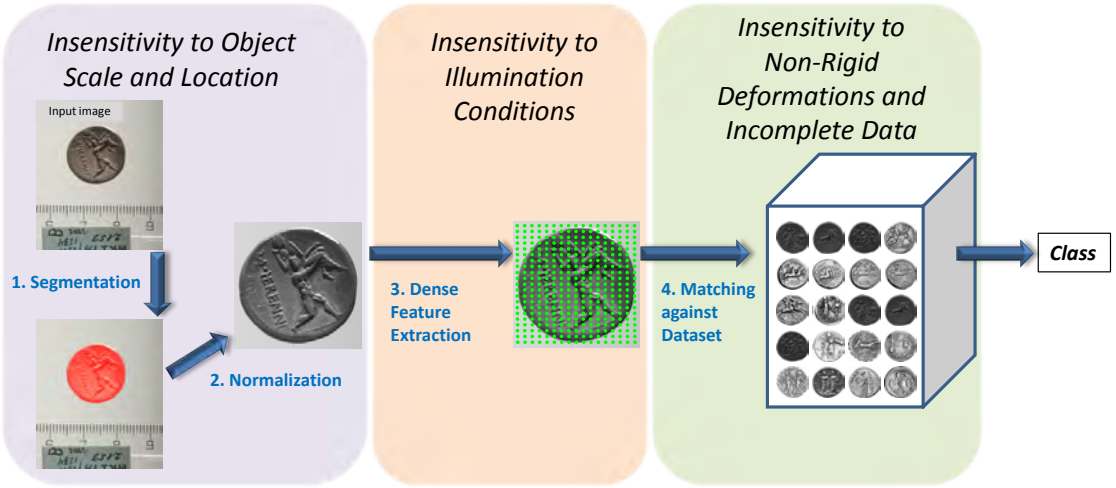


Figure 1.2: The coin classification pipeline that is provided by the methods presented in this thesis, each one achieving insensitivity to different intra-class appearance variations. Segmenting the coin and normalizing the coin region to a standard resolution achieves *insensitivity to location and scale*, dense LIDRIC feature extraction achieves *insensitivity to illumination conditions* and computing correspondence-based image similarities achieves *insensitivity to non-rigid deformations and incomplete data*.

From a more general perspective, automatic image-based coin classification contributes to the research field of *ICT and cultural heritage* which aims to capture, analyze, manage and deliver cultural information [Stanco et al., 2011]. The importance of this field is indicated by its inclusion in the activities of the EU Horizon 2020 research programme [H2020, 2013] as well as by several periodic scientific conferences[5] and journals[6]. The EU-funded COINS project (2007-2009) [Zaharieva et al., 2007c] aimed at developing technologies to fight against illegal trade and theft of coins by means of standardized inventories, data management tools and an image-based web search. This thesis was embedded in the ILAC project (2010-2014) [Kavelar et al., 2013], which was funded by the Austrian Science Fund (FWF) and aimed at the image-based classification of ancient coins.

1.3 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 gives an overview of the state-of-the-art in computer vision fields related to the methods proposed in this thesis, namely invariance in visual object comparison, image seg-

[5]For instance, the *Computer Applications and Quantitative Methods in Archaeology Conference (CAA), International Congress on Digital Heritage* and *EUROGRAPHICS Workshop on Graphics and Cultural Heritage (GCH)*.

[6]For instance, the *Elsevier Journal of Cultural Heritage* and the *ACM Journal on Computing and Cultural Heritage*.

mentation, local image descriptors and correspondence-based image similarity. Additionally, an outline of relevant numismatic knowledge and methods for image-based coin analysis is provided.

Chapter 3 describes the method for achieving *insensitivity to object scale and location* by means of a shape-controlled object segmentation approach. Since the object of interest is known to have a specific shape (e.g. roughly circular for coins), a global threshold is optimized in a way such that the resulting shape is most similar to a circle. Results of the developed algorithm are shown for an image database of ancient coins from various sources and demonstrate the benefits of the approach in terms of robustness and speed.

Chapter 4 describes the method for achieving *insensitivity to illumination conditions* and starts with a comprehensive evaluation of the discriminative power of various low-level image features for a pixel-wise representation of the objects characteristics. For this purpose, a new dataset with rendered images of 3D models is used which allows to directly compare the influences of texture and material properties in an object recognition scenario. The results are further validated on a dataset of real object images and finally reveal that jets of single- and multi-scale even Gabor filter responses outperform other proposed features in scenarios with textureless objects and strong variations of illumination. In the second part of the chapter a new local image descriptor (LIDRIC) is proposed based on these findings. The descriptor is computed from histograms of oriented filter responses in various subcells of the local image region. For evaluation, a dataset of textured as well as textureless objects is used which introduces a greater challenge towards evaluating the robustness against illumination changes than conventional datasets used in the past. The experiments finally show the superiority of LIDRIC compared to existing descriptors under illumination changes.

Chapter 5 describes the method for achieving *insensitivity to non-rigid deformations and incomplete data*. It is demonstrated that learning-based methods are not practical and effective for the image classification problem in the case of a high number of classes with a limited number of training samples and complex intra-class variations. As a solution, a similarity metric based on feature correspondence is proposed which is designed to be robust against the possible intra-class coin variations like degraded parts and non-rigid deformations. The similarity metric is used in an exemplar-based ancient coin classification scheme which shows to outperform previously proposed methods for ancient coin recognition. Comparative experiments are conducted on a dataset of 60 Roman Republican coin classes where the presented method achieves superior classification rates ranging from 72.7% for the case of one training sample per class up to 97.2% when nine training samples per class are used. Additionally, a coarse-to-fine classification scheme is introduced to decrease runtime which would be otherwise linear to the number of classes in the training set.

Chapter 6 finally concludes the thesis by summarizing its main outcomes and implications for computer vision research. The chapter also includes a discussion of the limitations of the presented methods and highlights future research directions.

All the methods described in the chapters 3-5 have already been partially published in conference proceedings and journal articles. At the beginning of each chapter the corresponding publications are denoted.

CHAPTER 2

Background and Related Work

In this chapter background information and related work within the scope of this thesis' research is summarized. In Section 2.1, a general overview of methodologies for dealing with different types of appearance variations is given. The subsequent sections describe the state-of-the-art in the computer vision fields where the methods presented in this thesis add a contribution: image segmentation (Section 2.2), local image descriptors (Section 2.3) and correspondence-based image similarity (Section 2.4). Section 2.5 gives an overview of relevant numismatic background knowledge and image-based coin analysis methods, as this is the main practical focus of the thesis. The chapter is concluded by a summary in Section 2.6 where the key points of the related literature are highlighted and the consequential design choices and innovative aspects of the proposed object comparison methodology are described.

2.1 Invariance in Visual Object Comparison

Invariance in visual object comparison means that we want to compare objects such that non-informative variations of the images are not taken into account. This implies that the outcome of feature extraction is not affected by any environmental or object-specific conditions. This section gives an outline of existing methods for overcoming the kinds of appearance variations that are also treated by the methods proposed in this thesis: object location, object size, illumination conditions, non-rigid object deformations and missing object data.

2.1.1 Translation Invariance

Translation invariance in visual object comparison means to obtain a similarity measure that is unaffected by the location of the object in the images. Early recognition approaches in the computer vision field were dominated by simple global methods [Smeulders et al., 2000], which extract a description from the entire image, commonly being inherently translation-invariant. For instance, Swain and Ballard [Swain and Ballard, 1991] proposed to use color histograms to

describe and compare images. Other features used for image similarity have been shape [Mehtre et al., 1997] and texture [Manjunath and Ma, 1996, Ojala et al., 1996].

A problem with these methods is that all image data is equally considered for comparison which makes it vulnerable to background clutter. To locate the object as a whole in the image, a general model of its appearance or shape can be used and searched for via template matching [Brunelli, 2009]. The idea is to compare every image position to the model/template and locate the object at the image point with the highest correlation. The methodology has been successfully applied to problems like face detection [Viola and Jones, 2004] or character recognition [Due Trier et al., 1996]. However, one of its disadvantages is speed, as the model has to be tested for the complete parameter space of variations, i.e. for every image pixel, but also for different window sizes and orientations if object scaling and rotation is encountered. Furthermore, if the object variability is to complex a general template of the object class cannot be defined.

State-of-the-art algorithms for image description and comparison are based on interest points and thus represent an image as a set of detected points with corresponding local descriptors [Nixon and Aguado, 2012]. Hence, two parts are involved: a detector and a descriptor. Within this feature extraction scheme translation invariance follows directly from the interest point detection that aims to detect points with a high degree of saliency. An important performance criterion of interest point detectors is the repeatability [Schmid et al., 2000] which is the fraction of corresponding detected keypoints among all detected keypoints of two images. An interest point that is not redetected in the other image hinders the image comparison process and thus interest point detectors focus on well-defined shapes like corners [Moravec, 1980, Harris and Stephens, 1988]. Modern interest point detectors [Mikolajczyk and Schmid, 2004] provide not only invariance to translation and rotation but also to scaling, as described in the next section.

2.1.2 Scale Invariance

Scale differences of objects are differences in their pixel dimensions, e.g. due to varying object distances, camera focal length or image resolution. Whenever these differences are unknown, feature extraction needs to be scale-invariant. As a solution, scale selection techniques have been proposed that seek for a characteristic scale of each feature point. This procedure has been established mainly by the work of Lindeberg [Lindeberg, 1998] who first proposed to use extrema in the Laplacian of Gaussian (LoG) function computed over various image scales as interest point locations. Lowe [Lowe, 2004] proposes to use the Difference-of-Gaussians (DoG) operator as a more efficient approximation of LoG. The 3D scale space (2D image space plus scale) is then scanned for local maxima, followed by subpixel refinement using a quadratic fit and rejection of unstable points in low contrast areas or near edges. Another examples for interest point detectors with scale selection are Harris-Laplace [Mikolajczyk and Schmid, 2004], SURF [Bay et al., 2008], MSER [Matas et al., 2004], principle curvature [Deng et al., 2007] and SIFER [Mainali et al., 2013].

The problem of feature extraction with scale selection is that stable scales can be detected only for a sparse set of image points, typically on average on around 0.05-0.5% of all image pixels [Aanæs et al., 2012]. Furthermore, scale selection becomes more and more unreliable with increasing scale differences. For instance, experiments in [Mikolajczyk, 2002] reveal that for a scale change factor of 4.4 only 13.3% of the scales selected by LoG are correct. To overcome

these problems, local feature extraction methods have been proposed that are inherently scale-invariant without requiring to normalize image patches by the selected scale [Hassner et al., 2012, Kokkinos and Yuille, 2008], but apparently at the price of decreased discriminative power due to the discriminative power/invariance trade-off [Varma and Ray, 2007].

The problem of interest point detection with scale selection is demonstrated in Figure 2.1 for comparing images of ancient coins with unknown scales. Two coin images of the same class with a scale difference of 2 are compared by extracting DoG interest points [Lowe, 2004]. A manual inspection reveals that only $\sim 32\%$ of the interest points in Figure 2.1b have a corresponding interest point in Figure 2.1a at the same coin location and correct scale difference (green circles), due to the image structure and scale changes between the two images. Moreover, the repeatable interest points span only over certain subareas of the coin which means that the other areas are excluded from comparison.

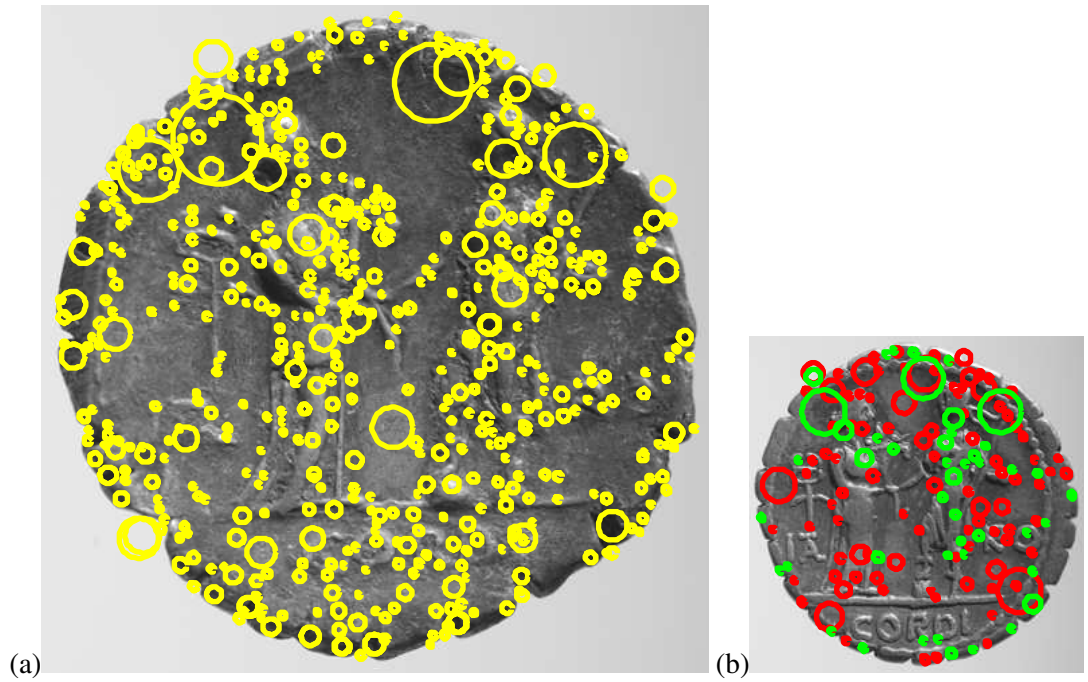
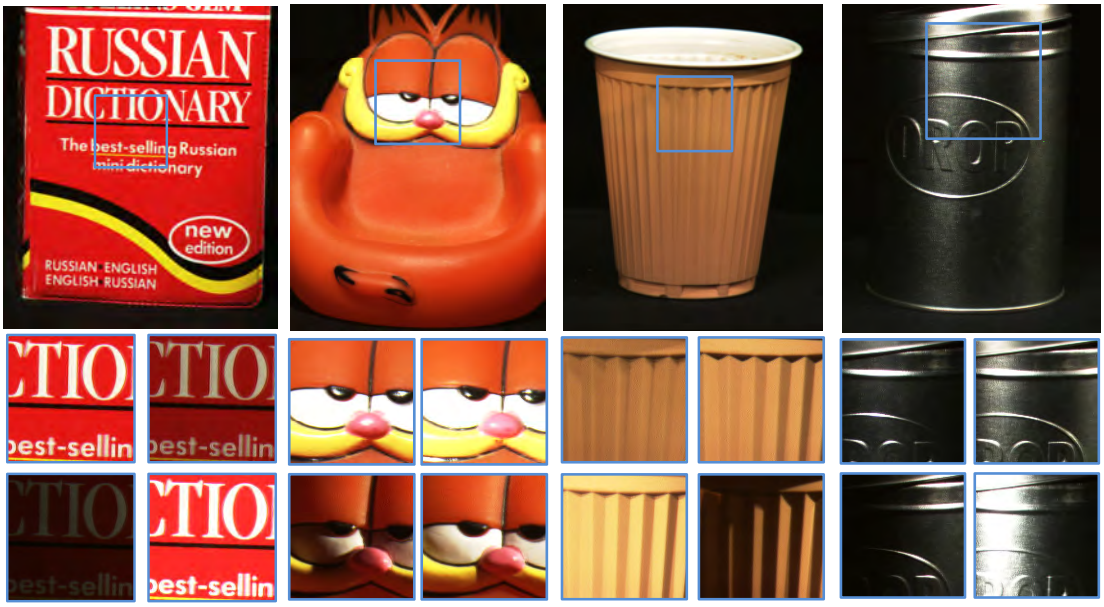


Figure 2.1: Detected interest points with selected scale (radius of circles) for two coin images of the same class. The coin image (b) has half the scale of the coin image (a) and interest points with a correspondence in (a) are marked in green, while interest points without a correspondence are marked in red.

To conclude, scale differences have to be mastered for application scenarios like coin recognition, but scale selection has to be guaranteed to be robust. Therefore, instead of selecting scales on a local level, in this thesis scale invariance is achieved on a global level by segmenting the coin area and normalizing the images to the same scale (see Chapter 3). As an additional benefit, the segmented coin area provides the region of interest for the further classification process and thus an elegant way to ignore background clutter is given.



(a) Flat, textured (b) Non-flat, textured (c) Textureless (d) Textureless, shiny

Figure 2.2: Four different kinds of objects from the ALOI dataset [Geusebroek et al., 2005]. The patches at the bottom show the appearance variations under illumination changes.

2.1.3 Illumination Invariance

When an object is imaged by a camera, its properties are communicated by the light that is reflected from the object's surface and projected onto the camera's image plane. Hence, without light no information about the object can be gathered, but the information is generally not unique, as an object can produce a variety of different images. The reason is that the light spectrum which is reflected from an object point depends on the light sources (i.e. the direction(s) and intensity of light that hits the object point) and is not necessarily uniformly distributed for all outgoing directions. Therefore, the level of difficulty to make feature extraction unaffected by the lighting conditions is influenced by the possible amount of illumination change between images, but also by the constitution of the objects. The influence of the object constitution is demonstrated in Figure 2.2 by means of different kinds of objects which exhibit an increasing degree of appearance variations under illumination changes (from left to right). Issues to be considered are:

- **Object Texturedness:** Within this thesis, *texture* refers to changes in the reflectance properties of an object, i.e. changes in the intrinsic color (a.k.a. albedo) of an object. Therefore, a *textured object* has a varying intrinsic color (Figure 2.2a-b), whereas a *textureless object* has only one, constant intrinsic color (Figure 2.2c-d). Obviously, texture is a rich source of information for recognition, e.g. considering the eyes and lips of a face or the stripes of a tiger. It also has the advantage that it is robust with respect to illumination changes. For instance, changing the illumination conditions for the flat, textured object

shown in Figure 2.2a induces only changes of the global brightness and does not affect the discontinuities of the intrinsic object color. In contrast, a textureless object does not contain texture information and thus information about its 3D shape becomes more important for recognition.

- **Object Depth Discontinuities:** Similar to texture discontinuities (edges), object depth discontinuities can be exploited for recognition but these object properties cannot be directly exploited in 2D images and have to be derived from the image appearance. This is generally an ill-posed problem [Belhumeur et al., 1999] in unconstrained conditions and the space of image appearance variation due to lighting changes is much wider than for flat objects, as can be seen in Figure 2.2b-d. If the object is non-flat and textured as the one shown in Figure 2.2b, also an increasing degree of appearance variation can be spotted but texture information can still be extracted robustly and used for recognition. In contrast, without texture additional effects of illumination change have to be considered: for instance, opposite lighting directions change the polarity of image edges at depth discontinuities like the ridges of the cup in Figure 2.2c. Image contrast can also vary more locally than for flat, textured objects due to shadows and the uneven intensity of light reflected by the surface with varying surface normals.

- **Object Material:** The material an object is made of also influences its reflective properties which are usually described by the Bidirectional Reflectance Distribution Function (BRDF) [Koenderink, 2010]. The BRDF defines for any incoming ray of light the light intensity that is emitted in a particular direction. Matte objects with a constant BRDF are called Lambertian, an assumption which is often made in computer vision to simplify the computational model [Basri and Jacobs, 2003, Wang et al., 2004, Drbohlav and Chantler, 2005, Osadchy et al., 2007, Liu and Dai, 2009]. However, objects with a shiny surface violate this assumption and exhibit an increased variability under illumination changes owing to local highlights, as shown in Figure 2.2d.

The examples in Figure 2.2 show that there are manifold ways how the appearance of an object can change under illumination variations, which leads to the overall crux of illumination invariance: it is generally impossible without any a-priori knowledge about the objects. This is theoretically proven by [Chen et al., 2000], meaning there is no function that maps every image to an illumination-invariant and object-specific representation. Even for Lambertian surfaces and point light sources it is always possible for any two images to find an object that produces these two images under different lighting directions, as shown by the example in Figure 2.3. Therefore, the best one can do is to find representations that are more insensitive than the pure gray values, that is representations whose likelihood of variation under illumination change is minimal. [Chen et al., 2000] propose to use image gradient directions and show that a simple recognition system that uses learned probabilities of gradient angle differences as dissimilarity measure is effective on face recognition. In general, image gradient directions are not affected by global brightness changes of the image as occurring on flat, textured objects like the one shown in Figure 2.2a.

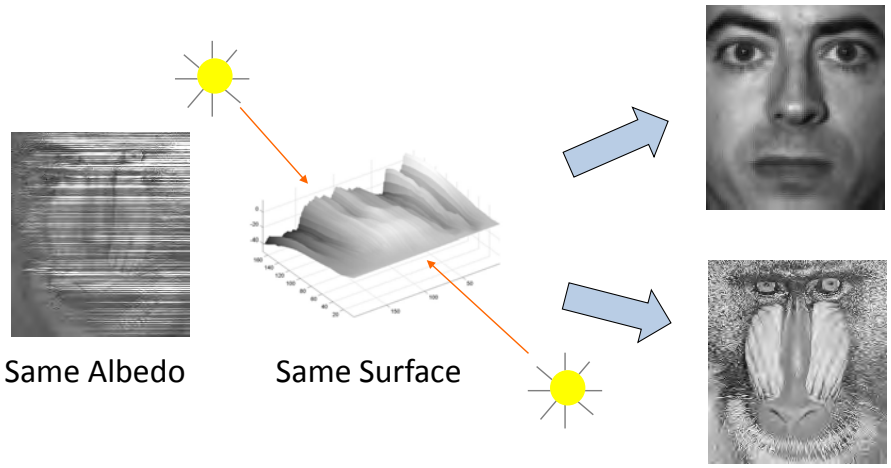


Figure 2.3: Illustration of the prove for the non-existence of illumination invariants [Chen et al., 2000]. The two different images on the right are produced by the same 3D object illuminated from two different directions (figure adapted from CVPR slides of Hansen F. Chen, 2000).

2.1.3.1 The Difference between Textured and Textureless Objects

Illumination insensitivity has been studied for various subfields of computer vision like face recognition [Gopalan and Jacobs, 2010], stereo vision [Hirschmüller and Scharstein, 2009] or object tracking [Gouiffès et al., 2012]. However, the research focuses on textured objects as texture is a common property of objects and illumination-insensitive feature extraction is less complex, as demonstrated in Figure 2.2. This goes back to the influential work of [Barrow and Tenenbaum, 1978] who postulated to extract *intrinsic images* of a scene - object-specific properties like depth, surface normals or color. More recently the term *intrinsic images* is widely understood as the decomposition of an image $I(x, y)$ into a reflectance image $R(x, y)$ of texture information and an image $L(x, y)$ of the illumination effects [Weiss, 2001, Tappen et al., 2006, Xie et al., 2011]. Therefore, the standard model used relates these quantities by

$$I(x, y) = R(x, y) \cdot L(x, y) \tag{2.1}$$

as illustrated in Figure 2.4a. We can see that the reflectance image bears the texture information without the illumination effects, hence an illumination-insensitive representation of the scene has been achieved. However, such a decomposition is of limited use for textureless objects like coins, statues, building facades etc. For instance, decomposing a coin image as in Figure 2.4b gives a representation which might be useful to segment the image into coin and background region, but does not help to recover the relief-like structures (i.e. object depth discontinuities) on the coin which are mandatory for classification.

2.1.3.2 Textured Objects

As the vast majority of papers dealing with illumination insensitivity follow the intrinsic image model, they attempt to extract the reflectance component from the image. A reasonable way to

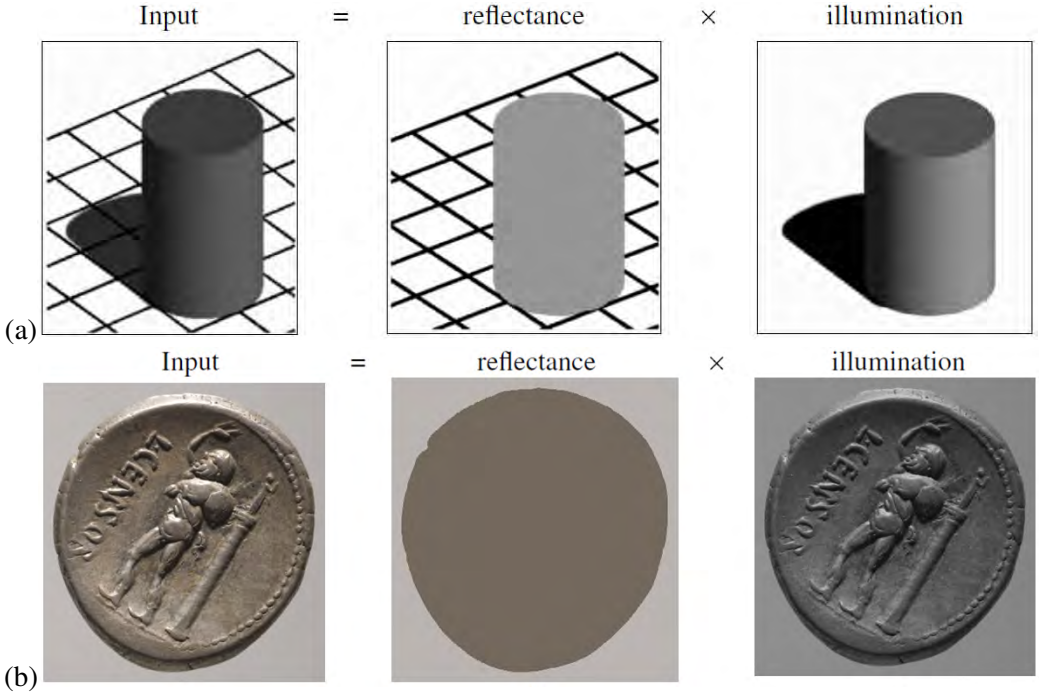


Figure 2.4: (a) Illustration of intrinsic image decomposition (figure taken from [Weiss, 2001]), (b) theoretical solution of intrinsic image decomposition for an image of a textureless ancient coin.

do so is to extract the high-frequency parts of the image as on Lambertian surfaces shading is smooth and thus contained in the low-frequency parts of the image [Shashua and Riklin-Raviv, 2001]. Hence, the idea of self-quotient images is followed by various authors [Wang et al., 2004, Chen et al., 2006, Arandjelovic, 2013] where the illumination is represented by a low-pass filtered image version $\hat{I}(x, y)$ and the intrinsic image can be obtained by simple image division:

$$R(x, y) = \frac{I(x, y)}{\hat{I}(x, y)}. \tag{2.2}$$

In a similar direction lie frequency transformations which are used to obtain high-frequency subbands for illumination-insensitive recognition, like wavelets [Garcia et al., 1998, Liu and Dai, 2009], discrete cosine transform [Hafed and Levine, 2001] or Fourier transform [Lai et al., 2001]. The image gradient directions proposed by [Chen et al., 2000] are also regularly used as they are stable for the high-frequency parts of an image, e.g. for illumination-insensitive face recognition [Zhang et al., 2009] or local image descriptors [Lowe, 2004, Chen et al., 2010, Tola et al., 2010]. Some works deal with extracting color as reflectance information rather than texture discontinuities. Such color invariants are proposed for local image descriptors [Van De Sande et al., 2010], edge detection [Gevers and Stokman, 2003] or keypoint detection [Van De Weijer et al., 2006].

These general-purpose low-level methods have the advantage of simple closed-form computation, but have a limited level of insensitivity due to their universal nature. Hence, methods making use of different kinds of a-priori information about the objects of interest or the scene conditions have been introduced to provide a more powerful image representation for specific domains. [Drbohlav and Chantler, 2005] exploit the known illumination direction of two images to obtain a comparable image representation by means of derivative filters oriented in reciprocal directions. The ambiguity of illumination effects can also be dissolved by using 3D models of the objects to be recognized. [Basri and Jacobs, 2003] show that the manifold of image appearance of Lambertian objects lies in a 9-parameter space of spherical harmonics which can be derived from a 3D model. This model-based approach is also extended to handle specular objects [Shirdhonkar and Jacobs, 2005, Netz and Osadchy, 2011]. Besides 3D models, the variability of image appearance can also be learned from training images taken under different illumination conditions [Georghiades et al., 2001,Leung and Malik, 2001,Cula and Dana, 2004], but the real world practicability is equally limited due to this requirement.

2.1.3.3 Textureless Objects

When it comes to textureless objects there is actually little research on efficient feature extraction in general scenarios without using any a-priori available information like material BRDF, illumination conditions, object shape in form of 3D models or illumination-induced object variability in form of training images. As the main source of information on such objects is the 3D shape (depth), shape from shading techniques [Horn, 1989] can be considered which attempt to derive depth information from the shading pattern on the objects. However, the shape from shading problem is generally ill-posed even for controlled imaging conditions, as the higher number of variables than measurements precludes a unique solution [Prados and Faugeras, 2006]. Still, different constraints like smoothness of the result can be used to obtain reasonable results, but a high degree of computational effort is needed and the robustness of the result is limited for arbitrary, unknown conditions (e.g. unknown lighting direction, unknown BRDF, unknown shape priors etc.) [Forsyth, 2011].

[Osadchy et al., 2007] investigate low-level illumination insensitive feature extraction in a general sense. They do not explicitly differentiate between textured and textureless objects, but rather between *non-isotropic* and *isotropic* surfaces. Non-isotropic surfaces are surfaces whose characteristics change in one direction and less in another, i.e. surfaces with discontinuities in depth or albedo. In contrast, isotropic surfaces are smooth in both directions, i.e. textureless and with slow variation in depth. Isotropic surfaces are more challenging, because in this case the gradients depend more on the illumination. The authors showed that the main problem of these surfaces is their high correlation, hence a whitening filter [Jain, 1989] helps to decorrelate the image intensities and makes the result more discriminative. As an approximation for whitening, the LoG filter is suggested. To handle both isotropic and non-isotropic surface types the orientational information of image gradients and the whitening effects of the LoG can be effectively combined by a Jet of Oriented Second Derivative filters (JOSD).

Although the problem of illumination-insensitive low-level feature extraction was studied by [Chen et al., 2000] and [Osadchy et al., 2007], research on illumination invariance lacks a comprehensive investigation and dataset of features for textureless objects. Both [Chen et al.,

2000] and [Osadchy et al., 2007] do not explicitly separate the cases of textured and textureless objects and thus cannot give a well-founded statement about the performance of the investigated representations on textureless objects. It is also unclear how the performances of low-level features are related to the material properties and the amount of illumination change. Therefore, a comprehensive evaluation on synthetic datasets with varying degrees of specularity and texturedness and on real images of textured and textureless objects is done in this thesis (see Section 4.1).

2.1.4 Invariance to Non-Rigid Deformations

Non-rigid deformations are spatial transformations between corresponding image features that cannot be described by a single global transformation. This problem has to be tackled whenever the shape of an object is not fixed and changes over time (e.g. a human, a tiger,...) or the class of objects exhibits non-rigid deformations (e.g. an ancient coin class, a person's signature, object categories like chair etc.). Therefore, it is of interest in areas like image registration [Crum et al., 2004], object tracking [Comaniciu et al., 2000] or object detection [Ferrari et al., 2010].

The simplest way to handle non-rigid deformations is to completely neglect the location of local parts in an image. The well-known Bag Of Visual Words (BOVW) model [Csurka et al., 2004] uses only the local image descriptions for recognition and is thus invariant to non-rigid deformations as long as the keypoint detector and descriptor are not affected. However, although the model is able to deal with large deformations between images, it is too general for a fine-grained classification and needs to be equipped with additional constraints considering the spatial locations of local parts. If local deformations can be considered to be small, spatial pooling can be applied to summarize the local descriptions over defined image regions, an operation that happens also in the visual cortex [Hubel and Wiesel, 1962]. Spatial pooling schemes proposed in the past are for instance spatial pyramids [Lazebnik et al., 2006] or object-centric spatial pooling [Russakovsky et al., 2012].

When harder constraints for object deformation are needed to obtain a more distinctive similarity metric, as for instance for human pose recognition [Moeslund et al., 2006], an object can be regarded as a deformed version of a template. This way of modeling non-rigid deformations goes back to the definition of pictorial structures [Fischler and Elschlager, 1973, Felzenszwalb and Huttenlocher, 2005], where "parts" represent local visual properties and "springs" encode spatial relations. Matching the model to an image involves the joint optimization of local appearance similarity as well as the compatibility of the local parts with the spatial model. The object template can also be a statistical model where the likelihoods of landmark deformations are learned from training images, as done by active shape models [Cootes et al., 1995] or their generalization - active appearance models [Cootes et al., 2001] - which additionally model statistics of appearance. These methods have in common that they search for the model parameters that provide the best description of the given data by solving an optimization problem. A similar idea is followed in the traditional graph matching problem [Conte et al., 2004]: the feature points of an image are modeled as a graph and a cost function involving first-order (local feature similarity) and higher-order (regularization) constraints is used to match graphs between images. The output of this cost function can be used as a dissimilarity metric which is discussed in detail in Section 2.4.

For image comparison non-rigid deformations happen to be an additional challenge as recognition schemes using a global transformation model like the RANdom SAmple Consensus (RANSAC) [Fischler and Bolles, 1981] are not applicable. Nonetheless, they occur frequently in computer vision applications. For instance, in ancient coin classification intra-class coin deformations originating from the non-industrial manufacturing of coins need to be mastered (see Section 2.5). An appropriate way of handling non-rigid deformations in visual object comparison is proposed in Section 5.2.

2.1.5 Invariance to Incomplete Image Data

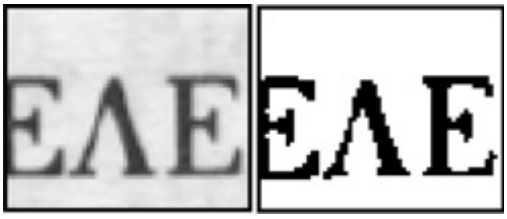
Incomplete image data in the context of recognition and image similarity means that certain parts of the object/image are not visible. Obviously, a general invariant for missing image data does not exist as in the worst case the whole object or the essential image parts are missing. Hence, methods are aiming at being insensitive to missing image data either by explicitly recovering it or by using image representations that are inherently robust to missing image data.

Recovering the missing image regions involves to find the most likely completion of the image data given its context. Image inpainting methods [Bertalmio et al., 2000] estimate the content of lost or deteriorated regions by means of the remaining image content, either locally by continuing the image structures [Tschumperlé and Deriche, 2003], non-locally by filling the regions with fitting exemplars of the whole image area [Criminisi et al., 2004] or by combinations thereof [Arias et al., 2011]. Context information outside the given image has also been leveraged, e.g. nearby video frames [Ling et al., 2011], multiple images of the same physical scene [Agarwala et al., 2004] or large ($>$ 1 million) image collections [Hays and Efros, 2008].

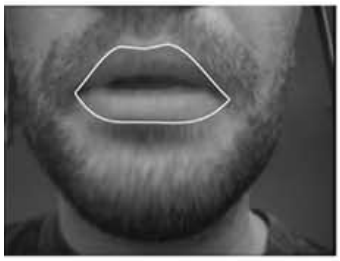
However, for the purpose of image-to-image similarity methods that recover the missing image data are barely helpful, as it is difficult to automatically identify the image regions which demand completion. Therefore, researchers tend to use insensitive image representations instead of explicitly detecting the missing parts. Schmid and Mohr's seminal work [Schmid and Mohr, 1997] stimulated the use of local features for recognition with partial visibility: in contrast to a global image representation, which suffers proportionally from the missing image regions, local representations are more stable as all features in the visible image regions remain unaffected. This is another motivation for using local image representations, in addition to the ones already given in Sections 2.1.1 and 2.1.4.

2.2 Image Segmentation

Image segmentation refers to the process of dividing an image into disjoint regions that belong together based on color, texture or semantic properties [Szeliski, 2010]. According to this broad definition, this computer vision field has a wide area of applications with different levels of complexity. Typically, image segmentation is a preprocessing step for image analysis methods, for example the binarization of document scans for optical character recognition [Gatos et al., 2006], the segmentation of melanoma for automated diagnosis [Celebi et al., 2007] ore more generally the oversegmentation of an image by means of so-called superpixels [Ren and Malik, 2003] to ease further processing.

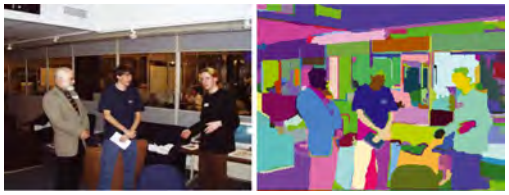


(a) **Thresholding:** binarization of document scans (taken from [Shi et al., 2012])



(b) **Edge-Based Segmentation:** lip tracking (taken from [Eveno et al., 2004])

(c) **Region-Based Segmentation:** color segmentation by mean-shift (taken from [Comaniciu and Meer, 2002])



(d) **Graph-Based Segmentation:** color segmentation by graph-partitioning (taken from [Felzenszwalb and Huttenlocher, 2004])

(e) **Semantic Segmentation:** semantic pixel labelling using harmony potentials (taken from [Boix et al., 2012])

Figure 2.5: Examples of applications and techniques in image segmentation.

Segmentation methods can be classified according to the scope of data they rely on: unsupervised methods use only the given input image for segmentation, whereas supervised methods include a-priori knowledge in form of segmented training images into the segmentation process. From a methodological perspective, the categorization listed below and exemplified in Figure 2.5 can be made.

Thresholding: Thresholding methods belong to the earliest techniques for image segmentation [Otsu, 1975] and automatically define a range of brightness values (the *thresholds*) in the original image. The pixels within this range are selected as belonging to the foreground, whereas the remaining pixels are rejected to the background. The basic assumption of thresholding methods is that the gray levels of the object are significantly different from the gray values of the background. Thresholding techniques can either work globally, where a single threshold is applied to the whole image, or locally, where the image is divided into regions and each re-

gion has its own threshold. Besides that, thresholding techniques differ in the way of finding optimal threshold values for a given image, e.g. by the use of histogram information [Glasbey, 1993, Dong et al., 2008], entropy of gray level distribution [Shanbhag, 1994] or shape information [Shi et al., 2012]. A survey is given in [Sezgin and Sankur, 2004].

Edge-Based Segmentation: This category of segmentation methods partitions an image based on abrupt changes in the intensity, i.e. edges found in an image by edge detectors [Heath et al., 1998]. For the segmentation of parameterizable shapes like circles the Hough transform [Hough, 1962] can be applied. In edge relaxation a global relaxation process based on edge properties is used to form continuous boundaries of objects [Duncan and Birkhölzer, 1992]. Border tracing methods are used to follow the object borders from known starting points [Zhang et al., 2000]. Active contours [Blake and Isard, 1998] like snakes or level sets are boundary detectors which iteratively move towards their final solution, with applications like lip tracking [Eveno et al., 2004] or scene segmentation [Adam et al., 2009].

Region-Based Segmentation: Region-based segmentation methods partition or group regions according to common image properties, like color or texture. Region growing techniques start with seed points that are iteratively increased based on defined region homogeneity criteria [Adams and Bischof, 1994, Mičušík and Hanbury, 2006]. Split-and-Merge [Duarte et al., 2006, Ning et al., 2010] combines two operations to segment an image: splitting, where the image is divided into a set of regions which are coherent within themselves, and merging, where adjacent split regions are merged together based on a similarity criterion. In the watershed segmentation method [Vincent and Soille, 1991] the image is considered as a topographic surface. According to that analogy, the watershed transform finds "catchment basins" and "watershed ridge lines" where the catchment basins theoretically correspond to the homogeneous gray level regions of this image. Clustering-based methods such as mean shift [Comaniciu and Meer, 2002] transform the image pixels into a feature space (e.g., color and position) to find clusters of the image data.

Graph-Based Segmentation: Graph-based methods model the image as graph where the pixels are connected by the graph edges. Every pixel and edge has a cost representing some measure of confidence that the corresponding pixels belong to the same segment. The segmentation goal is then to find an optimal partitioning of the graph. Graph partitioning methods exploited for image segmentation include graph-cuts [Shi and Malik, 2000, Peng et al., 2011], shortest path [Falcão et al., 2004], minimum spanning trees [Felzenszwalb and Huttenlocher, 2004] and random walks [Grady, 2006, Collins et al., 2012].

Semantic Segmentation: Semantic segmentation aims not at the segmentation of image regions for any further processing but rather at a holistic recognition and segmentation process [Arbeláez et al., 2012]. Hence, semantic labels are inferred for each image pixel by joining top-down object knowledge and bottom-up segmentation cues. This line of research has been initialized by the works of [Kumar et al., 2005] and [Borenstein and Ullman, 2008]. More recent methods model the problem as a conditional random field [Boix et al., 2012], refine rectangular object detections [Brox et al., 2011] or classify region proposals [Arbeláez et al., 2012]. The

accuracies (intersection vs. union score [Everingham et al., 2010]) of these methods are in the order of 50 % for a 20-category-problem[1].

The short review of segmentation methods given above shows the variety of applications and complexity of the problem, but also the historical evolution of image segmentation, from the simple thresholding into binary regions to holistic semantic segmentation. Anyhow, which method should be chosen for a given problem should be answered in application-specific manner regarding issues like robustness, accuracy and computational time. Consequently, there is no single state-of-the-art segmentation algorithm which produces the best results for every application.

2.3 Local Image Descriptors

It has been argued in the previous sections that by relying on local parts of the image various aspects of invariance can be handled in a straight-forward manner. However, this requires to use local features which are able to describe the local image appearance in such a way that both a high degree of distinctiveness and insensitivity to noise and other imaging conditions is given. The success of local feature-based methods has been mainly initialized and supported by the introduction of Lowe's Scale Invariant Feature Transform (SIFT) descriptor [Lowe, 1999] which can be regarded as one of the most influential works in computer vision[2]. Since its publication in 1999, SIFT influenced the development of many other local image descriptor, as we see later in this section. Nevertheless, SIFT shows an outstanding performance in comprehensive experimental evaluation papers [Mikolajczyk and Schmid, 2005, Moreels and Perona, 2007] and can still be seen as a powerful general-purpose descriptor which is constantly used in current state-of-the-art work[3].

The computation of image descriptors typically follows the workflow depicted in Figure 2.6. The input is an image patch whose location has been obtained by dense feature sampling or sparse keypoint detection (see Section 2.4) and eventually normalized by the detected scale, orientation or affine transformation [Mikolajczyk and Schmid, 2004]. The input image is first transformed to a more appropriate feature representation where instead of the raw gray values a stack of feature values is computed. For instance, SIFT uses the gradient direction of each image pixel and determines its bilinearly weighted membership to eight equally spaced directions. Next, the descriptor is enriched with spatial information by spatial encoding, e.g. by pooling the features within each cell arranged in a 4×4 grid as in SIFT. The output of this step is a feature vector of numerical values which is finally subject to a set of postprocessing steps. For instance, in SIFT values above a certain threshold are clipped and the feature vector is normalized to unit length to account for brightness changes of the image patch. The final outcome of descriptor

[1] see the results of the PASCAL VOC 2012 segmentation challenge (http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/results/index.html).

[2] according to http://academic.research.microsoft.com/RankList?entitytype=1&topdomainid=2&subdomainid=11&last=0 (accessed on June 8th, 2014), SIFT is the mostly cited computer vision publication of all time with 2229 citations of the original paper [Lowe, 1999] and 7806 citations of the extended journal version [Lowe, 2004].

[3] see for instance recent papers at the 2014 IEEE Conference on Computer Vision and Pattern Recognition such as [Jegou and Zisserman, 2014, Paulin et al., 2014, Trulls et al., 2014].

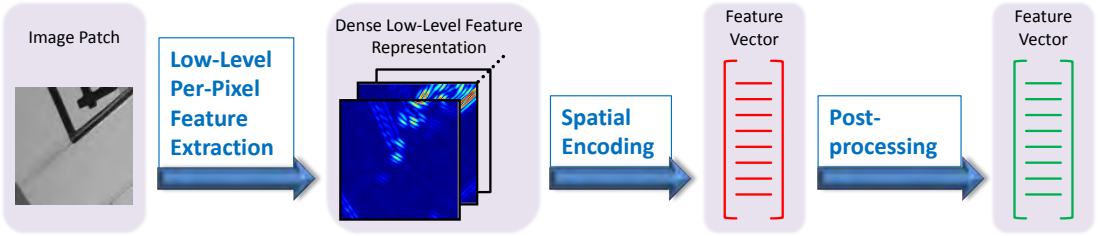


Figure 2.6: The image descriptor construction pipeline.

computation is a feature vector which describes a point in the feature space and the similarity of descriptors is defined by their proximity in this space.

A chronological overview of local image descriptors is given in Table 2.1. Typically, these descriptors are hand-crafted, except for the descriptors BESTDAISY [Brown et al., 2011] and PR [Simonyan et al., 2012] which are the result of research aiming to automatically learn optimal parameters and building blocks of local image descriptors.

2.3.1 Low-Level Per-Pixel Feature Extraction

The first step of descriptor construction is to transform the pixel intensities to a more robust representation of the image structure. Image gradient directions are a popular low-level feature which are leveraged by means of histogram binning to describe their distribution in the image patch, e.g. used by SIFT, DAISY, WLD, GLOH, or MROGH. Another type of low-level features are filter bank responses, e.g. oriented edge filters (Geometric Blur), Haar wavelets (SURF), higher-order derivatives (JBLD) or steerable filters (BESTDAISY). Encoding the local image structures by means of a set of basic patterns has also been proposed, e.g. local binary patterns (CS-LBP, MRRID), local ternary patterns (HRI-CSLTP) or local intensity order patterns (LIOP). These patterns are usually detected on a smaller scale than the overall image patch and the distribution of patterns is encoded as local image descriptor. An ordinal labeling of image intensities is used in the OSID descriptor and the HRI-part of the HRI-CSLTP descriptor.

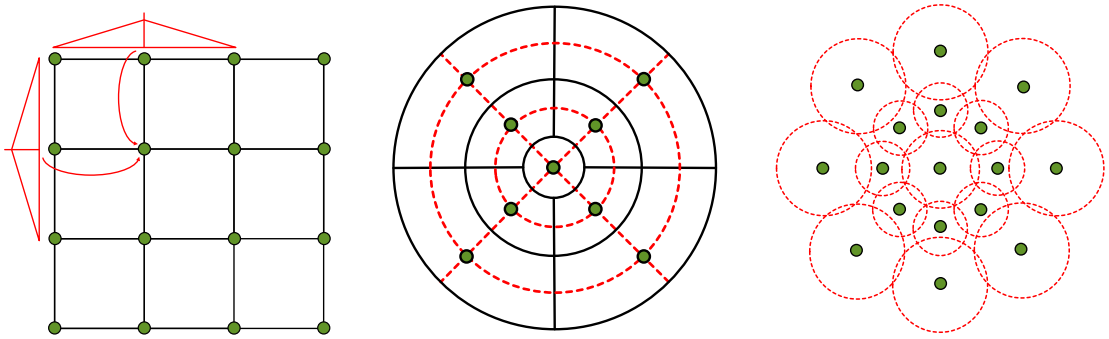
2.3.2 Spatial Encoding

The role of the spatial encoding stage is to describe the spatial configuration of the per-pixel features. For instance, instead of summarizing all features of the overall image patch one can downscale this process to specific subregions in the image patch. This operation is called spatial pooling as the features of one spatial subregion are pooled together, for instance by summation or the max-operation. Similar to the decision of which low-level features are used, the choice of the pooling scheme is application-dependent, as the goal is to find an optimal trade-off between a maximal descriptor distinctiveness and a maximal insensitivity to variations of the spatial distribution caused by inaccurate keypoint detection, varying viewpoints and non-rigid deformations. In the past, several pooling schemes have been proposed of which the most popular ones are depicted in Figure 2.7. For increased robustness, spatial pooling includes also a weighting of

Reference	Descriptor	Low-Level Feature	Spatial Encoding	Feature Vector Post-processing
[Lowe, 1999]	**SIFT**	Binned gradient directions	Spatial pooling (4×4 squared grid cells)	Clipping and L2 normalization
[Berg and Malik, 2001]	**Geometric Blur**	Oriented edge responses	Circularly arranged sample points with increasing amount of smoothing	-
[Belongie et al., 2002]	**Shape Context**	Edge points	Spatial pooling (log-polar grid)	-
[Lazebnik et al., 2003]	**Spin Image**	Binned image intensities	Spatial pooling (concentric rings)	-
[Ke and Sukthankar, 2004]	**PCA-SIFT**	Gradient vectors	-	Dimensionality reduction by Principal Component Analysis (PCA)
[Mikolajczyk and Schmid, 2005]	**GLOH**	Binned gradient directions	Spatial pooling (log-polar grid)	Dimensionality reduction by PCA
[Bay et al., 2006]	**SURF**	Haar wavelet responses	Spatial pooling (4×4 squared grid cells)	L2 normalization
[Shechtman and Irani, 2007]	**SSDESC**	L*a*b* colors	Similarity to center pixel on log-polar grid	Transformation of values to range $[0, 1]$
[Tola et al., 2008]	**DAISY**	Binned gradient directions	Spatial pooling (circularly arranged cells)	L2 normalization
[Kobayashi and Otsu, 2008]	**GLAC**	Binned gradient directions	Correlation of neighboring histograms, spatial pooling (4×5 squared grid cells)	Clipping and L2 normalization
[Chen et al., 2008]	**WLD**	Binned differential excitations and gradient directions	-	2D to 1D histogram transformation
[Heikkilä et al., 2009]	**CS-LBP**	Local Binary Pattern	Spatial pooling (4×4 squared grid cells)	Clipping and L2 normalization
[Tang et al., 2009]	**OSID**	Ordinal labeling	Spatial pooling (16 pie cells)	Clipping and L2 normalization
[Gupta et al., 2010]	**HRI-CSLTP**	Ordinal labeling and local ternary patterns	Spatial pooling (4×4 squared grid cells)	-
[Calonder et al., 2010]	**BRIEF**	Original image intensities	Pairwise intensity comparisons between random pixel locations	-
[Brown et al., 2011]	**BEST-DAISY**	Steerable filter responses	Spatial pooling (circularly arranged cells)	Clipping and L2 normalization
[Leutenegger et al., 2011]	**BRISK**	Gaussian smoothed image intensities	Pairwise intensity comparisons between circularly arranged pixel locations	-
[Wang et al., 2011]	**LIOP**	Local intensity order patterns	Spatial pooling (intensity order regions)	-
[Alahi et al., 2012]	**FREAK**	Gaussian smoothed image intensities	Pairwise intensity comparisons on retinal sampling grid	-
[Fan et al., 2012]	**MROGH**	Binned gradient directions	Spatial pooling (intensity order regions)	L2 normalization
[Fan et al., 2012]	**MRRID**	Local Binary Pattern	Spatial pooling (intensity order regions)	L2 normalization
[Larsen et al., 2012]	**JBLD**	Higher-order derivative filter responses	Spatial sampling ($n \times n$ squared grid cells)	Whitening and L2 normalization
[Simonyan et al., 2012]	**PR**	Binned gradient directions	Spatial pooling (circularly arranged cells)	Dimensionality reduction by linear projection
[Seidenari et al., 2014]	**P-SIFT**	Binned gradient directions	Spatial pooling (multiple layers of squared grid cells with increasing cell sizes)	Clipping and L2 normalization

Table 2.1: Chronological overview of image descriptors proposed in literature.

features where the contribution of a feature near the region center is higher than the contribution of a feature further away from the center.



(a) Regular grid with squared cells and bilinear weighting.

(b) Log-polar grid with bilinear radial and angular weighting.

(c) Circularly arranged cells with Gaussian weighting.

Figure 2.7: Examples of pooling schemes (adapted from [Brown et al., 2011]). Green points depict the centers of the spatial cells and weighting is illustrated in red.

The squared 4×4 grid used in SIFT (Figure 2.7a) is adopted by other descriptors like SURF, OSID, GLAC, CSLBP, and HRI-CSLTP. Shape Context and GLOH use a log-polar grid for spatial pooling (Figure 2.7b). Similarly, the DAISY descriptor uses circular cells of varying size arranged on concentric rings for spatial pooling (Figure 2.7c) and [Brown et al., 2011] propose a learning scheme to obtain optimal configurations of such cell arrangements. MROGH, MRRID and LIOP do not use fixed spatial cells but define cell locations by tiling the image patch based on pixel-intensity orders. This has the advantage that the resulting descriptor is inherently invariant to image rotations.

Another way of encoding spatial information is to describe the spatial correlation between defined image regions. Shechtman and Irani [Shechtman and Irani, 2007] pioneered this idea with the Self-Similarity DESCriptor (SSDESC) in order to represent the structure of an image patch independently from its color or texture. For this purpose, the color similarities between the center region and the surrounding regions located on a log-polar grid are measured. The GLAC descriptor also exploits correlation but uses gradient directions instead of color values whose local correlation patterns are measured and histogrammed. The concept of spatial correlation is also used by binary descriptors such as BRIEF, BRISK or FREAK which describe an image patch by comparing its image intensities at specified locations. Each comparison can be quantized to 1 bit which enables a compact description for scenarios with limited hardware resources (e.g. a mobile phone), as binary descriptors can be computed and matched faster and need less storage than traditional descriptors [Calonder et al., 2010].

2.3.3 Feature Vector Postprocessing

Postprocessing steps are applied to the feature vector in order to increase its robustness and/or decrease its dimensionality. Histogram-based descriptors, which compute a joint 2D histogram of feature values and their locations, typically weight the histogram values by their saliency in

the image patch, e.g. gradient directions are weighted by the gradient magnitude. In order to reduce the influence of relatively high saliency values and to be invariant to contrast changes the feature vector values above a certain threshold are clipped and the final feature vector is normalized to unit length (e.g. SIFT, GLAC, OSID, BESTDAISY). Dimensionality reduction techniques are applied to reduce the effects of the "curse of dimensionality" [Aggarwal et al., 2001] and to save storage space, e.g. by PCA [Ke and Sukthankar, 2004], Linear Discriminant Analysis (LDA) [Brown et al., 2011] or learned linear projections (PR) [Simonyan et al., 2012].

2.3.4 Illumination Insensitivity of Local Descriptors

Image descriptors are developed based on specific requirements, e.g. MROGH and MRRID were proposed for improved rotation insensitivity and binary descriptors for low computational effort. However, when it comes to illumination insensitivity of local image descriptors, there is a lack of research both on the methodological and the evaluation level. The Oxford dataset presented in [Mikolajczyk and Schmid, 2005] covers illumination insensitivity by means of 6 images of one single scene taken with different camera apertures (see Figure 2.8a), and this evaluation image data is also used by others to test the illumination insensitivity of their descriptors (e.g. BRIEF, BRISK, FREAK, LIOP, MROGH, SURF, CSLBP, OSID, HRI-CSLTP). However, changing the camera aperture leads only to global brightness changes in the image and more complex local effects on 3D objects as shown in Figure 2.2b-d are not considered. Some evaluations are additionally conducted on more image data (LIOP, MROGH, MRRID, CSLBP, OSID), but still only global brightness changes are simulated. A more comprehensive evaluation of illumination insensitivity is given in [Moreels and Perona, 2007]. Their study uses images with three different lighting conditions (see Figure 2.8b) from 100 objects. The experiments couple the descriptor performance evaluation with the interest point detection step and reveal that the combination of Harris-affine detector and SIFT descriptor performs best. However, the test data and evaluation protocol does not allow to restrict the experiments of descriptor performance on textured/textureless objects. Moreover, the study is from 2007 and modern descriptors are therefore not included.

Due to the fixation of treating illumination insensitivity as a contrast normalization problem induced by the commonly used Oxford dataset [Mikolajczyk and Schmid, 2005], current descriptors are generally only invariant to illumination changes on flat, textured objects (see Figure 2.2). Hence, they assume that illumination changes have only a global linear effect on the intensity values which can be compensated by feature vector normalization (SIFT, SURF, DAISY, GLAC, CS-LBP, BESTDAISY, JBLD, P-SIFT). The descriptors OSID, LIOP, MRRID and HRI-CSLTP use the relative order of pixel intensities to construct descriptors invariant to the wider class of generally monotonic intensity changes, but still complex illumination changes on 3D objects are not handled.

2.4 Correspondence-Based Image Similarity

Comparing images by means of sets of local features has benefits in terms of invariance: translation invariance is inherently achieved and rotation-, scale- and illumination-insensitivity can be obtained by using appropriate local detectors and descriptors. Moreover, the locality pro-

(a) Oxford dataset [Mikolajczyk and Schmid, 2005]: 6 shots with different camera apertures.

(b) 3D object dataset [Moreels and Perona, 2007]: 3 shots with different illumination conditions for a given camera angle.

Figure 2.8: Previously used datasets for the evaluation of illumination insensitivity of local image descriptors.

vides robustness to missing image data and background clutter, as only the local features in the disturbing image regions are affected.

Therefore, modern approaches for complex real world recognition scenarios with a high degree of variability (e.g. scene classification, object recognition) are commonly based on local features[4]. However, describing an image by a set of features (i.e. a set of fixed-size vectors) hinders direct image comparison, as images produce a variable number of unordered features. Hence, establishing correspondences is the core step for revealing similarities between images: if many parts in an image can be associated with similar-looking parts in another image, they are likely to show similar content.

The output of local feature extraction are image coordinates $\mathbf{p}_i$ describing the spatial location of the features as well as a vector $\mathbf{d}_i$ for each feature location describing its local appearance. The spatial locations are either determined by interest point detectors such as DoG [Lowe, 2004] or Harris-Laplace [Mikolajczyk and Schmid, 2004] to identify the most salient regions in an image

[4] see for instance recent papers on recognition presented at the 2014 IEEE Conference on Computer Vision and Pattern Recognition such as [Jegou and Zisserman, 2014, Lai et al., 2014, Xie et al., 2014, Yao et al., 2014].

or by taking regular samples (this is often referred to *dense* vs. *sparse* sampling, as the former typically produces fewer interest points and features [Nowak et al., 2006]). For determining the correspondences and estimating image similarities the information about feature location can either be ignored or used to guide this process by means of geometric constraints.

2.4.1 Without Geometric Constraints

Ignoring the locations and geometric dependencies of the features has the advantage that geometric deformations between images have not to be taken into account for comparison and a wide range of geometric invariances is covered. Hence, the remaining questions are how to detect the corresponding features and how can this information be used to derive an image similarity measure? As the features themselves have a fixed size they can easily be compared by vector distance metrics such as L_2 [Lowe, 2004] or χ^2 [Belongie et al., 2002]. More sophisticated metrics for non-aligned vector values such as the Earth Mover's Distance [Rabin et al., 2008, Pele and Werman, 2009] have also been proposed.

Given such a metric, correspondences can be established, either in a *one-to-many* scheme, where a feature can have a correspondence with multiple other features, or in a *one-to-one* scheme, where each feature is allowed only to have a correspondence with one feature from the other set. The simplest one-to-many matching scheme is to assign each feature to its nearest neighbor in the other feature set. However, [Lowe, 2004] claims that this is a error-prone procedure when similar features are present in the two images and suggests to accept a correspondence only if the distance ratio from the nearest to the second nearest neighbor is under a given threshold. The Hungarian algorithm [Kuhn, 1955] used in [Kumar et al., 2001, Belongie et al., 2002] finds the optimal one-to-one correspondences between feature sets but has the disadvantage that outliers are not rejected and interfere with the matching process. Similar approaches with more robustness to outliers are proposed by [Scott and Longuet-Higgins, 1991] and [Gold and Rangarajan, 1996]. The one-to-one symmetric search proposed by [Zhao et al., 2007] accepts correspondences only if a feature in the first image is the nearest neighbor of a feature in the second image and vice versa.

Correspondences of local features are also used as similarity measures in the form of kernel functions to make use of kernel-based classifiers like the Support Vector Machine (SVM) [Cortes and Vapnik, 1995]. The match kernel [Wallraven et al., 2003] uses the average vector distance of the optimal one-to-many correspondences whereas the pyramid match kernel [Grauman and Darrell, 2005] uses an approximation of the sum of vector distances of the optimal one-to-one correspondences.

A prominently used methodology to describe and compare images by local features without considering their geometric arrangement is the BOVW model [Csurka et al., 2004] which has shown to achieve state of the art performance on benchmarks like the PASCAL Visual Object Classes Challenge 2012[5] [Dong et al., 2013] or the ImageCLEF 2013 Photo Annotation Task[6] [Grana et al., 2013]. In this approach, correspondences are not established between images but between images and a vocabulary of codewords for sparse feature coding: vector distance

[5] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/ (accessed on June 8th, 2014).

[6] http://www.imageclef.org/2013/photo (accessed on June 8th, 2014).

metrics between local features of the image and codewords are computed to obtain a coding vector for each local feature which indicates the active codewords. Statistics such as the histogram of activated codewords are then used as the final image representation and can be used for image comparison. The power of this methodology lies in the compact description of images by means of the most relevant image structures and its invariance to geometric variations, as feature locations are not considered. However, neglecting spatial information also limits the discriminative power of the model as the appearance of local parts alone might be ambiguous and does not give enough information about the object. Therefore, correspondence-based image similarity can be enriched with the spatial information about the extracted local features, as discussed in the next section.

2.4.2 With Geometric Constraints

Adding geometric constraints helps to identify false correspondences and thus to make the similarity measure more reliable. A possible constraint is to assume a rough alignment of image structures and to compare features only for the same image region. In the context of BOVW this concept is known as *spatial pooling* [Boureau et al., 2010] and shows to improve the recognition rate of BOVW for scene categories [Lazebnik et al., 2006, Jia et al., 2012], object categories [Zhou et al., 2010b] or characters [Jia et al., 2012]. Other researchers pursue a more local encoding of spatial information by means of co-occurrence statistics of visual words [Agarwal and Triggs, 2006, Savarese et al., 2006, Arandjelović, 2010].

Geometric constraints can also be applied to feature matching as a postprocessing step. The task of this *geometric verification* step is to check the geometric consistency of the initial correspondences which are determined from feature vector distances as described in Section 2.4.1. A possible verification constraint is that the correspondences follow a common global geometric transformation. The RANdom SAmple Consensus (RANSAC) [Fischler and Bolles, 1981] scheme repeatedly takes random subselections of the feature correspondences and checks how many correspondences support the global transformation estimated from the chosen samples. The total number of these so-called inliers serves as a measure of trust if the estimated image transformation represents the true transformation between the two images. As the false matches (outliers) can be assumed to be randomly distributed and do not follow a common global transformation, they are effectively ruled out. RANSAC or weaker geometric verification constraints are widely used in the literature, e.g. by [Sivic and Zisserman, 2003, Lowe, 2004, Philbin et al., 2007, Jegou et al., 2008, Dreuw et al., 2009, Wu et al., 2009, Zhou et al., 2010a]. This shows that geometric verification can be a strong tool in correspondence-based image similarity, and even simple similarity measures like the number of matched features after geometric verification have been effectively used for tasks like face recognition [Dreuw et al., 2009] or duplicate image search [Zhou et al., 2010a].

However, geometric verification cannot be easily adapted to non-rigid deformations, as every corresponding feature pair is defined by its own local transformation and not by a common global transformation. Hence, for non-rigid deformations only relative geometric constraints can be used, such as that two neighboring points will likely have neighboring correspondences in the other images. In computer vision such a constraint is typically cast as a graph matching problem [Conte et al., 2004], as illustrated in Figure 2.9: a cost function consisting of first-order

(local feature similarity) and second-order (regularization) constraints is defined and solved by numerical optimization [Velho et al., 2011]. Graph matching can be used to find correspondences between images (e.g. the frames of a video stream), but the output of the cost function has also been used in the past as similarity metric for image comparison as it can be assumed that the costs of matching similar objects are higher than that of matching dissimilar objects [Berg et al., 2005, Duchenne et al., 2011b, Jorstad et al., 2011, Liu et al., 2011, Kim et al., 2013]. These methods show superior performance in recognition scenarios with non-rigid intra-class deformations and low number of training samples like face recognition [Jorstad et al., 2011, Liu et al., 2011]. The reason is that for low-number of training samples the variability is better handled "online" during image matching than "offline" by machine learning.

The second-order constraints of graph matching have the task of regularizing the matching process by assessing the pairwise costs of matching two features in an images connected by a edge of the graph. For instance, the matching schemes proposed in [Berg et al., 2005, Leordeanu and Hebert, 2005, Torresani et al., 2013] are regularized by a weighted sum of edge length difference and edge angle difference. The dense matching schemes proposed by [Liu et al., 2011, Kim et al., 2013] penalize differences in neighboring pixel correspondences as well as global displacements. The model of [Duchenne et al., 2011b] enforces smoothness as well as monotonicity of matched features.

Using such regularization constraints does not make the image comparison truly invariant to non-rigid deformations, as for instance the distance between two features can vary in the two images. However, under the assumption that the deformation is smooth a reasonable trade-off between insensitivity to non-rigid deformations and discriminability can be achieved. For the case where global deformations have to be taken into account also higher-order constraints can be used. For instance, [Chertok and Keller, 2010, Duchenne et al., 2011a, Cheng et al., 2013] use the inner angles of triples of feature points to obtain a regularization which is locally invariant

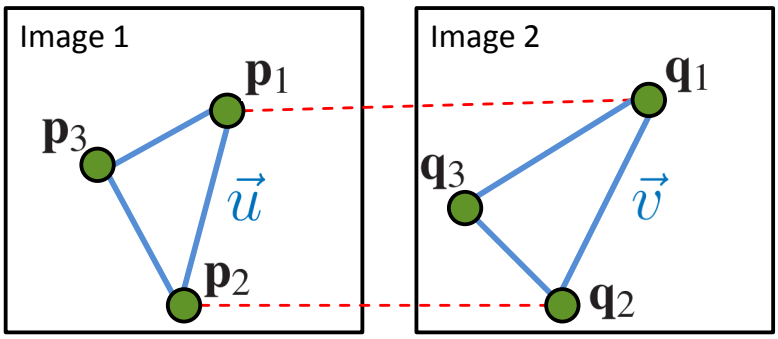


Figure 2.9: Using graph matching to determine the correspondences between the local features extracted from two images. The detected feature points represent the vertices of the graph connected by edges. Matching two point pairs, e.g. $\mathbf{p}_1$ with $\mathbf{q}_1$ and $\mathbf{p}_2$ with $\mathbf{q}_2$, produces first-order costs (difference of the feature descriptions, dotted red lines) as well as second-order regularization costs (difference of the vectors $\vec{u}$ and $\vec{v}$). Higher-order costs (e.g. the difference of the matched triangles $\mathbf{p}_1\mathbf{p}_2\mathbf{p}_3$ and $\mathbf{q}_1\mathbf{q}_2\mathbf{q}_3$) can also be used. The overall matching costs are minimized to obtain the optimal correspondences.

to a similarity transformation, i.e. rotation and scale. Projective invariance can be achieved by using the cross-ratios of points along the edges of the triples [Duchenne et al., 2011a].

2.5 Image-Based Coin Analysis

In this section an introduction to the field of image-based coin analysis is given. The practical focus of the image comparison research conducted in this thesis lies on ancient coins and thus an overview of ancient coinage and a definition of the most relevant numismatic terms is given in Section 2.5.1. In Section 2.5.2 the specific challenges of ancient coin analysis compared to present-day coins are specified. Finally, related work in the fields of coin image segmentation (Section 2.5.3) and coin image classification (Section 2.5.4) is reviewed.

2.5.1 Introduction to Ancient Coinage

The production of coins has begun in the 7th century BC in Greece and has become the central embodiment of money in this as well as other ancient cultures like the Roman empire, Byzantium, India or China [Grierson, 1975]. Like their present-day counterparts, ancient coins have a front and back side, which are referred to as *obverse* and *reverse* side [Jones, 1990]. The appearance of coins has been designed by artisans known as engravers who manufactured the *dies* for the obverse and reverse coins sides. For coin production a metal flan was placed between the two dies: the obverse die was held stationary on an anvil while the reverse die was placed over the flan and struck with a hammer (see Figure 2.10).

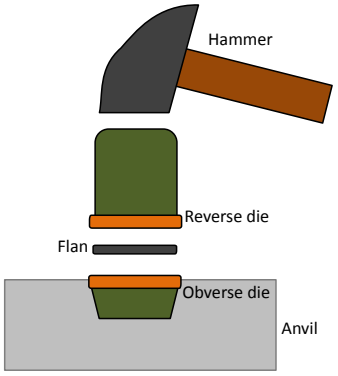


Figure 2.10: The striking of an ancient coin.

The basic elements of ancient coin design are *type*, *inscription* or *legend*, and *accessory symbols* [Grierson, 1975], as depicted in Figure 2.11 for a Roman Republican coin. The type is the central person, object or device represented on a coin. The legend is the writing placed upon a coin in order to give information about the person(s) by whose authority the coin was minted, to describe the shown type or to convey a general message. Accessory symbols are smaller features like minting date or control marks. Anyhow, none of the basic elements are necessarily present on an ancient coin.

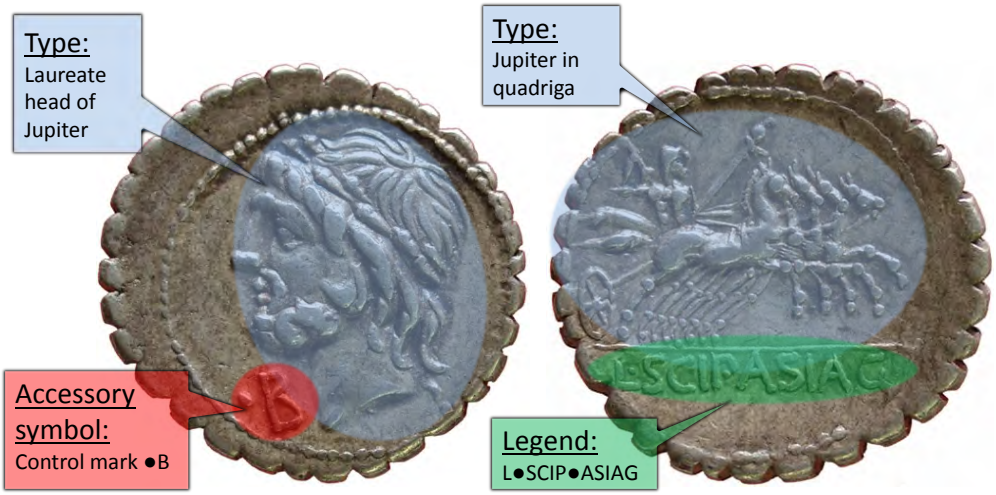


Figure 2.11: An example for the basic coin elements on the obverse (left) and reverse side (right) of a Roman Republican Denarius.

(a) Cra 344/1 (b) Cra 344/3 (c) Cra 410/1

Figure 2.12: Three Roman Republican coin classes with Crawford number.

Within this thesis the coin image datasets contain Roman Republican coins, as there exists a clear definition of coin classes based on Crawford's reference book [Crawford, 1974]. Hence, the task of automatic classification is to assign a coin to the correct number in this reference book. A coin class is basically defined as a distinct combination of coin type and legend and Crawford defined classes as well as subclasses, i.e. a complete reference is given by the class number and subclass number separated by a '/'. For instance, the coins shown in Figure 2.12a and 2.12b are from the same general class (344) but from different subclasses (/1 and /3). The obverse of both coins shows the head of the moneyer Lucius Titurius Sabinus while the reverse of 344/1 shows the abduction of the Sabine women and the one of 344/3 the goddess Victory in a 2-horse chariot. The coin in Figure 2.12c is from a different class and shows the diademed head of Apollo on the obverse and a standing Hercules on the reverse. In total, Crawford defined 550 main classes and over 1900 subclasses. The examples of Figure 2.12 demonstrate the common design of Roman Republican coin types [Jones, 1990]: the obverse displays the portrait of a historical or mythological person while the reverse depicts certain scenes or objects. Consequently, the reverse side shows more variation between classes and thus gives more discriminative information for classification.

(a) The difference between modern (top) and ancient coins (bottom)

(b) Challenges for coin segmentation

(c) Challenges for coin recognition: illumination changes.

(d) Challenges for coin recognition: non-rigid deformations.

(e) Challenges for coin recognition: abrasions.

Figure 2.13: The various challenges of image-based ancient coin recognition.

2.5.2 The Challenges of Image-Based Ancient Coin Analysis

In contrast to present-day coins, ancient coins offer particular challenges to image-based recognition, as shown in Figure 2.13a. For the automatic segmentation of coins, difficulties are posed by the non-roundness of coins, shadows at the coin borders and background clutter. These issues are demonstrated in Figure 2.13b: ancient coins are not necessarily completely circular (shown by the red circle around the coin), thus simple geometric structure detection methods like the Hough transform [Hough, 1962] are not applicable. An improper image acquisition setup can lead to shadow casts at coin borders which deteriorate their correct identification. And finally, background objects like a ruler complicate the detection of the coin object in the image.

For the classification of ancient coins, the complexity of the problem is given by the high number of classes (e.g. over 1900 subclasses for the Roman Republican age [Crawford, 1974]). Furthermore, all the kinds of variations illustrated in the introductory Figure 1.1 arise, as de-

scribed also in [Zambanini and Kampel, 2011]:

- **Illumination**: illumination variations significantly affect the surface appearance and hence make coin comparison difficult. With reference to Figure 2.2, coins with their metallic relief-like structures belong to the challenging group of shiny, textureless objects. An example for the illumination variation is given in Figure 2.13c, where two images of the same coin specimen under different lighting conditions are shown. A detailed look at the man's upper body reveals how the local appearance is affected by the lighting direction.
- **Non-rigid deformations**: Non-rigid deformations within a coin class result from the manual manufacturing of the coins. In ancient times, the manually made coin dies were used only for a limited amount of coins. Therefore, die variations and thus finally coin variations as the ones shown in Figure 2.13d can occur. Here a detail of two Roman coins from the same class depicting the goddess Roma is shown. The green arrows illustrate that the facial features have a different arrangement due to the different dies used for striking these coin specimens.
- **Incomplete data**. Parts of the visual information on a coin are possibly lost or distorted due to abrasions, caused both by use and by exposure to environmental influences like chemicals in the soil (see Figure 2.13e).

All these particular challenges sum up to a demanding relationship between intra-class and inter-class variations, as demonstrated in Figure 2.14. One the one hand, inter-class variations can be low: the coins of Figure 2.14a-b all show the goddess Calliope and the two classes can only be differentiated by the coin legend. On the other hand, non-rigid intra-class deformations due to the non-industrial manufacturing can be spotted within the classes (e.g., Calliope's hand and the legend structure in Figure 2.14b). The conditions of the coins lead to further intra-class variations: missing and worn parts due to the coins' age (Figure 2.14c) as well as appearance variations due to lighting variations (e.g., the two coins in Figure 2.14d are illuminated from opposite directions).

2.5.3 Coin Image Segmentation

Separation of the coin from its background is done in any coin recognition system to obtain a region-of-interest where the recognition procedure can be applied to. For systems designed for the recognition of modern coins the problem is simplified by the circularity of the coin and a controlled acquisition setup. For instance, the acquisition setup of the Dagobert coin recognition and sorting system [Nölle et al., 2003] ensures that the background (conveyor belt) is darker than the photographed coin and thus simple global thresholding is sufficient. Circularity of modern coins is exploited in [Reisert et al., 2006] by means of the Generalized Hough Transform [Ballard, 1981]. The segmentation step of [Van Der Maaten and Poon, 2006] uses a pipeline of global thresholding, edge detection and morphological operations.

Due to the higher complexity of the segmentation problem on ancient coins (see Figure 2.13b), researchers developed own segmentation strategies for ancient coin recognition methods that are more appropriate for this kind of objects. The first recognition system dedicated exclusively to ancient coins is presented by [Zaharieva et al., 2007a] and uses a modification of

(a) Cra 410/3 (b) Cra 442/1

(c) Cra 346/1 (d) Cra 393/1

Figure 2.14: Reverse side of Roman Republican coins from four different coin classes.

the adaptive thresholding approach originally proposed by [Yanowitz and Bruckstein, 1989] for coin segmentation. The method uses variable thresholds which are computed by sampling points at detected edges and interpolating them over the image. The modified version of [Zaharieva et al., 2007a] uses zero crossings of the second image derivatives [Marr and Hildreth, 1980] instead of gradient magnitudes for edge detection. The coin classification method presented by [Arandjelović, 2010] uses an improved circular Hough transform [Atherton and Kerbyson, 1999] for coarse coin location. Finally, the accurate coin border is found by inferring the states of a hidden Markov chain [Blake et al., 2011] representing the radial distances at uniformly discretized directions.

The method proposed in Chapter 3 for ancient coin segmentation uses global thresholding and exploits the approximately circular shape of coins as a cue to find an optimal threshold. Hence, the method remains simple and fast to compute, as no costly boundary tracing or optimization like in [Arandjelović, 2010] is needed. Still, the proposed method shows to be insensitive to wide variations of coin images in the experiments.

2.5.4 Image-Based Coin Classification

The first methods for image-based coin classification have been developed for modern coins, hence they all assume rigidity of objects and thus that an alignment of the coin structures can be achieved by solving for the global translation and rotation differences between a query coin and reference coins. The Dagobert system [Nölle et al., 2003] uses global rotation-invariant features

extracted from the detected coin to derive a list of subselected candidate reference coins. The image similarity to the candidate coin images is then established by finding the maximum correlation of the query edge image and reference images under rotation. The method is evaluated on 913 coin classes from over 30 currencies and achieves a classification rate of 99.24 %. Compared to this results, the multistage classifier based on eigenspaces proposed in [Huber et al., 2005] shows a lower performance with a classification rate of 92.23 % on a similar dataset. Motivated by the effectivity the Dagobert system, [Reisert et al., 2006] propose to use the Fast Fourier Transform on binary images encoding the pixels' gradient directions to register coin images and estimate their similarity. The method was the winner of the MUSCLE CIS Coin Competition 2006 [Nölle et al., 2006] with a classification rate of 97.24 % on a dataset of 2270 modern coin classes. It outperformed the COIN-O-MATIC system proposed by [Van Der Maaten and Poon, 2006] which uses a global descriptor of edge distributions on a log-polar grid (see Figure 2.7b). Since the descriptor summates all edges in a spatial cell it does not make full use of the rigidity of modern coins. In contrast, the method of [Reisert et al., 2006] assesses coin similarity on a per-pixel basis and is thus more distinctive.

As described in Section 2.5.2, ancient coins show a higher intra-class variability than modern coins, which permits the reverse conclusion that each specimen has its own individual characteristics that allow for coin identification. This task is important in the context of finding stolen coins in the internet as highlighted by [Huber-Mörk et al., 2011]. Their experiments reveal that the segmented coin border is a strong characteristic feature which can be leveraged for identification by shape matching. In combination with type matching on both coin sides 98.83 % of 2400 coin images representing 240 individual coin specimens could be identified. However, evidently the coin border does not assist in classifying the coin, hence it can be concluded that classification is a more difficult task than identification on ancient coins. The level of object individuality is in general strongly correlated with the complexity of classification, and consequently on modern coins the opposite case is true: classification is less complex than identification.

It is experimentally shown by [Zaharieva et al., 2007b] that the success of classification methods for modern coins cannot be transferred to the domain of ancient coins. While the method of [Van Der Maaten and Poon, 2006] achieves a classification rate of up to 76 % on their testset of 100 modern coin classes, only 6 % of coins in a testset of 106 Roman Imperial coin classes can be correctly classified. The first method exclusively dedicated to ancient coins is proposed by [Kampel and Zaharieva, 2008]. The non-rigid deformations of ancient coins are handled by using detected local SIFT features and a similarity measure is simply obtained by counting the number of features that can be matched by means of Lowe's distance-ratio rule (see Section 2.4.1). In their experiments this similarity measure is used for exemplar-based classification and achieves a classification rate of 90 %, but only on a very limited dataset of only three coin classes.

Learning-based methods for ancient coin classification are proposed by [Anwar et al., 2013] and [Arandjelović, 2010]. Both methods also rely on local SIFT features which are quantized into a fixed vocabulary of visual words. In [Anwar et al., 2013] the image is tiled into spatial regions and the concatenated single histograms of visual words of each region are used as image feature. This approach is rather used for a coarse classification of the types shown on the coin's reverse side than for a fine-grained classification based on reference numbers. The method achieves a classification rate of up to 90% for eight common Roman Republican types.

Arandjelović's method [Arandjelović, 2010] exploits the spatial configuration of visual words in a different way: Locally-Biased Directional Histograms (LBDH) are introduced for encoding the distribution of visual words around a detected keypoint in eight directions relative to its canonical orientation. The LBDH features are then again subject to vocabulary creation and the histogram of LBDH words serves as final image feature. This method achieves a classification rate of around 57% on 65 classes of the Roman Imperial age.

Learning-based methods are also exploited to support coin classification by means of legend recognition. [Arandjelović, 2012] describes a system which subselects Roman Imperial coin classes based on the recognized legends on the obverse sides. The system assumes that the legend is arranged along the coin border and thus can be normalized for orientation by means of a log-polar transformation. Obviously, known legend orientation eases the problem and the author reports a correct legend recognition for 24 of 25 test coins with a lexicon of 1478 known coin legends. This is achieved by encoding letter appearance by single local image descriptors and casting the search for words in the lexicon as a weighted-graph optimization problem which can be efficiently solved via dynamic programming [Szeliski, 2010]. The same approach is followed by Kavelar et al. [Kavelar et al., 2012, Kavelar et al., 2014], but on Roman Republican coins which requires to make the local descriptors and word search procedure orientation invariant. Consequently, the results are worse compared to [Arandjelović, 2012] with a detection rate of 53 % among the top five words found for a 35-word-lexicon [Kavelar et al., 2014]. Nevertheless, it is shown in [Zambanini et al., 2013] that legend recognition is still able to improve performance, as it exploits a different source of background information by means of a known lexicon. In [Zambanini et al., 2013] the exemplar-based classification proposed in Section 5.1.3 is fused with legend recognition which improves the classification rate from 78.9% to 81.0% for a dataset of 60 Roman Republican coin classes. Also the method of [Arandjelović, 2012] combines legend recognition with exemplar-based matching: global translation differences between the reverse coin types are first corrected by RANSAC-based registration of SIFT features. The detected SIFT features are then checked for consistent location, orientation and scale and the total number of verified correspondences is used as similarity measure.

2.6 Summary and Innovative Aspects of the Thesis

Invariance or insensitivity to certain conditions of the objects or the imaging procedure can be treated in manifold ways. Leveraging a-priori information to handle variations has led to successful methods, but with limited real world practicability due to their dependency on a high amount of training data. Without using offline training, recognition is reduced to an image similarity estimation problem, where the determination of local correspondences has proven to be the common methodology, due to the beneficial properties of comparing many local parts instead of the images in their entirety: translation, rotation and scale invariance by means of interest point detection with rotation and scale selection, the opportunity to flexibly match local parts in case of non-rigid deformations, and robustness against image clutter and missing content. Nevertheless, there are aspects in this image-to-image comparison problem which have not been tackled accordingly in existing works and are treated in this thesis:

- **Shape-Controlled Object Segmentation**: invariance to scale changes and image back-

ground clutter can be achieved by interest point detection, but this is an error-prone process, especially for textureless objects like ancient coins under illumination changes. Segmentation of the object of interest does not only select a proper image subregion for the visual comparison, but also allows for a scale normalization and dense sampling of local descriptors in order to have a more robust correspondence-based comparison. The manifold research in the wide area of image segmentation shows that application-specific solutions work best and commonly dominate general-purpose segmentation methods within a given field. Hence, in this thesis a method is proposed which exploits the known nearly circular shape of coins to achieve a fast, robust and accurate segmentation.

- **Illumination-Insensitive Feature Extraction**: Establishing correspondences between images demands both distinctive and insensitive local descriptors, but insensitivity to changes of the illumination conditions has only been marginally treated. Existing descriptors do only handle monotonic brightness changes but not the more complex effects on textureless and non-flat objects. Hence, there is a need for robust image descriptors in scenarios where strong illumination changes can be expected and texturedness as well as flatness of objects are not necessarily given. This shortcoming is tackled in this thesis by means of a fundamental evaluation of the illumination insensitivity of low-level per-pixel image features. Based on the achieved results, a new descriptor (LIDRIC) with a higher degree of illumination-insensitivity is proposed. The problem of unsatisfactory datasets and evaluation protocols for this kind of problem is overcome by new datasets allowing to intensively test the influence of textured/textureless 3D objects on descriptor performance.

- **Correspondence-Based Image Similarity**: in the past, the correspondences of local image parts have been exploited to derive similarity measures between images. A contribution to this research is given by means of a coarse-to-fine classification scheme which shows that a subselection of reference objects on coarser scales does not decrease classification performance while substantially speeding up the overall process. The method uses a dense correspondence search with spatial regularization and stimulates the construction of a more powerful correspondence-based similarity. Similar to geometric verification, the correspondences are checked for geometric plausibility, but in this case not to discern true from false correspondences, but rather to derive a similarity measure from it assuming a higher fraction of geometric plausible correspondences between similar images than between dissimilar ones.

- **Image-Based Coin Classification**: ancient coins prove to be a challenging type of objects for image classification and learning-based approaches [Csurka et al., 2004, Arandjelović, 2010] suffer from the training data problem given for this domain. Previously published similarity measures that can be exploited for exemplar-based classification either ignore spatial information and are thus not distinctive enough [Kampel and Zaharieva, 2008] or are too strict by assuming a rigid-body transformation between coins [Arandjelović, 2012]. Therefore, it is first shown that a correspondence search with spatial optimization provides a better framework for similarity estimation for the non-rigid deformations of ancient coins. On this basis, an improved similarity measure with lower computational

complexity is proposed which uses pairwise geometric consistency evaluation of matched features and consequently outperforms the previously proposed methods.

CHAPTER 3

Shape-Controlled Object Segmentation

In this chapter the method for ancient coin segmentation originally published in [Zambanini and Kampel, 2009] is described. Its goal is to segment the image in two areas: the area depicting the coin and the area belonging to the background. Within the context of image-based coin analysis, such a segmentation answers a twofold purpose: first, for coin classification, it provides a region-of-interest and allows for scale normalization. The proposed segmentation procedure has thus been applied as a preprocessing step in the coin recognition methods described in [Zambanini and Kampel, 2011,Zambanini and Kampel, 2012,Zambanini et al., 2013,Zambanini et al., 2014, Kavelar et al., 2014]. And second, it provides not only a region-of-interest for coin identification but also the opportunity the leverage the coin border as a characteristic coin-specific feature. Hence, the method is used to obtain the coin border in the coin identification system presented in [Huber-Mörk et al., 2011].

The proposed coin segmentation strategy is described in Section 3.1. Results of empirical evaluations are presented in Section 3.2. Section 3.2.1 reports experiments on a set of 92 manually segmented ground truth images, whereas Section 3.2.2 reports results on identification based on coin shape determined by the proposed segmentation method.

3.1 Methodology

An essential requirement for the usability of coin segmentation is to provide a ready-to-use method without the need for parameter tuning. Therefore, the objective of the presented methodology is a robust and fast segmentation for a large variety of coin image styles. Coin images from different sources (e.g. museum collections or public online databases) are photographed with different image acquisition setups. Therefore, no assumptions about image quality can be made and major challenges to be faced in the segmentation of coins are caused by an improper image acquisition procedure. Especially shadow casts caused by an insufficient illumination setup impede the correct determination of the coin border. Furthermore, tests have shown that image compression with chroma subsampling (e.g. JPEG image compression [Gonzalez and Woods, 2002]) is widely used when storing images of coins. The resulting compression artifacts preclude

the use of color information, thus only the luminance can be used for a reliable segmentation of the coins.

The method presented in this chapter is based on the assumption that the coin is the most circular object in the image and possesses more local information content and details than the rest of the image. Although this assumption can not be guaranteed to hold in practice, it is reasonable from experience and also supported by the experiments. Thus, the method consists of two steps:

1. **Saliency Extraction**: The image is filtered with a local entropy and range filter in order to obtain a more meaningful image representation for thresholding. After this operation each pixel's value can be seen as the likelihood of belonging to the coin region.

2. **Shape-Controlled Thresholding**: An optimal global threshold is found by maximizing an objective function describing the circularity of the resultant shape.

In the following the two steps are described in detail.

3.1.1 Saliency Extraction

Saliency extraction is done by two filters providing a local measurement of information content in the image: the *local entropy* and the *local range of gray values*.

Local entropy filter: entropy is the measure of the information content in a probability distribution. For digital images, the probability distribution is represented by the histogram of gray values [Kapur et al., 1985]. Given an image point $\mathbf{p}$ and its local neighborhood $\Omega_\mathbf{p}$, all image intensity values $I(\mathbf{p})$, $\mathbf{p} \in \Omega_\mathbf{p}$, are transformed to normalized frequency values $f_1, f_2, ..., f_N$, where N is the number of different values and $\sum_{i=1}^{N} f_i = 1$. Then, the local entropy of $\mathbf{p}$ is defined as:

$$X(\mathbf{p}) = -\sum_{i=1}^{N} f_i \cdot \log_2(f_i). \tag{3.1}$$

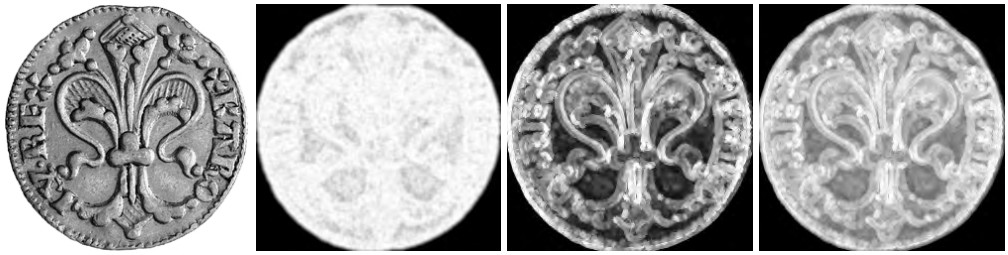
Local range filter: the local range of gray values is defined as the difference of the maximum and minimum gray value of a local neighborhood:

$$Y(\mathbf{p}) = \max_{\mathbf{p}\in\Omega_\mathbf{p}} I(\mathbf{p}) - \min_{\mathbf{p}\in\Omega_\mathbf{p}} I(\mathbf{p}). \tag{3.2}$$

For both filters a circular neighborhood with an empirically determined radius of 3 pixels is used. In order to bring both filter outputs to the same value range they are normalized with respect to the maximum value in the image. The final saliency measure is then given by the sum of both normalized filter outputs:

$$Z(\mathbf{p}) = \frac{X(\mathbf{p})}{\max_{\mathbf{p}\in I} X(\mathbf{p})} + \frac{Y(\mathbf{p})}{\max_{\mathbf{p}\in I} Y(\mathbf{p})}. \tag{3.3}$$

For illustration on a simple example, in Figure 3.1 the particular results of the entropy filter X, the range filter Y and their summation Z are shown. The output of both filters is higher for the region of the coin than for the region of the background, especially at the coin border.



(a) Original image. (b) Output of local entropy filter X. (c) Output of local range filter Y. (d) Final saliency measure Z.

Figure 3.1: Saliency extraction on a simple, illustrative example of a coin image.

3.1.2 Shape-Controlled Thresholding

In order to obtain the final coin segmentation from the saliency image shown in Figure 3.1d, a simple way would be to a apply a global threshold and close all holes in the binary mask caused by homogeneous regions inside the coin. However, the optimal threshold has to detected separately for each image. Defining the optimality of a threshold is in general a difficult task due to the unknown correct result, but in the case of ancient coins we can use the known approximately circular shape as strong constraint. Hence, what is needed is a measure of confidence that estimates the closeness of the resulting binary mask to a perfect circle. Consequently, the *formfactor* [Russ, 2011] turns out to be a suitable choice for this task, defined as follows for a binary connected component $\mathcal{C}$:

$$\mathrm{FF}(\mathcal{C}) = \frac{4\pi A_{\mathcal{C}}}{P_{\mathcal{C}}^2} \tag{3.4}$$

where $A_{\mathcal{C}}$ is the area of the shape and $P_{\mathcal{C}}$ its perimeter. For a connected component the area is typically computed by its number of pixels and the perimeter by tracing the border pixels where vertical and horizontal steps have a length of 1 and diagonal steps a length of $\sqrt{2}$ [Sonka et al., 2007]. The formfactor has adequate properties for measuring the closeness to a circle. First of all, it is invariant to the rotation and size of the shape [Russ, 2011]. And second, the formfactor provides a measurement which is sensitive to both the elongation of the shape and the jaggedness of its border. The higher the jaggedness or elongation of a border, the less the formfactor, which is equal to 1 for a circle and close to 0 for a straight line. The influence of the shape elongation and jaggedness to the formfactor is also demonstrated in Figure 3.2. It can be seen that the formfactor decreases both with increasing elongation and jaggedness of the synthetic shapes[1].

[1]Please note that the formfactor of the circle shape shown in Figure 3.2 is not 1 due to the approximative area and perimeter measurements on binary connected components.

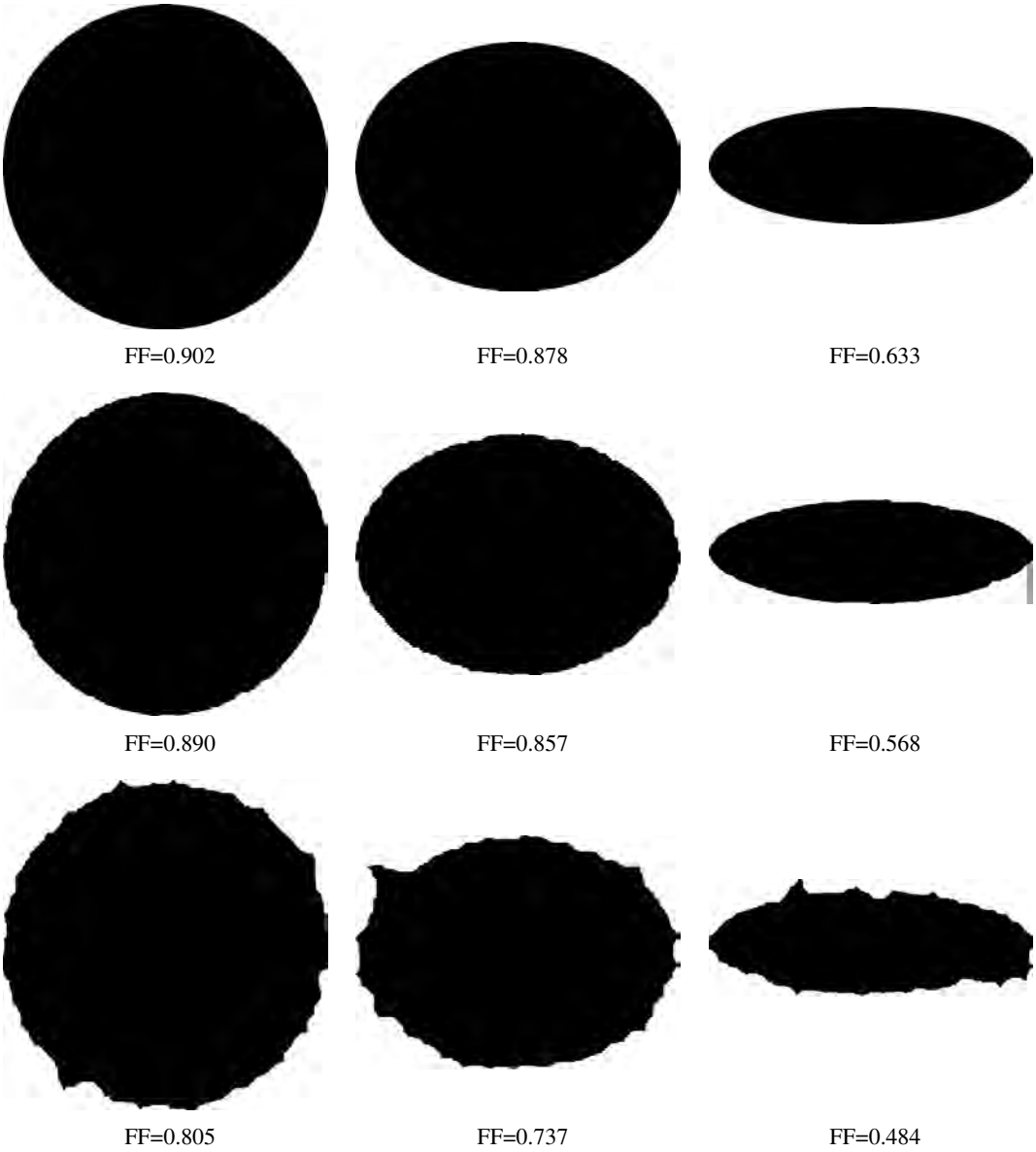


Figure 3.2: Formfactor (FF) of different synthetic shapes with varying elongation and jaggedness. Elongation increases from left to right and jaggedness from top to bottom.

Since the final shape of the segmentation should be close to circle with a regular border, the formfactor provides a convenient measure for the confidence of the segmentation.

As it is assumed that the coin is the most circular object in the image, in the presented method the connected component C_t to be used for the confidence measurement is the one with the highest formfactor in the binary image resulting from thresholding with t. Additionally, it

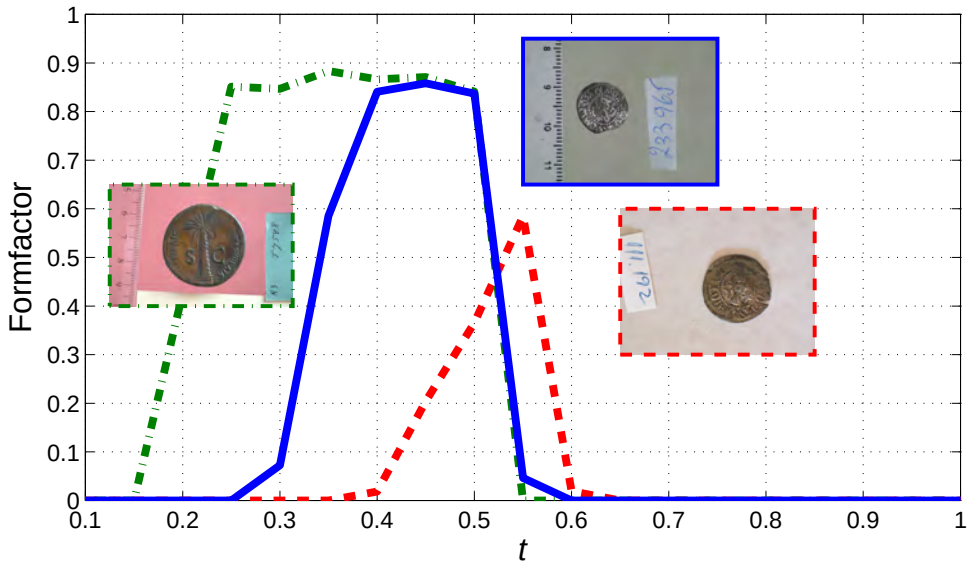


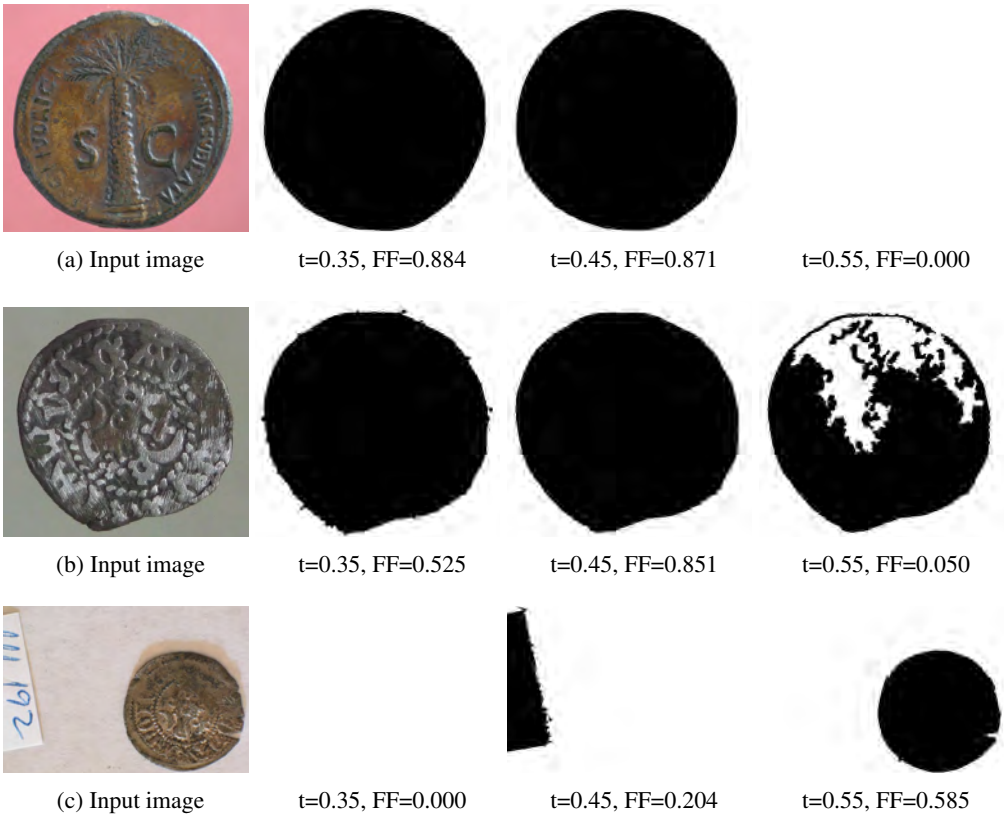
Figure 3.3: Plot of formfactors of three coin images when thresholded with t.

is expected that the size of the coin is between 5 % and 95 % of the overall image area, hence objects above and below are rejected. The binary segmentation mask is also shrunk by removing its 3-pixel-wide border in order to compensate for high saliency values outside the actual coin boundary caused by the applied filter kernels with a radius of 3 pixels.

However, as mentioned above, optimal thresholds have to be individually chosen for each image. This is also illustrated in Figure 3.3 where the confidence measures (formfactors) of three coin images are plotted as a function of the threshold t used to binarize the saliency image. It can be seen that for all three images the confidence measures have clear peaks that indicate an optimal threshold for this image, as the resulting binary shape has the highest formfactor and thus shows the highest confidence that the shape represents the actual coin region. However, these peaks are at different t and formfactor values. Therefore, for the optimal threshold t^* the maximum value of the objective function $\mathrm{FF}(\mathcal{C}_t)$ needs to be found, i.e.

$$t^* = \arg\max_t \mathrm{FF}(\mathcal{C}_t). \tag{3.5}$$

Optimizing this function numerically is straightforward as it consists of only one variable with a specified value range, hence the values of t must be just regularly sampled in the search space $[0, 1]$ and the maximum value obtained must be recorded. For the given task, it has been empirically determined that seven thresholds $t \in \{0.3, 0.35, \ldots, 0.6\}$ are sufficient and that a finer discretization does not improve the accuracy of the method. In Figure 3.4 the segmentation masks along with their respective formfactors for the three coin images depicted in Figure 3.3 and thresholds of 0.35, 0.45 and 0.55 are shown. It can be seen that the confidence values of the images shown in Figure 3.3 are in accordance with the obtained segmentation masks, i.e. the



(a) Input image t=0.35, FF=0.884 t=0.45, FF=0.871 t=0.55, FF=0.000

(b) Input image t=0.35, FF=0.525 t=0.45, FF=0.851 t=0.55, FF=0.050

(c) Input image t=0.35, FF=0.000 t=0.45, FF=0.204 t=0.55, FF=0.585

Figure 3.4: The three images of Figure 3.3 along with the segmentation masks obtained by thresholding with t values of 0.35, 0.45 and 0.55. The input images and segmentation masks are cropped for better visualization.

visually best segmentation correlates with the maximum formfactor. The segmentation masks of Figure 3.4a for $t = 0.35$ and $t = 0.45$ are very similar and $t = 0.35$ is chosen just due to a slightly more regular border. Similarly, the mask of $t = 0.45$ is chosen over the mask of $t = 0.35$ for Figure 3.4b, but the difference in border jaggedness is more pronounced. For Figure 3.4c, the mask of $t = 0.45$ depicts the sheet of paper indicating the coin ID and the correct segmentation with highest formfactor is obtained with $t = 0.55$.

3.2 Experiments

In this section the results of empirical evaluations of the presented method are reported. In Section 3.2.1 a comparison to manual ground truth segmentations is done for 92 coin images from different sources. The suitability of the method for extracting the exact coin border for shape-based coin identification is investigated in Section 3.2.2. The results of Section 3.2.1 and Section 3.2.2 have been originally published in [Zambanini and Kampel, 2009] and [Huber-

Mörk et al., 2011], respectively.

3.2.1 Comparison with Manual Coin Segmentations

The proposed method is tested on a set of 92 images acquired at the *Museum of Fine Arts, Vienna*, the *Fitzwilliam Museum, Cambridge*, and the *Romanian National History Museum, Bucharest*, representing a wide range of different coin images. Six of the images used for evaluation are shown in Figure 3.5. The images of the evaluation set differ in various ways:

- **Resolution**: from 178×184 up to 1154×866.
- **Background**: images with uniform background (Figure 3.5a,b,e,f) and images with less uniform background (Figure 3.5c-d).
- **Image clutter**: images where the coin is the only visible object (Figure 3.5b) and images where other objects like rulers or signs are present (Figure 3.5a,c-f).
- **Coin size relative to image size**: images where the coin perfectly fits into the image frame (Figure 3.5b) or images where the coin region makes only $\sim 15\%$ of the image (Figure 3.5c).
- **Coin roundness**: images with a nearly perfectly circular border (Figure 3.5a,f) and images with a more irregular border due to fragmentation (Figure 3.5b-c)
- **Illumination conditions**: images with (Figure 3.5b-c) and without shadow casts (Figure 3.5a,d-f).

For the experiments presented here, all color images were converted to gray-level images. Compression artifacts due to chroma subsampling as done in JPEG compression [Gonzalez and Woods, 2002] are highly present in the image data and make the use of color information infeasible.

3.2.1.1 Evaluation Procedure

For each image a ground truth segmentation was manually obtained by means of an image editing software. For the evaluation of a single segmentation the dice coefficient (DC) [Dice, 1945], also known as mutual overlap, is measured:

$$\mathrm{DC} = \frac{2 \cdot |\mathcal{C}_s \cap \mathcal{C}_g|}{|\mathcal{C}_s| + |\mathcal{C}_g|} \tag{3.6}$$

where $\mathcal{C}_s$ is the set of pixels in the segmented region and $\mathcal{C}_g$ the set of pixels in the ground truth segmentation. The formula measures the set agreement by the size of the union of two sets divided by the average size of the two sets. Hence, a dice coefficient of 0 indicates no overlap, whereas a dice coefficient of 1 indicates perfect agreement. The dice coefficient is a commonly used evaluation metric for image segmentation [Crum et al., 2006, Shattuck et al., 2009].

To demonstrate the appropriateness of the proposed method for the segmentation of coin images, the results are compared to the outputs of other segmentation methods: (1) the adaptive

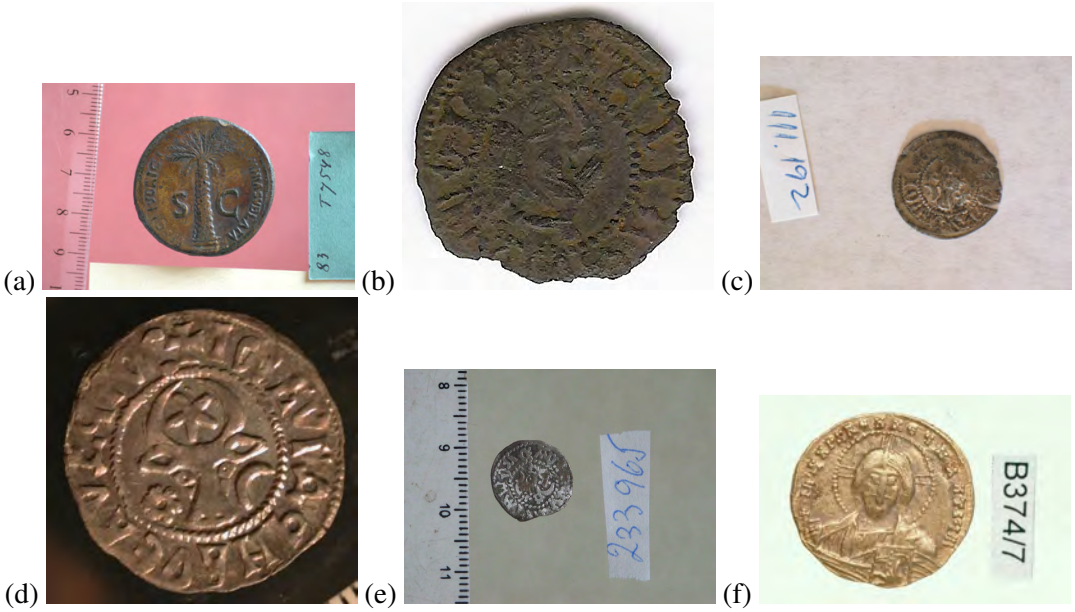


Figure 3.5: Six of the 92 coin images used for evaluation.

thresholding method used in [Zaharieva et al., 2007a] for the segmentation of ancient coins (see Section 2.5.3), (2) the mean shift method proposed by [Comaniciu and Meer, 2002] for a comparison with a state-of-the-art method in image segmentation (see Section 2.2) and (3) the presented method when the thresholding is directly applied to gray values instead of the saliency values.

It must be noted that the output of the mean shift segmentation method is not implicitly a partition into foreground and background, as needed here. Mean shift is a general unsupervised segmentation method and thus partitions the image in a set of disjoint regions without labeling the foreground and background. For the given task, the segmentation has to extract the single most salient object in the image, i.e. the coin. Therefore, to make the mean shift segmentation results comparable, the parameter M for the minimum allowable region area (see [Comaniciu and Meer, 2002] for details) has to be manually adapted for each image to produce a two-segment partition of the image. Evaluation was performed on the mean shift implementation of the EDISON system[2].

3.2.1.2 Results and Discussion

In Table 3.1 the average and median DC of the different methods are listed. The average DC of 0.517 and median DC of 0.720 of the adaptive thresholding method indicate its low robustness. Although the parameters of the method can be adjusted to perform well on a given type of coin image, it is not able to handle the wide range of different images contained in the test set.

[2] http://coewww.rutgers.edu/riul/research/code/EDISON/ (accessed on June 8th, 2014).

	Average	Median
Adaptive Thresholding [Zaharieva et al., 2007a]	0.517	0.720
Mean Shift [Comaniciu and Meer, 2002]	0.983	0.988
Presented method on original gray values	0.923	0.980
Presented method	**0.983**	**0.993**

Table 3.1: Average and median DC achieved on the 92 test images.

A second conclusion of the results is that the local entropy and range filtering is a reasonable preprocessing step to provide a more appropriate intensity image for the thresholding. This can be seen by the lower average and median DC when the original gray values are used. From the results in Table 3.1 it can also be seen that the presented method achieves a similar performance than mean shift segmentation. The average DC is equal (0.983) and the median DC of the presented method is even higher (0.993 to 0.988 of the mean shift method). However, shape-controlled thresholding has two advantages: firstly, in contrast to mean shift no parameter has to be adapted manually. And secondly, the method is computationally faster: while written in MATLAB, it takes 0.11s for a 178×184 image and 1.46s for a 1154×866 image, whereas the mean shift implementation (written in C++) takes 0.24s for the 178×184 image and 8.95s for the 1154×866 image on the same machine.

Figure 3.6 shows results on selected images where the obtained coin border is outlined by a black or white line. Figure 3.6a-c belong to the best segmentation results with a DC of 0.9973, 0.9981 and 0.9970, respectively. Figure 3.6d-e show the two worst results with a DC of 0.9441 and 0.9500, respectively. You see that shadows pose a problem to the method since they produce a strong edge which does not belong to the actual coin border. Nevertheless, although on these images the shadow prevents the correct detection of the coin borders, the coin area is still correctly identified and a further classification of the coin would be only marginally affected by the non-perfect segmentation. Anyhow, on the image of Figure 3.6f the method correctly excludes the shadow from the segmentation, producing a DC of 0.9904.

3.2.2 Evaluation for Shape-Based Coin Identification

The dataset used to evaluate coin identification consists of 2400 images of 240 different coins from the same ancient Greek coin class, provided by the *Fitzwilliam Museum, Cambridge, UK*. In Figure 3.7 four coins of the dataset are shown, where each row represents the same coin specimen and each column represents a different sensor used for coin acquisition and/or orientation of the coin. The two sensor types are a digital camera and a flatbed scanner, which were used to acquire three and two different coin orientations, respectively. By taking images of both coin sides, hence in summation 10 images are available per coin specimen. It can be seen in Figure 3.7 that all images show the same coin type: on the obverse side the head of Heracles in a lion skin in depicted. The reverse side shows the God Zeus seated on a throne. However, each coin specimen's border has its own individual shape, although differences can be subtle.

The coin identification dataset does not include manually annotated ground truth segmentations which prohibits the direct evaluation of segmentation performance. Instead, segmentation performance is indirectly measured by applying shape-based coin identification, i.e. the accu-

(a) DC = 0.9973 (b) DC = 0.9981 (c) DC = 0.9970

(d) DC = 0.9441 (e) DC = 0.9500 (f) DC = 0.9904

(g) DC = 0.9876 (h) DC = 0.9966 (i) DC = 0.9798

Figure 3.6: Results of the proposed segmentation method with respective dice coefficients (DC).

racy is evaluated in the terms of the identification rate when the segmented coin border is used to identify a coin. In [Huber-Mörk et al., 2011] the deviation from a circular shape is used to describe the coin border in order to perform the shape matching for coin identification. The shape descriptor samples the distances of border points to the center of gravity of the segmented coin region at equiangular intervals. For the final shape matching a global and local dissimilarity is computed. The global metric uses the minimum mean squared error of all shifts between two descriptors, whereas the local metric uses the squared distance of their Fourier magnitude values.

Figure 3.7: Four examples of coin specimens used for shape-based identification. Images with a white background have been acquired with a flatbed scanner and images with a gray background with a digital camera.

Tests performed with various training set sizes reveal that the coin identification rates are ranging from 90.3 %, when only one coin image per specimen is available, up to 99.0 %, when 9 images per class are available for comparison. Given the high number of classes (240) and the low inter-class variations (all coins are roughly circular), these results show the effectiveness of the shape-based coin identification method. As the shape matching relies on the presented coin segmentation method, the high identification rates indicate its suitability for obtaining an accurate description of the coin border.

3.3 Summary

In this chapter a shape-controlled segmentation method is described. The method exploits the approximately known shape of the target object and leverages a scalar shape confidence measure to achieve a both robust and fast segmentation. The method starts with the transformation of the image to a saliency map by means of combining the outputs of a local entropy and range filter. For approximately round objects like ancient coins, the formfactor of the shape resulting from thresholding the saliency map is proposed as confidence measure.

It is shown in the experiments that seven regularly sampled thresholds are sufficient to find the best segmentation for images of ancient coins. Hence, although based on optimization, the method is faster than general-purpose unsupervised segmentation methods like mean shift which do not take a-priori knowledge of shape into account. Additionally, for ancient coins the method shows higher segmentation accuracy than a previously proposed adaptive thresholding approach. On the testset of 92 images representing a wide variety of coin image conditions (resolution, relative coin size, illumination, background clutter etc.), the method proves its robustness with dice coefficients of no less than 0.9441. The highest errors are caused by shadow casts at the coin borders, but on the overall dataset there are no substantially wrong segmentations which emphasizes the usefulness of the method as a preprocessing step for coin classification. For coin classification, the purpose of an initial segmentation is to provide a scale-normalized region-of-interest for the subsequent feature extraction step (see Chapter 4). This goal is highly fulfilled, as the possible enlargement of the segmentation due to shadow casts is not more than 15% (see Figure 3.6). Hence, the scale difference is marginal and the introduced small image clutter is weakened by the locality of the similarity metric (see Chapter 5).

If proper arrangements for coin image acquisition are made and no shadows are present at the coin border (e.g. by placing the coin on a sheet of glass or by using an adequate illumination setup), the method can be assumed to give a highly accurate segmentation. This is shown by the experiments for shape-based coin identification, where with the presented segmentation method an identification rate of up to 99% for a dataset of 240 coins can be achieved, despite the generally high shape similarity of the coins.

CHAPTER 4

Illumination-Insensitive Feature Extraction

This chapter deals with illumination-insensitive extraction of image features. The feature should be able to (1) identify the underlying surface characteristics within the shading pattern on the one hand and (2) ignore effects resulting from unknown illumination and material conditions on the other hand. As there exists no fully invariant and distinctive representation for this kind of problem (see Section 2.1.3), it rather has to be aimed to maximize insensitivity and distinctiveness in a joint fashion. This is especially challenging for textureless objects, but this type of objects is widely ignored in existing research and thus the focus of the presented work.

In Section 4.1, first a general evaluation of low-level features is conducted. The objective is to comparatively investigate image representations with respect to their ability for illumination-insensitive recognition. The presented evaluation is comprehensive in the sense that object texturedness, material and the amount of light source change are manipulated by means of a synthetic dataset. Parts of Section 4.1 have been originally published in [Zambanini and Kampel, 2013a].

In Section 4.2, the new local image descriptor LIDRIC is presented. This descriptor is the result of the insights provided by previous feature evaluation and hence shows to outperform existing descriptors on real-world image data with illumination changes. This work has been originally published in [Zambanini and Kampel, 2013b].

4.1 Evaluation of Low-Level Image Representations

The purpose of the presented evaluation is illustrated in Figure 4.1. Given an image patch and a particular representation (i.e. a value or a set of values for each image pixel), it is evaluated how distinctive and insensitive to illumination changes the representation is. The less the distance between two representations of a the same imaged object under illumination changes, the more insensitivity is given. On the contrary, the more the distance between representation of two different imaged objects, the more distinctiveness is given.

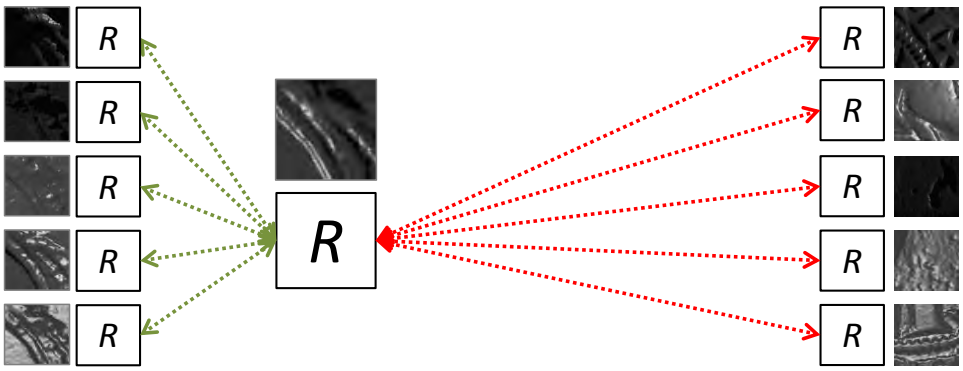


Figure 4.1: The scope of the presented evaluation. An image representation R is evaluated in terms of its distances to same object patches (left, green arrows) under different illumination conditions as well as to other object patches (right, red arrows). The green distances should be low, whereas the red distances should be high.

The remainder of this section is organized as follows. The various image representations used for this evaluation are described in Subsection 4.1.1. In Section 4.1.2 the test data and evaluation scheme is described and the final results are reported and discussed.

4.1.1 Low-Level Image Representations

In this study eight representations are compared. These representations are chosen since they have proposed as being insensitive to illumination changes in the past and are constantly used for illumination-insensitive feature extraction in computer vision. In the following, all representations are described in detail. Additionally, a visualization of the representations is exemplarily shown for the four object types shown in Figure 2.2, i.e. flat and textured, non-flat and textured, textureless as well as textureless and shiny. The patches shown for each object type are again depicted in Figure 4.2.

Gradient Direction (GD)

Image gradient directions have been identified by [Chen et al., 2000] as an illumination-insensitive image feature. The direction $\mathrm{GD}(\mathbf{p})$ of the gradient at the image point $\mathbf{p} = (x, y)$ in the image I is defined as

$$\mathrm{GD}(\mathbf{p}) = \arg\left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right) \tag{4.1}$$

where $\arg(x, y)$ is the angle (in radians) from the x-axis to the point (x, y). In the circular domain, the distance $d_{\mathrm{GD}}(\mathrm{GD}'(\mathbf{p}), \mathrm{GD}''(\mathbf{p}))$ between two gradient directions $\mathrm{GD}'(\mathbf{p})$ and $\mathrm{GD}''(\mathbf{p})$ is then computed as

$$d_{\mathrm{GD}}(\mathrm{GD}'(\mathbf{p}), \mathrm{GD}''(\mathbf{p})) = \min(\left|\mathrm{GD}'(\mathbf{p}) - \mathrm{GD}''(\mathbf{p})\right|, 2\pi - \left|\mathrm{GD}'(\mathbf{p}) - \mathrm{GD}''(\mathbf{p})\right|) \tag{4.2}$$

(a) Flat, textured (b) Non-flat, textured

(c) Textureless (d) Textureless, shiny

Figure 4.2: The image patches used for illustrating the representations used in this evaluation.

Using this distance metric for the individual pixels, the Sum of Squared Distances (SSD) is taken to compare two images,

$$\mathrm{SSD}(\mathrm{GD}', \mathrm{GD}'') = \sum_{\mathbf{p} \in I} (d_{\mathrm{GD}}(\mathrm{GD}'(\mathbf{p}), \mathrm{GD}''(\mathbf{p})))^2. \tag{4.3}$$

Gradient Orientation (GO)

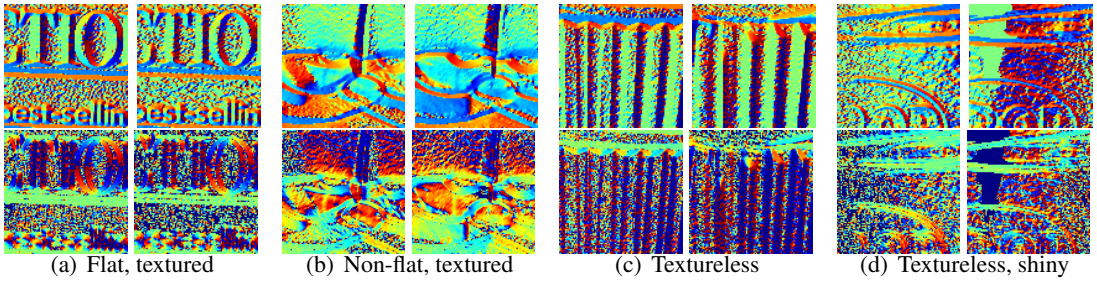
Instead of representing image gradients in a *signed* version (directions between 0-360 degrees), an *unsigned* version (orientations between 0-180 degrees) of gradients can also be used. In the following, the term *direction* is used for *signed* gradients and the term *orientation* for *unsigned* gradients. Gradient orientations are in theory less sensitive to the lighting directions than gradient directions, as opposite lighting directions tend to produce opposite gradient directions at depth discontinuities on the surface [Osadchy et al., 2007]. From the gradient direction $\mathrm{GD}(\mathbf{p})$ the gradient orientation $\mathrm{GO}(\mathbf{p})$ can be simply computed as

$$\mathrm{GO}(\mathbf{p}) = \mathrm{mod}(\mathrm{GD}(\mathbf{p}), \pi). \tag{4.4}$$

To compare two images, the SSD is used where the pixel difference $d_{\mathrm{GO}}(\mathrm{GO}'(\mathbf{p}), \mathrm{GO}''(\mathbf{p}))$ is defined as

$$d_{\mathrm{GO}}(\mathrm{GO}'(\mathbf{p}), \mathrm{GO}''(\mathbf{p})) = \min(\left|\mathrm{GO}'(\mathbf{p}) - \mathrm{GO}''(\mathbf{p})\right|, \pi - \left|\mathrm{GO}'(\mathbf{p}) - \mathrm{GO}''(\mathbf{p})\right|). \tag{4.5}$$

Figure 4.3 shows both the representations GD and GO. One can see that GD is very stable for textured objects, but is affected by edge polarity changes (Figure 4.3c), in contrast to GO.



(a) Flat, textured (b) Non-flat, textured (c) Textureless (d) Textureless, shiny

Figure 4.3: Output of GD (top row) and GO (bottom row).

Laplacian of Gaussian (LOG)

The Laplacian of Gaussian is an approximation of the whitening filter tending to decorrelate the images which makes the filter appropriate for isotropic surfaces [Osadchy et al., 2007]. The LOG filter kernel is computed by applying the Laplacian operator to a Gaussian function with standard deviation σ,

$$\mathrm{LOG}(x,y) = -\frac{1}{\pi\sigma^4} e^{-\frac{x^2+y^2}{2\sigma^2}} \left(1 - \frac{x^2+y^2}{2\sigma^2}\right). \tag{4.6}$$

An example of the filter is shown in Figure 4.4. The LOG filter is used by convolving the image and normalizing the absolute responses to unit length (see Figure 4.5). The distance between two images is then again determined by the SSD.

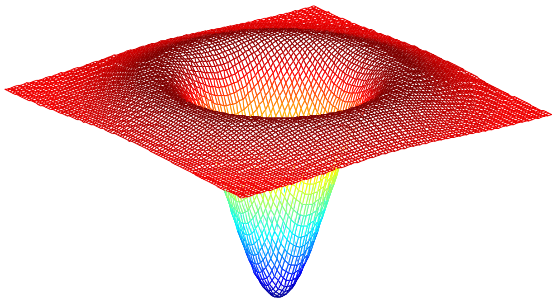
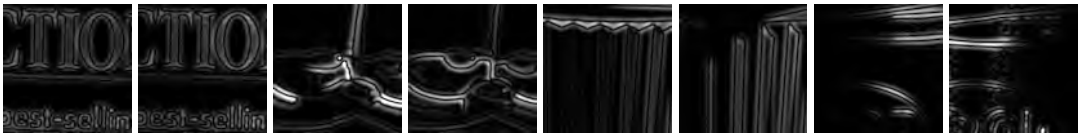


Figure 4.4: A Laplacian of Gaussian filter.

Jets of Gabor Filter Responses (JG)

Gabor filters refer to the work of Dennis Gabor [Gabor, 1946] in which he proposes to represent a signal as a combination of elementary functions. Daugman [Daugman, 1980] extended his



(a) Flat, textured (b) Non-flat, textured (c) Textureless (d) Textureless, shiny

Figure 4.5: Output of LOG.

theory to two dimensions and proposed to use Gabor filters for image feature extraction [Daugman, 1988]. Gabor filters are widely mentioned to be insensitive against illumination conditions [Adini et al., 1997, Kamarainen et al., 2006, Osadchy et al., 2007] due to their invariance against additive and multiplicative intensity changes, which makes them a popular low-level feature for applications like face recognition [Adini et al., 1997, Tan and Triggs, 2010]. A Gabor filter G has complex coefficients and can thus be defined in terms of a real/even part G_e and an imaginary/odd part G_o,

$$G_e(x,y) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\omega}\right), \tag{4.7}$$

$$G_o(x,y) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\sin\left(2\pi\frac{x'}{\omega}\right), \tag{4.8}$$

with $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$. The parameter σ defines the standard deviation of the Gaussian envelope whereas ω represents the wavelength of the sinusoidal plane wave. To construct Gabor filters of different sizes but equal shapes, one can define σ as a linear function of ω, $\sigma = c \cdot \omega$. The parameter θ defines the orientation of the filter and γ is the spatial aspect ratio. Gabor filters can be thought of as quadrature bandpass filters where the combination of a real and imaginary component allows for the phase-invariant detection of oriented frequencies in the image. Figure 4.6 demonstrates the influence of the parameters c and γ for a filter with horizontal orientation.

To construct a Gabor filter bank, the parameters σ, c and γ are kept fixed, N equally spaced orientations $\theta_1 \dots \theta_N$ are used and the image is filtered with the corresponding N Gabor filters $G_e^{\theta_i}$ and $G_o^{\theta_i}$. The jet $\widetilde{\mathrm{JG}}(\mathbf{p})$ is a vector of the magnitude responses of the filtered images $I_e^{\theta_i} = I \star G_e^{\theta_i}$ and $I_o^{\theta_i} = I \star G_o^{\theta_i}$,

$$\widetilde{\mathrm{JG}}(\mathbf{p}) = [\sqrt{(I_e^{\theta_1}(\mathbf{p}))^2 + (I_o^{\theta_1}(\mathbf{p}))^2}, \dots, \sqrt{(I_e^{\theta_N}(\mathbf{p}))^2 + (I_o^{\theta_N}(\mathbf{p}))^2}] \tag{4.9}$$

In addition to complex shading patterns, illumination variations can also induce simple multiplicative changes of image intensities which can be compensated by normalizing the jet to unit length [Kamarainen et al., 2006, Osadchy et al., 2007]. The final feature is thus given by the normalized jet $\mathrm{JG}(\mathbf{p})$. Figure 4.7 shows the elements of JG that correspond to the Gabor filter with $\theta = 0$. The distance between two jets $\mathrm{JG}'(\mathbf{p})$ and $\mathrm{JG}''(\mathbf{p})$ is computed as the L2-norm of their vector difference. Image distances are computed by taking the SSD of $\mathrm{JG}'(\mathbf{p})$ and $\mathrm{JG}''(\mathbf{p})$ for all image points $\mathbf{p}$.

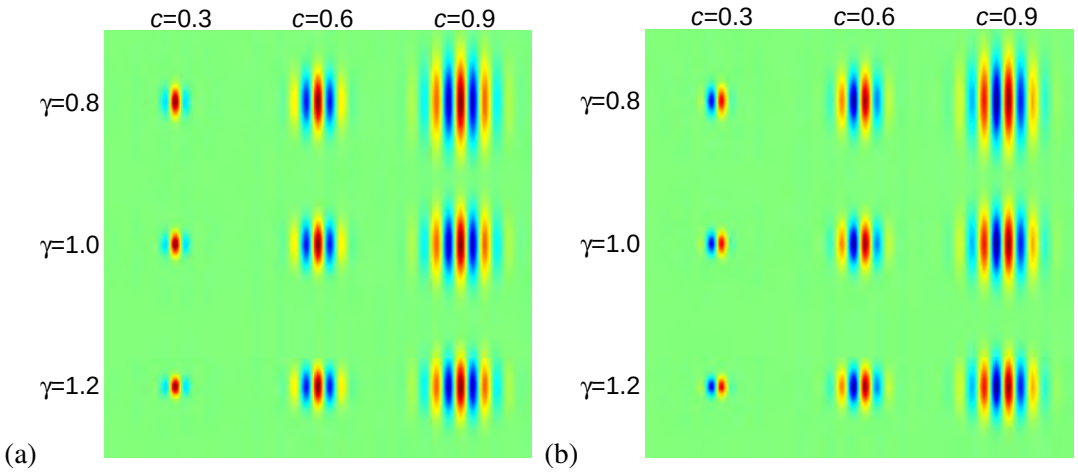
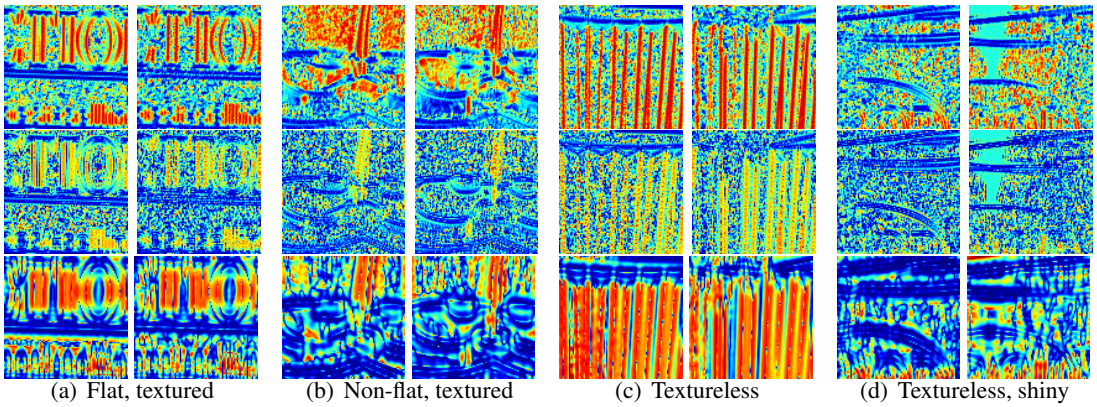


Figure 4.6: The influence of the parameters c and γ on the shape of the (a) even Gabor filter G_e and (b) odd Gabor filter G_o.

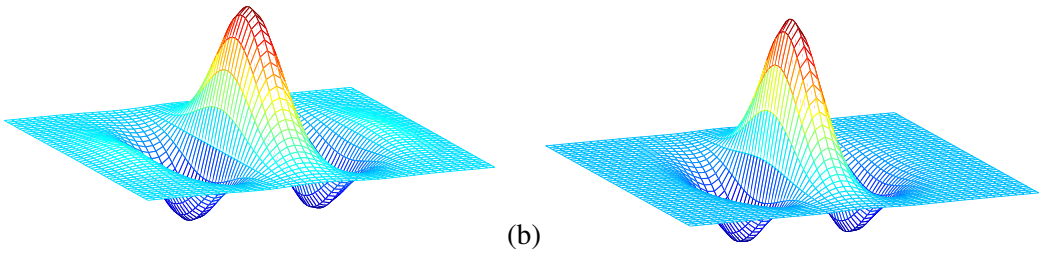


(a) Flat, textured (b) Non-flat, textured (c) Textureless (d) Textureless, shiny

Figure 4.7: Output of the jet element corresponding to a filter orientation of $\theta = 0$ for JG (top row) and JEG (middle row). The bottom row shows the jet element of JMSEG with the same orientation but a higher filter scale ω.

Jets of Even Gabor Filter Responses (JEG)

Besides Gabor jets, jets of oriented second derivatives of Gaussians [Freeman and Adelson, 1991] have also been proposed as an effective way of combining LOG and GO to produce a representation which is appropriate for both isotropic and anisotropic surfaces [Osadchy et al., 2007]. Even Gabor filters have a very similar shape to oriented second derivatives of Gaussians if the cosine bandwidth is chosen such that the Gaussian envelope roughly covers the cosine range of $[-1.5\pi, 1.5\pi]$ (i.e., $c \approx 0.4$) [Kamarainen et al., 2006, Osadchy et al., 2007]. This similarity is depicted in Figure 4.8. Figure 4.8a shows an even Gabor filter with parameters



(a) (b)

Figure 4.8: (a) Even Gabor filter and (b) second derivative of a Gaussian with reversed sign.

$c = 0.4$ and $\gamma = 1.0$ and Figure 4.8b a second derivative of Gaussian with reversed sign.

In this study, even Gabor filters are used as they provide a higher flexibility in the definition of the filter shape, due to a more general set of parameters. However, it is clear from the high similarity of the filters that substantially the same performance can be achieved by the use of second derivatives of Gaussians. In contrast to JG, the jet $\widetilde{\mathrm{JEG}}$ is formed only from the absolute values of $I_e^{\theta_i}$,

$$\widetilde{\mathrm{JEG}}(\mathbf{p}) = [\left|I_e^{\theta_1}(\mathbf{p})\right|, \ldots, \left|I_e^{\theta_N}(\mathbf{p})\right|] \tag{4.10}$$

The final feature is again given by the normalized jet $\mathrm{JEG}(\mathbf{p})$ (see Figure 4.7).

Jets of Multi-Scale Even Gabor Filter Responses (JMSEG)

The optimal size of the filters depends on the surface characteristics, as for smoother surface parts a wider filter is needed than for less smooth surface parts [Osadchy et al., 2007]. One can learn the optimal filter size for a given application domain by means of training data as done in [Osadchy et al., 2007], but nonetheless the variation of surface smoothness is disregarded if only one single filter size is used. As in a general scenario the surface characteristics are usually unknown and varying, it is beneficial to extend the single-scale jet JEG towards a multi-scale representation JMSEG. For this jet the single-scale jets JEG_{ω_i}, obtained by filtering with Gabor filters of scales $\omega_1 \ldots \omega_M$, are simply concatenated,

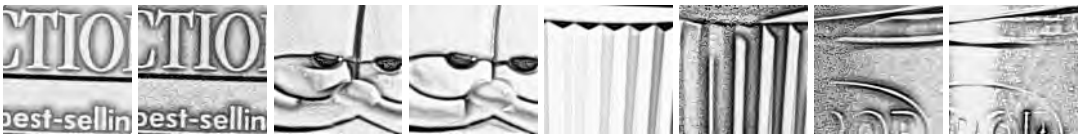
$$\mathrm{JMSEG}(\mathbf{p}) = [\mathrm{JEG}_{\omega_1}, \ldots, \mathrm{JEG}_{\omega_M}] \tag{4.11}$$

The output of a larger scale filter is shown in Figure 4.7. In the presented experiments, the multi-scale representation is only used for even Gabor filter responses due to the higher performance of JEG compared to JG on single scales (see Section 4.1.2.3).

Self-Quotient Image (SQI)

The SQI was introduced by Wang et al. [Wang et al., 2004] as a method to separate the albedo information $R(x, y)$ from images. Similar to other works in this area (see Section 2.1.3.2), the idea is - based on the Lambertian assumption - that the illumination effects mainly appear in the low-frequency components of the image and that they can therefore be eliminated by dividing the image by a smoothed version of it. For increased robustness, several anisotropic smoothing

kernels with different scales are used and integrated to the final self-quotient image. In Figure 4.9 the illumination normalization effect of SQI is shown. It can be seen that the low-frequency parts of the image originating from Lambertian shading (e.g. Figure 4.9b) are suppressed while the high-frequency parts originating from texture are preserved.



(a) Flat, textured (b) Non-flat, textured (c) Textureless (d) Textureless, shiny

Figure 4.9: Output for SQI.

The method is intentionally designed for textured objects, but also showed superior performance in a study of illumination invariance for face recognition [Gopalan and Jacobs, 2010] on the nearly textureless face parts cheek, chin and nose. Another motivation for including SQI in the evaluation is to assess the performance of the vast amount of methods dedicated to textured objects by evaluating one representative method. For the experiments, the implementation of SQI provided by the INFace[1] toolbox is used and the SSD of the SQIs is taken as distance measure.

Gray Value (GV)

In order to have a baseline performance, results for simple image differencing are also reported. In other words, the SSD between the original gray values of the two images is taken as image distance.

4.1.2 Experiments

Experiments are conducted on synthetic image datasets built from 3D historical coin models as well as on real datasets of textureless and textured objects. Synthetic images are used because this way the parameters of image formation can be freely changed to produce images with different illumination conditions and material properties with or without texture. In this manner, it is possible to directly compare the performance of the features under different conditions without introducing a bias due to different objects used between datasets. The real dataset is used to validate the results for various real-world material properties and illumination conditions. Both datasets are specifically described in Section 4.1.2.1. Section 4.1.2.2 details the general evaluation procedure. Various aspects of the evaluation like parameter selection, influence of object texturedness and specularity, influence of amount of light source change as well as real data performance are treated in the Sections 4.1.2.3-4.1.2.6.

[1] http://luks.fe.uni-lj.si/sl/osebje/vitomir/face_tools/INFace/ (accessed on June 8th, 2014).

4.1.2.1 Datasets

Synthetic Datasets

The synthetic datasets consist of images of 14 coin models which were rendered using the open-source graphics software *Blender*[2]. For each model, twelve sets of 500×500 images with 65 illumination directions were rendered where each set represents one out of four material BRDFs and one out of three texture density levels. Material BRDFs are intended to represent different levels of specularity starting from a Lambertian material with zero specularity up to specular intensity values of 0.25, 0.50 and 1.00. Three texture levels were chosen to show the correlation of the features' performances to the amount of texture on the objects. The first level shows no texture and thus represents the set of textureless objects. For the remaining two levels synthetically generated textures were used. For each coin such a texture was generated by placing random characters from different fonts at positions described by a quasi-random spatial distribution. Texture density was measured by computing the mean gradient magnitude in the texture images (for texture density level 1 and texture density level 2 the threshold was set to 0.04 and 0.08, respectively).

For each model and dataset, 65 images with varying illumination directions were rendered. The camera image plane was placed parallel to the coin and light source positions were defined by their azimuth angle φ and elevation angle λ, as illustrated in Figure 4.10. Eight levels of $\lambda \in \{10°, 20°, \ldots, 80°\}$ with eight levels of $\varphi \in \{0°, 45°, \ldots, 315°\}$ each were used to produce 64 images. The 65th image was rendered with the light placed at the camera position (i.e. $\lambda = 90°$). Figure 4.11 shows images of one model rendered with the same illumination parameters of $\varphi = 315°$ and $\lambda = 60°$ for the twelve synthetic datasets. Figure 4.12 shows the appearance variation induced by changing the light sources on the same textureless model and a specular intensity of 0.50.

In Figure 4.13 all coin models are shown. It can be recognized that the coin models exhibit, on a local level, smooth isotropic as well as non-isotropic surface parts and thus cover the wide range of surface characteristics desired for the purpose of this evaluation. As a contribution to other researchers in this field, the overall dataset is available for download[3].

Amsterdam Library of Object Images

The Amsterdam Library of Object Images[4] (ALOI) [Geusebroek et al., 2005] is an image database of 1 000 objects that were photographed from three viewpoints and with eight illumination configurations each. The database contains a wide variety of textureless objects (e.g., a nut, a sponge, white cotton, a metal elephant, a plastic cup...) as well as textured objects (e.g., labeled boxes, an alarm clock, a calendar, a cream tube, a shoe ...), as shown in Figure 4.14. Therefore, the ALOI images provide a realistic and challenging database due to the high variation of material BRDF and surface smoothness among the objects.

[2] http://www.blender.org/ (accessed on June 8th, 2014).
[3] http://www.caa.tuwien.ac.at/cvl/people/zamba/sidire/ (accessed on June 8th, 2014).
[4] http://staff.science.uva.nl/~aloi/ (accessed on June 8th, 2014).

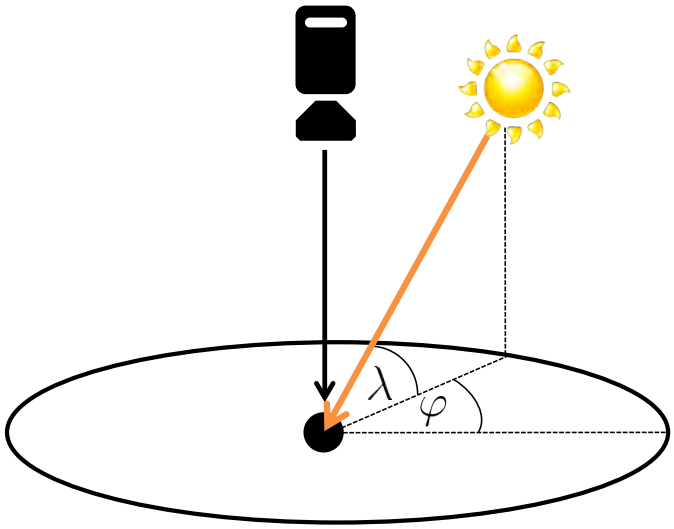


Figure 4.10: Camera and illumination setup for the synthetic datasets.

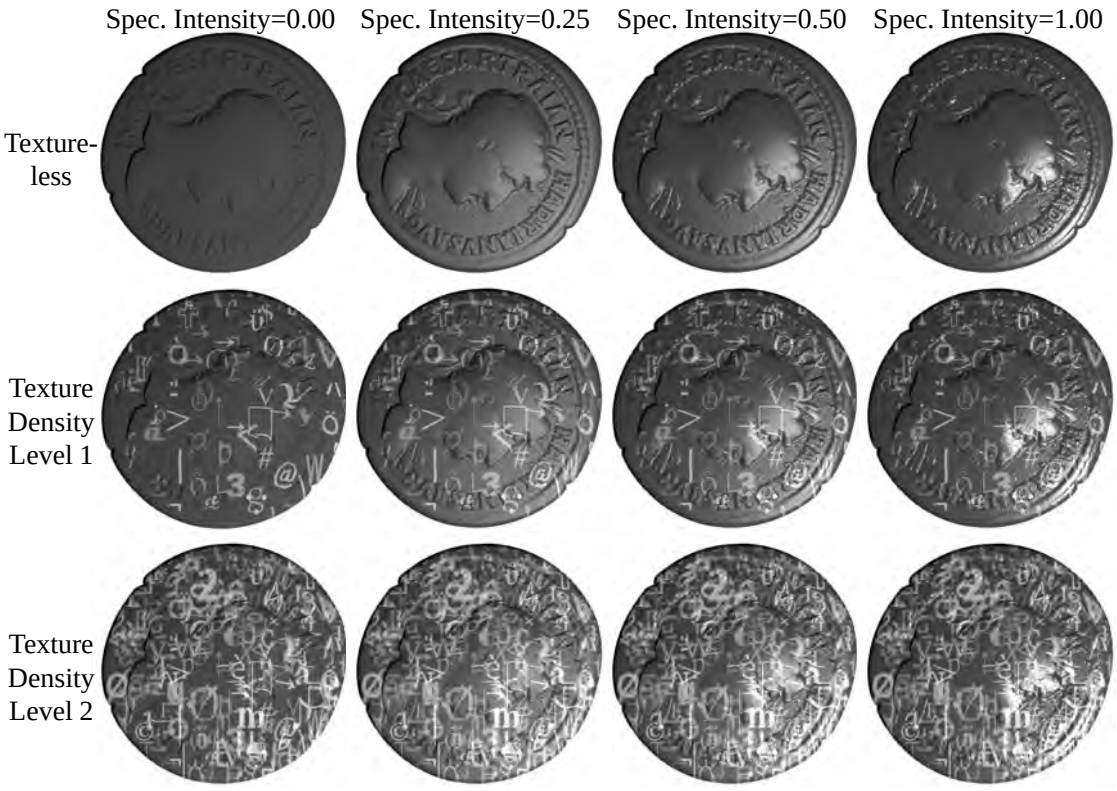


Figure 4.11: Coin model rendered with different material properties and texture densities.

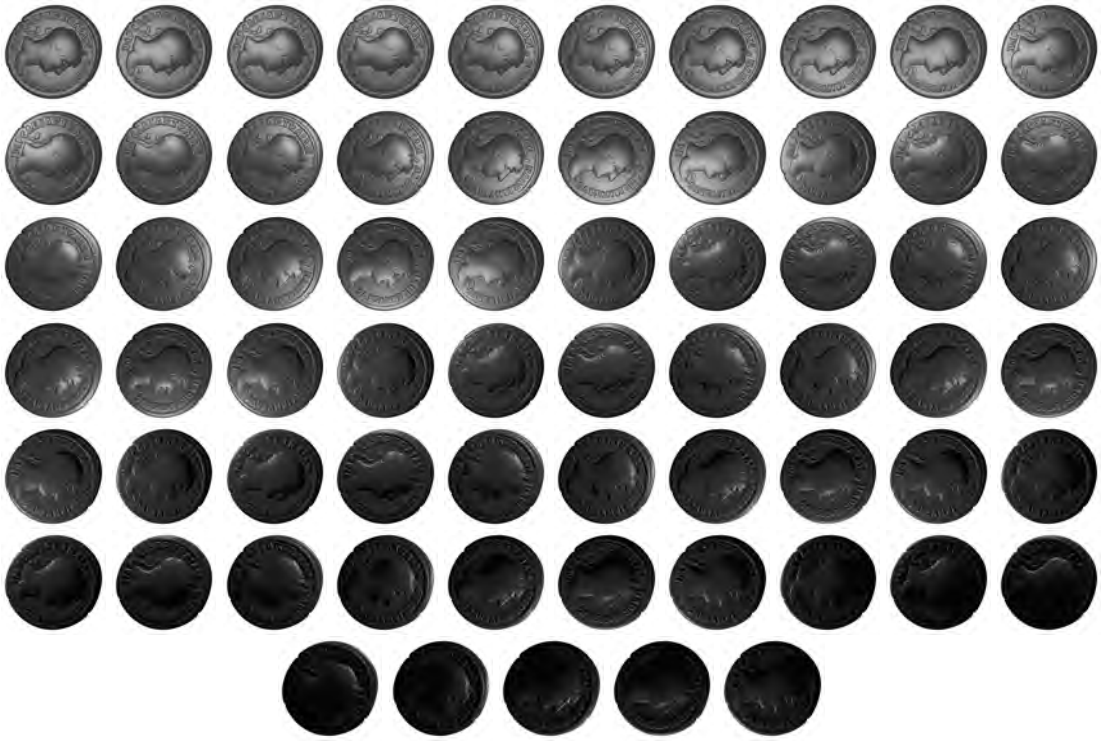


Figure 4.12: All 65 images of the coin model shown in Figure 4.11 rendered without texture and specular intensity of 0.50.

.

Figure 4.13: All 14 coin models used for creating the synthetic datasets. For this figure all models were rendered without texture, specular intensity of 0.25, light source azimuth angle $\varphi = 135°$ and elevation angle $\lambda = 50°$.

4.1.2.2 Evaluation Procedure

For evaluation an empirical performance measure is needed in order to assess the quality of a image representation by means of its distances to true and false images, as depicted in Fig-

Figure 4.14: Examples of (a) textureless and (b) textured objects in the ALOI dataset.

ure 4.1: a "good" feature will minimize the distance between image patches showing the same object part and maximize the distance between image patches showing different object parts. Hence, inspired by the evaluation scheme presented in [Brown et al., 2011], these two groups of distances are measured for a given feature by sets of true and false image patch pairs. True patch pairs show the same object patch but with different illumination conditions, whereas false patch pairs show different object patches. Figure 4.15a-c shows examples of true patch pairs from the synthetic datasets, the ALOI textureless dataset and the ALOI textured dataset, respectively.

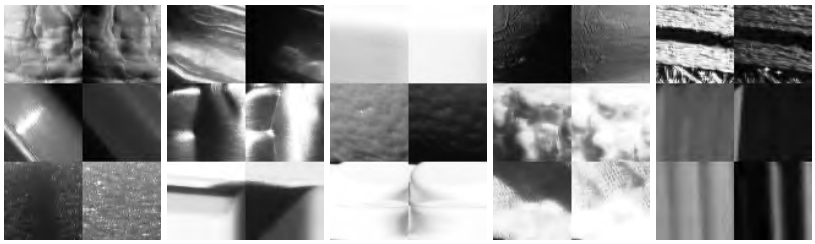
For a given representation and set of true and false pairs, the distances form two histograms, as depicted in Figure 4.16. These two histograms of distances are integrated to build a Receiver Operating Characteristic (ROC) curve [Fawcett, 2006] of which the Area Under Curve (AUC) is computed as performance measure. The ROC analysis allows to investigate how reliably one can discern true pairs from false pairs by evaluating the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR), defined as

$$\text{TPR} = \frac{\text{number of true positives}}{\text{number of positives}}, \text{FPR} = \frac{\text{number of false positives}}{\text{number of negatives}}. \tag{4.12}$$

Here, the number of true positives is the number of true patch pairs whose distance is below a given threshold and the number of positives is the total number of true patch pairs in the dataset. Likewise, the number of false positives is the number of false patch pairs whose distance is below the same threshold and the number of negatives is the total number of false patch pairs in the dataset. The ROC curve is then formed by varying the threshold from zero up to the maximum distance occurring in the overall set and measuring the corresponding TPRs and FPRs. The AUC is a simple scalar measure allowing to compare the performance of ROCs and their underlying classifiers, with 1.0 being a perfect result and 0.5 being the statistical outcome of random guessing. Intuitively, a highly illumination-insensitive feature will have less overlap between the two histograms shown in Figure 4.16, and thus will produce values nearer to the top left corner of the ROC space and a higher AUC than a less illumination-insensitive feature.

For generating the patch pairs, the same number of true patch pairs and false patch pairs were randomly extracted from the images of a dataset. A patch size of 16×16 pixels was used, but in general the patch size has no significant impact on the results, as has been observed in initial tests. To generate patch pairs from the ALOI datasets, 80 textureless objects and 80 textured objects were manually identified in the dataset and non-overlapping true and false patch pairs

(a) Synthetic Datasets



(b) ALOI textureless dataset

(c) ALOI textured dataset

Figure 4.15: True patch pairs used for evaluating the image representations (of size 64×64 for better illustration).

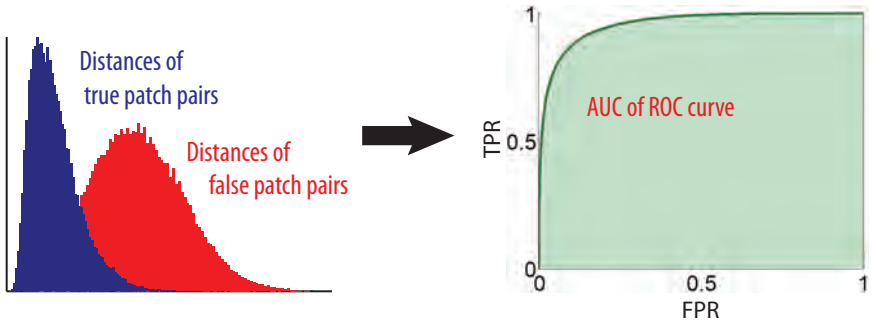


Figure 4.16: Histograms of distances between true and false patch pairs and derived ROC curve with measured AUC.

were randomly picked from images taken from the same viewpoint (12 000 from the textureless objects and 18 000 from the textured objects).

4.1.2.3 Parameter Selection

As the main purpose is the study of the features' behavior on textureless objects with varying material properties, tests for parameter selection were conducted on a mixed patch pair set extracted from the four synthetic datasets of textureless objects. More particularly, for the true patch pairs 50 000 pairs were randomly extracted, each showing a 16×16 region from the same object and at the same image position, but illuminated from different directions. For the false patch pairs, 50 000 pairs were randomly extracted, each showing a 16×16 region from different objects at different positions. Parameter selection was then achieved by an exhaustive search over the parameter space.

For GD and GO, it was tested if a presmoothing of the patches or a larger Sobel filter than the standard 3×3 one are beneficial in terms of recognition performance, but no improvement could be detected. For LOG an exhaustive search was done to find an optimal standard deviation of 1.5 of the Gaussian. For SQI, no exhaustive parameter selection was conducted as this method is intended for textured objects and initial tests with several parameter settings were not successful in substantially improving the generally bad performance of SQI. Therefore, the standard settings defined in the INFace Toolbox were used.

For the features JG and JEG, the parameters defining the shape of the Gabor filters (c and γ) as well as the number N of orientations are of interest. Hence, a parameter space of $c \in \{0.30, 0.35, \ldots, 1.00\}$, $\gamma \in \{0.50, 0.60, \ldots, 1.50\}$ and $N \in \{2, 3, \ldots 12\}$ was defined. Figure 4.17 shows the maximum AUC achieved over the parameter space for various fixed values of c, γ and N. It can be derived from these results that the best performance is achieved when c is set in a range of $0.45 - 0.50$, i.e. the filters have a shape close to second derivatives of Gaussians (see also Figure 4.8). The optimal value for the aspect ratio of the filters defined by the parameter γ is around 0.9. The experiments also reveal that the number of orientations has only a minor influence on the overall performance for $N \geq 6$. Based on these results, for the further experiments parameter values of $\gamma = 0.9$ and $N = 6$ for JG, JEG and JMSEG are used, as well as $c = 0.50$ for JEG and JMSEG and $c = 0.45$ for JG. The resulting shape of the JEG and JMSEG filters is shown in Figure 4.18, whose similarity to second derivative of Gaussian filters becomes apparent by comparison with Figure 4.8. Optimal filter sizes $\omega_1 \ldots \omega_M$ for JMSEG were identified as $\omega_j = \omega_1 2^{(j-1)/2}$ with $\omega_1 = 1$ and $M = 8$.

4.1.2.4 Recognition Performance Depending on Object Specularity and Texturedness

To evaluate the recognition performance of the features for the twelve synthetic datasets, 50 000 true and false patch pairs were randomly extracted from each dataset. Patch pairs contained in the mixed set for parameter selection were not included into the four textureless datasets used for this evaluation. The results are plotted in Figure 4.19a-c. It can be clearly seen that on textureless objects the representations based on even Gabor responses (JEG and JMSEG) perform best. The multi-scale representation of JMSEG is beneficial especially on Lambertian surfaces where it shows a significant improvement of recognition performance over JEG (AUC of 0.933 against 0.899). Complex Gabor filter responses (JG) are better than the other remaining

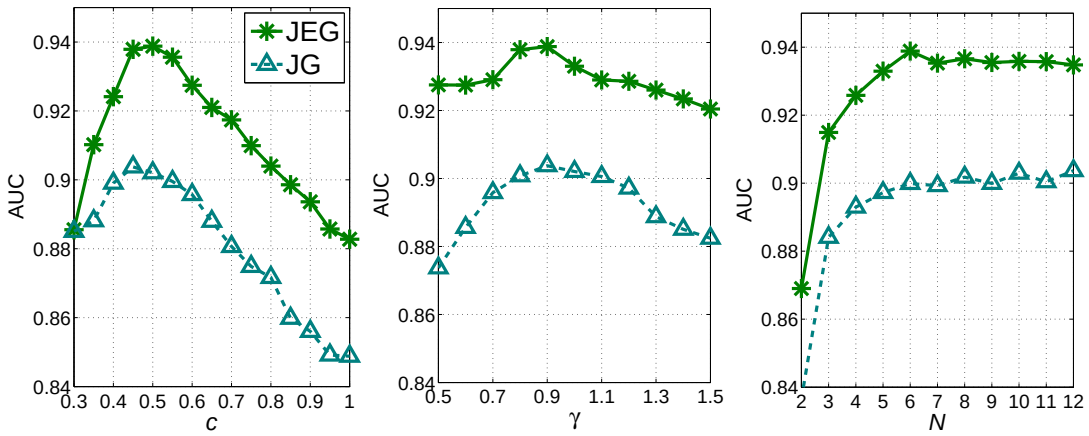


Figure 4.17: Recognition performance of JEG and JG dependent on parameters c, γ and N.

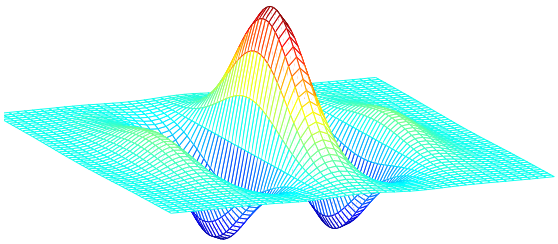
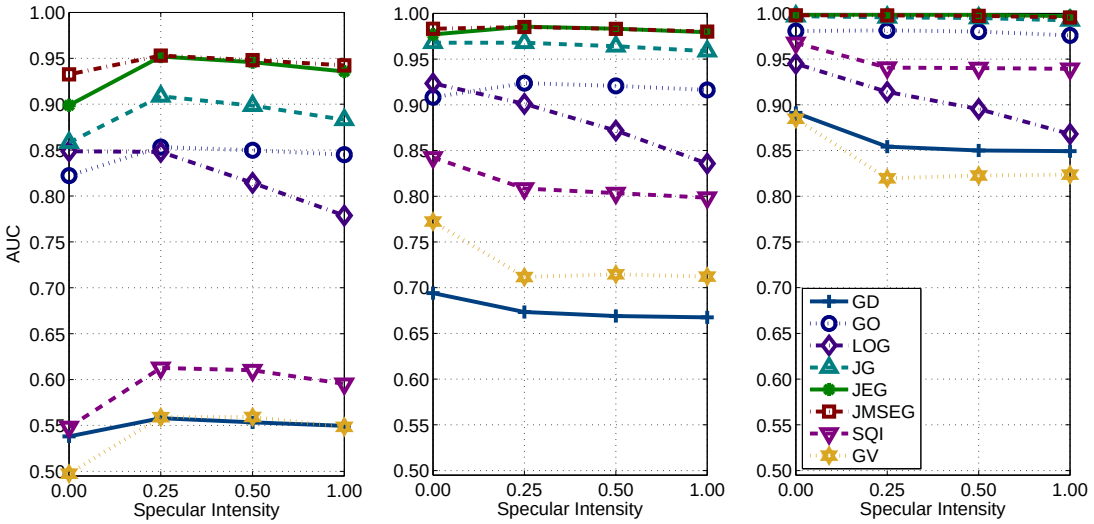


Figure 4.18: The even Gabor filter with $c = 0.5$ and $\gamma = 0.9$ used for JEG and JMSEG.

features but it can be concluded from the worse performance compared to JEG that the phase invariance of the complex filter decreases its recognition power. For image gradients, there is a large discrepancy between the use of gradient orientations (GO) and gradient directions (GD). GD is much less stable than GO as it is highly vulnerable to edge polarity changes induced by opposite lighting directions between true patch pairs. SQI is only slightly better and performs substantially worse than the top-performing features, as this representation is designed for textured objects and is thus highly affected by changes of the shading patterns on textureless objects. Therefore, the method achieves its best results on Lambertian objects with a high texture density (Figure 4.19c). Another conclusion from the results on textureless objects is that more specularity of the objects' material increases the performance. Although a specular BRDF causes more appearance variations from light source variations than a Lambertian BRDF, surface characteristics are also more accentuated by a specular surface, which in turn supports its recognition. The only exception of this effect is LOG which has been especially proposed for smooth, Lambertian objects [Osadchy et al., 2007].

The results on textured objects shown in Figure 4.19a-b show that texture increases the recognition performance of all features and in general that their performance is correlated to the degree of texture variation. Naturally, since changes of albedo are less affected by lighting variations than changes of object depth, the recognition of objects is more robust the more albedo



(a) Textureless objects.

(b) Objects with texture density level 1.

(c) Objects with texture density level 2.

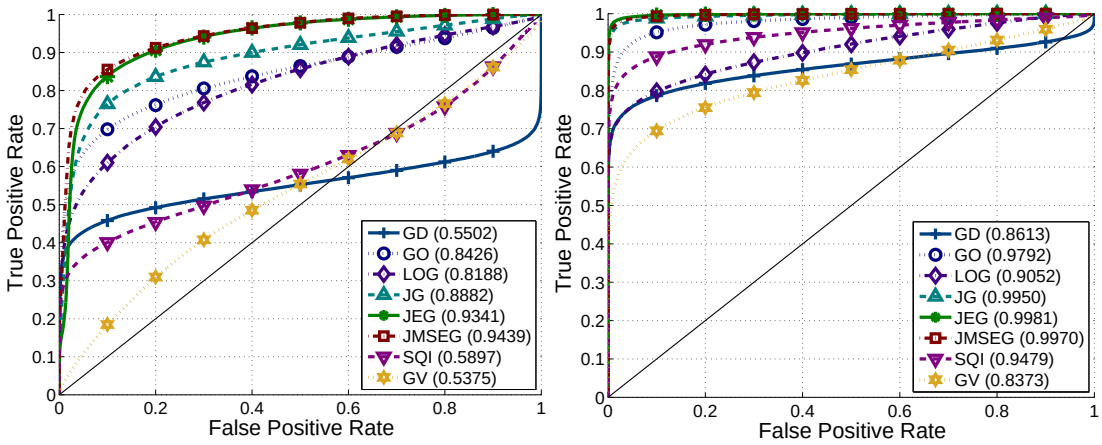
Figure 4.19: Comparison of recognition performance for different levels of texturedness and material specularity.

changes occur (i.e. a higher texture density). However, the representations based on Gabor filters are the best performing features for all scenarios, regardless of the texture density of the objects.

4.1.2.5 Influence of the Amount of Light Source Change

An interesting question in the context of this evaluation is how the amount of light source difference between the images to be compared has an influence on the discriminative power of the features. If one examines the ROC curves obtained from all textureless patch pairs shown in Figure 4.20a, it is evident that some curves cross with each other. This indicates that the relative orderings of recognition performance among the features are dependent on the amount of light source difference. The representations GD, SQI and GV, which generally perform badly due to their non-invariance to the polarity of image edges, have a lower false positive rate in a true positive range of around $0.0 - 0.4$ than the generally top-performing feature JEG. On the other hand, for higher false positive rates they have true positive rates below the line of no-discrimination. This effect is caused by the strong impact that opposite lighting directions between patch pairs of textureless objects have on the computed distance, as in such cases the polarity change makes the distance become larger than the distances between random false patch pairs. Albedo changes usually do not cause edge polarity changes for opposite lighting directions, and thus this effect is far less pronounced for textured objects (see Figure 4.20b)

To evaluate the features' performances with respect to the amount of light source differences, this issue is taken into account for the ROC curve generation by subselecting patch pairs from the textureless objects with a given difference of light source azimuth or elevation. Hence, only true patch pairs with a specified azimuth difference and no elevation difference, and vice versa, are



(a) Textureless objects

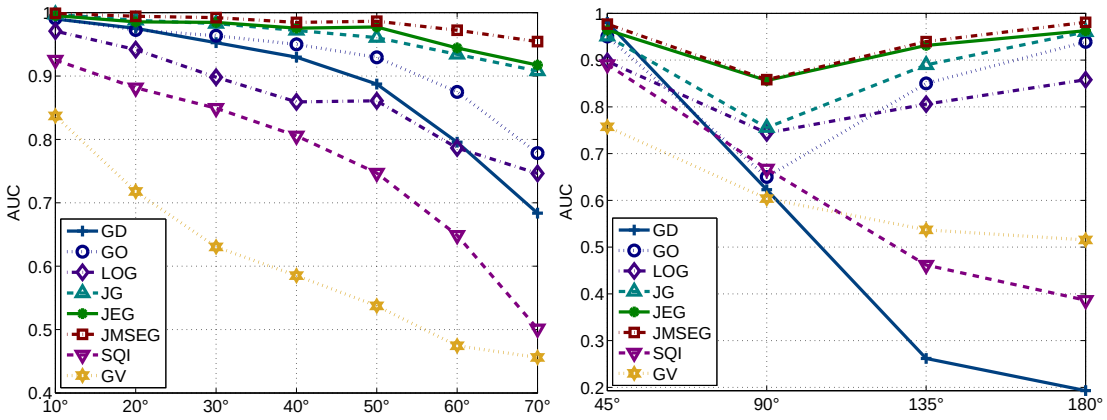
(b) Objects with texture density level 2

Figure 4.20: ROC curves for textureless and textured objects. The values in brackets are the AUCs of the features. The black solid line depicts the line of no-discrimination.

considered. The results of these tests are shown in Figure 4.21. The plotted curves demonstrate that for smaller light source changes the performances of the features are close together whereas for stronger changes there is also a higher difference in performance. GD is a competitive feature for small light source changes of $45°$ azimuth and $10° - 20°$ elevation, but its performance decreases stronger than that of other features for larger light source changes. GD, SQI and GV are especially vulnerable to changes of the light azimuth, in contrast to representations which are invariant to edge polarity. For these features azimuth changes of $90°$ represent the worst case scenario, whereas the recognition performance at changes of $180°$ lies in the same range as the performance at changes of $45°$. An important aspect of these experiments is that JMSEG shows the top performance for all levels of light source changes, but its dominance is more pronounced for higher levels of change. Therefore, for scenarios where only little variations of illumination conditions are expected, GD can still be considered as a powerful low-level representation, whereas for more extensive variations single- or multi-scale Gabor filter responses work remarkably better than all other representations.

4.1.2.6 Recognition Performance on Real Datasets

As can be seen in Figure 4.22, the results on the real datasets widely reflect the findings of the experiments on the synthetic datasets. JMSEG is again the best performing feature for textureless and textured objects, followed by JEG and JG. The generally lower performance on the real datasets is explained by image noise on the images as well as the acquisition setup used. There are more underexposed (i.e. completely black) and overexposed (i.e. completely white) objects parts which evidently hinders recognition. Nonetheless, the results show that the insights gained from the experiments on synthetic datasets can be transferred to the real world.



(a) Difference of light elevation angle λ (b) Difference of light azimuth angle φ

Figure 4.21: Recognition performance in relation to the amount of light source difference.

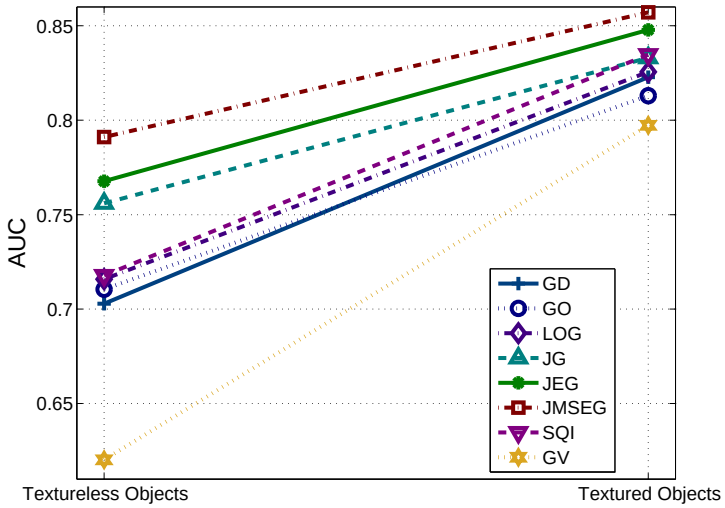


Figure 4.22: Recognition performance on the real ALOI datasets of textureless and textured objects.

4.2 LIDRIC: A Local Image Descriptor Robust to Illumination Changes

Due to the demonstrated superiority of the JMSEG feature transformation for illumination changes, in this section this low-level feature is used as a basis for image descriptor construction. It has been argued in Section 2.3.4 that typical effects of illumination variations like changes of edge polarity or spatially varying brightness changes are not taken into account by current descriptors. Gradient directions, which are a commonly used low-level feature for descriptor construction [Lowe, 1999, Ke and Sukthankar, 2004, Mikolajczyk and Schmid, 2005, Tola et al.,

2008, Chen et al., 2008, Simonyan et al., 2012, Seidenari et al., 2014], are not well-suited to handle these variations, as shown in Section 4.1.2. The same applies to other low-level features used like Haar wavelets [Bay et al., 2006] or pairwise image intensity comparisons [Calonder et al., 2010, Leutenegger et al., 2011, Alahi et al., 2012]. Hence, by relying on a more suitable low-level feature, a more illumination-insensitive image description is achieved. This description is enriched with spatial statistics in a spatial coding stage.

In Section 4.2.1 the methodology for constructing the LIDRIC descriptor is described in detail. In Section 4.2.2 the descriptor is empirically compared to existing local image descriptors. In this evaluation the shortcomings of past evaluations are omitted by using a more challenging dataset for testing insensitivity to illumination conditions. Past evaluations [Mikolajczyk and Schmid, 2005, Moreels and Perona, 2007, Van De Sande et al., 2010] have the problem of only marginally considering illumination changes (see also Section 2.3.4): [Mikolajczyk and Schmid, 2005] and [Van De Sande et al., 2010] only test image brightness changes and ignore changes of the light source direction. Moreels and Perona [Moreels and Perona, 2007] use only three different lighting configurations and couple the descriptor performance evaluation with the interest point detection step. The experiments presented in this thesis aim at a broader evaluation: testing the robustness of the small image patch descriptions against changes of the light source direction for textured as well as textureless objects, the latter being the more challenging type of objects.

4.2.1 LIDRIC Descriptor Construction

The LIDRIC descriptor is based on JMSEG described in Section 4.1.1, as this low-level feature showed the top-performance in the illumination insensitivity evaluation. Therefore, for given values of the filter shape parameters c and γ, N filters with equally spaced orientations $\theta_1 \dots \theta_N$, where $\theta_i \in [0, \pi[$ and $\theta_1 = 0$, are constructed. The M scales $\omega_1 \dots \omega_M$ are exponentially sampled to achieve homogeneous intervals, i.e. $\omega_j = k^{j-1}\omega_1$.

In order to build the LIDRIC descriptor, the absolute filter responses I^{θ_i,ω_j} are computed by convolving the image patch I with the $N \cdot M$ filters $G_e^{\theta_i,\omega_j}$,

$$I^{\theta_i,\omega_j} = \left| I \star G_e^{\theta_i,\omega_j} \right|. \tag{4.13}$$

These output images are arranged to a feature map F that contains for every image point $\mathbf{p}$ and discrete filter parameters θ_i and ω_j the absolute filter responses,

$$F(\mathbf{p}, \theta_i, \omega_j) = I^{\theta_i,\omega_j}(\mathbf{p}). \tag{4.14}$$

Obviously, the filter outputs I^{θ_i,ω_j} depend on the image contrast, e.g., stronger ridges produce higher values in $F(\mathbf{p})$. This gives a natural weighting of the local low-level features, in the same manner as, for instance, the gradient magnitude is used to weight the histogram inputs in the SIFT descriptor [Lowe, 2004]. Global linear brightness changes on the image patch can then be compensated by normalizing the final histogram vector to unit length. However, different light source directions can lead to brightness changes that vary locally, as demonstrated in Figure 4.2b-d. Therefore, similar to JMSEG, F is normalized on a per-pixel level,

$$\widetilde{F}(\mathbf{p}, \theta_i, \omega_j) = \frac{F(\mathbf{p}, \theta_i, \omega_j)}{\sqrt{\sum_{i=1}^{N} \sum_{j=1}^{M} F(\mathbf{p}, \theta_i, \omega_j)^2}}. \tag{4.15}$$

An option would also be to normalize only over the responses of different orientations, as done for JMSEG, but normalizing over all responses is more robust when parts of the image region are over- or undersaturated. On discrete images linear brightness changes lead to a clipping of values which are outside the dynamic range of the sensor. The normalized feature map $\widetilde{F}$ is not invariant to brightness changes when such effects occur, but normalizing over all orientation and scale responses is more robust in presence of partial over- and undersaturation as wider filters and thus more data samples are included. For illustration, the feature maps obtained for the right textureless image of Figure 4.2c are shown in Figure 4.23 for $N = 6$, $\omega_1 = 4$ and $k = 2$.

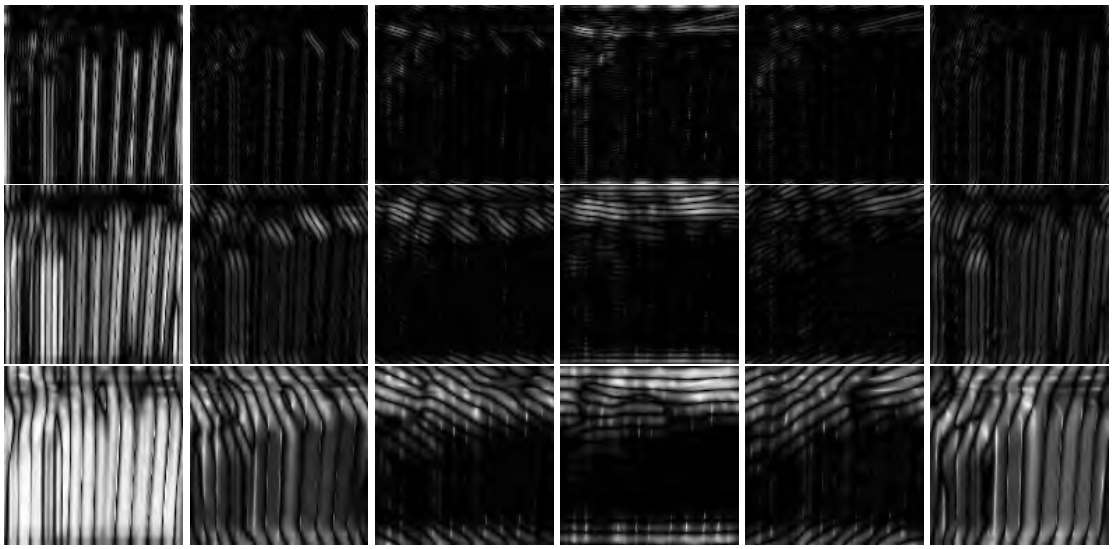


Figure 4.23: Feature maps for the right textureless image of Figure 4.2 with size 128×128. Filter orientations of $\{0, \frac{1}{6}\pi, \frac{2}{6}\pi, \frac{3}{6}\pi, \frac{4}{6}\pi, \frac{5}{6}\pi\}$ are shown from left to right and scales of $\{4, 8, 16\}$ from top to bottom.

The last step of the descriptor construction is to perform a spatial pooling on $\widetilde{F}$ to increase the descriptor's discriminative power by adding spatial information. Formally, L cells $C_l(\mathbf{p}), l = 1 \ldots L$ are defined that represent the weighting of the spatial location $\mathbf{p}$ for the cell's local subhistogram. The final descriptor is a 3D joint histogram $H(\theta_i, \omega_j, l)$ of the values in $\widetilde{F}$,

$$H(\theta_i, \omega_j, l) = \sum_{\mathbf{p} \in \widetilde{F}} C_l(\mathbf{p}) \cdot \widetilde{F}(\mathbf{p}, \theta_i, \omega_j). \tag{4.16}$$

The cells C_l can be, for instance, of Gaussian shape to achieve a DAISY-like pooling [Tola et al., 2010, Brown et al., 2011] or perform the bilinear weighting between squared cells as in the SIFT descriptor [Lowe, 2004]. However, in general the optimal pooling scheme depends on the application scenario, and a finer pooling increases the distinctiveness of the descriptor while

decreasing its robustness to deformations of the underlying image structure and vice versa. Thus, for object matching between viewpoints one can use smaller cells the less viewpoint differences are expected. In the presented experiments, the standard SIFT 4×4 squared cells with bilinear weighting [Lowe, 2004] for spatial pooling is used as it achieves reasonably good results on all datasets. The corresponding cells $C_1, \ldots, C_{16}$ are visualized in Figure 4.24. Please note that the cells are weighted with a Gaussian window with a standard deviation of half of the window size to make the descriptor less dependent on the exact keypoint positioning [Lowe, 2004].

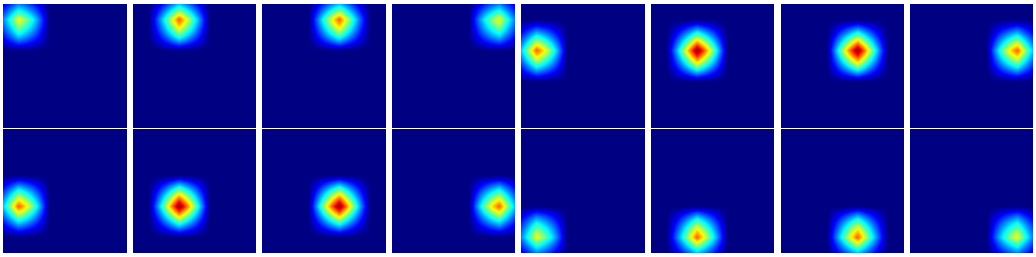


Figure 4.24: The 16 cells for the SIFT-like 4×4 pooling with bilinear weighting.

Although Gabor filters are well known and often used [Kamarainen et al., 2006], to the best of the author's knowledge they have never been used before in this manner for local image descriptors. Multi-scale and/or multi-oriented filter banks are used by others for image recognition tasks, e.g. by [Kyrki et al., 2004, Ahonen and Pietikäinen, 2009] with Gabor filters and by [Brown et al., 2011] with their single-scale Gaussian derivative counterparts, but the filters are usually used in quadrature, whereas in this work only their real part is used. It is shown in the general evaluation (see Section 4.1.2) that combining even and odd outputs has no positive effect on recognition performance. Additionally, it is shown in the LIDRIC experiments (see Section 4.2.2) that using only the real part for feature map construction achieves similar results under strong illumination variations while saving computation time. Larsen et al. [Larsen et al., 2012] also use filter bank responses to build a local descriptor, but they rely on higher-order derivative filters which are applied to single positions on the patch. In contrast, in this work statistics of the filter responses are established at all pixel positions. The statistics are well-founded for an illumination-insensitive descriptor by means of measuring the spatially varying frequency of occurrence of locally normalized responses.

4.2.2 Experiments

For the experiments the same evaluation procedure as described in the general evaluation (see Section 4.1.2.2) is applied, instead of combining the evaluation with interest point detection as done in [Mikolajczyk and Schmid, 2005, Moreels and Perona, 2007]. However, in lieu of the synthetic coin model datasets two additional data sets are used which are more appropriate for descriptor evaluation. All datasets consist of non-aligned image data for true patch pairs and thus simulate the common situation of inexact keypoint positioning and local deformations due to viewpoint differences when matching local image structures. The datasets are described in Section 4.2.2.1. Gabor filter parameter selection for LIDRIC construction is handled in Sec-

tion 4.2.2.2 and its performance compared to other descriptors is reported and discussed in Section 4.2.2.3.

Three configurations of the LIDRIC descriptor are used for evaluation: one that uses only single-scale even Gabor filters (SSEG) with dimensionality N, one that uses multi-scale even Gabor filters (MSEG) with dimensionality $N \cdot M$ and the full descriptor which uses a 4×4 grid for spatial pooling (MSEG4x4) with dimensionality $N \cdot M \cdot L$. These descriptors are compared to several descriptors proposed in literature where implementations are provided for download: SIFT [Lowe, 2004], SURF [Bay et al., 2008], DAISY [Tola et al., 2010], MROGH, MRRID [Fan et al., 2012], LIOP [Wang et al., 2011], FREAK [Alahi et al., 2012] and GLAC [Kobayashi and Otsu, 2008]. Additionally, the best DAISY descriptor reported in [Brown et al., 2011] (BESTDAISY) was implemented which uses second order steerable quadrature pair filters, a DAISY-like spatial pooling and a final vector normalization with range clipping. A modified version of the SIFT descriptor is also tested which uses unsigned gradients in the range $[0, \pi[$ to handle the polarity changes of edges on textureless surfaces under opposite lighting directions (UGSIFT).

4.2.2.1 Datasets

Experiments are conducted on four datasets of true and false image patch pairs. The ALOI datasets described in Section 4.1.2.1 are again used. To test the descriptors on scenarios with less lighting variations the existing image patch pair databases *Liberty* and *Virtual World* are also used for evaluation.

ALOI Textureless and *ALOI Textured*

The same 80 textureless objects as already selected for the patch pair generation described in Section 4.1.2.2 are used, but with larger patch sizes of 64×64. Correspondences for the true patch pairs were identified by manually estimating the homography between images from the three viewpoints. The viewpoint changes are small enough to describe the image correspondences by a homography and thus errors are also considered as being small enough to just simulate the uncertainty of interest point detection. In total, 60 000 true and 60 000 false patch pairs were extracted. Likewise, the same 80 textured objects were selected resulting in a set of 120 000 true and false patch pairs.

Liberty Dataset

The *Liberty* dataset[5] consists of true and false patches sampled from 3D reconstructions of the Statue of Liberty. This dataset has been used for descriptor learning [Brown et al., 2011] and represents an appropriate descriptor evaluation dataset for the scenario of multi-view reconstruction of large-scale outdoor objects. Hence, it also includes realistic outdoor lighting variations, although their amount and frequency is unknown. 50 000 patch pairs of the dataset are used for evaluation.

[5] http://www.cs.ubc.ca/~mbrown/patchdata/patchdata.html (accessed on June 8th, 2014).

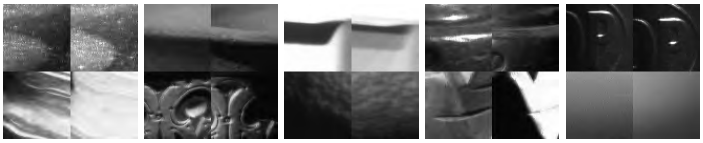
***Virtual World* Dataset**

This dataset[6] contains 3 000 photorealistic images of a virtual city model and has been used by [Kaneva et al., 2011] for image descriptor evaluation in the same manner as the real image patches in [Brown et al., 2011]. Likewise to [Kaneva et al., 2011], 120 000 true patch pairs were extracted by identifying corresponding Difference-of-Gaussians keypoints between viewpoints and different times of the day to introduce changing lighting conditions. Finally, the patches were resized to the standard size of 64×64 based on the detected scale. The advantage of this dataset over the *Liberty* dataset is that it exhibits a more controlled and evenly distributed variation of the lighting conditions, as each scene was rendered under five different lighting conditions (different times of the day), as shown in Figure 4.25.

Examples of true patch pairs contained in the datasets are shown in Figure 4.26. Please notice that, in contrast to the patches of the general evaluation (see Figure 4.15), the image structures of the patch pairs are not perfectly aligned. It also has to be noted that the correct patch pairs of all datasets show no rotation differences. Hence, to allow for a fair comparison, in the presented evaluation the rotation-variant versions of the descriptors are used, except for MROGH, MRRID and LIOP which are inherently rotation invariant. The other descriptors can be made rotation invariant by determining a canonical orientation per patch and describing the per-pixel features and cells relative to this orientation [Lowe, 2004]. The same principle can be used to make LIDRIC rotation-invariant, although it is not treated in this work.

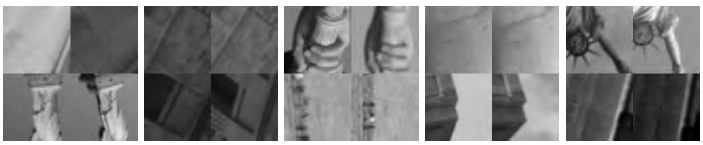
[6] http://people.csail.mit.edu/biliana/projects/iccv2011/ (accessed on June 8th, 2014).

Figure 4.25: Examples of scene renderings contained in the *Virtual World* dataset simulating five different times of the day.

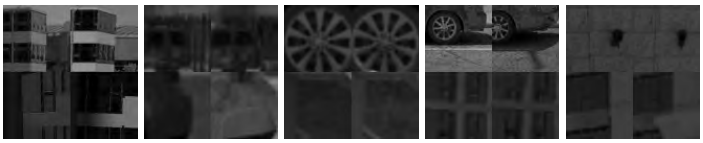


(a) *ALOI Textureless*

(b) *ALOI Textured*



(c) *Liberty*



(d) *Virtual World*

Figure 4.26: Examples of true patch pairs in the datasets.

4.2.2.2 Parameter Selection

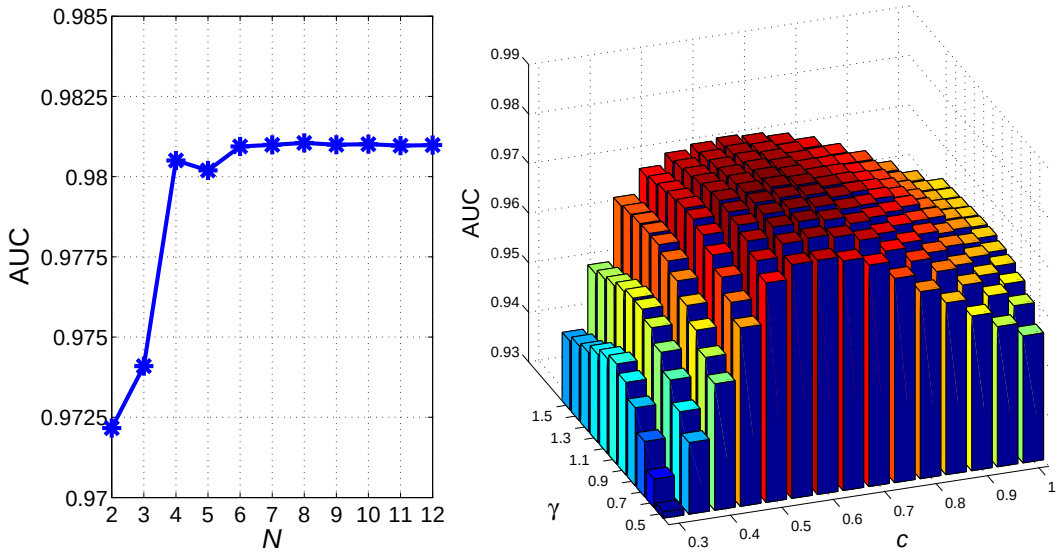
In order to investigate the relation of the LIDRIC parameters to the recognition performance, parameter selections were defined in discrete intervals and the AUC of all parameter combinations on a selected training set was determined. As the main goal of the LIDRIC descriptor is to achieve an optimal performance under strong lighting variations, a mixed dataset was built by randomly extracting 25 000 patches from the representative datasets *ALOI Textureless* and *ALOI Textured*. MSEG4x4 was then modified according to the filter shape parameters c and γ as well as the number of orientations N and tested on the mixed dataset.

Figure 4.27a shows the influence of the number of orientations by plotting the best AUC for a given value of N and all values of c and γ. It can be seen that for $N > 6$ no substantial improvement can be achieved. Therefore, a value of $N = 6$ is chosen for all further experiments. In Figure 4.27b the AUC values for the filter parameters c and γ are shown. It is evident that the best performance is not achieved for filter parameters that make the Gabor filter similar to the second derivative of Gaussian used in [Osadchy et al., 2007] ($c \approx 0.4$), but for values of c close to 0.6 where the Gaussian envelope is wider and thus a higher frequency and orientation resolution is provided [Kamarainen et al., 2006]. Compared to the second derivative of Gaussian as well as the even Gabor filter used in the general evaluation with $c = 0.5$ (see Figure 4.18),

this comes at the price of an increased spatial uncertainty of the filter. However, this seems not to be critical due to the subsequent pooling step. For values of c in this range the aspect ratio γ has only a minor influence on the performance. Based on these results, parameter values of $c = 0.6$ and $\gamma = 1$ are used for the experiments. Figure 4.28 shows the shape of the filter used. Optimal parameters for the multi-scale spacing have been determined on this dataset as $\omega_1 = 2$, $k = \sqrt{2}$ and $M = 8$.

4.2.2.3 Results and Discussion

At first, the illumination-insensitivity of the proposed descriptor is qualitatively demonstrated by analyzing the descriptor differences on the image pairs shown in Figure 4.2. In Figure 4.29 the LIDRIC descriptors using SSEG are shown on the left along with the absolute distances of vector values, whereas the SIFT equivalents are shown on the right. It can be seen that the LIDRIC descriptor is highly insensitive to the illumination variation effects on all object types, indicated by low difference values of corresponding descriptor vector elements depicted in red.



(a) N (b) c and γ

Figure 4.27: Performance of descriptor MSEG4x4 for different values of N, c and γ.

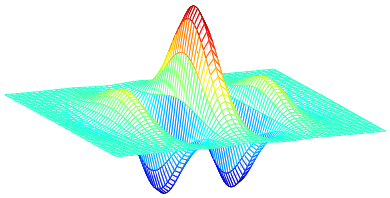


Figure 4.28: The even Gabor filter with $c = 0.6$ and $\gamma = 1.0$ used for LIDRIC.

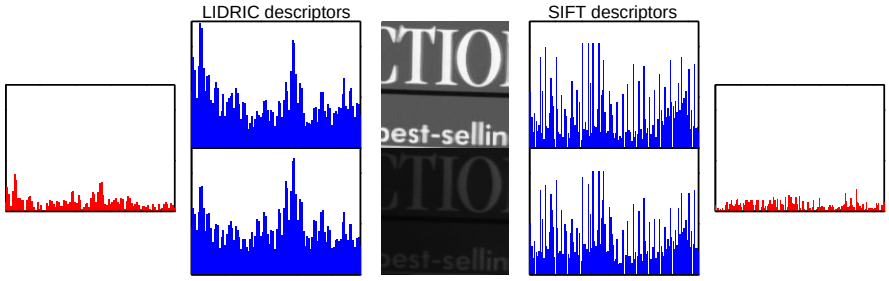
In contrast, SIFT is only robust against the global brightness changes on flat, textured objects.

Results on the four datasets were achieved by applying the descriptors to all patch pairs but excluding the patches used for parameter selection from the *ALOI Textureless* and *ALOI Textured* datasets. The obtained ROC plots are shown in Figure 4.30. The corresponding legends list the descriptors sorted by their achieved AUC given in brackets.

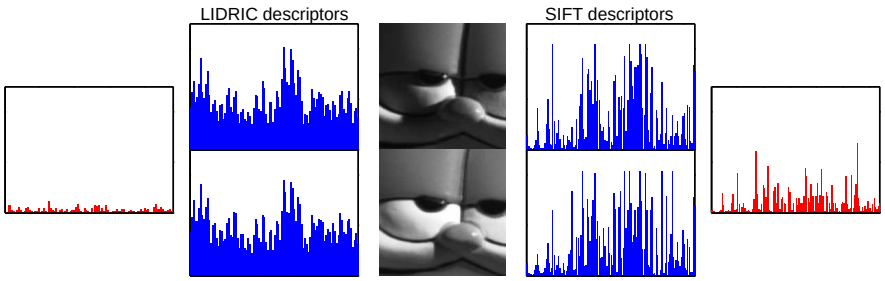
For the datasets representing strong illumination variations between patches (*ALOI Textureless* and *ALOI Textured*), all versions of the LIDRIC descriptor outperform the other descriptors, with a larger advance in performance for the more challenging textureless objects. On both ALOI datasets even the descriptor SSEG with a dimensionality of 6 achieves a better recognition performance than the remaining high-dimensional descriptors with dimensionalities of ≥ 64. MSEG4x4 clearly shows the best performance on these datasets as well as on the *Virtual World* dataset which is assumed to represent more lighting variations than the *Liberty* dataset. On the *Liberty* dataset MSEG4x4 is outperformed by BESTDAISY but shows nearly the same performance as SIFT and DAISY. However, BESTDAISY has been especially optimized for this dataset. It is worth noting that the best parameters for the LIDRIC descriptor were selected according to the datasets *ALOI Textureless* and *ALOI Textured*, but parameter tuning can also be used to improve the results of LIDRIC on the *Liberty* dataset. By using Gabor filters with a shape more similar to a second Gaussian derivative ($c = 0.4$) and $N = 8$, MSEG4x4 is competitive to BESTDAISY on the *Liberty* dataset (AUC=0.9582), while still achieving the best performance on *ALOI Textureless* and *ALOI Textured* (AUC of 0.9563 and 0.9807, respectively). In general, it can be concluded that MSEG4x4 shows the best performance under strong illumination changes for a wide range of filter shapes. Even the worst parameter combination with $c = 0.3$ and $\gamma = 0.5$ (see Figure 4.27b) achieves a better performance than the other descriptors on *ALOI Textureless* (AUC=0.9054, not shown in Figure 4.30).

Among the remaining descriptors, the gradient-based descriptors SIFT and DAISY show the best performance under illumination changes. It is also shown that using unsigned gradients (UGSIFT) is beneficial for the SIFT descriptor by making it invulnerable to edge polarity changes. However, this lowers also the discriminability of the SIFT descriptor and thus its recognition performance is decreased when less illumination changes are present in the data (AUC of 0.9301 compared to 0.9498 on the *Liberty* dataset). Image gradients have previously been mentioned to exhibit illumination-insensitivity properties [Chen et al., 2000, Osadchy et al., 2007] and thus descriptors, which rely on per-pixel features that are not well adapted to the problem of changing lighting conditions (GLAC, MROGH, MRRID, LIOP, SURF, FREAK), have a worse performance compared to SIFT and DAISY.

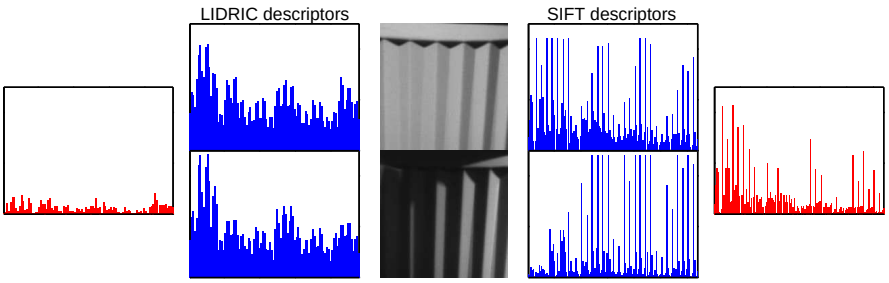
It has been shown in Section 4.1.2 that complex Gabor filters are less distinctive than just using the even filter part for aligned image data. In order to investigate this issue for the non-aligned image data used in this evaluation, the results achieved by either using the even Gabor filters or the entire complex filters for LIDRIC are compared in Table 4.1. It is shown that using the magnitude of both the even and odd filter as feature map does not contribute to considerably better results, while consuming twice the computational power. The advantage of the complex filters is in general that the response is invariant to the phase of the signal, but this does not help to improve illumination insensitivity, as has also been noted by Osadchy et al. [Osadchy et al., 2007].



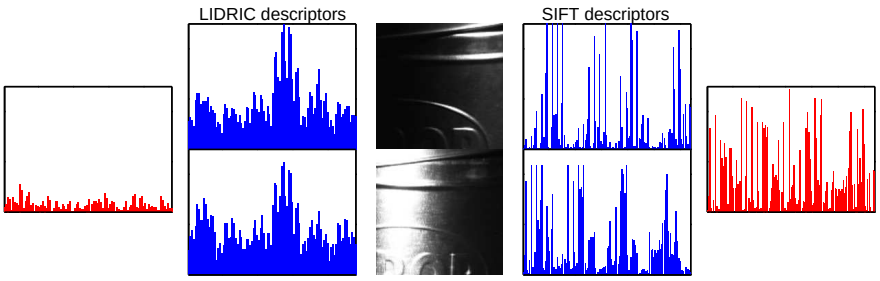
(a) Flat, textured



(b) Non-flat, textured

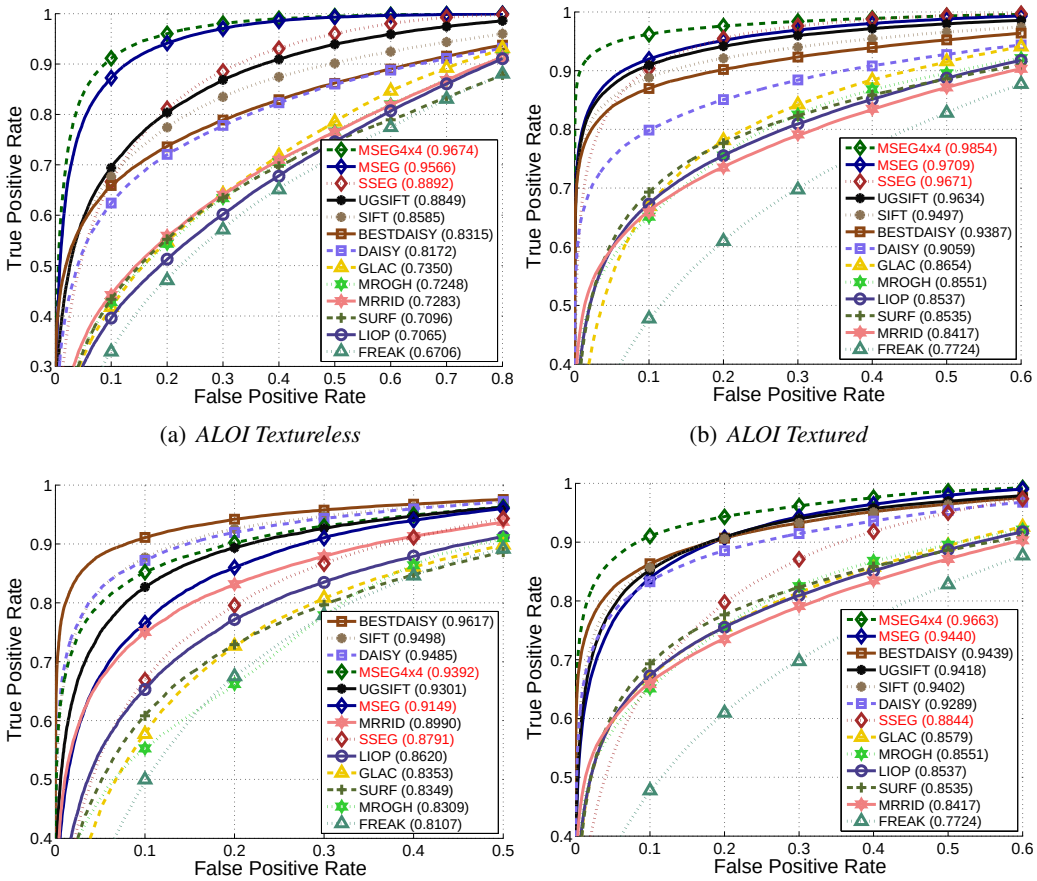


(c) Textureless



(d) Textureless, shiny

Figure 4.29: Comparison of the LIDRIC descriptor using SSEG4x4 and the SIFT descriptor on the image pairs of Figure 4.2. The red bar plots show the absolute differences of respective descriptor vector elements.



(a) *ALOI Textureless* (b) *ALOI Textured*

(c) *Liberty* (d) *Virtual World*

Figure 4.30: ROC curves of the descriptors on the datasets.

	ALOI Textureless	*ALOI Textured*	*Liberty*	*Virtual World*
Even Gabor Filters	0.9674	0.9854	0.9392	0.9663
Complex Gabor Filters	0.9666	0.9855	0.9397	0.9656

Table 4.1: Comparison of AUC values achieved when using even or complex Gabor filter responses for the LIDRIC descriptor.

4.3 Summary

In this chapter the problem of illumination-insensitive recognition of objects in unconstrained conditions is investigated. Therefore, in a preliminary study the discriminative power of various low-level image features for a pixel-wise representation of the underlying surface characteristics of the object. The emphasis of this study is on textureless objects, as this kind of objects

is an under-researched topic in existing literature, although they regularly occur in image data for computer vision tasks (e.g., the classification of coins or the matching of building facades). Hence, a new dataset with rendered images of 3D models is used which allows to directly compare the influences of texture and material properties in an object recognition scenario. The results are further validated on a dataset of real object images and finally reveal that jets of single- and multi-scale even Gabor filter responses are the most powerful low-level representation among the investigated ones. They outperform all other features regardless of object material conditions, levels of texturedness or amount of illumination change, but their superiority is more prominent for the textureless objects which naturally show a higher degree of variation under illumination changes. It is demonstrated that features claimed to be insensitive to illumination conditions based on previous studies, like gradient direction or the self-quotient image, perform substantially worse on textureless surfaces than on textured surfaces.

As a consequence, since popular image descriptors are based on such low-level features, they are likewise affected by illumination effects that are more complex than the simple monotonic brightness changes on textured, flat objects. It is demonstrated that existing descriptors, while performing reasonably well in scenarios with textured objects and only low changes of illumination conditions, show a tremendous decrease of performance in scenarios with strong changes of illumination conditions, especially when textureless objects are involved. The absence of texture on objects as well as strong illumination variations makes the recognition more challenging and this scenario has been neglected in descriptor design and evaluation in the past. Therefore, grounded on the findings of the preliminary study, the LIDRIC descriptor based on even Gabor filter responses is proposed. The descriptor shows to be more robust against the effects caused by changing lighting conditions on non-flat surfaces and thus a recognition performance boost in such scenarios is gained.

CHAPTER 5

Correspondence-Based Image Similarity

This chapter deals with the determination of image similarities based on local correspondences. Therefore, given two images, we want to derive an image similarity measure that tells us how "similar" the objects in the image are. In analogy to Figure 4.1, this means that the similarity metric should be minimized for similar objects and maximized for dissimilar objects. In the presented methods, similarity is defined in terms of class memberships of ancient coins, i.e. two images of coins are said to be similar if they show a coin specimen from the same class.

It has been argued in Sections 2.1.4 and 2.1.5 that the use of local features is subsidiary to overcome image clutter and non-rigid object deformations. However, it remains unclear how geometric constraints can be optimally incorporated for robust similarity measurements. The method presented in Section 5.1 proposes to use the minimal costs from a locally regularized dense correspondence method [Liu et al., 2011] as similarity. The method serves as a proof-of-concept that geometric constraints substantially improve the performance of the coin similarity metric. Consequently, it can be used in an exemplar-based classification framework that is not subject to the limitations of learning-based methods described in Section 1.1. Additionally, a hierarchical subselection scheme is proposed to reduce the classification runtime. This methodology has been originally published in [Zambanini and Kampel, 2011] and [Zambanini and Kampel, 2012].

In Section 5.2, image comparison based on location-aware correspondences is enhanced to an improved similarity measure with lower computational complexity. In this metric the correspondence search is not guided by the feature locations, which are instead utilized for the metric by means of evaluating the geometric plausibility of the matched features. This method has been originally published in [Zambanini et al., 2014], with an extended large-scale evaluation in [Zambanini and Kampel, 2014]. Experimental results for both methods are reported in Section 5.3.

5.1 Dense Feature Matching for Hierarchical Coarse-to-Fine Exemplar-Based Classification

In this section a method for automatically estimating the visual similarity between two coin images is presented and it is shown how this visual similarity can be used in a coarse-to-fine scheme for the ancient coin classification task. In Section 5.1.1 the SIFT flow method is described and the way of using it for ancient coin classification is presented. In Section 5.1.2 an adaptation of the SIFT flow algorithm for handling coin rotations is described. Section 5.1.3 presents the extension of hierarchical coarse-to-fine matching for classification speed-up.

5.1.1 Image Similarity from SIFT Flow

The measurement for coin image similarity is derived from a SIFT flow based image matching [Liu et al., 2011]. SIFT flow aligns images by minimizing an energy function defined over a dense grid of SIFT features. The main application field of this technique highlighted by the authors is scene image retrieval and alignment, i.e. for a given image similar scenes are found and densely aligned. Based on the specific challenges of ancient coin classification described in Section 2.5.2, it is argued that this method is also well suited for coin images as it allows a spatially coherent matching with local variations that is robust to image clutter. As coins from the same class show a similar spatial arrangement of local features, matching these images is assumed to produce a lower energy than matching images from two different classes.

SIFT flow is based on the SIFT descriptor [Lowe, 2004]. The descriptor is computed with a fixed scale and densely over the image, generating a 128-dimensional vector for every pixel, the so called *SIFT image* S. The SIFT images of two coins from the same class are shown in Figure 5.1. It can be seen that the shared image structures are also encoded in the local pixel-wise SIFT features. In order to find an image matching, corresponding SIFT features between two SIFT images S' and S'' have to be determined for each pixel location, represented as a field of flow vectors $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ at grid coordinates $\mathbf{p} = (x, y)$. This is achieved by minimizing the following energy function on $\mathbf{w}$:

$$E(\mathbf{w}) = \sum_{\mathbf{p}} \min(||S'(\mathbf{p}) - S''(\mathbf{p} + \mathbf{w}(\mathbf{p}))||_1, q) \tag{5.1}$$

$$+ \sum_{\mathbf{p}} \kappa(|u(\mathbf{p})| + |v(\mathbf{p})|) \tag{5.2}$$

$$+ \sum_{(\mathbf{p},\mathbf{q})\in\Phi} \min(\beta|u(\mathbf{p}) - u(\mathbf{q})|, d) + \min(\beta|v(\mathbf{p}) - v(\mathbf{q})|, d) \tag{5.3}$$

where Φ contains all four-connected pixel pairs. The energy function is composed of three terms. The *data term* (5.1) computes the L1-distances of all corresponding descriptors and thus measures how similar the local image structures are. The *small displacement term* (5.2) penalizes correspondences that are in different absolute image regions, as it is assumed to be more likely that corresponding image parts share the same image region and thus the flow vectors are small. And finally, the *smoothness term* (5.3) forces the algorithm to produce smooth alignments, i.e. flow vectors of adjacent pixels are similar. The parameters q and d are thresholds for clipping the

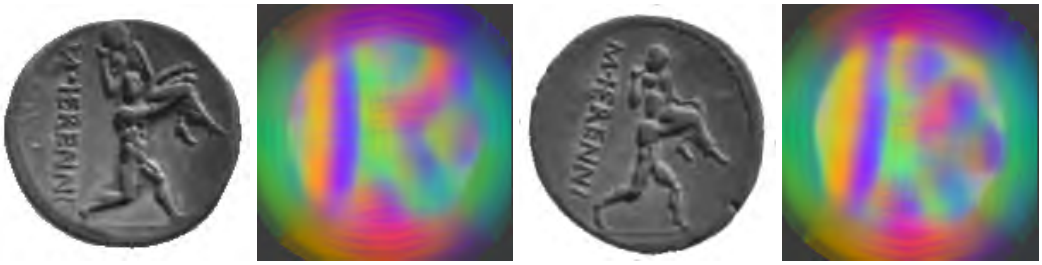


Figure 5.1: SIFT images of two coin images from the same class. The color values are obtained by projecting the 128-dimensional descriptors onto the three principal components with largest eigenvalues, previously determined from SIFT descriptors of a set of images [Liu et al., 2011].

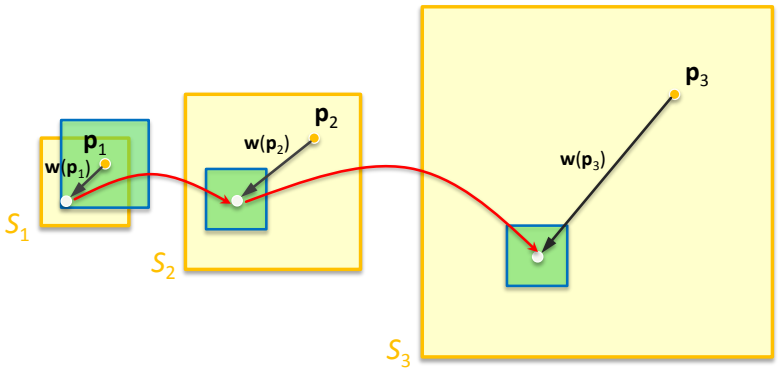
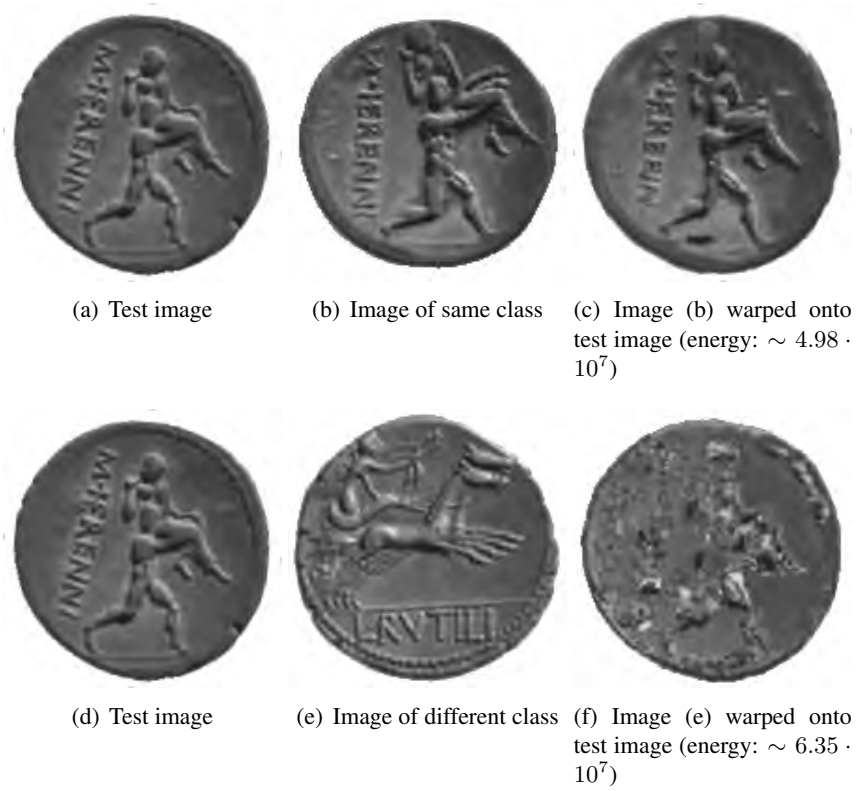


Figure 5.2: Illustration of the coarse-to-fine scheme for finding SIFT flow correspondences.

distances in order to reduce the influence of matched outliers. The parameters κ and β control the relative influences of the individual constraints.

In order to minimize the energy function and to obtain an optimal image matching, a dual-layer belief propagation [Shekhovtsov et al., 2008] is used. Additionally, a coarse-to-fine matching scheme is applied for speed-up and better matching results. This is needed due to the large number of variables and states to be optimized, as correspondences are searched for all image pixels and a pixel of one image can be possibly matched to all pixels of the other image. Therefore, a coarse-to-fine search is applied on a pyramid of SIFT images: initial correspondences are first searched on a coarse resolution and this information is iteratively propagated to the finer resolution layers of the pyramid where the flow vectors are only refined locally. The SIFT image pyramid is generated by consecutively smoothing and downsampling the SIFT images S_K to S_1, where S_{k-1} has half the size in pixel dimensions than S_k. This coarse-to-fine scheme for finding correspondences is illustrated in Figure 5.2. For a given point $\mathbf{p}$ to be matched at each pyramid level k, the best match is found by minimizing $E(\mathbf{w})$ and the found flow vector is used to center the search window for the level $k+1$. At the coarsest level the search window has the same size as the SIFT image of this level, while for the remaining levels the window size is fixed to 11×11. The authors describe that by this scheme the complexity is reduced from $O(h^4)$ to $O(h^2 \log h)$, where h is the width/height of the images.

The adoption of the SIFT flow algorithm for coin classification relies on the following idea: matching two coin images of the same class will likely produce a lower energy $E(\mathbf{w})$ than matching coin images from different classes, since a smooth matching can be more likely found in the former case. An example for this is shown in Figure 5.3. Matching the test coin image with a coin image from the same class produces a reasonable result, as can be seen in Figure 5.3c, where the result of warping the image back to the test image using the SIFT flow vectors is shown. In contrast, matching the test coin image with a coin image of a different class produces an unsuitable result and thus a higher energy. As a consequence, coin classification can be achieved by matching the coin image with all coin images in the database and finally choosing the class of the image with lowest energy. Note that the images have to be normalized with respect to the size of the coin, as the SIFT features are not extracted sparsely with scale detection, but densely with a fixed scale. This is accomplished by the segmentation method proposed in Chapter 3.



(a) Test image (b) Image of same class (c) Image (b) warped onto test image (energy: $\sim 4.98 \cdot 10^7$)

(d) Test image (e) Image of different class (f) Image (e) warped onto test image (energy: $\sim 6.35 \cdot 10^7$)

Figure 5.3: SIFT flow applied to coin images of the same class (top) and images of different classes (bottom).

5.1.2 Insensitivity to Coin Rotations

Although different coin rotations are a rare occasion (see Section 5.3.2), it is worthwhile noting that SIFT flow can be made insensitive to rotations with little adaptation. As coin rotations demand to allow large pixel displacements in the correspondence search, the *small displacement term* regulated by the parameter κ needs to be ignored. Therefore, by setting κ to 0 a rotation between image pairs only affects the correspondence search by producing a slightly larger energy in the smoothness term. It is quantitatively proofed in the experiments in Section 5.3.4 that the influence of the smoothness term in such cases is negligible and that classification performance is not affected by coin rotation differences.

Examples of correspondences found by using the energy function $E(\mathbf{w})$ with $\kappa = 0$ on coin images with rotation differences can be seen in Figure 5.4. Here the query image of Figure 5.4a is matched with an image of a coin from the same class (Figure 5.4b), which produces the correspondences visualized in Figure 5.4c. It can be seen that reasonable correspondences have been found despite the variations between the two coins. If SIFT flow is computed for a rotated version of the coin (Figure 5.4d), the result is almost identical (Figure 5.4e).

(a) Query image (b) Image of same class (c) Image b warped back to query image

(d) Image b rotated by 90 degree (e) Image d warped back to query image

Figure 5.4: Comparison of SIFT flow for coin images with and without rotation differences.

5.1.3 Hierarchical Coarse-To-Fine Classification

A disadvantage of using SIFT flow for example-based classification is that the runtime is linear to the amount of images in the database. However, the coarse-to-fine scheme for correspondence search described in Section 5.1.1 can be utilized by selecting only the most similar coin classes at each level for further processing and thus subsequently reducing the amount of possible target coin classes. This way, the computational effort of the whole classification process is reduced as the more costly computations at finer levels have to be conducted only on a subset of coin classes. An illustration of the proposed method is shown in Figure 5.5.

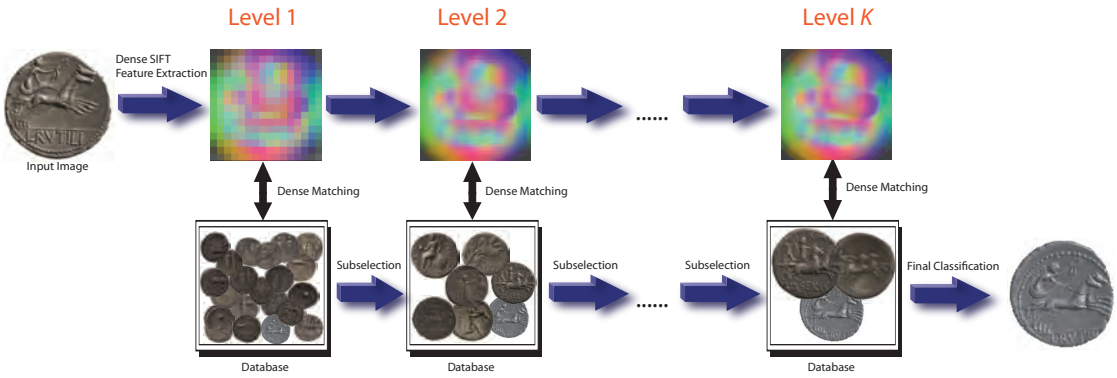


Figure 5.5: Schematic illustrating the proposed coarse-to-fine coin classification procedure. Given an input image, a dense set of SIFT features is extracted and matched against the database at the coarsest level. A defined amount of most similar coin images is selected and forwarded to the matching step of the next finer level. This process is continued until the finest level K is reached where the final classification decision is made.

More formally, a subselection scheme is applied within the SIFT pyramid consisting of K layers $\{S_1, \ldots, S_K\}$. If the set of coin target classes is denoted by $\mathcal{Z}$ and the SIFT flow energy obtained at level k by E_k, classification of a query SIFT image S is achieved in the following manner:

1. For all levels k, $k = 1...K$

   a) Compute SIFT flow energies E_k between S_k and all SIFT images of level k of classes $\mathcal{Z}$.

   b) For each class in $\mathcal{Z}$, compute the average energy $\bar{E}_k$ for all its SIFT images in the database.

   c) Sort all energies $\bar{E}_k$ and reduce $\mathcal{Z}$ by selecting only a percentage ξ_k of $\mathcal{Z}$ with lowest energy.

2. Finally, take the class with lowest energy.

5.2 Improved Similarity from Feature Correspondences by Evaluating Geometric Plausibility (GP)

SIFT flow offers a straightforward way of establishing a dense correspondence field between coin images, and the energy term composed of the feature similarities and their geometric likelihood provides an appropriate dissimilarity metric as non-rigid deformations are dissolved. This shows that it is beneficial to consider the location of features to increase the discriminative power of correspondence-based image similarity. However, SIFT flow enforces dense correspondences (i.e. every single pixel has to be matched to a pixel in the other image) and is thus vulnerable to outliers (non-matchable parts) due to abrasions, although their influence is diminished by the truncation of the energy terms. Additionally, in order to keep the computational complexity of the optimization process low, it only uses weak geometric constraints (i.e. the L1-norm of the 4-connected neighboring flow vectors and absolute displacements). Therefore, in the presented method a similarity metric is established in a different way: instead of regularizing the matching process by geometric constraints, a data-driven first-order matching is performed and constraints are used afterwards to reason about the geometric plausibility (GP) of the most stable correspondences found. This method implicitly excludes outliers and improves the discriminative power of the similarity measure while reducing the computation time. The higher discriminative power comes from the introduced potential of using stronger constraints with higher computational complexity, as the constraints have to be evaluated only once for the given correspondence configuration. Moreover, in contrast to optimization-based approaches like SIFT flow, the "freedom" of data-driven matching contributes to a statistically more meaningful way of using the matching costs as dissimilarity measure: the geometric plausibility of the matched features will be higher for similar coins than for dissimilar coins, as statistically more correspondences are correct. In contrast, in optimization-based approaches the correspondence search is highly forced by the geometric constraints in case of local appearance ambiguities, which consequently reduces the similarity metric's gap between similar and dissimilar image pairs, and hence the discriminative power.

The goal of the proposed exemplar-based coin classification methodology is to estimate the similarity of two coin images robustly against scale differences, illumination conditions, image background and non-rigid deformations. Therefore, it utilizes the previous achievements of this thesis and thus represents the coin classification pipeline shown in Figure 1.2. Robustness against scale differences and image background is achieved by segmenting the coin region in the image (see Chapter 3). Robustness against illumination conditions is accomplished by extracting illumination-insensitive LIDRIC features for matching (Section 5.2.1). Coin similarity insensitive to non-rigid deformations is finally enabled by first-order matching followed by an evaluation of the geometric consistency of the correspondences (Section 5.2.2).

5.2.1 Feature Extraction and First-Order Matching

Similar to SIFT flow, in the proposed method local features are regularly extracted from the image. However, the feature sampling is not done for every pixel but at positions $\mathbf{p}_i = (x_i, y_i)$ on a regular grid with pixel interval $\Delta\mathbf{p} = (\Delta x, \Delta y)$. Dense sampling is fundamental for the effectivity of the method, as more features and thus a statistically more valuable estimation of the

quality of feature correspondences is provided. In contrast to the SIFT flow method presented in Section 5.1 and other works [Kampel and Zaharieva, 2008, Arandjelović, 2012, Anwar et al., 2013], the method does not rely on SIFT features and uses the LIDRIC features instead. However, the original descriptor described in Section 4.2 has to be adapted to be more appropriate for the problem at hand. In contrast to the datasets used for the LIDRIC experiments presented in Section 4.2.2.1, matching is not done for identical objects under changing lighting conditions but rather similar objects, i.e. the underlying local object characteristics are not the same. Consequently, the use of multi-scale responses turned out to be not necessary anymore. In lieu thereof, responses are computed for only one single scale but for 8 orientations. Additionally, instead of dividing each filter response by the pure L2-norm of all 8 responses (denoted as $\|F\|$) for normalization, the power $\|F\|^r$ with $r > 1$ is taken of it before division. This reduces the relative influence of the highest responses in the image which likely arise from highlights on the metallic surface of the coin. Reducing the relative influences of the highest descriptor values has frequently been reported to be beneficial for performance, e.g. by clipping values above a threshold [Lowe, 2004] or by taking the square root of the descriptor values [Arandjelovic and Zisserman, 2012]. Finally, by performing a 4×4 spatial pooling, a 128-dimensional descriptor $\mathbf{d}_i$ for an image point $\mathbf{p}_i$ is obtained.

After the local descriptors $\mathbf{d}'_i \in \mathcal{D}'$ and $\mathbf{d}''_j \in \mathcal{D}''$ have been extracted from the two images I' and I'', it is aimed to find robust matchings between them. An option would be to accept only nearest neighbors with a certain distance to their second nearest neighbors as proposed in [Lowe, 2004], but a one-to-one symmetric search [Zhao et al., 2007] turned out to be the better choice. Two features $\mathbf{d}'_i$ and $\mathbf{d}''_j$ are matched only if $\mathbf{d}''_j$ is the nearest neighbor of $\mathbf{d}'_i$ in $\mathcal{D}''$ and $\mathbf{d}'_i$ is in turn the nearest neighbor of $\mathbf{d}''_j$ in $\mathcal{D}'$. The indices of the descriptors in $\mathcal{D}'$ with a match in $\mathcal{D}''$ are stored in the set $\mathcal{M}$ and the function $\phi(i)$ relates the indices of $\mathcal{D}'$ to the corresponding indices of $\mathcal{D}''$, i.e. $\phi(i) = j$ if $\mathbf{d}'_i$ corresponds to $\mathbf{d}''_j$. Figure 5.6 shows the result of the one-to-one symmetric correspondence search for two coins from the same class and two coins from different classes.

(a) Matching with an image from the same class. (b) Matching with an image from a different class.

Figure 5.6: Results of one-to-one symmetric matching. Only random 10% of the overall correspondences are shown for better illustration.

5.2.2 Similarity Estimation from First-Order Correspondences

The basic assumption of the presented approach is that by first-order matching more correct correspondences can be found for coins from the same class than for coins from different classes. By examining the matching results of similar and dissimilar coins as shown in Figure 5.6 we are able to identify three key observations that lead to the definition of the final similarity measure:

1. Number of Correspondences

The number of matched features is likely to be higher for similar coins than for dissimilar ones. This property is used for a similarity score by

$$\Theta_n = \frac{|\mathcal{M}|}{\min(|\mathcal{D}'|, |\mathcal{D}''|)}. \tag{5.4}$$

2. Displacement of Corresponding Feature Points

The displacement of correct correspondences is low which can be used as a dissimilarity score by

$$\Theta_d = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\mathbf{p}'_i - \mathbf{p}''_{\phi(i)}\|_2 \tag{5.5}$$

with $\|\cdot\|_2$ being the L2-norm.

3. Geometrical Consistency of Correspondences

Pairs of correct correspondences do not drastically change their relative position to each other. Hence, given two points $\mathbf{p}'_i$ and $\mathbf{p}'_j$ and their corresponding points $\mathbf{p}''_{\phi(i)}$ and $\mathbf{p}''_{\phi(j)}$, the vector $\vec{u} = \overline{\mathbf{p}'_i\mathbf{p}'_j}$ will be similar to the vector $\vec{v} = \overline{\mathbf{p}''_{\phi(i)}\mathbf{p}''_{\phi(j)}}$, as illustrated in Figure 5.7. Their difference is computed by

$$\Phi_{(\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}''_{\phi(i)},\mathbf{p}''_{\phi(j)})} = \eta \cdot \psi_{(\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}''_{\phi(i)},\mathbf{p}''_{\phi(j)})} + (1-\eta) \cdot \alpha_{(\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}''_{\phi(i)},\mathbf{p}''_{\phi(j)})} \tag{5.6}$$

$$\psi_{(\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}''_{\phi(i)},\mathbf{p}''_{\phi(j)})} = \frac{\left| \|\mathbf{p}'_i - \mathbf{p}'_j\|_2 - \|\mathbf{p}''_{\phi(i)} - \mathbf{p}''_{\phi(j)}\|_2 \right|}{\|\mathbf{p}'_i - \mathbf{p}'_j\|_2 + \|\mathbf{p}''_{\phi(i)} - \mathbf{p}''_{\phi(j)}\|_2} \tag{5.7}$$

$$\alpha_{(\mathbf{p}'_i,\mathbf{p}'_j,\mathbf{p}''_{\phi(i)},\mathbf{p}''_{\phi(j)})} = \frac{1}{\pi} \arccos\left(\frac{\mathbf{p}'_i - \mathbf{p}'_j}{\|\mathbf{p}'_i - \mathbf{p}'_j\|_2} \cdot \frac{\mathbf{p}''_{\phi(i)} - \mathbf{p}''_{\phi(j)}}{\|\mathbf{p}''_{\phi(i)} - \mathbf{p}''_{\phi(j)}\|_2} \right) \tag{5.8}$$

Intuitively, the terms ψ and α measure the vector difference in terms of length and orientation, respectively, where η serves as weighting parameter. This or a similar vector difference metric is typically used for regularization in optimization-based matching approaches [Berg et al., 2005, Duchenne et al., 2011b, Jorstad et al., 2011, Kim et al., 2013, Liu et al., 2011] in order to penalize matching discontinuities and prefer smooth results. However, for computational reasons only small neighborhoods can be considered (e.g., SIFT flow uses the L1-norm

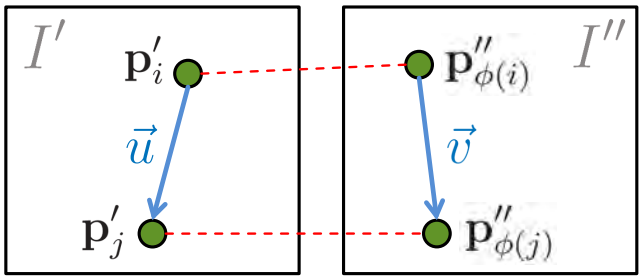


Figure 5.7: The geometric plausibility of the correspondences of the points $\mathbf{p}'_i$ and $\mathbf{p}'_j$ in I' with the points $\mathbf{p}''_{\phi(i)}$ and $\mathbf{p}''_{\phi(j)}$ in I'' is assessed by the comparing the vectors $\vec{u}$ and $\vec{v}$ in terms of length (Eq. 5.7) and orientation (Eq. 5.8).

of the 4-connected neighboring flow vectors). In the given case, these metrics have to evaluated only once for the given first order matching, which allows to use a larger neighborhood system $\mathcal{N}$ for the geometric dissimilarity score:

$$\Theta_g = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \Phi_{(\mathbf{p}'_i, \mathbf{p}'_j, \mathbf{p}''_{\phi(i)}, \mathbf{p}''_{\phi(j)})}. \tag{5.9}$$

In general, one can define all other feature points as the neighborhood $\mathcal{N}_i$ of a given feature point, but this unnecessarily increases the computational burden without substantially improving the quality of this similarity metric. Hence, in practice it turns out to be sufficient to compare every feature point to only a small subset of feature points. In this work it has been empirically chosen to compare every feature point to its neighboring features at the six distances $1\Delta\mathbf{p}$, $2\Delta\mathbf{p}$, $4\Delta\mathbf{p}$, $8\Delta\mathbf{p}$, $12\Delta\mathbf{p}$ and $16\Delta\mathbf{p}$, which on average leads to a comparison of a feature point with around 7.5% of the remaining feature points.

The final overall similarity score is computed from the three individual scores by

$$\Theta = (1 - g(\Theta_n; \sigma_n)) + g(\Theta_d; \sigma_d) + g(\Theta_g; \sigma_g) \tag{5.10}$$

where $g(x; \sigma) = \exp(-x^2/(2\sigma^2))$ is a Gaussian membership function that transforms the individual scores to the same value range.

5.3 Experiments

Empirical evaluation is conducted for coins of the Roman Republican age. Different datasets are used to serve the different requirements for the respective sub-experiments, which are described in Section 5.3.1. In Section 5.3.2 the SIFT flow-based method is compared to a simpler coin matching method that involves no geometric constraints [Kampel and Zaharieva, 2008] in order to demonstrate the usefulness of incorporating geometric constraints into the matching process. In Section 5.3.3 both the classification and runtime performances for various subselection values ξ_k of the coarse-to-fine SIFT flow classification method are reported. In Section 5.3.4 the SIFT flow method is applied to artificially rotated images to show the low vulnerability of the method

to this kind of image variation. A comparison of the SIFT flow method and the improved GP metric as well as of previously proposed learning-based methods is conducted in Section 5.3.5 on a challenging multi-source dataset. Finally, the GP similarity metric is tested for its behavior on a large scale dataset in Section 5.3.6.

5.3.1 Roman Republican Coin Datasets

In total, four coin datasets are used for the experiments which are listed in Table 5.1. All images in the datasets have been successfully scale-normalized in an automated fashion with the help of the proposed shape-controlled segmentation method by means of cropping the images at the coin borders and resizing the resulting region-of-interest to a standard size of 150×150. The *Single-Source Small-Scale* dataset acts as a small, initial demonstrator for the superiority of using geometric constraints for matching instead of feature similarity alone. The *Single-Source Medium-Scale* dataset consists of 60 classes where each class is represented by three coin images of the reverse side. It is used for testing the coarse-to-fine SIFT flow classification as well as for rotation insensitivity tests. Although only a small fraction of the theoretically over 1900 classes are contained in this test dataset, a wide period of minting dates from 199 to 39 B.C. is covered. Sample images of all classes' reverse sides are shown in Figure 5.8.

Dataset	**Classes**	**Images /class**	**Coin sides**	**Sources**	**Purpose**	**Sections**
Single-Source Small-Scale	24	3	obv. & rev.	*Museum of Fine Arts Vienna*	Proof-of-concept for usefulness of geometric matching constraints; comparison of classification performance w.r.t. coin side.	5.3.2
Single-Source Medium-Scale	60	3	rev.	*Museum of Fine Arts Vienna*	Test of SIFT flow classification performance and runtime analysis with hierarchical sub-selection; rotation insensitivity test.	5.3.3 5.3.4
Multi-Source Medium-Scale	60	10	rev.	*Museum of Fine Arts Vienna, British Museum London*, web resources	Comparison of proposed methods (SIFT flow and GP) and previously published learning-based methods for challenging real-world intra-class variations (abrasions, non-rigid deformations, illumination changes).	5.3.5
Single-Source Large-Scale	418	2	obv. & rev.	*Museum of Fine Arts Vienna*	Analysis of the scalability of the proposed GP coin classification method.	5.3.6

Table 5.1: Overview of the four datasets used for the evaluation of the coin classification methods.

For the comparison with other methods the *Multi-Source Medium-Scale* dataset of 60 Roman Republican coin classes is used. For each class 10 images were collected from different image sources to increase the diversity among the images and to mimic a more realistic scenario of coin classification under uncontrolled image acquisition conditions. Three images of each class were taken from the coin collection of the *Museum of Fine Arts, Vienna* (as for the other datasets),

Figure 5.8: Reverse side sample images of all 60 classes of the *Single-Source Medium-Scale* dataset.

another three images from the collection of the *British Museum, London*[1], and the remaining four from free online ancient coin search engines[2]. An example image of each class is shown in Figure 5.9.

The *Multi-Source Medium-Scale* dataset is well suited to comparatively evaluate the performance of coin classification methods for strong intra-class variations, but covers only on a small subset of coinage. Therefore, the real world practicability of the top-performing GP method is additionally investigated by means of the *Single-Source Large-Scale* dataset consisting of 418 classes. This dataset covers a much wider range of coinage from the Roman Republican era and thus allows to give a more qualified statement of achievable classification rates in practice. For this dataset two coin specimens are available per class. The images were gathered from the coin collection of the *Museum of Fine Arts, Vienna*, and all classes of the collections with at least two available coins have been included. This Museum collection belongs to the five largest collections in the world and hence the 418 classes can be seen as a cross section of the most common classes of the over 1900 (including all subclasses) defined in [Crawford, 1974]. Figure 5.10 shows the coin images of 30 classes in the dataset, which demonstrate the challenging nature of this image data for exemplar-based classification. Apart from the intra-class variations, which stem from variations in the manual manufacturing process and abrasions, the variety of the types shown on the coins is narrow due to the strong presence of popular coin themes. For instance, heads of mythological or historical persons are typically shown on the obverse sides [Jones, 1990], and also on the reverse sides motives are shared by multiple classes (e.g. horse teams, standing persons, jugs, animals etc.).

5.3.2 Comparison of Coin Matching with and without Geometric Constraints

In this section the SIFT flow method is compared to the correspondence-based method proposed by [Kampel and Zaharieva, 2008]. In their method, similarity between coins is measured by the number of matched interest points, extracted at Difference-of-Gaussians extrema and described by SIFT.

For this experiment the following empirically determined parameters for SIFT flow matching were used: dense SIFT features were computed for a local neighborhood of 12×12 pixels on the 150×150 images, the number K of pyramid levels was set to 4, and the parameters controlling the influence of the smoothness term were set to $\beta = 12$ and $d = 1200$. The small displacement term was ignored in order to make the similarity metric insensitive to coin rotations. However, it has to be noted that in the provided dataset only marginal rotation differences of less than $20°$ occur. Generally speaking, strong rotation differences between ancient coin images were found to be an uncommon situation, which has also not been encountered in gathering of coin images from the internet for the *Multi-Source Medium-Scale* dataset. The reason is that coins are typically imaged at a canonical orientation based on their type. Nevertheless, the issue of stronger coin rotations of more than $90°$ is treated in 5.3.4 on artificially rotated images to show the low vulnerability of the method to this kind of image variation.

[1] www.britishmuseum.org/research/publications/online_research_catalogues/rrc.aspx (accessed on June 8th, 2014).

[2] www.acsearch.info and www.coinarchives.com (accessed on June 8th, 2014).

Figure 5.9: An exemplar image for each of the 60 classes in the *Multi-Source Medium-Scale* dataset.

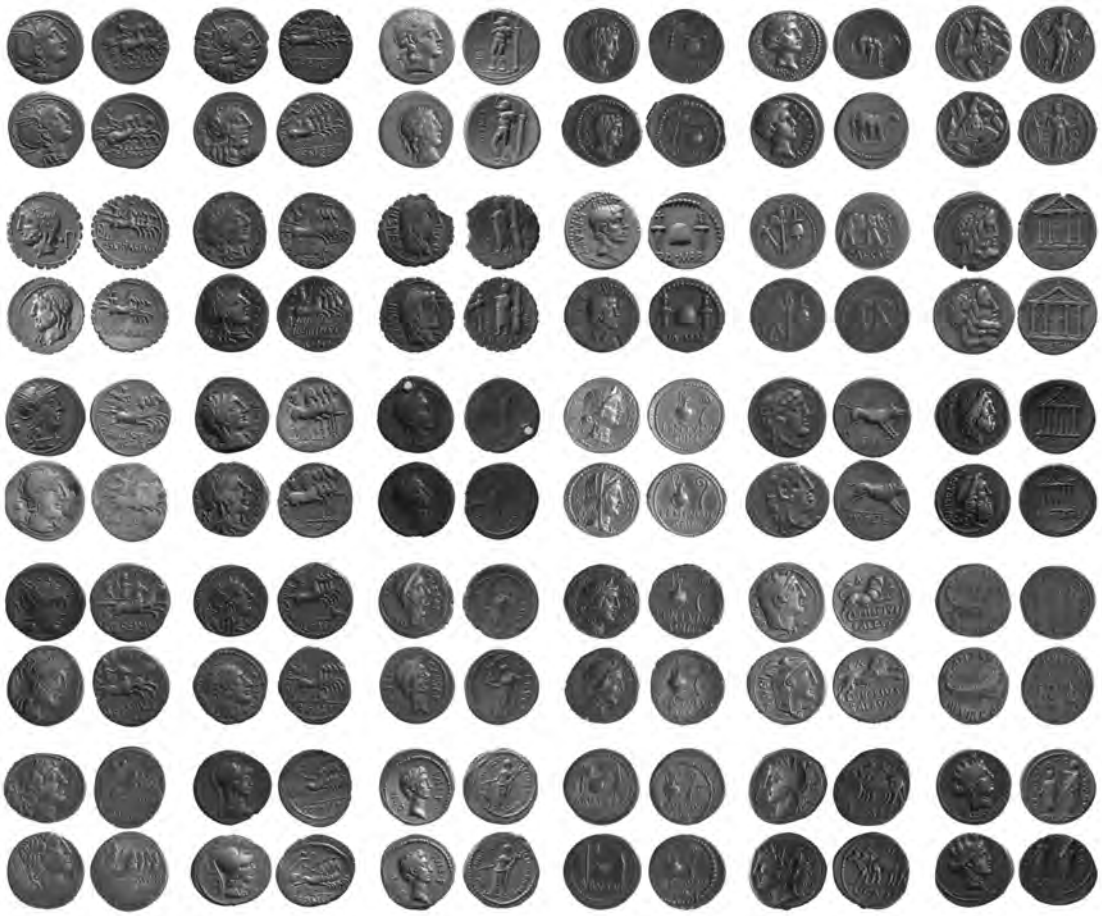


Figure 5.10: Example images from 30 classes of the overall 418 classes in the *Single-Source Large-Scale* dataset. Each image quadruple shows the obverse (left) and reverse (right) of the two coins of a class.

The comparison was performed on the *Single-Source Small-Scale* dataset where either the proposed SIFT flow energy or the number of matched features defined the coin-to-coin similarity. As three images are available per coin class, a 3-fold cross validation was used to test classification performance. The dataset was divided into three subsets, each set containing one image from each class. Three classification runs were executed whereas in each run one subset served as testset and the remaining two served as training set. An image from the testset was then matched with all images from the training set. The average of the similarity values of the two images of a class defined the class-similarity, and thus finally the image was assigned to the class with minimum class-similarity. Classification performance was tested on the obverse as well as the reverse sides.

The overall results are listed in Table 5.2. The SIFT flow method clearly outperforms the method of [Kampel and Zaharieva, 2008]. As the SIFT descriptor itself is potentially noisy on

ancient coins due to their challenging conditions, the simple matching of SIFT interest points is more vulnerable than SIFT flow matching. SIFT flow introduces an additional constraint for a spatially meaningful matching with local variations. Therefore, SIFT flow provides a more robust matching and coin similarity measure.

	SIFT Flow Matching	SIFT Matching
Obverse side	63.9%	25.0%
Reverse side	73.6%	33.3%
Total	**68.8%**	**29.2%**

Table 5.2: Classification rates of the proposed SIFT flow matching and standard SIFT matching [Kampel and Zaharieva, 2008] on the *Single-Source Small-Scale* dataset.

Another observation from the results is that classification rates are higher on the reverse sides of the coin. This is caused by the typical composition of Roman coins from the investigated period. Customarily, obverse sides show the heads of gods or historical persons. For instance, in the given evaluation dataset the obverse side of 15 of the 24 classes depicts the goddess Roma. Reverse sides depict certain scenes and thus have a higher inter-class variability.

5.3.3 Hierarchical Classification Performance and Runtime Analysis of SIFT Flow Method

For this experiment the *Single-Source Medium-Scale* dataset consisting of reverse side coin images was used. In each classification run one of the 180 coin images served as query image and one or two of the remaining images per class served as training images. This led to 180 (two training images per class) or 360 classification runs (one training image per class). For runtime evaluation, the average runtime of computing the SIFT flow between two coin images was measured by using the C++ implementation provided by the authors[3] on a standard machine with a quad-core 2.70 GHz processor. The resulting average SIFT flow matching time was 3.93s, where around 3%, 6%, 21% and 70% are needed for the first, second, third and fourth level, respectively. In Table 5.3 classification results for the two training set sizes as well as various values of ξ_k are shown. Runtimes are indicated as the time for classifying one coin against the dataset of 60 classes, without considering feature extraction of the query image. Subselection parameters of $\xi_1 = \xi_2 = \xi_3 = 100\%$ mean that no subselection was performed. Subselection parameters of $\xi_1 = 1\%$, $\xi_2 = 1\%$ and $\xi_3 = 1\%$ mean that only the energies of the first, second or third level, respectively, were used for classification.

One can see that, without subselection, over 70% of the images can be classified correctly with only one training image per class available. Adding a second training image brings a performance improvement of about $7 - 12\%$. Based on the results on this dataset, a reasonable choice for the subselection parameters is $\xi_1 = 10\%$ and $\xi_2 = \xi_3 = 50\%$. The classification rate is very close to the case without subselection (-2.2% for a training set size of 1 and -0.5% for a training set size of 2, respectively), whereas the runtime improvement is around 93%.

[3] http://people.csail.mit.edu/celiu/SIFTflow/ (accessed on June 8th, 2014).

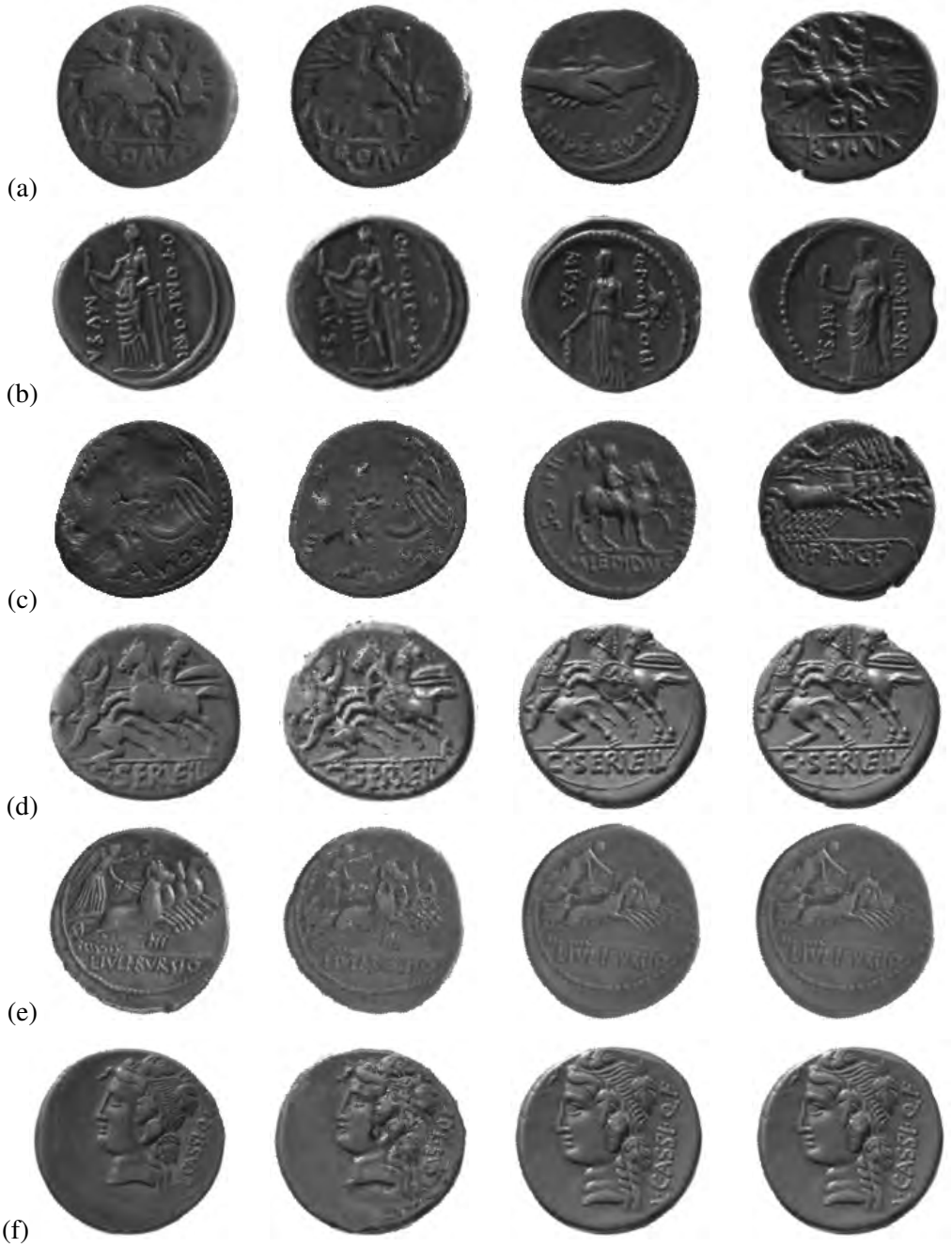
Training set size	ξ_1	ξ_2	ξ_3	Correct classifications	Classification rate	Average classification time
1	100%	100%	100%	257/360	71.4%	235.8s
1	1%	100%	100%	220/360	61.1%	7.1s
1	100%	1%	100%	234/360	65.0%	21.2s
1	100%	100%	1%	257/360	71.4%	70.7s
1	30%	50%	50%	258/360	71.7%	32.5s
1	10%	50%	50%	249/360	69.2%	16.5s
2	100%	100%	100%	150/180	83.3%	471.6s
2	1%	100%	100%	127/180	70.6%	14.1s
2	100%	1%	100%	133/180	73.9%	42.4s
2	100%	100%	1%	141/180	78.3%	141.5
2	30%	50%	50%	150/180	83.3%	65.0s
2	10%	50%	50%	149/180	82.8%	32.9s

Table 5.3: Classification results for training set sizes of 1 and 2 and various subselection values ξ_k on the *Single-Source Medium-Scale* dataset.

In Figure 5.11 some of the classification results are shown where Figure 5.11a-c depict incorrect classifications and Figure 5.11d-f depict correct classifications. It is obvious that strong abrasions, like in Figure 5.11a, as well as the low inter-class variability, like in Figure 5.11b, pose a problem to the method, since the SIFT flow energy becomes less reliable under such conditions. However, also the examples shown in Figure 5.11d-f exhibit strong abrasions and variations between the images which can be dissolved by SIFT flow. Figure 5.11c demonstrates the general limits of image-based ancient coin classification. The query image represents a misprint, which makes it impossible even for human experts to accurately classify the coin if only this coin side is available for examination.

5.3.4 Analysis of Coin Rotation Insensitivity of SIFT Flow Method

In order to assess the sensitivity of the SIFT flow matching to coin rotation differences, a random selection of a query and a training image from 20 coin classes of the *Single-Source Medium-Scale* dataset was done and different coin rotations were simulated by rotating the query image in 90 degree steps. Figure 5.12a shows the classification results of all four runs for the four pyramid levels of SIFT flow matching. In Figure 5.12b the average and maximum increase of energy due to the additional costs in the smoothness term is plotted. It is seen that at a coarser level the energy values are more sensitive to coin rotations, thus producing a decrease of classification performance and a higher relative increase of the energy value. Nevertheless, by using a coarse-to-fine classification with subselection parameters $\xi_1 = 10\%, \xi_2 = \xi_3 = 50\%$, 18 out of 20 classes can be classified correctly for all coin rotation differences. This shows that, although the method is in theory not invariant to coin rotation differences, a high degree of insensitivity is given.



(a)
(b)
(c)
(d)
(e)
(f)

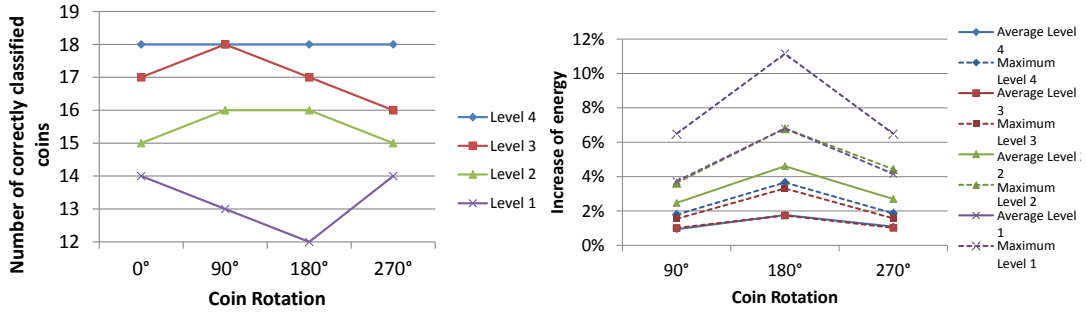
Query image.

Most similar image found warped back using the SIFT flow correspondences.

Original most similar image found.

Correct most similar image depicting a coin of the same class.

Figure 5.11: Six classification results on the *Single-Source Medium-Scale* dataset.



(a) Number of correct classifications.

(b) Average (solid lines) and maximum increase (dashed lines) of SIFT flow energy between images of the same class.

Figure 5.12: Results of evaluating sensitivity to coin rotations.

5.3.5 Comparison of Coin Classification Methods on Multi-Source Dataset

In this section, the GP similarity method is compared to the previous SIFT flow-based method as well as to the learning-based methods using locally biased directional histograms (LBDH) [Arandjelović, 2010] and bag of visual words (BOVW) [Csurka et al., 2004] on the *Multi-Source Medium-Scale* dataset.

For all methods compared suitable parameter choices were empirically determined. Local feature extraction is the first step of all methods and was accomplished either by using SIFT or LIDRIC features with $r = 1.3$ and $\omega = 5$. Dense sampling with $\Delta\mathbf{p} = 3$ for the proposed method and $\Delta\mathbf{p} = 1$ for SIFT flow was performed at a feature scale of 24×24 pixels. For LBDH and BOVW the standard Difference-of-Gaussian interest point detection [Lowe, 2004] was used. Features were only extracted from the coin region in the image provided by the initial coin segmentation step. As no rotation differences are present in the image dataset, for a fair comparison all features were extracted without rotation invariance, i.e. the canonical orientation of all features was automatically set to the same fixed value. Accordingly, the displacement term of SIFT flow controlled by the parameter κ was not set to 0 for these experiments.

For the presented coin similarity algorithm from one-to-one symmetric correspondences parameter values of $\eta = 0.7$, $\sigma_n = 0.1$, $\sigma_d = 50$ and $\sigma_g = 0.25$ were used. For SIFT flow parameter values of $\beta = 200$, $d = 20\,000$ and $\kappa = 12$ were used. For BOVW the descriptors were quantized to 100 visual words and the visual word histogram was computed as image feature. As in the original experiments [Arandjelović, 2010], a vocabulary size of 500 was used for the implementation of LBDH. However, the bandwidth R of the directional kernels was set to 200 instead of 1 000, as this showed superior results on the dataset.

For the final class decision a 5-nearest neighbors classifier was used for all methods compared. The distance of test and training samples was thereby determined by the proposed class similarity or the SIFT flow energy. For BOVW and LBDH the Euclidean distance of the visual word or LBDH histograms was used.

The intention of the proposed exemplar-based coin classification methodology is to achieve coin classification in scenarios with low number of training samples. Therefore, the influence of

	Training images per class		
	1	5	9
BOVW [Csurka et al., 2004] - SIFT	6.3%	11.2%	13.2%
BOVW [Csurka et al., 2004] - LIDRIC	7.6%	12.3%	14.8%
LBDH [Arandjelović, 2010] - SIFT	6.2%	11.9%	14.3%
LBDH [Arandjelović, 2010] - LIDRIC	8.8%	13.1%	16.8%
SIFT flow - SIFT	48.9%	81.0%	90.5%
SIFT flow - LIDRIC	68.6%	90.2%	95.8%
Proposed method - SIFT	68.0%	93.7%	97.1%
Proposed method - LIDRIC (full)	72.7%	94.1%	97.2%
Proposed method - LIDRIC (Θ_n + Θ_g)	70.5%	93.8%	97.2%
Proposed method - LIDRIC (Θ_n + Θ_d)	69.4%	92.9%	97.0%
Proposed method - LIDRIC (Θ_n)	56.0%	84.5%	91.1%

Table 5.4: Numerical classification results of all methods using SIFT/LIDRIC feature extraction on the *Multi-Source Medium-Scale* dataset.

the number of training samples per class to the methods' classification performances is analyzed. For this purpose, multiple classification runs were conducted for each image in the dataset with increasing number of training samples N per class, i.e. $N = 1 \ldots 9$. In each run, N randomly chosen images per class served as training set. This process was again repeated 10 times for each value of N and the overall classification rate out of the $60 \cdot 10 \cdot 10 = 6\,000$ classifications was recorded.

Comparison of Classification Rates

The classification results for the different training set sizes are shown in Figure 5.13 for all methods with LIDRIC feature extraction. Additionally, in Table 5.4 the numerical classification results of all methods with SIFT or LIDRIC feature extraction are listed. It can be seen that the correspondence-based methods dominate the learning-based ones and that the GP method outperforms all other methods for all training set sizes with classification rates from 72.7% ($N = 1$) to 97.2% ($N = 9$). The inclusion of spatial information provided by LBDH gives only a slight improvement over the general BOVW model and does not contribute to a performance comparable to the ones achieved by the correspondence-based methods. Due to the low number of training samples the learning-based methods are not able to sufficiently generalize over the intra-class variation. In the experiments presented in [Arandjelović, 2010] LBDH achieved a classification rate of 57.2% on a 65-class problem. However, the dataset used in Arandjelović's work shows a very uneven distribution of training samples among the classes which are represented by 10 up to 160 exemplars. It can be conjectured that the classification performance of LBDH on this dataset is mainly supported by the classes with a high number of training samples. Another reason for the low classification rate of LBDH is the erroneous interest point detection.

From the results shown in Table 5.4 it can also be concluded that LIDRIC represents a

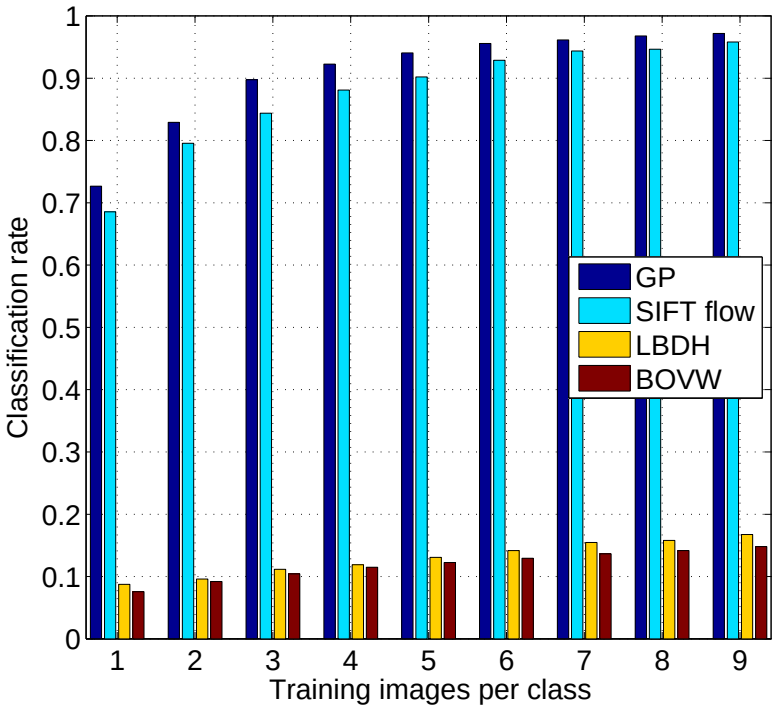


Figure 5.13: Bar plot of the classification results of all methods using LIDRIC feature extraction and training set sizes of 1 to 9 on the *Multi-Source Medium-Scale* dataset.

more powerful local descriptor for coin classification under uncontrolled conditions as its use improves the performance of each individual method. For the proposed method the performance is increased from 68.0% to 72.7% due to LIDRIC's lower sensitivity to illumination changes.

Influence of Individual Similarity Scores

As three single scores are combined for the overall similarity score, it is also of interest to assess their individual influence to the classification performance. It is evidently shown in Table 5.4 that all three scores have a contribution to the classification power of the method. By using only data-driven matching as similarity measure (Θ_n) and ignoring the geometric ones (Θ_d and Θ_g) only 56.0% correct classifications are achieved for $n = 1$, less than the SIFT flow method which also uses geometric information for finding the optimal correspondences (68.6%). Adding geometric information to the model either by the displacement similarity Θ_d or neighboring vector consistency Θ_g leads to classification rates that are higher than that of SIFT flow. The full model with all three terms achieves the highest classification rate of 72.7%.

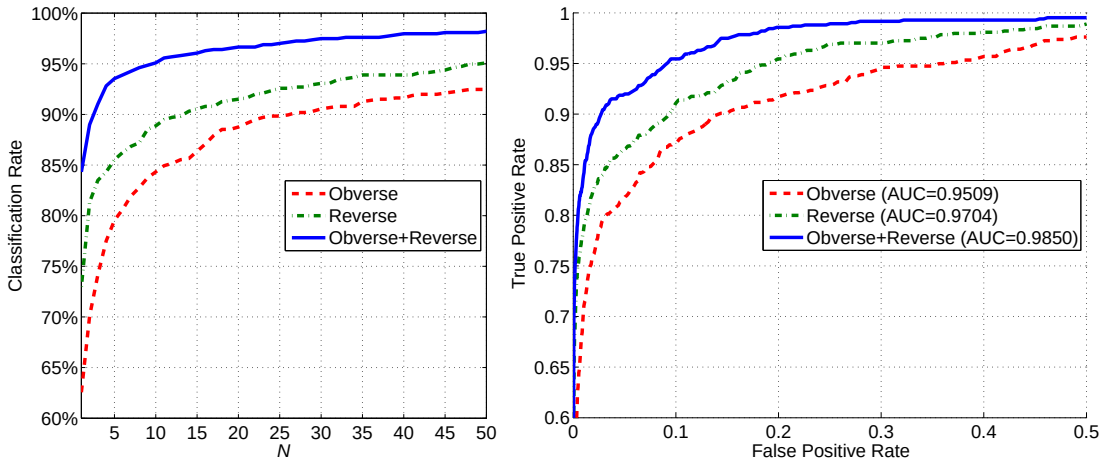
Runtime Analysis

An important issue of exemplar-based classification is the time it takes to compare two image samples. Without feature extraction, which takes around 1s, the MATLAB implementation of

the proposed method needs around $0.35s$ to compare two images whereas the C-implementation of SIFT flow takes around 2.2s. In practice, this means that it takes around 22s to classify a query image for this 60-class problem. However, it has been shown in Section 5.3.3 that the classification time of exemplar-based coin classification in conjunction with feature correspondence can be reduced to one-seventh without a loss of classification accuracy by applying a hierarchical subselection scheme. Generally, the same principle can be applied to the presented similarity metric for speeding up the classification process.

5.3.6 Classification Performance on Large-Scale Single-Source Dataset

The classification rates achieved by the GP similarity method on the large-scale single-source dataset are plotted in Figure 5.14a. The plot shows the respective percentage of coins where the correct class is within the top N similarities. First of all, like in a previous experiment (see Section 5.3.2), it can be observed that a classification based on the coins' reverse side has a better performance than a classification based on the obverse sides. This is caused by the higher variation of reverse side motives and leads to a $\sim 10\%$ higher performance (73.0% vs. 62.6% for $N = 1$). Combining the obverse and reverse side similarities boosts the classification rate by another $\sim 10\%$ (84.3% for $N = 1$). It can also be spotted in Figure 5.14a that there is a comparatively high increase in classification rates from $N = 1$ to $N = 5$: for the combined method the correct class is among the top 5 similarities in 93.5% of the cases. This shows that due to the high number of classes with low inter-class distances an exact classification is challenging, but the ranking is sensitive to the relevant class. The high discriminative power of the coin similarity metric is also demonstrated by the ROC curves shown in Figure 5.14b. These curves are obtained by applying increasing thresholds to the similarities as described in Section 4.1.2.2.



(a) Classification accuracy for the top N classes.

(b) ROC curve.

Figure 5.14: Classification results on the *Single-Source Large-Scale* dataset.

Nevertheless, the curve of 5.14a is flattened more and more for higher values of N which

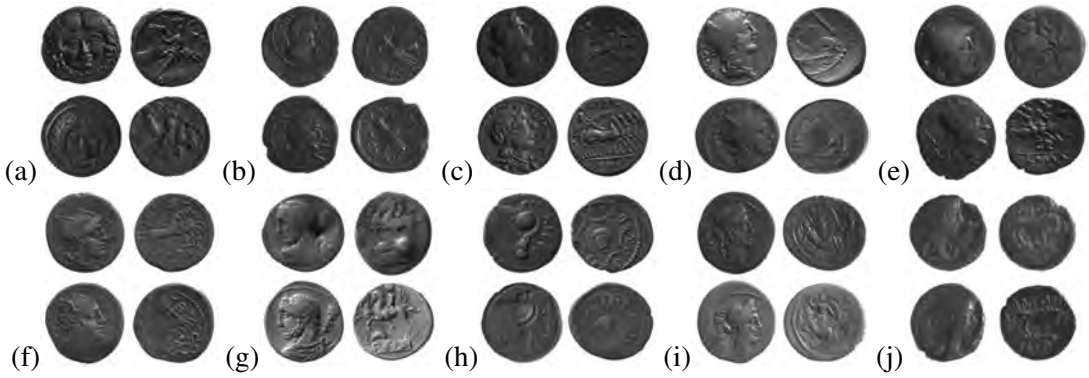


Figure 5.15: (a)-(e) Examples for misclassifications where the correct class is not ranked among the top 50 similarities, (f)-(j) examples for correct classifications.

indicates that there are particular coins where the classification goes completely wrong and the correct class is not ranked properly. For such cases the similarity to the correct class is not considerably higher or even lower than the mean similarity to all incorrect classes. For instance, even with the combined method for 15 out of the 836 coins (i.e. 1.8%) the correct class is not contained in the top 50 similarities. Some of these substantial misclassifications are shown in Figure 5.15a-e to exemplify their diverse causes. One source of error are scale differences between coin motives which are not sufficiently compensated by the scale normalization induced by the segmentation step. As the local features are extracted with a fixed scale, the correspondence search is disturbed by scale changes which is most evidently seen in the obverse side heads shown in Figure 5.15a-b. Strong changes in type appearance (e.g. the chariot of Figure 5.15c) and abrasions (Figure 5.15d-e) are further causes for misclassifications. Nevertheless, the correctly classified coin classes shown in Figure 5.15f-j also exhibit abrasions and motive variations which show that the method is generally able to cope with these types of image variations.

Compared to the results on the multi-source dataset presented in Section 5.3.5, on the large-scale dataset a similar classification rate is achieved (73.0% vs. 72.7% on the multi-source dataset), although seven times as many classes are considered. In contrast to the multi-source dataset, all images of the large-scale dataset come from one single source. This facilitates the classification process, as illumination changes between coin images can be assumed to be much less due to the invariable image acquisition setup. The particular challenge of this dataset is the high number of classes and it is shown that increasing the number of classes has no drastic effect on classification performance.

5.4 Summary

In this chapter the problem of assessing image-to-image similarities from local correspondences is treated. The proposed metrics are included in a exemplar-based coin classification framework to account for the limited availability of training data, which makes standard procedures based on an offline training phase inappropriate. Instead, it is aimed to make the similarity metrics robust against the intra-class variations to handle them directly in the image comparison.

In the first method proposed SIFT flow is exploited for robust dense correspondence between coin images. It is demonstrated that the matching costs are a powerful dissimilarity metric to establish coin classification for training set sizes of one or two images per class. This is shown by a higher classification rate when compared to a previously proposed correspondence-based method which does not include geometric constraints in the similarity metric. Additionally, a coarse-to-fine classification scheme is introduced to decrease runtime which would be otherwise linear to the number of classes in the dataset. It is shown in the experiments that this way the average classification time can be reduced from 471.6s to 32.9s, which is a reduction to 7% of the time needed without hierarchical classification. By using this subselection scheme, the method achieves a classification rate of 83.3% on a dataset of 60 Roman Republican coin classes.

The second method extends the idea of using geometric constraints due to the success of the SIFT flow and uses a new local correspondence-based image similarity metric that is both accurate and fast to compute. The method is designed to be robust against the possible intra-class coin variations like degraded parts, non-rigid deformations and illumination-induced appearance changes. It derives a similarity score by analyzing established data-driven correspondences for their geometric plausibility. Experiments are conducted on a dataset of 60 Roman Republican coin classes from various sources where the presented method achieves classification rates ranging from 72.7% for the case of one training sample per class up to 97.2% when nine training samples per class are used. Consequently, the method shows to deliver a less complex but more distinctive similarity measure than the SIFT flow-based method and outperforms also previously proposed learning-based methods for ancient coin classification. The method is further reviewed for its real-world applicability by evaluation on a large-scale dataset of 418 Roman Republican coin classes. As a result, the method achieves a classification rate of 84.3% when the information from the coins' obverse and reverse side is combined and the correct class is shown to be among the top 5 similarities in 93.5% of the cases. The superior classification performance of the presented method results also from the illumination-insensitive feature extraction by means of the LIDRIC descriptor. The descriptor provides the needed robustness against uncontrolled image acquisition conditions, but at the same time ensures enough discriminative power to establish correct correspondences between coins without needing to guide the correspondence search by regularization.

CHAPTER 6

Conclusions

In this chapter the thesis is concluded by highlighting its main achievements. Additionally, the limitations of the presented methods in their current form are reviewed in Section 6.1. Future work and possible implications for computer vision research and other application areas are discussed in Section 6.2.

The main motivation for the research conducted in this thesis stems from practical considerations with respect to available training data. Modelling image variabilities is a major issue for image recognition that has to be addressed for successful algorithms, and learning from large collections of training images is the widely followed paradigm. This trend has arisen from the increasingly eased access to - probably annotated - image data and the development of sophisticated machine learning methods to handle such large datasets, e.g. deep learning methods [Bengio, 2009]. However, for certain domains with a high number of classes and relatively low number of examples per class this approach is infeasible. In this thesis ancient coin recognition is identified as such a domain where the combined diversity and rarity of classes prevent the successful use of learning methods. As a consequence, in this thesis various aspects of handling image variability are addressed in an exemplar-based classification framework where a query image is compared to class reference images without the need for an offline training phase. This approach has also benefits from a practical point of view: the gathering of large datasets of annotated training images is not needed and class extensions are straightforward by just adding new reference images. In learning-based systems training can take days or weeks and must be repeated from scratch any time new training images or classes are added.

The purpose of the segmentation method presented in Chapter 3 is to be insensitive to image clutter in the background and the object scale. This allows to compute local features at a constant scale, and hence it is not needed to sacrifice a certain amount of discriminative power and reliability by scale-invariant feature detection. The contribution of the presented method to the field of object segmentation is that it points out the use of a simple, scalar confidence measurements to control the segmentation process. Therefore, prior knowledge about the approximate shape of objects can be exploited to achieve a both fast and robust segmentation. This is demonstrated on the annotated coin segmentation testset of 92 images as well as all other coin datasets used

for method evaluation in this thesis, where no segmentation results with less than 94% mutual overlap are encountered.

The research presented in Chapter 4 aims to obtain a deeper understanding of illumination-insensitive feature extraction. A basic evaluation of low-level pixel-wise features for their robustness and discriminative power under illumination changes illumination is conducted. The study is the first one conducted so far that is comprehensive with respect to evaluated features, influences of object material and object texturedness. Especially the type of textureless objects like coins has not been investigated accordingly in the past. The required comprehensiveness is reached by means of a new dataset of synthetically generated images. As a result, jets of oriented Gabor filter responses turn out to be most effective under illumination changes. In the study a comprehensive parameter analysis is conducted and it is further shown that for improved performance one can extend the single-scale representation towards multiple scales by concatenating the single-scale jets.

The outcome of this basic study leads to the development of the LIDRIC descriptor, which shows to outperform existing local descriptors on real images with illumination variations. Regarding the experimental evaluation of the illumination-insensitivity of local descriptors, the shortcomings of existing evaluations are emphasized. Therefore, a dataset of textured as well as textureless objects is used which introduces a greater challenge towards evaluating the robustness against illumination changes than conventional datasets used in the past.

Finally, non-rigid object deformations as for instance occurring between ancient coins of the same class are addressed in Chapter 5. It is deduced from a SIFT flow based method that geometric constraints are crucial for establishing correct dense correspondences between coin images. The proposed method is also inspired by correspondence-based methods for image classification, but does not intend to establish a dense field of correspondences but rather focuses on understanding what the most reliable correspondences tell us about the similarity of the two images. Hence, a similarity metric is derived from checking the geometric plausibility of matched features, instead of using geometric plausibility to guide the correspondence search.

All the single proposed methods treat a specific kind of image variation and can finally be integrated into a holistic ancient coin classification system: the query coin is segmented, dense LIDRIC features are extracted and the correspondence-based similarity to exemplars in the dataset is computed to classify the coin. Therefore, on the application side, the work presented in this thesis contributes to the research of image-based ancient coin classification, a quite new application field in the area of computer vision. Based on the experimental results, the system shows the top performance compared to existing correspondence-based as well as learning-based approaches. Moreover, it shows its real-world applicability by classifying over 400 classes of the Roman Republican period with high accuracy.

6.1 Limitations

Despite the diverse contributions of the proposed methods to the computer vision field, there are certain limitations which are discussed below:

- **Shape-Controlled Object Segmentation**
  - The presented segmentation method is quite application-driven and thus only applicable to nearly-circular objects like coins. However, within the broad field of object segmentation, an applied method accounts for the challenging diversity of the problem, as there is no general state-of-the-art segmentation method applicable for all scenarios. Therefore, designing a segmentation for a specific type of objects can be assumed to be more powerful than applying a general-purpose unsupervised segmentation method, which is also shown in the experiments.
  - The method might fail if its assumptions - the coin is less regular as the surrounding background and the most circular object in the image - are not fulfilled, although this has never been encountered during evaluation.
- **Illumination-Insensitive Feature Extraction**
  - The LIDRIC descriptor is designed to be insensitive to illumination conditions. Evidently, that means that it loses discriminative power when no illumination changes occur. More insensitivity means that more points in the space of all images are mapped to the same representation. Therefore, it is not advisable to use LIDRIC in scenarios where no illumination differences are present, in the same manner as it is not advisable to use the rotation-invariant SIFT descriptor when no image rotations are expected.
  - LIDRIC does only treat the description stage of local image features and not the interest point detection stage. Hence, it is most effectively used with regularly sampled interest points, although it can be used in conjunction with existing interest point detectors. However, existing detectors are not designed for scenes with strong illumination variations and the result is likely to be unsatisfactory.
- **Correspondence-Based Image Similarity**
  - The proposed GP similarity metric is based on data-driven matching and assumes that for similar objects more correspondences are correct. Hence, the method might fail when the objects to be compared are dominated by regular structures and thus no reliable correspondences can be detected without considering their spatial location.
  - In the presented form, the metric is not invariant to object rotations due to involvement of the absolute displacement of features and the rotation-variant comparison of neighboring correspondences. However, the individual terms can be flexibly adapted to extend the metric to other kinds of geometric variations between images.
- **Ancient Coin Classification System**
  - Naturally, the system relies solely on image data and does not use other information like the size and weight of the coin. Consequently, it fails in cases where the image data is also not sufficient for classification for a human expert, e.g. very strong abrasions. This includes also the detection of professional forgeries, which cannot be done from image data.

– Scale changes are compensated by normalizing the images with respect to the size of segmented coin, but scale changes within the motive shown on the coin can be handled only to a certain degree. Strong scale changes heavily affect the dense features which are extracted at a constant scale.

6.2 Future Work and Implications

Shape-Controlled Object Segmentation

The segmentation method is in theory not limited to circular objects but can be applied to all kinds of objects whose shape can be described by a confidence measure which is invariant to global transformations like rotation or scale as well as insensitive to local deformations. Given the proposed shape-controlled thresholding and an appropriate shape descriptor, the segmentation of other shapes like elongated or rectangular ones is possible. A further extension would be to use the formfactor to guide a locally adaptive thresholding method.

Illumination-Insensitive Feature Extraction

Generally, the presented research on illumination-insensitive feature extraction enhances the knowledge for this under-researched topic. The compared low-level features are the basis of many mid-level features and high-level systems, and thus it can also be derived from the presented evaluation how these features and systems act in scenarios with textureless objects and/or strong illumination changes. Consequently, it also helps researchers to select proper features for their application scenarios in the future.

The LIDRIC descriptor shows superior performance for strong appearance variations caused by different illumination directions. This includes one the one hand the matching of highly specular surfaces, as even small viewpoint changes can lead to strong appearance variations due to the inconstant BRDF for different viewing directions. Hence, the descriptor can be used for an improved matching result for applications like stereo vision [Hirschmüller and Scharstein, 2009] or structure-from-motion [Ozden et al., 2010] when such objects are involved. For structure-from-motion also the aspect of changing lighting conditions becomes more important as the images are often collected from different sources and hence were taken at different conditions (e.g. time of the day). The same applies to the problem of place recognition [Lategahn et al., 2013]. Although the majority of objects in the world might be textured, these methods would also benefit from a more reliable matching of textureless object (parts), like facades of buildings, statues etc.

Future work will focus on improving LIDRIC with respect to computational time, dimensionality and rotation invariance. Additionally, the detection of interest points under illumination changes will be investigated.

Correspondence-Based Image Similarity

The proposed GP similarity metric is evaluated on ancient coins, but is not restricted to this application. The method provides a fast and reliable similarity metric for non-rigid deformations of features as occurring, for instance, for human faces or within object categories like letters, cars

etc. The similarity metric can be effectively used for exemplar-based classification when a large number of classes has to be detected and it is impossible or impractical to gather large amounts of training images to handle the intra-class variations. This is not only the case for ancient coin classification but also for domains like writer identification [Bulacu and Schomaker, 2007] or plant leaf recognition [Kumar et al., 2012]. Hence, for future research it is planned to investigate and adapt the image similarity metric to such classes of problems as well as to extend it to other kinds of geometric variations between images. The method allows to flexibly adapt the similarity terms to account for other required geometric invariances. For instance, rotation invariance can be achieved by using rotation-invariant local features and a rotation-invariant evaluation of correspondence consistency, e.g. by using only the distance of pairs of correspondences or by using the length and angles between triplets of correspondences.

For further improvement of the similarity metric, the matching can be achieved on multiple scales to jointly consider smaller and larger structures for similarity. Matching from larger to smaller scales can be also done in a hierarchical subselection process to speed-up the classification process as done for the SIFT flow method.

Notably, the method provides similarity values between images but no dense correspondence field. Another research direction for the future is to derive dense correspondences from the similarity metric, e.g. by seeding the correspondence search on the most similar object parts and iteratively grow the correspondence field.

Ancient Coin Classification System

The ancient coin classification system demonstrates its practical usability by the high classification rate achieved on the large-scale dataset. Considering the fact that simple modifications will lead to further improved results, like a proper selection of reference coins as well as a subselection scheme, the system is ready-for-use and thus able to support numismatists in classifying coins from a certain era like the Roman Republican one. As due to the achieved empirical results the classification can be assumed to be correct or at least highly ranked for the majority of query coins, the presented automatic classification system bears the potential to speed-up and ease their daily work.

Potential is also seen in using the visual similarity estimation in other forms within the application field of numismatics. Visual similarity estimation can be combined with other methods like symbol or legend recognition for a more extensive classification process. It can also be used for automatic coin hoard grouping where a clustering of coins is performed based on the proposed similarity metric.

Bibliography

[Aanæs et al., 2012] Aanæs, H., Dahl, A. L., and Pedersen, K. S. (2012). Interesting interest points. *International Journal of Computer Vision*, 97(1):18–35.

[Adam et al., 2009] Adam, A., Kimmel, R., and Rivlin, E. (2009). On scene segmentation and histograms-based curve evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1708–1714.

[Adams and Bischof, 1994] Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647.

[Adini et al., 1997] Adini, Y., Moses, Y., and Ullman, S. (1997). Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732.

[Agarwal and Triggs, 2006] Agarwal, A. and Triggs, B. (2006). Hyperfeatures - multilevel local coding for visual recognition. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 30–43.

[Agarwala et al., 2004] Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., and Cohen, M. (2004). Interactive digital photomontage. *ACM Transactions on Graphics*, 23(3):294–302.

[Aggarwal et al., 2001] Aggarwal, C., Hinneburg, A., and Keim, D. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Proc. of International Conference on Database Theory (ICDT)*, pages 420–434.

[Ahonen and Pietikäinen, 2009] Ahonen, T. and Pietikäinen, M. (2009). Image description using joint distribution of filter bank responses. *Pattern Recognition Letters*, 30(4):368–376.

[Alahi et al., 2012] Alahi, A., Ortiz, R., and Vandergheynst, P. (2012). Freak: Fast retina keypoint. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–517.

[Anwar et al., 2013] Anwar, H., Zambanini, S., and Kampel, M. (2013). Supporting ancient coin classification by image-based reverse side symbol recognition. In *Proc. of International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 17–25.

[Arandjelović, 2010] Arandjelović, O. (2010). Automatic attribution of ancient roman imperial coins. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1728–1734.

[Arandjelović, 2012] Arandjelović, O. (2012). Reading ancient coins: automatically identifying denarii using obverse legend seeded retrieval. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 317–330.

[Arandjelovic, 2013] Arandjelovic, O. (2013). Making the most of the self-quotient image in face recognition. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7.

[Arandjelovic and Zisserman, 2012] Arandjelovic, R. and Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918.

[Arbeláez et al., 2012] Arbeláez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L., and Malik, J. (2012). Semantic segmentation using regions and parts. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3378–3385.

[Arias et al., 2011] Arias, P., Facciolo, G., Caselles, V., and Sapiro, G. (2011). A variational framework for exemplar-based image inpainting. *International Journal of Computer Vision*, 93(3):319–347.

[Atherton and Kerbyson, 1999] Atherton, T. J. and Kerbyson, D. J. (1999). Size invariant circle detection. *Image and Vision Computing*, 17(11):795–803.

[Ballard, 1981] Ballard, D. H. (1981). Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122.

[Barrow and Tenenbaum, 1978] Barrow, H. and Tenenbaum, J. (1978). Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, pages 3–26.

[Basri and Jacobs, 2003] Basri, R. and Jacobs, D. (2003). Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233.

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359.

[Bay et al., 2006] Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 404–417.

[Belhumeur et al., 1999] Belhumeur, P. N., Kriegman, D. J., and Yuille, A. L. (1999). The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44.

[Belongie et al., 2002] Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

[Bengio, 2009] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

[Berg et al., 2005] Berg, A. C., Berg, T. L., and Malik, J. (2005). Shape matching and object recognition using low distortion correspondences. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26–33.

[Berg and Malik, 2001] Berg, A. C. and Malik, J. (2001). Geometric blur for template matching. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 607–614.

[Bertalmio et al., 2000] Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *SIGGRAPH Conference Proceedings*, pages 417–424.

[Blake and Isard, 1998] Blake, A. and Isard, M. (1998). *Active contours: the application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*. Springer.

[Blake et al., 2011] Blake, A., Kohli, P., and Rother, C. (2011). *Markov random fields for vision and image processing*. MIT Press.

[Boiman et al., 2008] Boiman, O., Shechtman, E., and Irani, M. (2008). In defense of nearest-neighbor based image classification. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[Boix et al., 2012] Boix, X., Gonfaus, J. M., van de Weijer, J., Bagdanov, A. D., Serrat, J., and Gonzàlez, J. (2012). Harmony potentials. *International Journal of Computer Vision*, 96(1):83–102.

[Borenstein and Ullman, 2008] Borenstein, E. and Ullman, S. (2008). Combined top-down/bottom-up segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2109–2125.

[Boureau et al., 2010] Boureau, Y.-L., Ponce, J., and LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proc. of International Conference on Machine Learning (ICML)*, pages 111–118.

[Brown et al., 2011] Brown, M., Hua, G., and Winder, S. (2011). Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):43–57.

[Brox et al., 2011] Brox, T., Bourdev, L., Maji, S., and Malik, J. (2011). Object segmentation by alignment of poselet activations to image contours. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2225–2232.

[Brunelli, 2009] Brunelli, R. (2009). *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons.

[Bulacu and Schomaker, 2007] Bulacu, M. and Schomaker, L. (2007). Text-independent writer identification and verification using textural and allographic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):701–717.

[Calonder et al., 2010] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). Brief: Binary robust independent elementary features. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 778–792.

[Celebi et al., 2007] Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., and Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6):362 – 373.

[Chen et al., 2000] Chen, H., Belhumeur, P., and Jacobs, D. (2000). In search of illumination invariants. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 254–261.

[Chen et al., 2010] Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., and Gao, W. (2010). Wld: a robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720.

[Chen et al., 2008] Chen, J., Shan, S., Zhao, G., Chen, X., Gao, W., and Pietikainen, M. (2008). A robust descriptor based on weber's law. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7.

[Chen et al., 2006] Chen, T., Yin, W., Zhou, X. S., Comaniciu, D., and Huang, T. S. (2006). Total variation models for variable lighting face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1519–1524.

[Chen et al., 2013] Chen, X., Shrivastava, A., and Gupta, A. (2013). Neil: Extracting visual knowledge from web data. In *Proc. of International Conference on Computer Vision (ICCV)*.

[Cheng et al., 2013] Cheng, Z.-Q., Chen, Y., Martin, R., Lai, Y.-K., and Wang, A. (2013). Supermatching: Feature matching using supersymmetric geometric constraints. *IEEE Transactions on Visualization and Computer Graphics*, 19(11):1885–1894.

[Chertok and Keller, 2010] Chertok, M. and Keller, Y. (2010). Efficient high order matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2205–2215.

[Collins et al., 2012] Collins, M. D., Xu, J., Grady, L., and Singh, V. (2012). Random walks based multi-image segmentation: Quasiconvexity results and gpu-based solutions. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1656–1663.

[Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

[Comaniciu et al., 2000] Comaniciu, D., Ramesh, V., and Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 142–149.

[Conte et al., 2004] Conte, D., Foggia, P., Sansone, C., and Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298.

[Cootes et al., 2001] Cootes, T. F., Edwards, G. J., Taylor, C. J., et al. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Crawford, 1974] Crawford, M. H. (1974). *Roman Republican Coinage*. Cambridge University Press.

[Criminisi et al., 2004] Criminisi, A., Pérez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212.

[Crum et al., 2006] Crum, W., Camara, O., and Hill, D. L. G. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461.

[Crum et al., 2004] Crum, W. R., Hartkens, T., and Hill, D. L. G. (2004). Non-rigid image registration: theory and practice. *The British Journal of Radiology*, 77(2):140–153.

[Csurka et al., 2004] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Proc. of Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision*, pages 59–74.

[Cula and Dana, 2004] Cula, O. G. and Dana, K. J. (2004). 3d texture recognition using bidirectional feature histograms. *International Journal of Computer Vision*, 59(1):33–60.

[Datta et al., 2008] Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):5.

[Daugman, 1980] Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856.

[Daugman, 1988] Daugman, J. G. (1988). Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179.

[Deng et al., 2007] Deng, H., Zhang, W., Mortensen, E., Dietterich, T., and Shapiro, L. (2007). Principal curvature-based region detector for object recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

[Dice, 1945] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

[Dong et al., 2013] Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., and Yan, S. (2013). Subcategory-aware object classification. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 827–834.

[Dong et al., 2008] Dong, L., Yu, G., Ogunbona, P., and Li, W. (2008). An efficient iterative algorithm for image thresholding. *Pattern Recognition Letters*, 29(9):1311 – 1316.

[Drbohlav and Chantler, 2005] Drbohlav, O. and Chantler, M. (2005). Illumination-invariant texture classification using single training images. In *Proc. of International Workshop on Texture Analysis and Synthesis, International Conference on Computer Vision*, pages 31–36.

[Dreuw et al., 2009] Dreuw, P., Steingrube, P., Hanselmann, H., Ney, H., and Aachen, G. (2009). Surf-face: Face recognition under viewpoint consistency constraints. In *Proc. of British Machine Vision Conference (BMVC)*, pages 1–11.

[Duarte et al., 2006] Duarte, A., Sánchez, Á., Fernández, F., and Montemayor, A. S. (2006). Improving image segmentation quality through effective region merging using a hierarchical social metaheuristic. *Pattern Recognition Letters*, 27(11):1239–1251.

[Duchenne et al., 2011a] Duchenne, O., Bach, F., Kweon, I.-S., and Ponce, J. (2011a). A tensor-based algorithm for high-order graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2383–2395.

[Duchenne et al., 2011b] Duchenne, O., Joulin, A., and Ponce, J. (2011b). A graph-matching kernel for object categorization. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1792–1799.

[Duda et al., 2012] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern Classification*. John Wiley & Sons.

[Due Trier et al., 1996] Due Trier, i., Jain, A. K., and Taxt, T. (1996). Feature extraction methods for character recognition-a survey. *Pattern Recognition*, 29(4):641 – 662.

[Duncan and Birkhölzer, 1992] Duncan, J. S. and Birkhölzer, T. (1992). Reinforcement of linear structure using parametrized relaxation labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(5):502–515.

[Eveno et al., 2004] Eveno, N., Caplier, A., and Coulon, P.-Y. (2004). Accurate and quasi-automatic lip tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):706–715.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

[Falcão et al., 2004] Falcão, A. X., Stolfi, J., and de Alencar Lotufo, R. (2004). The image foresting transform: Theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):19–29.

[Fan et al., 2012] Fan, B., Wu, F., and Hu, Z. (2012). Rotationally invariant descriptors using intensity order pooling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2031–2045.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

[Felzenszwalb and Huttenlocher, 2004] Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181.

[Felzenszwalb and Huttenlocher, 2005] Felzenszwalb, P. F. and Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79.

[Ferrari et al., 2010] Ferrari, V., Jurie, F., and Schmid, C. (2010). From images to shape models for object detection. *International Journal of Computer Vision*, 87(3):284–303.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

[Fischler and Elschlager, 1973] Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92.

[Fisher et al., 2014] Fisher, R. B., Breckon, T. P., Dawson-Howe, K., Fitzgibbon, A., Robertson, C., Trucco, E., and Williams, C. K. (2014). *Dictionary of Computer Vision and Image Processing*. Wiley, 2nd edition.

[Forsyth, 2011] Forsyth, D. (2011). Variable-source shading analysis. *International Journal of Computer Vision*, 91(3):280–302.

[Freeman and Adelson, 1991] Freeman, W. and Adelson, E. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906.

[Frisby and Stone, 2010] Frisby, J. P. and Stone, J. V. (2010). *Seeing: The computational approach to biological vision*. MIT Press.

[Gabor, 1946] Gabor, D. (1946). Theory of communication. *Journal of the Institute of Electrical Engineers*, 93:429–457.

[Garcia et al., 1998] Garcia, C., Zikos, G., and Tziritas, G. (1998). A wavelet-based framework for face recognition. In *International Workshop on Advances in Facial Image Analysis and Recognition Technology, European Conference on Computer Vision*. Citeseer.

[Gatos et al., 2006] Gatos, B., Pratikakis, I., and Perantonis, S. (2006). Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317 – 327.

[Georghiades et al., 2001] Georghiades, A., Belhumeur, P., and Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660.

[Geusebroek et al., 2005] Geusebroek, J., Burghouts, G., and Smeulders, A. (2005). The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112.

[Gevers and Stokman, 2003] Gevers, T. and Stokman, H. (2003). Classifying color edges in video into shadow-geometry, highlight, or material transitions. *IEEE Transactions on Multimedia*, 5(2):237–243.

[Glasbey, 1993] Glasbey, C. A. (1993). An analysis of histogram-based thresholding algorithms. *Graphical Models and Image Processing*, 55(6):532–537.

[Gold and Rangarajan, 1996] Gold, S. and Rangarajan, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):377–388.

[Gonzalez and Woods, 2002] Gonzalez, R. C. and Woods, R. E. (2002). *Digital Image Processing*. Prentice Hall International.

[Gopalan and Jacobs, 2010] Gopalan, R. and Jacobs, D. (2010). Comparing and combining lighting insensitive approaches for face recognition. *Computer Vision and Image Understanding*, 114(1):135–145.

[Gouiffès et al., 2012] Gouiffès, M., Collewet, C., Fernandez-Maloigne, C., and Trémeau, A. (2012). A study on local photometric models and their application to robust tracking. *Computer Vision and Image Understanding*, 116(8):896–907.

[Grady, 2006] Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783.

[Grana et al., 2013] Grana, C., Serra, G., Manfredi, M., Cucchiara, R., Martoglia, R., and Mandreoli, F. (2013). Unimore at imageclef 2013: Scalable concept image annotation. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. [Cited on page 9]*.

[Grauman and Darrell, 2005] Grauman, K. and Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1458–1465.

[Grierson, 1975] Grierson, P. (1975). *Numismatics*. Oxford University Press.

[Gupta et al., 2010] Gupta, R., Patil, H., and Mittal, A. (2010). Robust order-based methods for feature description. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 334–341.

[H2020, 2013] H2020 (2013). Regulation (eu) no 1291/2013 of the european parliament and of the council of 11 december 2013 establishing horizon 2020 - the framework programme for research and innovation (2014-2020) and repealing decision no 1982/2006/ec (1). *Official Journal of the European Union*, L 347:104–173.

[Hafed and Levine, 2001] Hafed, Z. M. and Levine, M. D. (2001). Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, 43(3):167–188.

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Proc. of Alvey Vision Conference*, pages 147–151.

[Hassner et al., 2012] Hassner, T., Mayzels, V., and Zelnik-Manor, L. (2012). On sifts and their scales. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1528.

[Hays and Efros, 2008] Hays, J. and Efros, A. A. (2008). Scene completion using millions of photographs. *Communications of the ACM*, 51(10):87–94.

[Heath et al., 1998] Heath, M., Sarkar, S., Sanocki, T., and Bowyer, K. (1998). Comparison of edge detectors. A methodology and initial study. *Computer Vision and Image Understanding*, 69(1):38–54.

[Heikkilä et al., 2009] Heikkilä, M., Pietikäinen, M., and Schmid, C. (2009). Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425–436.

[Hirschmüller and Scharstein, 2009] Hirschmüller, H. and Scharstein, D. (2009). Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599.

[Horn, 1989] Horn, B. K. (1989). *Obtaining shape from shading information*. MIT Press.

[Hough, 1962] Hough, P. V. (1962). Method and means for recognizing complex patterns. US Patent 3,069,654.

[Howgego, 2005] Howgego, C. J. (2005). The potential for image analysis in numismatics. In *Images and Artefacts of the Ancient World*, pages 109–113.

[Hua and Akbarzadeh, 2009] Hua, G. and Akbarzadeh, A. (2009). A robust elastic and partial matching metric for face recognition. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 2082–2089.

[Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106.

[Huber et al., 2005] Huber, R., Ramoser, H., Mayer, K., Penz, H., and Rubik, M. (2005). Classification of coins using an eigenspace approach. *Pattern Recognition Letters*, 26(1):61–75.

[Huber-Mörk et al., 2011] Huber-Mörk, R., Zambanini, S., Zaharieva, M., and Kampel, M. (2011). Identification of ancient coins based on fusion of shape and local features. *Machine Vision and Applications*, 22:983–994.

[Jain, 1989] Jain, A. K. (1989). *Fundamentals of digital image processing*. Prentice-Hall, Inc.

[Jegou et al., 2008] Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 304–317.

[Jegou and Zisserman, 2014] Jegou, H. and Zisserman, A. (2014). Triangulation embedding and democratic aggregation for image search. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. accepted.

[Jia et al., 2012] Jia, Y., Huang, C., and Darrell, T. (2012). Beyond spatial pyramids: Receptive field learning for pooled image features. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3370–3377.

[Jones, 1990] Jones, J. M. (1990). *A dictionary of ancient Roman coins*. Spink & Son Ltd.

[Jorstad et al., 2011] Jorstad, A., Jacobs, D., and Trouvé, A. (2011). A deformation and lighting insensitive metric for face recognition based on dense correspondences. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2353–2360.

[Kamarainen et al., 2006] Kamarainen, J., Kyrki, V., and Kalviainen, H. (2006). Invariance properties of gabor filter-based features-overview and applications. *IEEE Transactions on Image Processing*, 15(5):1088–1099.

[Kampel and Zaharieva, 2008] Kampel, M. and Zaharieva, M. (2008). Recognizing ancient coins based on local features. In *Proc. of International Symposium on Visual Computing (ISVC)*, pages 11–22.

[Kaneva et al., 2011] Kaneva, B., Torralba, A., and Freeman, W. (2011). Evaluation of image features using a photorealistic virtual world. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 2282–2289.

[Kapur et al., 1985] Kapur, J., Sahoo, P. K., and Wong, A. (1985). A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image processing*, 29(3):273–285.

[Kavelar et al., 2012] Kavelar, A., Zambanini, S., and Kampel, M. (2012). Word detection applied to images of ancient roman coins. In *Proc. of International Conference on Virtual Systems and Multimedia (VSMM)*, pages 577–580.

[Kavelar et al., 2014] Kavelar, A., Zambanini, S., and Kampel, M. (2014). Reading the legends of roman republican coins. *Journal on Computing and Cultural Heritage*, 7(1).

[Kavelar et al., 2013] Kavelar, A., Zambanini, S., Kampel, M., Vondrovec, K., and Siegl, K. (2013). The ILAC project: Supporting ancient coin classification by means of image analysis. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1(2):373–378.

[Ke and Sukthankar, 2004] Ke, Y. and Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 506–513.

[Kim et al., 2013] Kim, J., Liu, C., Sha, F., and Grauman, K. (2013). Deformable spatial pyramid matching for fast dense correspondences. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2307–2314.

[Kobayashi and Otsu, 2008] Kobayashi, T. and Otsu, N. (2008). Image feature extraction using gradient local auto-correlations. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 346–358.

[Koenderink, 2010] Koenderink, J. J. (2010). *Color for the Sciences*. The MIT Press.

[Kokkinos and Yuille, 2008] Kokkinos, I. and Yuille, A. (2008). Scale invariance without scale selection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[Kuhn, 1955] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

[Kumar et al., 2005] Kumar, M. P., Ton, P., and Zisserman, A. (2005). Obj cut. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18–25.

[Kumar et al., 2012] Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. (2012). Leafsnap: A computer vision system for automatic plant species identification. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 502–516.

[Kumar et al., 2001] Kumar, S., Sallam, M., and Goldgof, D. (2001). Matching point features under small nonrigid motion. *Pattern Recognition*, 34(12):2353–2365.

[Kyrki et al., 2004] Kyrki, V., Kamarainen, J., and Kälviäinen, H. (2004). Simple gabor feature space for invariant object recognition. *Pattern Recognition Letters*, 25(3):311–318.

[Lai et al., 2001] Lai, J. H., Yuen, P. C., and Feng, G. C. (2001). Face recognition using holistic fourier invariant features. *Pattern Recognition*, 34(1):95–109.

[Lai et al., 2014] Lai, K.-T., Yu, F., Chen, M.-S., and Chang, S.-F. (2014). Video event detection by inferring temporal instance labels. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. accepted.

[Larsen et al., 2012] Larsen, A. B. L., Darkner, S., Dahl, A. L., and Pedersen, K. S. (2012). Jet-based local image descriptors. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 638–650.

[Lategahn et al., 2013] Lategahn, H., Beck, J., Kitt, B., and Stiller, C. (2013). How to learn an illumination robust image feature for place recognition. In *Proc. of IEEE Intelligent Vehicles Symposium (IV)*, pages 285–291.

[Lazebnik et al., 2003] Lazebnik, S., Schmid, C., and Ponce, J. (2003). A sparse texture representation using affine-invariant regions. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 319–324.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178.

[Leordeanu and Hebert, 2005] Leordeanu, M. and Hebert, M. (2005). A spectral technique for correspondence problems using pairwise constraints. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1482–1489.

[Leung and Malik, 2001] Leung, T. and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44.

[Leutenegger et al., 2011] Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 2548–2555.

[Li and Jain, 2011] Li, S. Z. and Jain, A. K. (2011). *Handbook of Face Recognition*. Springer.

[Lindeberg, 1998] Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116.

[Ling et al., 2011] Ling, C.-H., Lin, C.-W., Su, C.-W., Chen, Y.-S., and Liao, H.-Y. M. (2011). Virtual contour guided video object inpainting using posture mapping and retrieval. *IEEE Transactions on Multimedia*, 13(2):292–302.

[Liu et al., 2011] Liu, C., Yuen, J., and Torralba, A. (2011). Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994.

[Liu and Dai, 2009] Liu, C.-C. and Dai, D.-Q. (2009). Face recognition using dual-tree complex wavelet features. *IEEE Transactions on Image Processing*, 18(11):2593–2599.

[Liu et al., 2007] Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262 – 282.

[Lowe, 2004] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1150–1157.

[Mainali et al., 2013] Mainali, P., Lafruit, G., Yang, Q., Geelen, B., Van Gool, L., and Lauwereins, R. (2013). Sifer: Scale-invariant feature detector with error resilience. *International Journal of Computer Vision*, 104(2):172–197.

[Manjunath and Ma, 1996] Manjunath, B. S. and Ma, W.-Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842.

[Marr and Hildreth, 1980] Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217.

[Matas et al., 2004] Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767.

[Mehtre et al., 1997] Mehtre, B. M., Kankanhalli, M. S., and Lee, W. F. (1997). Shape measures for content based image retrieval: A comparison. *Information Processing and Management*, 33(3):319 – 337.

[Mikolajczyk, 2002] Mikolajczyk, K. (2002). *Detection of local features invariant to affine transformations*. PhD thesis, Institut National Polytechnique de Grenoble-INPG.

[Mikolajczyk and Schmid, 2004] Mikolajczyk, K. and Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86.

[Mikolajczyk and Schmid, 2005] Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630.

[Mičušík and Hanbury, 2006] Mičušík, B. and Hanbury, A. (2006). Automatic image segmentation by positioning a seed. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 468–480.

[Moeslund et al., 2006] Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2–3):90 – 126.

[Moravec, 1980] Moravec, H. P. (1980). Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, DTIC Document.

[Moreels and Perona, 2007] Moreels, P. and Perona, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73(3):263–284.

[Netz and Osadchy, 2011] Netz, A. and Osadchy, M. (2011). Using specular highlights as pose invariant features for 2d-3d pose estimation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 721–728.

[Ning et al., 2010] Ning, J., Zhang, L., Zhang, D., and Wu, C. (2010). Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition*, 43(2):445–456.

[Nixon and Aguado, 2012] Nixon, M. S. and Aguado, A. S. (2012). *Feature Extraction & Image Processing for Computer Vision*. Academic Press.

[Nölle et al., 2003] Nölle, M., Penz, H., Rubik, M., Mayer, K., Holländer, I., and Granec, R. (2003). Dagobert - a new coin recognition and sorting system. In *Proc. of the International Conference on Digital Image Computing - Techniques and Applications (DICTA)*, pages 329–338.

[Nölle et al., 2006] Nölle, M., Rubik, M., and Hanbury, A. (2006). Results of the muscle cis coin competition 2006. In *Proc. of the MUSCLE CIS Coin Competition Workshop*, pages 1–5.

[Nowak et al., 2006] Nowak, E., Jurie, F., and Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 490–503.

[Ojala et al., 1996] Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59.

[Osadchy et al., 2007] Osadchy, M., Jacobs, D., and Lindenbaum, M. (2007). Surface dependent representations for illumination insensitive image comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):98–111.

[Otsu, 1975] Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27.

[Ozden et al., 2010] Ozden, K. E., Schindler, K., and Van Gool, L. (2010). Multibody structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1134–1141.

[Paulin et al., 2014] Paulin, M., Revaud, J., Harchaoui, Z., Perronnin, F., and Schmid, C. (2014). Transformation pursuit for image classification. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. accepted.

[Pele and Werman, 2009] Pele, O. and Werman, M. (2009). Fast and robust earth mover's distances. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 460–467.

[Peng et al., 2011] Peng, B., Zhang, L., Zhang, D., and Yang, J. (2011). Image segmentation by iterated region merging with localized graph cuts. *Pattern Recognition*, 44(10):2527–2538.

[Philbin et al., 2007] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[Pishchulin et al., 2012] Pishchulin, L., Gass, T., Dreuw, P., and Ney, H. (2012). Image warping for face recognition: From local optimality towards global optimization. *Pattern Recognition*, 45(9):3131–3140.

[Prados and Faugeras, 2006] Prados, E. and Faugeras, O. (2006). Shape from shading. In Paragios, N., Chen, Y., and Faugeras, O., editors, *Handbook of Mathematical Models in Computer Vision*, pages 375–388. Springer.

[Rabin et al., 2008] Rabin, J., Delon, J., and Gousseau, Y. (2008). Circular earth mover's distance for the comparison of local features. In *Proc. of International Conference on Pattern Recognition (ICPR)*, pages 1–4.

[Reisert et al., 2006] Reisert, M., Ronneberger, O., and Burkhardt, H. (2006). An efficient gradient based registration technique for coin recognition. In *Proc. of the MUSCLE CIS Coin Competition Workshop*, pages 19–31.

[Ren and Malik, 2003] Ren, X. and Malik, J. (2003). Learning a classification model for segmentation. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 10–17.

[Russ, 2011] Russ, J. C. (2011). *The Image Processing Handbook*. CRC Press, 6th edition.

[Russakovsky et al., 2012] Russakovsky, O., Lin, Y., Yu, K., and Fei-Fei, L. (2012). Object-centric spatial pooling for image classification. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 1–15.

[Russell et al., 2008] Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173.

[Savarese et al., 2006] Savarese, S., Winn, J., and Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2033–2040.

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

[Schmid et al., 2000] Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer vVision*, 37(2):151–172.

[Scott and Longuet-Higgins, 1991] Scott, G. L. and Longuet-Higgins, H. C. (1991). An algorithm for associating the features of two images. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 244(1309):21–26.

[Seidenari et al., 2014] Seidenari, L., Serra, G., Bagdanov, A., and Del Bimbo, A. (2014). Local pyramidal descriptors for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):1033–1040.

[Sezgin and Sankur, 2004] Sezgin, M. and Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–165.

[Shanbhag, 1994] Shanbhag, A. G. (1994). Utilization of information measure as a means of image thresholding. *Graphical Models and Image Processing*, 56(5):414–419.

[Shashua and Riklin-Raviv, 2001] Shashua, A. and Riklin-Raviv, T. (2001). The quotient image: Class-based re-rendering and recognition with varying illuminations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):129–139.

[Shattuck et al., 2009] Shattuck, D. W., Prasad, G., Mirza, M., Narr, K. L., and Toga, A. W. (2009). Online resource for validation of brain segmentation methods. *NeuroImage*, 45(2):431 – 439.

[Shechtman and Irani, 2007] Shechtman, E. and Irani, M. (2007). Matching local self-similarities across images and videos. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[Shekhovtsov et al., 2008] Shekhovtsov, A., Kovtun, I., and Hlaváč, V. (2008). Efficient mrf deformation model for non-rigid image matching. *Computer Vision and Image Understanding*, 112(1):91–99.

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

[Shi et al., 2012] Shi, J., Ray, N., and Zhang, H. (2012). Shape based local thresholding for binarization of document images. *Pattern Recognition Letters*, 33(1):24 – 32.

[Shirdhonkar and Jacobs, 2005] Shirdhonkar, S. and Jacobs, D. W. (2005). Non-negative lighting and specular object recognition. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1323–1330.

[Simonyan et al., 2012] Simonyan, K., Vedaldi, A., and Zisserman, A. (2012). Descriptor learning using convex optimisation. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 243–256.

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1470–1477.

[Smeulders et al., 2000] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

[Sonka et al., 2007] Sonka, M., Hlavac, V., and Boyle, R. (2007). *Image Processing, Analysis, and Machine Vision*. Thomson-Engineering.

[Stanco et al., 2011] Stanco, F., Battiato, S., and Gallo, G. (2011). *Digital imaging for cultural heritage preservation: Analysis, restoration, and reconstruction of ancient artworks*. CRC Press.

[Swain and Ballard, 1991] Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

[Szeliski, 2010] Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer.

[Tan and Triggs, 2010] Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650.

[Tang et al., 2009] Tang, F., Lim, S., Chang, N., and Tao, H. (2009). A novel feature descriptor invariant to complex brightness changes. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2631–2638.

[Tappen et al., 2006] Tappen, M. F., Adelson, E. H., and Freeman, W. T. (2006). Estimating intrinsic component images using non-linear regression. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1992–1999.

[Tola et al., 2010] Tola, E., Lepetit, V., and Fua, P. (2010). Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815 –830.

[Tola et al., 2008] Tola, E., V.Lepetit, and Fua, P. (2008). A Fast Local Descriptor for Dense Matching. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

[Torralba et al., 2008] Torralba, A., Fergus, R., and Freeman, W. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970.

[Torresani et al., 2013] Torresani, L., Kolmogorov, V., and Rother, C. (2013). A dual decomposition approach to feature correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):259–271.

[Trulls et al., 2014] Trulls, E., Tsogkas, S., Kokkinos, I., Sanfeliu, A., and Moreno-Noguer, F. (2014). Segmentation-aware deformable part models. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. accepted.

[Tschumperlé and Deriche, 2003] Tschumperlé, D. and Deriche, R. (2003). Vector-valued image regularization with pdes: A common framework for different applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):506–517.

[Van De Sande et al., 2010] Van De Sande, K., Gevers, T., and Snoek, C. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596.

[Van De Weijer et al., 2006] Van De Weijer, J., Gevers, T., and Smeulders, A. W. (2006). Robust photometric invariant features from the color tensor. *IEEE Transactions on Image Processing*, 15(1):118–127.

[Van Der Maaten and Poon, 2006] Van Der Maaten, L. and Poon, P. (2006). Coin-o-matic: A fast system for reliable coin classification. In *Proc. of the MUSCLE CIS Coin Competition Workshop*, pages 7–18.

[Varma and Ray, 2007] Varma, M. and Ray, D. (2007). Learning the discriminative power-invariance trade-off. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1–8.

[Varma and Zisserman, 2005] Varma, M. and Zisserman, A. (2005). A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2):61–81.

[Velho et al., 2011] Velho, L., Carvalho, P., Gomes, J., and de Figueiredo, L. (2011). *Mathematical optimization in computer graphics and vision*. Morgan Kaufmann.

[Vincent and Soille, 1991] Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598.

[Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

[Wallraven et al., 2003] Wallraven, C., Caputo, B., and Graf, A. (2003). Recognition with local features: the kernel recipe. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 257–264.

[Wang et al., 2004] Wang, H., Li, S., and Wang, Y. (2004). Face recognition under varying lighting conditions using self quotient image. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 819–824.

[Wang et al., 2011] Wang, Z., Fan, B., and Wu, F. (2011). Local intensity order pattern for feature description. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 603–610.

[Weiss, 2001] Weiss, Y. (2001). Deriving intrinsic images from image sequences. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 68–75.

[Wright et al., 2009] Wright, J., Yang, A., Ganesh, A., Sastry, S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227.

[Wu et al., 2009] Wu, Z., Ke, Q., Isard, M., and Sun, J. (2009). Bundling features for large scale partial-duplicate web image search. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25–32.

[Xiang et al., 2014] Xiang, Y., Mottaghi, R., and Savarese, S. (2014). Pascal 3d: A benchmark for 3d pose estimation in the wild. In *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*. accepted.

[Xie et al., 2014] Xie, L., Wang, J., Guo, B., Zhang, B., and Tian, Q. (2014). Orientational pyramid matching for recognizing indoor scenes. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. accepted.

[Xie et al., 2011] Xie, X., Zheng, W.-S., Lai, J., Yuen, P., and Suen, C. (2011). Normalization of face illumination based on large-and small-scale features. *IEEE Transactions on Image Processing*, 20(7):1807–1821.

[Yanowitz and Bruckstein, 1989] Yanowitz, S. and Bruckstein, A. (1989). A new method for image segmentation. *Computer Vision, Graphics, and Image Processing*, 46(1):82 – 95.

[Yao and Fei-Fei, 2012] Yao, B. and Fei-Fei, L. (2012). Action recognition with exemplar based 2.5 d graph matching. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 173–186.

[Yao et al., 2014] Yao, C., Bai, X., Si, B. S., and Liu, W. (2014). Strokelets: A learned multi-scale representation for scene text recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. accepted.

[Yilmaz et al., 2006] Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys*, 38(4):13.

[Zaharieva et al., 2007a] Zaharieva, M., Huber-Mörk, R., Nölle, M., and Kampel, M. (2007a). On ancient coin classification. In *Proc. of International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*, pages 55–62.

[Zaharieva et al., 2008] Zaharieva, M., Kampel, M., and Vondrovec, K. (2008). From manual to automated optical recognition of ancient coins. In *Proc. of International Conference on Virtual Systems and Multimedia (VSMM)*, pages 88–99.

[Zaharieva et al., 2007b] Zaharieva, M., Kampel, M., and Zambanini, S. (2007b). Image based recognition of ancient coins. In *Proc. of International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 547–554.

[Zaharieva et al., 2007c] Zaharieva, M., Kampel, M., and Zambanini, S. (2007c). Image based recognition of coins–an overview of the coins project. In *Proc. of 31st AAPR/OAGM Workshop*, pages 57–64.

[Zambanini and Kampel, 2009] Zambanini, S. and Kampel, M. (2009). Robust automatic segmentation of ancient coins. In *Proc. of International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 273–276.

[Zambanini and Kampel, 2011] Zambanini, S. and Kampel, M. (2011). Automatic coin classification by image matching. In *Proc. of International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*, pages 65–72.

[Zambanini and Kampel, 2012] Zambanini, S. and Kampel, M. (2012). Coarse-to-fine correspondence search for classifying ancient coins. In *eHeritage Workshop, Asian Conference on Computer Vision*, pages 25–36.

[Zambanini and Kampel, 2013a] Zambanini, S. and Kampel, M. (2013a). Evaluation of low-level image representations for illumination-insensitive recognition of textureless objects. In *Proc. of International Conference on Image Analysis and Processing (ICIAP)*, pages 71–80.

[Zambanini and Kampel, 2013b] Zambanini, S. and Kampel, M. (2013b). A local image descriptor robust to illumination changes. In *Proc. of Scandinavian Conference on Image Analysis (SCIA)*, pages 11–21.

[Zambanini and Kampel, 2014] Zambanini, S. and Kampel, M. (2014). A large-scale evaluation of correspondence-based coin classification on roman republican coinage. In *EUROGRAPHICS Workshop on Graphics and Cultural Heritage (GCH)*. submitted.

[Zambanini et al., 2013] Zambanini, S., Kavelar, A., and Kampel, M. (2013). Improving ancient roman coin classification by fusing exemplar-based classification and legend recognition. In *Proc. of International Workshop on Multimedia for Cultural Heritage (MM4CH), International Conference on Image Analysis and Processing*, pages 149–158.

[Zambanini et al., 2014] Zambanini, S., Kavelar, A., and Kampel, M. (2014). Classifying ancient coins by local feature matching and pairwise geometric consistency evaluation. In *Proc. of International Conference on Pattern Recognition (ICPR)*. accepted.

[Zhang et al., 2009] Zhang, T., Tang, Y. Y., Fang, B., Shang, Z., and Liu, X. (2009). Face recognition under varying illumination using gradientfaces. *IEEE Transactions on Image Processing*, 18(11):2599–2606.

[Zhang et al., 2000] Zhang, Z., Stoecker, W. V., and Moss, R. H. (2000). Border detection on digitized skin tumor images. *IEEE Transactions on Medical Imaging*, 19(11):1128–1143.

[Zhao et al., 2007] Zhao, W.-L., Ngo, C.-W., Tan, H.-K., and Wu, X. (2007). Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Transactions on Multimedia*, 9(5):1037–1048.

[Zhou et al., 2010a] Zhou, W., Lu, Y., Li, H., Song, Y., and Tian, Q. (2010a). Spatial coding for large scale partial-duplicate web image search. In *Proc. of the International Conference on Multimedia*, pages 511–520.

[Zhou et al., 2010b] Zhou, X., Yu, K., Zhang, T., and Huang, T. (2010b). Image classification using super-vector coding of local image descriptors. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 141–154.

List of Acronyms

ALOI Amsterdam Library of Object Images

AUC Area Under Curve

BOVW Bag Of Visual Words

BRDF Bidirectional Reflectance Distribution Function

BRISK Binary Robust Invariant Scalable Keypoints

CBIR Content-Based Image Retrieval

COINS Combat On-line Illegal Numismatic Sales

CS-LBP Center Symmetric - Local BInary Pattern

DC Dice Coefficient

FPR False Positive Rate

FREAK Fast REtinA Keypoint

GD Gradient Direction

GLAC Gradient Local Auto-Correlation

GO Gradient Orientation

GP Geometric Plausibility

GV Gray Value

HRI-CSLTP Histogram of Relative Intensities - Center-Symmetric Local Ternary Pattern

ICT Information and Communication Technology

ILAC Image-based cLassification of Ancient Coins

JBLD Jet-Based Local Descriptor

JEG Jet of Even Gabor filter responses

JG Jet of Gabor filter responses

JMSEG Jet of Multi-Scale Even Gabor filter responses

JOSD Jet of Oriented Second Derivative filters

LBDH Locally-Biased Directional Histograms

LBP Local Binary Pattern

LDA Linear Discriminant Analysis

LIDRIC Local Image Descriptor Robust to Illumination Changes

LIOP Local Intensity Order Patterns

LoG Laplacian of Gaussian

MROGH Multi-support Region Order-based Gradient Histogram

MRRID Multi-support Region Rotation and Intensity monotonic invariant Descriptor

MSEG Multi-Scale Even Gabor filter

OSID Ordinal Spatial Intensity Distribution

P-SIFT Pyramidal SIFT

PCA Principal Component Analysis

PR Pooling Regions

RANSAC RANdom SAmple Consensus

ROC Receiver Operating Characteristic

SIDIRE Synthetic Image Dataset for Illumination Robustness Evaluation

SIFT Scale Invariant Feature Transform

SQI Self-Quotient Image

SSDESC Self-Similarity DESCriptor

SSD Sum of Squared Distances

SSEG Single-Scale Even Gabor filter

SURF Speeded-Up Robust Features

SVM Support Vector Machine

TPR True Positive Rate

UGSIFT Unsigned Gradient Scale Invariant Feature Transform

WLD Weber Local Descriptor

Sebastian Zambanini

Kreuzgasse 49/2/25
1040 Vienna, Austria
Tel. +43 664 9223093
email: zamba@caa.tuwien.ac.at
web: http://www.caa.tuwien.ac.at/cvl/people/zamba/

Personal Data

- **Date of Birth:** September 8, 1979
- **Place of Birth:** Bregenz, Austria
- **Marital Status:** Unmarried
- **Nationality:** Austrian

Education

- **MSc. Computer Science, Master program Computer Graphics and Digital Image Processing:** Vienna University of Technology 2004-2007 (graduated with distinction)
- **BSc. Computer Science, Bachelor program Media and Computer Science:** Vienna University of Technology 2000-2004
- **Secondary School:** General-education High-School (*AHS*) in Bregenz, Vorarlberg, graduated in 1998

Work Experience

- **University assistant** and **project assistant** for the FWF-funded project **ILAC** at the Computer Vision Lab, Vienna University of Technology, since February 2011
- **Project assistant** for the FFG-funded project **MuBisA**, Pattern Recognition and Image Processing Group, Vienna University of Technology, September 2009 – January 2011
- **University assistant** at the Pattern Recognition and Image Processing Group, Vienna University of Technology, March - August 2009
- **Project assistant** for the EU-funded project **COINS**, Pattern Recognition and Image Processing Group, Vienna University of Technology, February 2007- January 2009
- **National Service:** Community service at the *Caritas* (charity organization) in Feldkirch, Vorarlberg, from October 1998 to September 1999

Teaching

Lecturer for:

- VU "Einführung in Visual Computing" ("**Introduction to Visual Computing**"), summer terms 2012-2014
- VU "**Video Analysis**", Vienna University of Technology, summer terms 2013-2014
- VU "**Computer Vision**", Vienna University of Technology, winter term 2011-2013
- SE "Seminar aus Computer Vision und Mustererkennung" ("**Seminar in Computer Vision and Pattern Recognition**"), Vienna University of Technology, winter terms 2011-2013

- SE "Wissenschaftliches Arbeiten" ("**Scientific Work**"), Vienna University of Technology, winter term 2010-2013
- VU "**Visual Surveillance**", Vienna University of Technology, summer term 2012
- UE "Multimediale Systeme" ("**Multimedia Systems**"), UAS Technikum Wien, winter terms 2009-2012
- VU "Bildfolgen" ("**Analysis of Image Sequences**"), Vienna University of Technology, summer terms 2009-2011
- UE "Statistische Mustererkennung" ("**Statistical Pattern Recognition**"), Vienna University of Technology, winter terms 2010-2011

Tutor for:

- basic courses in **image processing** and **pattern recognition**, Vienna University of Technology, 2004-2007

Miscellaneous

- **Research interests**: Object Recognition, Image Features, Video and Image Retrieval, Image Matching
- **Winner** of the **PRIP prize 2005** for the bachelor thesis entitled “Automatic Measurement of Hemangiomas”