University of Basel

Model-based Image Analysis for Forensic Shoe Print Recognition

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Adam Kortylewski

Basel, 2017

What a strange pleasure it is to seek.
- Stuart Kauffman

Acknowledgments

I am very grateful for the support, guidance and encouragement of many individuals who contributed to this dissertation.

Prof. Thomas Vetter for giving me the opportunity to pursue this journey and for providing great support and the right ideas

Prof. Volker Roth for always asking the right questions and for advice during lab and committee meetings

Prof. Charless Fowlkes for agreeing to serve as an external examiner of this dissertation

Thomas Stadelmann for making this research project possible

I am glad to thank my colleagues who have taken the time to read parts of this manuscript: Mario Wieser, Sonali Parbhoo, Marcel Luethi, Sandro Schoenborn and Bernhard Egger

I want to especially thank Clemens Blumer and Andreas Morel-Forster for providing valuable feedback on every part of this manuscript

Vitali Nesterov for implementing the baseline algorithms on shoe print recognition

The people at NIST who provided valuable feedback on the algorithm and ideas on the evaluation: Martin Hermann, Hariharan Iyer, Steven Lund, Gunay Dogan, Yooyoung Lee

Nadine Callegaro for providing me with unfailing support. Without you, this would not have been possible.

Thank you

Contents

Abstract

This thesis is about automated forensic shoe print recognition. Recognizing a shoe print in an image is an inherently difficult task. Shoe prints vary in their pose, shape and appearance. They are surrounded and partially occluded by other objects and may be left on a wide range of diverse surfaces. We propose to formulate this task in a model-based image analysis framework.

Our framework is based on the Active Basis Model. A shoe print is represented as hierarchical composition of basis filters. The individual filters encode local information about the geometry and appearance of the shoe print pattern. The hierarchical composition encodes mid- and long-range geometric properties of the object. A statistical distribution is imposed on the parameters of this representation, in order to account for the variation in a shoe print‘s geometry and appearance.

Our work extends the Active Basis Model in various ways, in order to make it robustly applicable to the analysis of shoe print images. We propose an algorithm that automatically infers an efficient hierarchical dependency structure between the basis filters. The learned hierarchical dependencies are beneficial for our further extensions, while at the same time permitting an efficient optimization process. We introduce an occlusion model and propose to leverage the hierarchical dependencies to integrate contextual information efficiently into the reasoning process about occlusions. Finally, we study the effect of the basis filter on the discrimination of the object from the background. In this context, we highlight the role of the hierarchical model structure in terms of combining the locally ambiguous filter response into a sophisticated discriminator.

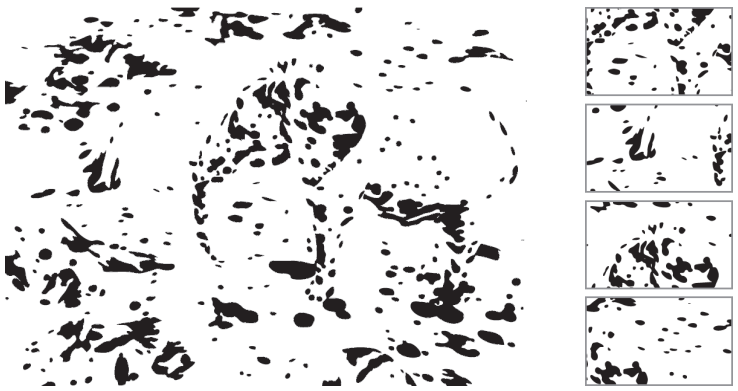
The main contribution of this work is a model-based image analysis framework which represents a planar object‘s variation in shape and appearance, it‘s partial occlusion as well as background clutter. The model parameters are optimized jointly in an efficient optimization scheme. Our extensions to the Active Basis Model lead to an improved discriminative ability and permit coherent occlusions and hierarchical deformations. The experimental results demonstrate a new state of the art performance at the task of forensic shoe print recognition.

Chapter 1

Introduction

> We are surrounded by objects.
> Our lives are spent identifying, classifying, using and judging objects.
> Objects are ... almost everything we know.
>
> R. L. Gregory - *The intelligent eye*

What do you see in Figure 1.1(a)? As soon as you recognize the elephant, you will be able to provide a detailed description of the image such as: The elephant is located in the right half of the image. It is standing sideways, while looking to the left. Some parts of its shape are not visible. The background is cluttered and locally similar to the elephants appearance.



(a) (b)

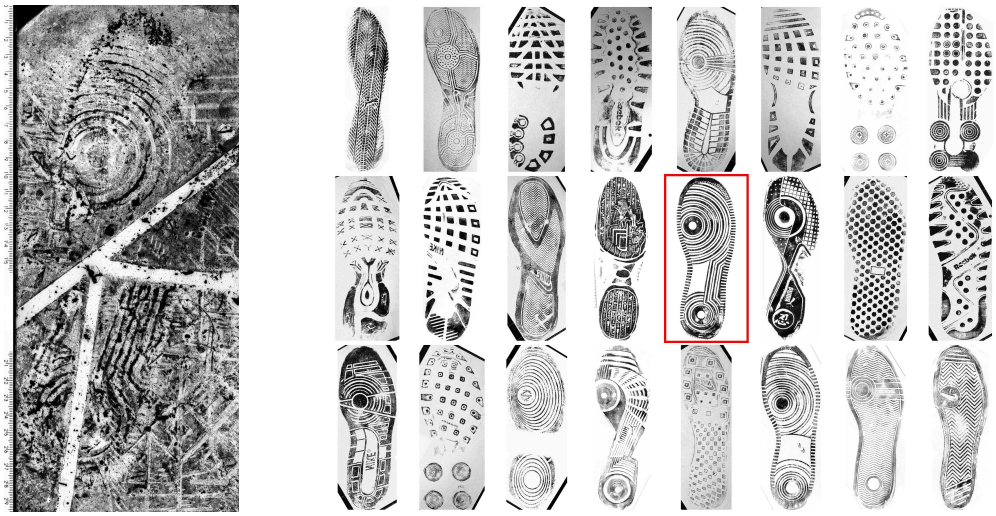
Figure 1.1: (a) Our ability to recognize the elephant in is remarkable. (b) Note that the elephant‘s shape is locally indistinguishable from the background clutter. (image source: [Mitra et al., 2009]).

This image interpretation process is inherently difficult. In order to be successful, our vision system must reason with prior knowledge about a complex interaction of

shape, pose, appearance, occlusion and background clutter. Remarkably, there is no local feature which suggests an elephant unambiguously. In fact, it's shape is locally indistinguishable from the background clutter, as we can observe from Figure 1.1(b). Despite the tremendous advances computer vision has made, it remains a fundamental challenge for computers to reliably recognize a general object in a cluttered image.

The fundamental question we address in this thesis is: How can local shape information be detected in a cluttered image, and how can this information be combined robustly into an object interpretation?

We study this question in the context of forensic shoe print recognition (Figure 1.2). In the following section, we introduce the shoe print recognition task. We then continue to embed the application into the general context of computer vision and highlight the major contributions of our work. We conclude this chapter with an overview about the structure of this thesis.



(a) (b)

Figure 1.2: Illustration of the forensic shoe print recognition task: Given a probe image (a), retrieve the corresponding gallery image from the database (b). Compared to the gallery image, the shoe print in the probe is subject to deformation, partially occlusion and appearance changes. In addition, the background is highly structured and difficult to distinguish from the actual shoe print.

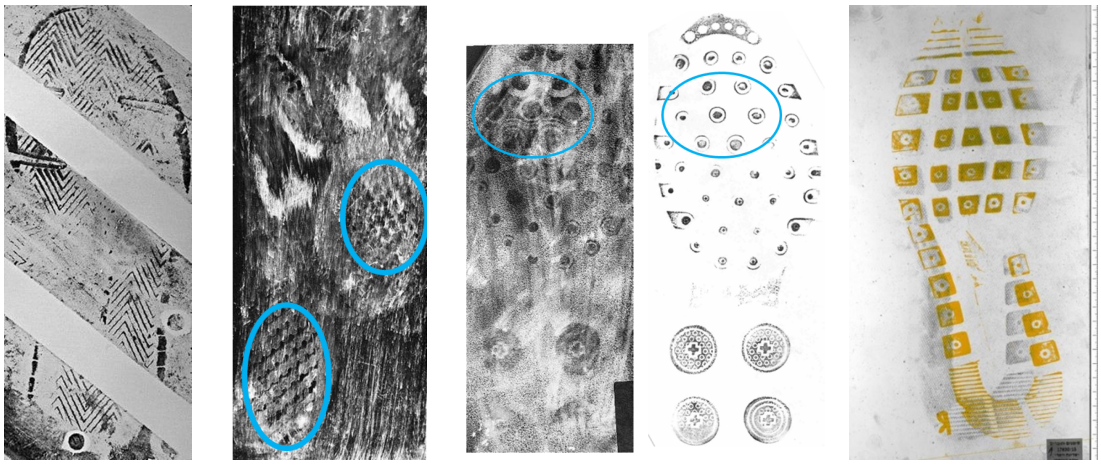
1.1 Forensic Shoe Print Recognition

Whenever we walk around, we leave traces on the ground in the form of shoe prints. A shoe print is a mark made when the tread of a shoe comes into contact with a surface. It can either be a three-dimensional surface deformation (e.g. left in snow) or a two-dimensional exchange of trace material [Bodziak, 1999]. The shoe print retains

the characteristics of a shoe's tread and therefore is a valuable piece of evidence in the crime investigation process. In forensic investigation, the shoe print is typically digitized either by photography or by lifting it from the ground with a sticky gel foil which is subsequently scanned. The task of forensic shoe print recognition is illustrated in Figure 1.2. Given an image of a crime scene print, we search for the corresponding gallery image in a database. In practice, this process is highly time-consuming as the forensic investigator has to search manually through a large database of gallery images. An automated identification based on computer vision techniques is therefore highly desirable. In the following section, we embed the application in the context of computer vision and show that it unifies a number of fundamental vision challenges.

1.2 Challenges for Automated Shoe Print Recognition

In this section, we highlight those properties of the shoe print recognition task, which render it highly challenging for computer vision systems.



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

Figure 1.3: Illustration of the fundamental vision challenges posed by the shoe print recognition task. (a) Shoe prints are often partially occluded. (b) They can be hardly distinguished from the structured background clutter. We highlight the pattern of interest in the blue circles. (c) The print in the probe can also be locally deformed. We highlight corresponding patterns in a probe image (left) and its gallery image (right) with blue circles. (d) Large non-rigid deformations of the shoe print can occur due to the forces acting on the tread while running. The probe pattern is depicted in gray. The corresponding gallery image is overlayed in orange. Note the deformation in the heel region.

Variation in Shape and Appearance. Depending on the characteristics of the tread

material and the movement of the wearer, the shoes tread deforms due to the forces acting upon contact with the surface. This deformation can happen on different scales either locally or more globally (Figure 1.3(c) & 1.3(d)). Therefore, the *shape* of a shoe print is subject to non-rigid deformation in the probe image relative to the reference image.

The *appearance* of a shoe print can also change significantly, as a result of the properties of the surface and the trace material (see Figure 1.3(c) left and right print). We can assume that a shoe print is planar. Thus, the appearance can be assumed to be independent of the lighting conditions. This is a major benefit over the common task of recognizing three-dimensional objects in natural scenes.

Clutter and partial occlusion. In natural images, objects are surrounded and partially occluded by other objects. Modeling all of these "other objects" explicitly is computationally infeasible, because of their sheer number and variability. Thus, they are often collectively treated as *clutter* - a structured background signal. However, this computational convenience is bought at a cost. Not being able to reason about the exact cause of a structure in the image renders the object recognition task even more difficult (Figure 1.4). In such a setting, human vision benefits from its reasoning capability, which makes it possible to resolve local ambiguities and to estimate what parts of the object are occluded.

(a) (b)

Figure 1.4: Can you find the bird (a) in the image on the right ? Object detection is even more difficult, when the background signal is structured, but not explainable (adapted from [Bach, 2017]).

Crime scene shoe prints are a perfect example of such cluttered scenes. The gallery image contains no background at all. However, in probe images the background is highly structured and difficult to distinguish form the actual pattern of interest (Figure 1.3(b)). In addition, it is not feasible to model all surfaces that a shoe print can occur on. Furthermore, parts that are present in the gallery images are often missing in the probe

image (Figure 1.3). This can happen because either the shoe did not touch the ground, or a lack of trace material, or an object being in between the tread and the ground. We refer to these missing parts collectively as *partial occlusion*.

Limited training data. So far, only small shoe print databases are available for research purposes. Thus, we assume that in the shoe print recognition setting the training data is limited to just one gallery image per class of shoe print. This introduces an additional type of complexity to the problem, because the amount of information about this highly complex variability between gallery and probe images is limited.

Relation to object recognition. In general, the task of *object recognition* is also not easy to define precisely, because of the ambiguity in the definition of the word "object" [Ullman et al., 1996]. Does it mean an individual entity ("my cat"), or a certain class of entities (cats in general); does the entity belong to a single or multiple categories (Siamese cat, a cat, an animal)?

In our work, we assume that a shoe is made by a single manufacturer and thus its tread is different from all other shoes. There exists no categorical hierarchy of shoes or ambiguity between treads, which makes the object recognition task very well defined. This assumption is not exactly fulfilled in practice. However, it is reasonable and enables us to focus on the general vision problem, without putting too much focus on details of the application.

Conclusive remarks. Shoe print recognition can be interpreted as object recognition of planer, non-rigid objects under partial occlusion in cluttered images. Based on this perspective, we will study the related computer vision literature in the next Chapter 2.

1.3 Contribution

The main contribution of this dissertation is the development of a holistic model-based image analysis framework which enables the computer to jointly reason about deformation, appearance change, partial occlusion and structured background clutter.

Our proposed framework builds on the Active Basis Model (ABM). It was originally proposed in [Wu et al., 2010] and later extended to be hierarchical [Dai et al., 2014]. ABMs are reference-based deformable models for describing object shapes. The basis filters provide unreliable local shape information while the deformation model allows for global geometric reasoning.
The original framework has several limitations, which we propose to overcome by extending the framework in numerous ways. These extension will result in a significant increase of shoe print recognition performance.

In the following we describe our contributions in detail.

1.3.1 Extensions to the Active Basis Model

Partial Occlusion. We extend the ABM with an independent occlusion model. During inference, parts can be deactivated if a general background model is better at explaining the image than the part model. We demonstrate that this model extension is highly effective in preventing the model from explaining image regions that do not depict the target object, thus making it possible to robustly react to partial occlusion during inference. In the context of hierarchical ABMs, the occlusion model can act at different levels of the hierarchy. We will study the effect of this choice on in detail.

Changing the Basis. A central element of ABMs is the pre-defined basis which encodes local shape information. In the earlier part of the thesis, we consider a Gabor filter bank with a fixed scale and frequency at different orientations. This setup was proposed in the original work by [Wu et al., 2010]. We will replace this filter bank with a bank of Laplacian-of-Gaussian filters of different scales. We demonstrate that this new basis improves the object model's ability to represent characteristic shape information about the target object. Ultimately, this will results in a significantly higher discriminative ability compared to the original Gabor basis.

Compositional Active Basis Models. So far, the hierarchical structure of a Compositional Active Basis Models had to be pre-defined manually. We propose a greedy EM-type algorithm that automatically infers the complete hierarchical structure from data, including the number of parts and the number of layers. Furthermore, we study the role of these additional hierarchical dependencies on the models ability to discriminate foreground from background and in the context of hierarchical deformations and coherent occlusions.

1.3.2 Contribution to Automated Forensic Shoe Print Recognition

We also contribute to the field of automated forensic shoe print recognition:

The FID-300 Dataset. Together with the German State Criminal Police Offices of Niedersachsen and Bayern and the company Forensity AG, we have collected the first footwear impression database with real forensic shoe prints and have made it publicly available. Hence, for the first time it is possible to evaluate different algorithms on a standard performance benchmark.

Shoe Print Recognition Performance. We evaluate our shape-based object recognition algorithm on the FID-300 dataset and compare it to a reimplementation of the most recent approaches. The experimental results demonstrate state-of-the-art performance.

1.4 Overview

The remainder of this dissertation is organized as follows. In Chapter 2, we will review different approaches for object recognition with respect to their suitability for object recognition in cluttered images. Throughout this review, we will relate previous works on automated shoe print recognition with a number of classical computer vision methods. Chapter 3 introduces the theoretical background of the Active Basis Model.

In the Chapters 4 and 5 we present our main contribution to the theoretical framework of ABMs and CABMs. In Chapter 4, we begin by highlighting limitations of the ABM in the context of analyzing shoe print images. We will overcome these limitations, by the introduction of an independent occlusion model and by changing the basis to a bank Laplacian-of-Gaussian filters of different scales. Chapter 5 is entirely devoted to the topic of Compositional Active Basis Models. We will introduce an algorithm that is capable of learning the complete hierarchical compositional structure of a CABM. Throughout this chapter, we will work out in detail the benefits of the hierarchical dependency structure in CABMs over the original "flat" ABM in terms of shoe print recognition. We evaluate different realizations of CABMs with respect to their performance on the application of automated shoe print recognition in Chapter 6. The thesis is concluded in Chapter 7 with a summary and a discussion on limitations of our approach and interesting future research directions.

Chapter 2

Related Work on Recognition with Reference-based Models

In this chapter, we review previous work on shoe print recognition together with related work on general object recognition. On an abstract level, we can distinguish between image-based and reference-based approaches to object recognition. Image-based methods learn a data separating function directly from the image using methods from Statistical Learning Theory. In contrast, reference-based methods represent an object class as a prototypical "reference template", together with a model of possible variations of this reference. In the context of shoe print recognition, image-based methods have not been applied so far. Due to a lack of training data, the approach that we present in this thesis is also model-based. Therefore, we focus in the following on reviewing work on object recognition with reference-based models.

In the previous chapter, we have discussed the immense variability of objects due to changes in e.g. pose, shape or appearance. A central question in reference-based object recognition is: "How can we compare the reference template to a probe image regardless of the object's variability?". This is an inherently difficult question. In general, two complimentary types of approaches can be considered to account for an objects variability: Invariant mappings and parametric models. Any work on object recognition implements a trade-off between these two types of approaches. In the following, we categorize work on reference-based object recognition into four categories, depending on the their extent of invariant and parametric object representation (Figure 2.1). Within each category, the approaches share a common mathematical view of how to compare the reference template and a probe image. Based on this taxonomy, we can relate work on shoe print recognition with well known approaches for general object recognition. In this way, we can benefit from the large knowledge about general object recognition. Furthermore, it becomes possible that new insights in the context of automated shoe print recognition find an application in other computer vision areas.

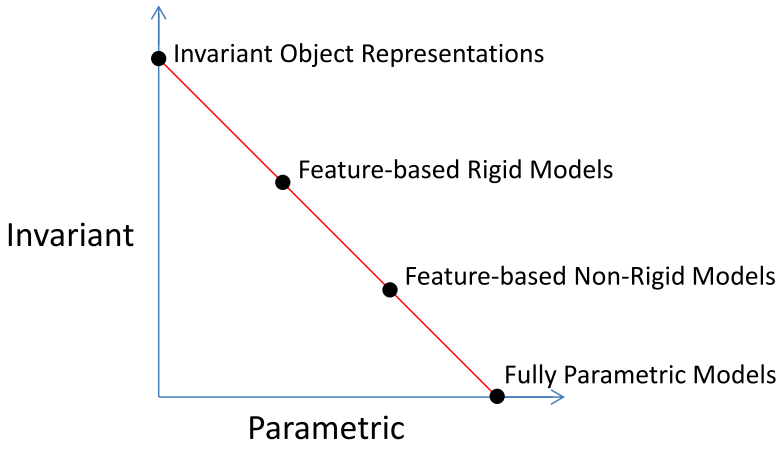


Figure 2.1: Schematic illustration of the trade-off between to extent of invariant and parametric representation in a reference-based object model. Let us assume a two-dimensional coordinate system. Each axis describes the extent to which the variability of an object is accounted for by a parametric or by an invariant representation. Any reference-based model can be roughly assigned a position in this coordinate frame. The total variability of an object class is a line. We illustrate four different categories of models as black dots on this line: Invariant object representations (Section 2.1), feature-based rigid and non-rigid models (Section 2.3 & 2.4) and fully parametric models (Section 2.5).

2.1 Invariant Object Representations

Invariant approaches leverage a priori knowledge about the space of all possible object transformations Ω, in order to design a function Φ whose output is invariant to any transformation $\Gamma \in \Omega_{inv}$. We introduce the variable Ω_{inv} at this point, since the "true" Ω can usually only be approximated. In Figure 2.1 we illustrate Ω as a red line. Invariant object representations are placed on the y-axis, since they account for the whole transformation space with an invariant mapping. A classic example in image processing is the Fourier Transform [Welch, 1967; Bracewell and Bracewell, 1986] whose power spectrum is invariant to the circular translation of the input signal. We focus here on communicating the general idea of invariant object representation. A detailed review of invariant methods can be e.g. found in [Wood, 1996].
In general, we can distinguish between integral invariants such as the Fourier-Mellin transform [Altmann and Reitbock, 1984] or the radon transform [Radon, 1917] and algebraic invariants such as Hu Moments [Hu, 1962] or Zernike Moments [Khotanzad and Hong, 1990]. All of these have been successfully applied in computer vision applications. For example [Freeman et al., 1998] use image moments for hand tracking in controlled environments. The position of the hand and the camera is fixed and the background was a uniformly colored. [Mercimek et al., 2005] tested moment invariants for the matching of image regions and report satisfying results.

The procedure of measuring the distance between two images with invariant transforms can mathematically be formulated as follows:

$$D(I, I') = \|\Phi(I) - \Phi(I')\|. \tag{2.1}$$

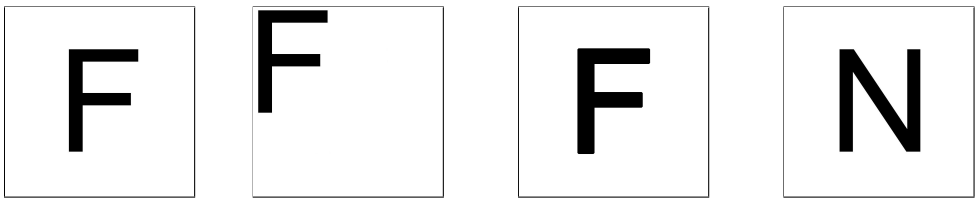
The mapping Φ is applied to both signals the reference I and the probe image I' before measuring their distance. In theory, any changes of the input signal induced by a transformation $\Gamma \in \Omega_{inv}$ should have no effect on the output of $\Phi(I)$. Thus, a major source of irrelevant image variation is factored out from the distance assessment. The recognition can then focus on deciding about an objects presence solely based on the intrinsic properties of an object. In controlled environments this assumption holds, reducing the influence of transformations $\Gamma \in \Omega_{inv}$ to a minimum.

Related work in shoe print recognition. Invariant transforms have been applied to automated shoe print recognition in the form of the Fourier Transform [De Chazal et al., 2005], the Fourier-Mellin transform [Gueham et al., 2008] or Hu-moments [AlGarni and Hamiane, 2008]. All of these works report excellent results when comparing gallery images with synthetically distorted gallery images. The synthetic distortion involves a transform $\Gamma \in \Omega_{inv}$ and the addition of locally independent noise. In our experiments (Chapter 6) we demonstrate however, that the performance of these approaches decreases significantly when tested on real data. This phenomenon is well known as instability of invariant transforms [Bruna and Mallat, 2013] and is of central importance to computer vision systems in general. Therefore, we will describe it in the following in more detail.

2.1.1 Instability of Invariant Transformations

We refer the interested reader to the work of [Bruna and Mallat, 2013] for an excellent mathematically rigorous analysis of this phenomenon. We report here just the intuition behind their line of thought, as a detailed analysis would be out of scope.
Let $\Gamma(I)$ be a small deformation of an input signal I. An invariant Φ is stable if the difference $\|\Phi(\Gamma(I)) - \Phi(I)\|$ is also small. In order to have any theoretical meaning, this statement presumes the ability to measure the amplitude of the deformation $|\Gamma|$ and to relate it with the distance between the signals. We will revisit this particular topic in section 2.4 and in Chapter 3.

In Figure 2.2 we illustrate the notion of stability visually. The reference template is depicted in Figure 2.2(a) and three probe images in the Figures 2.2(b)-2.2(d). We measure the distance between these images with the distance function as introduced in Equation 2.1. We set Φ to be the log power spectrum of the Fourier transformation and $\|.\|$ to be the euclidean distance. As expected, we observe that Φ is translation invariant (Figure 2.2(b)). However, Figure 2.2(c) shows that the distance measure is highly instable under local dilations in the reference. The influence is so big, that a completely different object is even more similar (Figure 2.2(d)). Stability is relevant in practice because the invariant space most often covers only a subset of all possible object transformations Ω. Thus, in order to be practically useful a computer vision system must be stable under those transformations that it does not account for.



(a) (b) $D(a,b) = 0.0$ (c) $D(a,c) = 724.3$ (d) $D(a,d) = 627.2$

Figure 2.2: Illustration of the instability of the Fourier transform to perturbations of the input signal. We compare the reference template in (a) with three other images (b-d) in terms of the euclidean distance between the log power spectra. The images show in (b) the reference template translated, in (c) the reference template dilated and in (d) an image of a different letter. From $D(a,b)$ we can observe that the invariant mapping is truly invariant to translations of the input. However, the invariant mapping is highly sensitive to small erosions of the pattern. Although, the pattern in (c) is much more similar to (a) than (d), this is not the case in the feature space.

2.1.2 Conclusive Remarks

In conclusion, invariant mappings are attractive because they factor out irrelevant image variations in a computationally efficient manner. However, their instability restricts their practical usability to highly controlled environments.

2.2 Introduction to Parametric Object Representations

Another major branch has developed around the approach of object recognition by reference alignment. Again, the principle is to compensate possible transformational changes of an object in order to perform the recognition solely based on intrinsic object properties. However, instead of requiring a single mapping to compensate for the whole space of possible transformations Ω, the idea is to model Ω explicitly with a parametric function $\Gamma_\alpha \in \Omega_{model}$. In Figure 2.1 fully parametric models are located on the x-axis, since any possible object variation is represented explicitly.
Computing the distance between the reference and a probe involves the search for those parameters α which minimize the functional:

$$D(I, I', \Gamma_\alpha) = \|\Gamma_\alpha[I] - I'\|. \tag{2.2}$$

Compared to invariant transforms, this is a fundamentally different way of compensating possible object variations. In the following, we will work out the advantages and difficulties of such alignment methods. We start by discussing the basic ideas and needs of alignment methods based on template matching, which is the simplest possible alignment method. We then continue by introducing hybrid methods which combine local invariants with parametric transformations. We classify these approaches into three dif-

ferent categories, depending on the complexity of the parametric transformation. The different categories are rigid transformations (Section 2.3), non-rigid shape deformations (Section 2.4) and fully parametric models (Section 2.5). In each category, we present relevant work covering literature from the automated shoe print recognition as well as other computer vision applications.

2.2.1 Template Matching

A very popular realization of the alignment approach is template matching. In its simplest form, the reference I is set to be an image of the target object and the parametric transformation Γ_α performs a 2D shift across the probe image. The evaluation metric is typically set to be the euclidean distance.

Despite its simplicity, this approach already shows the essential elements of general alignment techniques: a prototypical representation of the target object, prior knowledge about how the reference can change in probe images and a mechanism for comparing the transformed reference to the probe image. As the space of parameters α in a standard setup is small, the optimal parameter can be determined by a brute force search in the parameter space.

Compared to the Fourier matching as illustrated in Figure 2.2, template matching is more stable if the reference template is a tight bounding box around the object (distances: $D(a,b) = 0.0; D(a,c) = 53.5; D(a,d) = 102.7$). However, this comes at a cost. We need supervision for designing the reference template and the parametric function of relevant object transformations. For natural objects and complex transformations, such as shape deformations and appearance changes, this is difficult. Also as the transformation Γ_α gets more complex, a brute force search becomes computationally infeasible and different optimization techniques have to be used.

In the following section, we discuss models which combine invariant local features with rigid models of the objects geometry. In this way, complex shape deformations and appearance changes can be accounted for while keeping the function space Γ_α simple.

2.3 Feature-based Rigid Models

Feature-based rigid alignment combines the ideas of invariant and alignment approaches. We illustrate this property in Figure 2.1. The idea is to increase the stability of invariant transforms Φ by restricting their domain to a local part of the image, instead of transforming the image globally. However, in the local representation the global context is lost. In order to recover the context, the joint spatial configuration between features is also represented as a set of points in the image frame $X = \{X_i | i = 1, \ldots, N\}$. Geometric transformations are modeled with a parametric transformation $\Gamma_\alpha[X] \in \Omega_{model}$. The space of possible transformations is chosen to be the space of rigid transformation $\Omega_{model} = SO(2)$.

This strategy can be found in early works such as [Brooks, 1981; Ayache and Faugeras, 1986; Lowe, 1987; Huttenlocher and Ullman, 1990], where rigid objects are detected in images by searching for edges in a certain spatial configuration. More recent approaches use more complex invariants with larger domains. The most popular local invariant is certainly the SIFT transform [Lowe, 1999, 2004]. SIFT is a very thoughtfully designed local invariant which is highly stable under rigid and even affine transformations. Its special detection procedure allows for a scale and rotation normalized localization of its domain. Additionally, the representation with local gradient histograms induces invariance to small changes in the objects shape and appearance. The success of the SIFT transform inspired the development of other feature transformations with similar properties (see [Dalal and Triggs, 2005; Ahonen et al., 2004; Mikolajczyk and Schmid, 2005; Bay et al., 2006; Calonder et al., 2010; Alahi et al., 2012]). The important idea underlying these works is to model the space of possible object transformations as a union of invariant transformation and a parametric transformation (Figure 2.1). The invariant accounts for local transformations of the geometry and appearance which are difficult to model, whereas the geometric model accounts for the global positioning of the object. The distance between two images is computed by searching for the geometric transformation Γ_α that minimizes:

$$D(I, I', \Gamma_\alpha) = \sum_i \|\Phi(I, X_i) - \Phi(I', \Gamma_\alpha[X_i])\|. \tag{2.3}$$

The above distance measure sums over the distances between the local feature representations $\Phi(a, b)$, where a is the input image and b is the position at which the feature is extracted in the input.

Related work in shoe print recognition. Feature descriptors have been applied to automated shoe print recognition in a bag of words manner [Pavlou and Allinson, 2006; Su et al., 2007; Pavlou and Allinson, 2009] and in combination with geometric constraints [Patil and Kulkarni, 2009; Nibouche et al., 2009; Dardi et al., 2009; Cervelli et al., 2010]. These approaches offer excellent performance on synthetically generated data. However, on real data the performance of these approaches degrades significantly ([Luostarinen and Lehmussola, 2014] & experiments in Chapter 6). We believe the reason for this degradation is that the methods do not account for partial occlusion nor explicitly to non-rigid deformations of the shoe print.

A major disadvantage of locally invariant features is that it is mathematically not clear what transformations the space Ω_{inv} in fact encompasses. Efforts have been taken to visualize the preserved information by approximate inversion ([Oliva and Torralba, 2001; Weinzaepfel et al., 2011; Alahi et al., 2012; Vondrick et al., 2013]). However, these approaches only provide a rough visual reconstruction of the input which is only of limited use for diagnosing and predicting failure cases.

In the following section, we discuss non-rigid alignment methods, which aim at al-

lowing more complex geometric deformation spaces Ω_{model} in turn for reducing the complexity of Ω_{inv} (Figure 2.1).

2.4 Feature-based Non-Rigid Models

A natural further development of feature-based rigid alignment methods, is to increase the complexity of transformations $\Gamma_\alpha \in \Omega_{model}$ to cover also non-rigid geometric transformations. These, so called shape deformations relief the burden from the local invariant to model geometric transformations (Figure 2.1). Such deformable models have been proposed e.g. as deformable-template models in [Grenander, 1970, 1976], as "rubber mask" technique of [Widrow, 1973a,b] and as pictorial structures in [Fischler and Elschlager, 1973]. However, objects are highly variable as they can e.g. change their pose significantly. It has proven to be difficult to increase the complexity of Ω_{model} without also allowing for unnatural or unlikely shape deformations. The work by [Grenander, 1976] and [Kendall, 1989] proposes to account for the fact that some transformations are statistically more common than others by imposing a complexity measure $|.|$ on the space Ω_{model}. In particular, they propose to impose a statistical distribution on the parameter set α. This is typically incorporated into the distance measure by simply adding it to the part based distance:

$$D(I, I', \Gamma_\alpha) = \sum_i \|\Phi(I, X_i) - \Phi(I', \Gamma_\alpha[X_i])\| + |\alpha|. \tag{2.4}$$

In this way, a trade-off mechanism is induced between goodness of fit of the features and the complexity of the geometric transformation Γ_α needed to match these features. Variants of this approach are e.g. known as deformable templates [Amit et al., 1991; Yuille et al., 1992], constellation models [Weber et al., 2000], active shape models [Cootes et al., 1995], graphical templates [Amit and Kong, 1996] or parts models [Burl et al., 1998; Amit, 2007]. These approaches differ in important ways of how they represent the two spaces Ω_{model} and Ω_{inv}, however what they have in common is that they must account for the complexity of different transformations in Ω_{model}. Deformable models are a key development in Computer Vision and have established as a standard technique in object recognition. The most successful variant are so called discriminative part models [Felzenszwalb et al., 2010]. Given enough training data, the feature distance for each part D_i can be learned from the data instead of being hand designed. At runtime, the distance is measured according to:

$$D(I, I', \Gamma_\alpha) = \sum_i S_i(I', \Gamma_\alpha[X_i]) + |\alpha|. \tag{2.5}$$

Discriminative part learning can be a valid approach for shoe print recognition. However, we have not explored this line of thought and therefore list this method for the sake of completeness. We refer the interested reader to [Felzenszwalb et al., 2010; Girshick, 2012] for an excellent review and discussion on discriminative part models.

Related work in shoe print recognition. The only work on automated shoe print recognition proposing a flexible geometric model was presented by [Tang et al., 2011]. The authors propose to represent a shoe print as a graph of primitives. The local geometry is captured by three types of geometric primitives: circles, ellipses and lines. These primitives are detected in images using a Hough transformation. The global geometry is encoded with an attributed relational graph between these primitives. The specially designed "Footwear Print Distance" between graphs permits global as well as local geometric flexibility. Again, on synthetically generated data, the results are promising, but on real data the performance decreases significantly (Chapter 6). The key reason is that the geometric primitives cannot be detected reliably in real data and that the geometric model has no mechanism for compensating this uncertainty.

Our proposed approach builds on the Active Basis Model [Wu et al., 2010], which also falls in this category (chapters 3). A shoe print is represented as hierarchical compositional of basis filters [Kortylewski and Vetter, 2016]. A basis filter captures the local geometry and appearance of the image, whereas the hierarchical model permits variations in the mid and high level geometry. The is augmented with a statistical distribution which accounts for the variability of the shoe print. By interpreting the filters as local invariants (Chapter 3) our approach perfectly fits into the mathematical setup as presented in Equation 2.4.

2.5 Fully Parametric Models

Fully parametric object models are at the end of the continuum between invariant and parametric methods. Any characteristic object variation is supposed to be modeled by the transformation $\Gamma_\alpha \in \Omega_{model}$. Invariant feature extractors are omitted (Figure 2.1). Thus, the model must offer an inverse mapping Φ^{-1} in order to render the object based on the parameters α down to the pixel level. This is the conceptually most important difference to feature-based approaches. Typically, prior knowledge about object modeling and computer graphics is used to define the reference template I' and the rendering function Φ^{-1}. Large parts of the transformation space Ω are learned from data. At runtime, the distance is computed by optimizing the distance in the image space:

$$D(I, I', \Gamma_\alpha) = \sum_i \|\Phi_i^{-1}(I, \Gamma_\alpha) - I'\| + |\alpha|. \tag{2.6}$$

Optimizing for the parameter set α is widely known as *analysis by synthesis* approach. Their main advantage is that the parameter set α typically offers a rich description of the target object in the probe image. This description can be leveraged for a more detailed analysis beyond the pure recognition task such as answering questions about the objects pose, or occluded parts. Furthermore, this description can be combined for different objects in a higher order reasoning process in order to resolve local ambiguities. Depending on the complexity of the invariant transformations and the induced information loss, this is not possible with the object models as presented so far. Well known generative object

models are e.g. Active Appearance Models [Cootes et al., 2001] or 3D Morphable Models [Blanz and Vetter, 1999]. However, these models are difficult to design since any possible source of image variation must be explicitly modeled. Additionally, it is also difficult to optimize these models globally.

Generative models have, to the best of our knowledge, not been proposed for shoe print recognition. As the feature extraction in the ABM (Chapter 3) is performed with simple linear basis filters, the extracted features can be inverted easily. Thus additionally to the feature-based interpretation presented in the last section, we will also be able to interpret our method as a fully generative model (see e.g. [Wu et al., 2010]). In Chapter 3 we will discuss this ambiguity shortly and explain why the mathematical setting as presented in Equation 2.4 is more convenient for us.

2.6 Reference-based Object Recognition in Cluttered Scenes

In natural images, objects are surrounded and partially occluded by other objects. Modeling all of these "other objects" explicitly is computationally infeasible, because of their sheer number and variability. Thus, they are often collectively modeled as clutter - a structured background signal.
Reference-based object models focus on modeling the target object and possible variations of it. However, for the task of object recognition in natural environments the clutter must also be considered. We have discussed already in Section 2.3, that an important mechanism for reducing the influence of clutter is to choose an object-centered reference representation in contrast to image-centered representations. The focus on modeling solely the target object, implicitly splits the probe image into foreground and background. This comes with the additional task of also representing the background in order to prevent unwanted side effects during inference [Amit, 2002; Wu et al., 2010; Schönborn et al., 2015]. Typically, a very simple background model already suffices to do so.
Another effect of clutter is that it induces local minima in the distance functional because on the part-level it is difficult to distinguish from the target object. Optimizing the functional w.r.t. α, therefore is often performed locally which in turn requires a good initial starting position of the optimization process [Amit, 2002]. An automated initialization would be highly desirable, which led to the development of a significant automated approaches such as multi-scale optimization [Jain et al., 1996].
Clutter might also partially occlude the target object in the probe image. Thus, even if the correct parameters α would be given, at some parts of the model the object will actually not be present. This might distort the distance measure significantly and should be accounted for with a robust distance measure [Huber, 2011; Amit, 2002].

2.6.1 Related Work on Occlusion Models

In computer vision, a common way of to account for partial occlusion is to restrict the parameter space of the object model. However, this assumes prior knowledge about the objects position in the probe image e.g. in the form of a precise manual model initialization [Cootes and Taylor, 1992; Schönborn et al., 2015]. For part-based models, another common approach is to bound the distance measure between a part model and the image with a fixed threshold. However, thresholds on the distance are arbitrary and often lack in interpretation. Another approach to account for occlusion is to augment the object model with an explicit binary occlusion variable. The state of the variable determines if a part is visible in the image or if it is occluded. These additional model parameters are also inferred during the optimization. This approach has been successfully applied for detecting self-occlusion in part-based deformable models of human poses [Sigal and Black, 2006]. [Azizpour and Laptev, 2012] implement an independent occlusion model. They learn a general appearance model for clutter which competes with the part model. If the clutter model explains the local image appearance better than the part model, the part will set to be occluded in the model.
Due to the independence assumption between the occlusion variables, each part can be occluded independently from its neighbors. However, in real data the states of occlusion variables are locally correlated and often can only be inferred using contextual information from nearby parts. This property is known as occlusion-coherence. Different approaches have been proposed to introduce a coherence between the occlusion states. In [Ying and Castañon, 2002] the authors propose to couple the occlusion variables z_i with a Markov Random Field. However, the introduced cycles in the dependency structure. Depending on the connection to the shape model, this results in a slow inference process with only an approximate solution. [Ghiasi and Fowlkes, 2014] use a hierarchical deformable parts model where the individual parts are connected to intermediate parents nodes, which in turn are connected to the root node. Occlusion coherence is enforced via the intermediate parent nodes. Depending on the state of the intermediate nodes, groups of their children can be occluded. In this way occlusion coherence can be induced while preserving an efficient tree-like graph structure.

2.7 Conclusion

In this chapter, we discussed the fundamental trade-off between an invariant and parametric representation of an objects variability in reference-based models. We categorized the continuous spectrum between purely invariant and purely alignment methods based on how these compute the distance between the reference and the probe image. The common mathematical setup in each category made possible to relate work on automated shoe print recognition with well known methods for general object recognition. Based on a discussion about the properties of invariant and alignment methods, several important properties for object representations become apparent. An object-centered view is

desirable because it separates the target object from the background in the representation. Deformable models provide a stable mechanism for combining local part-based representations with contextual constraints. Invariant part-based representations can be computed efficiently and provide a mechanism for compressing irrelevant information. However, their domain should be restricted in order to prevent instabilities in the representation. In order to account for clutter a model should be augmented with a robust occlusion mechanism. However, despite providing answers, the discussion also leaves us with important questions which we will study in detail throughout this thesis:

- How can we learn a model from a single training datum? (Chapter 3)
- How can clutter be represented? (Chapter 3)
- How can we account for partial occlusion? (Chapter 4)
- What is a good choice for a local invariant feature? (Chapter 4)
- How can we ensure the object models stability under large occlusions and deformations? (Chapter 5)
- What are the mechanisms in a reference-based model that permit the discrimination of the object from the background? (Chapter 4 & 5)
- Given an object model, how can we perform recognition? (Chapter 6)

In Chapter 3 we will introduce the Active Basis Model which fulfills many of the mentioned desirable properties of an object model. We will extend this model in several ways in Chapter 4 & 5, such that the final object model will fulfill all of the mentioned properties.

Chapter 3

Theoretical Background: The Active Basis Model

In this chapter we revisit the Active Basis Model (ABM) as presented in the series of work by [Wu et al., 2007, 2010; Hong et al., 2013]. We will present the theoretical ABM framework as proposed in the original work. Where suitable, we refer to our discussion about the trade-off between invariant and parametric model representation from the previous chapter. This will provide additional insights in terms of the invariance properties of the ABM.
In the following Section 3.1 we will discuss the advantages of the Active Basis Model over other object models. We then continue to present the three main elements of the ABM framework. We start by presenting the basis decomposition process which induces a part-based feature representation from the image (Section 3.2). In Section 3.3 we show how prior knowledge can be leveraged to build a statistical object model from the feature representation (Section 3.3). We explain how it can be applied to the estimate infer the optimal model parameters given a probe image in Section 3.4.

3.1 Advantages over related Object Models

In the previous chapter, we have discussed the benefits of an object-centered, part-based, deformable object representation. We will now, based on a historic overview, present different approaches of implementing such a representation in a planar object model.
In their seminal work, [Kass et al., 1988] proposed an approach for detecting deformable contours in images. A-priori knowledge about the structure of the contour was given as a set of 2D coordinates. Given an initial positioning, the Active Contour Model iteratively adapts in order to align with nearby image gradients. This *object-centered* representation based on a few coordinates is highly beneficial since it focuses on the target object and masks out other structures of the image (Section 2.6). A limitation of this approach is that the deformation of the contour points is only constrained by enforcing *local* smoothness between neighboring points. Therefore, "it can represent an

arbitrary shape as long as the continuity and smoothness constraints are satisfied" [Jain et al., 1996].
In order to overcome this limitation, the research on object models has focused on constraining the deformations specific to an object's class. One way of doing so is to learn a statistical model of possible deformations from data as proposed in the Active Shape Model by [Cootes et al., 1995]. The authors assume a given training set of different instances of the same object and manual annotations of the points X on each instance. After an initial alignment process, Principal Component Analysis is applied to learn a linear deformation space which is characteristic for the target object. In this way, an object-centered *global* object model can be build. However, due to the global dependence structure the model is difficult to optimize,
[Yuille et al., 1992] proposed a way of relaxing this global dependence between variables by introducing a hierarchical model structure. They hand designed a model of an eye by defining individual *parts* such as the bounding contour and the iris outline. The parts can move independently. However, a global energy term on the part centers keeps the global structure of the model. Thus, the major difference to global shape models is that small object variations can be accounted for independently at the part level. In addition, such a hierarchical model structure makes it possible to apply efficient dynamic programming techniques in order to adapt the model to data. In contrast, global shape models often are optimized locally.
The Active Basis Model is a deformable template. Thus, it is object-centered, part-based and deformable. In addition to [Yuille et al., 1992], it also models the objects local appearance and the background. Furthermore, it comes with an intuitive learning framework which makes it possible to learn the hierarchical model structure efficiently from a limited amount of data. In the following section, we will present the details of this learning framework.

3.2 Learning a Representation via Basis Decomposition

Mathematically, an ABM is a linear additive model in the form of the well-known sparse coding principle proposed by [Olshausen and Field, 1996]:

$$I = \sum_{i=1}^{N} c_i B_{\beta_i^1} + U = CB_{\mathcal{R}} + U. \tag{3.1}$$

Without loss of generality, the image I' is reconstructed by a linear combination of a basis $B_{\beta_i^1}$ with coefficients c_i and a residual image U. The individual parameters of each basis filter are its position and orientation $\beta_i^1 = \{X_i^1, \alpha_i^1\}$. The parameters denote the absolute position and orientation of the filter in the image frame. These parameters are encoded relative to the template center $\beta^2 = \{X^2, \alpha^2\}$ such that $\beta_i^1 = \Delta\beta_i^1 + \beta^2$. The variable $\Delta\beta_i^1$ encodes the relative spatial configuration to the template center. We denote the parameters of all filters and the template center collectively as $\mathcal{R} = \{\beta^2, \beta_i^1 | i = 1, \ldots, N\}$.

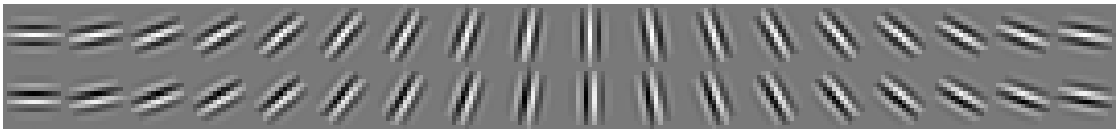


Figure 3.1: Illustration of the Gabor dictionary that we used to generate the over-complete basis. The top row shows the real part, whereas the bottom row shows the imaginary part. We apply Gabor filters always as pair of real and imaginary part.

As in the original work, we use the filter as pair of Gabor sine and cosine wavelet with fixed frequency. The filters have zero mean and unit L_2 norm. Figure 3.1 illustrates the Gabor basis.

The basis decomposition of an image I into the parameters C, $B_{\mathcal{R}}$ and U can be performed with the matching pursuit algorithm [Mallat and Zhang, 1993]. Thereby, the optimal basis is chosen from an over-complete dictionary of filters. The dictionary is obtained by translating and rotating a Gabor filter to any position in the image I in any rotation. During the matching pursuit, filters are selected one-by-one based on the amplitude of their coefficients $c_j = \langle I, B_j \rangle$. Importantly, in order to form a basis the filters $B_{\beta_i^1}$ must be independent. Therefore, after choosing a basis filter during matching pursuit, all filters which are not independent to the chosen filter are removed from the overcomplete dictionary. The process is repeated until the maximal coefficient does not exceed a previously fixed threshold. In Figure 3.2 we illustrate the result of this basis decomposition process schematically. Given a training image (Figure 3.2(a)), a basis is learned with matching pursuit. The learned basis is illustrated schematically in Figure 3.2(b). The overall shape of the basis template resembles the shape of the training image.

(a) (b)

Figure 3.2: Schematic illustration of the basis decomposition result. (a) A training image. (b) Schematic illustration of the learned basis $B_{\mathcal{R}}$. Each of the ellipsoids represents a basis filter at a certain orientation and location. The overall shape of the basis template resembles the shape of the training image (adapted from [Wu et al., 2010]).

Although very simple in its nature, the basis decomposition has far-reaching implications:

Object-centered representation. The encoding of the part parameters relative to the object center makes it possible to enforce a global constraint on their overall spatial configuration. This is an important difference compared to the local constraints

in the Active Contour Model. Due to this global dependence, whenever the position and orientation of the object changes, the parts must also change. In this way globally rigid transformations can be separated from local shape deformations in the spirit of the deformable template approach [Yuille et al., 1992] (Section 3.3).

Separation of foreground and background. The linear additive model (Equation 3.1) separates the image into two distinct entities. The linear combination of basis filters represents the object of interest (the foreground), whereas any irrelevant information is captured in the residual image U (the background). This separation permits the already mentioned object-centric representation. In addition, it also presents an explicit representation for the background. Most object models focus on modeling the foreground only. However, ignoring the background can have severe side effects [Schönborn et al., 2015], which often prevent the model from recovering the correct image interpretation. Hence, making the background explicit is a desirable feature, because it offers more control about the interdependence between the foreground and the background.

Separation of shape and appearance. An additional advantage of the linear additive model is that it separates the target objects shape from its appearance. The spatial configuration of the basis encodes the geometric properties of the object, whereas the filter coefficients encode the appearance. Having a separate representation for the shape and the appearance is desirable, as it will allows the model to reason about both properties of the object independently.

Feature extraction. Each filter coefficient is an interpretation of a small region of the input image in terms of a summary statistics, which compresses information and thus acts as a feature extractor (Section 2.3). In Section 4.3 we will study in detail what information is lost in this feature transformation.

Image generation. Due to the linearity of the filtering operation, a coefficient c_i can be used together with the corresponding filter B_{β_i} to approximately invert the feature transform. In this way images can be generated from the feature representation. Thus, by perturbing the representation parameters $\mathcal{R}$ and/or the coefficients C a bit, new images can be generated. Such invertible or generative feature representations of an object are highly desirable [Grenander, 1976; Mumford and Desolneux, 2010] since they make possible for humans to look at the model‘s internal object representation.

In summary, we have learned a parametric image model $O(\mathcal{R}, C, U) = O(\Theta)$. The parameters of the model represent the shape, appearance of a foreground object, as well as structured background clutter. By changing the parameters, we can generate new images that depict the target object. Our ultimate goal is to use the learned image model in order to recognize the object in a probe image. However, not every possible parameter setting Θ will generate a "valid" image of the object. In the next section, we will discuss how the parameters can be constrained with a statistical model. In addition, the statistical interpretation will allow us to compare results of different object models, which is problematic in a non-probabilistic setting [Amit, 2002].

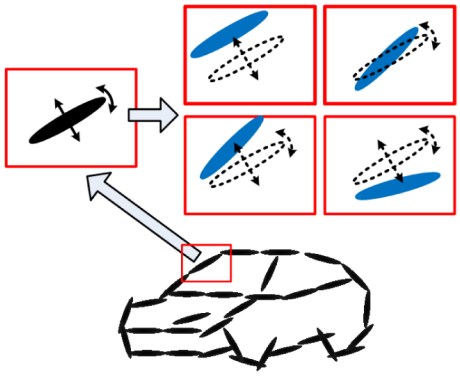


Figure 3.3: Schematic illustration of the active perturbation of basis elements in the Active Basis Model. Each of the basis filters (black ellipsoids) is allowed to perturb its parameters according to a pre-specified statistical distribution. Thus, it can take on new positions or orientations (blue ellipsoids) (image copied from [Wu et al., 2010]).

3.3 Statistical Variations of the Object Representation

In this section, we will build a statistical object model from the learned representation, by imposing a statistical distribution on the parameters $\Theta = \{\mathcal{R}, C, U\}$.

Variability in the Geometry. The shape of an object can be deformed by varying the position and orientation of the individual filters. A possible choice is to vary a part's parameters according to a uniform distribution [Wu et al., 2010]:

$$p(\beta_i^1|\beta^2) = \mathcal{U}(\beta_i^1 - \delta\beta\ , \beta_i^1 + \delta\beta). \tag{3.2}$$

The uniform distribution is defined in a range $\delta\beta = \{\delta X, \delta\alpha\}$ around a parts original position β_i^1. Thereby, δX describes the distance from its mean position and $\delta\alpha$ the difference in the angular parameter. The parameters of a part β_i^1 are conditioned on the position and orientation of the overall object β^2 thus reflecting the relative parameter encoding as introduced in the previous Section 3.2. This local variations of a filter are illustrated schematically in Figure 3.3. The statistical distribution on the templates center position and orientation $p(\beta^2)$ is modeled as a uniform prior over location an rotation. This reflects the assumption that an object can occur in an image at any position and in any orientation. By assuming independence between the perturbations of individual parts, the overall deformation model for the complete object is modeled as:

$$p(\mathcal{R}) = p(\beta^2, \beta_1^1, \ldots, \beta_N^1) = p(\beta^2) \prod_{i=1}^{N} p(\beta_i^1|\beta^2). \tag{3.3}$$

From this equation, we can observe the advantage of tree-like model structures over global dependence structures (see dicussion in Section 3.1). Given the central position and orientation of the object β^2, the parts are independent of each other. Hence, they

can change their position without affecting the other parts. This also implies that the parameters of the parts can locally be optimized independently of each other. This property allows for an efficient optimization during inference (Section 3.4.1).

Variability in the Appearance. In [Wu et al., 2010], the authors propose to use a combination of even and odd Gabor filters as feature extractors. Thus each coefficient has two components $c_i = \{c_{i,0}, c_{i,1}\}$. The final appearance features are computed by the energy of the coefficients $|c_i|^2 = c_{i,0}^2 + c_{i,1}^2$ followed by a sigmoid transform $\eta(v, \tau) = \tau[2/(1+e^{-2v/\tau})-1]$ that saturates at value τ. In order to emphasize that this corresponds to a non-linear feature transformation, we denote the final feature as $f_i = \eta(|c_i|^2)$. The authors in [Wu et al., 2010] set $\tau = 6$. In order to account for variations in the appearance feature the authors define a statistical distribution $p_i(c_i|\lambda_i)$ on the coefficients in the form of an exponential family model:

$$p(c_i|\lambda_i) = \frac{exp(\lambda_i f_i)q(c_i)}{Z(\lambda_i)}. \tag{3.4}$$

The distribution $q(c_i)$ models the expected distribution of coefficient energies in the background. It is estimated by computing the empirical distribution of feature responses from a set of images that do not depict the target object. The normalizing constant $Z(\lambda_i)$ is estimated for a range of different values of λ_i by numerical integration. The practical purpose of multiplying the exponential distribution with $q(c_i)$ is that due to the shape of $q(c_i)$ (Figure 3.4(a)) the likelihood of low energy values increases compared to a pure exponential distribution (Figure 3.4(c)). This property is beneficial as parts that are occluded by clutter will likely observe low energies. Thus, the negative influence of occlusion on the model is lowered. In the next paragraph we will discuss the estimation of $q(c_i)$ in more detail in the context of modeling background clutter. The variable λ_i controls the skewness of the expected distribution (Figure 3.4) of the features and can be learned from data if multiple images of the same object are available.

In Active Basis Models, the variation in appearance is accounted for in three ways:
1.) The filter acts as a local invariant transformation and thus already accounts for appearance variation on the pixel level.
2.) In the feature transformation, the Gabor coefficients are squared and again non-linearly distorted with a sigmoid transformation. The sigmoid compresses the feature space in the sense that low feature values are preserved whereas large feature values are dampened significantly (Figure 3.5).
3.) The statistical distribution $p(c_i|\lambda_i)$ accounts for appearance variation on the abstract feature level.

In summary, the design of the appearance model focuses on capturing "edge-like" properties of an object. However, due to multiple application of linear and non-linear transformations in the feature representation, a clear intuition of which patterns are similar is already lost.

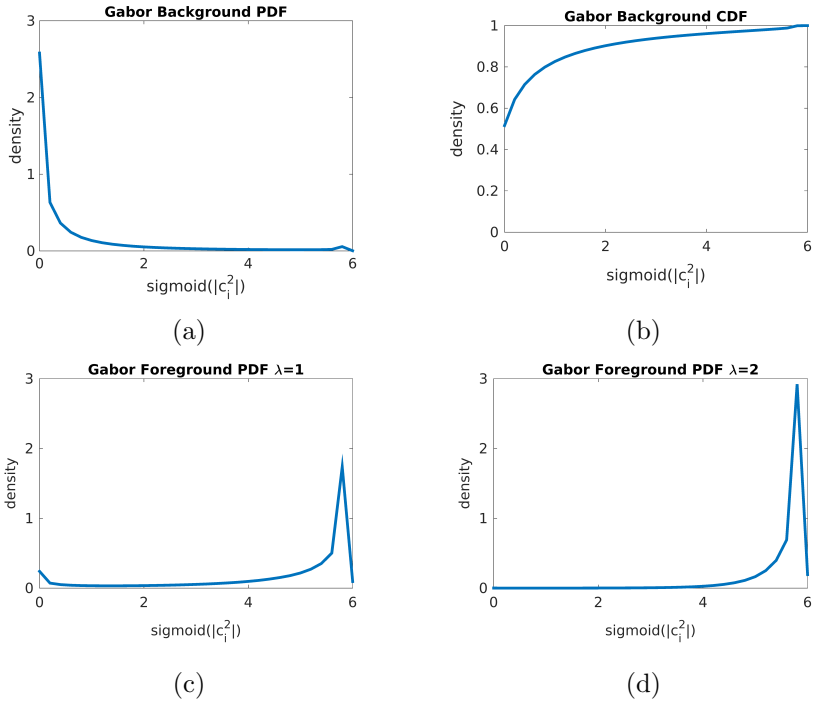


Figure 3.4: Illustration of different energy densities. (a) The density of the background distribution $q(c)$. (b) The CDF of $q(c)$. (c) The foreground density with $\lambda = 1$. (c) The foreground density with $\lambda = 2$. Due to the multiplication with the exponential function (equation) in the foreground model, low energies have higher likelihoods as in a purely exponential model. Increasing the value of λ, increases the influence of the exponential function.

Variability in the Background. In order to account for variation in the background, we must choose a representation for the background. A common assumption is to represent the background as pixels which are distributed according to Gaussian white noise [Hong et al., 2013]. However, this is an unnatural assumption as the background in images is much more likely to be structured. We can enforce more structure in the background by representing it also as composition of Gabor filters:

$$U = \sum_{k=1}^{M} c_k B_k \ . \tag{3.5}$$

Here, the variable k indexes those positions in the image, that are not covered with the foreground model. We model the variation in the background as product of independent background likelihoods:

$$p(U|C) = p(U|c_1, \ldots, c_N) = \prod_{k=1}^{M} q(c_k) \ . \tag{3.6}$$

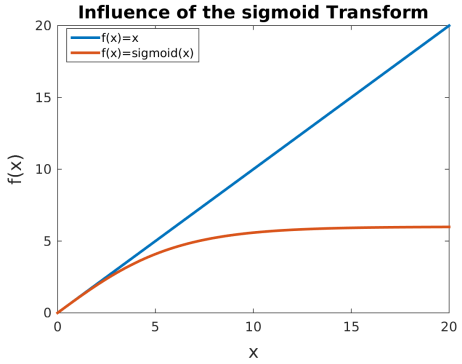


Figure 3.5: Influence of the sigmoid function on the feature values. Large values are significantly dampened in the output space.

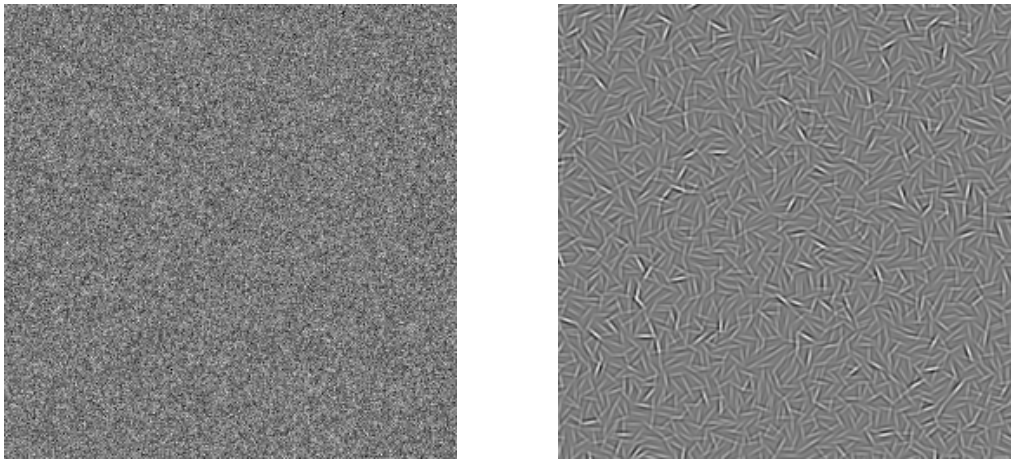
As mentioned earlier, the distribution $q(c)$ can be estimated from a set of images which do not depict the target object. Representing the background with a dictionary of Gabor filters has the advantage, that the local correlation of pixels is taken into account. In Figure 3.6 we compare samples from a Gaussian white noise model and the structured background model as defined in Equation 3.6. We can observe that the structured background model is capable of generating strong edges. As we will see in the next Section 3.4.1, this is beneficial because the background model will compete with the foreground model when explaining the image. Thus, it will prevent the foreground model from being too much attracted by strong edges in the background, while at the same time penalizing the explanation of areas without edges. We can combine the different models for variations in the geometry (Equation 3.3), the appearance (Equation 3.4) and the background (Equation 3.6) into a statistical image model:

$$\begin{aligned} p(\Theta|O) &= p(R, C, U|O) \\ &= p(\beta^2) \prod_{i=1}^{N} p(\beta_i^1|\beta^2) p(c_i|\lambda_i) \prod_{k=1}^{M} q(c_k) \end{aligned} \tag{3.7}$$

The model is generative, in the sense that we can sample from this prior distribution and generate images by inverting the Gabor features. In order to generate images at the pixel level we divide the energy equally on the even and odd Gabor filters. Thus we get $c_{i,0/1} = \sqrt{\frac{\eta^{-1}(e)}{2}}$. Importantly, the contribution of the individual filters cannot be recovered, thus this information is lost in the feature transformation process.

3.4 Parameter Estimation

At runtime, we are given a probe image I' and must optimize for the parameters Θ^* of our image model. We do so by maximizing the ratio between the foreground and



(a) (b)

Figure 3.6: Samples from different background models. (a) Gaussian white noise. (b) The background model as proposed in the ABM framework. An independent composition of Gabor filters. The coefficients are sampled from $q(c)$. Compared to (a), the sample in (b) is much more structured.

background model:

$$\frac{p(\Theta|I', O)}{q(I')} = p(\beta^2) \prod_{i=1}^{N} p(\beta_i^1|\beta^2) \frac{p(c_i|\lambda_i)}{q(c_i)}. \tag{3.8}$$

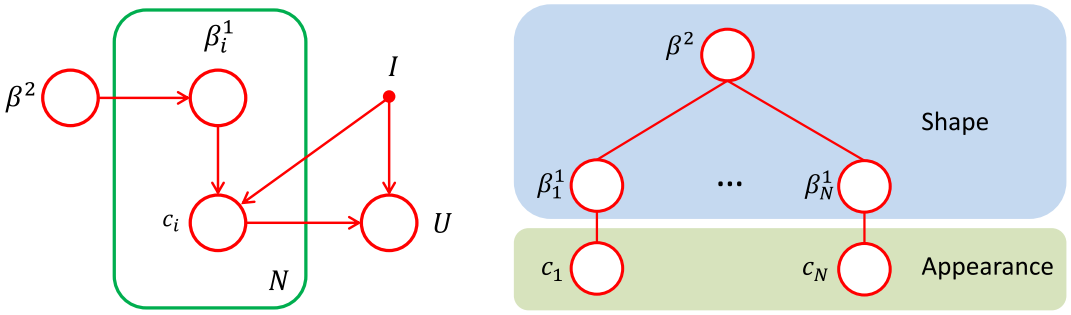
In this equation the background has canceled out since it is the same for both models. We can observe, that the appearance likelihoods of the foreground and the background act as competing hypothesis in a likelihood ratio. This term can be interpreted as a classification mechanism. If the ratio is > 1 the local appearance of the image is more likely to be part of the target object, accordingly if it is < 1 it is more likely to be background clutter. Substituting $p(c_i|\lambda_i)$ as defined in Equation 3.4 and taking the logarithm, we arrive at the log likelihood ratio:

$$log(\frac{p(\Theta|I', O)}{q(I')}) = log(\beta^2) + \sum_{i=1}^{N} \lambda_i f_i - log(Z(\lambda_i)) + log(p(\beta_i^1|\beta^2) \tag{3.9}$$

$$= \sum_{i=1}^{N} \lambda_i f_i - log(Z(\lambda_i)) + constant \tag{3.10}$$

$$= \lambda \sum_{i=1}^{N} f_i + constant. \tag{3.11}$$

From Equation 3.10, we can nicely observe the influence of the background model. It acts implicitly via the normalization factor $Z(\lambda_i)$ on the feature values. Equation 3.10 is used in the original works on ABMs. In the last line we use the assumption that λ_i is the



(a) (b)

Figure 3.7: Dependence structure between random variables of an Active Basis Model illustrated graphically. (a) The full graphical model; (b) The common way of illustrating tree-structured models, by focusing on the object representation. The tree-structured Markov random field allows for an efficient bottom-up optimization.

same for all parts. At this point the computation reduces to a sum over feature values f_i, which can be efficiently computed using a convolution operation. We will leverage this property during the optimization.

3.4.1 Bottom-up Inference

Because of the strong independence assumptions between the random variables, the model is tree structured (Figure 3.7(b)). Tree structured models can be optimized with dynamic programming. In [Wu et al., 2010] the authors propose a bottom-up optimization scheme which we will sketch in the following in a slightly adapted form.

Intuitively, during inference each filter of the ABM will be evaluated at one position in the image. Depending on its rotation it will observe a certain feature value f_i. The goal is to find a positioning of the ABM such that the sum of feature values over all filters is maximal. Computing this global optimum for most object models would require a global brute force testing of any parameter setting. However, for tree structured models this maximum can be computed with a cascade of convolutions.
First, in order to obtain the filter responses, the probe image is convolved with the dictionary of Gabor filters. Subsequently, for each coefficient c its feature value $f(c)$ is computed. This computation generates one feature map for each orientation of the filter. In order to account for the active perturbation of a filter a maximum-convolution is applied to the appearance score maps. In the original work the maximum is taken along the normal direction of the filter. We perform a window-based convolution because it can be computed more efficiently using the method proposed in [Lemire, 2006]. The computational speedup is more than a factor of two. Let us assume the active filter perturbation is $\delta\beta = \{\Delta X = 2\text{pixel}, \Delta\alpha = 0°\}$, then the max-kernel is 2D with a size of 5×5 pixels.

The final log-likelihood ratio of the deformed template can then be computed with a 3D convolution, where the 3D kernel encodes the global spatial configuration of the filters at different positions and orientations relative to the center of the template.

3.5 Conclusion

In this chapter, we have presented the Active Basis Model framework as presented in [Wu et al., 2010]. The ABM, has several desired properties of object representations that we discussed in the previous Chapter 2. It is an object-centered deformable part model which is capable of accounting for changes in the objects appearance. In addition, the model structure can be learned from data and optimized globally with an efficient bottom-up inference procedure. In the next chapter, we will study the properties of the ABM in the context of forensic shoeprint recognition.

Chapter 4

Shoe Print Analysis with the Active Basis Model

> To reason about an entity, we must first represent the entity, or at least the relevant aspects of it.
>
> V. S. Nalwa

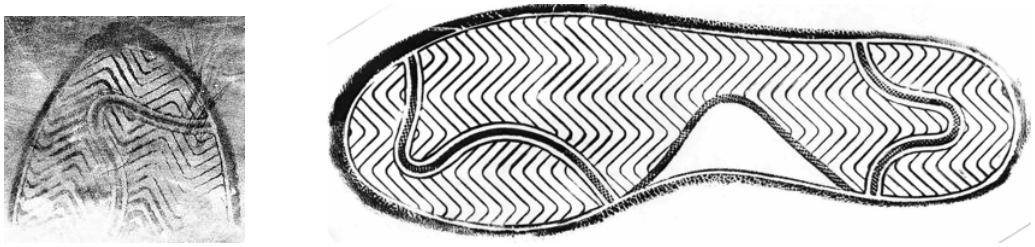
In this chapter, we study the Active Basis Model in the context of shoe print analysis. Thereby, we will build on and extend the theoretical framework as introduced in the last Chapter 3. Our study will provide additional insights about the ABM framework in terms of the interdependence between the foreground and background appearance models and the role of the basis filter. In this context, we will observe two limitations of the model which prevent it from robustly analyzing shoe prints. We propose to address these limitations by extending the model in terms of:

1. Occlusion-awareness: We extend the model with an occlusion mechanism which will improve its robustness to missing parts during inference
2. A basis change: We will exchange the Gabor filters with a dictionary Laplacian-of-Gaussian filters of different scales

In Section 4.1 we present how the ABM can be applied to compute the similarity between a probe image and a gallery image. We will then introduce our model extension in terms of an independent occlusion model (Section 4.2) and a Laplacian-of-Gaussian basis (Section 4.3).

4.1 Comparing Shoe prints with the Active Basis Model

Our ultimate goal is to recognize a shoe print in a probe image. This presumes the ability to compute the similarity between a gallery image and the probe image. Figure 4.1 shows a probe image and the corresponding gallery image. Compared to the gallery image, the print in the probe image is rotated, partially occluded, slightly deformed and



(a) (b)

Figure 4.1: A probe image (a) and its corresponding gallery image (b). The shoe print in the probe differs from the gallery image in terms of a global rotation, local deformation, partial occlusion and appearance variation.

has much lower contrast in its appearance. We propose to apply the ABM framework in order to compute the similarity between these two images. Figure 4.2 schematically illustrates our approach.
At training time, we learn an ABM from the gallery image I as presented in the previous Chapter 3. First, I will be projected onto a Gabor basis (Figure 4.2 - Basis Decomposition). Each of the orange ellipsoids represents one combination of even and odd Gabor filter (see Figure 3.1) in a certain position and orientation. This spatial configuration of filters is a part-based representation of the shoe print's shape. We can clearly observe, how the geometry of the curved structure in the toe region is preserved in the basis decomposition. An ABM $p(\Theta)$ is build by imposing a statistical model on the parameters of this decomposition (Figure 4.2 - blue box).

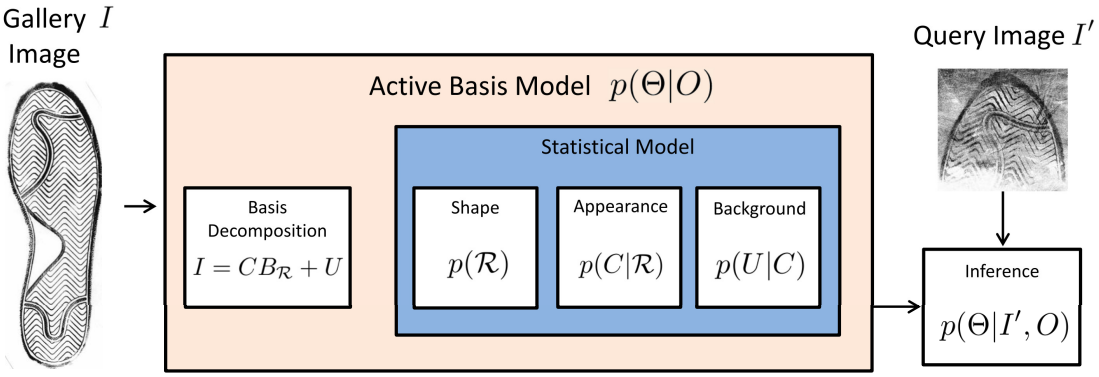


Figure 4.2: Computing the similarity between a gallery image and a probe image. The gallery image I is decomposed into a linear combination of basis filters($B_{\mathcal{R}}$), corresponding coefficients C and a residual image U (Section 3.2). An ABM is built by imposing a statistical distribution on the variables $\Theta = \{\mathcal{R}, C, U\}$ of the learned parametric image model $O(\Theta)$ (Section 3.3). Given a probe image I', the ABM is optimized with a bottom-up inference procedure (Section 3.4.1). The maximal posterior probability $p(\Theta^*|I', O)$ will be used to compute the similarity measure $S(I, I')$ between the two images.

4.1.1 Shoe print specific Appearance Model

The background appearance model $q(c)$ in the original ABM (Figure 3.4(a)) has been estimated from a set of general pictures [Wu et al., 2010]. We estimate a shoe print specific background appearance model from textures which are commonly observed in shoe print images (Figure 4.3).

Figure 4.3: Training images that we use to estimate the shoe print specific background appearance model. They depict textures that are commonly observed in probe images.

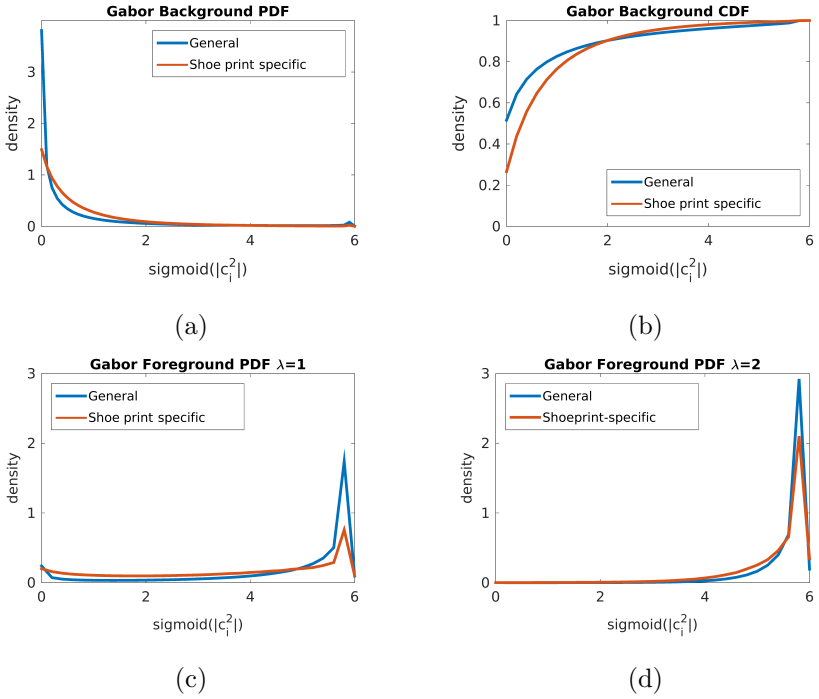


Figure 4.4: Comparison of the shoe print specific appearance models with those of the original ABM. (a) The background densities $q(c)$. (b) The CDFs . (c) & (d) show the foreground densities for $\lambda = 1$ and $\lambda = 2$.

In Figure 4.4 we compare the original original appearance models with the shoe print specific appearance models. Compared to the original model, the probability mass shoe

print specific background model is less concentrated on the extreme feature values (Figure 4.4(a)). As the computation of the foreground involves the background (Equation 3.4), its probability mass is also less concentrated. The main effect of the shoe print specific appearance, is that the differences between foreground and background models are smaller for extreme feature values. This effect is beneficial, as the influence of outliers is reduced.

4.1.2 Qualitative Experiment

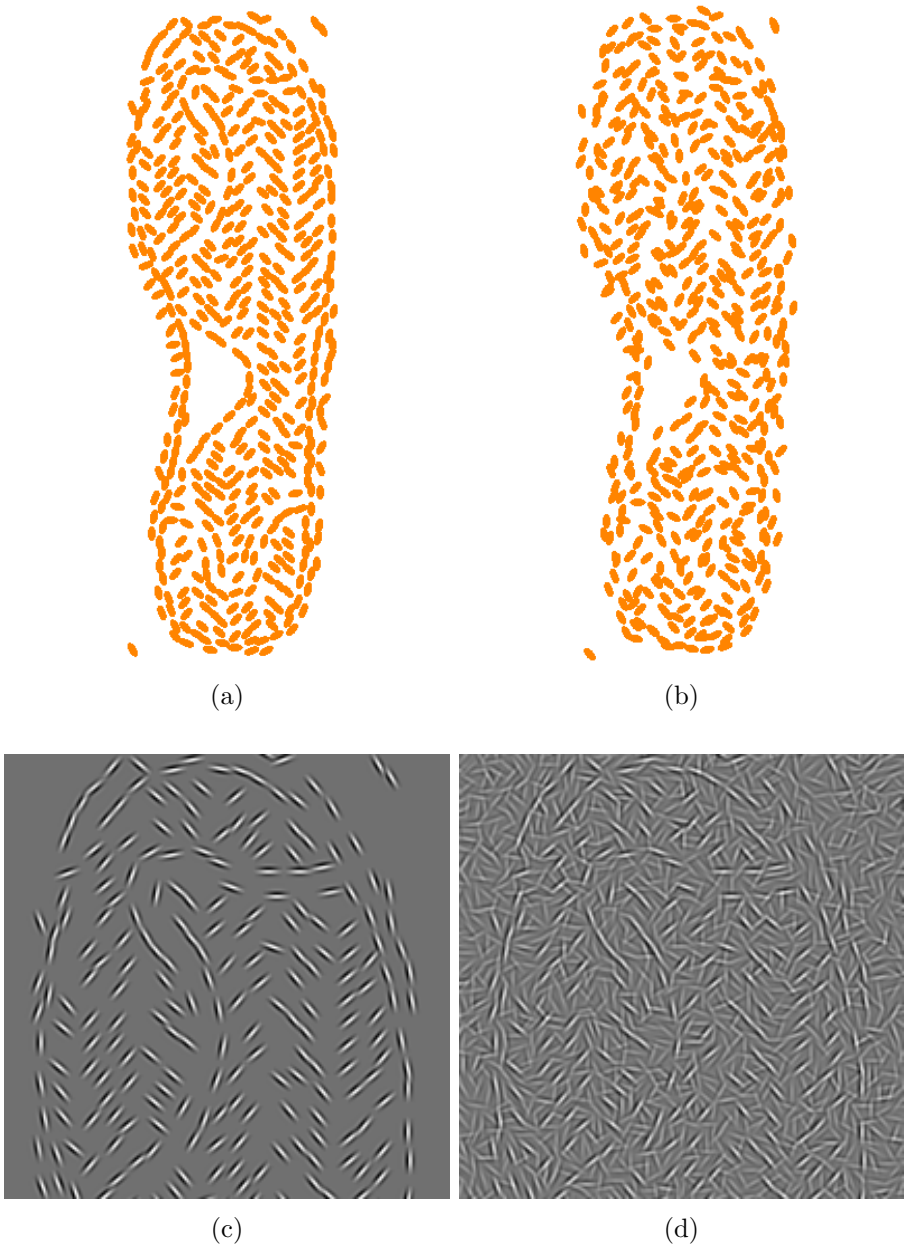
In Figure 4.5 we illustrate samples from the ABM. A sample from the shape model $p(\mathcal{R})$ with $\delta\beta = \{\delta X = 5 \text{ pixel }, \delta\alpha = 10°\}$ is depicted in Figure 4.5(b). We can observe that the shape has been perturbed locally, whereas the overall structure is preserved. A sample from the foreground appearance model $p(C|\mathcal{R})$ is shown in 4.5(c). The appearance of the Gabor filters can be clearly recognized in the image. We have already shown a sample from the background appearance model $p(U|C)$ in the last chapter and therefore do not include this here. From this collection of sub-figures, we can observe the models ability of representing the gallery image and possible variations of it. This is a nice feature of this generative representation over purely discriminative representations, as we can interpret the internal state of the model visually, which is much more intuitive than trying to interpret the numerical values of parameters.

We show a sample from the full ABM $p(\Theta|O)$ in Figure 4.5(d). In the following, we assume that the probe image (Figure 4.1(a)) was generated by this ABM. We will recover the parameters of the ABM given the probe image with a bottom-up inference procedure (Figure 4.2 - Inference). In Figure 4.6 we illustrate the inferred parameters. The optimal rigid alignment of the model is illustrated by overlaying the gallery image on the probe image with a rigid transformation that is parametrized by β^2 (Figure 4.6(a)). Despite the significant contrast change, partial occlusion and small deformations in the probe image, the model has been aligned successfully. Figure 4.6(b) illustrates the optimal positions of the individual filters. Each filter is color-coded according to the sign of the log-likelihood ratio between the foreground and background appearance model of its observed coefficient $\log(\frac{p(c_i|\lambda)}{q(c_i)})$. This is useful for judging the ability of the appearance models to discriminate between foreground (red) and background (blue). Filters colored in red have positive log-likelihood ratios, whereas the ratio is negative for blue filters. Those filters which lie outside of the image are colored in yellow. They are assumed to have a zero log-likelihood ratio. We can observed, that those filters which are positioned on an edge in the right orientation are classified as foreground.

We propose to compute the similarity of the gallery and a probe image with the log-likelihood ratio of the maximal posterior value and the likelihood of encoding the whole image with the background model:

$$S(I, I') = log\Big(\frac{p(\Theta^*|I', O)}{q(I')}\Big). \tag{4.1}$$

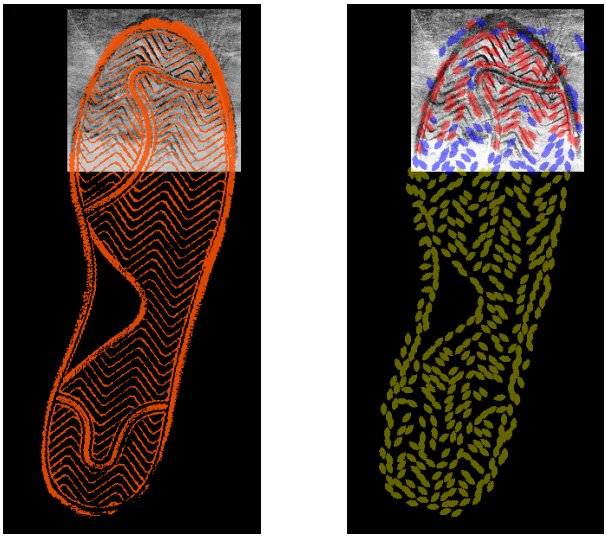
Based on this similarity measure, we will perform shoe print recognition in our experiments in Chapter 6.



(a) (b)

(c) (d)

Figure 4.5: Illustration of the ABM which was learned from the gallery image (Figure 4.1(b)). (a) A schematic illustration of the learned basis decomposition. Each ellipsoid represents a combination of even and odd Gabor filter in a certain position and orientation. (b) A sample from the shape model $p(\mathcal{R})$ with $\delta\beta = \{\delta X = 5 \text{ pixel }, \delta\alpha = 10^\circ\}$. (c) A sample from the appearance model $p(C|\mathcal{R})$ with fixed geometry. (d) A sample from the full ABM $p(\Theta|O)$.



(a) (b)

Figure 4.6: Visualization of the ABM inference result. (a) The optimal alignment of the ABM on the probe image. Each filter of the ABM is color-coded depending on its observed filter coefficient c_i. The filter is colored in red if the foreground likelihood is greater than the background likelihood $p(c_i|\lambda_i) > q(c_i)$, in blue if it is the other way round $q(c_i) > p(c_i|\lambda_i)$ and in yellow if it lies outside of the image. (b) Illustration of the global rigid alignment of the ABM. The gallery image (Figure 4.1(b)) is overlayed on the probe image (Figure 4.1(a)) with the inferred rigid model parameters β^2.

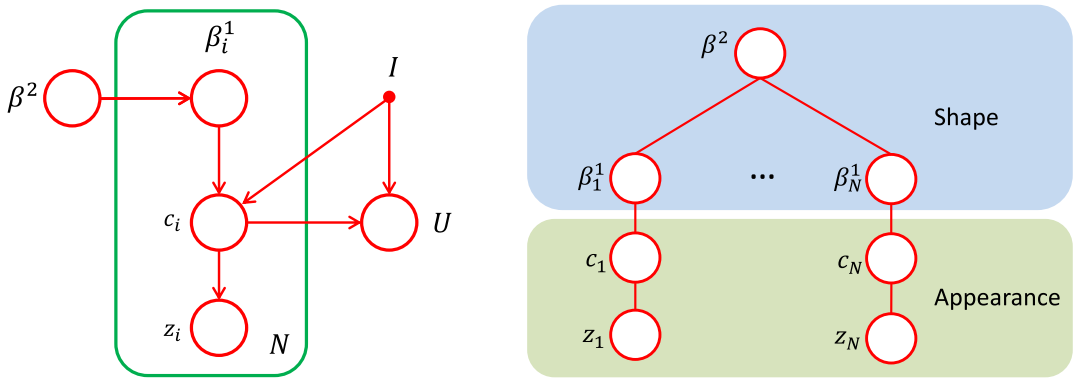
4.2 Handling Partial Occlusion

Occlusion means that an object is only partially visible in an image. This property is a challenge for model-based approaches to image analysis. If an object is occluded by another object, the model will explain parts of the image which actually do not belong to the target object. Very often, the correct image interpretation can therefore not be recovered (see example in Figure 4.9). In this section, we propose an ABM formulation with an independent binary occlusion model in the spirit of the work by [Azizpour and Laptev, 2012]. In the context of ABMs this is a novel contribution, which increases the models robustness to partial occlusion.

4.2.1 Occlusion-aware Active Basis Model

We handle occlusion on a per filter level. For this we introduce a binary random variable z_i that indicates if a part is visible in the image. The random variable follows a Bernoulli distribution:

$$p(z_i) = Ber(\rho). \tag{4.2}$$



(a) (b)

Figure 4.7: The dependence structure between random variables in the occlusion-aware ABM. (a) An illustration of the full graphical model. Compared to the original ABM (Figure 3.7(a)), a dependence of the appearance variable c_i on the occlusion variable z_i has been introduced. (b) The tree structure of the object model highlights that each appearance variable depends on a separate occlusion variable z_i.

The occlusion variable extends the part appearance likelihood as follows:

$$p(c_i|z_i, \lambda_i) = p(c_i|\lambda_i)^{z_i} q(c_i)^{1-z_i} p(z_i) \tag{4.3}$$

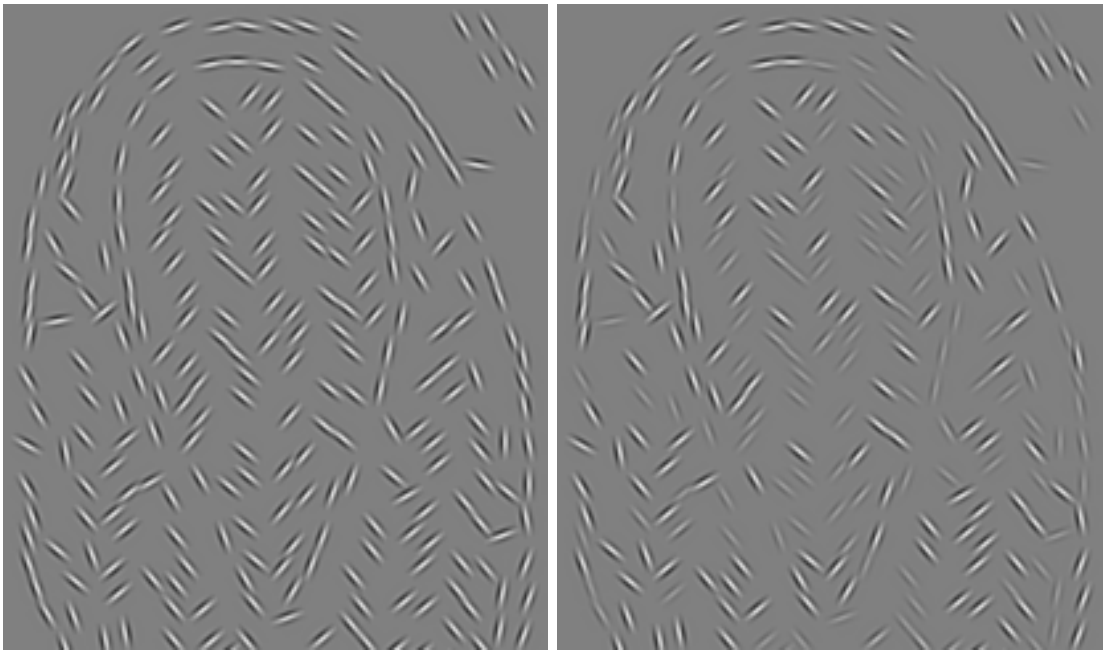
If $z_i = 1$, the original ABM part appearance model is active, whereas if $z_i = 0$ the background appearance is activate. We assume that each part can be occluded independently of the others as in [Ying and Castañon, 2002; Azizpour and Laptev, 2012]. Therefore, the maximum likelihood solution to the state of the occlusion variable z_i is:

$$z_i = \begin{cases} 1, & p(c_i|\lambda_i, \beta_i^1)p(z_i = 1) > q(c_i|\beta_i^1)p(z_i = 0) \\ 0, & else \end{cases} \tag{4.4}$$

The effect of this model extension is that the log likelihood ratio between foreground and background appearance will be bound. This implies that the negative effect of missing parts on the cost function is reduced. Equation 4.4 shows that the state of the occlusion variable in the end reduces to a thresholding operation. In contrast to other approaches, this threshold has a clear statistical interpretation, revealing the influence of the background model on the decision process via $Z(\lambda) = \int exp(\lambda f)q(f)df$. The implications on the models dependence structure is illustrated in Figure 4.7.

4.2.2 Qualitative Experiment

In this subsection, we compare the original ABM with the occlusion-aware ABM qualitatively. For our experiment we assume that every part of the model is equally likely to be visible or occluded. Hence, we set the parameter of the Bernoulli-distribution to $\rho = 0.5$. In a typical object recognition application the likelihood of being occluded



(a) (b)

Figure 4.8: Comparison between samples from the foreground appearance models of the original ABM and the proposed occlusion-aware ABM. The geometry of the model is fixed. (a) A sample from the original ABM. (b) A sample from the occlusion-aware ABM with $p(z_i) = Ber(0.5)$. Due to the occlusion, roughly half of the appearance variables are sampled from the background model. Therefore, they are barely visible in (b).

might be lower. However, our setting is justified by the fact that in shoe print recognition the probe images often strongly occluded.
In Figure 4.8 we compare samples from the foreground appearance model of the original ABM (Figure 4.8(a)) and the occlusion-aware ABM (Figure 4.8(b)). Due to the occlusion model some appearance variables are sampled from the background model and therefore are only latently visible in Figure 4.8(b).

The relevance of our extension is apparent in Figure 4.9 on a real world example. Figure 4.9(a) depicts a shoe print impression for which the toe and heel regions are visible but the central part is missing. Figure 4.9(b) depicts the corresponding gallery image and its schematic basis decomposition (Figure 4.9(c)). In order to be correctly registered to the image, the ABM must bridge the clutter between toe and heel region. Without a proper occlusion model this is too expensive, thus inducing a wrong optimum (Figure 4.9(d)). The proposed occlusion-aware model is aligned correctly.

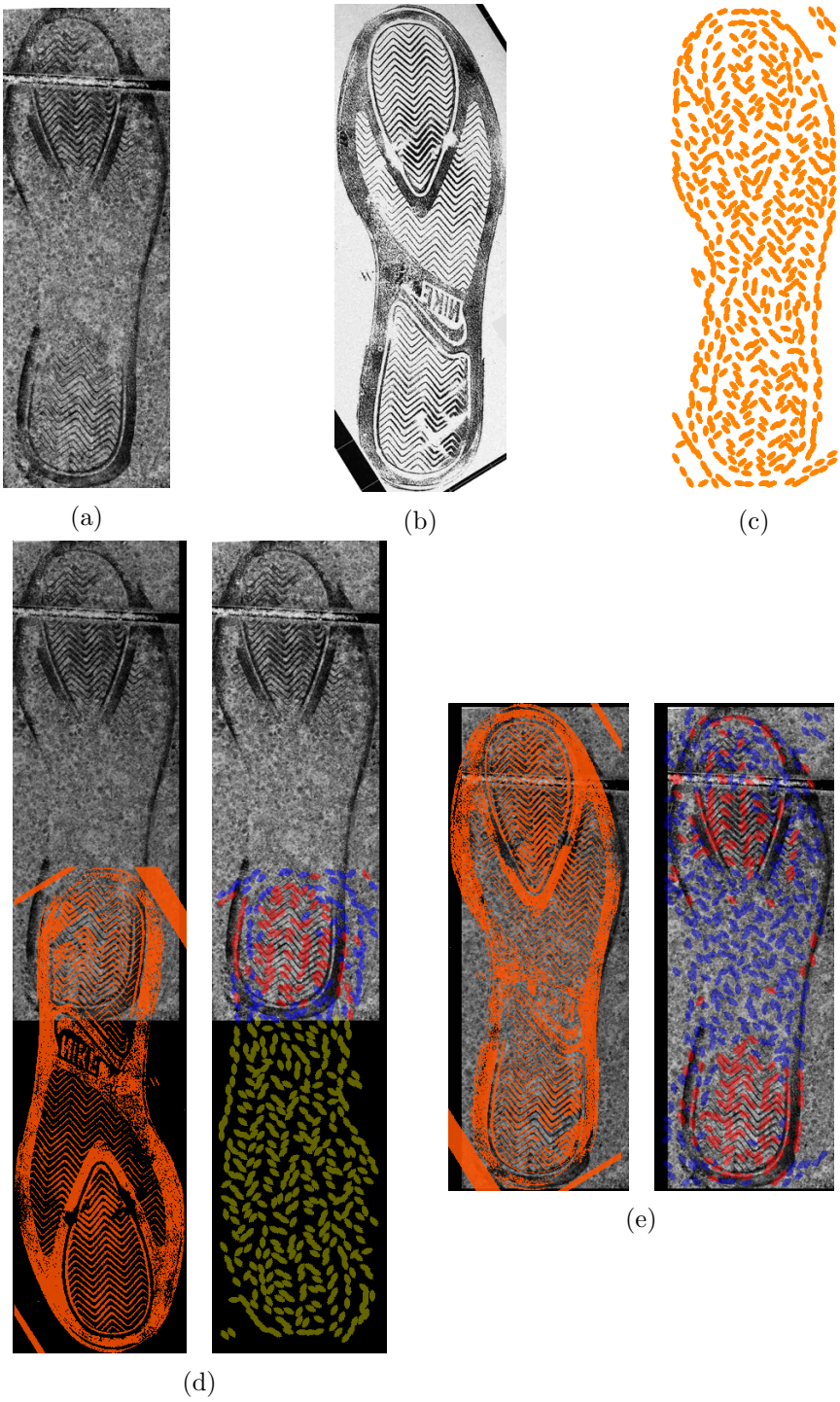


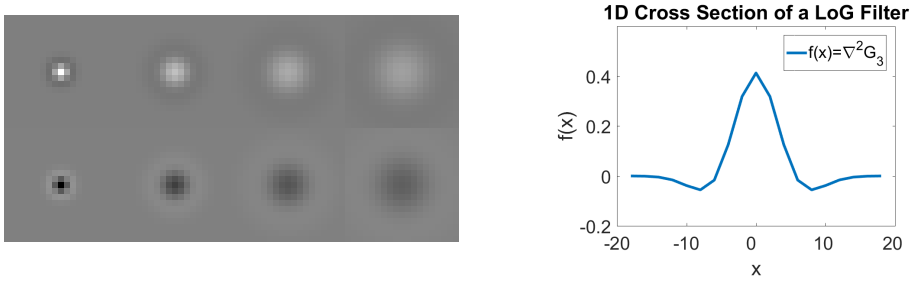
Figure 4.9: The effect of the proposed occlusion model on the inference result on real data. (a) The probe image. (b) The corresponding gallery image ofy the probe image. (c) Schematic illustration of the ABM learned from the gallery image. (d) Inference result with the original ABM. (e) Result with the proposed occlusion-aware ABM.

4.3 Changing the Basis in the Active Basis Model

After studying the role of the basis filter as a local feature extractor, we will focus in this section on studying what information is preserved in this feature space. We can visualize the preserved information by projecting the Gabor features back into the image space. From Figure 4.12(b) we can observe that the Gabor basis generally represents the location and orientation of sharp edges. Other useful information is not well captured, such as the direction of intensity change, point-like structures or the scale of an edge (Figure 4.12(b) & 4.12(e)). This additional information would be beneficial for the recognition process. One possible approach of encoding more information is to increase the dictionaries complexity by sampling the parameter space of Gabor filters extensively at different frequencies and scales. This allows to tune the filter properties, to better represent the pattern in the training image. However, this would increase computational cost at runtime significantly, as the inference process includes a 3D convolution on the feature map of each dictionary element (see Section 3.4.1). Furthermore, the direction of the intensity change could still not be recovered as the feature transformation involves a squaring of the filter coefficients.

4.3.1 The Laplacian-of-Gaussian Filter

We propose to capture more information by replacing the Gabor dictionary with a dictionary of Laplacian-of-Gaussian (LoG) filters at different scales (Figure 4.10(a)). LoG filters have been proposed by Marr and Hildreth [Marr and Hildreth, 1980] for the



(a) (b)

Figure 4.10: Illustration of the Laplacian-of-Gaussian (LoG) filter. (a) A dictionary of LoG filter of different scales $\sigma^2 = \{1, 2, 3, 4\}$ and with different signs. Top row: $\nabla^2 G_\sigma$; bottow row $-\nabla^2 G_\sigma$. (b) Cross section of the LoG filter from the dictionary with scale parameter $\sigma^2 = 3$.

detection of edges at multiple scales. In addition, we will leverage these filters properties for the detection and encoding of edge information. The LoG function is computed by taking the second-order derivative ∇^2 of the two dimensional Gaussian distribution. It

is parametrized in terms of radial distance r from the origin:

$$\nabla^2 G_\sigma(r) = \frac{-1}{\pi\sigma^4}\Big(1 - \frac{r^2}{2\sigma^2}\Big)\exp\Big(\frac{r^2}{2\sigma^2}\Big). \tag{4.5}$$

The function $\nabla^2 G_\sigma(r)$ is a circularly symmetric Mexican-hat-shaped operator which is depicted in Figure 4.10(b). Figure 4.10(a) illustrates the LoG-dictionary we use in our work. The top row of filters is generated by evaluating $\nabla^2 G_\sigma(r)$ for different variances $\sigma^2 = \{1, 2, 3, 4\}$ at equally spaced coordinates. This choice of variance parameters has been empirically validated to reconstruct shoe print patterns well. The bottom row are negative LoG filters $-\nabla^2 G_\sigma(r)$. A further difference of the proposed LoG features compared to the Gabor filters is that we use the filter coefficients directly as features. This preserves the sign of the intensity change. We restrict the filter coefficients to the range $c_i \in [0, \text{inf}]$, in order to still fit into the statistical setup of the ABM framework, which presumes positive feature values. This is why we also include negative LoG filters $-\nabla^2 G_\sigma$ into the dictionary (Figure 4.10(a), bottom row). All filters have zero mean and unit L_2 norm.

4.3.2 Adjustments to the Statistical Model

In this subsection, we discuss the effect of the basis change on the appearance model as well as on the shape model. The appearance models for the LoG filters mus be re-estimated. We follow the same procedure as presented in Section 3.3 for the Gabor filters. The background distribution $q(c_i)$ can be estimated by computing the empirical distribution of LoG filter coefficients. As training data for the background model we use the texture images as depicted in Figure 4.3. In order to reduce the influence of very strong filter responses, we apply a sigmoid function to the filter coefficients $f_i = \eta(c_i, \tau)$. We set the saturation value to $\tau = 3$ in order to induce a background distribution (Figure 4.11(a)) with similar properties as the one of the Gabor features (Figure 4.4(a)).

The distribution favors low feature values, however it also permit‘s strong responses. Compared to the CDF of the shoe print specific Gabor features (Figure 4.4(b)), the CDF of the LoG features has higher probability mass at low feature values (Figure 4.11(b)). Thus, when sampling from the background model, we can expect to generate textures with less structure. The foreground likelihood $p(c_i|\lambda_i)$ for the LoG features is computed according to Equation 3.4. Increasing the value of λ, changes the shape of the distribution as illustrated in Figure 4.11(c) & 4.11(d). We assume that all LoG filters follow the same foreground appearance model.

The impact on the shape model is less extensive. LoG filters have two parameters. Their scale and position. As LoG filters are circular symmetric, they are rotationally invariant. We assume that a basis filter does not change its scale. Thus, the only free parameter is the position $\beta_i^1 = \{X_i^1\}$. Hence, local perturbations of filters are restricted to positional changes $\delta\beta = \{\delta X\}$. The rest of the shape model remains as defined in Section 3.3.

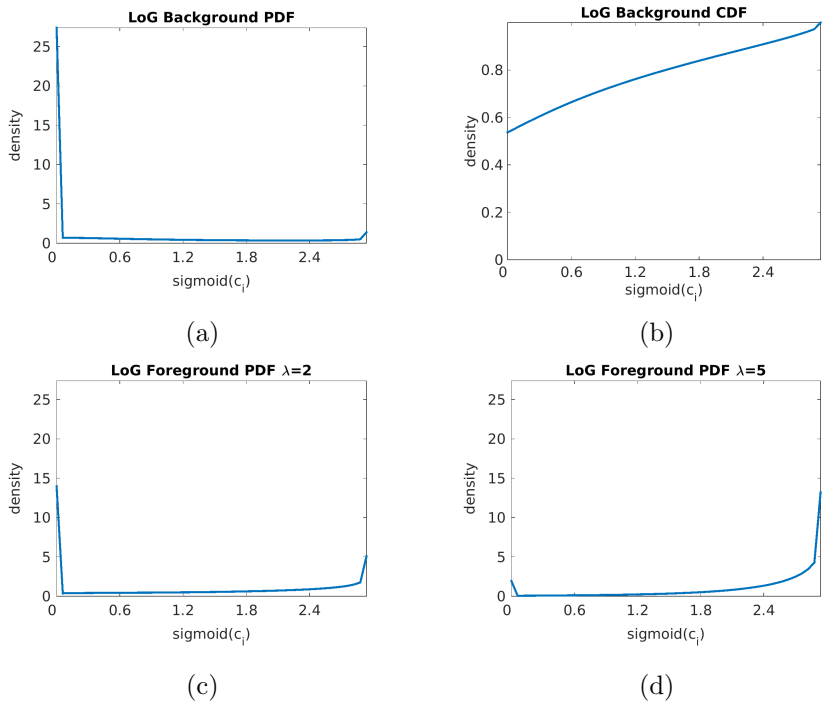


Figure 4.11: Illustration of the statistical appearance models for the LoG basis. (a) The background density $q_{LoG}(c)$. (b) The CDF of the density in (a). (c) The foreground density with $\lambda = 2$. (d) The foreground density with $\lambda = 5$.

4.3.3 Qualitative Experiments

The basis change has a significant effect on the properties of the ABM. In the following, we discuss the differences between the LoG-ABM and the Gabor-ABM in terms of their invariance properties.

Invariance properties. A highly relevant question is: "What information is preserved in a feature representation?". In Chapter 2 we have discussed, that the representation should be invariant to irrelevant transformations of the object. However, it is desirable that it also conserves as much relevant object information as possible. This is a fundamental trade-off for object representations.

In general, it is difficult to know exactly what information is preserved by a feature transformation. One possible way of examining the conservation property of a feature representation is by inverting it back to the image space. However, for most feature transformations such as SIFT ([Lowe, 2004]) or HOG ([Dalal and Triggs, 2005]) this inversion can only be approximated (see [Vondrick et al., 2013]. Even for the relatively simple Gabor feature transform as used in the original ABM (Section 3.3) we can not recover exactly which contribution the even and off filters have. For the LoG features an exact inversion is possible, since the sigmoid transform is bijective and thus can be inverted uniquely. We can then render the filter back into the image.

In Figure 4.12 we show two gallery images and their reconstructions from the Gabor and LoG feature spaces. We can observe, that the LoG representation (Figure 4.12(c) & 4.12(f)) better reconstructs the input image compared to the Gabor representation (Figure 4.12(b) & 4.12(e)). For example, the dotted pattern of the reference in Figure 4.12(d), as well as the direction of edges are much better preserved.

In Figure 4.13 we have illustrated the conservation property of the LoG feature transformation under Gaussian white noise with different variance. For noise with low variance (Figure 4.13(a) & 4.13(b)) the object is reconstructed well. Even under strong noise distortion (Figure 4.13(c)) the structure of the object is still preserved to a large extend. Under very strong noise (Figure 4.13(d)) the structure of the shoe print is lost to a large extent, whereas in the pixel representation, the structure of the shoe can still be guessed by a human observer. The cause of this phenomenon is that the local support of the filters does not capture long range correlations. As a consequence, the independent encoding of the image via matching pursuit fails to decide on the correct filter. This in turn distorts the reconstructed image, whereas in the orignal image the long range correlation is still present which the human vision can leverage to recover the true structure.

Samples from a LoG-ABM. In Figure 4.14 we illustrate samples from a LoG-ABM. The image we used to train the ABM is depicted in Figure 4.14(a). The learned basis decomposition is illustrated schematically in Figure 4.14(b). We illustrate LoG filters as green circles. The sign of the LoG filter is encoded by the color of the circle. Negative LoGs are colored in dark green, whereas positive ones are light green. The size of the circle depends on the scale parameter σ. Figure 4.14(c) shows a sample from the shape model with a local perturbation of $\delta\beta = \{\delta X = 5 \text{ pixel}\}$. Visually, it seems that the internal structure of the shoe print is distorted stronger compared to the Gabor-ABM (Figure 4.5(b)). Figure 4.14(d) shows a sample from the foreground model with fixed geometry. The LoG model preserves the valleys and ridges in between the edges, which is not the case in the Gabor appearance model (Figure 4.8(a))). A sample from the LoG background model is shown in Figure 4.14(e). Compared to the Gabor background (Figure 3.6(b)), the LoG background is less structured with larger regions of low gradients. This is expected, the background model has less probability mass at large feature values (Section 4.3.2). In addition, the filter does not encode directional information of the gradient. Which is a key factor for the perception of "structure".

4.4 Conclusion

In this chapter, we discussed how the ABM can be applied to compute the similarity between two shoe prints. We identified the models sensitivity under occlusion and introduced an independent occlusion model which resolved this sensitivity. Our study then focused on improving the specificity of the ABM. We proposed to change the basis filters and demonstrated a significant improvement in terms of the preserved object information

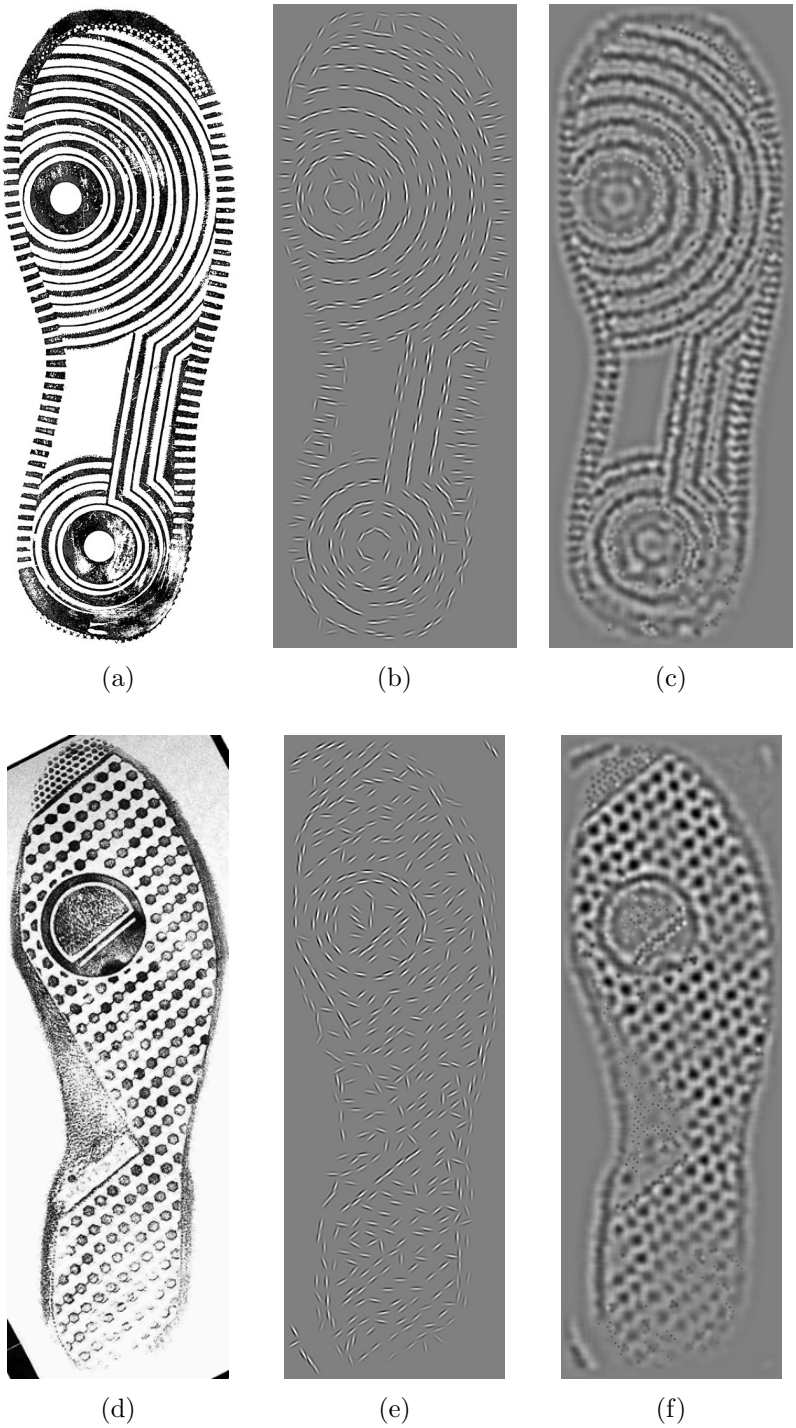


Figure 4.12: Visualization of the information which is preserved by the Gabor and LoG feature transformations. (a & d) Training images. (b & e) The images in (a & d) are projected onto the Gabor basis with matching pursuit and subsequently projection back into the image space. (c & f) The preserved information after a projection onto the basis of LoG filters.

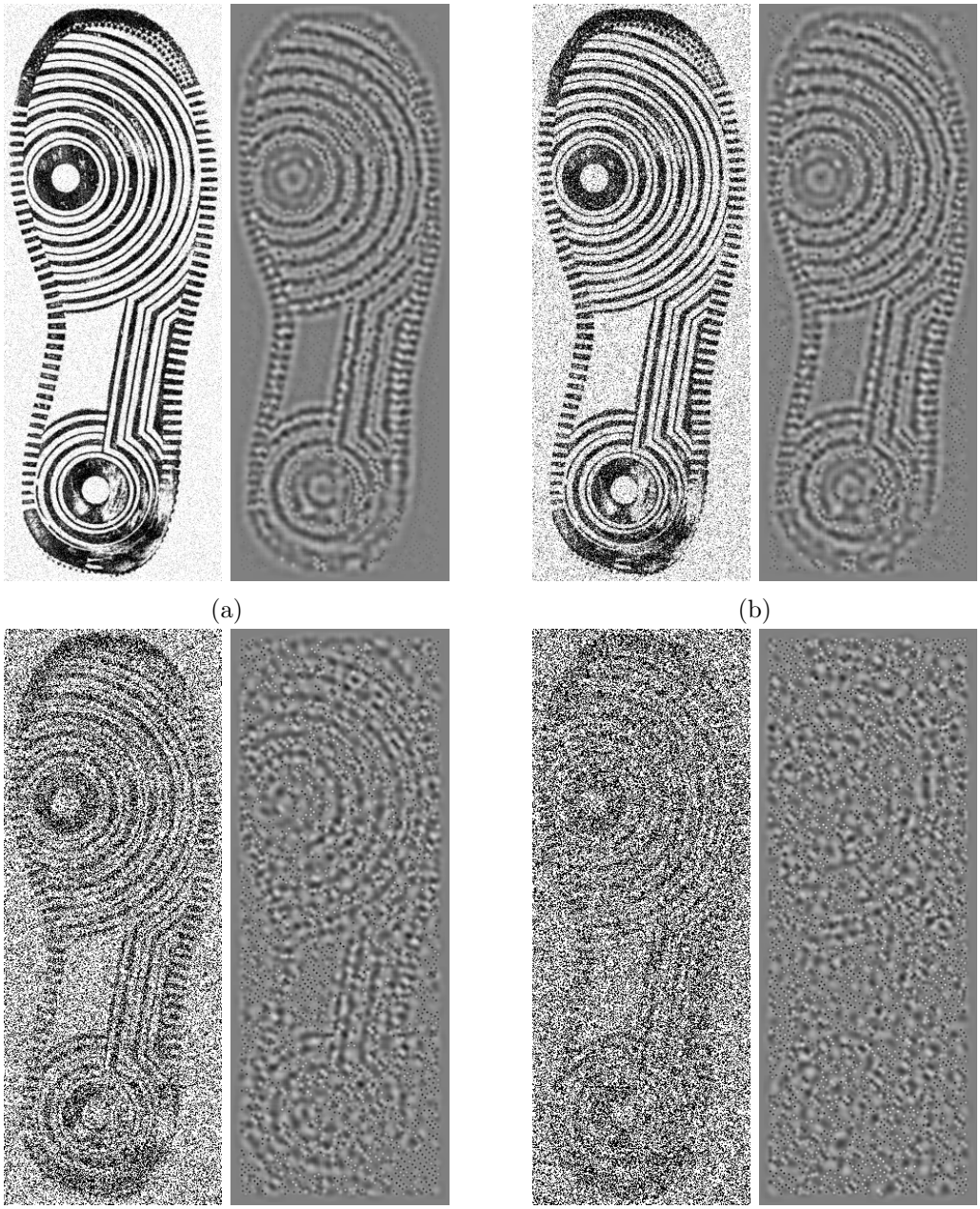


Figure 4.13: Preserved information in the LoG feature space. (a-d) Noisy images (left) are reconstructed with LoG features (right). The input images are generated by adding different amount of independent Gaussian noise to a gallery image with (a) $\sigma^2 = .01$, (b) $\sigma^2 = 0.1$, (c) $\sigma^2 = 1$ and (d) $\sigma^2 = 3$. From the reconstruction we can observe that the feature representation is insensitive to small amounts of independent noise (a) & (b). The main characteristics of the pattern are still preserved under strong noise (c). Under very strong noise (d), many details of the object are lost in the reconstruction which are still visible in the image.

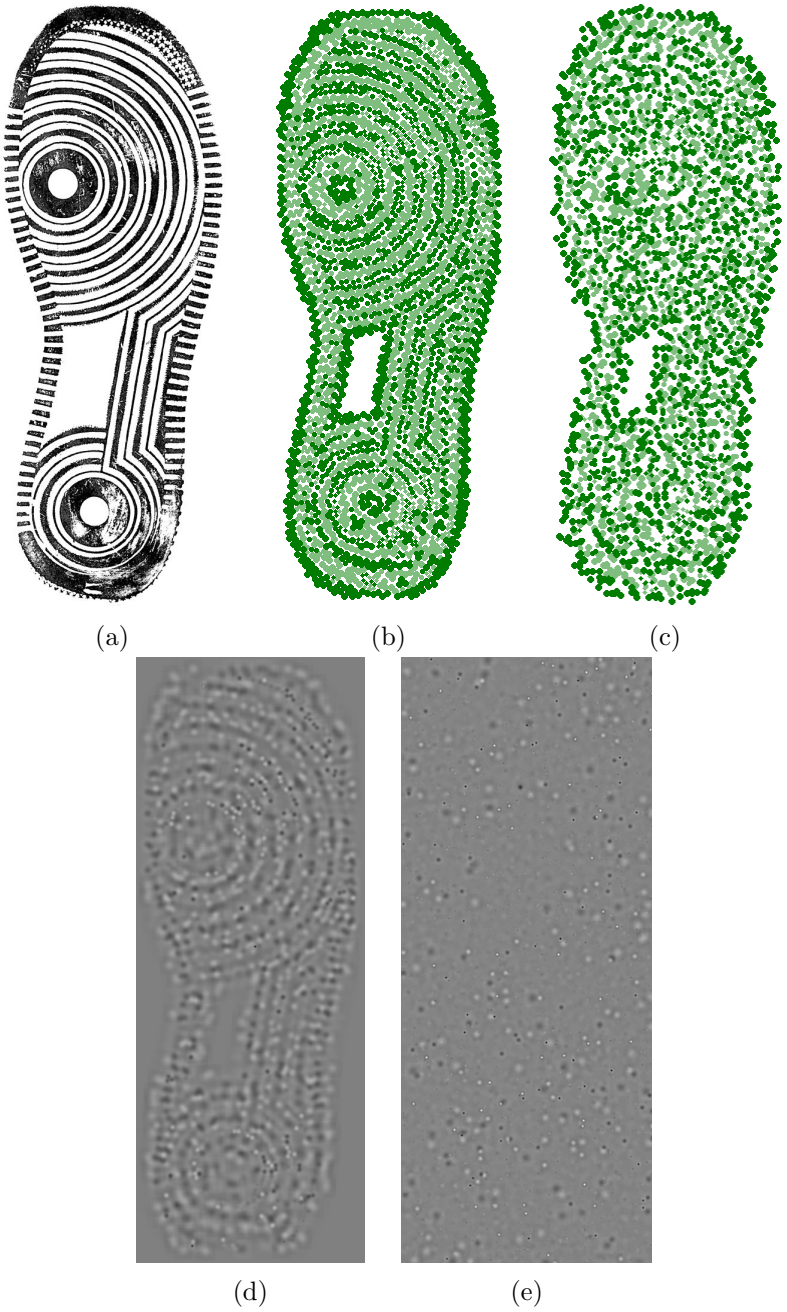


Figure 4.14: Samples from the LoG-ABM . (a) The training image. (b) The learned basis decomposition. LoG filters are illustrated as small circles. The size of a circle depends on the sale parameter σ. The color of a circle encodes the sign of the LoG filter. Negative LoGs are illustrated in dark green, positive ones in light green. (c) A sample from the shape model with $\delta\beta = \{\delta X = 5 \text{ pixel}\}$. (d) A sample from the foreground appearance model with fixed shape. (e) A sample from the background appearance model.

in the feature space. However, the improved representation of the specific object details is bought at a cost. The LoG basis filter is rotationally invariant. Hence, an important shape property is lost. In the next chapter, we will discuss how this information can be recovered by introducing hierarchical dependencies between the basis filters. In this way, the object model will have an increased specificity, while retaining the access to local directional information.

Chapter 5

Multi-Layer Compositional Active Basis Model

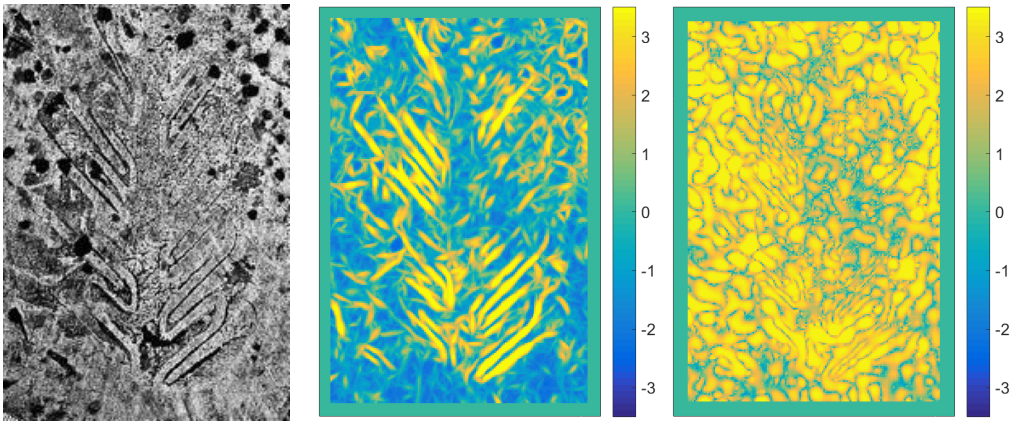
In this chapter, we will revisit and extend the Compositional Active Basis Model (CABM). In Section 5.1, we will identify several limitations of the presented LoG-ABM in the context of shoe print analysis. Subsequently, we will discuss how these limitations can be resolved by introducing a hierarchical dependence structure between the individual basis filters. This will bring us to the concept of a CABM, which we will revisit in Section 5.2. CABMs are hierarchical part-based models which offer major benefits over the LoG-ABM in terms of:

1. The modeling of large shape deformations
2. The ability to discriminate between foreground and background at the part level
3. The local coherence of occlusion states

In Section 5.3, we propose an algorithm that can learn the hierarchical dependency structure of a two-layered CABM from data. We will study the advantage of a hierarchical model structure over the flat ABM in Section 5.4 and propose to generalize our structure induction to learn multi-layered CABMs in Section 5.5. We will study the advantages and weaknesses of the resulting multi-layer CABM in detail. This study will reveal interesting insights about the role of the hierarchical abstraction in object representations.

5.1 Limitations of the LoG-ABM

Despite its significant advantages over the Gabor-ABM in terms of the generative ability, the LoG ABM also reveals major limitations:
1.) Extensive independence assumptions. Due to the extensive independence assumption in the ABM, the states of basis filters are independent given the root node. This leads to inconsistencies in the deformation and occlusion model. From Figure 4.14(b) & 4.14(c) we can observe, that the independence of basis filters induces a loss of



(a) (b) (c)

Figure 5.1: Classification of a shoe print image into foreground and background based on appearance. (a) The probe image. (b & c) Classification scores of the Gabor and LoG appearance models. Positive scores classify as foreground and negative scores as background. Compared to the LoG appearance the Gabor appearance is more selective.

structure under large deformations. However, this characteristic structural information is critical for the recognition process. In addition, the assumption that each part can be occluded independently from its neighbors ignores the fact that occlusions are locally correlated process.

2.) Poor discriminative ability at the part level. In Figure 5.1 we compare the LoG and Gabor appearance models in terms of the ability to classify the image into foreground and background based on the filter coefficients. The classification score can be computed as the log-likelihood ratio $\log(\frac{p(c_i|\lambda_i)}{q(c_i)})$ between the foreground and background appearance models. In Figure 5.1(a), we depict a typical probe image. Figure 5.1(b) shows the classification score with the Gabor dictionary (as illustrated in Figure 3.1). We illustrate at each position the maximal classification score of all filters in the dictionary. The Gabor dictionary shows a high discriminative ability, since most of the background has a negative log-likelihood ratio, while for the foreground it is mostly positive. In Figure 5.1(c) we illustrate the classification score for the LoG dictionary, which is far less discriminative than the Gabor dictionary. This can be attributed to the weak background model (Figure 4.11(a)) which even for small feature values f_i is outperformed by the foreground appearance model (Figure 4.11(c)). For this reason, the background model is not capable to explain structured background well. The task of discriminating foreground from background is therefore shifted to a large extent from the part-level to the shape model which combines the individual part scores. This is, in principle, a desirable mechanism as long range contextual constraints are available to the shape model. However, if parts of the image could be assigned to the background reliably, this would increase the overall reliability of the recognition process.

3.) Missing information about part orientation at the root-level. LoG filters are circular-symmetric and therefore do not represent the local gradient direction. Hence, a discriminative shape property is not explicitly available to the object model. This is a strong limitation with regard to our final goal of discriminating shoe prints.

In the remainder of this chapter, we will discuss how these limitations can be resolved by the introduction of hierarchical dependencies between the basis filters.

5.2 Prior Work on the Compositional Active Basis Model

In this section, we will present prior work on how hierarchical dependencies can be introduced between nearby basis filter in the ABM. For the illustration of these ideas, we will use our proposed LoG basis.
The concept of hierarchical part-based models is well known in Computer Vision and has been successfully applied in a diverse set of applications in the form of discriminative part-based models [Girshick, 2012; Felzenszwalb, 2005] and generative part-based models [Jin and Geman, 2006; Wu et al., 2007]. Along the line of this research the CABM has been developed. A hierarchical Active Basis Model was initially mentioned in [Hong et al., 2013] and then proposed in [Si and Zhu, 2013; Dai et al., 2014]. We will study this model in detail throughout this section. We begin with an overview about the theoretical extension to the original ABM framework (Section 5.2.1). We continue to study how the hierarchical dependency structure can be learned from data by first introducing the method proposed in [Dai et al., 2014] (Section 5.2.2). We highlight its weaknesses and propose an improved learning scheme in Section 5.3.3 which is based on a unsupervised clustering technique that we discuss in Section 5.3.2. Then we discuss the benefits of the learned CABMs over the original ABM representation in Section 5.4. Finally, we show how the proposed unsupervised learning scheme can be applied to learn higher-order CABMs with multiple levels of abstraction in Section 5.5.

5.2.1 Overview

Figure 5.2 graphically illustrates the dependency structure of a CABM. We depict the full image model of a two-layered CABM as a plate graph in Figure 5.2(a). Figure 5.2(b) focuses on illustrating the hierarchical tree structure of the object model. The CABM is composed of $N_2 = 4$ filters with geometric parameters $\{\beta_i^1 | i = 1, \dots, 4\}$. The novelty compared to the original ABM (Figure 3.7) is that an intermediate layer of dependency has been added to the shape model (Figure 5.2(b) Layer 2). The variables of "Layer 1" are grouped into $N_1 = 2$ groups $\{(\beta_1^1, \beta_2^1), (\beta_3^1, \beta_4^1)\}$. Each group depends on intermediate geometric parameters $\{(\beta_1^2), (\beta_2^2)\}$. These intermediate parameters in turn depend on the root node β^3. Following the principle of relative encoding as introduced in Section 3.2, the parameters β_i^1 are encoded relative to their parents β_j^2, whereas β_j^2 is encoded relative to the root β^3. The absolute parameters of a filter in the image frame therefore

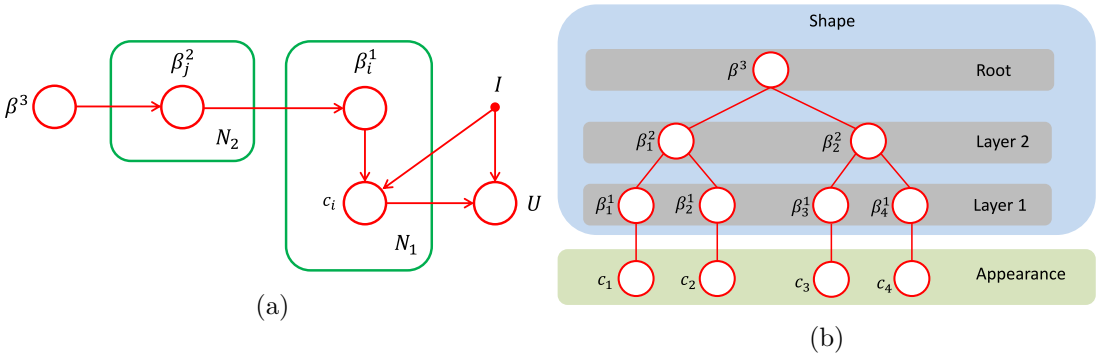


Figure 5.2: The dependence structure between random variables in a CABM. (a) An illustration of the full graphical model. Compared to the original ABM (Figure 3.7(a)), a new layer of variables has been introduced between the root node and the first layer. This new layer only has an effect on the shape model of the object. In (b) we depict the simplest possible CABM, a binary-tree structured Markov random field.

can be computed recursively e.g. as $\beta_1 = \beta^3 + \Delta\beta_1^2 + \Delta\beta_1^1$. The effect of this hierarchical dependency is that given β^3 the two *groups* of filters are independent, whereas within a group the filters are still depending on their parent β_j^2. This group-wise independence structure is beneficial compared to the element-wise independence of the original ABM. Parts in the same group can be forced to have similar positions and orientations. As a result, the flexibility of the deformation model can be increased, without loosing the overall structure of the object. At the same time, the model still has a tree structure. Hence, the efficient global bottom-up inference as discussed in Section 3.4.1 can be applied. Conceptually, this construct can be interpreted as an Active Basis Model which is in turn composed of Active Basis Models. Thus the name Compositional Active Basis Model.

According to the graphical model in Figure 5.2 the full probabilistic CABM can be written as:

$$p(\Theta|O^2) = p(\beta^3) \prod_{j \in ch(\beta^3)} p(\beta_j^2|\beta^3) \prod_{i \in ch(\beta_j^2)} p(\beta_i^1|\beta_j^2) p(c_i|\lambda_i) \prod_{k=1}^{M} q(c_k). \tag{5.1}$$

The superscript of the object variable O^2 highlights that it is a CABM with two layers of shape deformations. We will from now on denote a standard ABM as O^1. The operator $ch(\beta^3)$ selects the set of children nodes of the root β^3. Compared to the original Equation 3.7, an additional deformation model is introduced by the factor $p(\beta_i^1|\beta_j^2)$. It influences the geometric parameters of a whole group of the individual basis filters $(ch(\beta_j^2))$. In this way, the global dependency structure is broken into multiple conditionally independent groups. This two layer factorization defines a hierarchical deformation model. Following the original ABM framework, the conditional distributions of the geometric parameters

are defined as uniform perturbations:

$$p(\beta_j^2|\beta^3) = \mathcal{U}(\beta_j^2 - \delta\beta^2\ , \beta_j^2 + \delta\beta^2) \tag{5.2}$$

$$p(\beta_i^1|\beta_j^2) = \mathcal{U}(\beta_i^1 - \delta\beta^1\ , \beta_i^1 + \delta\beta^1) \tag{5.3}$$

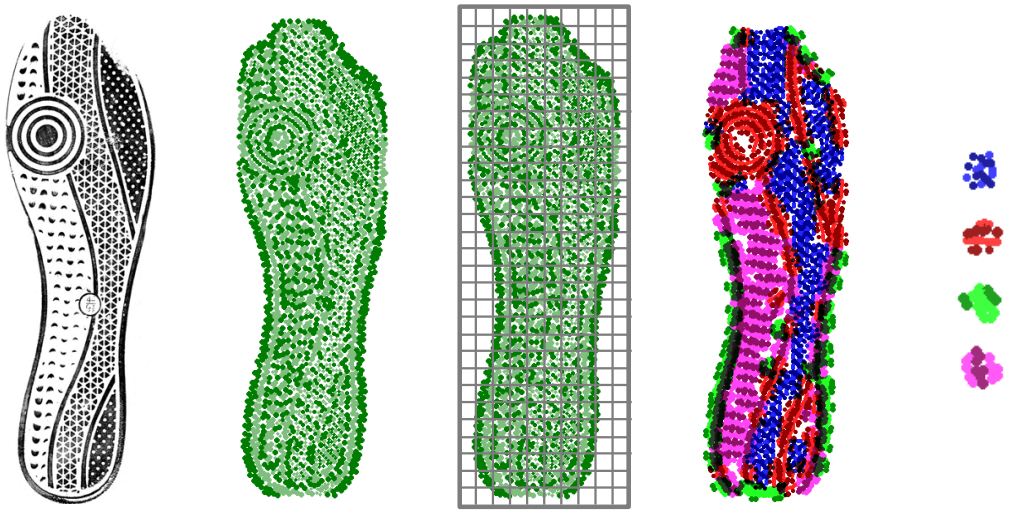
We will denote the deformation parameters of a CABM as $\delta\beta = \{\delta\beta^1, \delta\beta^2\}$. The idea is to cover large perturbations by $\delta\beta^2$ such that the groups of filters move together, whereas small perturbations should be covered by $\delta\beta^1$. Importantly, when LoG filters are used as basis, the parameters β_i^1 will only cover positional change, since LoG filters are circular symmetric. However, the groups of LoG filters in layer two will have a relative orientation to the root node β^3. Thus, local directional information will be accessible to the global shape model. This resolves the second point of the limitations of LoG-ABMs which we discussed in the previous Section 5.1.

Comment on the Uniform Deformation Prior: The assumption that the perturbation of parts follows a uniform distribution is common for work on ABMs [Wu et al., 2010] as well as CABMs [Dai et al., 2014]. However, parts of objects are naturally more likely to stay close to their initial position than to move far away. Therefore, in the literature on statistical shape models large deformations are typically penalized [Grenander et al., 1990]. The positive feature of the uniform deformation prior is that a parts deformation cost can be computed efficiently during the bottom-up inference. Since the uniform distribution is constant, during inference it just requires the addition of a constant number independent of the part position. A non uniform prior such as e.g. a Gaussian would require to keep track where the part moves to. Ultimately, this would require a point wise matrix multiplication at each possible part position for each possible orientation. We will therefore follow the uniform assumption throughout this work, however, we are aware that this assumption should be resolved in the future.

The additional complexity of the hierarchical model comes at the cost of having to learn the structure of the model from data. Essentially, the learning process must find an answer to the following three questions: How many filters should form a group? How many groups should be used to represent the target object? And how should the groups be arranged spatially? The authors in [Dai et al., 2014] design the hierarchical model structure manually. In the following, we discuss the downsides of this approach and propose a novel method to learn the structure automatically.

5.2.2 Grid-based Design of the CABM Structure

In the original work on CABMs [Dai et al., 2014], the authors propose to design the hierarchical structure of CABMs manually. Although, their focus is on learning CABMs from multiple natural images, it can be applied in a straight forward manner to the application of shoe print modeling. A CABM consists of an appearance model and a hierarchical shape model. Since the appearance model is the same as in the original ABM it can be estimated in the same way (Section 4.3.2). The hierarchical shape



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

Figure 5.3: Comparison of a grid-based composition of basis filters with our proposed learning-based method. (a) The training image. (b) The decomposition of (a) with a LoG dictionary as introduced in Section 4.3. (c) Schematic illustration of the grid-based grouping of filters as proposed in [Dai et al., 2014]. Basis filters which fall into the same square region are grouped into a local ABM. (d) & (e): Result of our learning-based method as presented in Section 5.3.3. The four ABMs which were learned from the training image in (a) are depicted in (e). The encoding of the image with the learned ABMs is illustrated in (d).

model has three types of shape parameters $\mathcal{R} = \{\beta^3, \beta_j^2, \beta_i^1\}$ (Fig5.2), which must be estimated from data. Given an image (Figure 5.3(a)), we can trivially set the models global position to be the center of the image X_c and its global orientation to be upright, thus $\beta^3 = \{X^3 = X_c, \alpha^3 = 0°\}$. [Dai et al., 2014] propose to learn the other parameters as follows.

First, the basis decomposition is applied as we have discussed in Section 3.2 (Figure 5.3(b)). In this way the absolute position of each filter in the image frame X_i is computed. This position gives rise to the shape parameters of the first layer $\beta_i^1 = \{X_i^1 = X_i\}$ (Section 5.2.1). In order to define the complete CABM structure, the variables in the intermediate layer $\{\beta_j^2 | j = 1, \ldots, M\}$ must be estimated together with the parent child relation. In order to do so, the authors propose to divide the image into a grid of M equally sized square region. Figure 5.3(c) illustrates schematically a grid of 30×10 squares. The position of each square in the image frame is denoted by X_m. The size of the of the squares is fixed a-priori. The filters which fall into the same square region form a local ABM. The parent node is set to be $\beta_j^2 = \{X_j^2 = X_m, \alpha_j^2 = 0°\}$. At this point, the complete CABM shape structure is defined.

The major limitation of this approach is that it does not account for the redundancy in the structure of the training image. Ignoring this redundancy has several drawbacks

in terms of computational efficiency, its discriminative ability and its efficiency when learning from additional data. In the next section, we will discuss these drawbacks and propose a novel approach to the induction of a CABMs structure.

5.3 Learning the CABM Structure

In this section, we propose a learning-based approach to the induction of a CABMs structure. In Figure 5.3(d) we illustrate the result of applying our algorithm to the image in Figure 5.3(a). The training image is encoded by translating and rotating four different local ABMs (Figure 5.3(e)). Each ABM is color-coded in order to be able to identify it in Figure 5.3(d). Compared to the grid-based result 5.3(c) the result is more meaningful and exploits the redundancy of the shoe print pattern. This representation offers three major advantages over the grid-based method:
1.) It is computationally more efficient. The total number of different local ABMs has been reduced from $10 \times 30 = 300$ in the grid-based result to just four. Thus, the memory consumption during inference is reduced by about a factor of 75.
2.) The training image is semantically interpreted. Similar elemental patterns of the image are encoded with the same ABM. This additional knowledge is highly useful for higher-order tasks such as e.g. the learning of a multi-layer CABM as we will propose in Section 5.5.
Another advantage that we will not explore in this thesis is:
3.) Efficient learning from additional data. By exploiting the redundancy in the representation, we can extract multiple training patches from a single new training image. We could leverage these training patches to improve the appearance or shape model of the local ABMs.
In the following we discuss related work on learning hierarchical deformable models from images.

5.3.1 Related Work on Learning Hierarchical Deformable Models

In the context of hierarchical deformable models only few works have been proposed for inducing hierarchical compositional shape models. In contrast to [Jin and Geman, 2006; Dai et al., 2014; Si and Zhu, 2013], where at least part of the structure of the models was manually set, we induce the whole structure of the model automatically. Some works are based purely on contours e.g. [Ferrari et al., 2010; Kokkinos and Yuille, 2011]. However, these suffer from the same drawbacks as the Gabor-based ABMs as they only conserve high frequent edge information about an object. So far, only Filder et al. [Fidler and Leonardis, 2007; Fidler et al., 2014] and L. Zhu et al. [Zhu et al., 2008] have successfully learned the complete structure of fully generative hierarchical compositional models from images.
In the following Subsection 5.3.2, we revisit an algorithm which is capable of learning the local ABMs depicted in Figure 5.3(e). The algorithm assumes that the number of different groups is known a-priori. We will extend the algorithm in Subsection 5.3.3 such

that the number of groups is also learned from the data. In this way, the full hierarchical structure of a CABM can be learned automatically from a single image.

5.3.2 EM-type Learning of Dictionaries of ABMs

The process underlying the EM-type algorithm as proposed in [Hong et al., 2013] is simple. First, the number of ABMs N_{ABM} is fixed. In order to initialize the learning process, patches of fixed size $d \times d$ pixels are sampled at random positions and orientations from the training image I and randomly assigned to one of the N_{ABM} clusters. The learning process then follows an EM-type learning scheme iteratively for a fixed number of iterations:

1. **Learning the Models**: Learn an Active Basis Model $O_n^1(\Theta)$ from each cluster of patches with *shared matching pursuit* (M-step).

2. **Detection**: The models $O_n^1(\Theta)$ are detected in the training image at different positions and orientations. Patches are cut out at the detected positions and serve as new training data for the next learning iteration (E-step).

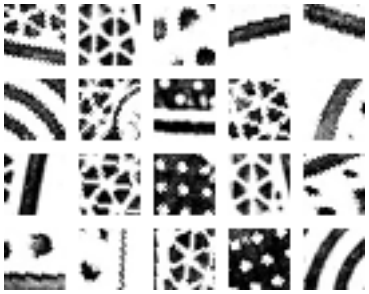
From the general principle, the process is similar to the Expectation Maximization algorithm for learning Gaussian mixture models. The main difference is that the data is not modeled with a Gaussian distribution but with an ABM. In the following we will explain the full learning process in detail. The gallery image depicted in Figure 5.5(a) will serve as training image. The parameters are set to $N_{ABM} = 4$ clusters with $N = 20$ bases per model. The patch size is set to $d = 41$ pixels. For reference, the height of the training image in Figure 5.5(a) is 580 pixels.

The EM-type learning is initialized by sampling patches randomly from the training image and assigned to one cluster. Subsequently, for each cluster of patches one ABM is learned with the shared matching pursuit algorithm as introduced in [Wu et al., 2007].

Shared Matching Pursuit. Shared matching pursuit is an extension of the matching pursuit algorithm of Section 3.2. It makes possible to learn an ABM from multiple images, which is the main difference to the single image decomposition with matching pursuit that we used so far. The linear additive model as proposed by [Olshausen and Field, 1996] (Equation 3.1) is extended to represent an ensemble of images:

$$I_m = C_m B + U_m = \sum_{i=1}^{N} c_{i,m} B_i + U_m, \tag{5.4}$$

where $\{I_m, m = 1, \dots M\}$ are image patches of size $d \times d$. They are linearly decomposed into a basis B_i, coefficients $c_{i,m}$ and a residual image U_m. The idea is that the basis must be shared across patches. The decomposition can be learned in the same way as in the original matching pursuit algorithm by selecting the optimal bases one by one. The difference is that the image patches must all be reconstructed with the chosen basis.



(a) (b)

Figure 5.4: Illustration of the initial learning stage in the EM-type learning scheme. (a) A set of 20 randomly sampled training patches. (b) A basis decomposition learned with shared matching pursuit on the training set in (a).

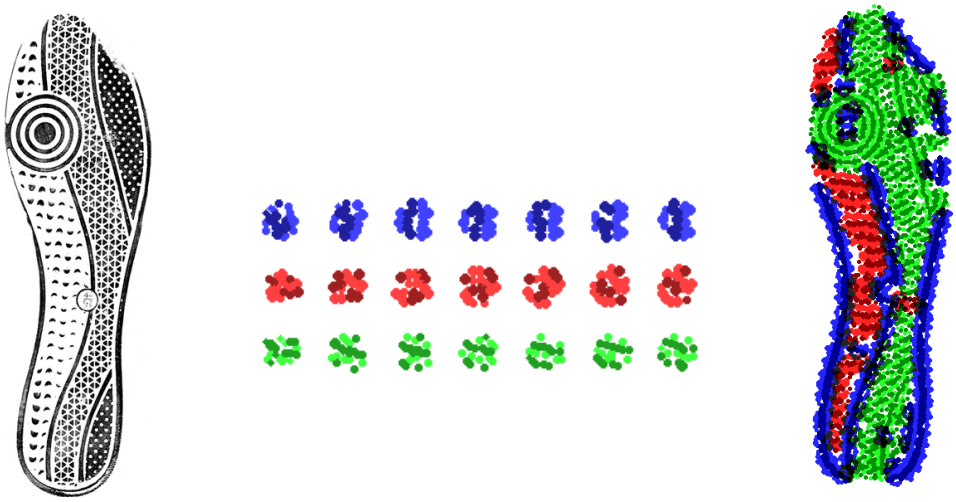
Figure 5.4(a) illustrates 20 randomly sampled patches from the training image and the learned basis decomposition 5.4(b). Given the learned basis decomposition, an ABM is produced by imposing a statistical model on the parameters of the basis as discussed in Section 3.3.

Detection. Based on the the learned ABMs $\{O_n^1(\Theta)|n = 1, \ldots, N_{ABM}\}$ the training data for the next learning iteration is gathered by detecting the ABMs in the training image at different positions and orientations as discussed in Section 3.4.1. The training patches are then cut out at the detected orientation and position. An important feature of the detection phase is that only the best matching ABM is allowed to occupy a part of the image. In this way, a competition process between the ABMs is induced which ultimately results in specialization of the ABMs for modeling different local patterns.

Figure 5.5(b) illustrates the specialization of the individual ABMs throughout the iterations of the EM-type learning procedure. Each column illustrates the result of one learning iteration. The process proceeds from left to right. Due to the initial random assignment of patches to clusters, the ABMs in the first column have no clear internal structure. However, after a few iterations, each ABM has specialized to represent one local pattern of the training image. Typically, the learning process is converged after about eight iterations.

The learned ABMs can be composed into a global CABM by encoding the training image as depicted in Figure 5.3(d). The full structure of the CABM is then given by the global spatial arrangement of the ABMs.

In the context of shoe print recognition it is difficult to set the number of clusters a-priori, since outsole patterns are highly diverse (Figure 5.6). In the following, we propose to extend the EM-type learning scheme with a greedy mechanism. The proposed algorithm will automatically infer the number of ABMs needed for encoding a given training image.



(a) (b) (c)

Figure 5.5: Illustration of the result of the EM-type learning as proposed by [Hong et al., 2013]. (a) The training image. (b) Each row illustrates the evolution of an ABM over time. Each column shows the result for each iteration in the learning process. In the first iteration, the ABMs have no clear internal structure. After a eight iterations, however, each ABM has specialized to represent a particular elemental pattern. (c) The result when detecting the learned ABMs in the training image at different positions and orientations.

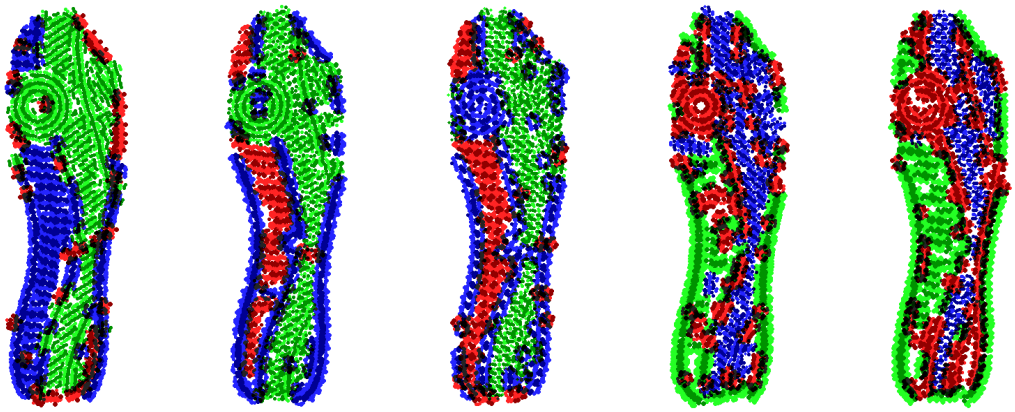
5.3.3 Greedy EM-type Learning of Dictionaries of ABMs

The authors in [Hong et al., 2013] propose to select the optimal number of ABMs for a training image based on a Bayesian Information Criterion(BIC). However, this requires multiple runs of the training procedure with different parameter settings and thus is inefficient. Furthermore, it is questionable if the trade-off parameter in the BIC can be set optimally for all possible object shapes.
An additional issue is that due to the random initialization the learning process is instable in the sense that learning results can be quite different for different trials (Figure 5.7). This issue is typical for randomly initialized clustering algorithms and can be observed e.g. also for k-means clustering. Note that the color of the ABMs is only important for separating the different ABMs in one particular image. There is no relation between the colors of different images. This is also the case for all other images which we will show throughout the rest of this thesis.

In the following, we propose a greedy-learning scheme that enables the automatic determination of the number of ABMs. Furthermore, we introduce a mechanism for spreading out the cluster centers through a careful initial seeding as in the work on the k-means++ algorithm [Arthur and Vassilvitskii, 2007].

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

Figure 5.6: Diversity of the outsole patterns of shoe prints. The shoe prints are sorted roughly according to the redundancy of the outsole patterns. (a) Shows two highly repetitive patterns. One is translational symmetric (the zigzag pattern), one is rotational symmetric (the black outline). Others show multiple translational symmetries (b) or mostly rotational symmetries (c). The print in (d) is only partially repetitive but also contains characteristic patterns which do not repeat at all. (e) Shows a print with hardly any redundant information in it.



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

Figure 5.7: Illustration of the instability of the EM-type algorithm. (a-e) Show five results which were obtained by different runs of the EM-type learning procedure. The results are visualized by detecting the learned ABMs in the training image. Two different local minima in the parameter space can be observed. In (a,b) the concentric circle is encoded with the same ABM as the repetitive patterns right of it. In (c,d,e) the concentric circle is encoded differently from these repetitive patterns. Beware that the color of the ABMs is only important for separating the different ABMs in one particular image. There is no relation between the colors of different images.

We propose to learn the models greedily one at a time instead of learning them all at once in parallel. Figure 5.8 illustrates the difference in the greedy specialization process compared to the parallel specialization as discussed in Section 5.3.2 (Figure 5.5(b)). In the beginning, the number of clusters is set to $N_{ABM} = 2$. As in the original algorithm, both models are initialized with randomly sampled patches. However, in the learning phase only one ABM is adapted. The other ABM only participates in the detection phase serving as a generic competing model (Figure 5.8, ABM with gray background). After a preset number of iterations, the first ABM $O_1^1(\Theta)$ is assumed to be converged. Then two new ABMs are added to the pool of models. This time, the location of the training patches is not sampled uniformly across the image, but inversely proportional to $p(X|I, O_1^1)$. In this way, those regions which are well explained by the learned ABM O_1^1, are less likely to be sampled as training data for the next ABMs. Again, only one of the models will be adapted during the learning phase. The other two serve as competitors in the detection phase. The EM-type scheme proceeds until the second ABM O_2^1 converged. Again, two new ABMs enter the pool of models. This time, the training patches are sampled inversely proportional to $p(X|I, O_1^1, O_2^1)$. In this way, ABMs are learned greedily until a new model is not able to explain some parts of the image better than any previously learned model. The learned models are illustrated in Figure 5.8(b). The encoding of the training images is depicted in Figure 5.9(a) together with four other encodings that were generated by repeating the experiment multiple times. Again, note that the color of the ABMs is only relevant for separating the ABMs in each particular image. There is no relation between the colors of different images. The algorithm always converges at five ABMs, which illustrates its stability. Compared to the result of the EM-type algorithm (Figure 5.7), our greedy extension always manages to separate the representation of the concentric circles from the one of the repetitive patterns in its surrounding. This separation property is beneficial, since each model can be tailored to the characteristics of one pattern and thus can be expected to be more discriminative in the inference process. However, the dotted pattern on the right of the gallery image (Figure 5.5(a)) is always represented with the same ABM as the triangle pattern in the middle. Most likely this is due to the fact that the dotted pattern is very similar to the triangle pattern in the feature space. Therefore, it becomes unlikely that during the initial seeding many of these patterns are chosen to be part of a new cluster.

Figure 5.10 illustrates the learning results and the corresponding image reconstruction for the gallery images depicted in 5.6. The clustering discovers between four and six clusters. For patterns which are repetitive across translation or rotation, the algorithm manages to discover the elemental pattern very well. Even if patterns cover only a small region of the overall image, the algorithm manages to develop a separate ABM (Figure 5.10(d)). Perceptually, the pattern segmentations are very similar to what one would expect as a human observer. However, high-frequent patterns which are non-repetitive within the training image (Figure 5.10(e)) are underrepresented. This is because these patterns do not fulfill the assumption that they can be merged into groups. The algorithm also tends to overestimate the number of models needed to represent an image. For

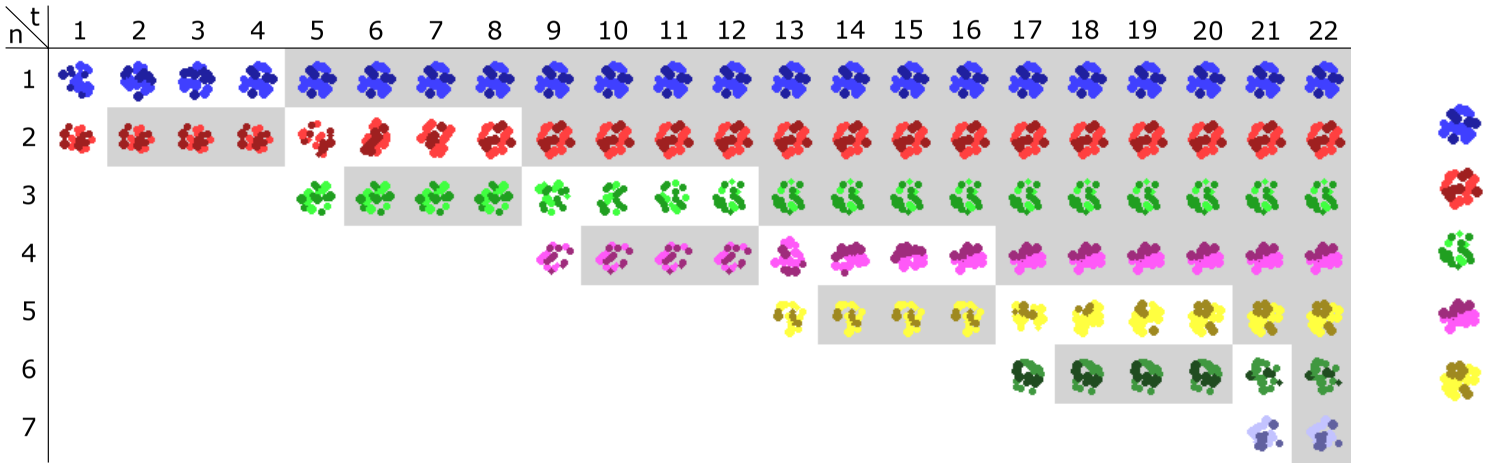
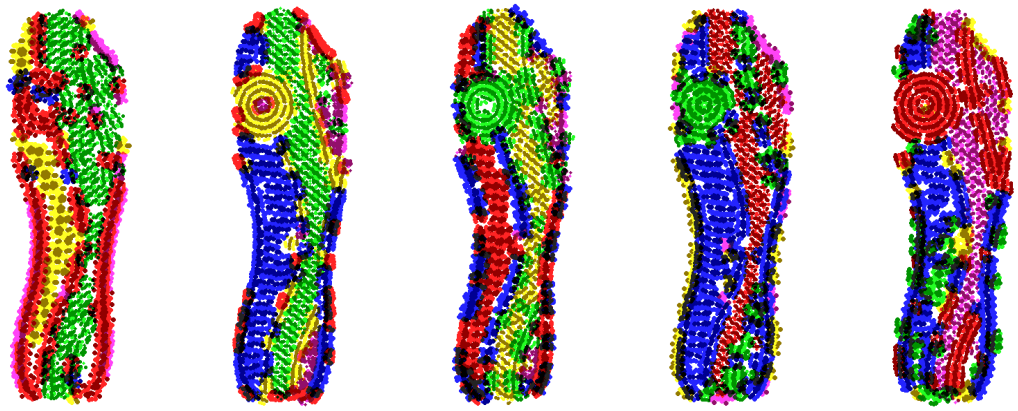


Figure 5.8: Illustration of the greedy EM-type learning scheme at different iterations $t = \{1, \ldots, 22\}$ of the learning process. Each of the $n = 7$ rows illustrates the evolution of an ABM over time. Each column shows the result at each of the 22 iterations of the learning process. In the first iteration $t = 1$, two ABMs are learned from randomly sampled image patches. In order to gather the patches for the next learning iteration, *both* ABMs are detected at different positions and orientations in the training image. From the detected locations patches are cropped and used as training data for the next iteration $t = 2$. However, in this iteration only the first ABM is relearned, whereas the second one is not. Hence, the second ABM only participates in the detection phase serving as a generic competing model. We mark this status by coloring its background in gray. This process of detection and learning is continued until iteration $t = 5$. There, the first ABM is assumed to be converged to a steady solution. Again, two ABMs are learned from randomly sampled patches in the image ($n = 2\&3$). We continue to adapt the second ABM while the first and third only act as competing models in the detection phase. We follow this process until iteration $t = 22$. At this point, the newly learned ABMs ($n = 6\&7$) were not able to be detected in the image. The final learned ABMs are illustrated in (b).



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

Figure 5.9: Illustration of the stability of the proposed greedy EM-type learning scheme. (a-e) Show five results which were obtained by different runs of our greedy EM-type learning procedure. The results are visualized be detecting the learned ABMs in the training image. Compared to the results in Figure 5.7 our algorithm always manages to separate the representation of the concentric circles from the one of the repetitive patterns in its surrounding.

example, in Figure 5.10(a) the green and pink cluster might also be dropped. However, this compression is an essential trade-off in the parametrization of any unsupervised clustering algorithm for which so far no solution has been found so far.
This clustering process could be extended in many ways. The ABMs could be learned on multiple scales, or could be forced to concentrate on edges of certain scales. Also the initial patch sampling could be improved by basing it on higher-order information such as saliency or local self-similarity. However, we will apply the process throughout our experiments as presented in this section.

The learned dictionary of ABMs can be used to induce the structure of a CABMs in the same simple manner as we proposed in the previous Section 5.3.2. First, the ABMs from the dictionary must be detected in the training image. Based on the position and orientation of the detected ABMs, the variables β_j^2 can be computed as in the grid-based algorithm (Section 5.2.2). The parameters β_i^1 can then be computed by projecting the basis filters of each ABM into the image frame.

5.4 Impact of the Hierarchical Dependence Structure

In this section, we study the benefits of the hierarchical dependence structure in a CABM compared to the flat representation in the ABM (Section 4.3). In particular, we will show that the discussed limitations of the LoG-ABM (Section 5.1) are largely resolved in the LoG-CABM.

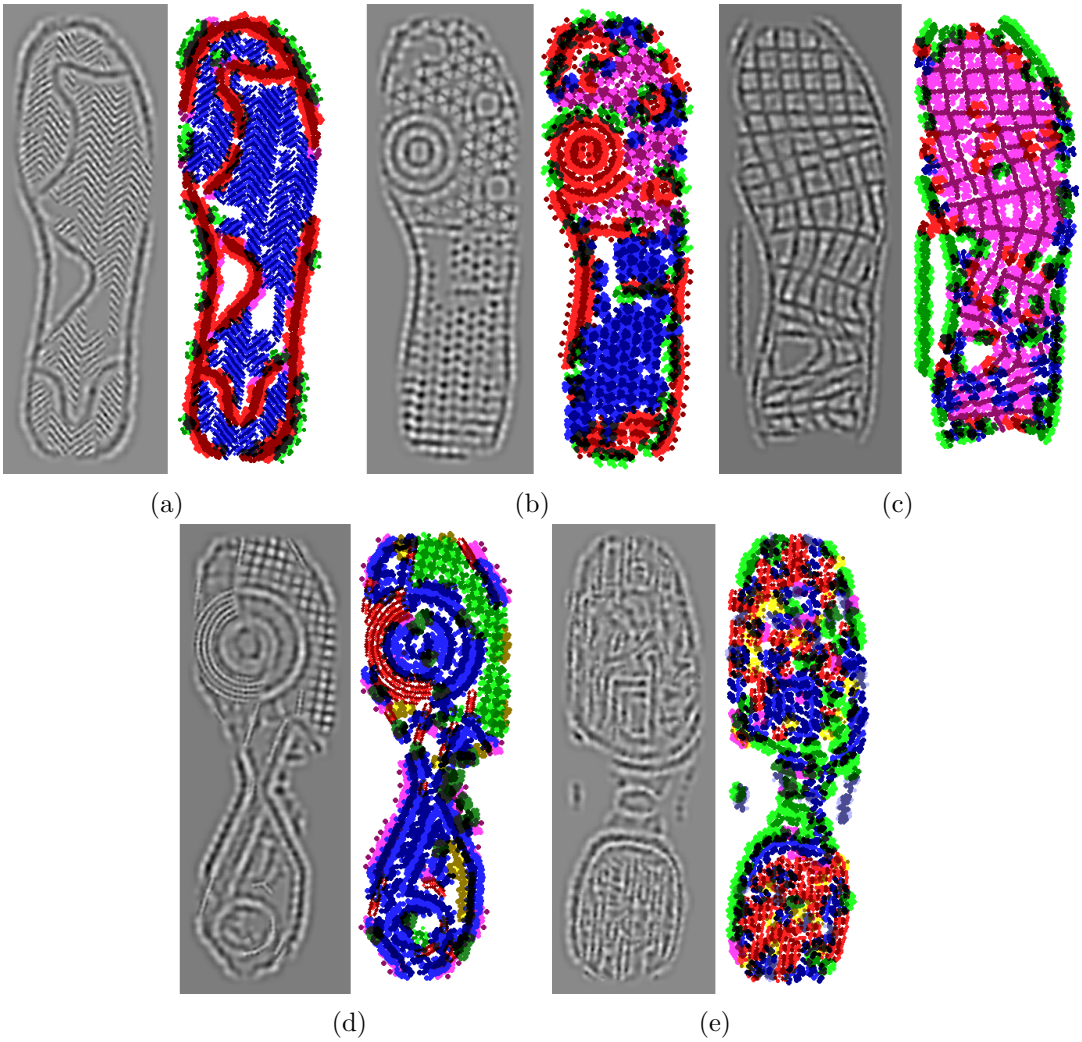
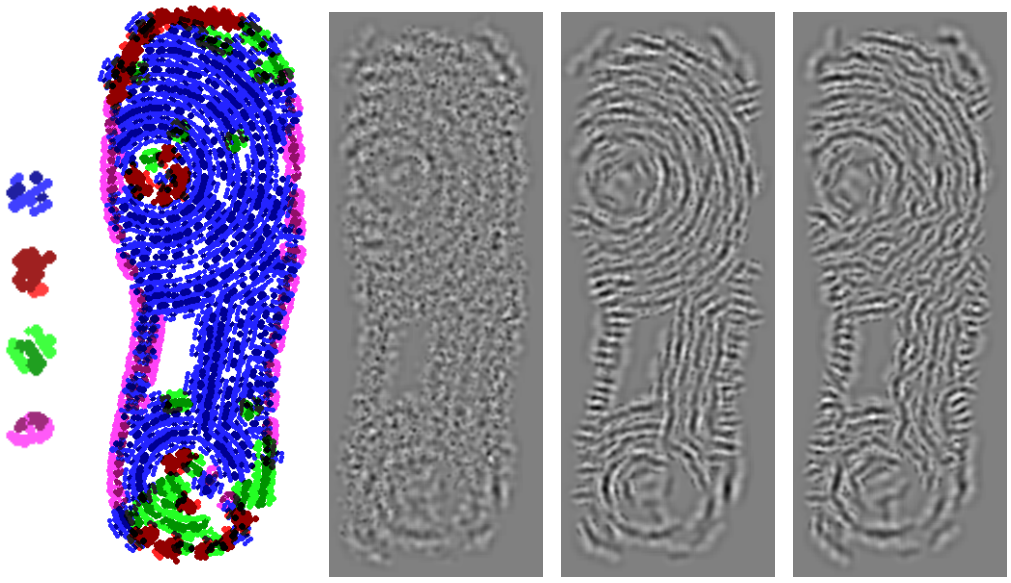


Figure 5.10: Application of the greedy EM-type learning scheme on the set of images depicted in Figure 5.6. Each sub-figure shows two images. On the right is the encoding of the training image with the learned ABM dictionary. On the left, the back projection from the feature space into the image space. The algorithm is capable of exploiting translational and rotational symmetries in the training patterns, while preserving most of the characteristic information of the patterns in the feature space.

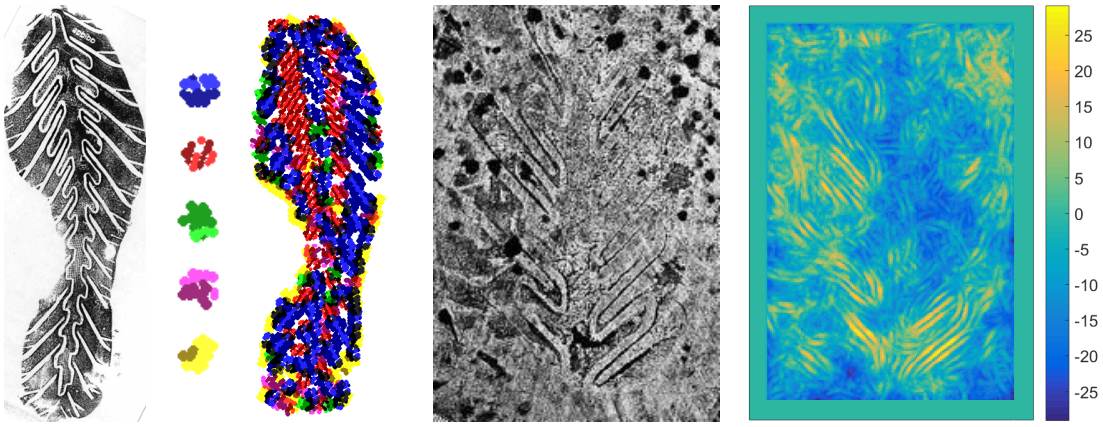
5.4.1 Hierarchical Deformation



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

Figure 5.11: Illustration of the effect of decomposing deformations hierarchically. (a) & (b) show a schematic illustration of a CABM. (a) The learned dictionary of ABMs. The elements are depicted in twice their original size. (b) The global spatial configuration of the ABMs in the CABM. (c-e) Show samples from the deformation model of the CABM. The appearance is fixed. (c) A deformation solely in the first layer by $\delta\beta^1 = \{\delta X^1 = 5$ pixel $\}$. (d) The same amount of deformation in the intermediate layer $\delta\beta^2 = \{\delta X^2 = 5$ pixel $, \delta\alpha^2 = 0°\}$. Due to the group-wise movement which is imposed in layer two, the structure of the object is better preserved. Due to the compositinal layer, the model also has gained control over local directional information. This is illustrated in (e) by sampling from the deformation model with $\delta\beta^2 = \{\delta X^2 = 0$ pixel $, \delta\alpha^2 = 30°\}$.

A major benefit of the hierarchical dependency structure in a CABM is that filters can be forced to stay together during shape deformations. In this way the internal structure of the object is better preserved for large deformations. Figure 5.11 illustrates this property for the CABM which is depicted in Figure 5.11(b). We fix the appearance of the model and sample only from the shape model at two different layers. In Figure 5.11(c) the intermediate layer is fixed and the deformation takes place solely in Layer 1 with $\delta\beta^1 = \{\delta X^1 = 5$ pixel $\}$. Due to the large, locally independent movement of the filters, the structure of the object is lost to a large extent. In Figure 5.11(d), the same amount of deformation is performed on the intermediate layer $\delta\beta^2 = \{\delta X^2 = 5$ pixel $, \delta\alpha^2 = 0°\}$ while keeping the bottom layer fixed. The potential extent of movement stays the same, however due to the group-wise constraint, the internal structure of the object is much better preserved. In addition, the group wise dependency of the LoG filters provides



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

Figure 5.12: Illustration of the discriminative ability of an ABM. (a) The training image. (b) The ABMs learned from the training image with the proposed greedy EM-type algorithm. (c) The encoding of the training image with the learned ABMs. We can observe that e.g. the red ABM is specialized to represent the characteristic geometry of the repeating ridge structure at the center of the training image. (d) A probe image which depicts this repeating ridge structure surrounded by strong clutter. (e) The appearance score of the red ABM on the probe image. Due to its specialization, it can discriminate the foreground from the background well.

access to directional information, which can also be varied with the deformation model. We show a sample with $\delta\beta^2 = \{\delta X^2 = 0 \text{ pixel }, \delta\alpha^2 = 30°\}$ in Figure 5.11(e).
Overall the hierarchical deformation model largely resolves the limitation of a loss of structure under large deformations (Section 5.1). This makes it possible to introduce more flexibility in the object model while better preserving the characteristic features of the object.

5.4.2 Discriminative Ability at the Part-level

In Section 4.3, we have discussed the deficit of LoG appearance models compared to Gabor appearance models in terms of the ability to discriminate foreground from background in a probe image (Figure 5.1(c)). As a reference, we depict the same probe image in Figure 5.12(d). Figure 5.12(a) shows its corresponding gallery image. The CABM which was learned with our greedy EM-type learning is illustrated in Figure 5.12(c). We can see that multiple ABMs have been learned (Figure 5.12(b)), each specialized to one particular elemental pattern of the training image. Figure 5.12(e) illustrates the maximal log-likelihood ratio of the ABM that is colored in red in Figure 5.12(c) evaluated in each position of the image (Equation 3.8). The ratio is maximized over all possible orientation of the ABM. We can observe that the foreground and background patterns

are much better classified compared to the individual LoG appearance models (Figure 5.1(c)). Thus the limitation that the LoG ABM lacks discriminative power at the part level is resolved to a large extent (Section 5.1).

5.4.3 Occlusion Coherence

(a) (b) (c)

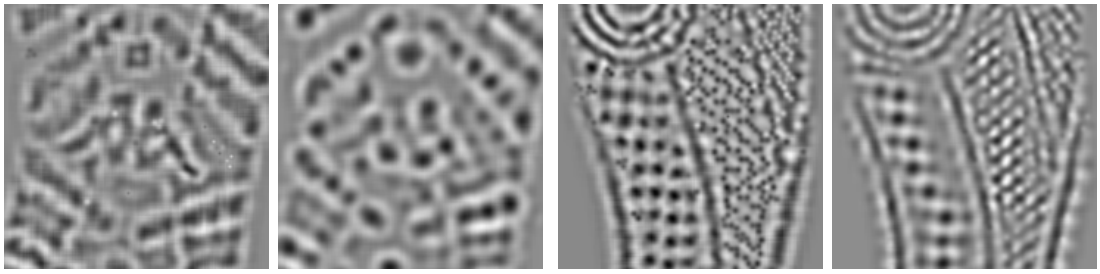
Figure 5.13: Coherent occlusion in the CABM. (a-c) Show samples of the CABM model with fixed geometry and appearance but with different occlusion models. (a) No occlusion. (b) Independent occlusion as introduced in Section 4.2. (c) The coherent occlusion model proposed in this section.

In Section 4.2 we have proposed an independent occlusion model for the ABM. CABMs can be augmented with the same occlusion model by changing the original appearance model with the one we proposed in Equation 4.3.

$$p(\Theta|O^2) = p(\beta^3) \prod_{j \in ch(\beta^3)} p(\beta_j^2|\beta^3) \prod_{i \in ch(\beta_j^2)} p(\beta_i^1|\beta_j^2) p(c_i|z_i, \lambda_i) \prod_{k=1}^{M} q(c_k). \tag{5.5}$$

Due to the hierarchical model structure, we can enforce coherence between the occlusion states of filters by letting all children of a node β_j^2 share a common occlusion variable z_j:

$$p(\Theta|O^2) = p(\beta^3) \prod_{j \in ch(\beta^3)} p(\beta_j^2|\beta^3) \prod_{i \in ch(\beta_j^2)} p(\beta_i^1|\beta_j^2) p(c_i|z_j, \lambda_i) \prod_{k=1}^{M} q(c_k). \tag{5.6}$$



(a) (b)

Figure 5.14: Information loss in the CABM representation compared to the ABM representation. In (a) & (b) the left image depicts the ABM appearance and the right image the CABM appearance of the training image. (a) Loss of information which is not repetitive across orientation or translation, e.g. the trademark in the center of the shoe. (b) Loss of high frequent patterns such as e.g. the dotted structures.

Note how the appearance model in the above equation depends on a common occlusion variable z_j, whereas in Equation 5.5 each model has its own variable z_i. Due to the sharing, occlusion coherence is enforced as either all filters of a subtree are jointly occluded or visible. The maximum likelihood solution to the state of z_j is:

$$z_j = \begin{cases} 1, & \prod_{i \in ch(\beta_j^2)} p(c_i|\lambda_i) > \prod_{i \in ch(\beta_j^2)} q(c_i) \\ 0, & else. \end{cases} \tag{5.7}$$

The benefit of this coherent occlusion model can be observed by comparing this equation with Equation 4.3. The decision about the visibility of individual parts is now based on the joint appearance of all group elements. In this way contextual information is integrated into the decision process about the occlusion state of individual filters. In Figure 5.13 the effect of the coherent occlusion model can be observed from samples of the image model without occlusion (Figure 5.13(a)), with independent occlusion (Figure 5.13(b)) and with the proposed coherent occlusion model (Figure 5.13(c)). The appearance and the geometry of the object model is the same for all samples depicted in Figure 5.13. Compared to the independent occlusion model, the coherent model is able to generate locally coherent occlusions which are much more realistic.
It might be beneficial to have multiple occlusion states which model different occlusion configurations per group as proposed by [Ghiasi and Fowlkes, 2014]. We think, however, that due to the small size of the groups w.r.t. the size of the object, an approximation with binary state variables is sufficient.

5.4.4 Loss of Characteristic Object Information

The compositional representation also has one downside. Some characteristic details of the training image which are present in the original ABM representation, are not

preserved in the CABM representation. The reason for this loss of information can be found in the learning procedure of the CABM. Particularly, in the assumption that the training image must be encoded with the learned dictionary of ABMs in order to induce the structure of the CABM. However, not all characteristic details of the training image can be represented by the dictionary of ABMs. Especially, those patterns that are non-repetitive cannot be represented well. Figure 5.14 illustrates this loss of information based on two examples. In both examples, the left image depicts the appearance of the original ABM and the right one illustrates the appearance of a CABM. Elemental patterns which are not repetitive across orientation or translation are not represented well by the CABM, such as the trademark in Figure 5.14(a). However, also information about repetitive patterns is lost in the CABM feature space as can be observed from Figure 5.14(b). Especially, the high frequent information which is captured by the LoG filters with small variance is lost in the CABM representation. This is also a result of the training procedure. When ABMs are learned via shared matching pursuit, the algorithm chooses those filters which can jointly reconstruct all training patches well. Small misalignments in the training patches immediately result a loss of high frequent patterns during the learning stage. Therefore, high frequent patterns such as these dotted structures are unlikely to be represented in an ABM.

Conclusion. In summary, the effect of the compositional object representation is threefold. The part-based interpretation of the image provides highly useful information in terms of the presence and orientation of possible foreground patterns in the probe image. Furthermore, the hierarchical deformation model allows to increase the flexibility of the shape model while better preserving the characteristic structural information compared to the original ABM. Lastly, the compositional structure makes it possible to model locally coherent occlusion. Nevertheless, the efficiency in terms of representation is bought at the cost of a loss of some characteristic information about the object.

In the next section we will introduce CABMs with more than two layers which encode long range structural dependencies of an object. We will introduce how these models can be learned from a single image and discuss their application in the context of shoe print recognition.

5.5 Multi-layer Compositional Active Basis Models

Considering the benefits of the hierarchical abstraction in the two-layered CABM over the flat ABM, a natural question is: What benefits would a further increase of hierarchical abstraction imply? In this section we present an algorithm for learning the multi-layer compositional structure of a CABM. We discuss the benefits and limitations of such *Hierarchical Compositional Models* (HCMs) in the context of shoe print recognition and for general object recognition.

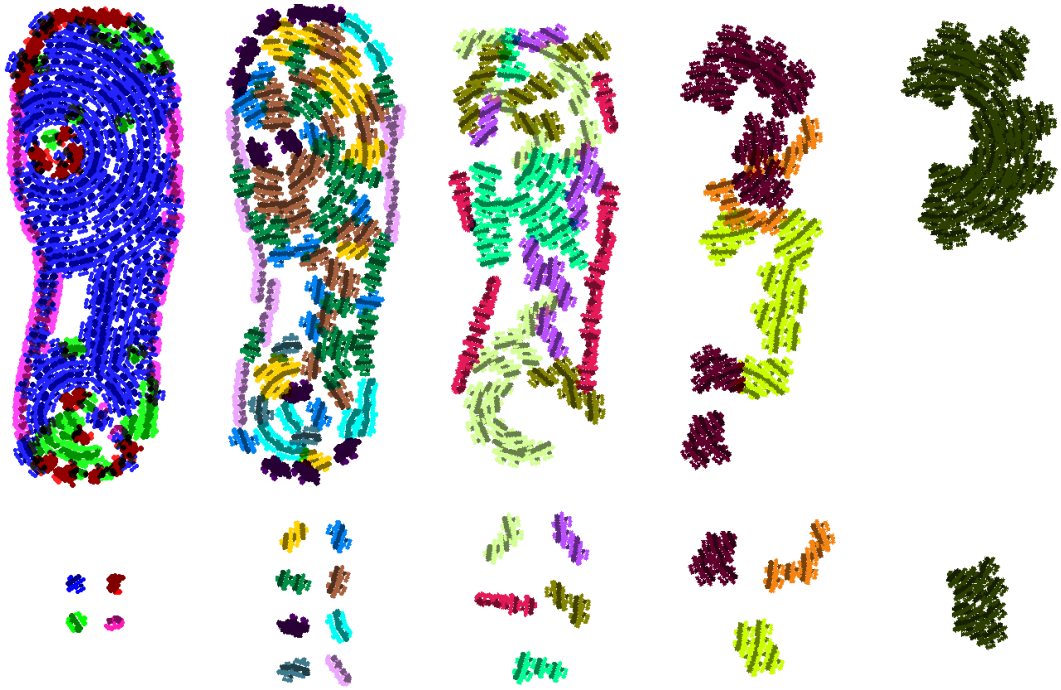
5.5.1 Related Work on Learning Hierarchical Compositional Models

HCMs with multiple layers of abstraction are known for their efficient representation and have successfully been applied in the context of object recognition e.g. by [Fidler and Leonardis, 2007; Zhu et al., 2008; Si and Zhu, 2013]. This principle of compositionality has been proposed in [Geman et al., 2002] and enables an efficient learning of HCMs. A common technique for higher-order compositional models is bottom-up compositional structure induction [Zhu et al., 2008; Fidler and Leonardis, 2007; Yuille, 2011]. The idea is to learn the hierarchy layer by layer in a bottom-up manner, with the constraint that the parts at each layer must be composed of the parts from the previous layer. In the following subsection, we will integrate our greedy EM-type algorithm in such a bottom-up structure induction process in order to learn multi-layer CABMs.

5.5.2 Learning a Multi-Layer CABM

We will now integrate our EM-type algorithm into a bottom-up structure induction process as proposed in our work on shoe print recognition [Kortylewski and Vetter, 2016]. Similar to the work by [Fidler and Leonardis, 2007] & [Zhu et al., 2008], we pursue a bottom-up compositional learning scheme, which learns a hierarchical dictionary of parts. These parts are then composed into a holistic object model in a top-down structure induction process.

Bottom-up learning. We start by learning the ABMs that encode the first layer of the hierarchy. We apply the greedy EM-type algorithm as explained in Section 5.3.3 to a gallery image. The learned dictionary of ABMs $D^1 = \{O_n^1 | n = 1, \dots, N_1\}$ is illustrated in Figure 5.15(a) together with an encoding of the gallery image. If we would connect the detected ABMs to a common root node, this would resemble exactly the global CABM as introduced in Section 5.3.3. Instead, we will compose the ABMs of D^1 into local CABMs $D^2 = \{O_n^2 | n = 1, \dots, N_2\}$ with the same greedy EM-type algorithm. The difference is that the LoG basis dictionary will be replaced with the ABM dictionary D^1.For our experiment, we define that the elements in D^2 must be composed of two elements from D^1. We also define, that in the E-step of the algorithm the local CABMs can deform their shape according to $\delta\beta^2 = \{\delta X^2 = 2 \text{ pixel }, \delta\alpha = 0°\}$. In this way, we formulate that a whole class of local patterns are described by the same model. Since we learn from a single image, this ensures that the model finds data for the M-step during the E-step.
The learning process exploits compositional structure of the model, as it allows us to first learn the level-one models, before composing them into a level-two model. We repeated the compositional learning recursively layer by layer until no further compositions are found, thus generating dictionaries of CABMs at multiple levels $\{D^1, \dots, D^{N_L}\}$. Figure 5.15(b)-5.15(e) illustrates the results of this learning process. Note how the number of elements increases in the second layer and then decreases again in the following layers. This phenomenon that the number of parts of intermediate complexity is maximal in HCMs has similarly been observed in [Zhu et al., 2008; Fidler and Leonardis, 2007].



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

Figure 5.15: Result of the bottom-up compositional learning process. In every subfigure (a-e) the top image illustrates the encoding of the training image with the learned dictionary of models. Below this image, the elements of the dictionary are illustrated in their original size. (a) The dictionary D^1. It is learned with the greedy EM-type procedure as introduced in Section 5.3.3. (b) The dictionary D^2. Its elements are local CABMs which are composed of two elements from D^1. (c-d) The dictionaries D^{3-5}. Each dictionary element is composed of elements from the dictionary of the level below. The elements are CABMs with an increasing number of compositional layers.

Top-down structure induction. In order to obtain a holistic object model, we must build a global object model from the elements of the learned dictionaries $D^1, \ldots, D^L$. We achieve this with a top-down structure induction process 5.16. It starts by detecting the elements of the highest layer D^L in the training image 5.16(a). We continue by detecting those parts of the dictionary D^{L-1} which do not overlap with the already detected parts 5.16(b). These are also connected to the root node. We proceed in a top-down manner layer by layer until every part of the training image is encoded (Figure 5.16(c)-5.16(e)). At this point, we have learned a holistic multi-layer Compositional Active Basis Model $O^L(\Theta)$ from the training image. The number of layers L as well as the number of parts for each individual layer $N_{1,\ldots,L}$ have been inferred automatically. The proposed learning scheme is related to the ones presented by [Zhu et al., 2008] and [Fidler and Leonardis,

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

Figure 5.16: Illustration of the top-down detection process. The elements of the learned dictionaries of CABMs $D^1, \ldots, D^L$ (Figure 5.15) are detected in the training image. Thereby a top-down strategy is applied. (a) First, the elements of D^L are detected in the training image. (b) Detection result of the elements of D^{L-1} which do not overlap with the already detected ones from (a). This process is repeated with the dictionaries (c) D^{L-2}, (d) D^{L-3} and (e) D^{L-4}. A holistic multi-layer CABM is build by connecting the detected local CABMs to a common root node.

2007]. However, ours is more intuitive and less heuristic than the other algorithms. Another advantage compared to the work of [Zhu et al., 2008] is that we infer the number of dictionary elements in each layer automatically. We have also demonstrated the generality of our proposed learning framework by applying it in the context of learning from natural images [Kortylewski et al., 2017].

In Figure 5.17 we illustrate the structure of an occlusion-aware multi-layer CABM. Due to the top-down structure induction process, the root node is connected to subtrees of different depth. We will analyze the effect of this in the following discussion on the hierarchical deformation and occlusion models.

5.5.3 Hierarchical Deformation

In the multi-layer model structure large deformations can be decomposed hierarchically. In this subsection, we study the benefits of this property. We will restrict our discussion to changes in location only. However, the insights we provide are valid for rotational changes as well. Let us assume a given 6-layered CABM as illustrated in Figure 5.16 (the layer of the individual filters plus 5 compositional layers). We want to build the flexibility into the model to move a basis filter by a distance of 20 pixels from its mean position. In a two-layered CABM, this could be achieved by completely accounting

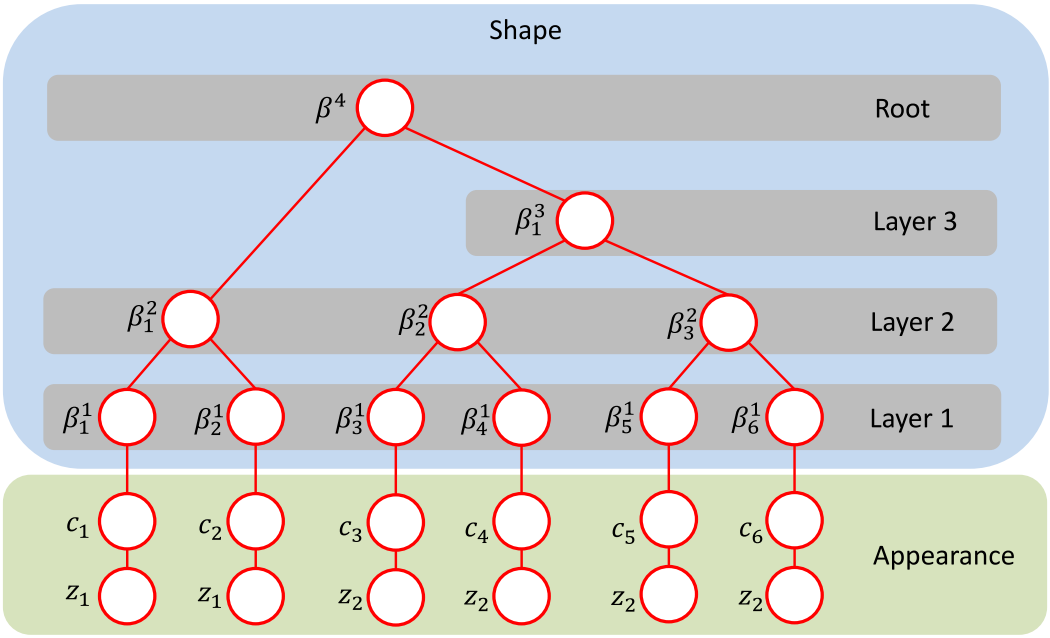


Figure 5.17: The dependence structure between random variables in an occlusion-aware multi-layer CABM. Compared to the CABM representation as depicted in Figure 5.2, the shape model is represented with multiple compositional layers. Due to the top-down structure induction process (Figure 5.16) nodes from lower layers can be connected directly to the root node. The multi-layer abstraction also makes it possible to share occlusion variables among large groups of appearance variables. In this way, occlusion coherence can be enforced within the bottom-up inference process.

for this movement in the compositional layer. Thus, each part in layer two could move independently by a distance of 20 pixels. A sample from the resulting model is illustrated in Figure 5.18(b). Similar to the observation we made in Section 5.1, we observe that large independent part movements distort the geometry of the object severely.

However, in order to introduce more dependence between the filters, we can also account for this deformation on different layers of the hierarchy. For example, a more reasonable strategy would be to increase the amount of movement linearly among the layers. Thus, performing small movements independently at low layers, while restricting large movements to be performed in higher layers by larger groups of filters. In this way, the hierarchical decomposition has the effect of constraining the space of possible deformations. Figure 5.18(d) illustrates a sample from the object model with a linear decomposition strategy.

In Figure 5.19 we illustrate in a bar plot different strategies of decomposing a deformation of 20 pixels among the different layers of the hierarchy. The independent deformation model is visualized in blue. The complete movement is performed in the

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

Figure 5.18: Samples from shape models which follow different strategies of decomposing a total deformation of 20 pixels per basis element across the layers in the hierarchy (Fig 5.19). The appearance is fixed in all samples. (a) No deformation. (b) Independent deformation at the second layer of the hierarchy. The geometry of the object is distorted severely. (c) Shows a uniform decomposition and (d) a linear decomposition of the deformation across all layers. The result in (c) and (d) look very similar. The objects structure is much better preserved compared to (b), although the theoretical total movement is the same.

first compositional layer (Layer 2). The linear decomposition strategy is illustrated by the red bars. Other strategies, such as a uniform decomposition (illustrated in black) might also be reasonable. It would certainly be interesting to study the influence of the different decomposition strategies on the properties of deformations which they imply. We however restrict our model to the linear decomposition strategy throughout the remainder of this work.

One issue which has to be resolved is the fact that some subtrees, which are connected to the root node have a smaller depth than others (see illustration in Figure 5.17). We account for this issue by defining, that each subtree must model the same total amount of flexibility. Hence, according to Figure 5.17 $p(\beta_1^2|\beta^4)$ would have to permit larger movements than $p(\beta_2^2|\beta_1^3)$, although both nodes come from the same layer. In this way, we ensure that each basis filter can move by the same distance in the model.

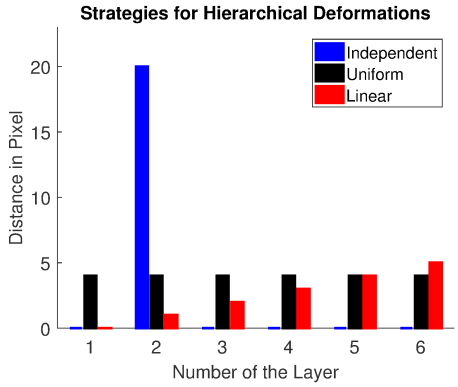
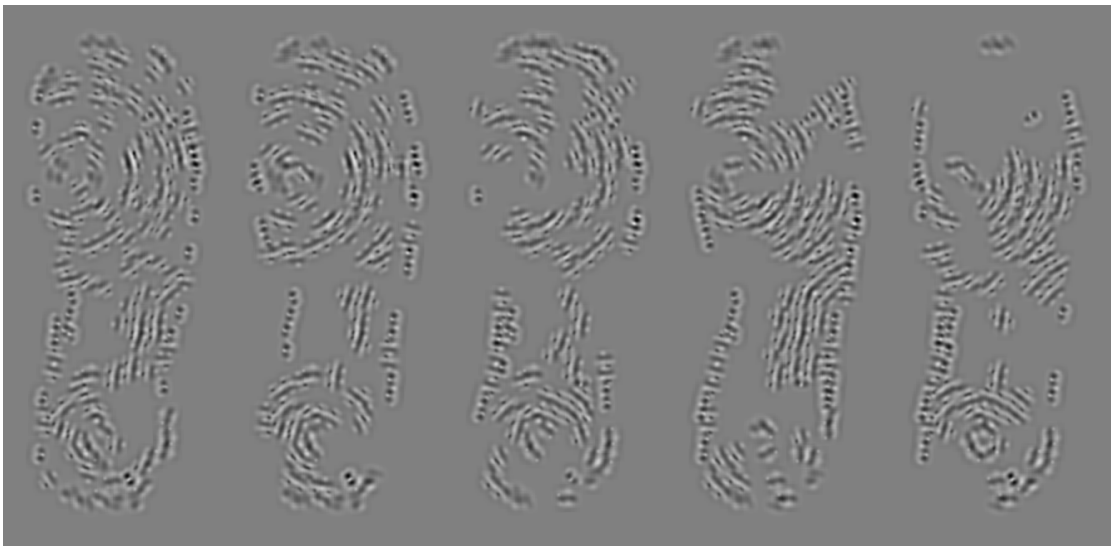


Figure 5.19: Three different strategies for decomposing a deformation among the layers of a hierarchical shape model (6-layered model). Each strategy allows the basis filters in a CABM to move a total distance of 20 pixels from their mean position (sum over). The blue bar illustrates a model which accounts for this deformation completely in the second layer of the hierarchy. The uniform strategy (black bars) decomposes the deformation into equal movements of at each layer. A linearly increasing decomposition across the layers is illustrated by the red bars. This strategy permits only small independent movements of the basis elements, whereas large movements must be performed in higher layers.

Relation to Global Shape Models. In this section, the benefits of multi-layer CABMs over traditional global shape models became evident. Similar to global shape models, CABMs can represent complex deformations at different scales. However, in contrast to global models, the tree-like independence structure of the shape model in CABMs can be optimized globally very efficiently. Thus omitting any form of initialization or bottom-up guidance during optimization.

5.5.4 Hierarchical Occlusion

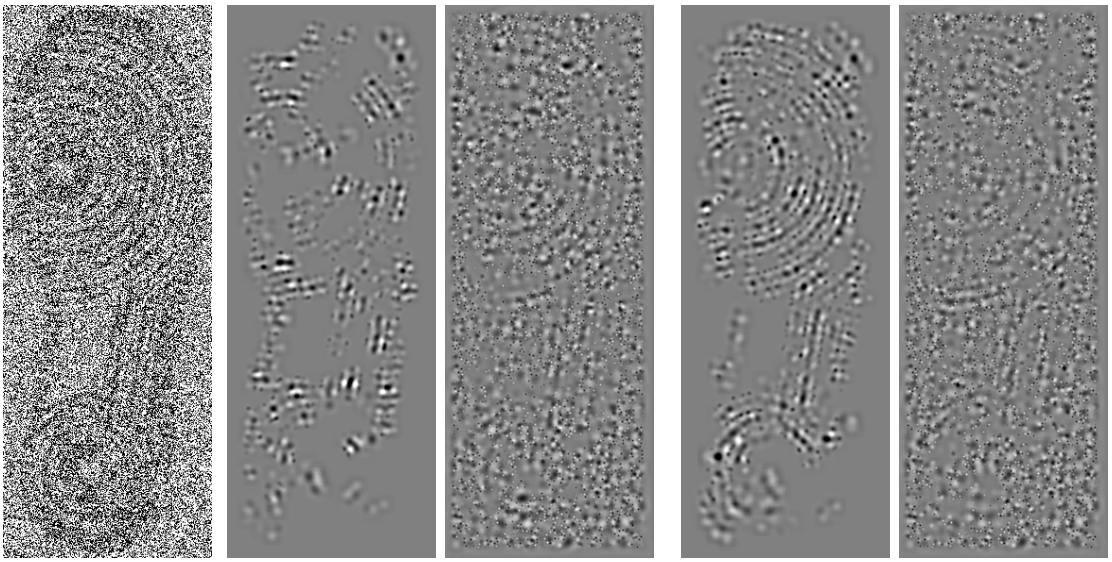
We have discussed the ability to efficiently enforce locally coherent occlusions with CABMs in Section 5.4. Thereby, different appearance variables c_i which were part of the same sub-tree shared a common occlusion variable z. Now, due to the multi-layer structure, we can also share variables among higher-order subtrees. We have illustrated this property in Figure 5.17. The appearance variables $\{c_3, c_4, c_5, c_6\}$ share the same occlusion variable z_2 and thus will be visible or occluded jointly. However, another possibility would be to share the occlusion variable of all sub-trees in the second layer. Hence, $\{c_1, c_2\}$, $\{c_3, c_4\}$ & $\{c_5, c_6\}$ each would share one variable z_1, z_2 & z_3. In Figure 5.20 we illustrate the effect of sharing the occlusion variables among sub-trees at different layers in the hierarchy. What can be observed, is that if subtrees of layer 5 (Figure 5.20(d)) or 6 (Figure 5.20(e)) are chosen, larger regions are coherently occluded. For lower layers such as Layer 2 (Figure 5.20(a)) and Layer 3 (Figure 5.20(b)) the occlusion is more fractured.



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

Figure 5.20: Influence of the layer at which occlusion variables are shared on the spatial coherency in the occlusion model. (a-e) Show samples from a multi-layer CABM with increasing layers at which occlusion variables are shared. The appearance and geometry of the object is fixed. (a) Sharing variables among subtrees of layer two. Occlusions in the sample are local. The synthesized shoe print is fractured. Increasing the number of the layer at which occlusion variables are shared to (b) three, (c) four (d) five and (e) six also increases the spatial coherency of the occlusions in the synthesized shoe prints.

Relation to other coherent occlusion models. Compared to models which enforce occlusion coherence with Markov random field (MRF) priors, our proposed methodology has two advantages: It can be computed optimally in a single bottom-up inference pass, whereas MRFs on the occlusion variables must be optimized iteratively. Additionally, long range dependencies are built directly into the model structure, whereas typically in MRF models long range correlation are encoded implicitly by the parametrization of the MRF. [Ghiasi and Fowlkes, 2014] also enforce coherency in the occlusion states of facial landmarks via a hierarchical dependency structure. They use a hand designed two-layered hierarchical model which limits the possible influence of contextual information on the decision about the occlusion states. However, an interesting feature of their approach is that higher order parts have multiple occlusion states which imply different occlusion patterns at the basis elements. In contrast, our approach only implements binary occlusion states that must be shared among all basis elements of a subtree. Therefore, we think an interesting future research direction would be to permit multiple occlusion states at higher order parts of a CABM.



(a) (b) (c)

Figure 5.21: Image interpretation with CABMs. The input image (a) is interpreted by a two-layered and a multi-layered occlusion-aware CABM. (b) The result obtained with the two-layered CABM as proposed in Section 5.3. The image is interpreted in terms of foreground (left) and background (right). (c) The result obtained with the multi-layered CABM as proposed in this Section 5.5. Also here the image is partitioned into foreground (left) and background (right). We can observe that the hierarchical occlusion model enforces a more coherent image interpretation.

5.5.5 Interpreting Images with CABMs

Given an image, we can optimize the model parameters of a CABM with the bottom-up inference procedure as proposed in [Dai et al., 2014]. Thereby, the convolution-based optimization as we discussed in Section 3.4.1 is applied in a bottom-up manner for each layer in the hierarchy. Based on the optimal global positioning of the model, the individual states of all parameters can be computed by traversing the tree-structure backwards from the root node down to the leafs, while keeping track of the states of the variables. The result of this optimization procedure is an interpretation of the image in terms of which parts of the image belong the foreground and background, the global position of the object, the state of the occlusion variables and the extent of deformation of each part. Inferring these hidden factors that have generated the image describes the ultimate goal of vision [Yuille and Kersten, 2006]. In general, the better these factors can be inferred, the better the quality of the object model.
Figure 5.21 illustrates the advantage of higher-order model based knowledge for the image interpretation process. We analyze the noisy image in Figure 5.21(a) with the two-layered CABM as introduced in Section 5.3 and the multi-layered CABM as presented in the previous Section 5.5. Each model provides an interpretation of the image

in terms of which parts belong to the foreground (Figure 5.21(b) & 5.21(c), left) and which to the background (Figure 5.21(b) & 5.21(c), right). From these interpretations, we can clearly observe that the multi-layer model is much better at extracting which parts of the image depict foreground and which background. Although, both models have been learned from the same data and have the same flexibility. The mechanism which makes this possible is the hierarchical occlusion model. In contrast to the more independent occlusion model of the two-layered CABM, the hierarchical occlusion model first gathers local evidences about the likelihood of parts and then takes a decision about the occlusion of large groups of parts based on their joint statistics. This is more robust since more contextual information is taken into account.

5.5.6 Limitation of the Tree-structured model

Throughout this chapter, we have pointed out the benefits of a hierarchical dependence structure between random variables in an object model. However, the tree-structured model also has one important limitation. Dependency between variables must not be spatially correlated. Parts which are spatially next to each other but do not belong to the same subtree of the hierarchy, can deform and occlude independently of each other. This is a fundamental limitation of tree-structured models in general, which is still to be resolved.

5.6 Conclusion

In this chapter, we have studied Compositional Active Basis Models in detail. In the beginning, we have worked out limitations of the LoG ABM in terms of a loss of structure under large deformations, a poor discriminative ability at the part level and the invariance to rotation at the part level. We demonstrated how to overcome these limitations by the introduction of a hierarchical dependence structure between the basis filters. In this context, we developed a greedy EM-type algorithm and demonstrated its ability to learn the hierarchical structure of a CABM. We showed that the CABM offers coherence in occlusion and deformation and an increased ability to discriminate foreground from background at the part-level, while preserving the beneficial properties of an ABM.
In the following section, we will extensively evaluate the CABM in the context of object recognition in cluttered images.

Chapter 6

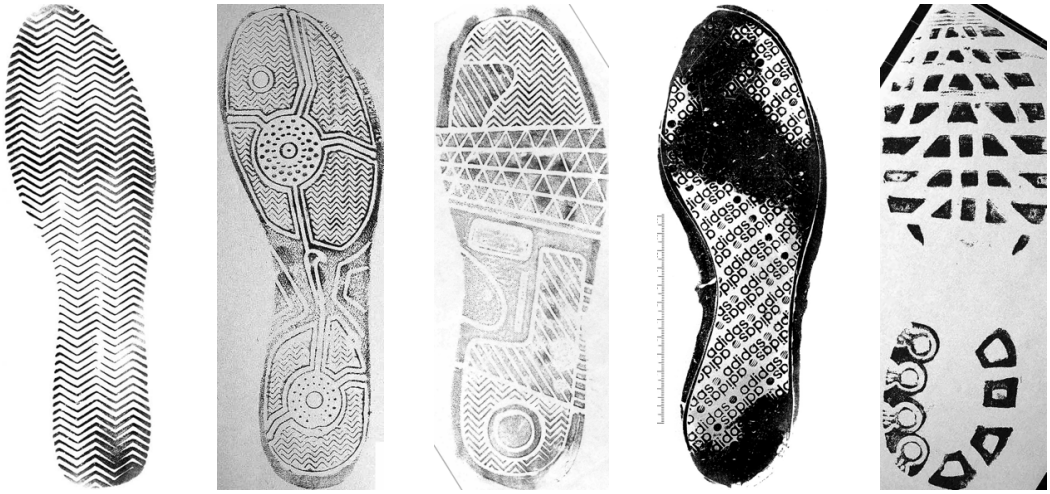
Shoe Print Recognition Experiments

> In the computer field, the moment of truth is a running program; all else is prophecy.
>
> Herbert Simon

In this chapter we apply our proposed modeling framework to the task of forensic shoe print recognition. We start by introducing the FID-300 database (Section 6.1). The database contains more than a thousand gallery images and 300 probe images which have been collected by forensic experts from crime scenes. We then continue to evaluate our model-based recognition approach (Section 6.2). In order to provide a baseline for our evaluation, we re-implement several well known shoe print recognition algorithms and test them against the database (Section 6.2.1). We shortly present how our modeling framework can be applied to object recognition in Section 6.2.2. In Section 6.2.3 we evaluate the original ABM and study the effect of our proposed independent occlusion model and the basis basis change on the recognition performance. The two-layered CABM is evaluated in Section 6.2.4. In particular, we study the effects of the coherent occlusion and the hierarchical deformation. Section 6.2.5 evaluates the multi-layered CABM and its properties under large deformations. We show qualitative retrieval results in Section 6.3.

6.1 The FID-300 Database

A major issue in any past work on shoe print recognition is the lack of a standardized evaluation benchmark. Most of the published algorithms have been evaluated on probe images which were synthetically generated from gallery images e.g [De Chazal et al., 2005; Gueham et al., 2008; Nibouche et al., 2009]. Neither the gallery images, nor the synthetic probe images have been made publicly available. Thus, a direct comparison of the reported recognition performances was not possible. [Dardi et al., 2009; Tang et al., 2011] have evaluated their approaches on real data. However, the authors have also not publicly released their data.



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

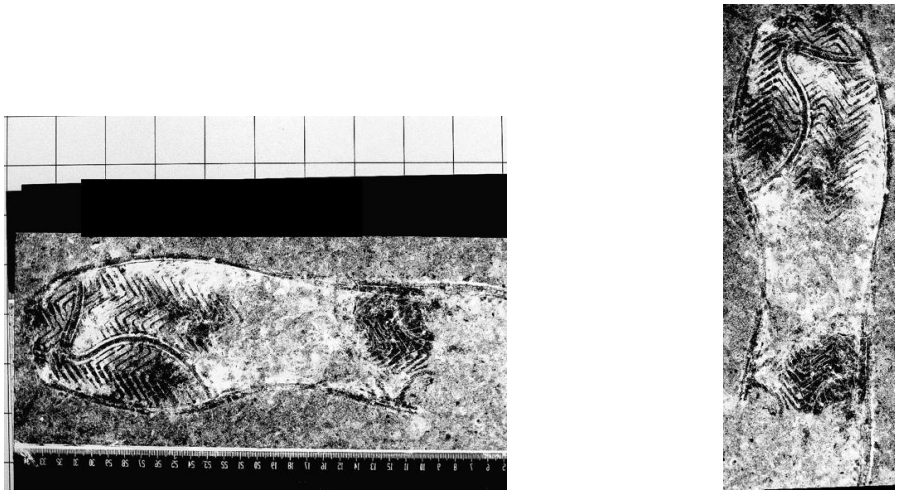
Figure 6.1: Diversity in the appearance of gallery images. (a) An ideal gallery image. The image appearance is nearly binary and contains only relevant information about the shoe print. Most gallery images are not ideal. They can either have gray backgrounds (b), or might be of low contrast (c). Additionally, some information might be lost due to wear (d). Some impressions also contain structures in the background which do not belong the shoe print (c, d & e).

Together with the German State Criminal Police Offices and the company Forensity AG we have collected a database of real data, including probe images and gallery impressions. The database is publicly available under http://fid.cs.unibas.ch/. In the following, we will in detail describe the characteristics of the database.

6.1.1 Gallery Images

The database contains 1175 gallery images (see examples in Figure 6.1). The impressions have been scaled to 20 pixel per centimeter. Those impressions which did not show a ruler were scaled to a height 586 pixels. This is the mean height of all references which had a ruler in the image. We aligned the images upright. If needed, we flipped them along the vertical axis in order to depict only right shoes.

In general, one would assume that the gallery images are mostly binary and depict solely shoe print information (Figure 6.1(a)). However, in practice this assumption is wrong. The collection process of gallery impressions is not standardized and therefore different techniques exist. Hence, some gallery images may have a darker background (Figure 6.1(b)), or may have a very low contrast (Figure 6.1(c)). Often, shoes have been worn before taking the gallery impression. This results in a loss of information in some regions of the print (Figure 6.1(d)). Furthermore, additional structures may be in the image which do not belong to the impression itself, such as rulers (Figure 6.1(d)) or



(a) (b)

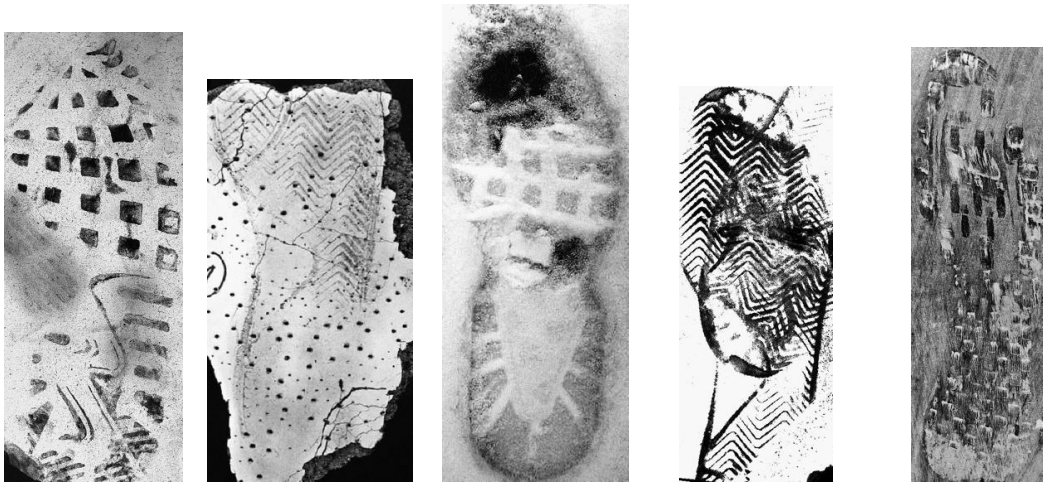
Figure 6.2: Illustration of the manual preprocessing of a probe image. (a) The original probe. (b) The result of the preprocessing step. The probe has been cropped and rotated to be roughly upright. In addition, we normalize it's scale to 20 pixels per centimeter.

other strong edges (Figure 6.1(c) & 6.1(e)). The mentioned changes in the appearance of the gallery impressions, render the shoe print recognition task even more complex than we initially described in the beginning of this thesis (Section 1.2).

6.1.2 Probe Images

The database contains 300 probe images. All of them depict a ruler and therefore could also be scaled to 20 pixels per centimeter. The original images often depict some background from the laboratory in which they were gathered (see the squares in Figure 6.2(a)), in order to reduce the computational load and to ease the recognition process, we preprocessed the probe images. The preprocessing involved a cropping of the shoe print information, followed by a manual rotation of the cropped image such that it is oriented roughly upright (Figure 6.2(b)). Both, the original and the preprocessed images are publicly available. Most of the probe images were lifted with a gel foil from the ground and were subsequently digitized with a scanner (Figure 6.3(a)). Some were digitized by photographing them directly from the surface on which they were placed (Figure 6.3(b)). About 20 of the impressions are photos of three dimensional impressions in snow (Figure 6.3(c)).

Additional challenges. In Chapter 1, we have described the main challenges of shoeprint recognition to be: limited training data, appearance change and deformation, clutter and partial occlusion. In rare cases, probe images can also be smeared (Figure 6.3(e)), or the pattern of the print might change, due to a second print which is overlayed (Figure 6.3(d)). Another source of variation between a gallery and probes images is scale change. Although, we have normalized the resolutions of the images, the



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

Figure 6.3: Diversity in the appearance of the probe images. The majority of the probe images in our database has been digitized with a scanner (a). About 10% have been photographed (b & c). In addition to the complexities discussed in Chapter 1, some probe images depict double prints (d) or can be smeared significantly (e).

gallery image and the probe image might have a different size.

In the following section, we will evaluate different methods in terms of recognition performance on this dataset. All of the experiments will be performed on the preprocessed data.

6.2 Shoe Print Recognition

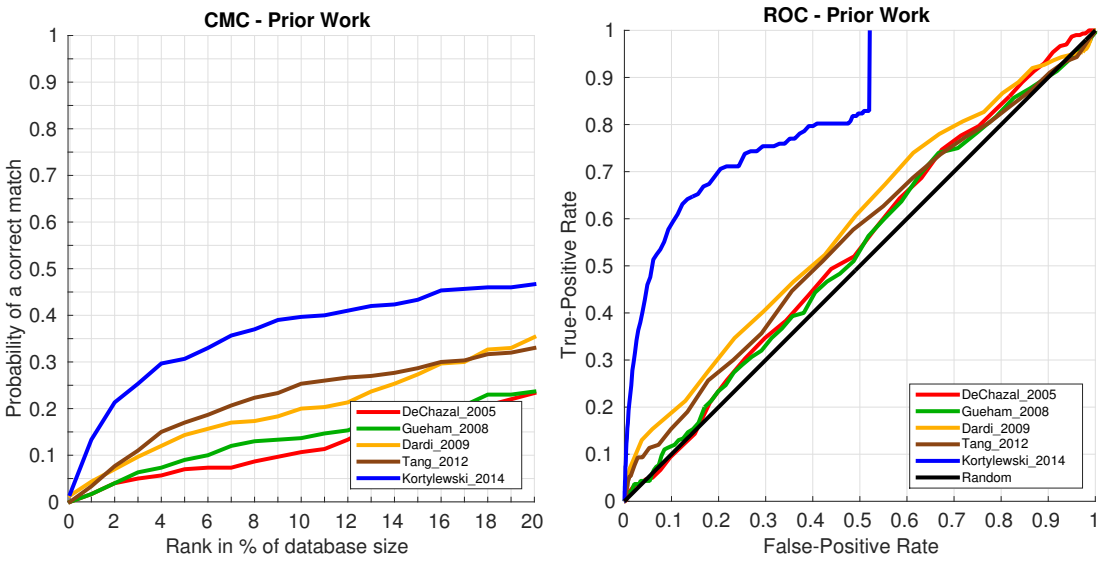
In object recognition, two different setups of recognition are common. One is object retrieval, also known as relative object recognition, and one is classification, which is also known as absolute object recognition. In an object retrieval setup, we assume that a complete database of objects exists. Thus, any object which can be depicted in the probe image is also present in the gallery. Hence, for recognition it is sufficient to rank the database elements relative to each other according to some similarity measure. The gallery element with the highest rank is supposed to be the correct object. The results in this setup are reported with the Cumulative Match Characteristic (CMC). It measures the likelihood of finding the correct match at a certain rank. Typically, a point on a CMC curve is referred to as recall@X. For example a 90% chance of finding the correct gallery object in the first 5% of the ranked list, is referred to as 90% recall@5%. Throughout this section we will also use this notation. The y-intercept of a CMC curve marks the rank-1 performance.
In a classification setup, object recognition is performed by defining an absolute threshold on the similarity between two objects. A gallery element for which the similarity measure

exceeds this threshold, is supposed to be the correct object. The results in this setup are reported with the Receiver Operating Characteristic (ROC). It measures the relation between true matches and false matches for a certain similarity threshold. The interested reader is referred to the work by [DeCann and Ross, 2013] for a discussion on the relation between the CMC and ROC measures.
In summary, CMC is a rank-based measure which is used in a relative recognition setup. ROC is used in an absolute setup and reflects an aggregate statistics [Ross, 2017]. In order to get a complete impression about the performance of a recognition system it is recommend to report both of these measures, which we will do in the following. However, the CMC performance might be of higher relevance in the forensic practice because the ultimate decision about an identification is performed by a human expert. In the experiments with ABMs and CABMs we therefore first focus on interpreting the CMC performance and then mention only additional insights from the ROC results.

6.2.1 Benchmarking of Prior Work

We have reimplemented four of the most well known algorithms in the field, in order to relate our results to the previous work on shoe print recognition. We have evaluated these approaches on the FID-300 database. The results are illustrated in Figure 6.4.

Approaches that measure the similarity between images with global invariants such as the Fourier Transformation [De Chazal et al., 2005] (red curve) and the Fourier-Mellin Transformation [Gueham et al., 2008] (green curve) show a poor recognition performance. In fact, the results are close to random. We have discussed in Chapter 2 that such global representations, fail to account for clutter in the background as well as for partial occlusion. Furthermore, these representations also do not account for local deformations of the target object.
[Dardi et al., 2009] (orange curve) divide the probe and the reference image into a fixed grid and compute local appearance features for each grid patch. These features are then compared under the constraint that the global geometry of the grid is exactly preserved. This approach performs significantly better than the global invariant approaches in terms of CMC. The ROC curve shows a short rise in the true-positive rate (TPR) for a false positive rate (FPR) between 0 and 0.05. This implies that for large similarity scores, the recognition is significantly better than a random performance. However, after this short rise, the recognition becomes random. A major drawback of the approach is that the probe and the gallery image are assumed to be roughly aligned. Furthermore, it does not account for missing features. The approach presented by [Tang et al., 2011] (brown curve) is the most similar to our approach. A gallery shoe print is represented with geometric primitives (lines, circles and ellipses). Their spatial configuration is captured in a graph. At runtime, the primitives are detected in the probe image with a hough transformation. The spatial relation between the primitives in the probe image is then compared to the stored graph of each gallery image. The approach has multiple limitations. In the probe images, the primitives are mostly partial. Depending on the detection threshold, either only very few primitives are detected or primitives will be



(a) (b)

Figure 6.4: Shoe print recognition results of prior work (see text for details). The performance is visualized in terms of CMC (a) and ROC curves (b). *Comment:* The blue curve shows the result of our previous work on classification based on periodic patterns [Kortylewski et al., 2014]. The algorithm detected a periodic pattern in 628 out of 1175 gallery images. Hence, a similarity measure can be computed for these. For the other images, a similarity value of 0 is fixed. This induces the steep jump in the ROC for an FPR at $\frac{628}{1175} \approx .53$.

found all over the probe image. From our experience, finding a global threshold for all images is difficult. Furthermore, the applied similarity measure is highly sensitive to partial occlusion. The recognition performance is very similar to the approach proposed by [Dardi et al., 2009].

Our previous work on invariant representations of periodic patterns [Kortylewski et al., 2014] (blue curve) performs significantly better than the other approaches. We have developed an interest point detector which extracts the symmetry properties of periodic patterns based on local autocorrelation. The extracted properties include the direction and frequency of the periodicity. Based on these, we define a local window from which a Fourier descriptor is computed. The extracted periodicity direction makes it possible to normalize for rotations in the pattern. The recognition based on the extracted features proves to be highly robust. This can be observed from the ROC curve which is significantly better compared to all other previous works. This means that if a periodic pattern is detected in the probe and in the gallery image, the computed similarity measure is highly discriminative. The algorithm detected a periodic pattern in 628 out of 1175 gallery images. Hence, a similarity measure can be computed for these. For the other images, a similarity value of 0 is fixed. This induces the steep jump in the ROC

for an FPR at $\frac{628}{1175} \approx .53$. For some probes, the algorithm detects a periodic pattern, where in fact in the corresponding gallery image no periodicity has been detected. We can observe from the maximal TPR before the jump that this is the case for 17% of all probe images. The CMC curve is about 15% above the approach proposed by [Tang et al., 2011]. Despite the excellent ROC curve, the CMC is nevertheless comparatively low. This highlights the major drawback of this method. Periodic patterns have only been detected in 179 out of the 300 probe images. the remaining 131 probe images could not be matched against the database. In this case, we assigned a rank of 1175.

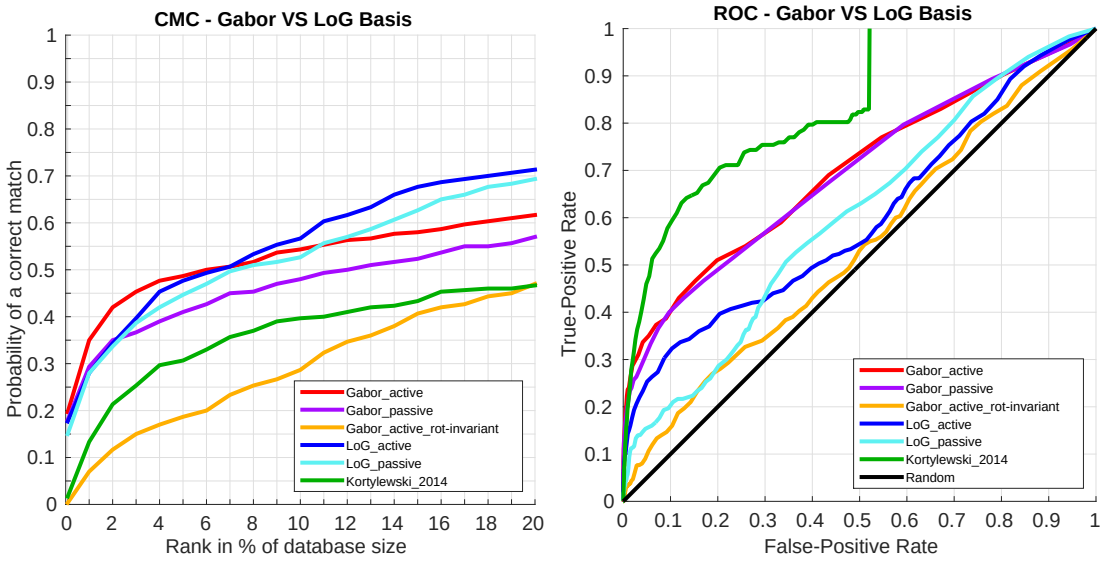
6.2.2 Recognition Setup

In order to perform shoe print recognition, we learn a statistical object model $p(\Theta|O_k)$ for each of the 1175 gallery images $\{I_k|k = 1, \ldots, 1175\}$. Given a probe image I', we infer the optimal parameters Θ^* in order to measure the maximal likelihood ratio between the posterior of the object model $p(\Theta|I', O_k)$ and the background model $q(I')$ (Equation 3.8). This ratio measures the similarity between the probe image and the gallery image $S(I_k, I')$ (Equation 4.1). Based on this similarity, we compute the performance in terms of CMC and ROC.
In our experiments, we do not account for scale transformations explicitly, for the benefit of computational efficiency. We therefore assume that large scale changes are not present in the data, while small scale changes can be, to a certain extent, accounted for by local deformations.
Throughout our experiments, we fix the value to $\lambda = 5$ for LoG appearance models (Figure 4.11(d)) and $\lambda = 3$ for Gabor appearance models (Figure 3.4(d)). This setting induces a foreground likelihood which favors strong feature responses and thus can act as competitor to the background model during inference.
We set the parameter of the Bernoulli distribution in the occlusion model to $\rho = 0.5$. Thus, we assume that every part is equally likely to be visible or occluded. In a typical object recognition application the likelihood of being occluded might be lower. However, in shoe print recognition the patterns in the probe image are very often strongly occluded. Therefore, we think this assumption is justified.

6.2.3 Gabor ABM VS LoG ABM

In this section, we compare the recognition performance of the standard ABM. We test the influence of the LoG basis as well as of the independent occlusion model. In the ABM framework, we can also reduce the effect of missing parts by choosing a smaller value for the parameter λ in the foreground appearance model (see Figure 3.4 for an effect of this parameter on the appearance model). In this way, the cost of explaining the background will be reduced. Hence, the model might still be able to recover the correct image interpretation. We refer to this approach as "passive" occlusion model, whereas we refer to our proposed approach as "active" occlusion.

For the LoG ABM the parameters were set to $\lambda = 2$ and $\lambda = 5$ (Figure 4.11(c) & 4.11(d)). If not otherwise specified, the ABMs are parametrized with $\delta\beta = \{\delta X = 2$ pixels $, \delta\alpha = 0°\}$. As we will observe in the experiments, permitting a rotational perturbation in the ABM reduces its discriminative ability.



(a) (b)

Figure 6.5: Evaluation of the standard ABM with different bases and different occlusion mechanisms: a weak background model ("passive") and an independent occlusion model ("active"). The performance is visualized in terms of CMC (a) and ROC curves (b).

CMC: All of the ABM methods perform significantly better compared to any of the evaluated previous works in terms of CMC score (we plot our work on periodic patterns [Kortylewski et al., 2014] in green as a reference). Importantly, the rank-1 performance has raised from practically zero to between $15 - 20\%$.

Occlusion: The active occlusion mechanism generally improves the recognition performance. This highlights the benefit of having a strong appearance based classification already at the filter level. Such a strong classification is however only possible due to the proposed occlusion mechanism (see discussion in Section 4.2).
For the LoG ABMs the occlusion mechanism has only little effect. This can be attributed to the background model of the LoG filters, which is too weak to be able to explain structured background well (see discussion in Section 5.1). Thus the log-likelihood ratio between foreground and background model tends to be positive most of the time. This in turn renders the occlusion model superfluous.
The active Gabor model (red curve) performs best until the recall@7% and is then outperformed by the active LoG model (dark blue curve). However, in practice a good performance in the early ranks is important. Thus, we can state that the active Gabor

ABM is more beneficial compared to the the LoG ABMs.

Orientation of parts: A limitation of the LoG basis is that it has no orientation. However, from the orange CMC curve we can observe that this information is important for the models recognition performance. In this experiment, we tested the active Gabor ABM with a perturbation of $\delta\beta = \{\delta X = 2 \text{ pixels }, \delta\alpha = 180°\}$ (orange curve). Thus, effectively making the Gabor filters rotational invariant. The effect on the CMC curve is significant. The performance decreases by 20% at rank-1 and about 30% at recall@5%. Compared to the LoG ABMs, the invariant Gabor ABM performs significantly worse. This difference in performance can be attributed to the higher specificity of the LoG ABM, which encodes more information about the gallery image.

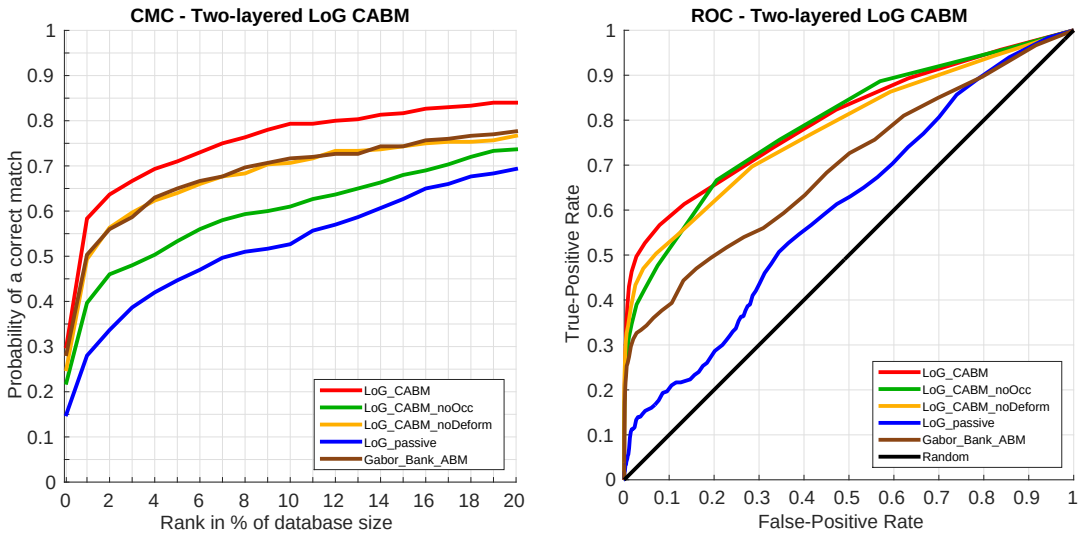
ROC: The ROC curve reveals additional properties about the applied models. In terms of classification accuracy, our work on invariant description of periodic patterns is the most accurate. This is expected, since it focuses on representing a particular class of patterns which can be reliably detected. Those patterns which are classified as non-periodic are not matched and thus also are not reflected in the plot. Furthermore, both LoG ABMs perform significantly worse than the Gabor ABMs in terms of absolute recognition. This, again highlights the importance of a part‘s orientation for the discriminative ability of the model. Interestingly, the passive and the active Gabor ABMs have a similar ROC curve, despite having quite different CMC curves. This shows that absolute and relative recognition reflect different properties of a recognition system.

Conclusion: In summary, we have observed the role of the occlusion model in terms of permitting a stronger discrimination between foreground and background at the part level. This in turn improves the overall discriminative ability of the model. In a standard setting, the Gabor ABM outperforms the LoG ABM. However, when constraining Gabor filters to be rotation invariant, the LoG basis outperforms the Gabor basis significantly. This highlights that orientation at the part level is highly beneficial for the discriminative ability of an ABM.

6.2.4 Two-layered LoG CABM

In this section, we evaluate the recognition performance of the two-layered LoG CABM as presented in Section 5.3. We analyze the effect of deactivating the occlusion model as well as the deformation model on the recognition performance (Figure 6.6). In addition, we compare the results to a Gabor ABM which was learned from an extended bank of Gabor filters (Figure 6.7). The two-layered CABMs are parametrized with $\delta\beta^1 = \{\delta X^1 = 1 \text{ pixel }, \delta\alpha^1 = 0°\}$ an $\delta\beta^2 = \{\delta X^2 = 3 \text{ pixel }, \delta\alpha^2 = 0°\}$. We do not permit rotational perturbation as this is an important information for the discrimination process (see previous Section 6.2.3).

CMC. The two-layered CABM (red curve) performs significantly better than the ABMs in the previous section. The rank-1 performance increases by about 10%. The recall@1% increases by 30% to a total of 58%. Overall, this is a considerable increase in recognition



(a) (b)

Figure 6.6: Evaluation of the two-layered CABM. We evaluate the influence of the hierarchical abstraction, the coherent occlusion and the hierarchical deformation on the recognition performance. Additionally, we test a Gabor ABM which was learned from an extensive Gabor bank (Figure 6.7). We show a passive LoG ABM as reference. The performance is visualized in terms of CMC (a) and ROC curves (b).

performance, which can be attributed to the newly available orientation at the part level, the coherent occlusion model and the hierarchical deformation.

Deactivating Deformation: Deactivating the deformation model decreases the recognition performance by about 10% starting from recall@1% (orange curve). This indicates, that the patterns in the probe images are subject to small deformation and that the invariance to these is beneficial for the recognition process.

Deactivating Occlusion: The coherent occlusion model has an important effect on the performance. This can be observed from the green performance curve, which shows the recognition result of the same CABM with the occlusion model turned off.

Comparison to passive Gabor ABM: The CABM without occlusion also performs better than the passive Gabor ABM in Figure 6.5(a), although both have access to rotational information at the part level. We suppose that this performance increase is induced by the better conservation of the characteristic object information in the LoG model (see discussion in Section 4.3). However, it is difficult to compare these models as the balance between the foreground and background model is different.

Comparison to active Gabor bank ABM: In order to confirm our hypothesis about the benefit of specific representations, we learn a Gabor ABM based on a whole bank of Gabor filters with different frequency (Figure 6.7). During basis decomposition, the learning algorithm can thus reconstruct the details of the target object much better compared the basis with fixed frequency.

We apply the Gabor bank ABM with occlusion model (brown curve) in order to

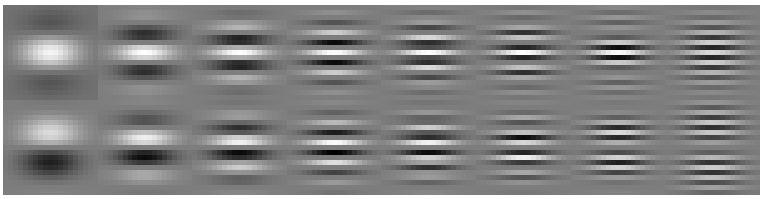


Figure 6.7: An extended bank of even (top row) and odd (bottom row) Gabor filters with a fixed scale $\sigma = 4$ and different frequencies $\omega = \{t\frac{\pi}{8} | t = 1, \ldots, 8\}$.

compare it to the two-layered LoG CABM (red curve). The tuning of the representation increases the recognition performance considerably compared to the active Gabor ABM with the original Gabor basis (see performance in Figure 6.5(a)). Hence, our hypothesis is confirmed. The reason for this difference must be further investigated. We suppose, that the invariance of the Gabor feature transformation (Section 4.3) cause of this effect.

Comparison to passive LoG ABM: Compared to the passive LoG ABM (blue curve), the recognition performance of the CABM without occlusion is constantly higher. We suppose that this is an effect of the additional flexibility due to the hierarchical deformation.

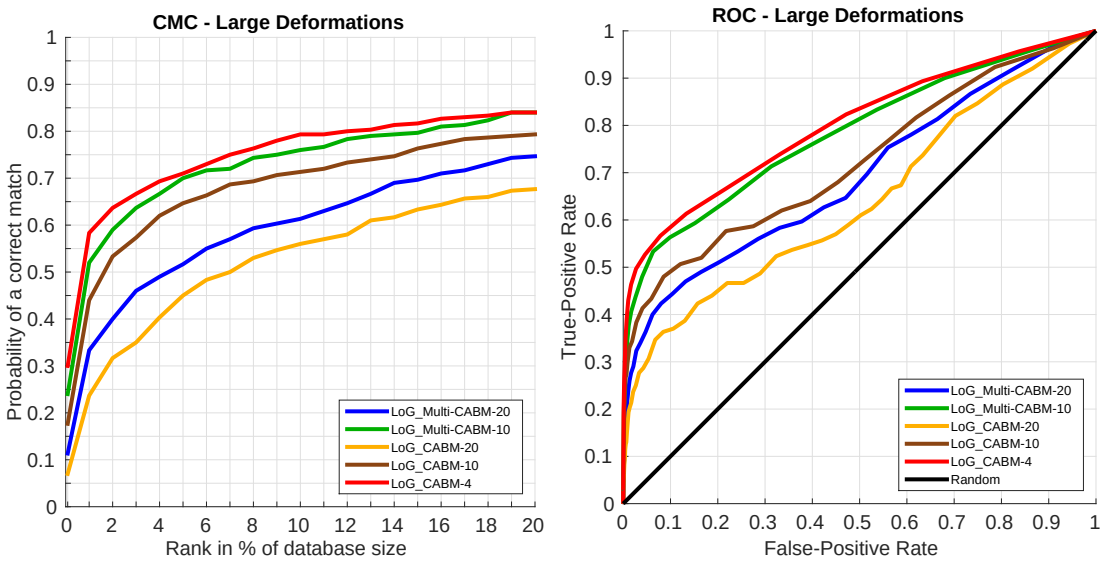
ROC. In terms of absolute classification performance, the two-layered LoG CABM (red curve) performs best. Turning off the occlusion model (green curve), results in a performance drop for FPRs between $0 - 0.2$. Interestingly, compared to the Gabor-bank ABM (brown curve) the green curve is lower in terms of CMC performance, however the ROC performance is significantly higher. Thus, the posterior of the two-layered LoG CABM is a much better indicator for the class membership. This observation is also confirmed by the ROC curve of the two-layered LoG CABM without deformation (orange curve). We would also attribute this to the invariance of the Gabor features which prevent a comparison of the object details.

Conclusion. In summary, we have observed that the hierarchical abstraction of the two-layered LoG CABM greatly improves the recognition performance compared to the ABMs in the previous layer. We have shown, that this can be attributed to the newly available orientation at the part level, the coherent occlusion model and the hierarchical deformation.

6.2.5 Large Deformations with CABMs

In the introduction of this thesis we have discussed the challenge of large-non rigid deformations in the probe images (Figure 1.3(d)). In this section, we increase the flexibility of the model, in order to be able to represent such large deformations. We study the effect of this additional flexibility on two-layered and multi-layered LoG CABMs. Since we only work with LoG CABMs in this section, we drop this label in the following discussion. For the multi-layered model, we apply a linear decomposition of the deformations

across the CABM layers (see Section 5.5). In the two-layered model, the flexibility will be modeled in the second layer. In this way, we allow the most possible dependence between individual basis filters during deformations of the two-layered model. As a reference, we depict the result of the two-layered CABM as presented in the previous section (total flexibility of 4 pixels per basis filter).



(a) (b)

Figure 6.8: Evaluation of CABMs with increased flexibility. We compare the two-layered with the multi-layered LoG CABM. Both models have the same theoretical flexibility. Two model setups are tested: A total flexibility of 10 pixels and 20 pixels. As a reference, we show the result of the two-layered CABM as presented in the previous section (total flexibility of 4 pixels). The performance is visualized in terms of CMC (a) and ROC curves (b).

CMC. The performance of a two-layered CABM with a flexibility of 10 pixels (brown curve) is significantly lower compared to the two-layered CABM from the previous section (red curve). At this point, a limitation of the model becomes apparent (see discussion in Section 5.2). The uniform distribution in the shape model does not penalize large deformations. Therefore, the large flexibility induces an invariance in the model to a large number of patterns. This in turn reduces its discriminative ability.
Increasing the flexibility of the two-layered model to 20 pixels, results in a further significant performance drop (orange curve). Again, the invariance class is increased, which in turn reduces the discriminative ability of the model. The multi-layered CABM with a total flexibility of 10 pixels shows a performance drop which is only about half of the magnitude as in the two-layered CABM (green curve). In the multi-layered model, the perturbation of parts is still modeled with a uniform distribution, however the hierarchical decomposition of the deformation enforces group-wise movement. Thus, the

hierarchical dependence structure introduces additional constraints between the parts. We have observed this property in the samples from the deformation model (Figure 5.18). It also excludes extreme deformations from the space of possible deformations, which are possible in the two-layered model (see discussion in Section 5.5). These additional constraints lead to the improved recognition performance. For a total flexibility of 20 pixels, the recognition performance of the multi-layered CABM remains superior to the two-layered CABM by about 5% (blue curve). Nevertheless, the performance is also decreased significantly compared to the model with less flexibility.

ROC. Overall, our observations in terms of CMC performance are confirmed by the ROC performance. The multi-layer CABM is superior to the two-layered CABM in terms of recognition performance with increased flexibility. Interestingly, the two-layered CABM with flexibility of 10 pixels (brown curve), is only slightly better than the multi-layered CABM with 20 pixels flexibility (blue curve). In terms of CMC performance, the difference between these models is more distinct. Furthermore, the multi-layered CABM with 10 pixels flexibility (green curve) performs just slightly worse than the two-layered CABM with a total flexibility of 4 pixels (red curve). These observations confirm the constraining property of the hierarchical dependence structure.

Conclusion. In summary, the spatial configuration of parts is an important information for the discrimination of patterns. Increasing the flexibility of the CABM results in decrease in recognition performance. The reason for this decrease is the models property that any deformation of the shape model is equally likely. The hierarchical decomposition of a deformation has a positive effect on the recognition performance under large deformations, because it introduces additional constraints compared to the two-layered CABM.

6.3 Qualitative Retrieval Results

In this section, we discuss two retrieval results that illustrate the quality of of the two-layered LoG CABM at the task of forensic shoe print recognition. We set the parameters of the model as in the experiments in Section 6.2.4. Figure 6.9 depicts two rows of images. Each row shows first the target probe image, followed by the six best results of the ranked list of gallery images. Below each gallery image, we show its similarity to the probe image in terms of the maximal log-likelihood ratio between the object model and the background model. The score of the correct gallery image is colored in red. The similarity was computed with a (Figure 6.6, red curve).
The probe image in the top row shows a double-print. The shoe print pattern is clearly visible on the left and at the bottom of the probe image, whereas in the central part of the image a second shoe print distorts the target pattern. The corresponding reference is found at the top position of the ranked list. The following gallery images in the list are perceptually similar to the correct image. This property is highly desirable as it illustrates the stability of the proposed approach (see discussion on instability of repre-

senations in Section 2.1.1)
The probe image in the second row is highly difficult to analyze due to its strong contrast change and the local deformations in the probe image compared to the gallery image (Figure 1.3(c)). The algorithm manages to find the pattern which was labeled as ground truth at the second position. At the first position, it discovered a duplicate print in the database which we were not aware of.
The likelihood ratios in the top row are about three times as big as in the bottom row. This reflects the higher confidence of the algorithm. Due to the design of the appearance model, a major factor which influences this ratio is the contrast of edges. As the probe in the top row is mainly black on white, the CABMs will in general achieve a higher ratio than in low contrast images.

6.4 Conclusion

In this chapter, we evaluated our model-based image analysis framework at the task of forensic shoeprint recognition. We showed that the hierarchical dependency structure is of critical importance for the recognition performance. It enables the combination of a more sophisticated appearance based classification with an occlusion model and a hierarchical decomposition of deformations. Furthermore, the hierarchical composition of LoG filters makes it possible to leverage part-level orientation in the recognition process. We achieved state of the art performance compared to a reimplementation of previous work. The best results are obtained with the two-layered LoG CABM. Compared to "flat" ABMs it benefits from the mentioned hierarchical dependency structure. Multi-layer CABMs have an advantage over two-layered CABMs for large deformations. However, these deformations are rare and therefore the loss of discriminative ability, which is induced by the uniform deformation prior, has a negative impact on the recognition performance.

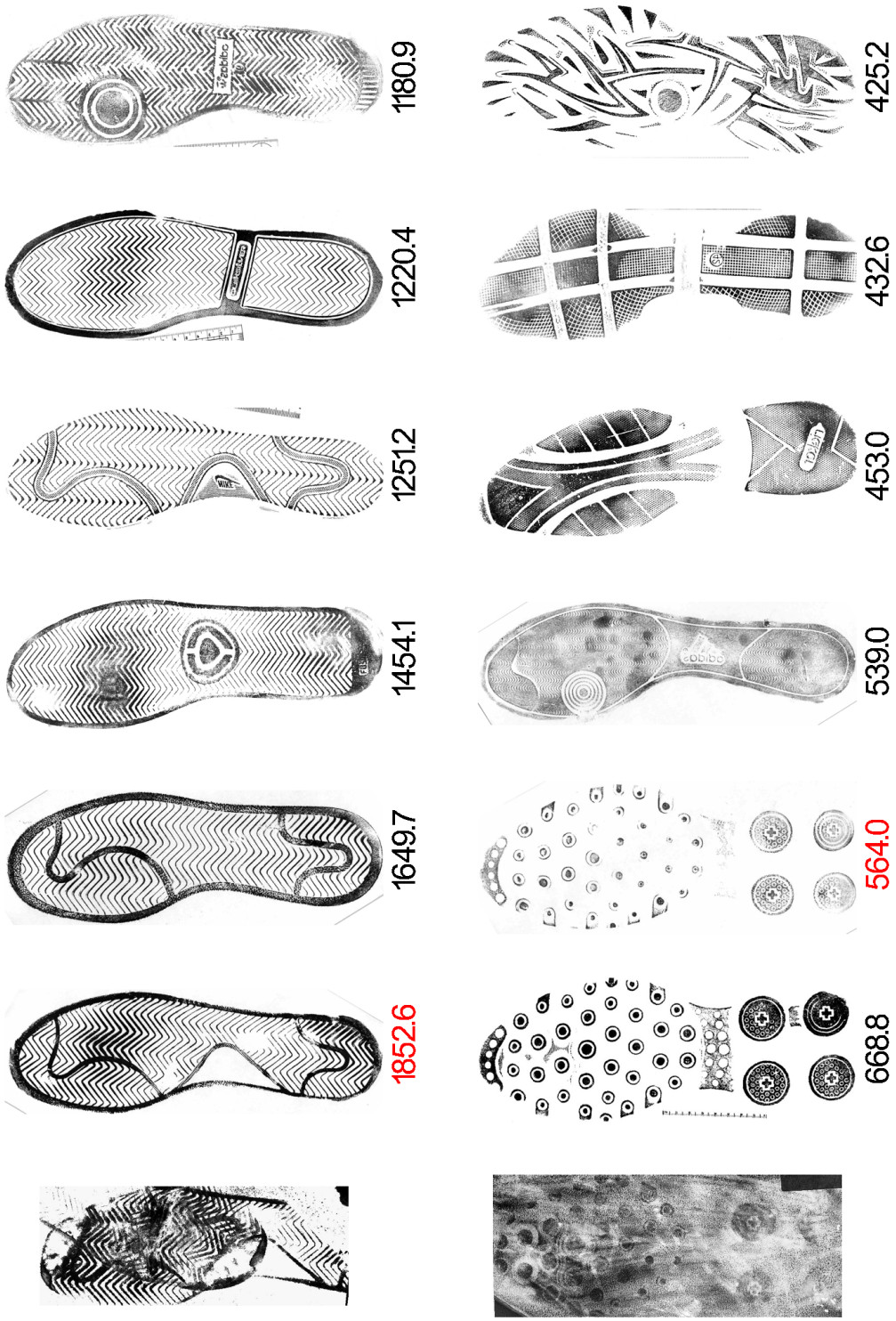


Figure 6.9: Illustration of the retrieval results for two probe images with the proposed two-layered CABM. Each row shows from left to right the probe image and the six most similar gallery images. Below each gallery image we visualize its similarity in terms of the maximal log-likelihood ratio between the object model and the background model (Equation 4.1). The score of the correct gallery image is colored in red. Despite the challenging pattern recognition conditions, the algorithm finds the correct gallery image. For the second probe it even discovered a duplicate in the database.

Chapter 7

Conclusion

The fundamental question we addressed in this thesis is: "How can local shape information be detected in a cluttered image, and how can this information be combined robustly into an object interpretation?"

We studied this question in the context of forensic shoe print recognition. Our answer is a holistic model-based image analysis framework. We represent shoe prints as a hierarchical composition of a Laplacian-of-Gaussian basis. The basis elements detect the local geometry and appearance of a shoe print in an image, whereas the hierarchical model structure enables a global reasoning about an object interpretation, based on this ambiguous local information. A coherent occlusion model facilitates a robust reaction to the effects of clutter in images. The final model holistically accounts for shape and appearance variation, partial occlusion as well as background clutter. It's parameters can be efficiently inferred globally optimally. The application of the proposed framework to forensic shoe print recognition led to a major gain in terms of recognition performance. We belief that our framework is valuable for object recognition beyond this particular application.

7.1 Summary

A significant part of this work was dedicated to a reflection on the trade-off between the invariant and parametric representation in an object model. In this context, we analyzed previous work on object recognition under this perspective. We discussed the major difficulties that prevent a machine from computing the similarity between an object and an image reliably under clutter. We worked out desirable properties of an object representation in order to overcome these difficulties. The Active Basis Model meets many of theses properties. It is object-centered part-based representation that can account for an object's deformation and appearance. Furthermore, it can be learned from a single training image. We have therefore chosen this model to serve as object representation throughout our study. A thorough analysis of this model at the task of shoe print recognition revealed its sensitivity to partial occlusion and a lack in the representation of

characteristic object information. We proposed to overcome these limitations with an extension of the model in terms of a coherent occlusion model and a new basis representation.
Our study continued by analyzing the role of hierarchical abstractions in object representations. We proposed an algorithm for the induction of a multi-layered compositional object representations. We demonstrated the algorithms ability to induce the hierarchical model structure automatically for highly diverse object classes. A detailed evaluation of the effect of the learned hierarchical abstraction showed that the occlusion and the shape model benefit from the additional hierarchical dependency structure. At the end of our theoretical study, we demonstrated the importance of model-based knowledge on the ability to combine ambiguous local information robustly into an image interpretation.
For our experiments, we collected real shoe print data with forensic experts and made it publicly available. We reimplemented prior work on shoe print recognition and demonstrated that our approach delivers a new state of the art performance at the recognition of forensic shoe prints from images. A detailed analysis of the results revealed that each of our extensions in the form of the Laplacian-of-Gaussian basis, a hierarchical occlusion model and hierarchical decomposition of deformations contribute significantly to the discriminative ability of the model.

7.2 Limitations & Future Work

Although, we can perform complex object recognition tasks with the proposed object modeling framework, some limitations remain. We will discuss these in the following section and propose promising directions of future work.

7.2.1 Top-Down Reasoning

As we learn our models form a single training datum, most parameters of the model are set manually based on our prior beliefs. Figure 7.1 illustrates a probe image for which our design of the appearance models is not optimal. The foreground model as we designed it would rather like to explain the strong edge in the background than the low contrast shoe print. Updating the models prior beliefs based on the data which is observed at runtime is therefore a highly promising direction of future research.

Appearance Models. We observed in our experiments, that the balance between foreground and background appearance model is an important factor for the recognition abilities of the system. Setting this balance universally for all probe images is not possible. The appearance of the foreground and background vary significantly between different probe images. We therefore suggest to learn these models from the probe image. The inference process provides us with an estimate about regions of foreground and background in the probe. We can leverage this information to re-learn the appearance models. Subsequently, the inference could be performed again with the newly adapted

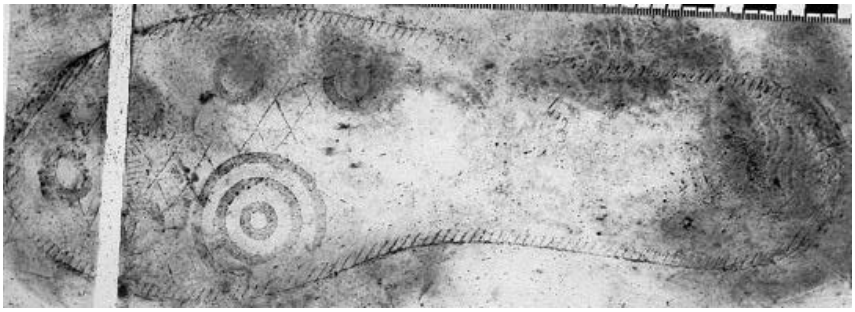


Figure 7.1: In this probe image the foreground and background appearance models in the CABM are not optimal. The foreground is rather of low contrast, whereas the background shows a strong edge. The preference in our appearance models is the other way round. During inference this can mislead the shape model.

appearance models.

Occlusion. No matter how well the appearance models are adjusted to the probe image, the bottom-up decision about the visibility of object parts lacks a full global context. With a top-down occlusion reasoning mechanism this global context could be taken into account. Given a bottom-up inference result, the algorithm would traverse the model tree in a top-down manner and re-evaluate the occlusion states. Thereby, the foreground model could be weakened by lowering its parameter λ. The weaker foreground model would permit previously occluded parts to reappear from the background an thus to serve as evidence for the recognition process.

7.2.2 Discriminative Improvements

We have so far not taken into account that some patterns are more common among different gallery images than others. For example the zigzag pattern in the probe image in Figure 7.2 can be commonly observed in gallery images. However, the curved structure in the shoe print is much more unique. Thus, the curved structure delivers a much stronger evidence than the zigzag pattern only.

In order to estimate this rarity, we could compute how frequently different parts of a CABM occur in other gallery images. At runtime we could simply combine the visible parts in a naive Bayes classifier. Alternatively, one might try to integrate these as importance weights directly into the model. In this way, the model would be guided during inference to actively put more weight on its rare parts.

7.2.3 Model Improvements

Representing non-redundant parts. Our proposed structure induction algorithm relies on the assumption that the parts of a shoe print pattern are repetitive. In fact, some parts are unique and therefore are not represented well in the model (see discussion

Figure 7.2: Different patterns of a shoe print are more common than others. The zigzag lines in the probe can be frequently observed in different gallery images, whereas the curved structure is much more unique. Our recognition system does not take this into account.

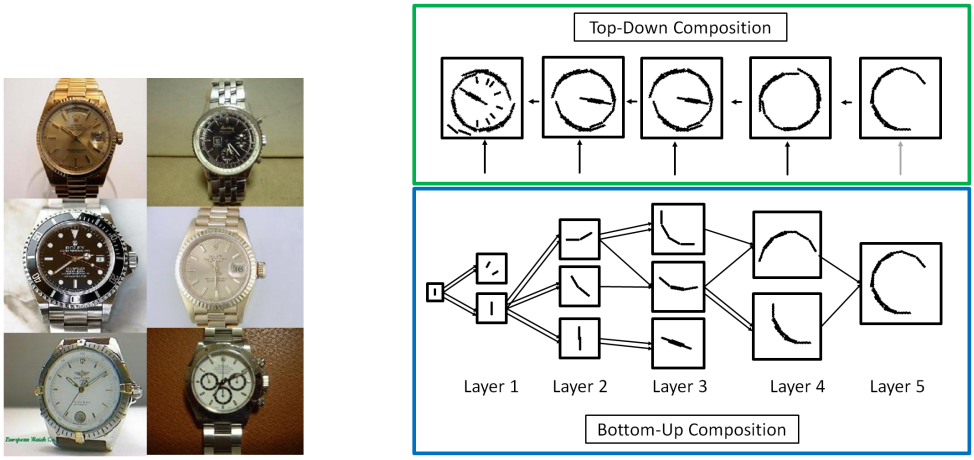
in Section 5.3.3). These underrepresented image regions can be detected easily, since they cannot be reconstructed well from the feature representation. We could then reconstruct these separately with individual filters. However, one would need to find a mechanism to integrate these individual filters into the hierarchical structure efficiently.

Non-uniform deformations. In the experiments we have observed that large deformations decrease the recognition performance significantly. We have identified the uniform deformation prior as the cause of this performance decrease, because it does not penalize large deformations. The choice of this prior is motivated by its computational efficiency during inference (see discussion in Section 5.2). Non-uniform priors require a point-wise matrix multiplication during the max-convolution in the inference procedure and thus are computationally much more expensive. This additional computational load would have to be reduced somehow in order to keep our method as efficient as it is.

7.2.4 Applications beyond Shoe Print Recognition

We have already demonstrated that the proposed CABM learning algorithm is not restricted to learn models from single shape masks, but can also be applied to learn object models form natural images [Kortylewski et al., 2017]. Based on a set of images that depict the same object (Figure 7.3(a)) we can apply essentially the same greedy EM-type bottom-up learning and top-down structure induction processes as presented in Section 5.3.3 (Figure 7.3(b)). Thus, the presented modeling framework is applicable to general object recognition tasks. Many interesting questions arise in this context such as: "How can we model articulated objects?" or "How can we represent a colored appearance coherently despite the hierarchical independence structure?"

Our vision is to combine the presented simple 2D object models with complex 3D object models into an efficient yet powerful holistic object representation.



(a) (b)

Figure 7.3: Our proposed CABM learning algorithm applied to a set of natural images. (a) A subset of the training images. (b) Illustration of the result of our structure induction process. In this example we used a Gabor basis (black lines). During the bottom-up process (blue box). Higher order parts are built by the compositional learning (similar as in Figure 5.15). As soon as no further compositions can be found, the top-down process is initialized (green box). The part from the highest layer is greedily augmented with parts from lower layers based on the proposed top-down structure induction (Figure 5.16).

List of Abbreviations

ABM	Active Basis Model
BIC	Bayesian Information Criterion
CABM	Compositional Active Basis Model
CDF	Cumulative Distribution Function
CMC	Cumulative Match Characteristic
EM	Expectation–Maximization
FID	Footwear Impression Database
FPR	False Positive Rate
HCM	Hierarchical Compositional Model
HOG	Histogram of Oriented Gradients
LoG	Laplacian-of-Gaussian
MRF	Markov Random Field
ROC	Receiver Operating Characteristic
SIFT	Scale-Invariant Feature Transform
TPR	True Positive Rate

Bibliography

T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European conference on computer vision*, pages 469–481. Springer, 2004. 14

A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*, pages 510–517. Ieee, 2012. 14

G. AlGarni and M. Hamiane. A novel technique for automatic shoeprint image retrieval. *Forensic science international*, 181(1):10–14, 2008. 11

J. Altmann and H. J. Reitbock. A fast correlation method for scale-and translation-invariant pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, (1):46–57, 1984. 10

Y. Amit. *2D object detection and recognition: Models, algorithms, and networks*. MIT Press, 2002. 17, 24

Y. Amit. Pop: Patchwork of parts models for object recognition. *International Journal of Computer Vision*, 75(2):267–282, 2007. 15

Y. Amit and A. Kong. Graphical templates for model registration. *IEEE Transactions on pattern analysis and machine intelligence*, 18(3):225–236, 1996. 15

Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, 86(414):376–387, 1991. 15

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. 60

N. Ayache and O. D. Faugeras. Hyper: A new approach for the recognition and positioning of two-dimensional objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):44–54, 1986. 14

H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. *Computer Vision–ECCV 2012*, pages 836–849, 2012. 18, 38, 39

M. Bach. Visual phenomena & optical illusions. http://www.michaelbach.de/ot/cog-hiddenBird/index.html, 2017. 4

H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 14

V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 17

W. J. Bodziak. *Footwear impression evidence: detection, recovery and examination.* CRC Press, 1999. 2

R. N. Bracewell and R. N. Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986. 10

R. A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial intelligence*, 17(1-3):285–348, 1981. 14

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013. 11

M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. *Computer Vision—ECCV'98*, pages 628–641, 1998. 15

M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer, 2010. 14

F. Cervelli, F. Dardi, and S. Carrato. A translational and rotational invariant descriptor for automatic footwear retrieval of real cases shoe marks. Eusipco, 2010. 14

T. F. Cootes and C. J. Taylor. Active shape models—'smart snakes'. In *BMVC92*, pages 266–275. Springer, 1992. 18

T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 1995. 15, 22

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 17

J. Dai, Y. Hong, W. Hu, S.-C. Zhu, and Y. N. Wu. Unsupervised learning of dictionaries of hierarchical compositional models. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2505–2512. IEEE, 2014. 5, 53, 55, 56, 57, 78

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 14, 44

F. Dardi, F. Cervelli, and S. Carrato. A texture based shoe retrieval system for shoe marks of real crime scenes. In *Image Analysis and Processing–ICIAP 2009*, pages 384–393. Springer, 2009. 14, 81, 85, 86

P. De Chazal, J. Flynn, and R. B. Reilly. Automated processing of shoeprint images based on the fourier transform for use in forensic science. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):341–350, 2005. 11, 81, 85

B. DeCann and A. Ross. Relating roc and cmc curves via the biometric menagerie. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8. IEEE, 2013. 85

P. F. Felzenszwalb. Representation and detection of deformable shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005. 53

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 15

V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *International journal of computer vision*, 2010. 57

S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 57, 71, 72

S. Fidler, M. Boben, and A. Leonardis. Learning a hierarchical compositional shape vocabulary for multi-class object representation. *arXiv preprint arXiv:1408.5516*, 2014. 57

M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 1973. 15

W. T. Freeman, D. B. Anderson, P. Beardsley, C. N. Dodge, M. Roth, C. D. Weissman, W. S. Yerazunis, H. Kage, I. Kyuma, Y. Miyake, et al. Computer vision for interactive computer graphics. *IEEE Computer Graphics and Applications*, 18(3):42–53, 1998. 10

S. Geman, D. F. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, 60(4):707–736, 2002. 71

G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2385–2392, 2014. 18, 69, 77

R. B. Girshick. *From rigid templates to grammars: Object detection with structured models*. Citeseer, 2012. 15, 53

U. Grenander. A unified approach to pattern analysis. *Advances in Computers*, 10: 175–216, 1970. 15

U. Grenander. Lectures in pattern theory i, ii and iii: Pattern analysis, pattern synthesis and regular structures, 1976. 15, 24

U. Grenander, Y.-s. Chow, and D. M. Keenan. *Hands: A pattern theoretic study of biological shapes*. 1990. 55

M. Gueham, A. Bouridane, D. Crookes, and O. Nibouche. Automatic recognition of shoeprints using fourier-mellin transform. In *Adaptive Hardware and Systems, 2008. AHS'08. NASA/ESA Conference on*, pages 487–491. IEEE, 2008. 11, 81, 85

Y. Hong, Z. Si, W. Hu, S.-C. Zhu, and Y. N. Wu. Unsupervised learning of compositional sparse code for natural image representation. *Quarterly of Applied Mathematics*, 72: 373–406, 2013. 21, 27, 53, 58, 60

M.-K. Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187, 1962. 10

P. J. Huber. *Robust statistics*. Springer, 2011. 17

D. P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990. 14

A. K. Jain, Y. Zhong, and S. Lakshmanan. Object matching using deformable templates. *IEEE Transactions on pattern analysis and machine intelligence*, 1996. 17, 22

Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2145–2152. IEEE, 2006. 53, 57

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. 21

D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99, 1989. 15

A. Khotanzad and Y. H. Hong. Invariant image recognition by zernike moments. *IEEE Transactions on pattern analysis and machine intelligence*, 12(5):489–497, 1990. 10

I. Kokkinos and A. Yuille. Inference and learning with hierarchical shape models. *International Journal of Computer Vision*, 93(2):201–225, 2011. 57

A. Kortylewski and T. Vetter. Probabilistic compositional active basis models for robust pattern recognition. In *British Machine Vision Conference*, 2016. 16, 71

A. Kortylewski, T. Albrecht, and T. Vetter. Unsupervised footwear impression analysis and retrieval from crime scene data. In *Asian Conference on Computer Vision*, pages 644–658. Springer, 2014. 86, 88

A. Kortylewski, C. Blumer, and T. Vetter. Greedy compositional clustering for unsupervised learning of hierarchical compositional models. *arXiv preprint arXiv:1701.06171*, 2017. 73, 100

D. Lemire. Streaming maximum-minimum filter using no more than three comparisons per element. *arXiv preprint cs/0610046*, 2006. 30

D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial intelligence*, 31(3):355–395, 1987. 14

D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision, 1999*, pages 1150–1157. IEEE, 1999. 14

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 14, 44

T. Luostarinen and A. Lehmussola. Measuring the accuracy of automatic shoeprint recognition methods. *Journal of forensic sciences*, 2014. 14

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993. 23

D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, 207(1167):187–217, 1980. 42

M. Mercimek, K. Gulez, and T. V. Mumcu. Real object recognition using moment invariants. *Sadhana*, 30(6):765–775, 2005. 10

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005. 14

N. J. Mitra, H.-K. Chu, T.-Y. Lee, L. Wolf, H. Yeshurun, and D. Cohen-Or. Emerging images. In *ACM Transactions on Graphics (TOG)*, 2009. 1

D. Mumford and A. Desolneux. *Pattern theory: the stochastic analysis of real-world signals*. CRC Press, 2010. 24

O. Nibouche, A. Bouridane, D. Crookes, M. Gueham, et al. Rotation invariant matching of partial shoeprints. In *Machine Vision and Image Processing Conference, 2009. IMVIP'09. 13th International*, pages 94–98. IEEE, 2009. 14, 81

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 14

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996. 22, 58

P. M. Patil and J. V. Kulkarni. Rotation and intensity invariant shoeprint matching using gabor transform with application to forensic science. *Pattern Recognition*, 42 (7):1308–1317, 2009. 14

M. Pavlou and N. M. Allinson. Automatic extraction and classification of footwear patterns. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, pages 721–728. Springer, 2006. 14

M. Pavlou and N. M. Allinson. Automated encoding of footwear patterns for fast indexing. *Image and Vision Computing*, 27(4):402–409, 2009. 14

J. Radon. Uber die bestimmung von funktionen durch ihre integralwerte langs gewisser mannigfal-ligkeiten, ber. verh. säcks. akad. 69 (1917), 262-277. *Radon26269Ber. Verh. Sächs. Akad*, 1917. 10

A. Ross. Relating roc and cmc curves. https://www.nist.gov/sites/default/files/documents/2016/12/06/12_ross_cmc-roc_ibpc2016.pdf, 2017. 85

S. Schönborn, B. Egger, A. Forster, and T. Vetter. Background modeling for generative image models. *Computer Vision and Image Understanding*, 136:117–127, 2015. 17, 18, 24

Z. Si and S.-C. Zhu. Learning and-or templates for object recognition and detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(9):2189–2205, 2013. 53, 57, 71

L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2041–2048. IEEE, 2006. 18

H. Su, D. Crookes, A. Bouridane, and M. Gueham. Local image features for shoeprint image retrieval. In *British Machine Vision Conference*, volume 2007, 2007. 14

Y. Tang, S. N. Srihari, H. Kasiviswanathan, and J. J. Corso. Footwear print retrieval system for real crime scene marks. In *Computational Forensics*, pages 88–100. Springer, 2011. 16, 81, 85, 87

S. Ullman et al. *High-level vision: Object recognition and visual cognition.* MIT press Cambridge, MA, 1996. 5

C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2013. 14, 44

M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *Computer Vision-ECCV 2000*, pages 18–32, 2000. 15

P. Weinzaepfel, H. Jégou, and P. Pérez. Reconstructing an image from its local descriptors. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 337–344. IEEE, 2011. 14

P. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967. 10

B. Widrow. The rubber-mask technique-i. pattern measurement and analysis. *Pattern recognition*, 1973a. 15

B. Widrow. The rubber-mask technique-ii. pattern storage and recognition. *Pattern Recognition*, 1973b. 15

J. Wood. Invariant pattern recognition: a review. *Pattern recognition*, 29(1):1–17, 1996. 10

Y. N. Wu, Z. Si, C. Fleming, and S.-C. Zhu. Deformable template as active basis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 21, 53, 58

Y. N. Wu, Z. Si, H. Gong, and S.-C. Zhu. Learning active basis model for object detection and recognition. *International journal of computer vision*, 90(2):198–235, 2010. 5, 6, 16, 17, 21, 23, 25, 26, 30, 31, 35, 55

Z. Ying and D. Castañon. Partially occluded object recognition using statistical models. *International Journal of Computer Vision*, 49(1):57–78, 2002. 18, 39

A. Yuille and D. Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006. 78

A. L. Yuille. Towards a theory of compositional learning and encoding of objects. In *ICCV Workshops*, pages 1448–1455. Citeseer, 2011. 71

A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision*, 8(2):99–111, 1992. 15, 22, 24

L. L. Zhu, C. Lin, H. Huang, Y. Chen, and A. Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *Computer vision–eccv 2008*, pages 759–773. Springer, 2008. 57, 71, 72, 73