Probabilistic Compositional Active Basis Models for Robust Pattern Recognition

Adam Kortylewski
adam.kortylewski@unibas.ch
Thomas Vetter
thomas.vetter@unibas.ch

Department of Mathematics and Computer Science
University of Basel
Basel, Switzerland

Abstract

Hierarchical compositional models (HCMs) have recently shown impressive generalisation capabilities, especially with small amounts of training data. However, regarding occlusion and background clutter experimental setups have been relatively controlled so far. The contribution of this paper is two-fold. First, we introduce a greedy EM-type algorithm to automatically infer the complete structure of HCMs. Second, we demonstrate how HCMs can be applied to the robust analysis of patterns under structured distortions. The proposed compositional active basis models (CABM) are embedded into a probabilistic formulation of the learning and inference processes. Building on the statistical framework, we enhance the CABM with an implicit geometric background model that reduces the models sensitivity to outliers due to occlusions and background clutter.

In order to demonstrate the robustness of the proposed object representation, we evaluate it on a complex forensic image analysis task. We demonstrate that probabilistic CABMs are capable of recognising patterns under complex non-linear distortions that can hardly be represented by a finite set of training data. Experimental results show that the forensic image analysis task is processed with unprecedented quality.

1 Introduction

A critical property for computer vision systems is the robustness against pattern distortions and structured background. Recently, hierarchical compositional models (HCMs) have shown impressive generalisation ability in standard classification [7], transfer learning [3] and one-shot learning [28]. However, regarding occlusions and non-linear distortions of patterns, experimental setups have been controlled so far. An important question is how robust these models are under more challenging pattern recognition conditions. The automated analysis of forensic images is highly suitable for studying this question. The task of forensic footwear impression recognition is particularly interesting because it unifies many computer vision questions in a well-defined application scenario (Figure 1). Some of the most interesting properties of this application are: 1) The patterns in probe images are significantly occluded and subject to non-linear distortions that interfere with the pattern. 2) The background signal contains structured geometry that is difficult to distinguish from the actual pattern of interest. 3) The geometry of the patterns is diverse and complex. 4) Probe images are scarce compared to the number of reference impressions, thus learning has to be performed in one shot.

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

Figure 1: Overview over the process of automated footwear impression analysis. (a) A typical probe image. The pattern is distorted by occlusion and background clutter; (b) The corresponding reference impression; (c) Sketch of the learned CABM for the reference impression. Pixels that share the same colour are explained by the same type of part. (d) An overlay of the learned CABM over the probe image with the spatial transformation that maximises the posterior probability. Despite complex structured background and missing parts, the correct spatial transformation has been recovered.

We propose to formulate this pattern recognition task in a statistical estimation setting by representing a reference impression with a generative model. We will estimate the posterior distribution of the model parameters given the probe image. The advantages of a generative approach to pattern analysis have been argued in detail *e.g.* in [11]. We will show that a thorough mathematical formulation of the pattern recognition process is important for overcoming the challenging properties 1) - 4) mentioned in the previous paragraph.

We propose to represent the pattern as a hierarchical composition of active basis models. Active basis models are generative HCMs. During learning, the model is composed hierarchically from groups of active basis models in a bottom-up manner. The model structure is learned with a greedy EM-type clustering process (Sections 3.1 & 3.2). The resulting compositional active basis model (CABM) encodes local as well as long-range geometric constraints of the pattern. In this way it forms a powerful prior for the distinction of the actual pattern of interest from the structured background patterns.

We present a fully probabilistic formulation of the learning and inference process for CABMs. Building on the statistical framework, we enhance the CABM with an implicit geometric background model that increases the robustness against occlusion and clutter. The main novelties of this work are:

i) A greedy EM-type algorithm that can infer the full structure of general HCMs

ii) A fully probabilistic formulation of compositional active basis models

iii) An implicit geometric background model that increases the CABMs robustness to occlusion and structured background clutter

iv) A significant improvement of the recognition performance in footwear impression analysis

Prior Work on HCMs: Hierarchical compositional models have been successfully applied in computer vision applications e.g. in [7, 8, 9, 19, 24, 28, 32]. However, the models are usually applied in a relatively controlled experimental setup with respect to distortions of the patterns and occlusions and/or trained with a lot of data. In this work, we learn a hierarchical compositional representation from just one training sample and perform pattern recognition under highly unconstrained conditions. Our work builds on the compositional sparse coding procedure proposed in [3, 29]. However, we do not stop after the dictionary learning phase, but encode higher structural relationships between dictionary templates. Probabilistic HCMs have been proposed for representing faces in [26, 30] and for general objects in [3]. However, in contrast to these, we automatically learn the structure of the hierarchy based on the greedy EM-type algorithm proposed. This facilitates the automated selection of the number of dictionary templates and hierarchical layers. Unsupervised learning of HCMs has been successfully performed in *e.g.* [6, 32]. However, [6] is not probabilistically formulated. The work in [32] is most related to our work. The main differences are that we use fully generative compositional units instead of invariant features. Furthermore, we do not make hard decisions on the detection of parts instead the full part likelihoods are used throughout the bottom-up learning process. Finally, our model is enhanced with an implicit geometric background model, which makes it more robust to occlusions and background clutter. Despite the popularity of hierarchical compositional models, to the best of our knowledge, this is the first time they are shown to achieve state-of-the-art recognition performance in a highly unconstrained vision task.

Prior Work on Footwear Impression Analysis: Earlier attempts in footwear impression recognition learn global [1, 5, 11] or local [13, 21, 22, 25] hand-crafted feature representations. However, it was shown that the application scenario of these works is limited [13, 15] (see also experiments in Section 4). The main reasons are that pure local as well as pure global representations are sensitive to local distortions of patterns. Several works enrich local features with global constraints [2, 4, 17, 20, 27]. However, the main assumption in all works is that the features can be detected successfully by a purely local process. Thus, local ambiguities as well as structured backgrounds and local pattern distortions have not been taken into account. In this work, we propose to encode both the local and global structure in a joint pattern model. The hierarchical representation renders it possible to localise the matching cost during model fitting. We augment the pattern model with a geometric background model that increases robustness to structured clutter and missing parts.

Experiments. Experimental comparison is performed on the FID-300 database [13] (http://fid.cs.unibas.ch/). We demonstrate an increase in recognition performance by a wide margin compared to previous works [2, 5, 11, 13, 27].

In Section 2, we will introduce the theoretical background of traditional active basis models. Section 3 introduces a detailed probabilistic formulation of compositional active basis models and the implicit geometric background model. Experimental results are presented in Section 4.

2 Theoretical Background - Active Basis Models

In this Section we shall introduce active basis models (ABMs). Detailed information concerning ABMs can be found in the original work [29]. We concentrate on the results that are relevant for understanding our contribution. We adapt the notation used in [29] at some points such that it fits into the theoretical framework presented in Section 3.

ABMs are a type of deformable template for describing object shapes under small local shape deformations. An ABM is composed of a set of basis filters at positions $X_i = \{x_i, y_i\}$ with orientations α_i. Throughout this work, we use combinations of even and odd Gabor wavelets B as basis filters. We keep the frequency fixed. The set of parameters per filter is denoted by $\beta_i^0 = \{X_i^0, \alpha_i^0\}$. The spatial parameters are encoded relative to the position of the overall template β_1^1, which is, for now, assumed to be given. The position of an individual basis filter in the image frame therefore is $\beta_i = \{X_i = X_1^1 + X_i^0, \alpha_i = \alpha_1^1 + \alpha_i^0\}$. The parameters of an ABM are denoted by $\Pi = \{\beta_i^0 | i = 1 \dots N\}$. The global spatial configuration of the basis filters is rigid. However, each filter can perturb its location and orientation independently of the other filters within a small specified range $\Delta\beta = \{\Delta X, \Delta\alpha\}$. This active deformation enables the model to compensate small changes in the object shape without the need for re-optimising the state of all variables, as would be the case when using a global shape model.
An ABM is a linear additive model in the form of the well-known sparse coding principle proposed by Olshausen and Field [18]. An important difference, however, is that the ABM is applied to represent a whole ensemble of image patches $\{I_m, m = 1, \dots M\}$. Each patch is represented by:

$$I_m = C_m B_\Pi + U_m = \sum_{i=1}^{N} c_{i,m} B_{\beta_i} + U_m. \tag{1}$$

The patches I_m are linearly decomposed into a set of orthogonal basis filters B_Π with coefficients C_m and the residual image U_m. The individual coefficients are calculated by $c_{i,m} = \langle I_m, B_{\beta_i} \rangle$. The basis filters have zero mean and unit l_2 norm. The probability density of a patch I_m given the template Π is modelled by:

$$p(I_m|\Pi) = p(U_m|C_m)p(C_m|\Pi) = p(U_m|C_m) \prod_{i=1}^{N} p(\beta_i^0|\beta_1^1) p(c_{m,i}|\beta_i^0) \tag{2}$$

The factorization in Equation 2 is based on the assumption that the model has a tree structure and that parts do not overlap. In the original equation as introduced in [29], the factor $p(\beta_i^0|\beta_1^1)$ is omitted. This is equivalent to assuming that the patches $\{I_m | m = 1, \dots, M\}$ are aligned and depict an object that is exactly in the same pose. This assumption is a major weakness of the active basis model approach. In Section 3.1 we will show that the model can be learned from unaligned training images as proposed in [12]. A more challenging task is to resolve the assumption about the fixed pose of the object. This is, however, beyond the scope of this work as footwear impressions can be approximated by rigid objects.
The template Π can be learned based on a set of training patches I_m with a matching pursuit process [16]. Subsequently, the composition of filters B_Π could be directly applied as an object detector. However, in order to be less sensitive to strong edges in the background clutter we estimate the expected distribution of filter responses in a background image $q(c_{m,i}|\beta_i^0)$ and compare it to the one we observe in the training patches $p(c_{m,i}|\beta_i^0)$. Let $q(I|\Pi) = q(C,U|\Pi) = q(U|C)q(C|\Pi)$ model the distribution of filter responses and residual images as they occur in natural images. The ratio between the foreground and the background model is:

$$\frac{p(I_m|\Pi)}{q(I_m|\Pi)} = \frac{p(U_m|C_m) \prod_{i=1}^{N} p(\beta_i^0|\beta_1^1) p(c_{m,i}|\beta_i^0)}{q(U_m|C_m) \prod_{i=1}^{N} q(\beta_i^0|\beta_1^1) q(c_{m,i}|\beta_i^0)} = \prod_{i=1}^{N} \frac{p(\beta_i^0|\beta_1^1) p(c_{m,i}|\beta_i^0)}{q(\beta_i^0|\beta_1^1) q(c_{m,i}|\beta_i^0)}. \tag{3}$$

An important assumption is that the probability densities of the residual background are the same $q(U_m|C_m) = p(U_m|C_m)$ [12, 29], thus they cancel out of the equation. This means that those parts of the image that cannot be explained by the basis filters follow the same distribution. Furthermore, $p(\beta_i^0|\beta_1^1)$ can be modelled by a uniform distribution over the range of active perturbation $U_{\beta_i^0}(\Delta\beta)$ around β_i^0. The background model $q(\beta_i^0|\beta_1^1) = U(D,\alpha)$ is uniform over the orientations α and the patch domain $D = d \times d$, where d is the size of the patch. We assume $q(c_{m,i}|\beta_i^0)$ is stationary and therefore translation-, rotation- and scale-invariant. The distribution $q(c_{m,i}|\beta_i^0)$ can be estimated by pooling a histogram of basis filter responses from a random set of natural images. In contrast to the standard assumption of a Gaussian distribution, $q(c_{m,i}|\beta_i^0)$ is much more heavy-tailed and can therefore better explain strong edges that occur in the cluttered background. This approach of reducing the sensitivity to clutter was introduced in [29]. We will introduce an additional implicit background model on compositions of filters in Section 3.3.

The foreground distribution $p_i(c_{m,i}|\beta_i^0)$ is modelled in the form of an exponential family model:

$$p(c_{m,i}|\lambda_i,\beta_i^0) = \frac{exp(\lambda_i\sigma(|c_{m,i}|^2))q(c_{m,i}|\beta_i^0)}{Z(\lambda_i)}, \tag{4}$$

As proposed in [29], we apply a sigmoid transform $\sigma(r) = \tau[2/(1+e^{-2r/\tau})-1]$ that saturates at value τ. The normalising constant $Z(\lambda_i)$ as well as the mean of the model $\mu(\lambda_i)$ can be estimated for a range of λ values on a set of natural training images by numerical integration. Following the maximum entropy principle [23], the maximum likelihood estimate for $\lambda_i = \mu^{-1}(\sum_{m=1}^{M}\sigma(|c_{m,i}|^2)/M)$. The coupling of the matching pursuit process with the modelling of the expected distribution of the coefficients is generally referred to as shared matching pursuit [29]. We denote the final ABM by $\Theta = \{\Pi,\Lambda\}$, where $\Lambda = \{\lambda_i | i = 1,\dots,N\}$.

In the next Section 3, we will introduce a hierarchical extension of ABMs called compositional active basis models (CABMs). We propose a greedy EM-type learning scheme that makes it possible to induce the hierarchical model structure in an unsupervised manner. Furthermore, we embed the methodology in a fully probabilistic theoretical framework.

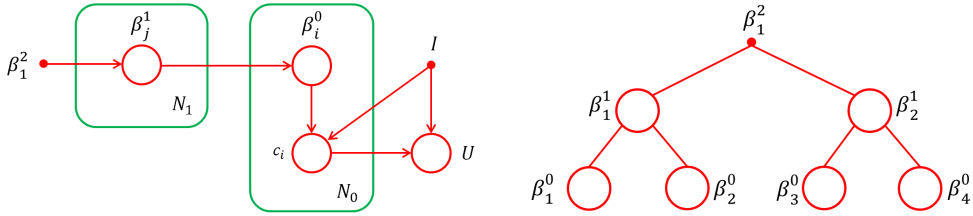
3 Compositional Active Basis Models

In this Section we will extend the active basis model framework to encompass hierarchic compositions of ABMs (Section 3.1 & 3.2). The advantages of hierarchical compositional models have been argued in detail in *e.g.* [8, 31, 33]. Regarding the traditional flat ABM, a hierarchical model makes it possible to decouple the globally rigid dependence structure between the random variables into localised group-wise dependencies. The hierarchical decoupling will allow us to compensate missing object parts with a robust geometric background model during inference and will thus lead to a more robust pattern recognition in Section 4.

For ease of notation, we will use in all equations the example of a level-two CABM. A graphical model with $N_1 = 2$ level-one groups is depicted in Figure 2. This is the simplest possible CABM. However, the presented results fully generalise to arbitrary numbers of layers and compositions per node. Note that the standard ABM can be regarded as a special case of CABM with no compositional layer.

The probability density of an image patch given a level-two CABM factorises in the following way:

$$p(I_m|\Theta) = p(U_m|C_m) \prod_{j\in ch(\beta_1^2)} p(\beta_j^1|\beta_1^2) \prod_{i\in ch(\beta_j^1)} p(\beta_i^0|\beta_j^1)p(c_{m,i}|\beta_i^0), \tag{5}$$



(a) (b)

Figure 2: Graphical model of a level-two compositional active basis model. (a) The full graphical model; (b) The common way of illustrating hierarchical models, by focusing on the model structure. We depict the simplest possible compositional active basis model, a binary-tree structured Markov random field.

where the term $ch(\beta_j^1)$ denotes the set of child nodes of the node β_j^1. The compositional layer introduces the factor $p(\beta_i^0|\beta_j^1)$, which conditions the spatial configuration of the individual basis filters β_i^0 on different parent nodes β_j^1. In this way, the global dependence structure is broken into multiple conditionally independent groups. However, the increased power of the model comes at the cost of having to estimate more parameters. In this work, we present an algorithm that is capable of estimating the number of independent level-one groups N_1 (Section 3.1) as well as the number of layers L (Section 3.2). During learning, we benefit from the compositional structure of the model, as it allows us to first learn the level-one models, before composing them into a level-two model. This property facilitates the efficient learning of complex hierarchical structures as demonstrated in [6, 32]. We manually set the number of parts that are composed to two. However, the proposed learning scheme can be applied with any number of compositions. Following the standard active basis model framework, we assume that the geometric relation between ABMs can be modelled as uniform distribution over the range of active perturbation. Therefore we define $p(\beta_j^1|\beta_1^2) = U_{\beta_j^1}(\Delta\beta)$.

In the following Section 3.1 we will introduce an algorithm that will infer the number of nodes for a layer N_l given the parts of the previous layer automatically.

3.1 Greedy EM-type Clustering

In order to learn the level-one models with shared matching pursuit [29], we must first gather the training patches for the individual models. This can be done by standard K-Means clustering as proposed in [31, 32]. However, in an unsupervised learning setup it is desirable to automatically determine the optimal number of clusters. We therefore introduce a greedy EM-type clustering scheme. We start by learning the first level-one model Θ_1^1 according to the following procedure: In the first iteration $t = 1$, we sample an initial set of patches I_1^t according to an initial distribution Q. We will define Q to be uniform on the image lattice $Q(x,y) = U(x,y)$. However, alternative distributions that are based on prior measures could be possible (*e.g.* based on saliency or on the gradient information). We learn an initial ABM θ_1^t from I_1^t with the shared matching pursuit algorithm. For the next learning iteration, we gather all image patches for which the prediction of the object model is better than a default background model:

$$p(I_1^{t+1}|\theta_1^t) > d(I_1^{t+1}) \tag{6a}$$

$$\max \prod_{i \in ch(\beta_1^1)} p(\beta_i^0|\beta_1^1) \frac{exp(\lambda_i \sigma(|c_{m,i}|^2))q(c_{m,i}|\beta_i^0)}{Z(\lambda_i)} > \max \prod_{i \in ch(\beta_1^1)} U(\beta_i^0) q(c_{m,i}|\beta_i^0). \tag{6b}$$

The default model $d(I_1^{t+1})$ simply assumes that the parts follow independent uniform distributions over the domain of the patch. Note that the parameters β_i^0 can be different for the two sides of the inequality. Alternatively, a fixed detection threshold could also be applied. The set of patches that satisfies Equation 6b serves as training data for the next iteration of shared matching pursuit. We terminate the iterative learning process when $p(I_1^{t+1}|\theta_1^t)$ does not change significantly between consecutive iterations. Finally, we set $\Theta_1^1 = \theta_1^t$.

We repeat the above procedure for the second level-one model θ_2^t, but this time the object model θ_2^t must achieve a better prediction on the training patches I_2^t than all previously learned models:

$$p(I_2^{t+1}|\theta_2^t) > max(d(I_2^{t+1}), p(I_2^{t+1}|\Theta_1^1)). \tag{7}$$

In this way, ABMs are learned greedily until a new model is unable to explain some parts of the image better than previously learned models. This process is inspired by the EM-type learning as proposed in [12]. However, the important difference is that by introducing the default model, we are in addition able to infer the number of clusters from the data.

Given a set of level-one ABMs, we shall in Section 3.2 compose these into higher-order models that encode long-range structural dependencies of the trainingpattern.

3.2 Compositional Structure Induction

A common way of learning higher-order compositional models is to detect the learned level-one models Θ_i^1 based on a fixed threshold η, and to subsequently learn part compositions using standard clustering techniques [6, 31, 32]. However, we propose to follow the same greedy EM-type clustering as introduced in Section 3.1 in order to learn compositions of ABMs. Hence, we replace the Gabor wavelets as basis filters with the learned level-one models Θ_i^1. The main advantage compared to other approaches is that we can avoid to take an early decision on the presence of level-one models. Thus, we can leverage additional knowledge from the level-two model when deciding on the presence of level-one models. This late commitment is possible because of $p(I_m|\Theta_j^2)$ is a weighted summary of low level statistics $p(I_m|\Theta_i^1)$ (Equation 5). Therefore, if one of these $p(I_m|\Theta_i^1)$ is a bit too low, the compositional distribution $p(I_m|\Theta_j^2)$ can still compensate for this in order to outperform the default model. In this way, parts can be recovered that would have been classified as background before. This process can be observed in Figure 3 multiple times, whenever image regions that are not encoded in one layer get encoded in the layer above. The selection process for the training patches I_2^t can again be guided by the independence principle as in Equation 6. The procedure is repeated for multiple levels until no further compositions are found, thus generating a probabilistic compositional active basis model Θ^L. The results of the learning process are illustrated in Figure 3.

In order to build a model for the whole reference impression, we do not need a complex top-down process as *e.g.* [32]. We can assume that the structure in the training image is generated by the object of interest. Therefore, the full CABM can be built by connecting all detected parts to the root node that are not explained away by a part from a higher layer

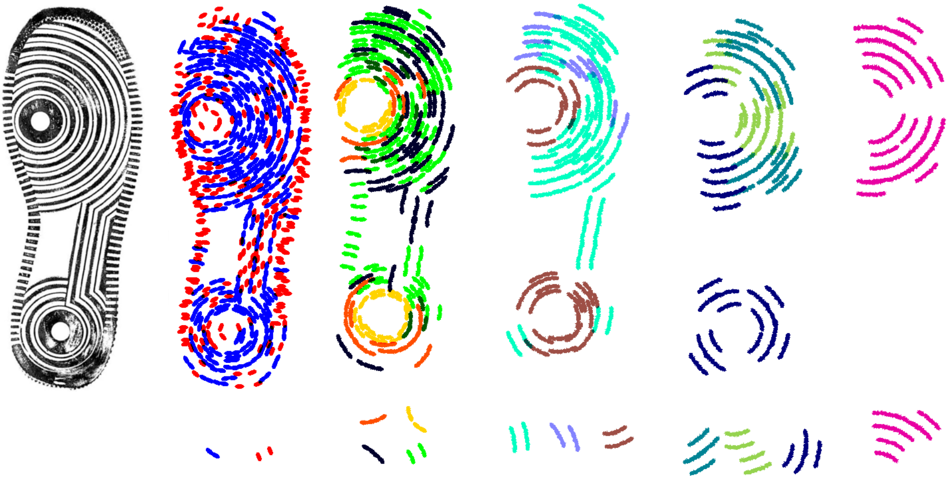


Figure 3: The results of the compositional learning procedure when applied to a reference impression. (a) The input image. (b-f) The learning result for each layer $(1-5)$ of the hierarchy. **Bottom row**: Illustration if the learned CABMs with different colours in their mean position. The individual Gabor wavelets are represented by small strokes. **Top row**: The input image when encoded with the learned models of each layer.

(Figure 1).
At this point, we have learned the number of layers L as well as the number of parts for each individual layer $N_{1,...,L}$. Furthermore, we have formulated the pattern model as well as the learning process in a fully probabilistic manner. These achievements mark the main contribution of this work. In the following Section 3.3, we further propose to augment the pattern model with an implicit background model that reduces the sensitivity to outliers due to occlusions or structured clutter.

3.3 Robust Inference

Let assume we are given a two-level CABM Θ_1^2. We want to infer its optimal spatial configuration for a test image I. Thus, we want to maximise the posterior $p(\Pi|I,\Theta_1^2)$. According to Bayes' rule the posterior can be written as:

$$p(\Pi|I,\Theta_1^2) \propto P(I|\Pi,\Theta_1^2)P(\Pi|\Theta_1^2). \tag{8}$$

We can infer the parameters with a standard recursive bottom-up inference procedure as *e.g.* presented in [3, 6, 32]. A main issue is, however, that in the probe images some parts of the reference impression are missing (Figure 1). Without adjustments to the standard model (Equation 5), missing parts are evaluated at the background and thus disproportionately decrease the posterior probability at the correct position. As we do not have prior information on what parts are occluded or on the appearance of the background, we cannot pre-learn an explicit occlusion model as *e.g.* in [9, 14]. Instead, we augment the distribution that models the geometry between parts with an implicit background model:

$$\hat{p}(\beta_i^0|\beta_j^1) = \frac{p(\beta_i^0|\beta_j^1) + U_r}{2}. \tag{9}$$

The distribution U_r is defined over the whole patch domain and is greater zero where $p(\beta_i^0|\beta_j^1) = 0$. In this way, part configurations that could not be explained by $p(\beta_i^0|\beta_j^1)$ at all are assigned a small probability in $\hat{p}(\beta_i^0|\beta_j^1)$. Thus the CABM becomes more robust to locally unlikely part configurations if the other parts of the model still fit well with the data.

4 Experiments

We evaluate the proposed methodology on the FID-300 dataset [13] (http://fid.cs.unibas.ch/). The footwear impression dataset contains 300 probe images I_P and 1175 reference impressions. During training we learn a pattern model Θ_R for each of the reference impressions. At testing time we calculate the posterior $p(\Pi_R|I_P,\Theta_R)$ for each model. An important aspect of the probabilistic embedding of compositional active basis models is that we can compare inference results for models with different numbers of layers and nodes.

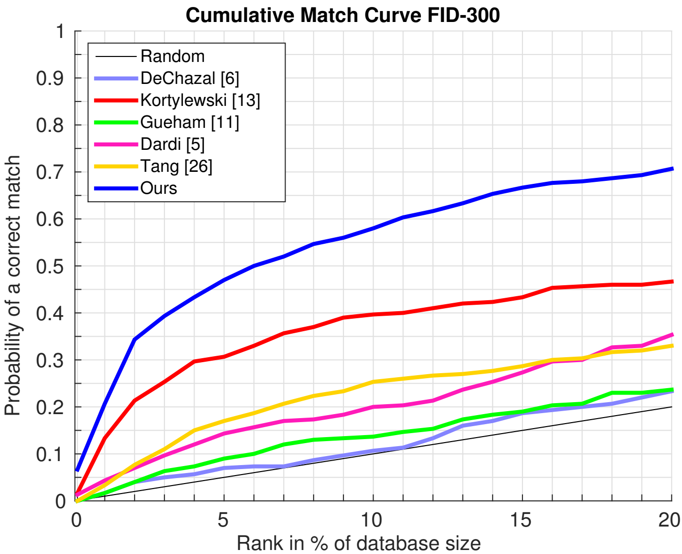


Figure 4: Image retrieval results on the FID-300 dataset.

According to the standard evaluation procedure, we sort the models based on their posterior probability and record the position of the correct reference from the ranked list. Afterwards, we calculate the cumulative distribution of the rank histogram. Figure 4 shows the cumulative match curves of our method compared to a reimplementation of five other approaches [2, 5, 11, 13, 27]. The section on the y-axis marks rank-1 performance. Compared to the other approaches the proposed method is able to increase the performance by a wide margin. We constantly outperform the state-of-the-art by approximately 15% starting from 3% of the ranked list.

5 Conclusion & Future Work

In this paper we propose an approach for learning the structure of compositional active basis models. We infer the number of layers per model as well as the number of parts in each layer with a greedy EM-type clustering process. Furthermore, we formulate the pattern model as well as the learning process in a fully probabilistic manner. Finally, based on the statistical framework, we augment the pattern model with an implicit background model that reduces the models sensitivity to pattern occlusions or structured clutter. We show that the proposed methodology is capable of solving the complex pattern analysis task of footwear impression recognition with yet unseen quality.
We think that part sharing between pattern models would open promising directions for future research, facilitating the learning of semantic regularities between patterns.

Acknowledgements. Part of this project was supported by the Swiss Commission for Technology and Innovation (CTI) project 16424.2 PFES-ES. We gratefully acknowledge the support of forensity ag and the German State Criminal Police Offices of Bavaria and Lower Saxony.

References

[1] Gharsa AlGarni and Madina Hamiane. A novel technique for automatic shoeprint image retrieval. *Forensic science international*, 181(1):10–14, 2008.

[2] Federico Cervelli, Francesca Dardi, and Sergio Carrato. A translational and rotational invariant descriptor for automatic footwear retrieval of real cases shoe marks. Eusipco, 2010.

[3] Jifeng Dai, Yi Hong, Wenze Hu, Song-Chun Zhu, and Ying Nian 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.

[4] Francesca Dardi, Federico Cervelli, and Sergio 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.

[5] Philip De Chazal, John Flynn, and Richard 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.

[6] Sanja Fidler and Aleš 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.

[7] Sanja Fidler, Marko Boben, and Ales Leonardis. Learning a hierarchical compositional shape vocabulary for multi-class object representation. *arXiv preprint arXiv:1408.5516*, 2014.

[8] Stuart Geman, Daniel F Potter, and Zhiyi Chi. Composition systems. *Quarterly of Applied Mathematics*, 60(4):707–736, 2002.

[9] Ross B Girshick, Pedro F Felzenszwalb, and David A Mcallester. Object detection with grammar models. In *Advances in Neural Information Processing Systems*, pages 442–450, 2011.

[10] Ulf Grenander. *Elements of pattern theory*. JHU Press, 1996.

[11] Mourad Gueham, Ahmed Bouridane, Danny Crookes, and Omar 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.

[12] Yi Hong, Zhangzhang Si, Wenze Hu, Song-Chun Zhu, and YING NIAN Wu. Unsupervised learning of compositional sparse code for natural image representation. *Quarterly of Applied Mathematics*, 72:373–406, 2013.

[13] Adam Kortylewski, Thomas Albrecht, and Thomas Vetter. Unsupervised footwear impression analysis and retrieval from crime scene data. In *Computer Vision-ACCV 2014 Workshops*, pages 644–658. Springer, 2014.

[14] Bo Li, Wenze Hu, Tianfu Wu, and Song-Chun Zhu. Modeling occlusion by discriminative and-or structures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2560–2567, 2013.

[15] Tapio Luostarinen and Antti Lehmussola. Measuring the accuracy of automatic shoeprint recognition methods. *Journal of forensic sciences*, 2014.

[16] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[17] Omar Nibouche, Ahmed 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.

[18] Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[19] Bjorn Ommer and Joachim M Buhmann. Learning the compositional nature of visual object categories for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):501–516, 2010.

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

[21] Maria Pavlou and Nigel M Allinson. Automatic extraction and classification of footwear patterns. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, pages 721–728. Springer, 2006.

[22] Maria Pavlou and Nigel M Allinson. Automated encoding of footwear patterns for fast indexing. *Image and Vision Computing*, 27(4):402–409, 2009.

[23] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19 (4):380–393, 1997.

[24] Zhangzhang Si and Song-Chun Zhu. Learning and-or templates for object recognition and detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35 (9):2189–2205, 2013.

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

[26] Jinli Suo, Song-Chun Zhu, Shiguang Shan, and Xilin Chen. A compositional and dynamic model for face aging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):385–401, 2010.

[27] Yi Tang, Sargur N Srihari, Harish Kasiviswanathan, and Jason J Corso. Footwear print retrieval system for real crime scene marks. In *Computational Forensics*, pages 88–100. Springer, 2011.

[28] Alex Wong and Alan Yuille. One shot learning via compositions of meaningful patches. In *ICCV*. 2015.

[29] Ying Nian Wu, Zhangzhang Si, Haifeng Gong, and Song-Chun Zhu. Learning active basis model for object detection and recognition. *International journal of computer vision*, 90(2):198–235, 2010.

[30] Zijian Xu, Hong Chen, Song-Chun Zhu, and Jiebo Luo. A hierarchical compositional model for face representation and sketching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):955–969, 2008.

[31] Alan L Yuille. Towards a theory of compositional learning and encoding of objects. In *ICCV Workshops*, pages 1448–1455. Citeseer, 2011.

[32] Long Leo Zhu, Chenxi Lin, Haoda Huang, Yuanhao Chen, and Alan Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *Computer vision–eccv 2008*, pages 759–773. Springer, 2008.

[33] Song-Chun Zhu and David Mumford. *A stochastic grammar of images*. Now Publishers Inc, 2007.