Learning the Compositional Nature of Visual Object Categories for Recognition

Björn Ommer, *Member*, *IEEE*, and Joachim M. Buhmann, *Senior Member*, *IEEE*

Abstract—Real-world scene understanding requires recognizing object categories in novel visual scenes. This paper describes a composition system that automatically learns structured, hierarchical object representations in an unsupervised manner without requiring manual segmentation or manual object localization. A central concept for learning object models in the challenging, general case of unconstrained scenes, large intraclass variations, large numbers of categories, and lacking supervision information is to exploit the compositional nature of our (visual) world. The compositional nature of visual objects significantly limits their representation complexity and renders learning of structured object models statistically and computationally tractable. We propose a robust descriptor for local image parts and show how characteristic compositions of parts can be learned that are based on an unspecific part vocabulary shared between all categories. Moreover, a Bayesian network is presented that comprises all the compositional constituents together with scene context and object shape. Object recognition is then formulated as a statistical inference problem in this probabilistic model.

Index Terms—Image categorization, object recognition, compositionality, graphical models, visual learning.

1 Introduction

Learning object models for detection and recognition poses one of the key challenges of computer vision. The complexity of this problem depends on several factors, such as the level of supervision during training, the degree of intraclass variabilities, and the constraints (e.g., constraints on variation in scale or viewpoint) that can be imposed on scenes. Despite these problems, learning of object representations from a small number of samples is possible due to the *compositional nature* of our (visual) world. As Attneave [1] points out, the visual stimulus is highly redundant in the sense that there exist significant spatial interdependencies in visual scenes. *Compositionality* (c.f. Geman's work [2]) serves as a fundamental principle in cognition and especially in human vision [3] that exploits these dependencies. It refers to the prominent ability of perception to represent complex entities by means of comparably few, simple, and widely usable parts. Additional information that is missing in the individual parts is added by incorporating relations between them. To this end, *perceptual organization* [4] provides a number of *Gestalt laws* that establish a basis for perceptually founded relations between parts. In contrast to modeling an object directly based on a constellation of its parts (e.g., [5]), the compositional approach learns intermediate groupings of parts—possibly even forming a hierarchy of recursive compositions. As a consequence, compositions bridge the semantic gap between low-level features and high-level object recognition by establishing intermediate representations.

In this paper, we investigate models for learning the compositional structure of objects and integrate them in a category-level object recognition system. The approach detects characteristic compositions of atomic parts for each category without supervision, requiring neither hand segmentations nor object localization. Learning higher level compositions of compositions then proceeds in a recursive manner. Finally, a Bayesian network serves as a coherent statistical model that comprises all the compositions together with object shape. Inference based on this probabilistic model yields a decomposition of a scene into relevant compositions, and finally, enables localization and recognition of objects. Moreover, the generative model of compositions can be used to sample object representations and explain away background clutter.

The *compositional object recognition model* from [6] realizes feature sharing on the lowest level where robust statistics are available. Therefore, edge and color distributions of small image patches are computed. A generic set of atomic parts that is shared among categories is established by forming a small codebook of these features. Category-specific relations between parts are used to build compositions, which are represented by probability distributions over their constituent parts, i.e., distributions over atomic parts yield compositions, distributions over compositions yield higher level compositions of compositions, and so on. Finally, a statistical, hierarchical scene representation is obtained which captures the spatial arrangement of all compositions. This *compositional shape model* couples all compositions by means of: 1) their spatial arrangement; 2) by forming relations between compositions that yield higher level compositions; and 3) by a co-occurrence of all compositions that roughly describes the context of the scene.

- *B. Ommer is with the Department of Electrical Engineering and Computer Science, University of California at Berkeley, 527 Soda Hall, Berkeley, CA 94720-1776. E-mail: ommer@eecs.berkeley.edu.*
- *J.M. Buhmann is with the Department of Computer Science, ETH Zurich, CAB G 69.2, Universitaetstrasse 6, 8092 Zurich, Switzerland. E-mail: jbuhmann@inf.ethz.ch.*

Manuscript received 18 Jan. 2008; revised 16 Aug. 2008; accepted 29 Dec. 2008; published online 15 Jan. 2009.
Recommended for acceptance by M. Figueiredo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-01-0040.
Digital Object Identifier no. 10.1109/TPAMI.2009.22.

2 Related Work and Key Modeling Decisions

Visual recognition can be pursued on different levels of semantic granularity. One extreme strategy is exemplar detection (e.g., [7]), where exactly the same query object is sought in scenes with different environmental conditions such as background, lighting, occlusion, viewpoint, etc. The other extreme is category-level object recognition, where all instances of a category are to be recognized. Therefore, the granularity of the set of categories controls the complexity of the recognition task as it defines the within-class variability (e.g., compare the task of finding all sports cars with that of finding cars of all types). Influential papers such as [8], [9] have focused research in the field of category-level object recognition on principled probabilistic object models with semilocal feature descriptors. The general goal is to represent objects by learning local appearance features and their spatial configuration and comprising both in a common model.

Within this coarse fundamental modeling framework, the current approaches to object categorization can be characterized by the core modeling decisions they make.

1. *Local descriptors.* A classical way to capture image region information is the use of *appearance patches*, i.e., subsampled image patches that are vector-quantized into a large codebook (e.g., [10], [5], [11], [12]). Another popular choice is *SIFT* features [7]. These are complex edge histogram features that have been proposed for exemplar detection. Nevertheless, they have also shown to yield acceptable performance in the task of categorization. In [13], we have proposed a low-dimensional representation of image patches that is based on compact local edge and color histograms of subpatches. The lack of specificity is compensated by capturing relations between the local descriptors. We use these *localized feature histograms* in this contribution. Such edge-based features are the limit case of neurophysiologically motivated *Gabor filters* (used, for instance, in [14]) with spatial filter width approaching zero. Another popular descriptor is *geometric blur* [15]. This feature weights edge orientations around a feature point using a spatially varying kernel. Moreover, edge contour-based methods have been proposed in [16], [17]. Opelt et al. [16] extract curve fragments from training images and they apply Adaboost to learn strong object detectors.
2. *Spatial model.* A second choice concerns the model that combines all local features with their spatial distribution to represent object shape. It should be emphasized that this notion of shape is not based on the object boundary but on the geometry of object parts that are distributed all over the object. This view of shape is common in the field of object categorization (e.g., [5], [18]). Object models have to deal with two problems simultaneously. On the one hand, individual local appearance descriptors in a test image are to be matched against those from a learned model. On the other hand, the co-occurence and the spatial relations between individual features have to be taken into account to represent the global object geometry. The simplest approach is therefore to histogram over all local descriptors found in an image (e.g., [19]) and categorize the image directly based on the overall feature frequencies. On the one hand, such *bag of features* methods offer robustness with respect to alteration of individual parts of an object (e.g., due to occlusion) at low computational costs. On the other hand, they fail to capture any spatial relations between local image patches and they often adapt to background features. By making the restricting assumption that the spatial structure of objects is limited in its variation with respect to the image, Lazebnik et al. [20] can improve the performance of the bag of features approach using a spatially fixed grid of feature bags. At the other end of the modeling spectrum, we find *constellation models*: Originally, Fischler and Elschlager [21] have proposed a spring model for coupling local features. Inspired by the *Dynamic Link Architecture* for cognitive processes, Lades et al. [22] followed the same fundamental idea when proposing their face recognizer. Lately, increasingly complex models for capturing part constellations have been proposed, e.g., [23], [5], [11], [24]. However, the complexity of such a joint model of all parts causes only small numbers of parts to be feasible. To incorporate larger numbers of parts, Agarwal et al. [10] and Leibe and Schiele [12] use a simpler object model and a comparably large codebook of distinctive parts. Leibe and Schiele [12] use a probabilistic Hough voting strategy to distinguish one category from the background. In [6], we advance the idea of large numbers of parts by grouping parts prior to spatially coupling the resulting compositions in a graphical model. Conflicting categorization hypotheses proposed by compositions and the spatial model are then reconciled using probabilistic inference in the underlying Bayesian network. Finally, Berg et al. [15] describe and regularize the spatial distortion resulting from matching an image to a training sample using thin-plate splines.
3. *Hierarchies.* For a long time, research on object recognition has aimed at building hierarchical models [25]. Despite this effort, many popular current methods, such as [12], [5], [19], are single layered. Recently, *probabilistic latent semantic analysis* (pLSA) [26] and *latent dirichlet allocation* [27] have become popular (e.g., [28], [29]), where a hidden representation layer of abstract concepts is introduced. Fergus et al. [29] have extended pLSA by incorporating spatial information. Other examples for hierarchical approaches are the feature hierarchies of [30], the hierarchical parts and structure model of [31], or the deep compositional hierarchies of [32].
4. *Learning paradigm.* Another modeling decision is related to the learning paradigm, i.e., pursuing a generative versus a discriminative approach. Although discriminative approaches have been shown to yield superior performance in the limited case of large training sets [33], generative models have been very popular in the vision community, e.g., [34], [35], [12], [36], [37], [10], [13], [15]. They naturally establish correspondences between model components and image features. Discriminative approaches are, for instance, [19] and [38]. To

TABLE 1
Summary of the Main Building Blocks of the Compositional Model and Their Mathematical Definition

Entity	Covered image region	Representation	Domain of Rand. Var.	Definition
• Atomic parts		40-dim edge & color hist (local descriptor/feature)	$\mathbf{e}_i \in [0,1]^{40}$	Eq. (1)
• Compositions		distrib over codebook $\mathcal{V}$ of atoms (bag of parts)	$\mathbf{g}_j \in [0,1]^k, k=\lvert\mathcal{V}\rvert$	Eq. (3)
• Compositions of compositions		tuples of comps. with spatial relation $\mathbf{r}_{kl} = \mathbf{x}_k - \mathbf{x}_l \in \mathbb{R}^2$, $\mathbf{x}_k$ denotes location of $\mathbf{g}_k$	$(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}) \in [0,1]^{\lvert\mathcal{V}\rvert} \times [0,1]^{\lvert\mathcal{V}\rvert} \times \mathbb{R}^2$	Sect. V-D
• Context		mixture of all comp. distributions (bag of comps.)	$\mathbf{g}^I \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$	Eq. (9)

recognize faces in real time, Viola and Jones [38] use boosting to learn simple features which are based on local intensity differences. Holub et al. [24] propose a hybrid approach using Fisher kernels.

5. *Degree of supervision:* Similar to the influential paper by Fergus et al. [5], several other approaches (e.g., [10], [19], [6]) have been proposed that only need training images (showing objects and even background clutter) and the overall category label of an image. The restriction of user assistance is desirable for scaling methods up to large numbers of categories with large training sets. A system that can be trained in an unsupervised manner is, for instance, that of [39], whereas Felzenszwalb and Huttenlocher have taken a supervised approach to object detection in [40]. Furthermore, Jin and Geman [41] present a compositional architecture with manually built structure for license plate reading. In their conclusion, they emphasize the complexity of the future challenge of learning such a compositional model. This contribution deals with exactly this problem in the even less constraint case of large numbers of natural object classes.

3 Terminology and Outline of the Framework for Recognition

Part-based object models represent objects using a set of *parts*. These parts represent local regions of an image and are therefore commonly referred to as *descriptors* or *features*. The spatial arrangement of these parts can be either fixed [20] or flexible [40] and the number of parts can also be predefined [5] or variable [18]. All of these approaches have in common that they strive for very specific part descriptors that act as fingerprints of objects—local regions that are highly discriminative for object classes. In this paper, we follow an orthogonal approach: We start off with generic parts that can be shared among categories and are thus not category-specific. To compensate for discriminative information that is lacking in the individual parts, object representations are based on *compositions* of parts. Compositions are carrying additional information by virtue of the relations they establish between parts. Thus, the underlying idea is to avoid complex, highly specific part descriptors and rather learn to automatically form characteristic groupings of generic parts. Since compositions are themselves also grouped to form compositions of compositions (a higher level in the resulting hierarchical object representation), we call the initial, local descriptors *atomic* parts to express that they are not further decomposable. Table 1 summarizes the building blocks of the compositional model.

Fig. 2 shows examples where the compositional approach succeeds, while simpler models would exhibit limitations: Fig. 2a shows that recognition is possible based on generic parts when they are used in compositions. Fig. 2b visualizes the effectiveness of compositions of compositions, whereas Fig. 2c demonstrates the value of context combined with local appearance. Finally, in Fig. 2d, the flexible compositional shape model can handle changes in viewpoint, while rigid representations such as [20] would fail to model the shape of the object.

Let us now briefly overview the approach to compositional scene analysis (illustrated in Fig. 1) before presenting the individual processing steps of the recognition algorithm (Algorithm 2) and the learning of the underlying model (Algorithm 3) in detail in later sections. For a novel image, a set $\mathcal{E}^{(\sigma)}$ of small patches, the atomic parts $\mathbf{e}_i \in \mathcal{E}^{(\sigma)}$, is extracted on different scales σ at interest points. For each of them, localized feature histograms [13] are computed as local descriptors. Thereafter, vector quantization is performed to represent parts by distributions over a small codebook of feature prototypes which is shared by all objects. As patches are only local features and the codebook is shared by all categories, these atomic parts alone are far from being category-specific. To compensate for this lack of information, compositions $\mathbf{g}_j \in \mathcal{G}^{(\sigma)}$ of these parts are established subsequently.

We develop a learning algorithm that automatically learns to establish relevant compositions rather than

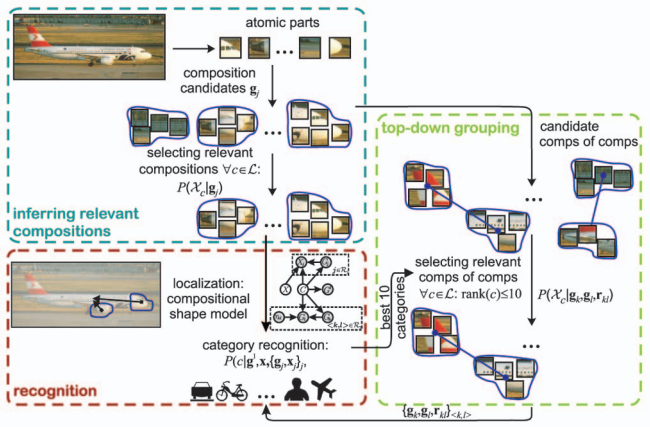

Fig. 1. Processing pipeline for automatic scene analysis. Key steps: Feature extraction, perceptual grouping to form compositions, selection of relevant compositions, object localization and recognition, and top-down grouping to form compositions of compositions which then yield an update of object hypotheses.

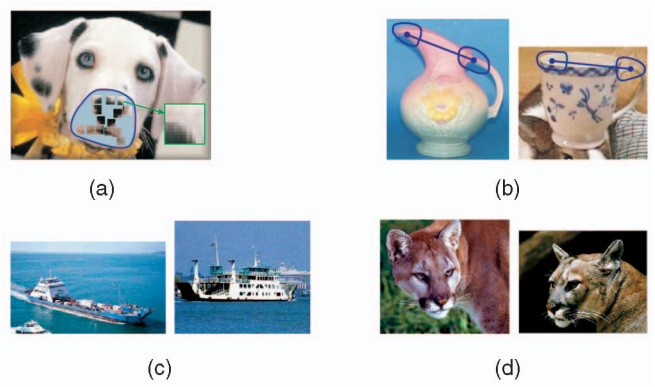

Fig. 2. Advantages of the compositional approach over other representation schemes. (a) Simple, generic, unspecific parts suffice when used in compositions. (b) Compositions of compositions help to distinguish categories based on constituents that are not characteristic in isolation. (c) Context provides valuable information when object classes are visually diverse. (d) The flexible compositional shape model helps to recognize the shape of both cougar faces where a rigid object model such as a rigid grid of histograms [20] fails to capture that structure.

manually modeling a set of grouping laws (c.f. [32]) that lead to characteristic compositions. Hence, we employ a simple, proximity-based grouping to form candidate compositions $\mathcal{G}^{(\sigma)}$. Thereby, the search space for the subsequent relevance learning is restricted. Out of these candidate compositions, relevant compositions are selected by a relevance model that has been learned during the training phase (Algorithm 3). The relevant compositions are then coupled in a Bayesian network (the *compositional shape model*) and the category posterior is computed.

At this stage, a large fraction of all possible object categorization hypotheses can already be rejected with high confidence. Conditioned on each of the remaining hypotheses, we seek a set $\widetilde{\mathcal{R}}$ of relevant compositions of compositions (the constituents are compositions themselves) to accumulate additional evidence for the correct hypothesis. This procedure defines a top-down grouping process which is guided by previously inferred object information. The newly generated compositions enter then into the Bayesian network together with the compositions from before to refine the categorization hypothesis.

4 Atomic Parts for Compositionality

The compositional approach learns hierarchical object representations that are based on groupings of constituent parts. The features representing the atomic parts in the initial layer of the compositional hierarchy should exhibit:

1. good localization,
2. robustness to local image changes,
3. low dimensionality, and
4. they should be shareable by object categories.

4.1 Localized Feature Histograms

A classical representation of image content is given by local appearance patches which have been widely used in the field of object categorization, e.g., [10], [18], [5]. However, due to the global subsampling and intensity normalization, translations or local distortions still corrupt the feature. Moreover, the low-pass filtering retains only the strongest edges while blurring the remaining patch content. An alternative approach at the other end of the modeling spectrum is that of using histograms over complete images (c.f. [42]). In summary, the former approach facilitates almost perfect localization, while the latter one offers maximal invariance with respect to local distortions. As a compromise between these two opposing goals, we aim at a representation whose invariance properties are transparently adjusted between these two classical extremes and add the specificity lost by invariance through the relations incorporated in compositions.

To process an image, quadratic patches of size 20×20 pixels are extracted at interest points (obtained using the Harris detector of [43]). Each patch is divided up into four equally sized subpatches with locations fixed relative to the patch center, see Fig. 3. In each of these subwindows, $l = 1, \ldots, 4$, marginal histograms over edge orientation and edge strength are computed (allocating four bins to each of them), denoted as $\mathbf{e}_i^{(o_l)}, \mathbf{e}_i^{(s_l)} \in [0,1]^4$. Furthermore, an 8-bin color histogram $\mathbf{e}_i^{(c)} \in [0,1]^8$ over all subpatches is extracted. All these histograms are then combined in a 40D ($4 \times 4 + 4 \times 4 + 8 = 40$) feature vector:

$$\mathbf{e}_i := \left(\mathbf{e}_i^{(o_1)}, \ldots, \mathbf{e}_i^{(o_4)}, \mathbf{e}_i^{(s_1)}, \ldots, \mathbf{e}_i^{(s_4)}, \mathbf{e}_i^{(c)}\right)^T. \quad (1)$$

The proposed representation differs from SIFT features [7] not only in that color is used. Whereas SIFT features aim at distinguishing different instances of the same object from another, we seek a representation that is invariant to the specificities of individual object instances and environment configurations. To obtain a small codebook of atomic representatives for compositionality (Section 4.2), we propose local descriptors of reduced complexity (40D), whereas the other approach would have to perform this reduction of complexity indirectly by clustering in a high-dimensional space (128D) with few prototypes. An experimental comparison of localized histograms with SIFT and an analysis of the effective dimensionality of both descriptor types is presented in Section 7.2.

Features on multiple image scales. Features on several image scales $\sigma = \{1, \frac{1}{2}, \frac{1}{4}\}$ are computed by rescaling the image and extracting the features in the resized image as described above. For $\sigma = 1/2$, for instance, this corresponds to subsampling to half the original scale. All features on the same scale are summarized in a set $\mathcal{E}^{(\sigma)}$ (see line 4 of Algorithm 1).

Algorithm 1. Extracting composition candidates from images

ALLCOMPOSITIONCANDIDATES(I) ▷ *I: a novel test image*
1 $\mathcal{P} \leftarrow$ INTERESTPOINTS(I) ▷ *Harris detector of [43]*
2 **for** all scales $\sigma \in \mathcal{S}$
3 **do** $\mathcal{E}^{(\sigma)} \leftarrow$ ATOMICPARTS($I, \mathcal{P}, \sigma$)
▷ $\mathcal{E}^{(\sigma)}$ *is a set of* $\mathbf{e}_i$ *on scale* σ, *(1)*
4 $\mathcal{G}^{(\sigma)} \leftarrow$ COMPOSITIONCANDIDATES ($\mathcal{E}^{(\sigma)}$)
▷ $\mathcal{G}^{(\sigma)} = \{\mathbf{g}_j : \mathbf{g}_j$ from scale $\sigma\}$, *c.f. Section 5.1*
5 $\mathcal{G} \leftarrow \bigcup_{\sigma \in \mathcal{S}} \mathcal{G}^{(\sigma)}$
6 $\mathbf{g}^I \leftarrow \frac{1}{|\mathcal{G}|} \sum_{\mathbf{g}_j \in \mathcal{G}} \mathbf{g}_j$ ▷ *context descriptor from (9)*
7 **return** $\mathcal{G}, \mathbf{g}^I$

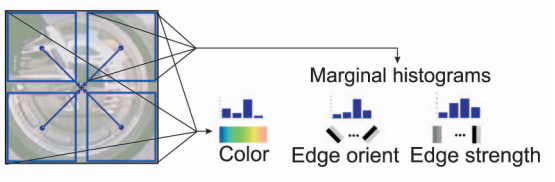

Fig. 3. Sketch of localized feature histograms.

4.2 A Codebook of Atomic Parts

To obtain a common representation for compositions consisting of different numbers of parts, a part codebook is established that is shared by all categories. Therefore, part descriptors $\mathbf{e}_i$ detected in the training images of all object classes are vector-quantized using k-means clustering yielding a $k = 200$D codebook $\mathcal{V}$, $k = |\mathcal{V}|$. To make the representation robust, each part is described by a Gibbs distribution [44] over the codebook. Let $d_\nu(\mathbf{e}_i)$ denote the squared euclidean distance of $\mathbf{e}_i$ to a centroid $\mathbf{a}_\nu \in \mathcal{V}$. The local region is then encoded by the following distribution of its cluster assignment random variable F_i:

$$P(F_i = \nu|\mathbf{e}_i) := Z(\mathbf{e}_i)^{-1} \exp(-d_\nu(\mathbf{e}_i)),$$
$$Z(\mathbf{e}_i) := \sum_{1\le\nu\le k} \exp(-d_\nu(\mathbf{e}_i)). \tag{2}$$

In summary, we are first extracting edge and color histograms for local regions. Thereafter, these descriptors are vector-quantized based on a small codebook of part descriptors. Thus, parts as well as compositions can be encoded by distributions over the codebook.

Algorithm 2. Compositional scene analysis (single object per image)

OBJECTRECOGNITION(I) ▷ *I: a novel query image*
▷ *extract atomic parts and form composition candidates:*
1 $(\mathcal{G}, \mathbf{g}^I) \leftarrow$ ALLCOMPOSITIONCANDIDATES(I) ▷*Algo. 1*
2 $\mathbf{x} \leftarrow \frac{\sum_{j:\mathbf{g}_j\in\mathcal{G}} \mathbf{x}_j \sum_{c\in\mathcal{L}} p(\mathbf{g}_j|c,\mathbf{g}^I)P(c|\mathbf{g}^I)}{\sum_{j:\mathbf{g}_j\in\mathcal{G},c\in\mathcal{L}} p(\mathbf{g}_j|c,\mathbf{g}^I)\,P(c|\mathbf{g}^I)}$
▷ *estimate object center using (17)*
3 $\mathcal{R} \leftarrow$ RELEVANTCOMPOSITIONS $(\mathcal{G}, \mathbf{x})$
▷ *select relevant compositions using (11)*
4 $P(c|E) \leftarrow$ COMPOSTTIONALSHAPEMODEL $(\mathcal{R}, \mathbf{g}^I, \mathbf{x}, \emptyset)$
▷ *recognition using (16)*
▷ *draw compositions of compositions for most likely categories:*
5 $\widetilde{\mathcal{G}} \leftarrow$ COMPOFCOMPCANDIDATES $(\mathcal{G})$
▷ *random subset of $\mathcal{G} \times \mathcal{G}$, Section 5.4*
6 $\widetilde{\mathcal{R}} \leftarrow$ RELEVANTCOMPSOFCOMPS $(\widetilde{\mathcal{G}}, P(c|E))$
▷ *select relevant comps of comps, (12)*
▷ *update previous object category hypothesis by using comps. $\mathcal{R}$ and higher order comps. $\widetilde{\mathcal{R}}$ for recognition:*
7 $P(c|E) \leftarrow$ COMPOSITIONALSHAPEMODEL $(\mathcal{R}, \mathbf{g}^I, \mathbf{x}, \widetilde{\mathcal{R}})$
▷ *recognition, (16)*
8 **return** $P(c|E), \mathbf{x}$ ▷ *object category posterior and location*

5 Learning a Hierarchy of Relevant Compositions on Multiple Scales

Subsequently, we present an approach that automatically learns to build category-specific compositions without requiring prior knowledge regarding the compositional nature of objects.

5.1 Composition Candidates

On each scale σ of an image, 40 parts are selected from the set $\mathcal{E}^{(\sigma)}$ of all atomic parts that have been extracted as described in Section 4. Around each of these parts, all parts in a local neighborhood of three local patches are grouped together. This size offers a good trade-off between localization of compositions and the statistical robustness of their estimation during training. The proximity grouping yields unordered agglomerations of atomic parts. We call these sets of parts *composition candidates* $\mathbf{g}_j$ and the set of all candidates on scale σ is $\mathcal{G}^{(\sigma)}$ (line 4 of Algorithm 1).

Equation (2) provides a representation of atomic parts $\mathbf{e}_i$ based on a probability distribution $(P(F_i = 1|\mathbf{e}_i), \ldots, P(F_i = k|\mathbf{e}_i)) \in [0,1]^k$ over the k-dimensional part codebook $\mathcal{V}$. Let $\Gamma_j = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ denote the grouping of parts $\mathbf{e}_1, \ldots, \mathbf{e}_m$. The number of constituents $|\Gamma_j| = m$ is not predefined since the grouping is defined by the grouping diameter and not by the number of parts. Compositions are then represented by the multivariate random variable G_j. Realizations $\mathbf{g}_j \in [0,1]^k$ of this random variable are again distributions over the part codebook $\mathcal{V}$. The distribution that represents a composition $\mathbf{g}_j$ is a mixture of the distributions of all the parts $\mathbf{e}_i \in \Gamma_j$:

$$\mathbf{g}_j = \sum_{\mathbf{e}_i\in\Gamma_j} \frac{1}{|\Gamma_j|} (P(F_i = 1 \mid \mathbf{e}_i), \ldots, P(F_i = k \mid \mathbf{e}_i))^\top. \tag{3}$$

Finally, each of the k dimensions of compositions is independently standardized to zero mean and unit variance across the whole training set, giving *z-scores*. This mixture model is robust w.r.t. missed or corrupted individual local parts, since it leverages an ensemble of parts. Second, it is invariant to shifts of the local parts in the composition, which result from fluctuations in the interest point detection. Third, the mixture model exhibits low dimensionality irrespective of the number of constituents that are grouped together.

5.2 Learning Relevant Compositions of Parts

A significant number of composition candidates do actually only capture clutter such as background or other unspecific regions of the scene and should thus be discarded. Subsequently, an approach will be developed that automatically learns to retrieve those compositions which are actually relevant for our task of recognizing object categories. Therefore, we present a Bayesian criterion that defines what relevant compositions are and we propose a statistically feasible learning algorithm.

From a Bayesian point of view, a composition $\mathbf{g}_j$ that is drawn from an image I is relevant for representing objects of some category c if it has a high likelihood $p(\mathbf{g}_j \mid \chi_c)$. The indicator function $\chi_c(I) \in \{0,1\}$ specifies whether an image I contains an object of category c:

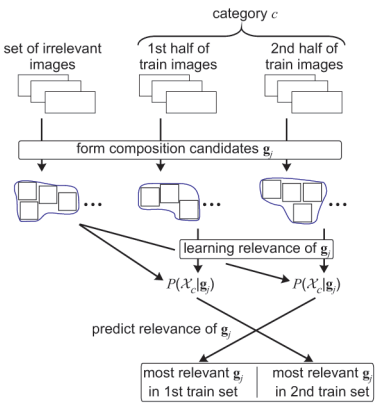

Fig. 4. Learning relevant compositions.

$$\chi_c(I) := \begin{cases} 1, & I \text{ contains an object of category } c, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

By applying Bayes' theorem, the likelihood factorizes

$$P(\chi_c \mid \mathbf{g}_j) = \frac{p(\mathbf{g}_j \mid \chi_c)\, P(\chi_c)}{p(\mathbf{g}_j)}. \quad (5)$$

Since all categories are a priori equally likely, $P(\chi_c)$ can be absorbed in a normalization constant:

$$p(\mathbf{g}_j \mid \chi_c) \propto P(\chi_c \mid \mathbf{g}_j)\, p(\mathbf{g}_j). \quad (6)$$

Assume for the moment that we have already estimated the object location $\mathbf{x}$ (c.f. (17)). In this case, the estimate of compositional relevance from (6) can be refined by incorporating the relative position of the object center w.r.t. to the location $\mathbf{x}_j$ of a composition, $S_j = \mathbf{x} - \mathbf{x}_j$:

$$\begin{aligned} p(\mathbf{g}_j \mid \chi_c, S_j = \mathbf{x} - \mathbf{x}_j) \propto\ & P(\chi_c \mid \mathbf{g}_j, S_j = \mathbf{x} - \mathbf{x}_j) \\ & \times\ p(\mathbf{g}_j \mid S_j = \mathbf{x} - \mathbf{x}_j). \end{aligned} \quad (7)$$

Compositional relevance as defined in (6) and (7) factorizes into two distributions. The first expresses the discriminative power of a composition $\mathbf{g}_j$, whereas the second indicates how reliably $\mathbf{g}_j$ can be detected. While the first distribution can be estimated using discriminative learning, the second one requires a density estimation in a high-dimensional feature space which we avoid by means of a cross-validation-based approach (illustrated in Fig. 4). Therefore, the posterior $P(\chi_c \mid \mathbf{g}_j, \mathbf{s}_j)$ is learned on one part of the training data before using it to predict the relevance of compositions in the other part. Unfavorable compositions with low prior $p(\mathbf{g}_j \mid \mathbf{s}_j)$ also have a low probability of appearing in the validation set so that validation prevents overfitting to the training set. The learning algorithm, which is summarized in Algorithm 3, starts by randomly splitting $\mathcal{T}_c^{(+)}$, the training images of category c, into two disjoint subsets $\mathcal{T}_c^{(1)}$ and $\mathcal{T}_c^{(2)}$ of equal size. Moreover, a set of irrelevant compositions $\mathcal{T}_c^{(0)}$ has to be established by taking a random sample of compositions from all categories other than c. Then, a probabilistic classifier is trained to distinguish compositions in $\mathcal{T}_c^{(1)}$ from the irrelevant ones in $\mathcal{T}_c^{(0)}$ and it yields an estimate of $P(\chi_c \mid \mathbf{g}_j, \mathbf{s}_j)$, *line 7*. For classification, we use *nonlinear kernel discriminant analysis* (NKDA) [45], a kernelized version of linear discriminant analysis. However, experiments with SVMs have shown the same performance so that the choice of this particular classifier is not crucial. To discard erroneously detected compositions in $\mathcal{T}_c^{(1)}$ (e.g., background or other objects), *line 8*, we use this classifier to predict the relevance of compositions from the other training subset $\mathcal{T}_c^{(2)}$, which thereby act as a validation set. Given this ranking, the subset $\mathcal{R}_2 \subset \mathcal{G}_2$ of cardinality ρ with the highest relevance is selected from all compositions of the validation set (ρ is set to retain 50 percent of the original compositions). In *line 11*, the relevant subsets of each half of the training data are merged to train a single NKDA classifier that is used to measure compositional relevance in novel images.

Algorithm 3. Algorithm for learning relevant compositions from cluttered images

RELEVANCELEARNING $(\mathcal{T}_c^{(+)}, \mathcal{T}_c^{(0)})$
▷ $\mathcal{T}_c^{(+)}$: *set of training images for category c.*
▷ $\mathcal{T}_c^{(0)}$: *irrelevant images of other categories*
1 $(\mathcal{T}_c^{(1)}, \mathcal{T}_c^{(2)}) \leftarrow$ SPLITTRAINSET $(\mathcal{T}_c^{(+)})$
2 **for** $i \in \{0,1,2\}$ ▷ *collect comps for images in the train subsets:*
3 **do** $\mathcal{G}_i = \emptyset$
4 **for** $I \in \mathcal{T}_c^{(i)}$
5 **do** $\mathcal{G}_i \leftarrow \mathcal{G}_i \cup$ ALLCOMPOSITIONCANDIDATES(I) ▷ *Algo. 1*
6 **for** $i \in \{1,2\}$
7 **do** $p(\chi_c|\mathbf{g}_j, \mathbf{s}_j) \leftarrow$ LEARNPROBCLASSIFIER $(\mathcal{G}_0, \mathcal{G}_i)$ ▷ *Train on i-th train set*
8 $p(\chi_c|\mathbf{g}_j, \mathbf{s}_j)\big|_{\mathbf{g}_j \in \mathcal{G}_{\{1,2\}-i}} \leftarrow$ PREDICT $(p(\chi_c|\mathbf{g}_j, \mathbf{s}_j), \mathcal{G}_{\{1,2\}-i})$ ▷ *Predict on other half*
9 $\mathcal{R}_{\{1,2\}-i} \leftarrow$ subset (cardinality ρ) of $\mathcal{G}_{\{1,2\}-i}$ with highest $p(\chi_c|\mathbf{g}_j, \mathbf{s}_j)\big|_{\mathbf{g}_j \in \mathcal{G}_{\{1,2\}-i}}$
10 $\mathcal{R} \leftarrow \mathcal{R}_1 \cup \mathcal{R}_2$ ▷ *Learn relevance on both parts of the train set:*
11 $p(\chi_c|\mathbf{g}_j, \mathbf{s}_j) \leftarrow$ LEARNPROBCLASSIFIER $(\mathcal{G}_0, \mathcal{R})$
12 **return** $p(\chi_c|\mathbf{g}_j, \mathbf{s}_j), \mathcal{R}$

5.3 Visualizing Relevant Compositions

Now, we visualize the compositions that have been learned to be relevant. Therefore, all candidate compositions of the training data that have been predicted to be relevant for a category c are clustered using histogram clustering (using the histogram clustering of [46]). The relevances of all compositions that are assigned to a centroid are averaged and the centroids with the highest relevance for the category are presented in Fig. 5. We depict each centroid by plotting the three closest representatives that have been assigned to it during clustering. Each of these three compositions is visualized by displaying the image patches from the training image that constituted the respective composition.

5.4 Learning Higher Order Compositions

To incorporate additional information into object models, direct dependencies between compositions can be captured by learning groupings of compositions. To build these higher order compositions, we select random tuples of compositions $(\mathbf{g}_k, \mathbf{g}_l)$ and measure relations $\mathbf{r}_{kl}$ between them. Our current model uses only the distance vector $\mathbf{r}_{kl} = \mathbf{x}_k - \mathbf{x}_l$, although various other kinds of perceptual relationships are conceivable. The relevance score (6) is then adapted:

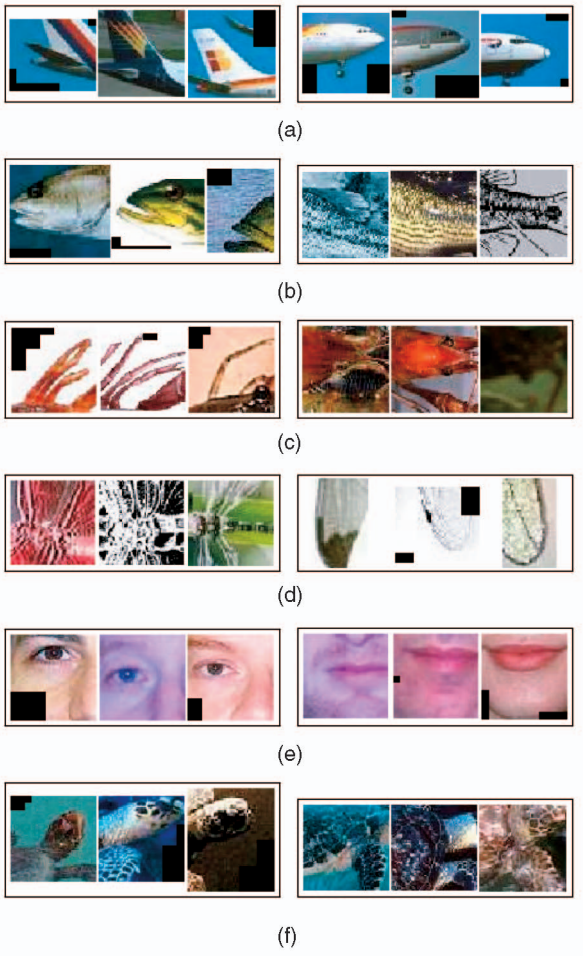

Fig. 5. Clustering of relevant compositions. For each category, the two centroids with the highest relevance are shown by visualizing the three closest compositions to that prototype. (a) Airplanes. (b) Bass. (c) Crayfish. (d) Dragonfly. (e) Faces. (f) Hawksbill.

$$p(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl} \mid \chi_c) \propto P(\chi_c \mid \mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl})\, p(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}), \quad (8)$$

and plugged into Algorithm 3 to learn the relevance of higher order compositions as in Section 5.2.

6 Object Classification and Detection Using a Compositional Shape Model

Subsequently, object classification and detection will be formulated as two coupled inference problems that are combined in a single statistical model and are solved alternatingly.

6.1 Binding Compositions in a Compositional Shape Model for Object Classification

Classification of an object requires that all compositions are combined in a single object model so that a concerted categorization hypothesis can be derived. This structured object model should couple: 1) the appearance of salient object regions, 2) object shape (that is the geometry of the local features), and 3) the scene context in which the object appears (c.f. Fig. 6a). To this end, we utilize a Bayesian network, which is illustrated in Fig. 6b. In this compositional shape model, the appearance of object regions is covered by compositions, whereas the shape is represented

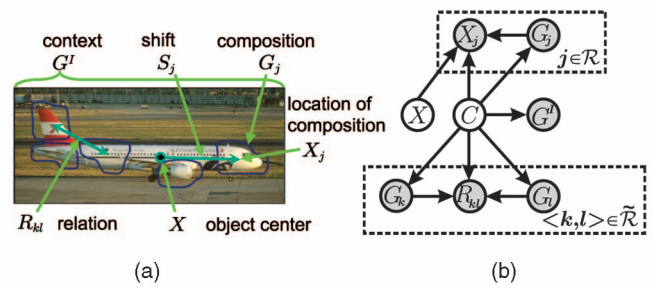

Fig. 6. (a) Illustration of the different constituents of the compositional shape model. (b) Bayesian network that couples compositions G_j by means of scene context G^I, object shape (represented by shifts between composition locations X_j and object center position X), relations R_{kl} between compositions, and finally, object categorization C.

through spatial relations between compositions and the object center and by means of higher order compositions. Last, scene context is captured by the co-occurence of all compositions $\mathbf{g}_j \in \mathcal{G}$. Therefore, a mixture $\mathbf{g}^I$ of the distributions of all $\mathbf{g}_j$ is computed:

$$\mathbf{g}^I := \frac{1}{|\mathcal{G}|} \sum_{\mathbf{g}_j \in \mathcal{G}} \mathbf{g}_j. \quad (9)$$

Assume for the moment that an estimate of the object center $\mathbf{x}$ has been established. Object recognition does then amount to finding the category $c \in \mathcal{L}$ that maximizes the posterior

$$P(c \mid \mathbf{g}^I, \mathbf{x}, \{\mathbf{g}_j, \mathbf{x}_j\}_{\mathbf{g}_j \in \mathcal{R}}). \quad (10)$$

6.1.1 Selecting Relevant Compositions

The set of relevant compositions $\mathcal{R}$ is formed by retaining those $\varrho = 50\%$ of all compositions $\mathbf{g}_j \in \mathcal{G}$ with the highest relevance score (7), i.e.,

$$\begin{aligned} \mathcal{R} := \mathcal{A} : \mathcal{A} \subset \mathcal{G} \ \wedge \ |\mathcal{A}| \approx \varrho \cdot |\mathcal{G}| \\ \wedge \forall \mathbf{g}_j \in \mathcal{A}, \mathbf{g}_{j'} \in \mathcal{G} - \mathcal{A} : \\ \max_{c \in \mathcal{L}} P(\chi_c \mid \mathbf{g}_j, \mathbf{s}_j) \geq \max_{c \in \mathcal{L}} P(\chi_c \mid \mathbf{g}_{j'}, \mathbf{s}_{j'}). \end{aligned} \quad (11)$$

Relevant higher order compositions are selected using a top-down grouping. Candidate composition tuples are formed by randomly drawing from $\mathcal{G} \times \mathcal{G}$. The most relevant ones are then selected by maximizing the relevance (8). This time, however, we can make use of the category hypotheses that have been established based on singletons (10). Therefore, we look only for those tuples $(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}$ that are most relevant for the 10 categories with the highest posterior (10). Let $\mathcal{L}_\mathcal{R} \subset \mathcal{L}, |\mathcal{L}_\mathcal{R}| = 10$ denote the 10 most likely categories given the posterior (10):

$$\begin{aligned} \widetilde{\mathcal{R}} := \mathcal{A} : \ \mathcal{A} \subset \mathcal{G} \times \mathcal{G} \ \wedge \ |\mathcal{A}| \approx \frac{1}{2}|\mathcal{G}| \\ \wedge \forall (\mathbf{g}_k, \mathbf{g}_l) \in \mathcal{A}, (\mathbf{g}_{k'}, \mathbf{g}_{l'}) \in (\mathcal{G} \times \mathcal{G}) - \mathcal{A} : \\ \max_{c \in \mathcal{L}_\mathcal{R}} P(\chi_c \mid \mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}) \geq \max_{c \in \mathcal{L}_\mathcal{R}} P(\chi_c \mid \mathbf{g}_{k'}, \mathbf{g}_{l'}, \mathbf{r}_{k'l'}). \end{aligned} \quad (12)$$

Consequently, singleton compositions focus the search for higher order compositions. The higher order compositions are then used to confirm the correct category hypothesis.

We have chosen a subset of 10 categories since the correct class is among this set in more than 90 percent of all cases.

6.1.2 Object Recognition Using Statistical Inference

Now, the category posterior conditioned on all compositions and higher order compositions is derived—the posterior for singletons is then a special case of this distribution. Let us abbreviate the posterior by using E as a shorthand notation for the collected evidence. We start by applying Bayes' formula:

$$P(c|E) = P(c \mid \mathbf{g}^I, \mathbf{x}, \{\mathbf{g}_j, \mathbf{x}_j\}_{\mathbf{g}_j \in \mathcal{R}}, \{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}}) = \frac{p(\{\mathbf{g}_j, \mathbf{x}_j\}_{\mathbf{g}_j \in \mathcal{R}}, \{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}}, \mathbf{g}^I \mid \mathbf{x}, c) P(c \mid \mathbf{x})}{p(\{\mathbf{g}_j, \mathbf{x}_j\}_{\mathbf{g}_j \in \mathcal{R}}, \{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}}, \mathbf{g}^I \mid \mathbf{x})}. \tag{13}$$

Now, we neglect the evidence in the denominator as it is independent of c. Moreover, we factorize the enumerator by exploiting that compositions are independent conditioned on parameters $c, \mathbf{x}$:

$$\begin{aligned} P(c \mid E) &\propto p\big(\{\mathbf{g}_j, \mathbf{x}_j\}_{\mathbf{g}_j \in \mathcal{R}} \mid \mathbf{x}, c\big) \\ &\quad \times p\big(\{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}} \mid \mathbf{x}, c\big)\; p(\mathbf{g}^I \mid \mathbf{x}, c) P(c \mid \mathbf{x}) \\ &= P(c \mid \mathbf{x}) \cdot p(\mathbf{g}^I \mid \mathbf{x}, c) \times \prod_{\mathbf{g}_j \in \mathcal{R}} p(\mathbf{g}_j, \mathbf{x}_j \mid \mathbf{x}, c) \\ &\quad \times \prod_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}} p(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl} \mid \mathbf{x}, c). \end{aligned} \tag{14}$$

Combining the first two factors and applying Bayes' rule to the individual likelihoods yields

$$\begin{aligned} P(c \mid E) &\propto P(c, \mathbf{g}^I \mid \mathbf{x}) \times \prod_{\mathbf{g}_j \in \mathcal{R}} \frac{P(c \mid \mathbf{x}, \mathbf{g}_j, \mathbf{x}_j) p(\mathbf{g}_j, \mathbf{x}_j \mid \mathbf{x})}{p(c \mid \mathbf{x})} \\ &\quad \times \prod_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}} \frac{P(c \mid \mathbf{x}, \mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}) p(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl} \mid \mathbf{x})}{p\big(c \mid \mathbf{x}\big)}. \end{aligned} \tag{15}$$

Factors that are independent of c can be neglected. Moreover, the object class is independent of the absolute position of the object in the image. Therefore, only the relative positions of compositions w.r.t. the object center are retained since these shifts $\mathbf{s}_j = \mathbf{x} - \mathbf{x}_j$ represent object shape. Thus, we obtain,

$$\begin{aligned} P(c \mid E) &\propto P(c \mid \mathbf{g}^I, \mathbf{x}) \cdot p(\mathbf{g}^I \mid \mathbf{x}) \times \prod_{\mathbf{g}_j \in \mathcal{R}} P(c \mid \mathbf{x}, \mathbf{g}_j, \mathbf{x}_j) \\ &\quad \times \prod_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}} P(c \mid \mathbf{x}, \mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}) \\ &\propto \exp\Bigg[\ln P(c \mid \mathbf{g}^I) + \sum_{\mathbf{g}_j \in \mathcal{R}} \ln P(c \mid \mathbf{g}_j, S_j = \mathbf{x} - \mathbf{x}_j) \\ &\quad + \sum_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}} \ln P(c \mid \mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl})\Bigg]. \end{aligned} \tag{16}$$

Here, the logarithm has been introduced for numerical stability. Computing (10) is then a special case of this formula where the last sum over $\widetilde{\mathcal{R}}$ has been omitted. The individual distributions of (13) are all estimated using a probabilistic classifier (we use NKDA).

6.2 An Initial Estimate of Object Location

Subsequently, a first estimate of the object center is computed. Therefore, $\mathbf{g}^I$ from (9) is employed to bind compositions based on their co-occurrence. The object center is then estimated by weighing the contribution of each composition with the probability that it should be observed:

$$\mathbf{x} = \frac{\sum_j \mathbf{x}_j \sum_{c \in \mathcal{L}} p(\mathbf{g}_j \mid c, \mathbf{g}^I)\; P(c \mid \mathbf{g}^I)}{\sum_{j,c} p(\mathbf{g}_j \mid c, \mathbf{g}^I) P(c \mid \mathbf{g}^I)}. \tag{17}$$

The first distribution is estimated using Parzen windows and the second one using NKDA. In the training phase, when the true category label is available for images, the second sum reduces to the true category c and the distribution over categories degenerates to a discrete Dirac distribution:

$$\mathbf{x} = \frac{\sum_j \mathbf{x}_j \cdot p(\mathbf{g}_j \mid c_{\text{true}}, \mathbf{g}^I)}{\sum_j p(\mathbf{g}_j \mid c_{\text{true}}, \mathbf{g}^I)}. \tag{18}$$

An evaluation on the Caltech-101 database shows that the estimate for the object center in (17) deviates from the true center (provided by the hand annotations) by 8.8 ± 3.8 percent of the bounding box diagonal (averaged over all categories). This is roughly the size of the atomic parts, and therefore, exact enough to couple compositions in the compositional shape model.

6.3 Simultaneous Classification and Detection of Multiple Objects

6.3.1 Recognition Phase

Now, we extend the compositional approach so that it can localize the bounding boxes of multiple objects in heavily cluttered scenes, i.e., the scenario defined in the *PASCAL Visual Object Classes Challenge 2006 (VOC '06)* [47]. The following extends the recognition procedure from Algorithm 2 so that multiple objects per image can be handled under the assumption that there is maximally one instance per category. Bounding box hypotheses are inferred for each category and a confidence in each hypothesis is computed—this is carried out in the main loop starting at line 2 of Algorithm 4. To start the alternating localization and classification of objects, an initial bounding box estimate is computed by weighting the location of each composition with the compositional relevance $P(\chi_c \mid \mathbf{g}_j)$ from (6). The bounding box is a square which is determined by its center $B^c_{\mathbf{x}} \in \mathbb{R}^2$ and by its side length $2 \cdot B^c_\sigma \in \mathbb{R}_+$:

$$\begin{aligned} B^c_{\mathbf{x}} &= \frac{\sum_{j:\mathbf{g}_j \in \mathcal{G}} \mathbf{x}_j \cdot P(\chi_c \mid \mathbf{g}_j)}{\sum_{j:\mathbf{g}_j \in \mathcal{G}} P(\chi_c \mid \mathbf{g}_j)}, \\ B^c_\sigma &= \sqrt{\frac{\sum_{j:\mathbf{g}_j \in \mathcal{G}} \|B^c_{\mathbf{x}} - \mathbf{x}_j\|^2 \cdot P(\chi_c \mid \mathbf{g}_j)}{\sum_{j:\mathbf{g}_j \in \mathcal{G}} P(\chi_c \mid \mathbf{g}_j)}}. \end{aligned} \tag{19}$$

The relevance score is computed according to (7), which in this case has the form $p(\mathbf{g}_j \mid \chi_c, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma})$. Based on the set $\mathcal{R}$ of relevant compositions, a new estimate of the bounding box can be obtained. The locations $\mathbf{x}_j$ of relevant compositions are weighted with the compositional relevance $P(\chi_c \mid \mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma})$ from Section 5.2, giving an updated estimate of B^c (see line 8). This alternating estimation of relevance $\mathcal{R}$ and bounding box B^c is repeated iteratively.

The remainder of the recognition algorithm proceeds then according to Algorithm 2: First, candidates for higher order compositions are established within the estimated bounding box (i.e., for $(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{G}} \subset \mathcal{G} \times \mathcal{G}$ both constituents must lie within the bounding box).

Finally, the object localized by the bounding box is to be recognized. The underlying inference algorithm has been derived in Section 6.1.2. To make it applicable to the extended approach of this chapter, only the shifts S_j have to be adapted in (16):

$$\begin{aligned}\varphi_c :&= P(C = c|E)\\ &= P(c \mid \mathbf{g}^I, \mathbf{x}, \{\mathbf{g}_j, \mathbf{x}_j\}_{\mathbf{g}_j \in \mathcal{R}}, \{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}})\\ &\propto \exp\left[\ln P(c \mid \mathbf{g}^I) + \sum_{\mathbf{g}_j \in \mathcal{R}} \ln P\left(c \mid \mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma}\right)\right.\\ &\quad \left. + \sum_{(\mathbf{g}_k, \mathbf{g}_l) \in \widetilde{\mathcal{R}}} \ln P(c|\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl})\right].\end{aligned} \tag{20}$$

The object hypothesis (category c, localization B^c) is then rated with the confidence score φ_c.

Algorithm 4. Simultaneous Classification and Detection of Multiple Objects

OBJECTRECOGNITIONANDDETECTION(I) ▷ *I: query image*
1 $(\mathcal{G}, \mathbf{g}^I) \leftarrow$ ALLCOMPOSITIONCANDIDATES(I) ▷ *Algo. 1*
2 **for** all $c \in \mathcal{L}$
3 **do** ▷ *obtain initial estimate of bounding box B^c:*
4 $B^c_{\mathbf{x}} \leftarrow \frac{\sum_{j:\mathbf{g}_j \in \mathcal{G}} \mathbf{x}_j \cdot P(\chi_c|\mathbf{g}_j)}{\sum_{j:\mathbf{g}_j \in \mathcal{G}} P(\chi_c|\mathbf{g}_j)}$
5 $B^c_\sigma \leftarrow \sqrt{\frac{\sum_{j:\mathbf{g}_j \in \mathcal{G}} \|B^c_{\mathbf{x}} - \mathbf{x}_j\|^2 \cdot P(\chi_c|\mathbf{g}_j)}{\sum_{j:\mathbf{g}_j \in \mathcal{G}} P(\chi_c|\mathbf{g}_j)}}$
6 **repeat** ▷ *update $\mathcal{R}$ and B^c alternatingly for cat. c:*
7 $\mathcal{R} \leftarrow$ RELEVANTCOMPOSITIONS $(\mathcal{G}, B^c, c)$ ▷ *find relevant comps in B^c, (11)*
8 $B^c_{\mathbf{x}} \leftarrow \frac{\sum_{j:\mathbf{g}_j \in \mathcal{R}} \mathbf{x}_j \cdot P\left(\chi_c|\mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma}\right)}{\sum_{j:\mathbf{g}_j \in \mathcal{R}} P\left(\chi_c|\mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma}\right)}$
9 $B^c_\sigma \leftarrow \sqrt{\frac{\sum_{j:\mathbf{g}_j \in \mathcal{R}} \|B^c_{\mathbf{x}} - \mathbf{x}_j\|^2 \cdot P\left(\chi_c|\mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma}\right)}{\sum_{j:\mathbf{g}_j \in \mathcal{R}} P\left(\chi_c|\mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma}\right)}}$ ▷ *update B^c for given $\mathcal{R}$*
10 **until** convergence
▷ *draw compositions of compositions for category c:*
11 $\widetilde{\mathcal{G}} \leftarrow$ COMPOFCOMPCANDIDATES $(\mathcal{G}, B^c)$ ▷ *random subset of $\mathcal{G} \times \mathcal{G}$, Section 5.4*
12 $\widetilde{\mathcal{R}} \leftarrow$ RELEVANTCOMPSOFCOMPS $(\widetilde{\mathcal{G}}, c)$ ▷ *select relevant comps of comps, (12)*
▷ *confidence in cat. hypothesis c for obj. within B^c:*
14 $\varphi_c = P(C = c|E)$ ▷ *(20)*
$\leftarrow$ COMPOSITIONALSHAPEMODEL $(\mathcal{R}, \mathbf{g}^I, B^c, \widetilde{\mathcal{R}})$
15 **return** $\{\varphi_c, B^c\}_{c \in \mathcal{L}}$ ▷ *obj. cat. posterior & B.Box*

6.3.2 Learning Phase

Learning object models from unsegmented, cluttered training images implies that objects need to be localized automatically. This task is challenging since we are not provided any object models in this phase. Therefore, we apply an alternating update scheme (Algorithm 5): Object hypotheses are established for all training images before using these models to localize objects in the training images. Thereafter, the models are updated again based on the localization. The algorithm requires a set of images $\mathcal{T}^{(1)}_c$ that show objects of some category c and a set of irrelevant images $\mathcal{T}^{(0)}_c$ from other categories. The learning algorithm starts by extracting candidate compositions out of all the training images. In line 5, the distribution $P(\chi_c|\mathbf{g}_j)$ is learned using NKDA before computing an initial bounding box estimate in line 6. Now, an alternating update of bounding boxes and compositional relevance starts. The final bounding boxes are then used to select the relevant compositions for category c by employing Algorithm 3 and discarding all compositions that lie outside of a bounding box. Based on the set of relevant compositions and the estimated bounding boxes, the remaining distributions in (20) can be learned using multiclass NKDA as has been previously done.

Algorithm 5. Learning relevant compositions from images with multiple objects by alternating bounding box detection and the estimation of compositional relevance

LEARNINGRELEVANCEANDOBJECTDETECTION $(\mathcal{T}^{(1)}_c, \mathcal{T}^{(0)}_c)$
▷ $\mathcal{T}^{(1)}_c$: *set of training images for category c.*
▷ $\mathcal{T}^{(0)}_c$: *irrelevant images of other categories.*
1 **for** $i \in \{0, 1\}$ ▷ *collect comps. for imgs. in $\mathcal{T}^{(1)}_c, \mathcal{T}^{(0)}_c$:*
2 **do** $\mathcal{G}_i = \emptyset$
3 **for** $I \in \mathcal{T}^{(i)}_c$ ▷*Algo. 1:*
4 **do** $\mathcal{G}_i \leftarrow \mathcal{G}_i \cup$ ALLCOMPOSITIONCANDIDATES (I)
5 $P(\chi_c|\mathbf{g}_j) \leftarrow$ LEARNPROBCLASSIFIER $(\mathcal{G}_0, \mathcal{G}_1)$
6 **for** $I \in \mathcal{T}^{(1)}_c$ ▷ *obtain initial estimate of B.Box B^c:*
7 **do** $B^c_{\mathbf{x}}(I) \leftarrow \frac{\sum_{j:\mathbf{g}_j \in \mathcal{G}} \mathbf{x}_j \cdot P(\chi_c|\mathbf{g}_j)}{\sum_{j:\mathbf{g}_j \in \mathcal{G}} P(\chi_c|\mathbf{g}_j)}$
8 $B^c_\sigma(I) \leftarrow \sqrt{\frac{\sum_{j:\mathbf{g}_j \in \mathcal{G}} \|B^c_{\mathbf{x}} - \mathbf{x}_j\|^2 \cdot P(\chi_c|\mathbf{g}_j)}{\sum_{j:\mathbf{g}_j \in \mathcal{G}} P(\chi_c|\mathbf{g}_j)}}$
9 **for** h $= 1$ **to** 3 ▷ *alternat. update of relevance & B.Box:*
10 **do** $P(\chi_c|\mathbf{g}_j, S_j) \leftarrow$ LEARNPROBCLASSIFIER $(\mathcal{G}_0, \mathcal{G}_1, B^c)$
11 **for** $I \in \mathcal{T}^{(1)}_c$
12 **do** $B^c_{\mathbf{x}}(I) \leftarrow \frac{\sum_{j:\mathbf{g}_j \in \mathcal{R}} \mathbf{x}_j \cdot P\left(\chi_c\left|\mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}}(I) - \mathbf{x}_j\|}{B^c_\sigma(I)}\right.\right)}{\sum_{j:\mathbf{g}_j \in \mathcal{R}} P\left(\chi_c\left|\mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}}(I) - \mathbf{x}_j\|}{B^c_\sigma(I)}\right.\right)}$
13 $B^c_\sigma(I) \leftarrow \sqrt{\frac{\sum_{j:\mathbf{g}_j \in \mathcal{R}} \|B^c_{\mathbf{x}} - \mathbf{x}_j\|^2 \cdot P(\chi_c|\mathbf{g}_j, S_j)}{\sum_{j:\mathbf{g}_j \in \mathcal{R}} P\left(\chi_c|\mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}}(I) - \mathbf{x}_j\|}{B^c_\sigma(I)}\right)}}$

TABLE 2
Retrieval Rates of Current Approaches on Caltech-101 Using 30 Training Images Per Category

Approach	[49]	[20]	[39]	[50]	[32]
Retrieval rate [%]	$66.23 \pm .48$	$64.6 \pm .8$	58.23	56	53.0 ± 0.49

▷ *selection of relevant compositions using Algo. 3:*
14 $(p(\chi_c|\mathbf{g}_j, \mathbf{s}_j), \mathcal{R})$
← RELEVANCELEARNING $(\mathcal{T}_c^{(1)}, \mathcal{T}_c^{(0)}, \{B^c(I)\}_{I \in \mathcal{T}_c^{(1)}})$
15 **return** $p(\chi_c|\mathbf{g}_j, \mathbf{s}_j), \mathcal{R}, \{B^c(I)\}_{I \in \mathcal{T}_c^{(1)}}$

7 Evaluation of the Compositional Approach

7.1 Results on Caltech-101

We first evaluate the compositional approach on the challenging Caltech-101 database consisting of 101 object categories and a background category. Categories have varying numbers of samples (between 30 and 800) and range from photos with clutter to line drawings. The large intracategory variations in this database render object recognition a challenging task. However, there are only limited variations in pose. Following common practice, retrieval rates (fraction of correctly predicted test images) are averaged per class to avoid a bias toward the easy classes with many images. The overall retrieval rate ζ is, therefore, defined as

$$\zeta := \frac{1}{|\mathcal{L}|} \sum_{c \in \mathcal{L}} \{\text{true positive rate for category } c\}. \quad (21)$$

Moreover, five-fold cross validation is performed to obtain error bars.

Berg et al. [15] have calculated a reasonable baseline performance of 16 percent using texton histograms (random classification is below 1 percent). Table 2 summarizes the retrieval rates achieved by the state-of-the-art approaches on this database using 30 training images per category. Note that the top-ranked methods exploit the peculiarity of this specific database that the spatial structure of objects is limited in its variation with respect to the image, e.g., [20], and split the image into a regular grid and concatenate the individual descriptors to a joint one. In contrast to this, our approach aims at *learning the compositional structure* of objects. Recently, the approach of [20] has been extended [48] by further adapting the features, including an ROI search, and testing different types of classifiers. However, judging from the authors' discussion of their results, the performance gain to around 80 percent on Caltech-101 is mainly due to the adapted features.

Gain of compositionality over a baseline model. In the following experiments, different aspects of our compositional approach will be investigated. To evaluate the gain of compositionality, we start with a model that discards the compositional structure completely and uses features from a single scale. Recognition is then based on the bag representation $\mathbf{g}^I$ by maximizing $P(c|\mathbf{g}^I)$. This model achieves a retrieval rate of $\mathbf{35.3 \pm 0.8}$ percent for 30 training images. In contrast to this, the full compositional model increases performance to $\mathbf{58.8 \pm 0.9}$ percent using compositionality. A two-tail Student's t-test underlines the significance of this improvement (p-value of 0.005). Removing context $\mathbf{g}^I$ from this model decreases the performance to 54.2 ± 0.8 percent. Finally, we investigate an even simpler model: Images are tiled (a 5×5 grid showed the best performance), each grid cell is represented using a bag-of-features (same codebook as before), and an image is classified by maximizing the product of the individual posteriors. This model performs at 50.9 ± 1.1 percent.

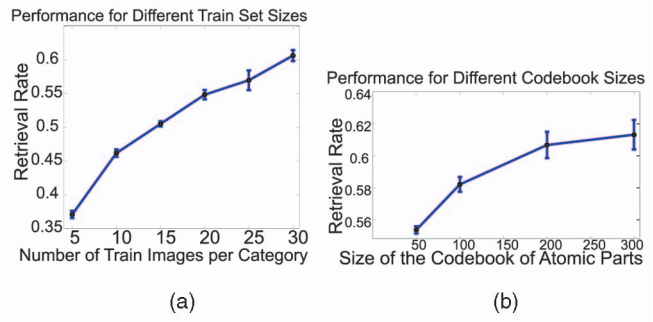

Fig. 7. (a) Retrieval rates of the full, multiscale compositional approach for different training set sizes and a 200D codebook of atomic parts (retrieval rate for 30 training images is 60.7 ± 0.8 percent). (b) Retrieval rates of the full compositional approach for different sizes of the part codebook. The algorithm is trained on 30 images per category.

MultiScale approach and different codebook sizes. In the previous experiment, recognition has been only conducted on a single scale. When processing images on three scales $\sigma_1 = 1, \sigma_2 = 1/2, \sigma_3 = 1/4$ (where $\sigma = 1$ corresponds to the original image scale), the retrieval rate is further improved to $\mathbf{60.7 \pm 0.8}$ **percent**. This shows that although the individual scales alone yield weaker performance (the individual performances are 58.8 ± 0.9 percent, 56.3 ± 1.3 percent, and 52.5 ± 0.7 percent), combining the image representations from multiple scales is effective in boosting the performance.

In Fig. 7b, different codebook sizes are investigated. Increasing the codebook from 200 prototypes in the previous experiments to 300 improves the performance to $\mathbf{61.3 \pm 0.9}$ **percent**.

Analyzing the established category hierarchy. From the category confusion table, it can be judged how similar the different categories are. Therefore, the confusion probabilities are used to establish a class hierarchy which reveals the degree of relatedness of categories.

The probability that a test image of category $c_{\text{true}} \in \mathcal{L}$ is classified by our architecture as belonging to class $c_{\text{pred}} \in \mathcal{L}$ is given by $P(c_{\text{pred}} \mid c_{\text{true}})$ and the confusion table is then represented by the matrix $\mathbf{M}_{c_{\text{true}}, c_{\text{pred}}} := P(c_{\text{pred}} \mid c_{\text{true}})$. The matrix is symmetrized by adding its transpose:

$$\widetilde{\mathbf{M}} := \eta \mathbf{E} - \left(\mathbf{M} + \mathbf{M}^T - 2\,\text{diag}[\mathbf{M}]\right). \quad (22)$$

Here, $\mathbf{E}$ denotes the matrix of only ones, η is a constant, and $\text{diag}\,[\mathbf{M}]$ is $\mathbf{M}$ with its off-diagonal entries set to zero. $\widetilde{\mathbf{M}}$ is used as a distance matrix between categories for a subsequent hierarchical clustering of categories (using *Ward's Method*). The resulting cluster tree (Fig. 8) has categories at its leaves and the length of a path between two classes is proportional to their dissimilarity. The confusion table is presented with permuted rows and columns so that it fits to the leafs of the adjacent hierarchy tree. The categories that are judged to be most similar are "water lily"

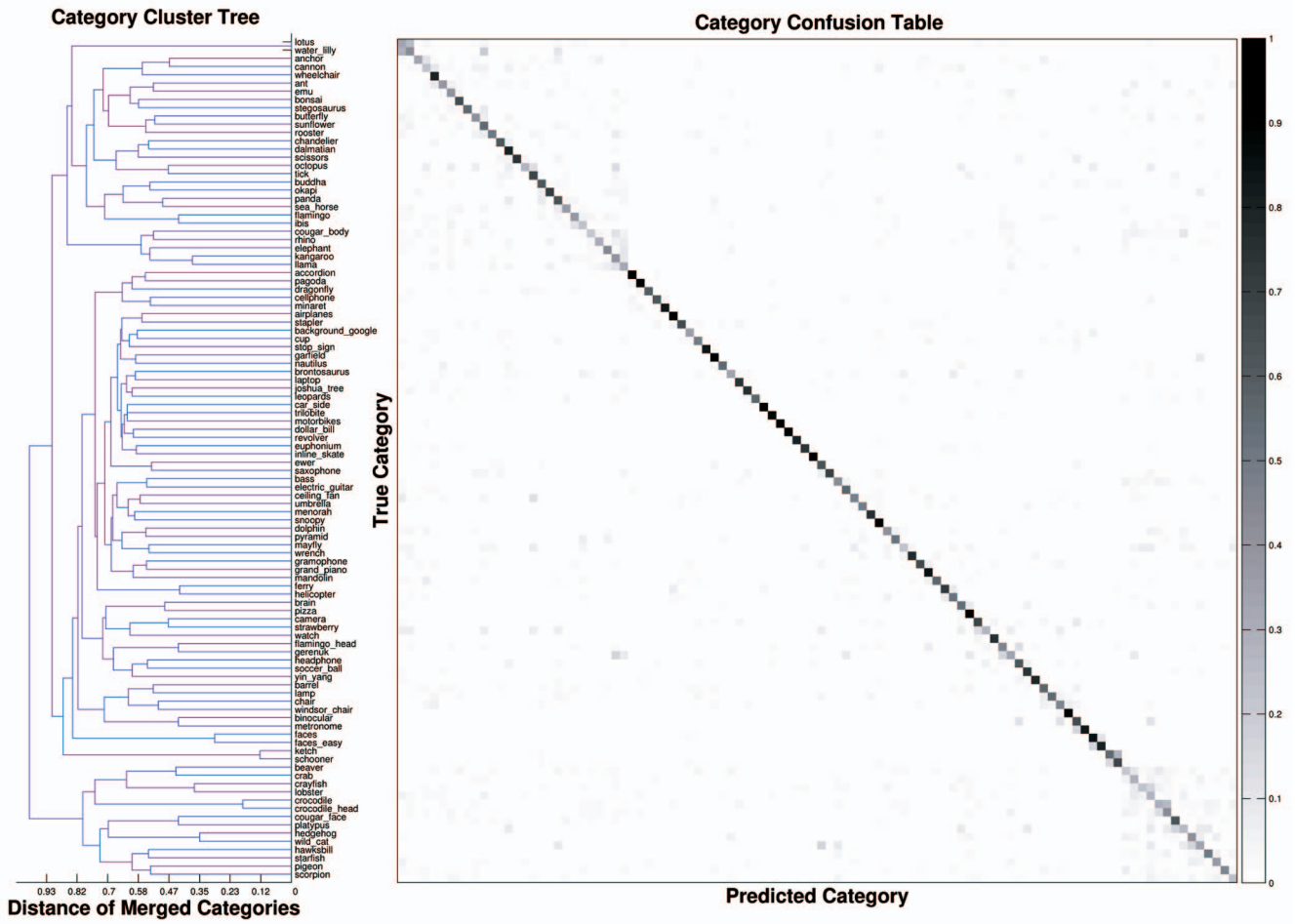

Fig. 8. Category hierarchy and category confusion table permuted to fit to the class tree. The retrieval rate is 61.3 ± 0.9 percent.

and "lotus," "ketch" and "schooner," and "crocodile" and "crocodile head." These similarities are intuitive since they are between pairs that are semantically close.

7.2 A Comparison of Feature Descriptors for Atomic Parts

The low dimensionality of localized feature histograms is crucial to render the learning of compositional object models statistically feasible. Subsequently, we contrast this feature with a common representation, the SIFT features [7]. Both representations are compared by plugging either of them as local feature $\mathbf{e}_i$ into the single-scale version (running on scale $\sigma_2 = 1/2$, using a 200D codebook and 30 training images per category) of our compositional system. However, we first analyze the effective dimensionality of both descriptors by randomly drawing 10,000 features from all Caltech-101 categories and applying PCA to the resulting features. Fig. 9 shows the eigenvalue spectrum for both descriptors. For the 40D localized feature histograms, a small subspace of 20 dimensions captures more than 90 percent of the total variance. Similarly, for SIFT, half of all eigenvalues represent 90 percent of the variability. Now, we plug either descriptor type into our system and summarize the results in Table 3. This evaluation underlines that localized feature histograms are effective in capturing local object particularities in a low-dimensional representation and clearly outperform SIFT as a representation basis for atomic compositional parts.

7.3 Influence of Color in Localized Feature Histograms

To analyze how much color attributes to the performance of localized feature histograms, we again use the compositional approach from Section 7.2 (scale $\sigma_2 = 1/2$ and 200D codebook). Two different versions of local features $\mathbf{e}_i$ are set up. The original 40D features described in Section 4.1 and a variant where the 8-bin color histogram is replaced by a 4-bin

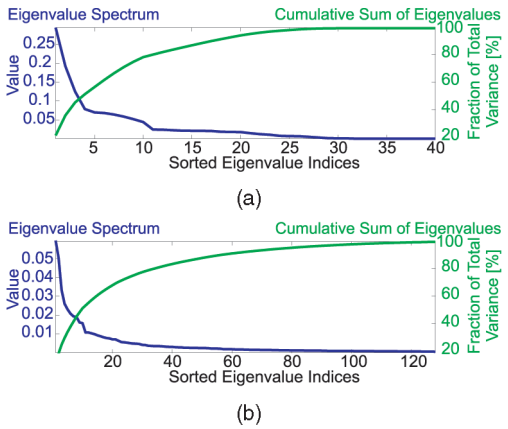

Fig. 9. Eigenvalue spectrum of (a) localized feature histograms and (b) SIFT features. The effective dimensionality of localized feature histograms is roughly a third of SIFT.

TABLE 3
Performance of the Compositional Approach (Single Scale $\sigma_2 = 1/2$, 200D Codebook, 30 Training Images Per Category) Using Different Local Descriptors and After Dimensionality Reduction Using PCA

Loc. Feat. Hists.	10-D	20-D	Orig. 40-D
Retrieval [%]	55.5 ± 1.0	56.8 ± 1.0	56.3 ± 1.3

SIFT	20-D	40-D	60-D	Orig. 128-D
Retrieval [%]	38.8 ± 0.6	39.6 ± 0.7	40.8 ± 1.1	40.6 ± 0.2

gray-scale histogram. Fig. 10 compares the retrieval rates of both approaches achieve for different training set sizes. On average, the color version yields a retrieval rate that is roughly 2 percent higher than that of the gray-scale version. Therefore, the difference in performance between our features and SIFT is only to a small extend caused by the additional color information but rather due to the lower dimensionality which renders learning statistically feasible.

7.4 Sampling a Compositional Representation from the Generative Object Model

During recognition, inference propagates information from local image features over intermediate compositions to an object category label. However, the graphical model in Fig. 6b can also be applied in a generative manner: Given object category c and object position $\mathbf{x}$, compositions and, finally, image patches can be inferred. It should be noted that this generative process infers atomic parts (the image patches), rather than directly generating pixels. Since the mapping from image patches to atomic parts is not invertible in general, we identify a synthesized atomic part with the image patch in the training data whose atomic part is the nearest neighbor of the synthesized part. To obtain the image representation in a region around $\mathbf{x}_j$, compositions $\mathbf{g}_j$ have to be sampled from the likelihood

$$p(\mathbf{g}_j \mid c, \mathbf{x}, \mathbf{x}_j) = \frac{P(c \mid \mathbf{g}_j, S_j = \mathbf{x} - \mathbf{x}_j) \cdot p(\mathbf{g}_j \mid S_j = \mathbf{x} - \mathbf{x}_j)}{P(c \mid \mathbf{x}, \mathbf{x}_j)}. \tag{23}$$

The denominator can be dropped since it only depends on evidence variables. Moreover, all candidate compositions that are established in the training images are distributed according to the composition prior, $\mathbf{g}_j \sim p(\mathbf{g}_j \mid S_j = \mathbf{x} - \mathbf{x}_j)$. Compositions can, therefore, be sampled by evaluating the

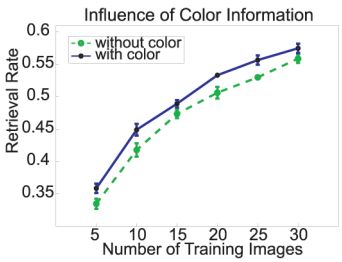

Fig. 10. Performance of the compositional approach (scale $\sigma_2 = 1/2$ and 200D codebook) based on the original 40D localized feature histograms that use color and based on a 36D variant of these features that discards all color information.

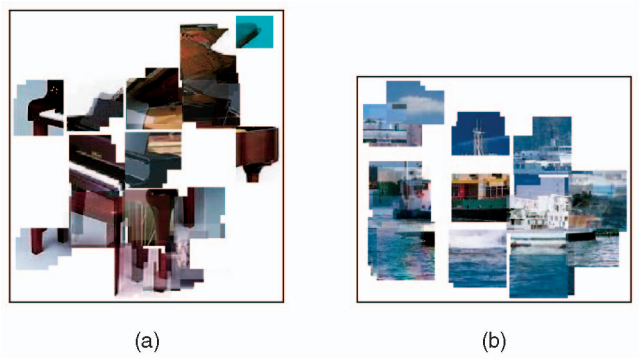

Fig. 11. Compositional image puzzles obtained by sampling compositions for (a) grand piano and (b) ferry. Given the position of the image center and a category label, compositions are sampled from the generative model.

category posterior $P(c \mid \mathbf{g}_j, S_j = \mathbf{x} - \mathbf{x}_j)$ (which has been learned for (16)) on compositions $\mathbf{g}_j$ that have been drawn from the training data:

$$p(\mathbf{g}_j \mid c, \mathbf{x}, \mathbf{x}_j) \propto P(c \mid \mathbf{g}_j, S_j = \mathbf{x} - \mathbf{x}_j)|_{\mathbf{g}_j \text{ from training}}. \tag{24}$$

The resulting compositional image puzzles in Fig. 11 provide insights into this generative process. Here, compositions have been inferred at points $\mathbf{x}_j$ on a regular grid (five compositions have been drawn at each point). We allow the sampled compositions to shift a short distance by performing gradient ascent on the likelihood (24) over $\mathbf{x}_j$ in a local neighborhood to reduce artifacts that result from sampling on a regular grid. This experiment reveals that the composition system has learned relevant compositions and their spatial relations to the object.

7.5 Inferring Missing Object Components

The higher order compositions which have been introduced in Section 5.4 can be used to infer missing compositions of an object. Given a composition $\mathbf{g}_k$, the remainder of an object can be inferred by drawing compositions $\mathbf{g}_j$ from the likelihood:

$$\begin{aligned} p(\mathbf{g}_j \mid \mathbf{g}_k, \mathbf{x}_k, c, \mathbf{x}_j) &= \frac{P(c \mid \mathbf{g}_j, \mathbf{g}_k, \mathbf{r}_{jk}) \cdot p(\mathbf{g}_j \mid \mathbf{g}_k, \mathbf{r}_{jk})}{P(c \mid \mathbf{g}_k, \mathbf{x}_k, \mathbf{x}_j)} \\ &\propto P(c \mid \mathbf{g}_j, \mathbf{g}_k, \mathbf{r}_{jk})|_{\mathbf{g}_j \text{ from training}}. \end{aligned} \tag{25}$$

In Fig. 12, a single composition is given together with the object category label. This information is used to infer a

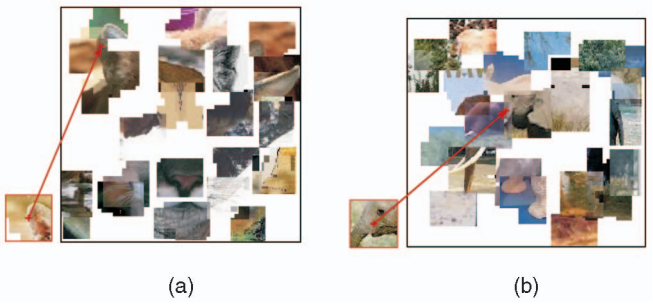

Fig. 12. Inferring compositions for (a) a cougar face and (b) an elephant. Given only the composition displayed in the box at the bottom left and the true category label, image patches corresponding to the inferred compositions are shown. The location of the conditioned composition is marked by a cross in the inferred image.

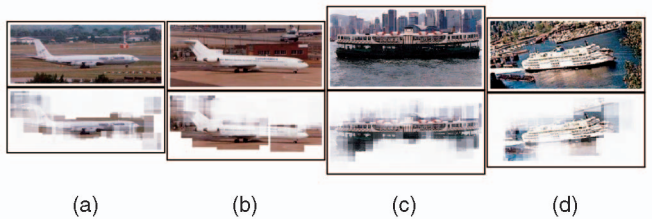

Fig. 13. Relevance of a composition is illustrated for recognition of objects. The opaqueness encodes the category posterior evaluated at compositions. The visualization shows which image patches contributed to recognizing the object.

maximum likelihood solution on the basis of compositions derived from the training set. The spatial structure of the reconstructed objects underlines that the compositional model has learned characteristic relationships between compositions.

7.6 Analyzing the Relevance of Compositions

Subsequently, the *relevance* of individual compositions for categorizing a test image is evaluated. Therefore, the category posterior of the true category,

$$P(c \mid \mathbf{g}^I, \mathbf{x}, \mathbf{g}_j, \mathbf{x}_j, \mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl})|_{c=\text{True Category}}, \tag{26}$$

is computed for individual pairs of compositions. In Fig. 13, the resulting probability is then encoded in the opaqueness of the underlying image parts. We obtain probabilistic segmentations of scenes based on the relevance of compositions. The visualization shows that relevant compositions cover meaningful object parts. Note that this segmentation is learned from *unsegmented* training images and for multiple categories simultaneously, as opposed to [18].

7.7 Detection and Classification of Multiple Objects: PASCAL VOC'06

7.7.1 Classification Performance

For each of the 10 classes, the system must decide if an instance of that class is present in a test image. The confidence scores are then used to rank the individual predictions so that a *receiver operator characteristic* (ROC curve) can be drawn. Fig. 14a shows the ROC curves for all 10 categories. The corresponding AUCs are in the range of 0.81 to 0.96 except for one outlier, category *person* with an AUC of 0.66. This class is particularly complicated as it features panoramic pictures that show full persons as well as close-up views that show only parts of a person. Altogether, the model is competitive compared with the average of all submissions to the VOC '06 challenge which is displayed in Fig. 14b. Since the compositional approach is not specifically tailored to any of these classes and does—unlike most of the PASCAL competitors—not use location information in training, achieving comparable performance to these models is remarkable. The largest performance gap can be observed for class *person* for which a lot of specifically designed approaches exist. The manual tuning toward this category and the rich supervision used by these methods explain their advantage.

In addition to the PASCAL challenge, we can also test our approach in a more complicated multiclass setting. The confusion table is presented in Fig. 14c and it shows a retrieval rate of 46.3 percent. The best performance is achieved for man-made objects, whereas retrieval rates for animal and person categories are generally lower. The dominant off-diagonal entries are confusions between *cat* and *dog* (23 percent) and *cows* mistaken for *sheep* (22 percent of all cows).

7.7.2 Detection Performance

The localization accuracy is measured in VOC'06 by predicting object bounding boxes together with a confidence score for test images. For a detection to be correct, VOC '06 demands that the area $A(\bullet)$ of overlap between a predicted bounding box B^c and a ground truth bounding box B^{gt} must be more than half the union of both areas:

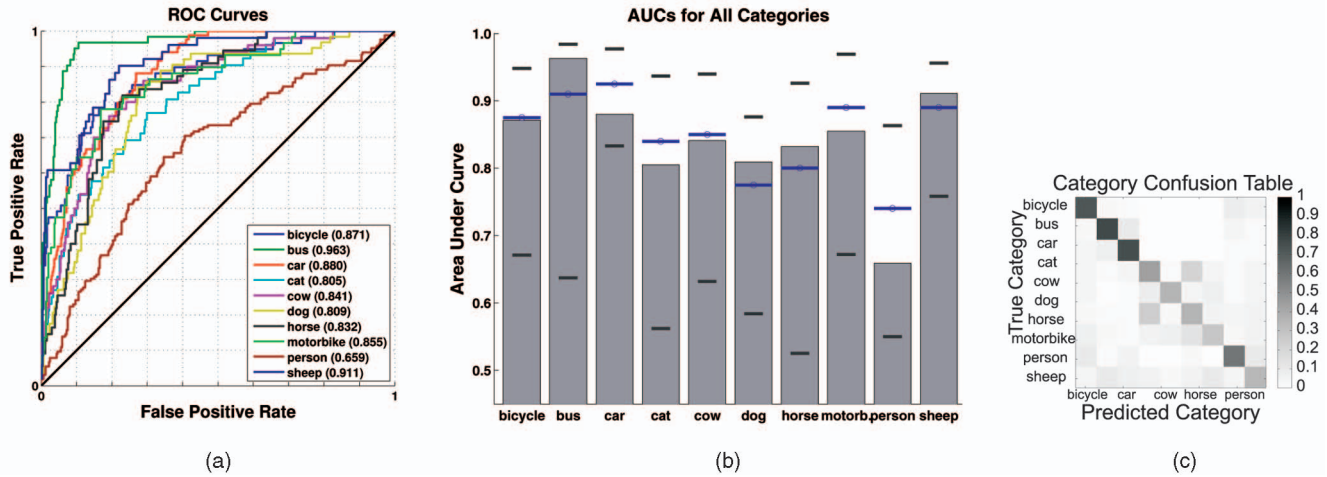

Fig. 14. Classification performance. (a) ROC curves for all categories of the PASCAL database, computed according to the guidelines of the PASCAL competition. The legend provides the categories names and the corresponding area under ROC curve (AUC). (b) The bars show the AUCs for all 10 categories. The lines indicate the average/min/max performance over the 20 submission of the PASCAL 2006 challenge. These methods include highly supervised approaches trained with bounding box information. (c) Multiclass recognition of all 10 categories. The overall retrieval rate is 46.3 percent.

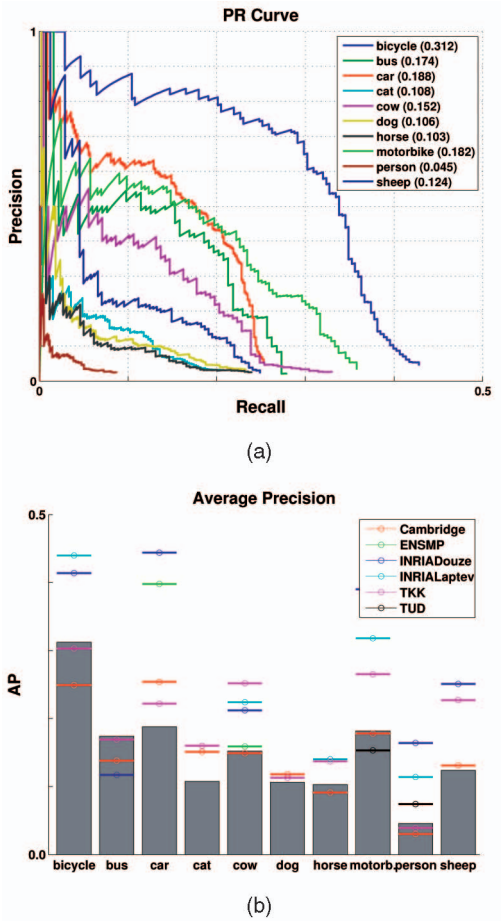

Fig. 15. Detection performance. (a) Precision recall curves for detection as defined in VOC '06. The legend provides next to each category label the average precision of that class. (b) The bars show the average precision (AP) for each category. The colored markers denote the performance of entries in the PASCAL competition which are mostly trained using heavy supervision.

$$\text{bounding box hypothesis } B^c \text{ correct} \Leftrightarrow \frac{A(B^c \cap B^{gt})}{A(B^c \cup B^{gt})} > \frac{1}{2}. \tag{27}$$

Fig. 15a shows the precision recall curves and compares against all PASCAL entries. Although these methods use heavy supervision and are mostly tailored to just a few categories, the compositional approach outperforms all of them for category *bus* and its performance is well within the range of competition entries for categories *dog*, *bicycle*, and *horse*. It is remarkable that a model trained without any localization information on cluttered training data is nevertheless competitive to state-of-the-art detection algorithms that have been trained in a supervised manner.

An example that depicts object localization is presented in Fig. 16. The blue bounding box on the left is the ground truth, whereas the green one in Fig. 16b is the predicted box. The probability map for category *dog* visualizes $P(\chi_{c=\text{"dog"}} \mid \mathbf{g}_j, S_j = \frac{\|B^c_{\mathbf{x}} - \mathbf{x}_j\|}{B^c_\sigma})$ for all compositions $\mathbf{g}_j$. This map shows that compositions on the object have a high confidence for *dog* and faithfully cover the animal. The example also indicates that the choice of bounding box

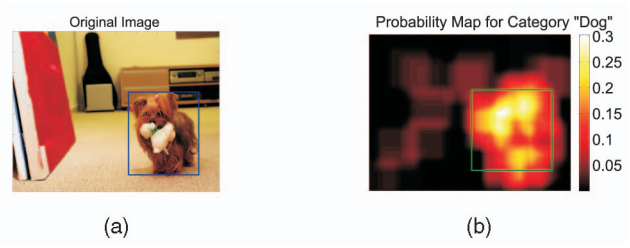

Fig. 16. Detection. (a) Test image with ground truth bounding box. (b) Probability map for category *dog*. This map depicts how confidently compositions vote for category *dog*.

localization may underrepresent the model's true, pixel-level localization capabilities.

7.8 Analyzing the Performance of the Compositional Model

In Section 7.1, it has been shown that the compositional approach significantly outperforms a baseline bag representation. To further investigate the performance gain achieved by compositionality over the baseline model, we focus on only the classification task: The correct object bounding box is given and only the object category is queried. Now, we can compare the performance of the full compositional model from Section 6.3.1 with that of the baseline model, which performs recognition by maximizing $P(c \mid \mathbf{g}^I)$. As an additional, intermediate experiment, we neglect the spatial information in the compositional model so that the gain of localization information can be measured separately. Therefore, recognition is based on maximizing

$$P(c \mid \mathbf{g}^I, \{\mathbf{g}_j\}_{\mathbf{g}_j \in \mathcal{R}}) \propto \exp\left[\sum_{\mathbf{g}_j \in \mathcal{R}} \ln P(c \mid \mathbf{g}_j)\right]. \tag{28}$$

Fig. 17 compares the retrieval rates achieved by the three models. Since the evaluation is restricted to a classification task of limited complexity, the performance differences of the approaches are also reduced. Nevertheless, it shows that even the compositional representation without spatial information (28) performs significantly better than the bag model. In conclusion, the experiment has underlined that compositions play a crucial role in building powerful vision systems.

8 Conclusion

Composition systems for real-world object recognition have been developed in this contribution. We have, in particular,

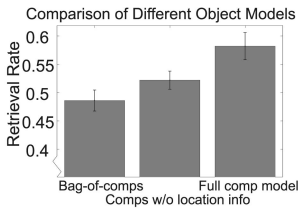

Fig. 17. Comparing the performance of a bag representation $\mathbf{g}^I$, with that of a simple compositional approach without spatial information (28), and the full compositional model from Section 6.3.1.

investigated how the compositional nature of objects can be learned automatically without requiring manual supervision. Therefore, relevant object structure is automatically discovered in cluttered training images without requiring hand segmentations or other manual localization information. Moreover, it has been shown how compositions can bridge the large semantic gap between robust, but unspecific local descriptors and high-level categories. A small codebook of these local descriptors, which is shared by all categories, is therefore sufficient to represent large numbers of diverse categories and the system automatically compensates for the information that is lacking in the local features by learning characteristic relations between them. Finally, a Bayesian network has been presented that couples all compositions with relations between them, object shape, and scene context to provide a concerted object hypothesis. Thus, our compositions system efficiently learns structured object models to infer complex scene interpretations.

ACKNOWLEDGMENTS

This work was supported in part by the Swiss National Science Foundation under contract no. 200021-107636.

REFERENCES

[1] F. Attneave, "Some Informational Aspects of Visual Perception," *Psychological Rev.,* vol. 61, no. 3, pp. 183-193, 1954.
[2] S. Geman, D.F. Potter, and Z. Chi, "Composition Systems," *Quarterly of Applied Math.,* vol. 60, pp. 707-736, 2002.
[3] I. Biederman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Rev.,* vol. 94, no. 2, pp. 115-147, 1987.
[4] D.G. Lowe, *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, 1985.
[5] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale Invariant Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 264-271, 2003.
[6] B. Ommer and J.M. Buhmann, "Learning Compositional Categorization Models," *Proc. European Conf. Computer Vision,* pp. 316-329, 2006.
[7] D.G. Lowe, "Distinctive Image Features from Scale Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
[8] Y. Amit and D. Geman, "A Computational Model for Visual Selection," *Neural Computation,* vol. 11, no. 7, pp. 1691-1715, 1998.
[9] M.C. Burl, M. Weber, and P. Perona, "A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry," *Proc. European Conf. Computer Vision,* pp. 628-641, 1998.
[10] S. Agarwal, A. Awan, and D. Roth, "Learning to Detect Objects in Images via a Sparse, Part-Based Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 11, pp. 1475-1490, Nov. 2004.
[11] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop Generative Model Based Vision,* 2004.
[12] B. Leibe and B. Schiele, "Scale Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search," *Proc. Pattern Recognition Symp.,* pp. 145-153, 2004.
[13] B. Ommer and J.M. Buhmann, "Object Categorization by Compositional Graphical Models," *Proc. Int'l Workshop Energy Minimization Methods in Computer Vision and Pattern Recognition,* pp. 235-250, 2005.
[14] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 3, pp. 411-426, Mar. 2007.
[15] A.C. Berg, T.L. Berg, and J. Malik, "Shape Matching and Object Recognition Using Low Distortion Correspondence," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 26-33, 2005.
[16] A. Opelt, A. Pinz, and A. Zisserman, "Incremental Learning of Object Detectors Using a Visual Shape Alphabet," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 3-10, 2006.
[17] V. Ferrari, T. Tuytelaars, and L.J.V. Gool, "Object Detection by Contour Segment Networks," *Proc. European Conf. Computer Vision,* pp. 14-28, 2006.
[18] B. Leibe, A. Leonardis, and B. Schiele, "Combined Object Categorization and Segmentation with an Implicit Shape Model," *Proc. European Conf. Computer Vision Workshop Statistical Learning in Computer Vision,* 2004.
[19] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual Categorization with Bags of Keypoints," *Proc. European Conf. Computer Vision Workshop Statistical Learning in Computer Vision,* 2004.
[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2169-2178, 2006.
[21] M.A. Fischler and R.A. Elschlager, "The Representation and Matching of Pictorial Structures," *IEEE Trans. Computers,* vol. 22, no. 1, pp. 67-92, Jan. 1973.
[22] M. Lades, J.C. Vorbrüggen, J.M. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers,* vol. 42, no. 3, pp. 300-311, Mar. 1993.
[23] M. Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," *Proc. European Conf. Computer Vision,* pp. 18-32, 2000.
[24] A. Holub, M. Welling, and P. Perona, "Combining Generative Models and Fisher Kernels for Object Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 136-143, 2005.
[25] K. Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics,* vol. 36, no. 4, pp. 193-202, 1980.
[26] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning,* vol. 42, no. 1, pp. 177-196, 2001.
[27] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.
[28] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering Objects and Their Localization in Images," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 370-377, 2005.
[29] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1816-1823, 2005.
[30] B. Epshtein and S. Ullman, "Feature Hierarchies for Object Classification," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 220-227, 2005.
[31] G. Bouchard and B. Triggs, "Hierarchical Part-Based Visual Object Categorization," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 710-715, 2005.
[32] B. Ommer, M. Sauter, and J.M. Buhmann, "Learning Top-Down Grouping of Compositional Hierarchies for Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop Percept Organization in Computer Vision,* 2006.
[33] A.Y. Ng and M.I. Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," *Proc. Advances in Neural Information Processing Systems,* pp. 841-848, 2002.
[34] E.B. Sudderth, A.B. Torralba, W.T. Freeman, and A.S. Willsky, "Learning Hierarchical Models of Scenes, Objects, and Parts," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1331-1338, 2005.
[35] Z. Tu, X. Chen, A. Yuille, and S. Zhu, "Image Parsing: Unifying Segmentation, Detection and Recognition," *Int'l J. Computer Vision,* vol. 63, no. 2, pp. 113-140, 2005.
[36] R. Fergus, P. Perona, and A. Zisserman, "A Visual Category Filter for Google Images," *Proc. European Conf. Computer Vision,* pp. 242-256, 2004.
[37] E. Borenstein, E. Sharon, and S. Ullman, "Combining Top-Down and Bottom-Up Segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop Percept Organization in Computer Vision,* 2004.
[38] P.A. Viola and M.J. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 511-518, 2001.

[39] K. Grauman and T. Darrell, "Pyramid Match Kernels: Discriminative Classification with Sets of Image Features," Technical Report MIT-CSAIL-TR-2006-020, 2006.
[40] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial Structures for Object Recognition," *Int'l J. Computer Vision*, vol. 61, no. 1, pp. 55-79, 2005.
[41] Y. Jin and S. Geman, "Context and Hierarchy in a Probabilistic Image Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2145-2152, 2006.
[42] R. Veltkamp and M. Tanase, "Content-Based Image Retrieval Systems: A Survey," Technical Report UU-CS-2000-34, Information and Computing Sciences, Utrecht Univ., 2000.
[43] K. Mikolajczyk and C. Schmid, "Scale & Affine Invariant Interest Point Detectors," *Int'l J. Computer Vision*, vol. 60, no. 1, pp. 63-86, 2004.
[44] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods—A Mathematical Introduction*, second ed. Springer, 2003.
[45] V. Roth and K. Tsuda, "Pairwise Coupling for Machine Recognition of Hand-Printed Japanese Characters," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1120-1125, 2001.
[46] J. Puzicha, T. Hofmann, and J.M. Buhmann, "Histogram Clustering for Unsupervised Segmentation and Image Retrieval," *Pattern Recognition Letters*, vol. 20, pp. 899-909, 1999.
[47] M. Everingham, A. Zisserman, C.K.I. Williams, and L. VanGool, "The PASCAL Visual Object Classes Challenge 2006 (VOC '06)," http://www.pascal-network.org/challenges/VOC/voc2006, 2006.
[48] A. Bosch, A. Zisserman, and X. Munoz, "Image Classification Using Random Forests and Ferns," *Proc. IEEE Int'l Conf. Computer Vision*, 2007.
[49] H. Zhang, A.C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2126-2133, 2006.
[50] J. Mutch and D.G. Lowe, "Multiclass Object Recognition with Sparse, Localized Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 11-18, 2006.

Björn Ommer received the diploma in computer science from the University of Bonn, Germany, in 2003 and the PhD degree in computer science from ETH Zurich, Switzerland, in 2007. Afterward, he held a postdoctoral position in the Computer Vision Group headed by Professor J. Malik at the University of California, Berkeley. In 2009, he joined the University of Heidelberg, Germany, as an assistant professor for scientific computing. His research interests include computer vision, machine learning, and cognitive science. Special research topics are object recognition in images and video, biomedical image analysis, graphical models, and perceptual organization. He is a member of the IEEE.

Joachim M. Buhmann received the PhD degree in theoretical physics from the Technical University of Munich, Germany, in 1988. He has held postdoctoral and research faculty positions at the University of Southern California, Los Angeles, and the Lawrence Livermore National Laboratory, California, between 1988 and 1992. Until October 2003, he headed the Research Group on Pattern Recognition, Computer Vision and Bioinformatics in the Computer Science Department, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany. In October 2003, he joined the Computer Science Department of the Swiss Federal Institute of Technology (ETH) in Zurich as a professor for information science and engineering. His current research interests cover statistical learning theory and its applications to image understanding and signal processing. Special research topics include exploratory data analysis and data mining in bioinformatics, stochastic optimization, graphical models, and computer vision. He has served as an associate editor for the IEEE-TNN, the IEEE-TIP, and the IEEE-TPAMI. He is a senior member of the IEEE.