Learning the Compositional Nature of Visual Objects

Björn Ommer Joachim M. Buhmann*
Institute of Computational Science, ETH Zurich
8092 Zurich, Switzerland
bjoern.ommer@inf.ethz.ch jbuhmann@inf.ethz.ch

Abstract

The compositional nature of visual objects significantly limits their representation complexity and renders learning of structured object models tractable. Adopting this modeling strategy we both (i) automatically decompose objects into a hierarchy of relevant compositions and we (ii) learn such a compositional representation for each category without supervision. The compositional structure supports feature sharing already on the lowest level of small image patches. Compositions are represented as probability distributions over their constituent parts and the relations between them. The global shape of objects is captured by a graphical model which combines all compositions. Inference based on the underlying statistical model is then employed to obtain a category level object recognition system. Experiments on large standard benchmark datasets underline the competitive recognition performance of this approach and they provide insights into the learned compositional structure of objects.

1. Introduction

Learning object representations for detection and recognition poses one of the key challenges of computer vision. This problem becomes especially complex and difficult in the limit of unconstrained scenes, large intra-class variations and weak supervision during training. A central concept for learning object models despite these problems is to exploit the compositional nature of our (visual) world.

In this contribution we investigate methods for learning the compositional structure of objects and we present an integration in a category level object recognition system. The approach learns characteristic compositions of atomic parts for each category in an unsupervised manner, requiring neither hand segmentations nor object localization during training. In the same way higher level compositions of compositions are learned. Finally, a Bayesian network serves as a coherent model that comprises all the compositional constituents together with object shape. Inference based on this probabilistic model yields a decomposition of a scene into a hierarchy of relevant compositions and, finally, enables localization and recognition of objects. Our main theme of *learning a compositional architecture for object recognition* substantially extends the recognition system of Jin & Geman [14] for license plates, who focused their study on structural aspects of compositionality and explicitly excluded the question of learning such systems.

Compositionality (e.g. [10]), which serves as a foundation for this contribution, is a general principle in cognition and can be especially observed in human vision [3]. Perception has a high tendency to represent complex entities by means of comparably few, simple, and widely usable parts together with relations between them. In contrast to modeling an object directly based on a constellation of its parts (e.g. [8]), the compositional approach learns intermediate groupings of parts. As a consequence, compositions bridge the semantic gap between low level features and high level object recognition by establishing intermediate hidden layer representations. In conclusion, compositions model category-distinctive subregions of an object, which show small intra-category variations compared to the whole object.

A key idea of the *compositional object recognition model* from [21] is to realize feature sharing on the lowest level on which robust statistics are available. To this end, edge and color distributions of small image patches, *i.e. localized feature histograms* [20], have proved to be a feasible choice. A small codebook of these features is then shared by all categories and serves as a set of atomic parts. Therefore, this initial representation layer alone is generic and far from being category specific. The information that is relevant for delineating object classes from another comes from learning relations between parts and using them to build higher level compositions. These compositions are then represented by probability distributions over their constituent parts, thereby leading to a probabilistic, hierarchical scene representation: Distributions over atomic parts yield

*This work was supported in part by the Swiss national fund under contract no. 200021-107636.

in Proceedings of CVPR'07,

compositions, distributions over compositions yield higher level compositions of compositions, and so on. Finally, the spatial arrangement of all compositions is captured in a probabilistic model, *i.e*. the *compositional shape model*, which yields a statistical scene interpretation. In this model, all compositions which are present in a scene are coupled by (i) their spatial arrangement, (ii) by establishing relations between compositions, which yields higher level compositions, and (iii) by scene context, *i.e*. by the co-occurrence of all compositions. The problem of learning object models is therefore decomposed into learning the individual constituent distributions that represent compositions.

We extend the approaches [21, 22] by revising the learning and inference of compositions. Foremost, a training stage is integrated that automatically learns the relevancy of compositions for the task of discriminating object categories from another. Moreover, higher level compositions of compositions are inferred using top-down information and the coupling of compositions in the graphical model is extended. Finally, the probabilistic model can be used in a generative manner so that compositions and thereby an image representation can be estimated given a categorization.

2. Related Work

Typically, the problem of object representation has been addressed by using local descriptors and modeling their configuration in a flexible way, e.g. [9, 16, 8, 6, 1, 20, 2]. A common choice of local image features are template-based *appearance patches* (e.g. [1, 8, 6, 16]) and histogram-based descriptors such as *SIFT* features [17]. *Geometric blur* [2] and *localized feature histograms* [20] fall in the latter category. Moreover, Serre *et al*. [25] have proposed neurophysiologically motivated descriptors and hierarchical decompositions of features have been studied in [5].

A simple and robust way to model the configuration of descriptors are *bag of features* methods such as [4] that establish a histogram over all image features. This representation, however, discards the spatial structure of a scene. By making the assumption that the spatial structure of objects is limited in its variation with respect to the image, Lazebnik *et al*. [15] can improve the performance of the bag of features approach using a spatially fixed grid of feature bags. In this paper we do however address the problem of automatically learning the structure of objects. At the other end of the modeling spectrum are *constellation models*, e.g. [8, 6, 13], which code spatial relations according to the original approach of Fischler and Elschlager [9]. In contrast to such joint models of all image parts (which are limited in the number of parts for complexity reasons), [1, 16, 20, 21, 22] aim at utilizing greater numbers of image constituents. Opelt *et al*. [23] extract curve fragments from training images and apply Adaboost to learn strong object detectors. Feature sharing is conducted in the joint space

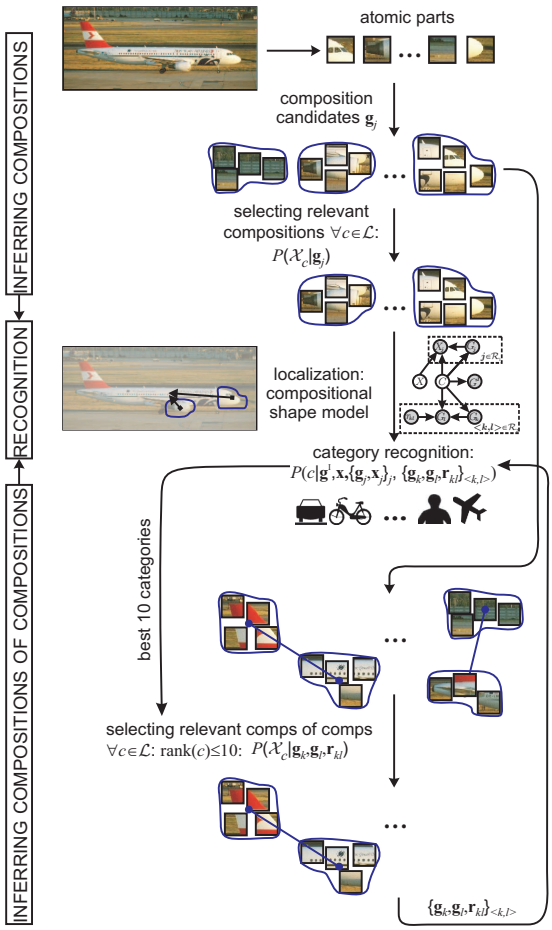


Figure 1. Processing pipeline for scene analysis.

of curves and their relative position to the object center. In contrast to this, we follow an approach that shares features already on the lowest level of small image patches and learns characteristic combinations of these. By virtue of relations between the parts, such compositions are capable of representing texture as well as boundary curves [21]. An example of a supervised approach to modeling configurations of parts is given by Felzenszwalb and Huttenlocher in [7]. Furthermore, Jin and Geman [14] present a compositional architecture with manually built structure for license plate reading. In their conclusion they put emphasis on the complexity of the future challenge of learning such a compositional model. In this contribution we deal with exactly this problem in the even less constraint case of large numbers of natural object classes. Finally, an approach that is based on establishing coherent spatial mappings between a probe image and all training images has been taken in [2, 26].

3. Learning a Composition System

Subsequently, we give a brief overview over our approach to compositional scene analysis (illustrated in Figure

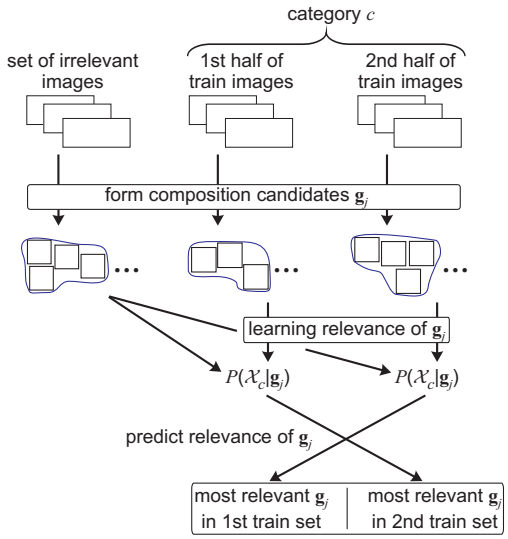


Figure 2. Learning relevant compositions.

1) before presenting the processing pipeline in detail in later sections. Given a novel image, small patches are extracted. They serve as atomic parts in the compositional hierarchy and for all of them localized feature histograms [20] are computed. Each image region is then represented by a discrete probability distribution 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. Therefore, compositions of these parts are established subsequently.

The aim of this paper is not to manually model a set of grouping laws that lead to characteristic compositions (e.g. using perceptual grouping), but to develop a learning strategy that automatically learns to establish relevant compositions. Hence we employ a simple, proximity based grouping strategy to form candidate compositions. Out of these, relevant compositions are selected by a relevance model that has been learned during the training phase. The relevant compositions enter into a Bayesian network where the individual object hypotheses are coupled by means of the co-occurrence and spatial distribution of all compositions. 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 relevant compositions of compositions (the constituents are compositions themselves) to accumulate additional evidence for the correct hypothesis. This is a top-down grouping process which is guided by previously inferred object information. The idea is that for the true hypothesis many compositions can be found which exhibit a peaked distribution over all categories. In contrast to this, incorrect object hypotheses are very likely to yield additional compositions with close to uniform class distributions. The newly generated compositions enter then into the Bayesian network together with the compositions from before to refine the categorization hypothesis.

The learning of relevant compositions is in the spirit of [21]. However, in this contribution we follow an improved learning strategy. Moreover, the approach differs from the one by Opelt *et al*. in [23] in that a uniform probabilistic model is used to determine the relevance of compositions. In addition, our system learns to build a hierarchy of compositions and performs feature sharing already on the lowest level of atomic parts.

3.1. A Shared Representation of Atomic Compositional Parts

The features representing the atomic parts of the compositional hierarchy should exhibit (i) good localization, (ii) robustness to local image changes, (iii) low dimensionality, and (iv) they should be shareable among the different object categories. Localized feature histograms [20] have been shown to provide a satisfactory trade-off between these requirements [21]. Let us therefore give a brief summary of these features: At salient image locations, which are detected by a scale invariant version of the Harris interest point detector [18], quadratic patches of size 20×20 pixels are extracted. Each patch is divided up into four equally sized subpatches with locations fixed relative to the patch center. In each of these subwindows marginal histograms over edge orientation and edge strength are computed (allocating four bins to each of them). Furthermore, an eight bin color histogram over all subpatches is extracted. All histograms are then combined in a common feature vector $\mathbf{e}_i$.

By performing a k-means clustering on all feature vectors detected in the training data a $k = 200$ dimensional codebook is assembled. For increased robustness of the representation, each feature is described by a Gibbs distribution over the codebook rather than by the closest codebook entry: Let $d_\nu(\mathbf{e}_i)$ denote the squared Euclidean distance of a measured feature $\mathbf{e}_i$ to a centroid $\mathbf{a}_\nu$. The local descriptor is then represented by the following distribution of its cluster assignment random variable F_i,

$$P(F_i = \nu|\mathbf{e}_i) := \frac{\exp\left(-d_\nu(\mathbf{e}_i)\right)}{\sum_\nu \exp\left(-d_\nu(\mathbf{e}_i)\right)} \quad . \tag{1}$$

3.2. Composition Candidates

One approach to obtaining category specific compositions of parts is to combine a set of grouping laws in a carefully designed, complex grouping algorithm (e.g. [22]). In this paper we do however follow the idea of [21] and form a large number of candidate compositions using a simple proximity based grouping and remove all irrelevant ones afterwards. Therefore, a learning algorithm can automati-

cally carry out the tedious task of retrieving relevant compositions, as will be shown in Section 3.4.

From all image patches that have been extracted as outlined in Section 3.1, a subset of 120 is randomly selected. Each of these parts is then grouped with the parts in its local neighborhood. We have extracted compositions with various neighborhood sizes to be less prone to scale changes. A validation has however shown that for the Caltech-101 image database most of the relevant compositions originated from a single grouping radius (30 pixel). This seems reasonable as most objects show characteristic compositions (such as rudders of airplanes) on this scale (see Figure 4).

All the constituent parts of a composition are then combined and a histogram over the part codebook is established. Therefore, a composition is represented as a mixture over the distributions of its parts, Eq. (1). Let $\Gamma_j = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ denote the grouping of parts represented by features $\mathbf{e}_1, \ldots, \mathbf{e}_m$. The composition is then represented by the random variable G_j which is a bag of features, i.e. its value $\mathbf{g}_j$ is a multivariate distribution over the k-dimensional feature codebook

$$\mathbf{g}_j \propto \sum_{i=1}^{m} \Big(P(F_i = 1|\mathbf{e}_i), \ldots, P(F_i = k|\mathbf{e}_i) \Big)^T \quad . \tag{2}$$

Each of the k dimensions is independently standardized to zero mean and unit variance across the whole training set. This mixture model exhibits the favorable property of robustness with respect to variations in the individual parts.

3.3. Using Compositions for Object Localization

Subsequently, all composition candidates $\mathbf{g}_j$ and the *scene context* $\mathbf{g}^I$ are used to obtain a first estimate of the object center. $\mathbf{g}^I$ captures the context of the scene by representing the co-occurrence of all compositions that are present in an image I. Therefore, we use a bag of compositions which is a mixture of all the composition distributions, *i.e.* $\mathbf{g}^I \propto \sum_j \mathbf{g}_j$. To determine the object location $\mathbf{x}$, the positions $\mathbf{x}_j$ of all compositions $\mathbf{g}_j$ are considered as proposed in [21]. Moreover, $c \in \mathcal{L}$ denotes a category label and $\mathcal{L}$ is the set of all labels. The position of the object center is then estimated by weighting the contribution of each composition with the probability that it should be observed

$$\mathbf{x} = \frac{\sum_j \mathbf{x}_j \sum_{c \in \mathcal{L}} p(\mathbf{g}_j|c, \mathbf{g}^I)\ P(c|\mathbf{g}^I)}{\sum_{j, c \in \mathcal{L}} p(\mathbf{g}_j|c, \mathbf{g}^I)\ P(c|\mathbf{g}^I)} \quad . \tag{3}$$

The first distribution is estimated using Parzen windows and the second one using *nonlinear kernel discriminant analysis* (NKDA) [24]. NKDA uses probabilistic two-class kernel classifiers and performs pairwise coupling to solve the multi-class problem. 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.

An evaluation on the Caltech-101 database shows that the estimate of the object center in (3) deviates from the true center (taking the center of the object bounding box from hand annotations) by $8.8 \pm 3.8\%$ of the bounding box diagonal (averaged over all categories). This is roughly the size of the atomic images patches and, therefore, exact enough to couple compositions in the compositional shape model.

3.4. Learning Relevant Compositions of Parts

Subsequently, we present an approach that automatically learns to retrieve those compositions which are relevant to distinguish a category from the rest. This sampling is valuable for both training and recognition as it discards distracting compositions such as irrelevant background clutter. We present a Bayesian criterion that defines what relevant compositions are and we show how they can be learned in the training phase. However, to learn the relevant compositions for category c we first need a set of irrelevant compositions as a negative set. As there is no proper background set available, we take a random sample of compositions from all images of all other categories.

Adopting the Bayesian view point, a composition $\mathbf{g}_j$ is relevant for representing objects of some category c if it has a high likelihood $p(\mathbf{g}_j|\chi_c)$. The indicator function χ_c is defined by $\chi_c = 1$ iff $\mathbf{g}_j$ is from an image I of category c,

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

Bayes' theorem implies that

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

Since a priori all categories should be equally likely, $P(\chi_c)$ can be dropped and we obtain for the likelihood

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

Now we also incorporate the estimate of the object center $\mathbf{x}$ from Section 3.3 and the position $\mathbf{x}_j$ of the composition,

$$p(\mathbf{g}_j|\chi_c, \mathbf{x}_j, \mathbf{x}) = p(\mathbf{g}_j|\chi_c, S_j = \mathbf{x} - \mathbf{x}_j) \tag{7}$$

$$\propto P(\chi_c|\mathbf{g}_j, S_j = \mathbf{x} - \mathbf{x}_j) \times p(\mathbf{g}_j|S_j = \mathbf{x} - \mathbf{x}_j) \quad . \tag{8}$$

Here the relative position of a composition with respect to the object center is represented by the shift $\mathbf{s}_j = \mathbf{x} - \mathbf{x}_j$. We therefore exploit the fact that compositions are not dependent on their absolute position in an image but that their probability only depends on their shifts relative to the object center. In Equation (6) as well as in (8), the relevance of compositions factorizes into two distributions. The first

one captures the discriminative power of $\mathbf{g}_j$, whereas the second indicates how reliably it can be detected. To avoid density estimation of $p(\mathbf{g}_j)$ and to render learning of compositions less prone to overfitting we choose an approach based on cross-validation (see Figure 2). The idea is to learn the posterior distribution $P(\chi_c|\mathbf{g}_j, \mathbf{s}_j)$ on one part of the training data and use it to predict the relevance of compositions in the other part. Unfavorable compositions with low prior $p(\mathbf{g}_j|\mathbf{s}_j)$ have a low probability to also appear in the validation set. As a consequence, validation prevents the learning algorithm from overfitting to the compositions extracted from the training set. Splitting up the set of training images for category c into m subsets yields m rounds of cross-validation. In each round the posterior is estimated on all compositions from $m-1$ image subsets against the set of irrelevant compositions. We solve this two-class classification task using NKDA. The classifier is then used to predict the relevant compositions from the remaining validation set. By computing the distance to the separating hyperplane of the kernel classifier, we can also estimate the probability that a composition is relevant. These estimates yield a ranking of compositions in each validation image and we retain the top 50% from an image. Section 4.2 presents a visualization of the learned compositions.

3.5. Learning Compositions of Compositions

Direct dependencies between compositions can be incorporated by learning groupings of compositions. Therefore, random tuples of compositions $\mathbf{g}_k$, $\mathbf{g}_l$ are considered and relations $\mathbf{r}_{kl}$ between them are measured (currently we only use the distance vector, *i.e.* $\mathbf{r}_{kl} = \mathbf{x}_k - \mathbf{x}_l$). Learning which compositions of compositions are particularly relevant proceeds along the lines of Section 3.4. We only have to adapt the relevance score (6) leading to

$$p(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}|\chi_c) \propto P(\chi_c|\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl})\, p(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}) \ . \quad (9)$$

The new posterior of the category indicator function $P(\chi_c|\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl})$ is then plugged into the learning algorithm derived in Section 3.4 and illustrated in Figure 2.

3.6. Compositional Shape Model for Binding Compositions

Object recognition in novel test images (see Figure 1) proceeds by forming candidate compositions as described above and by selecting the most relevant candidates for each potential category using the classifier of Section 3.4. Based on this set $\mathcal{R}_1$ of relevant compositions $\mathbf{g}_j$, the object is localized using (3). Subsequently, all compositions and the context descriptor $\mathbf{g}^I$ from Section 3.3 enter into an extended version of the compositional shape model [21] (see Figure 3) to yield a joint hypothesis of the object category

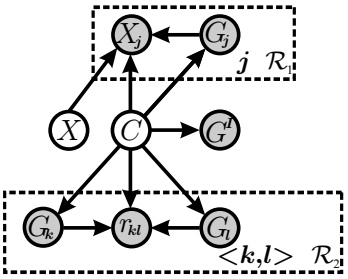


Figure 3. Bayesian network that couples compositions, shape, and image categorization. Shaded nodes denote evidence variables.

$c \in \mathcal{L}$. This initial hypothesis is given by the posterior

$$P(c|\mathbf{g}^I, \mathbf{x}, \{\mathbf{g}_j, \mathbf{x}_j\}_{j\in\mathcal{R}_1}) \quad (10)$$

and will be derived below. For the 10 most likely categories, a set $\mathcal{R}_2$ of compositions of compositions is formed by means of a *top-down grouping*: Random candidates are drawn from the image and those which are not relevant (cf. Section 3.5) for one of these 10 categories are discarded. Therefore, the initial hypothesis from (10) controls the grouping of higher order compositions. The category posterior of all compositions can then be derived by applying Bayes' rule and exploiting the conditional independences expressed in the graphical model of Figure 3,

$$\begin{aligned} &P\big(c|\mathbf{g}^I, \mathbf{x}, \{\mathbf{g}_j, \mathbf{x}_j\}_{j\in\mathcal{R}_1}, \{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{<k,l>\in\mathcal{R}_2}\big) \\ &= \frac{p\big(\{\mathbf{g}_j, \mathbf{x}_j\}_j|\mathbf{x}, c\big)\, p\big(\{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{<k,l>}|\mathbf{x}, c\big)\, p\big(\mathbf{g}^I|\mathbf{x}, c\big)}{p\big(\{\mathbf{g}_j, \mathbf{x}_j\}_j, \{\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\}_{<k,l>}, \mathbf{g}^I|\mathbf{x}\big)} \\ &\quad \times P\big(c|\mathbf{x}\big) \ . \end{aligned} \quad (11)$$

Now we can neglect the evidence in the denominator as it is independent of c. Again we exploit the conditional independence between compositions conditioned on c and $\mathbf{x}$,

$$\begin{aligned} \ldots \propto P\big(c|\mathbf{x}\big) \cdot p\big(\mathbf{g}^I|\mathbf{x}, c\big) \times \prod_{j\in\mathcal{R}_1} p\big(\mathbf{g}_j, \mathbf{x}_j|\mathbf{x}, c\big) \\ \times \prod_{<k,l>\in\mathcal{R}_2} p\big(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}|\mathbf{x}, c\big) \ . \end{aligned} \quad (12)$$

Applying Bayes' rule to the likelihoods yields

$$\begin{aligned} \cdots = P\big(c, \mathbf{g}^I|\mathbf{x}\big) \times \prod_{j\in\mathcal{R}_1} \frac{P\big(c|\mathbf{x}, \mathbf{g}_j, \mathbf{x}_j\big) \cdot p\big(\mathbf{g}_j, \mathbf{x}_j|\mathbf{x}\big)}{p\big(c|\mathbf{x}\big)} \\ \times \prod_{<k,l>\in\mathcal{R}_2} \frac{P\big(c|\mathbf{x}, \mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\big) \cdot p\big(\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}|\mathbf{x}\big)}{p\big(c|\mathbf{x}\big)} \ . \end{aligned} \quad (13)$$

Neglecting factors that are independent of c and exploiting the fact that object categories are independent of object location yields

$$\begin{aligned} \ldots \propto \exp\Big[\ln P\big(c|\mathbf{g}^I\big) + \sum_{j\in\mathcal{R}_1} \ln P\big(c|\mathbf{g}_j, S_j = \mathbf{x} - \mathbf{x}_j\big) \\ + \sum_{<k,l>\in\mathcal{R}_2} \ln P\big(c|\mathbf{g}_k, \mathbf{g}_l, \mathbf{r}_{kl}\big)\Big] \ . \end{aligned} \quad (14)$$

The first term represents scene context, the second shape based on compositions, and the third relations between object components. The last two distributions are estimated with NKDA on the training data, whereas the first one has already been computed for (3). This model does not only recognize an object, but it also returns a confidence in this prediction.

4. Evaluation of the Compositional Approach

4.1. Results on Caltech-101

The Caltech-101 image database [6] consists of 101 object categories and an additional background class. It contains approximately 30 to 800 images per category which range from line drawings to photos with clutter. The large intra-category variations in this database render the categorization of images a challenging task. However, it features only limited variations in pose. For evaluation we use the standard experimental setup, namely, we train on 30 samples per class and test on the rest. We follow the common practice of averaging retrieval rates per class to avoid a bias to the easier classes with more samples. Moreover, 5-fold cross-validation is used to obtain error bars, *i.e*. the same algorithm is run on 5 different splits of the data into training and test set.

To evaluate the gain of our compositional approach, we restrict the categorization system to the bag representation $\mathbf{g}^I$ in a baseline experiment. Recognition is then based on maximizing $P(c|\mathbf{g}^I)$. This model discards the learned compositional structure and achieves a retrieval rate of $35.3 \pm 0.8\%$. In contrast to this, the full compositional model performs at $\mathbf{58.8 \pm 0.9\%}$. Recently we have extended the model to incorporate multiple scales by additionally establishing compositions of atomic parts from half and a fourth of the original image scale. This multi-scale extension has increased retrieval rate to $\mathbf{61.3 \pm 0.9\%}$. Table 1 gives an overview over the state of the art. 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. [15] 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.[1]

[26]	[15]	[11]	[19]	[22]
66.2 ± 0.5	64.6 ± 0.8	58.2	56	53.0 ± 0.5

Table 1. Retrieval rates (in percentage) of current approaches on the Caltech-101 database using 30 training images per category.

[1] **Caltech-256:** A preliminary experiment on the newly released Caltech-256 database [12] (similar in style to Caltech-101 but consisting of 256 object categories and a class for clutter) achieved 12% retrieval rate for 5 training images per category compared with roughly 18% reported by Griffin *et al*. using the method from [15].

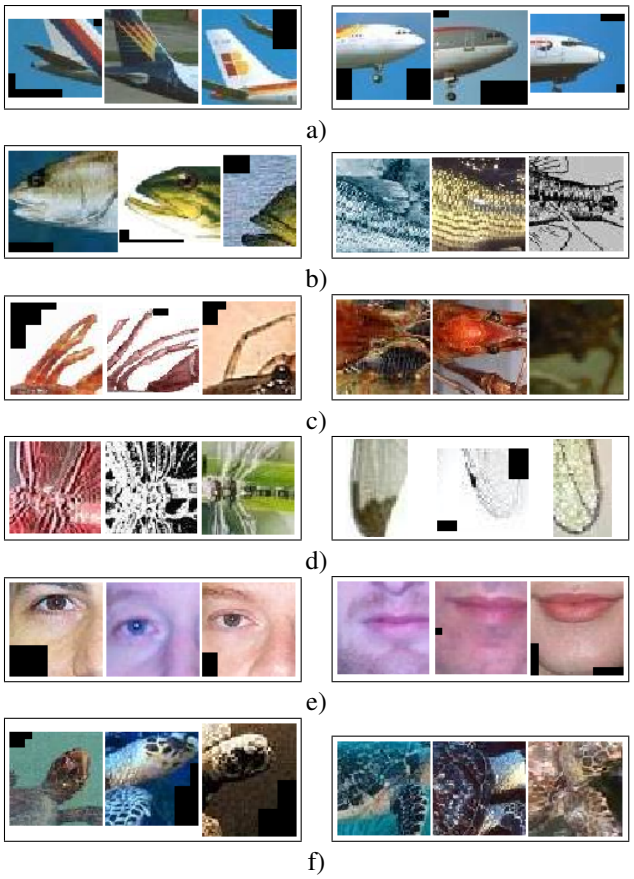


Figure 4. Clustering of relevant compositions. For each category, the two centroids with highest relevance are illustrated by visualizing the closest compositions to that prototype. a) airplanes, b) bass, c) crayfish, d) dragonfly, e) faces, and f) hawksbill.

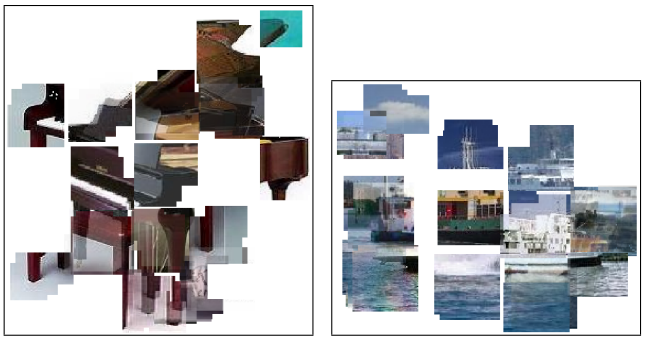
In terms of computational cost, training our algorithm on roughly 3000 and testing on 6000 Caltech-101 images takes about 15 hours on an ordinary PC. To our knowledge this is a competitive speed for a structured object model. Moreover, restricting the localized feature histograms to only grayscale decreases the retrieval rate by roughly 1.5%.

4.2. Analyzing the Learning of Relevant Compositions

The following experiment analyzes the learning of relevant compositions. Therefore, all compositions from the training data that have been predicted to be relevant for a category c are clustered. The centroids that contain compositions with highest relevance (averaged over all cluster members) are presented in Figure 4.

4.3. Sampling Compositions Using the Generative Model

During recognition, inference propagates information from image features over compositions to an object cate-



a) b)

Figure 5. 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. Image patches corresponding to the inferred compositions are then displayed.

gory label. However, the graphical model from Figure 3 can also be used in a generative mode: Given object category c and object position $\mathbf{x}$, compositions and, finally, image patches can be inferred. To obtain the image representation in a region around $\mathbf{x}_j$, compositions $\mathbf{g}_j$ have to be sampled from the likelihood

$$p\big(\mathbf{g}_j|c,\mathbf{x},\mathbf{x}_j\big) = \frac{P\big(c|\mathbf{g}_j,\mathbf{x}-\mathbf{x}_j\big)\cdot p\big(\mathbf{g}_j|\mathbf{x}-\mathbf{x}_j\big)}{P\big(c|\mathbf{x},\mathbf{x}_j\big)} \ . \quad (15)$$

The denominator can be dropped since c is an evidence variable in this experiment. Moreover, all compositions that are established in the training images using the approach from Section 3.2 are distributed according to the composition prior $\mathbf{g}_j \sim p(\mathbf{g}_j|\mathbf{x}-\mathbf{x}_j)$. Compositions can therefore be sampled by evaluating the category posterior $P(c|\mathbf{g}_j,\mathbf{x}-\mathbf{x}_j)$ (which has been learned for (14)) on compositions of the training data,

$$p\big(\mathbf{g}_j|c,\mathbf{x},\mathbf{x}_j\big) \propto P\big(c|\mathbf{g}_j,\mathbf{x}-\mathbf{x}_j\big)\big|_{\mathbf{g}_j \text{ from training}} \ . \quad (16)$$

The resulting compositional image puzzles in Figure 5 provide insights into this generative process. Here compositions have been inferred at points $\mathbf{x}_j$ on a regular grid (5 compositions have been drawn at each point). The image patches from the training image that constituted a specific composition are then displayed (the sampled compositions can shift a short distance by performing gradient ascent on the likelihood (16) over $\mathbf{x}_j$ in a local neighborhood). This experiment reveals that the composition system has learned relevant compositions and their spatial relation to the object, and that it can be used as a generative model for inferring compositional representations.

4.4. Inferring Missing Object Components

The compositions of compositions which have been introduced in Section 3.5 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

$$p\big(\mathbf{g}_j|\mathbf{g}_k,\mathbf{x}_k,c,\mathbf{x}_j\big) = \frac{P\big(c|\mathbf{g}_j,\mathbf{g}_k,\mathbf{r}_{jk}\big)\cdot p\big(\mathbf{g}_j|\mathbf{g}_k,\mathbf{r}_{jk}\big)}{P\big(c|\mathbf{g}_k,\mathbf{x}_k,\mathbf{x}_j\big)} \ . \quad (17)$$

Following the line of reasoning from Section 4.3 we obtain

$$p\big(\mathbf{g}_j|\mathbf{g}_k,\mathbf{x}_k,c,\mathbf{x}_j\big) \propto P\big(c|\mathbf{g}_j,\mathbf{g}_k,\mathbf{r}_{jk}\big)\big|_{\mathbf{g}_j \text{ from training}} \ . \quad (18)$$

In Figure 6, a single composition is fixed together with the object category label. This information is used to infer a maximum likelihood solution for the remainder of the object on the basis of compositions derived from the training set. As already done in Section 4.3, compositions are shifted in a local neighborhood using gradient ascent to reduce the artifacts that result from sampling on a regular grid. The spatial structure of objects, which can be observed in the reconstructions, demonstrates that the compositional model has learned characteristic relationships between compositions.

a) b)

Figure 6. 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.

4.5. Towards Learning Category Level Segmentation from Unsegmented Images

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

$$P\big(c|\mathbf{g}^I,\mathbf{x},\mathbf{g}_j,\mathbf{x}_j,\mathbf{g}_k,\mathbf{g}_l,\mathbf{r}_{kl}\big)\big|_{c=\text{True Category}} \ , \quad (19)$$

is computed for individual pairs of compositions. In Figure 7 the resulting probability is then encoded in the opaqueness of the underlying image parts, *i.e.* alpha blending is used for visualization of the category relevance of compositions.

5. Discussion

In this contribution, we have presented a composition system that automatically learns the compositional structure

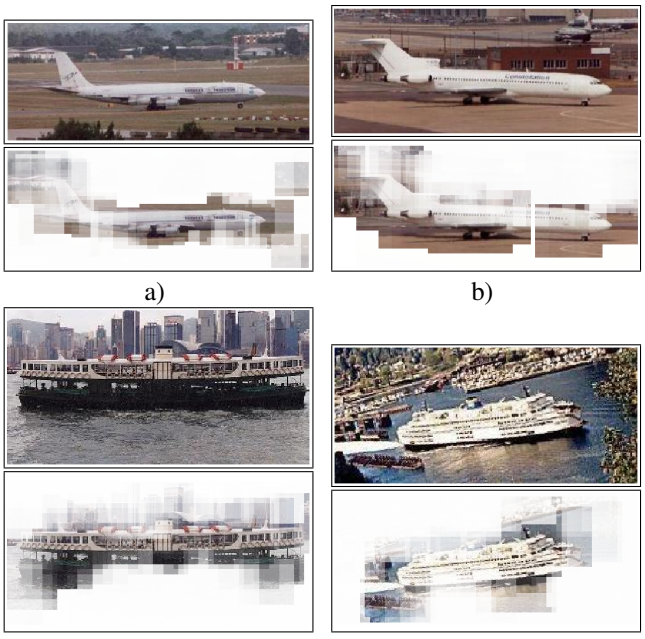


Figure 7. Illustration of compositional relevance. The opaqueness encodes the category posterior evaluated at compositions. The visualization shows which image patches contributed to recognizing the object.

of visual objects. The induced representation is based on a shared codebook of local parts. The semantic gap between these low level features and high level object recognition is bridged by establishing intermediate compositions. We have taken a Bayesian approach to derive a criterion for the relevance of compositions. A feasible learning algorithm has then been presented that is controlled by this criterion. Moreover, compositions have been represented as probability distributions over their constituents and all compositions have been integrated in a Bayesian network. Recognition is then formulated as an inference problem in this statistical model. The experimental validation has shown that our composition system is successful in learning the compositional structure of objects and that it shows competitive recognition performance.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 26(11), 2004.

[2] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, 2005.

[3] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Rev.*, 94(2), 1987.

[4] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV Workshop on Stat. Learn. in Comp. Vis.*, 2004.

[5] B. Epshtein and S. Ullman. Feature hierarchies for object classification. In *ICCV*, 2005.

[6] 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. In *CVPR Workshop on GMBV*, 2004.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Computer Vision*, 61(1), 2005.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[9] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1), 1973.

[10] S. Geman, D. F. Potter, and Z. Chi. *Composition Systems. Quarterly of Applied Mathematics*, 60, 2002.

[11] K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. Technical Report MIT-CSAIL-TR-2006-020, 2006.

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

[13] A. D. Holub, M. Welling, and P. Perona. Combining generative models and fisher kernels for object class recognition. In *ICCV*, 2005.

[14] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *CVPR*, 2006.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] B. Leibe and B. Schiele. Scale-invariant object categorization using a scale-adaptive mean-shift search. In *Pattern Recognition, DAGM*, 2004.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2), 2004.

[18] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Computer Vision*, 60(1), 2004.

[19] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *CVPR*, 2006.

[20] B. Ommer and J. M. Buhmann. Object categorization by compositional graphical models. In *EMMCVPR*, 2005.

[21] B. Ommer and J. M. Buhmann. Learning compositional categorization models. In *ECCV*, 2006.

[22] B. Ommer, M. Sauter, and J. M. Buhmann. Learning top-down grouping of compositional hierarchies for recognition. In *CVPR Workshop on Percept. Org. in Comp. Vis.*, 2006.

[23] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *CVPR*, 2006.

[24] V. Roth and K. Tsuda. Pairwise coupling for machine recognition of hand-printed japanese characters. In *CVPR*, 2001.

[25] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, 2005.

[26] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.