Learning Top-Down Grouping of Compositional Hierarchies for Recognition

Björn Ommer, Michael Sauter, and Joachim M. Buhmann*
Institute of Computational Science, ETH Zurich
8092 Zurich, Switzerland
{bjoern.ommer, misauter, jbuhmann}@inf.ethz.ch

Abstract

The complexity of real world image categorization and scene analysis requires compositional strategies for object representation. This contribution establishes a compositional hierarchy by first performing a perceptual bottom-up grouping of edge pixels to generate salient contour curves. A subsequent recursive top-down grouping yields a hierarchy of compositions. All entities in the compositional hierarchy are incorporated in a Bayesian network that couples them together by means of a shape model. The probabilistic model underlying top-down grouping as well as the shape model is learned automatically from a set of training images for the given categories. As a consequence, compositionality simplifies the learning of complex category models by building them from simple, frequently used compositions. The architecture is evaluated on the highly challenging Caltech 101 database[1] *which exhibits large intra-category variations. The proposed compositional approach shows competitive retrieval rates in the range of* $53.0 \pm 0.49\%$.

1. Introduction

Object categorization, which has received increasing attention over the last years, aims at recognizing visual objects of some general class in scenes. Categorization is widely considered as a subtask of the long standing, major goal of computer vision to automatically detect and recognize objects in unconstraint scenes. Large intra-category variations of appearances and instantiations of the same object category turn representing and learning category models into a difficult challenge. Learning algorithms have to capture common characteristics of a category while simultaneously providing invariance with respect to variations or absence of features.

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

[1] www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

Learning Compositional Hierarchies: This contribution proposes a system that learns to group parts of a scene into a hierarchy of category-specific compositions, and binds them together using a probabilistic shape model to categorize scenes. The parts used for this top-down grouping are in turn agglomerations of atomic local features that are grouped using a bottom-up, perceptual organization strategy.

The principle of *compositionality* [10] lays the foundation for the approach presented in this contribution: As observable in cognition in general and especially in human vision (see [3]), complex entities are perceived as compositions of comparably few, simple, and widely usable parts. Objects are then represented based on their components and the relations between them. In contrast to modeling the constellation of parts directly (*e.g.* [8]), the compositionality approach learns intermediate groupings of parts—possibly even forming a hierarchy of recursive compositions [18]. As a result compositions bridge the semantic gap between low level image features and high level scene categorizations [19, 20] by establishing an intermediate hidden layer representation. The fundamental concept is then to find a trade-off between two extremes: On the one hand objects have high intra-category variations so that learning representations for whole objects directly becomes infeasible. On the other hand local part descriptors fail to capture reliable information on the overall object category. Therefore compositions represent category-distinctive subregions of an object, which show minor intra-category variations compared to the whole object and turn learning them into a feasible problem.

Learning a compositional hierarchy is divided into two subproblems. Firstly, image parts are grouped in a data-driven, bottom-up manner. Secondly, these intermediate compositions are then grouped in a top-down fashion dependent on the specific category models. The underlying rationale for the two stage approach is that grouping in the first stage is basically driven by similarity of parts and by the minimum description length principle. As this approach is mainly controlled by local observations in an image, it

0-7695-2646-2/06 $20.00 (c) 2006 IEEE

has a high tendency to group constituents with similar local statistics. Therefore, such a grouping process can enrich the descriptiveness of the shape of compositions (their spatial structure). However, compositions are likely to have homogeneous feature distributions which provide no additional information compared to their parts, since the descriptors of the constituent image regions tend to be fairly similar. In contrast to this similarity driven structure extraction, a top-down grouping process forms agglomerations of constituents based on their distinctiveness for a category. Such groupings constitute characteristic combinations of dissimilar object parts to cover the heterogeneity of real world objects. The final challenge is then to automatically learn top-down grouping models for great numbers of categories without any explicit information about the compositional structure of objects in the training data. This problem is tackled by first approximating category dependent co-occurrence statistics on the training data and using them to form a hierarchy of potential grouping candidates. Using this compositional hierarchy the grouping model is then being refined. In other words, we start with simple and robust category statistics which are then used to guide the system during its investigation of increasingly complicated compositions that are in turn utilized to refine the statistics.

In this contribution the first problem of bottom-up grouping is addressed by *perceptual organization* [15]. Therefore, edge contours are grouped on the basis of *Gestalt laws* to yield salient contour curves which are then represented by localized feature histograms [19]. Top-down grouping is then conducted by forming compositions that are most likely, given the previously learned category models. This process is applied recursively to construct a hierarchy of compositions.

2. Related Work

Typically, the problem of object categorization has been addressed by representing a scene with local descriptors and modeling their configuration in a flexible/adaptive way, *e.g*. [9, 13, 8, 6, 1, 19, 2]. A common choice of local image features are template-based *appearance patches* (e.g. [1, 8, 6, 13]) and histogram-based descriptors such as *SIFT* features [16]. *Geometric blur* by Berg *et al*. [2] and *localized feature histograms* [19] fall in the latter category. Moreover, Serre *et al*. [22] have proposed features that are neuro-physiologically motivated.

A simple and robust way to model the configuration of descriptors are *bag of features* methods such as [5] that establish a histogram over all image features. By this feature extraction step, however, the spatial structure of a scene is discarded. At the other end of the modeling spectrum are *constellation models*, e.g. [24, 8, 6, 12], 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, 13, 19, 20] aim at utilizing greater numbers of image constituents. The compositional approach that has been taken in [20] differs from our method in that it only learns a single layer of part agglomerations using a spatially fixed grouping strategy. In contrast to this, the present work proposes a framework that automatically learns to build hierarchies of compositions for a large number of categories. Therefore, it also substantially differs from supervised methods to model configurations of parts, such as [7]. Another way to construct abstraction hierarchies has been pursued in [14] using many-to-many correspondences between blobs. Moreover, bottom-up and top-down groupings have been applied in [4] to refine figure-ground segmentation of objects from a single class and in [23] to recognize text and faces in images. Finally, an approach that is based on establishing coherent spatial mappings between a probe image and all training images has been taken in [2].

3. Approach

Subsequently, we give an overview over our approach by first considering recognition. The involved processing steps are then covered in detail by later sections. Given a novel image, Canny edge detection is performed. The resulting edge pixels are grouped using a purely bottom-up, perceptual grouping strategy that yields nearly closed arcs or fairly straight curves. Each of these edge curves is then represented by a bag of features. These features (we use the localized feature histograms from [19]) are computed for patches on a regular grid and the bag is then formed by histogramming over all patches that lie on the curve. In a second stage top-down grouping is performed recursively. This step yields agglomerations of curves with increased discriminative power compared to their original constituents. Assume for the moment that groupings which are distinctive for categories have already been learned automatically from the training data. The objective of top-down grouping is then to form a hierarchy of compositions by recursively combining those pairs of constituents whose composition has highest category posterior. Finally, all the groupings are coupled together by means of a shape model.

The rationale underlying these two grouping stages is the following. Bottom-up grouping condenses the information present in an image by forming few salient curves. The underlying perceptual criteria of simplicity and homogeneity however hardly capture heterogeneous compositions that are truly category specific. The main objective of this processing is therefore to condense relevant information in few entities which in turn increases computational feasibility. In contrast to this, top-down grouping is guided by the familiarity of compositions. Heterogeneous agglomerations of constituents that are characteristic for categories have a high saliency according to this principle. The final chal-

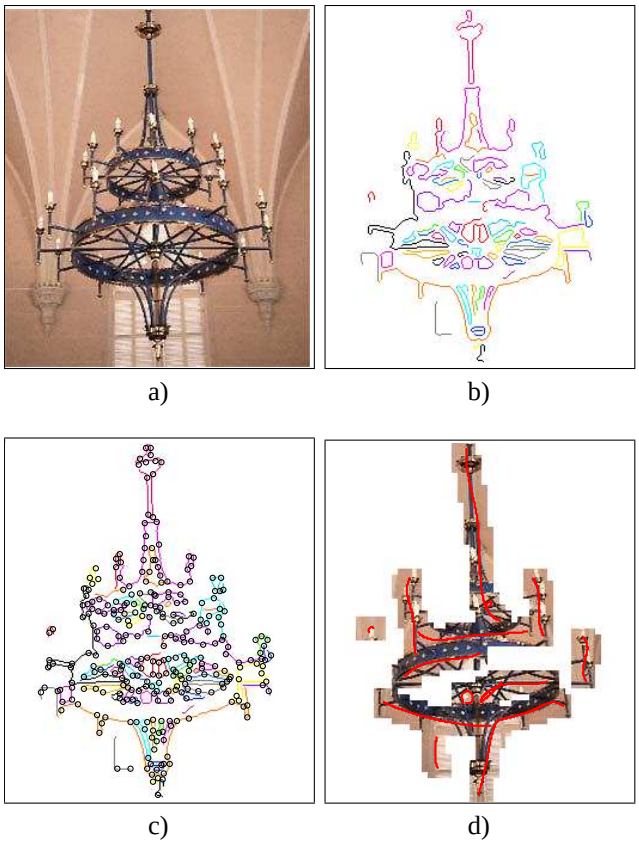


Figure 1. Perceptual bottop-up grouping. a) Original image. b) Connected edge curves from Canny edge detection depicted in the same color. c) Potential high curvature break points. d) Splines fitted to salient curves and illustration of the corresponding image regions.

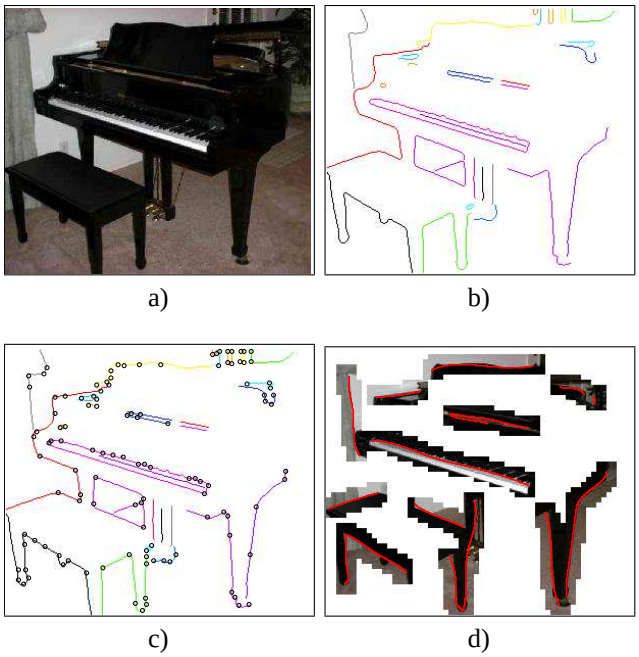


Figure 2. Perceptual bottop-up grouping. a) Original image. b) Connected Canny edge curves. c) Potential high curvature break points. d) Splines fitted to salient curves and visualization of the underlying image regions.

lenge is then to automatically learn and represent models for top-down grouping in the case of large numbers of object classes without extensive user supervision. In other words, how can the system learn which compositions are relevant for a category without being told about the compositional structure of objects? We tackle this problem by first estimating category dependent co-occurrence statistics of bottom-up grouped curves in training images of given categories. Using this distribution the curves are then grouped in a recursive manner, thereby giving rise to a hierarchy of compositions. This hierarchy is finally used to update the previously estimated category dependent statistics with probabilities of higher level groupings and to learn the global shape model. The complexity of the underlying category model is adjusted on the training data using cross-validation.

3.1. Perceptual Bottom-Up Grouping

The primary objective of bottom-up grouping is to find a comprehensive image representation based on salient edge curves that is yet compact. Processing of an image starts by performing Canny edge detection and finding connected edge curves as illustrated in Figure 1 b). This step, however, yields curves of any degree of complexity. To find salient edge curves we therefore continue by first breaking contours into fairly simple parts before grouping them again in a perceptually controlled manner.

The complexity of a curve grouping is examined with respect to the following *Gestalt laws* of perceptual organization [15]: *good continuation* (preferring curves with smooth continuity), *proximity* (avoiding large gaps), and *convexity* (short curves circumscribing large areas). The underlying idea is to look for curves that remain stable and prominent over different realizations of an object category despite the large intra-category variations. Roughly speaking we are interested in smooth elongated curves (no convexity, but maximal smoothness) or nearly circular arcs (maximal convexity). These two cases constitute the extrema of a criterion function $\zeta(\gamma)$ for curves γ. Let $A(\gamma)$ denote the area circumscribed by the curve and $l(\gamma)$ be its length then

$$\zeta(\gamma) := \frac{A(\gamma)}{A(\text{circle with perimeter } l(\gamma))} \tag{1}$$

$$= 4\pi \frac{A(\gamma)}{l^2(\gamma)} \quad . \tag{2}$$

For straight lines, ζ is zero and for circles it is one. There-

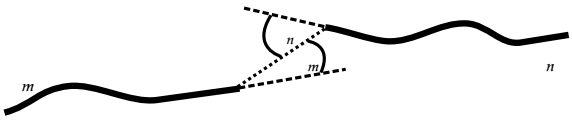


Figure 3. Measuring the smoothness of a grouping that merges curves γ_m and γ_n.

fore it is suitable to maximize the criterion function

$$\widetilde{\zeta}(\gamma) := \left|\zeta(\gamma) - \tfrac{1}{2}\right| \quad . \tag{3}$$

Breaking contours into simple parts is then carried out as follows: Curves with almost maximal criterion function $\widetilde{\zeta}$ are kept unaltered as they are already nearly straight or circular. Otherwise they are split at the point of highest curvature as shown in Figure 1 c). The resulting two segments are then in turn processed recursively. As the resulting curvelets will be merged subsequently, a splitting into too short segments is not critical.

A cubic B-spline is fitted to each of the curvelets to remove small wiggles. The spline curves are then ranked in a queue according to $\widetilde{\zeta}$. For the curve γ_m with maximal $\widetilde{\zeta}$ its pairwise groupings with all other curves γ_n are evaluated by fitting a spline to each resulting composition γ_g and computing $\widetilde{\zeta}(\gamma_g)$. To enforce smoothness of the grouping those compositions will be removed for which at least one of the angles α_m and α_n between the tangent of a curve and the connection between the two curves is greater than $90°$ as illustrated in Figure 3. To follow the principle of proximity, a grouping is also discarded if the gap between the two curves is longer than $\min\{l(\gamma_m), l(\gamma_n)\}$, the length of the shortest of the two constituent curves. Finally, compositions γ_g are removed if they do not improve the criterion function $\widetilde{\zeta}$ in comparison to their constituents. To summarize, composition candidates γ_g formed from γ_m and γ_n will be discarded in the following cases

$$\text{discard } \gamma_g \Leftrightarrow \begin{cases} \min\{\alpha_m, \alpha_n\} > 90° \vee \\ \text{gap}(\gamma_m, \gamma_n) > \min\{l(\gamma_m), l(\gamma_n)\} \vee \\ \widetilde{\zeta}(\gamma_g) < \min\{\widetilde{\zeta}(\gamma_m), \widetilde{\zeta}(\gamma_n)\} \end{cases} \tag{4}$$

The grouping γ_g with maximal $\widetilde{\zeta}$ is chosen among the remaining candidates and it is added to the queue while both of its constituents are removed. If the set of candidates is empty, only γ_m will be removed. This curve merging continues with the currently best curve in the queue until there is only one left. In the subsequent stages of the architecture, all created groupings and those curves that could not be merged with another curve are processed further.

3.2. Forming Robust Descriptors for Salient Curves

Each contour that is generated by above bottom-up grouping has to be represented in such a way that curves of varying length and number of constituent curves are possible. Therefore we use a slight variation of bags over localized feature histograms proposed in [19]. On a regular grid (spacing of 5 pixels) quadratic patches with a side length of 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 these 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 obtained. To robustify the representation each feature is not merely described by its nearest prototype but by a Gibbs distribution over the codebook: 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,

$$P(F_i = \nu|\mathbf{e}_i) := Z(\mathbf{e}_i)^{-1} \exp\left(-d_\nu(\mathbf{e}_i)\right) \quad , \tag{5}$$

$$Z(\mathbf{e}_i) := \sum_\nu \exp\left(-d_\nu(\mathbf{e}_i)\right) \quad . \tag{6}$$

For each pixel on a curve generated in Section 3.1 the closest feature patch $\mathbf{e}_i$ is selected. All these patches are collected and duplicates are removed (see Figure 1 d)) before forming a bag of features as in [20]. Therefore, a curve is represented as a mixture over the distributions (5) of its parts. 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 curve is then represented by the vector valued random variable G_j which is a bag of features, i.e. its value $\mathbf{g}_j$ is a distribution over the k-dimensional codebook from above

$$\mathbf{g}_j \propto \sum_{i=1}^{m} \Big(P(F_i = 1|\mathbf{e}_i), \ldots, P(F_i = k|\mathbf{e}_i)\Big)^T \quad . \tag{7}$$

This mixture model has the favorable property of robustness with respect to variations in the individual parts.

3.3. Unsupervised Learning of Top-Down Grouping

In order to be able to group parts of a novel image in a top-down manner, the system first has to learn to do so from the training data. This step is carried out automatically with just the training images and their corresponding category label, but without any further supervision. How can the system then learn what to group without being instructed about the compositional nature of objects? The key idea is to estimate co-occurrence statistics of constituents and to produce groupings based on this information. These groupings are then used to update the compositional statistics.

Salient Base Compositions: Processing of a training image starts by selecting a subset of the curves generated in Section 3.1. Therefore, interest point detection is performed (using the scale invariant Harris interest point detector from [17]) and all those patches extracted in Section 3.2 are marked which cover at least a single interest point (IP). The idea is then to find the most salient curves γ using the score function

$$\xi(\gamma) := l(\gamma) \cdot \frac{\#\,\text{patches with IP on } \gamma}{\#\,\text{patches on } \gamma} \quad . \tag{8}$$

From the set of all grouped curves we choose the 7 with maximal score $\xi(\gamma)$. From the remainder, at most 4 curves with minimal $\zeta(\gamma)$ (curves that are most circular) are selected; all other curves are discarded. To cover regions not represented by curves, 3 seed points are chosen from the set of all interest points. All patches with interest points that are not farther than 50 pixels from such a seed point are combined to yield 3 additional groupings. The selected curves and additional groupings are collected in the set Γ_{C_0} and form atomic base compositions for the subsequent top-down grouping. They are, however, groupings themselves and each is represented by a bag of features $\mathbf{g}_j$ as described in Section 3.2.

Approximating Grouping Probabilities Using Initial Groupings of Base Compositions: For each pair of base compositions $\mathbf{g}_i, \mathbf{g}_j$ a grouping $\mathbf{g}_{ij}$ is established. It is represented by the mixture over its constituent feature histograms from (7). The advantage of such a representation is that all compositions are encoded in the same feature space, independently of their level in the compositional hierarchy and the number of atomic patches they cover. Let $\mathcal{L}$ denote the set of all category labels and $c \in \mathcal{L}$ be the category label of the image under consideration. For the initial training step all the groupings $\mathbf{g}_{ij}$ which have been formed in all the training images are combined. These samples are then used to learn a first approximation of the category posterior of groupings $\mathbf{g}_{ij}$

$$P\left(C = c|\mathbf{g}_{ij}\right) \quad . \tag{9}$$

This distribution is learned by training probabilistic two-class kernel classifiers on all the training samples. For the two-class classification we choose *nonlinear kernel discriminant analysis* (NKDA)[21] and perform a pairwise coupling to solve the multi-class problem (see [11, 21]). The rationale behind our choice is that a joint optimization over all classes (one vs. all classifiers) is unnecessarily hard and computationally much more costly than solving the simpler pairwise subproblems. The combined probabilistic classifier yields an estimate of the posterior (9) for the respective image category.

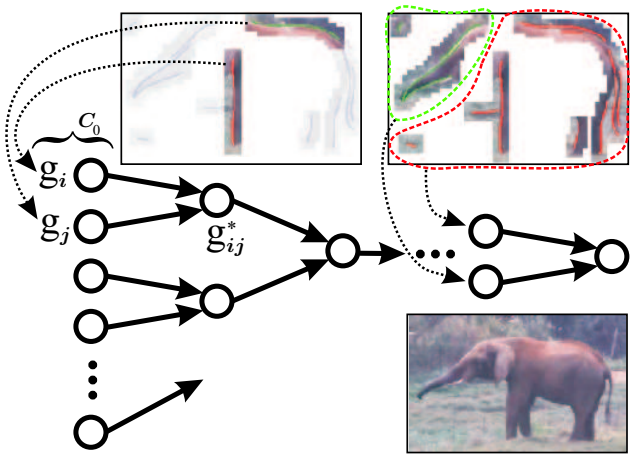


Figure 4. Sketch of the compositional hierarchy. The leafs constitute the set Γ_{C_0} of base compositions formed by bottom-up grouping. As an example the first and last grouping in the hierarchy for the given image of an elephant are illustrated.

Forming a Compositional Hierarchy: The goal is now to recursively form top-down groupings guided by the posterior (9) to obtain a hierarchy as illustrated in Figure 4. Firstly a list Γ_C of all grouping candidates is established by inserting all base compositions from Γ_{C_0}. Moreover, all base compositions $\mathbf{g}_i, \mathbf{g}_j$ are grouped in a pairwise manner, yielding compositions $\mathbf{g}_{ij}$. As discussed at the beginning of Section 3, among all $\mathbf{g}_{ij}$ the grouping with maximal posterior

$$\mathbf{g}^\star_{ij} = \underset{\mathbf{g}_{ij}:\mathbf{g}_i,\mathbf{g}_j \in \Gamma_C}{\operatorname{argmax}} P(c|\mathbf{g}_{ij}) \tag{10}$$

is selected and added to the list of candidates Γ_C and the constituents are removed

$$\Gamma_C \leftarrow \Gamma_C \cup \{\mathbf{g}^\star_{ij}\} - \{\mathbf{g}_i, \mathbf{g}_j\} \quad . \tag{11}$$

Now $\mathbf{g}^\star_{ij}$ is grouped in a pairwise manner with all remaining elements of Γ_C. Then recursive grouping continues again with (10) to find the next best composition until there is only one element in Γ_C. In conclusion a hierarchy of groupings in the form of a binary tree is established as illustrated in Figure 4. Base compositions constitute the leafs whereas the last remaining element of Γ_C forms the root.

Local Maxima of the Compositional Hierarchy: Subsequently, a subset of all groupings in the hierarchy is to be selected. For each leaf of the hierarchy (illustrated in Figure 4) the path to the root is followed and all locally optimal groupings on this path are collected. A grouping is locally optimal if its category posterior (9) is greater than that of its predecessor and successor node which lie on the path to the root. After removing duplicates this processing yields the set Γ_L of all compositions with locally maximal posterior.

Finally, the category posterior in (9) is updated by training the classifiers with all locally optimal compositions

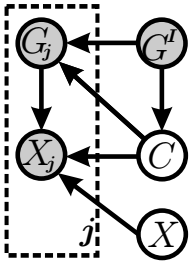


Figure 5. Bayesian network that couples compositions G_j using their locations X_j, the location of the object center X, a bag of features G^I, and image categorization C. Shaded nodes denote evidence variables. See text for details.

$\mathbf{g} \in \Gamma_L$ established for all the training images. This updated posterior guides top-down grouping in the recognition phase.

3.4. Applying Top-Down Grouping to Recognition

In the recognition phase a novel probe image is processed by bottom-up grouping and feature extraction as described in Section 3.1 and Section 3.2, respectively. Moreover, salient base compositions are selected as in Section 3.3. Thereafter, the final estimate of the category posterior (9) is used to form a hierarchy of compositions as in the previous section. In contrast to training, the correct category label of the image needed for eq. (10) is not given, during recognition. Therefore, (10) is replaced by

$$\mathbf{g}_{ij}^{\star} = \underset{\mathbf{g}_{ij}:\mathbf{g}_i,\mathbf{g}_j \in \Gamma_C}{\operatorname{argmax}} \max_{c \in \mathcal{L}} P(c|\mathbf{g}_{ij}) \tag{12}$$

in the recognition phase. Similarly, the set of local maxima Γ_L can only be found when using $\max_{c \in \mathcal{L}} P(c|\mathbf{g})$ as the criterion to maximize.

3.5. Shape Model for Composition Binding

Subsequently, all compositions $\mathbf{g}_j \in \Gamma_L$ of an image that have been selected as local maxima in the compositional hierarchy are coupled on the basis of a generalized version of the shape model in [20]. Moreover, a composition $\mathbf{g}^I$ of all parts $\mathbf{e}_i$ in the image, i.e. a bag of features descriptor for the whole image, is employed for binding the compositions. The underlying graphical model is depicted in Figure 5.

To determine the object location $\mathbf{x}$, the positions $\mathbf{x}_j$ of all previously generated, locally optimal compositions $\mathbf{g}_j \in \Gamma_L$ are considered for this estimate. The position of the object center is then estimated by weighing the contribution of each composition with the probability that it would be observed

$$\mathbf{x} = \sum_j \mathbf{x}_j \sum_{c \in \mathcal{L}} p(\mathbf{g}_j|c, \mathbf{g}^I)\ P(c|\mathbf{g}^I)\ . \tag{13}$$

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

In the following, the graphical model depicted in Figure 5 is used to couple all compositions $\mathbf{g}_j$, their locations $\mathbf{x}_j$, the previously estimated object center $\mathbf{x}$, and the bag of features descriptor $\mathbf{g}^I$ to infer the image category c:

$$P\big(c|\mathbf{g}^I, \mathbf{x}, \{\mathbf{g}_j, \mathbf{x}_j\}_{j=1:|\Gamma_L|}\big) \tag{14}$$

$$= \frac{p\left(\{\mathbf{g}_j, \mathbf{x}_j\}_j|c, \mathbf{g}^I, \mathbf{x}\right) P(c|\mathbf{g}^I, \mathbf{x})}{p\left(\{\mathbf{g}_j, \mathbf{x}_j\}_j|\mathbf{g}^I, \mathbf{x}\right)} \tag{15}$$

$$= P(c|\mathbf{g}^I, \mathbf{x}) \times \prod_{\mathbf{g}_j \in \Gamma_L} \frac{P\left(c|\mathbf{g}^I, \mathbf{x}, \mathbf{g}_j, \mathbf{x}_j\right) \cdot p\left(\mathbf{g}_j, \mathbf{x}_j|\mathbf{g}^I, \mathbf{x}\right)}{p\left(\mathbf{g}_j, \mathbf{x}_j|\mathbf{g}^I, \mathbf{x}\right) \cdot P(c|\mathbf{g}^I, \mathbf{x})} \tag{16}$$

$$= \left[P(c|\mathbf{g}^I)\right]^{1-|\Gamma_L|} \prod_{\mathbf{g}_j \in \Gamma_L} P\left(c|S_j = \mathbf{x} - \mathbf{x}_j, \mathbf{g}_j, \mathbf{g}^I\right) \tag{17}$$

$$= \exp\Big[(1 - |\Gamma_L|) \ln P(c|\mathbf{g}^I) + \sum_{\mathbf{g}_j \in \Gamma_L} \ln P\left(c|S_j = \mathbf{x} - \mathbf{x}_j, \mathbf{g}_j, \mathbf{g}^I\right)\Big] \tag{18}$$

Equation (17) relies on the assumption that categorization is translation invariant. Moreover, the relative location of a composition with respect to the object center is represented by the shift $\mathbf{s}_j := \mathbf{x} - \mathbf{x}_j$. Here, we exploit the fact that categorization is not depending on the actual position of the object center but that it only depends on relative shifts. Introducing the logarithm in the last step enhances numerical stability.

The first distribution in (18) has already been estimated for (13). The latter distribution is again estimated using NKDA from the training data. In conclusion, novel images cannot only be assigned a category label, but also a confidence value for this categorization.

4. Experiments

Our categorization model based on perceptual grouping is evaluated on the challenging Caltech 101 database. The dataset contains 101 object categories and a background category, each with approximately 30 to 800 samples. Images range from line drawings to photos with clutter, but they show only limited variations in pose. As categories are having different numbers of samples (the easier ones have larger numbers of images than the complicated ones), retrieval rates that are estimated over all test images tend to be too optimistic. A common practice is therefore to average over the retrieval rates computed for each category separately. Using texton histograms a reasonable baseline performance of 16% has been calculated by Berg *et al*. for this database in [2] (random classification by chance is below 1%). Moreover, they have proposed an approach based on shape correspondence that yields a retrieval rate of 48%. In [20], Ommer and Buhmann have learned a non-hierarchical compositional model which performs at $53.6 \pm 0.88\%$ for a

single scale approach and $57.8 \pm 0.79\%$ for multiple scales. Using a constellation model, Fei-Fei *et al*. [6] have reported a performance of 16%. This generative model has been enhanced by means of a discriminative classifier [12] and a fusion of multiple interest point detectors to yield a retrieval rate of 40.1%. Finally, Serre *et al*. [22] have introduced neuro-physiologically motivated features that achieve a performance of 42%.

4.1. A Baseline Model without Compositionality

The presented approach establishes a hierarchy of compositions between the initial feature representation of a scene and its final categorization. To estimate the gain of compositionality, this hidden representational layer is neglected, in the following. Therefore, images are categorized by combining all the descriptors in the bag of features representation $\mathbf{g}^I$ described in Section 3.5. Evaluation is then conducted by randomly collecting up to 30 training images per category and taking the remainder as test set. The retrieval rate and its error is estimated by performing 5-fold cross-validation, *i.e*. the same algorithm is run on splits of the data into five different training and test sets. For the $k = 200$ dimensional codebook introduced in Section 3.2 this base model achieves a retrieval rate of $41.3 \pm 0.38\%$.

4.2. Evaluation of the Learning Approach to Forming Compositional Hierarchies

In the following, the entire compositional approach is evaluated under 2-fold cross-validation. It yields a competitive retrieval rate of $\mathbf{53.0 \pm 0.49}\%$ (note that the current approach does not use the multi-scale features of [20]). Figure 6 a) shows the corresponding category confusion table after a permutation of the category labels which is described in Section 4.3. The observable gain in performance over the baseline model from above emphasizes the advantage of an intermediate compositional representation layer in contrast do a direct categorization. A further investigation of the full model shows that the best performing categories are "car", "motorbike", and "pagoda", the worst ones are "panda", "strawberry", and "ant". The most prominent pairwise confusions are "water-lilly" vs. "lotus", "crocodile" vs. "crocodile head", and "panda" vs. "soccer ball". These confusions are between pairs that are either semantically very close or visually similar.

Finally, Figure 6 b) shows an evaluation of the sparseness of the image representation induced by the grouped curves. Therefore, the fraction of all local image features that are used to describe the grouped curves in an image is measured over all test images. On average, 7.7% of all local features are used, yielding a fairly sparse representation.

4.3. Class Hierarchies for Analyzing Categorization

The categorization which has been established in Section 4.2 induces a hierarchical structure among the categories which reveals the degree of relatedness of categories. Therefore, the category confusion probabilities are used to measure the mutual similarities between categories in the database. The final goal is then to establish a hierarchy 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 $P(c_{\text{pred}}|c_{\text{true}})$. The complete category confusion table is then represented by the matrix

$$\mathbf{M}_{c_{\text{true}},c_{\text{pred}}} := P(c_{\text{pred}}|c_{\text{true}}) \ . \tag{19}$$

The matrix is symmetrized by adding its transpose

$$\widetilde{\mathbf{M}} := \eta\mathbf{E} - \left(\mathbf{M} + \mathbf{M}^T - 2\,\mathrm{diag}[\mathbf{M}]\right) \ . \tag{20}$$

Here $\mathbf{E}$ denotes the matrix of only ones, η is a constant and $\mathrm{diag}[\mathbf{M}]$ is $\mathbf{M}$ with its off-diagonal entries set to zero. The resulting matrix $\widetilde{\mathbf{M}}$ is used as a distance matrix between categories for a subsequent hierarchical clustering of categories. For this step, *Ward's Method* with its minimum variance concept is applied. As a result a hierarchical cluster tree is obtained with categories at the leafs and sets of similar categories at inner nodes. Dissimilar categories are connected by long paths over inner nodes near the root, whereas similar ones are connected over shorts paths close to the leafs. Figure 6 a) shows the hierarchical cluster tree. Moreover the category confusion table is presented after having permuted both its rows and columns in the same way so that they fit to the leafs of the adjacent hierarchy tree.

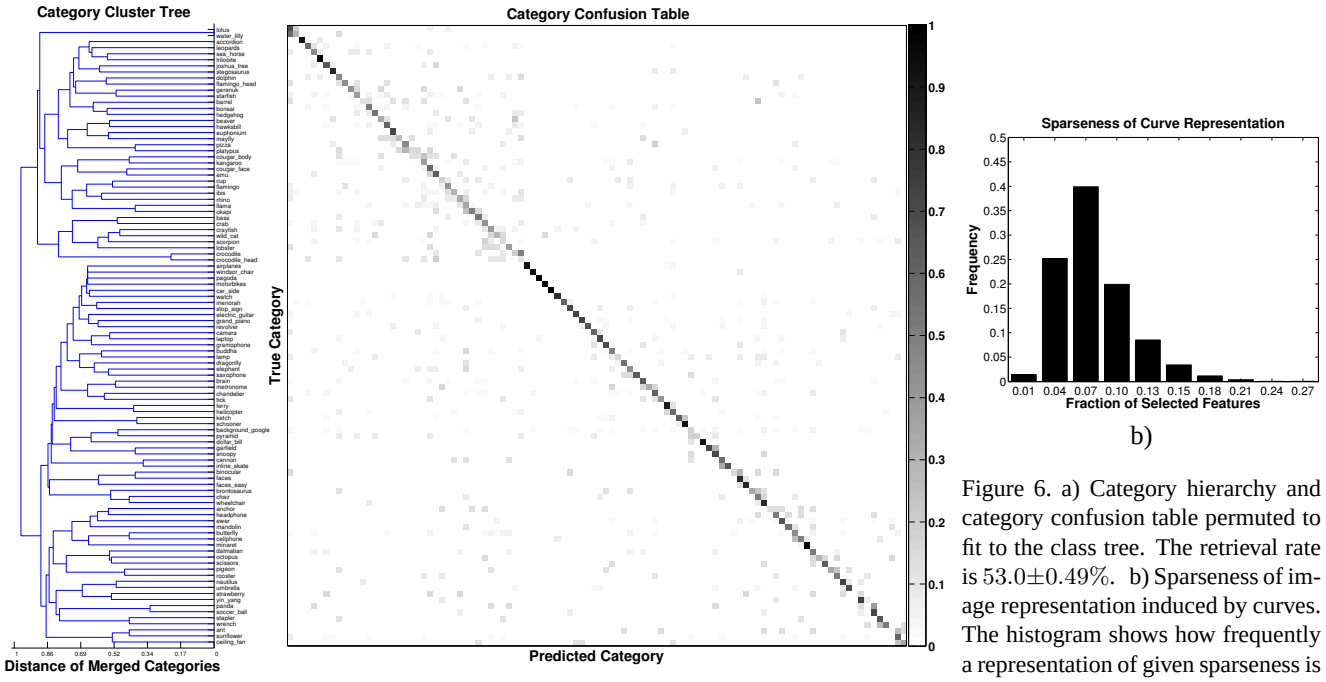
5. Discussion and Further Work

This contribution has combined a perceptual bottom-up grouping stage with a top-down agglomeration strategy to establish compositional hierarchies as intermediate scene representations. The latter grouping process, which is driven by object class models, is learned for a large number of categories, automatically. The architecture has been shown to be competitive compared to other current approaches on challenging test data for image categorization.

Among the many interesting future extensions of this model the most promising ones are to incorporate multiple scales (see also [20]) and to increase the representational power of the shape model to capture local warpings and perspective transformations.

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) b)

Figure 6. a) Category hierarchy and category confusion table permuted to fit to the class tree. The retrieval rate is $53.0 \pm 0.49\%$. b) Sparseness of image representation induced by curves. The histogram shows how frequently a representation of given sparseness is used.

[2] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, 2005. 2, 6

[3] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2), 1987. 1

[4] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *CVPR Workshop on Perceptual Organization in Comp. Vis.*, 2004. 2

[5] 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. 2

[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 GMBV*, 2004. 2, 7

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Computer Vision*, 61(1), 2005. 2

[8] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV*, 2004. 1, 2

[9] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1), 1973. 2

[10] S. Geman, D. F. Potter, and Z. Chi. *Composition Systems*. Technical report, Division of Applied Mathematics, Brown University, Providence, RI, 1998. 1

[11] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *NIPS*, 1998. 5

[12] A. D. Holub, M. Welling, and P. Perona. Combining generative models and fisher kernels for object class recognition. In *ICCV*, 2005. 2, 7

[13] B. Leibe and B. Schiele. Scale-invariant object categorization using a scale-adaptive mean-shift search. In *Pattern Recognition, DAGM*, 2004. 2

[14] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Learning hierarchical shape models from examples. In *EMMCVPR*, 2005. 2

[15] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, 1985. 2, 3

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

[17] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Computer Vision*, 60(1), 2004. 5

[18] B. Ommer and J. M. Buhmann. A compositionality architecture for perceptual feature grouping. In *EMMCVPR*, 2003. 1

[19] B. Ommer and J. M. Buhmann. Object categorization by compositional graphical models. In *EMMCVPR*, 2005. 1, 2, 4

[20] B. Ommer and J. M. Buhmann. Learning compositional categorization models. In *ECCV*, 2006. 1, 2, 4, 6, 7

[21] V. Roth and K. Tsuda. Pairwise coupling for machine recognition of hand-printed japanese characters. In *CVPR*, 2001. 5

[22] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, 2005. 2, 7

[23] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Segmentation, detection, and object recognition. In *ICCV*, 2003. 2

[24] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV*, 2000. 2