Learning Compositional Categorization Models

Björn Ommer and Joachim M. Buhmann*

Institute of Computational Science, ETH Zurich
8092 Zurich, Switzerland
{bjoern.ommer, jbuhmann}@inf.ethz.ch

Abstract. This contribution proposes a compositional approach to visual object categorization of scenes. Compositions are learned from the Caltech 101 database[1] and form intermediate abstractions of images that are semantically situated between low-level representations and the high-level categorization. Salient regions, which are described by localized feature histograms, are detected as image parts. Subsequently compositions are formed as bags of parts with a locality constraint. After performing a spatial binding of compositions by means of a shape model, coupled probabilistic kernel classifiers are applied thereupon to establish the final image categorization. In contrast to the discriminative training of the categorizer, intermediate compositions are learned in a generative manner yielding relevant part agglomerations, i.e. groupings which are frequently appearing in the dataset while simultaneously supporting the discrimination between sets of categories. Consequently, compositionality simplifies the learning of a complex categorization model for complete scenes by splitting it up into simpler, sharable compositions. The architecture is evaluated on the highly challenging Caltech 101 database which exhibits large intra-category variations. Our compositional approach shows competitive retrieval rates in the range of $53.6 \pm 0.88\%$ or, with a multi-scale feature set, rates of $57.8 \pm 0.79\%$.

1 Introduction

Automatically detecting and recognizing objects in images has been one of the major goals in computer vision for several decades. Recently, there has been significant interest in the subfield of object categorization, which aims at recognizing visual objects of some general class in scenes. The large intra-category variations which are observed in this setting turn learning and representing category models into a key challenge. Therefore, common characteristics of a category have to be captured while simultaneously offering invariance with respect to variabilities or absence of these features. Typically, this problem has been tackled by representing a scene with local descriptors and modeling their configuration in a more or less rigid way, e.g. [1, 2, 3, 4, 5, 6, 7, 8].

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

[1] www.vision.caltech.edu/feifeili/101_ObjectCategories

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part III, LNCS 3953, pp. 316–329, 2006.

Overview over the Compositional Approach to Categorization: This contribution proposes a system that learns category-dependent agglomerations of local features, i.e. localized histograms, and binds them together using a shape model to categorize scenes. It is evaluated on the challenging Caltech 101 image database and shows competitive performance compared to the current state of the art. Our approach has its foundation in the principle of *compositionality* [9]: It can be observed that in cognition in general and especially in human vision (see [10]) 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 (as [4]), the compositionality approach learns intermediate groupings of parts—possibly even forming a hierarchy of recursive compositions [11]. As a result compositions are establishing hidden layers between image features and scene categorization [7]. We do however restrict our system to a single layer of compositions as this already proves to be complex enough. 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. As a result the description length of the intermediate compositional representation is reduced. Therefore we propose methods for both, learning a set of compositions and establishing image categorization based on compositions detected in an image. The underlying training is conducted in a weakly supervised manner using only category labels for whole images.

Learning compositions is then guided by three modeling decisions: (i) Firstly, it has to be determined which parts to group to form potential candidate compositions. Here we follow the principles of *perceptual organization* [12]. (ii) Secondly, we aim at learning a fairly small set of compositions (currently 250) so that estimating category statistics on the training data becomes feasible. Therefore, the system cannot afford to learn compositions that are observed only rarely in the visual world. As an approximation on the training set we cluster potential composition candidates and estimate the priors of the different composition prototypes. (iii) Thirdly, each composition should be valuable for the task of discriminating sets of categories from another—not necessarily one category from all others. Compositions representing background that is present in many different categories or compositions that are only present in individual instances of a category are to be discarded. This discriminative relevance of a composition is estimated by the entropy of the category posterior distributions given the composition. Finally, the priors of composition prototypes and the entropy of the category posterior are combined in a single cost function. Based on this function relevant compositions are selected from the set of all prototypical compositions.

Crucial Modeling Decisions and Related Work: Methods in this field differ in the way they are approaching crucial modeling decisions: Firstly, various

local descriptors have been used. A classical way to capture image region information are *appearance patches* (e.g. [6, 4, 5, 3]). This method extracts image patches, converts them to grayscale, and subsamples them. As a result limited invariance with respect to minor variations in such patches is obtained. The resulting features are clustered to acquire a codebook of typically some thousand local patch representatives that are category specific. Another popular choice are *SIFT* features [13]. These are complex edge histogram features that have been proposed to distinguish different instances of an object class from another. Nevertheless they have also shown to perform reasonably well in the field of categorization. The high dimensionality and the specificity of these features with respect to individual visual realizations of an object require to cluster them into a large codebook representation. On the other end of the modeling spectrum are methods that compute histograms over complete images (cf. [14]). Such an approach offers utmost invariance with respect to changes of individual pixels at the cost of limited specificity. An approach which formulates a trade-off between these two classical extremes has been proposed in [7]. Here local edge and color histograms of subpatches are combined to obtain a low dimensional representation of an image patch. The lack of specificity is made up for by capturing relations between the local descriptors. We use these *localized histograms* in this contribution. Another approach that has shown to perform reasonably well is that of *geometric blur* [8]. This descriptor weights edge orientations around a feature point using a spatially varying kernel.

A second choice concerns the combination of all local features into a single model that captures the overall statistics of a scene. On the one hand individual local descriptors in a test image are to be matched against those from a learned model. On the other hand the co-occurence and spatial relation between individual features has to be taken into account. Here the simplest approach is to histogram over all local descriptors found in an image (e.g. [15]) 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 have a high chance to adapt to background features. At the other end of the modeling spectrum are *constellation models*. Originally, Fischler and Elschlager [1] have proposed a spring model for coupling local features. Inspired by the *Dynamic Link Architecture* for cognitive processes, Lades et al. [2] followed the same fundamental idea when proposing their face recognizer. Lately increasingly complex models for capturing part constellations have been proposed, e.g. [16, 4, 5, 17]. Finally Fergus et al. [4] estimate the joint Gaussian spatial, scale, appearance, and edge curve distributions of all detected patches. However the complexity of the joint model causes only small numbers of parts to be feasible. In contrast to this [6, 3] build a comparably large codebook of distinctive parts for a single category. Leibe and Schiele [3] estimate the mean of all shifts between the positions of codebook patches in training and test images. A probabilistic Hough voting strategy is then used to distinguish one category

from the background. [7] further refines this approach and groups 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 belief propagation. In this contribution we extend the shape model underlying [7] using probabilistic kernel classifiers. Finally, Berg et al. [8] describe and regularize the spatial distortion resulting from matching an image to a training sample using thin plate splines.

The approaches mentioned above are weekly supervised, that is they only need training images (showing objects and probably even background clutter) and the overall category label of an image. The restriction of user assistance is a desirable property for scaling methods up to large numbers of categories with huge training sets. In contrast to this a supervised approach to finding an object of a certain class in images is taken by Felzenszwalb and Huttenlocher in [18]. Given example images and the object configurations present in each image they explicitly model the appearance of a small number of parts separately and capture their spatial configuration with spring-like connections. Similarly, Heisele et al. [19] learn characteristic regions of faces and their spatial constellation. They create training faces from a textured 3-D head model by rendering and determine rectangular components by manually selecting specific points of a face (e.g. nose). Component sizes are estimated by reducing the error of a SVM.

Finally there are two broad categories of learning methods to choose from, generative and discriminative models. While the former aims at estimating the joint probability of category labels and features, the latter one calculates the category posterior directly from the data. Although discriminative approaches have, in principle, superior performance generative models have been very popular in the vision community, e.g. [3, 4, 20, 6, 7, 8]. One reason is that they naturally establish correspondence between model components and image features. Thereby the missing of features can be modeled intuitively. In contrast to this [15, 21] pursue a discriminative approach to object class recognition. To recognize faces in real-time Viola and Jones [21] use boosting to learn simple features in a fixed configuration that measure the intensity difference between small image regions. Holub et al. [17] propose a hybrid approach using Fisher kernels, thereby trying to get the best of both worlds.

The next section summarizes our compositional approach to categorization. Section 3 evaluates our architecture on the challenging Caltech 101 database and shows competitive performance compared to other current approaches. We conclude this presentation with a final discussion.

2 Categorization Using Compositional Models

The model can be best explained by considering recognition, see Figure 1(a). Given a novel image, salient image regions are detected in a first stage using a scale invariant Harris interest point detector [22]. Each region is then described by localized histograms [7]. In a next step a perceptual grouping of these local part descriptors is conducted to obtain a set of possible candidate compositions.

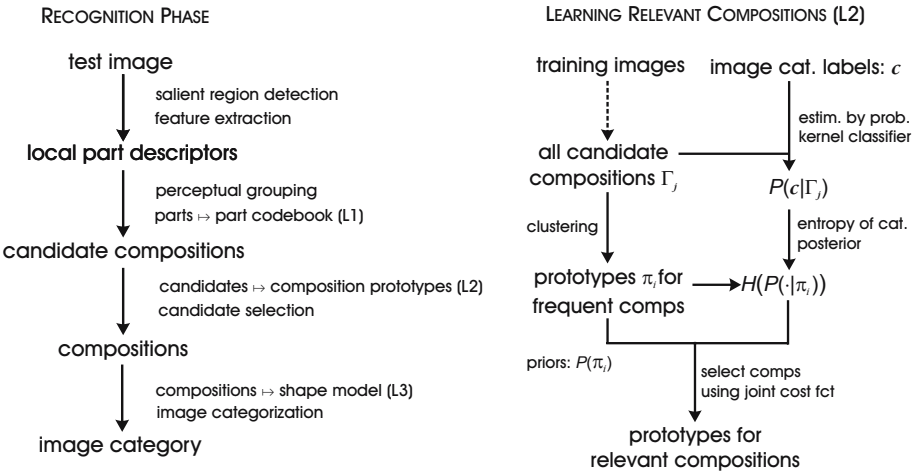


Fig. 1. (a) Recognition based on compositions. The three learning stages (L1–L3) which are involved are presented in Section 2.1, Section 2.3, and Section 2.4, respectively. (b) Learning relevant compositions (learning stage L2 from (a)), see text for details.

This grouping leads to a sparse image representation based on (probably overlapping) subregions, where each candidate represents an agglomeration of local parts. Consecutively, composition candidates have to be encoded. Therefore all detected local part descriptors are represented as probability distributions over a codebook which is obtained using histogram quantization in the learning stage. This codebook models locally typical configurations of the categories under consideration. A composition is then represented as a mixture distribution of all its part distributions, i.e. a *bag of parts*.

In a next stage relevant compositions have to be selected, discarding irrelevant candidates that represent background clutter. The set of relevant compositions has to be computed in the learning phase from the training data in a weakly supervised manner (see Figure 1(b)). As intermediate compositional representations should have limited description length, this learning obeys the following rationale: (i) Firstly, we aim at a set of compositions that occur frequently in the visual world of the categories under consideration. For that purpose all composition candidates found in all the training images are clustered and the prior assignment probabilities of candidates to these prototypes are estimated. (ii) Secondly, relevant compositions have to support the discrimination of sets of categories from another. Clutter that is present in many different categories or configurations that are only observed in few instances of a category are to be discarded to reduce the model complexity. In order to find a relevance measure the category posteriors of compositions are learned from the training data. The relevance of a composition for discriminating categories is then estimated by the entropy of its category posterior. By combining both the priors of the prototypes and the entropy, a single cost function is obtained that guides the selection of relevant compositions.

After discarding the irrelevant compositions from a new test image, the image category has to be inferred based on all the remaining relevant compositions. These compositions are spatially coupled by using a shape model similar to the one presented in [7].

2.1 Codebook Representation of Local Part Descriptors

In order to render the learning of compositions robust and feasible, low dimensional representations of local descriptors extracted from an image are sought. We choose a slight variation of *localized histograms* presented in [7]. At each interest point detected in an image a quadratic patch is extracted with a side length of 10 to 20 pixel, depending on the local scale estimate. Each patch is then divided up into four subpatches with locations fixed relative to the patch center. For each of these subwindows marginal histograms of edge orientation and edge strength are computed (allocating four bins to each of them). Moreover, 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 codebook (of currently $k = 100$ prototypes) is obtained. To robustify the representation each feature is not merely described by its nearest prototype but by a Gibbs distribution [23] 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,

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

$$Z(\mathbf{e}_i) := \sum_\nu \exp\left(-d_\nu(\mathbf{e}_i)\right). \tag{2}$$

2.2 Forming Candidate Compositions

Given all detected local part descriptors in an image, our categorization algorithm follows the principles of perceptual organization, i.e. *Gestalt laws*, to search for possible candidates for compositions. For the sake of simplicity, the current approach uses only the grouping principle of *proximity* although other agglomeration strategies could be invoked: From the set of all parts detected in an image, a subset (currently 30) is randomly selected. Each of these parts is then grouped with neighboring parts that are not farther away than 60-100 pixel (depending on the local scale estimate of the part mentioned in Section 2.1). Consequently compositions sparsely cover salient image regions.

The resulting candidate compositions are then represented as mixtures of the part distributions in (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 candidate composition is then represented by the vector valued random variable G_j which is a bag of parts, i.e. its value $\mathbf{g}_j$ is a distribution over the k-dimensional codebook from Section 2.1

$$\mathbf{g}_j \propto \sum_{i=1}^{m} \Big(P(F_i = 1|\mathbf{e}_i), \ldots, P(F_i = k|\mathbf{e}_i)\Big)^T. \tag{3}$$

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

2.3 Learning Relevant Compositions

Given all candidate compositions a selection has to be performed, retaining only the discriminative ones and discarding clutter. Learning such compositions is divided up into two stages, see Figure 1(b). First those groupings have to be retrieved which are representative for a large majority of objects observed among the considered categories. Thereby, the system avoids to memorize compositions that capture details of only specific instances of a category. Moreover, compositions should be shared among different categories. These concepts limit the description length of a compositional image representation and, thereby, reduce the risk of overfitting to specific object instances. In the learning phase the candidate compositions of all training images are therefore clustered (using k-means) into a comparably large set Π of prototypes $\pi_i \in \Pi$—currently 1000. Moreover, the prior assignment probabilities of candidates to clusters, $P(\pi_i)$, are computed.

In a second stage those prototypes have to be selected that help in distinguishing sets of categories from another. As the system combines multiple compositions found in one image, we do not have to solve the harder problem of finding groupings that are characteristic for a single category. In contrast to such an approach we pursue the robust setting of sharing compositions for multiple categories (cf. [24]). To begin with, the category posterior of compositions has to be estimated, i.e. the posterior of a categorization with label $c \in \mathcal{L}$ ($\mathcal{L}$ denotes the set of all category labels) given a composition Γ_j,

$$P_{\Gamma_j}(c) := P\left(C = c|\Gamma_j\right). \tag{4}$$

This distribution is learned by training probabilistic two-class kernel classifiers on all composition candidates found in the labeled training images. For the two-class classification we choose *nonlinear kernel discriminant analysis* (NKDA)[25] and perform a pairwise coupling to solve the multi-class problem (see [26, 25]). 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 (4) for the respective image category.

Subsequently the category posterior is used to calculate the relevance of a composition for discriminating categories. Groupings that are present in all categories are penalized by this idea, whereas combinations which are typical for only a few classes are fostered. The discriminative relevance measure is then modeled as the entropy of (4),

$$H(P_{\Gamma_j}) = -\sum_{c\in\mathcal{L}} P\left(C = c|\Gamma_j\right) \log P\left(C = c|\Gamma_j\right), \tag{5}$$

which should be minimized.

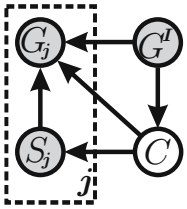


Fig. 2. Bayesian network that couples compositions G_j using their relative location S_j, a bag of features G^I, and image categorization C. Shaded nodes denote evidence variables. See text for details.

Finally a cost function can be formulated that measures the total relevance of a prototype π_i. It combines the prior assignment probabilities of clusters, $P(\pi_i)$, with the entropy (5),

$$\mathcal{S}(\pi_i) \propto -P(\pi_i) + \lambda H(P_{\pi_i}). \tag{6}$$

Both constituents of the cost function should be normalized to the same dynamic range, giving rise to an additional additive constant that can be discarded and to the parameter $\lambda > 0$. The latter trades the occurrence frequencies of compositions against their discriminative usefulness.

A set of 250 relevant composition prototypes that is shared by all categories can then be obtained by selecting the prototypes π_i with minimal cost $\mathcal{S}(\pi_i)$. An image is then represented by retaining only those composition candidates formed in Section 2.2 which are closer to one of the relevant prototypes than to any irrelevant one. However, at least the best 5 candidates are retained, thereby ensuring that images from the background category always yield a non-empty representation.

2.4 Binding Compositions Using a Shape Model

Subsequently, all relevant compositions which have been detected in an image are to be coupled with another using a shape model similar to that in [7]. First we have to estimate the object location $\mathbf{x}$. Therefore the positions $\mathbf{x}_j$ of all compositions $\mathbf{g}_j$ are considered. Moreover, we include 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.

$$\mathbf{x} = \sum_j \mathbf{x}_j \sum_{c \in \mathcal{L}} p(\mathbf{g}_j | c, \mathbf{g}^I)\ P(c|\mathbf{g}^I). \tag{7}$$

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 collapses to only the true category c and the distribution over categories is dropped. Following [7] the composition locations $\mathbf{x}_j$ are transformed into shifts, $\mathbf{s}_j := \mathbf{x} - \mathbf{x}_j$. Finally, the bag of features descriptor $\mathbf{g}^I$, the relative positions $\mathbf{s}_j$, and the image categorization c couple the compositions $\mathbf{g}_j$ with another as depicted in the graphical model in Figure 2. Using this model, the categorization posterior can be written as

$$P\big(c|\mathbf{g}^I, \{\mathbf{g}_j, \mathbf{s}_j\}_{j=1:n}\big) \propto \exp\Big[(1-n)\log P(c|\mathbf{g}^I) + \sum_j \log P\big(c|\mathbf{g}_j, \mathbf{s}_j, \mathbf{g}^I\big)\Big]. \tag{8}$$

As already mentioned previously, both distributions on the right hand side are estimated separately from the training data using NKDA. Consequently, novel images cannot only be assigned a category label, but also a confidence in this categorization.

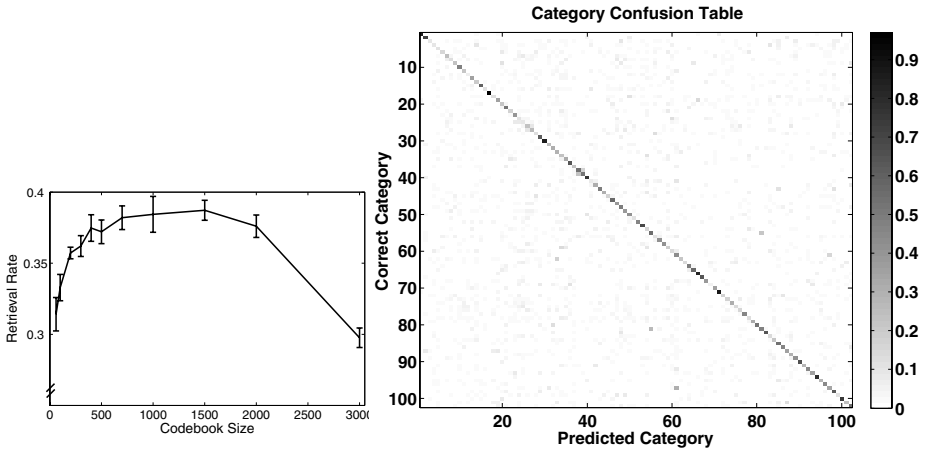
3 Experiments

We evaluate our approach on the challenging Caltech 101 database consisting of 101 object categories and a background category with varying numbers of samples (between about 30 and 800). The dataset contains the full spectrum of images ranging from photos with clutter to line drawings. However, there are only limited variations in pose. Subsequently, the retrieval rate is to be computed. As categories are having different sample sizes, we average over the retrieval rates that are measured for each category individually, thereby avoiding a bias towards classes with more images. Berg et al. [8] have calculated a reasonable baseline performance of 16% using texton histograms. Moreover their approach which is based on shape correspondence achieved a classification rate of 48%. Using a constellation model Fei-Fei et al. performed at about 16%. Finally, Holub et al. [17] extend the generative constellation model approaches with a discriminative method and a fusion of several interest point detectors to achieve 40.1%.

Baseline Performance Without Compositions: Object categorization is based on an intermediate compositional image representation in our approach. The following experiments estimate a baseline performance of the system without this hidden representational layer. Therefore we neglect all compositions and consider only the bag of features representation $\mathbf{g}^I$ of the whole image, introduced in Section 2.4.

The basic evaluation scenario is as follows: For each class up to 50 training images are randomly selected (the coupled classifiers are weighted to compensate for the unequal priors) and the remainder is taken as test set (minimally 10 images in a class and over 4000 in total). To estimate the retrieval rate and its error 5-fold cross-validation is performed, i.e. the same algorithm is applied to 5 different training and test set compositions. Figure 3(b) shows the resulting category confusion table for the case of a feature bag which consists of 100 prototypes. This simple model achieves a retrieval rate of $33.3 \pm 0.9\%$.

To evaluate its dependence on the size of the codebook from Section 2.1 the simple bag of features approach is now evaluated with different numbers of clusters. Figure 3(a) shows the retrieval rates under varying model complexity. In the case of 1000 prototypes this model yields a retrieval rate of $\mathbf{38.4 \pm 1.3}\%$. Compare this with the maximal performance of 29% that [17] obtain with their discriminative method on the basis of a single interest point detector (like our simple model presented in this section) and the 40.1% of the combination of all three detectors. As the localized histograms are fairly low-dimensional descriptors, comparably small codebooks do already yield considerable retrieval rates. This is advantageous for modeling compositions robustly which are obviously



(a) (b)

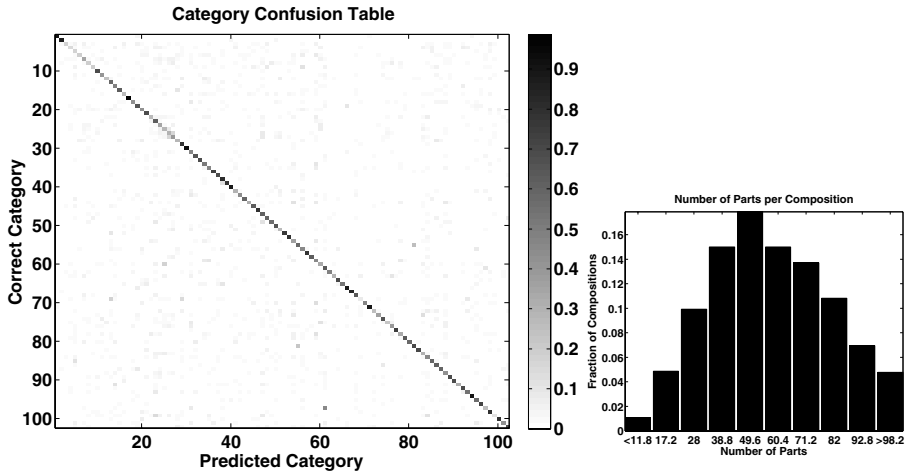
Fig. 3. (a) Retrieval rates for a bag of features approach with codebooks of different sizes. (b) Category confusion table for a bag of features approach with 100 prototypes. The retrieval rate is $33.3 \pm 0.9\%$.

consisting of fewer parts than a complete image and justifies our choice of a 100 prototype representation in the full compositional architecture.

Categorization Performance of the Compositional Model: Subsequently, the full compositional model is learned to categorize images. Evaluation under 5-fold cross-validation yields a retrieval rate of $\mathbf{53.6 \pm 0.88}\%$ which compares favorably with the 48% of Berg et al. [8]. Additionally, we note that the overall retrieval rate per image without averaging over categories is $67.3 \pm 2.1\%$. Figure 4(a) depicts the respective category confusion table. When comparing this plot with the one for the simple bag of features approach from above it is evident that the number of incorrectly classified images has significantly decreased. The categories with lowest performance are "octopus", "wildcat", and "ant", the best ones are "car", "dollar bill", and "accordion". Amongst the off-diagonal elements the confusions "water-lilly" vs. "lotus", "ketch" vs. "schooner", and "lobster" vs. "crayfish" are the most prominent ones. All of these confusions are between pairs that are either synonymous or at least semantically very close. To conclude, the observable gain in resolving ambiguities between classes emphasizes the advantage of an intermediate compositional image representation in contrast to a direct categorization.

Evaluating Compositions: The following evaluates the relevant compositions that have been learned. Firstly, Figure 4(b) plots the number of parts that are typically grouped to form a composition. On average there are approximately 57 parts coupled together. This is a significant increase compared to the tuple groupings formed in [7].

The next experiment intends to visualize the learned compositions. Since these are agglomerations of localized histograms that cannot be displayed



(a) (b)

Fig. 4. (a) Category confusion table of the compositional model. The retrieval rate is $53.6 \pm 0.88\%$. (b) Distribution of the number of parts assigned to each composition.

directly an indirect method has to be pursued. We therefore plot image regions from the test images that have been detected to contain a specific composition. A displayed region is then simply the rectangular hull of all parts that have been agglomerated to a composition. As space in this paper does not permit to present the full set, Figure 5 visualizes a subset of all learned compositions by showing 3 candidate regions for each. The zones are therefore scaled to equal sizes. Observe that compositions are reflecting quite different, abstract concepts: There are those that nicely correspond to salient structures in a single category Figure 5(a)-(c). In the latter case there are however also representatives from another category (motorbike) that show a visually similar pattern. Figure 5(d) and (e) exhibit more extended feature sharing. In (d) the triangular structures of airplane rudders and schooners are captured, while (e) combines sails of different boat categories and butterfly wings. The composition in (f) grasps roundish, metallic structures and (g) elongated, repetitive patterns of windsor chairs and menorahs. The next two compositions are an example of textures. The latter however also seems to model the presence of sharp edges, while (j) captures characteristic contours of pianos and staplers. An example for drawings is given in (k), while (l) seems to model the abstract concept of feet of chairs, pianos, and insects. In conclusion various kinds of low level properties are combined to represent fairly abstract concepts that help to discriminate between categories.

Localizing Object Constituents: Subsequently the relevance of individual compositions for the task of categorizing an image is to be evaluated. Therefore, the relevant object constituents are to be identified and localized. We measure how the categorization performance varies when a single composition is removed. Relevance is then proportional to the decrease in categorization probability of

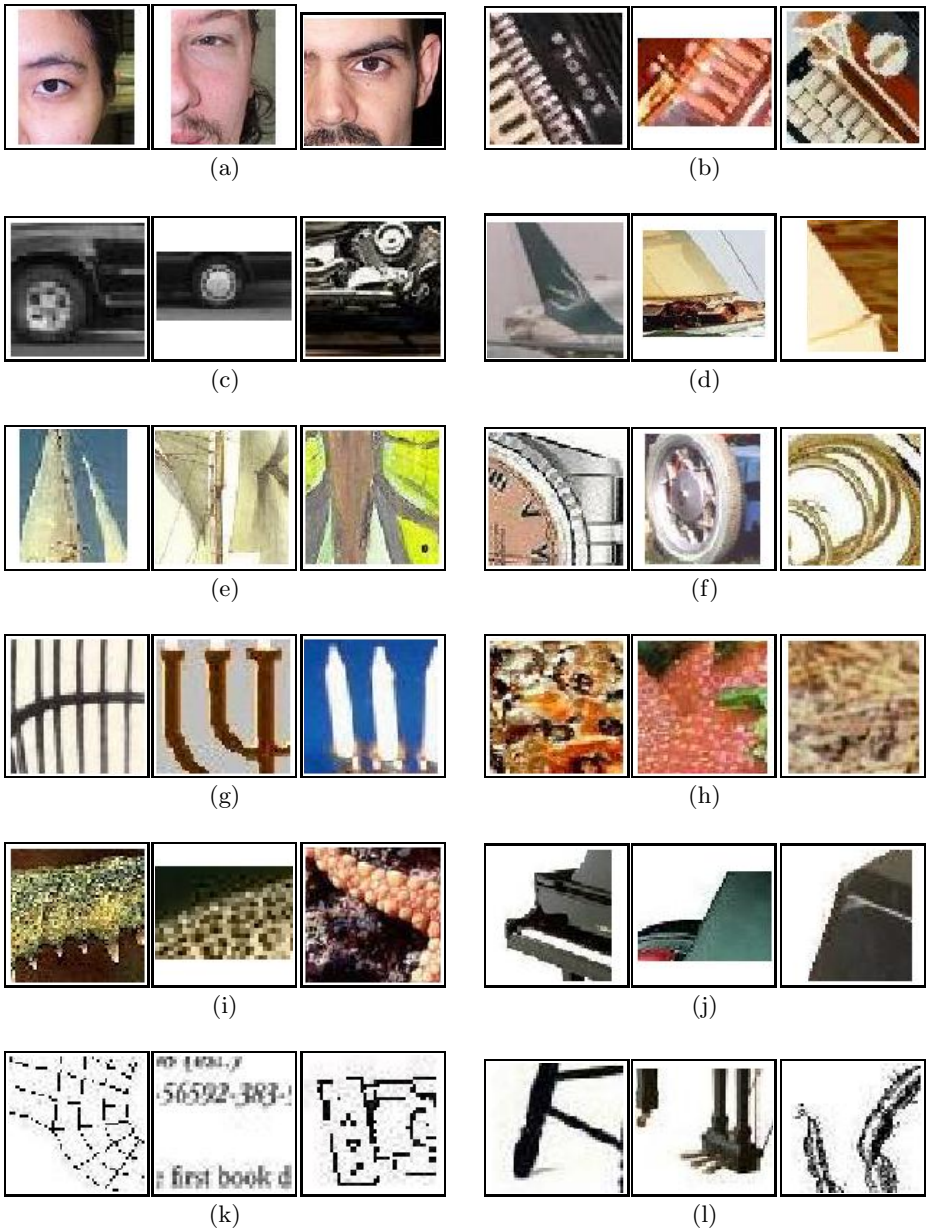


Fig. 5. Visualization of compositions: The pictures show the rectangular hulls of test image regions associated with different compositions. Different, abstract concepts captured by compositions: (a) Parts of faces, (b) accordions, and (c) cars, motorbikes. Feature sharing for complex structures of airplanes and schooners in (d), and of boat sails and butterfly wings in (e). (f) roundish structures. (g) elongated patterns of chairs and menorah. (h), (i) texture with and without a sharp edge, respectively. (j) contours. (k) drawings. (l) feet of chairs, pianos, and insects.

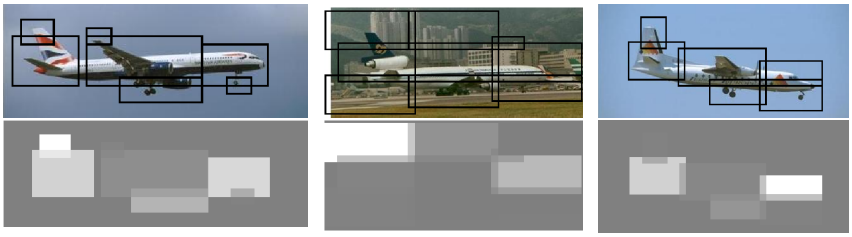


Fig. 6. Relevance of detected compositions (black boxes). Brighter patches than background indicate high relevance, darker ones indicate compositions are not useful.

the true category. Figure 6 shows examples for the airplane category. It is obvious that especially the noses and rudders are particularly relevant.

4 Discussion and Further Work

In this contribution we have successfully developed an architecture for categorizing scenes based on compositional models that are automatically learned. This intermediate, semantic abstraction layer has been shown to yield competitive performance compared to other current approaches on challenging test data.

Currently we are extending the system to incorporate multiple scales and hierarchies of compositions. The multi-scale extension alone, which incorporates additional features extracted on half the original scale, has boosted the retrieval rate to $\mathbf{57.8 \pm 0.79}\%$. Therefore we consider these system design decisions as a promising direction to further increase the robustness of our compositional model.

References

1. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Trans. Comput. **22** (1973)
2. Lades, M., Vorbrüggen, J.C., Buhmann, J.M., Lange, J., von der Malsburg, C., Würtz, R.P., Konen, W.: Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. Comput. **42** (1993)
3. Leibe, B., Schiele, B.: Scale-invariant object categorization using a scale-adaptive mean-shift search. In: Pattern Recognition, DAGM. (2004)
4. Fergus, R., Perona, P., Zisserman, A.: A visual category filter for google images. In: ECCV. (2004)
5. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR Workshop GMBV. (2004)
6. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. Pattern Anal. Machine Intell. **26** (2004)
7. Ommer, B., Buhmann, J.M.: Object categorization by compositional graphical models. In: EMMCVPR. (2005)

8. Berg, A.C., Berg, T.L., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: CVPR. (2005)
9. Geman, S., Potter, D.F., Chi, Z.: *Composition Systems*. Technical report, Division of Applied Mathematics, Brown University, Providence, RI (1998)
10. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review **94** (1987)
11. Ommer, B., Buhmann, J.M.: A compositionality architecture for perceptual feature grouping. In: EMMCVPR. (2003)
12. Lowe, D.G.: Perceptual Organization and Visual Recognition. Kluwer Academic Publishers, Norwell, MA (1985)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Computer Vision **60** (2004)
14. Veltkamp, R.C., Tanase, M.: Content-based image and video retrieval. In: A Survey of Content-Based Image Retrieval Systems. Kluwer (2002)
15. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV Workshop on Stat. Learn. in Comp. Vis. (2004)
16. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: ECCV. (2000)
17. Holub, A.D., Welling, M., Perona, P.: Combining generative models and fisher kernels for object class recognition. In: ICCV. (2005)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Computer Vision **61** (2005)
19. Heisele, B., Serre, T., Pontil, M., Vetter, T., Poggio, T.: Categorization by learning and combining object parts. In: NIPS. (2001)
20. Borenstein, E., Sharon, E., Ullman, S.: Combining top-down and bottom-up segmentation. In: CVPR Workshop on Perceptual Organization in Comp. Vis. (2004)
21. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. (2001)
22. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int. J. Computer Vision **60** (2004)
23. Winkler, G.: Image Analysis, Random Fields and Markov Chain Monte Carlo Methods—A Mathematical Introduction. 2nd edn. Springer (2003)
24. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. In: CVPR. (2004)
25. Roth, V., Tsuda, K.: Pairwise coupling for machine recognition of hand-printed japanese characters. In: CVPR. (2001)
26. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: NIPS. (1998)