Object Categorization by Compositional Graphical Models

Björn Ommer and Joachim M. Buhmann

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

Abstract. This contribution proposes a compositionality architecture for visual object categorization, i.e., learning and recognizing multiple visual object classes in unsegmented, cluttered real-world scenes. We propose a sparse image representation based on localized feature histograms of salient regions. Category specific information is then aggregated by using relations from perceptual organization to form compositions of these descriptors. The underlying concept of image region aggregation to condense semantic information advocates for a statistical representation founded on graphical models. On the basis of this structure, objects and their constituent parts are localized.

To complement the learned dependencies between compositions and categories, a global shape model of all compositions that form an object is trained. During inference, belief propagation reconciles bottom-up feature-driven categorization with top-down category models. The system achieves a competitive recognition performance on the standard CalTech database[1].

1 Introduction

The automatic detection and recognition of objects in images has been among the prime objectives of computer vision for several decades. There are several levels of semantic granularity on which classification of objects can be conducted, e.g. recognizing different appearances of the same object as opposed to different representations of the same category of objects. Object categorization aims at recognizing visual objects of some general class in scenes and labeling the images accordingly. Therefore, a given set of training samples is used to learn category-specific properties which are then represented in a common model. Based on this model, previously unknown instances of the learned categories are then to be recognized in new visual scenes.

The large variations among appearances and instantiations of the same visual object category turn learning and representing models for various categories into a key challenge. Therefore, common characteristics of objects in a category have to be captured while at the same time a great flexibility with respect to variability or absence of such features has to be offered. Consequently, we propose a system that infers scene labels based on learned category-dependent agglomerations

[1] http://www.vision.caltech.edu/html-files/archive.html

A. Rangarajan et al. (Eds.): EMMCVPR 2005, LNCS 3757, pp. 235–250, 2005.

of features which are robust with respect to intra-class variations and are thus reliably detectable. This approach to categorization has its origin in the principle of *compositionality* [9]. It is not only observed in human vision (see [2]) but also in cognition in general that complex entities are perceived as compositions of simpler, more unspecific, and widely usable parts. Objects are then defined by their components and the relations between those components. Therefore, the relationships between parts compensate for the limited information provided by each individual part. Moreover, a comparably small number of these lower-level constituents suffices to enable perception of various objects in diverse scenes. We like to emphasize that we see key contribution of our approach in the probabilistic coupling of different components which have been discussed in the literature. The homogeneity of this compositionality architecture and the common probabilistic framework for information processing yields the demonstrated robustness which we consider indispensible for object recognition.

Our architecture detects features of salient image regions and represents them using a small codebook that has been learned on the training set. Consecutively relations between the regions are acquired and used to establish compositions of these image parts. Therefore the part representations, the compositions, as well as the overall image categorization are all combined in a single graphical model, a Bayesian network [19]. Thus the probabilities of compositions and overall image categorization can be inferred from the observed evidence using model parameters that are learned from the training data. This learning is based on category labels of complete training images as the only information, e.g., images labeled as belonging to the car category contain a car somewhere in the image without marking the car region specifically. Therefore, the intermediate representation, that is the set of relevant compositions, is learned with no user supervision. Furthermore, the spatial configuration of all object components of a given category are learned and captured in a global, probabilistic shape model. Categorization hypotheses are then refined based on this information and objects can be localized in an image. The architecture has been trained and evaluated on the standard CalTech database (cars, faces, motorbikes, and airplanes). An additional background category of natural sceneries from the COREL image database has been incorporated for learning the hidden compositions. In summary, the architecture combines bottom-up, feature-driven recognition with top-down, category model driven hypotheses in a single Bayesian network and performs them simultaneously during belief propagation to infer image categorization.

This contribution outlines our approach in Section 3. An evaluation of the categorization model follows in Section 4 before we conclude this presentation with a final discussion. Related work is summarized in the next section.

2 Related Work

Object categorization has previously been mainly based on local appearance patches (e.g. [1, 12, 13, 6, 7]). That is, image regions are extracted, converted to grayscale and subsampled to obtain limited invariance with respect to minor variations in such patches. The resulting features are clustered to acquire a codebook of typically some thousand local patch representatives that are category specific.

To incorporate additional information beyond extracted local patches and features, a sequence of recognition models have been proposed in the class of *constellation models*. Originally, Fischler and Elschlager [8] described a spring model with local features to characterize objects. In the same spirit, Lades et al [11] proposed a face recognizer which has been inspired by the *Dynamic Link Architecture* for cognitive processes with a neurobiologically plausible dynamics. Similar to this model, Weber et al. [21] have introduced a joint model for all features present in an object. Fergus et al. [7] extend this approach and estimate the joint spatial, scale, appearance, and edge curve distributions of all detected patches which they normalize with respect to scale. However, due to the complexity of the joint models used by these approaches, only a small number of parts can be used. In contrast to this, Agarwal et al. [1] build a comparably large codebook of distinctive parts and learn spatial configurations of part tuples which belong to objects of a single category. However, since the individual appearance patches are highly specific the joint model is restricted in terms of its generalization ability and requires large training sets. To overcome these difficulties, Leibe et. al [12, 13] estimate the mean of all shifts between positions of codebook patches in training and test images. Using a probabilistic Hough voting strategy one object category is distinguished from a background category. Moreover, the spatial information is used to segment images and to take account of multiple objects in a scene. We further refine this approach and reconcile conflicting categorization hypotheses proposed by compositions of parts and those proposed by spatial models. Therefore, compositions and spatial models are coupled in a Bayesian network and beliefs are propagated between them in an alternating manner.

The approach in [12] to incorporate top-down information into segmentation has been proposed previously by Borenstein and Ullman in [5] where learned object fragments are aligned based on bottom-up coherence criteria. This improves especially segmentation boundaries, but they do not use this process for recognition. In [4] an extension is presented that uses the sum-product algorithm [19, 10] to solve local contradictions and to obtain a globally optimal segmentation.

The approach of forming an object representation based on compositions of unspecific, and reliably detectable features has strong support by visual cognition [2]. Geman et al. [9, 3] present this concept in the context of stochastic grammars and use it for recognizing handwritings. However, compositionality in the scenario of object categorization is a novel technique. In [18] we have proposed an architecture for forming compositional grouping hierarchies based on the psychological principles of perceptual organization [14]. Therefore different types of perceptual relations between parts are established to build salient compositions of reduced description length and increased robustness.

3 Categorization Based on Interacting Compositions

Our architecture which represents compositionality in a graphical model for performing object categorization has several stages. The following sketches the

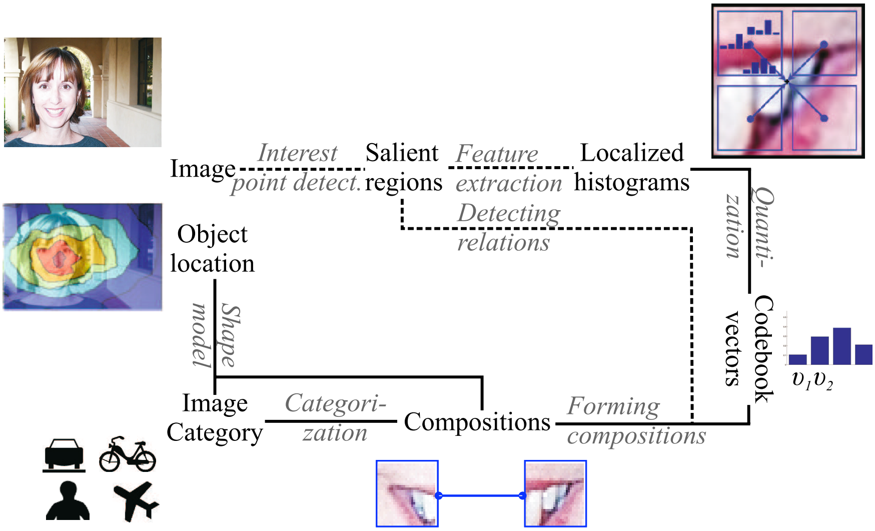


Fig. 1. Outline of the processing pipeline and information flow. Belief is being propagated along the solid lines, whereas the dotted ones indicate the capturing of evidence.

recognition process and states how learning is involved (see Figure 1): At first a scale invariant Harris interest point detector is used [16] to detect salient image regions. Every region is then captured by several feature histograms, each being localized with respect to the interest point. The features are represented using a probability distribution over a codebook that has been obtained by a histogram quantization in the learning stage. This codebook captures locally typical feature configurations of the categories under consideration. In a next step relations are detected between the regions and are being used to infer compositions. This inference is based on previously learned category specific grouping probabilities. Thereafter, all these compositions are taken into account to yield the overall categorization probability for the image. In addition to a maximum a-posteriori estimate for the category, this also yields a confidence in this classification. Finally, a learned model of object shapes is used to infer the object position in the image based on all compositions and the categorization hypothesis. This spatial probability distribution is in turn used to refine compositions and overall categorization. Thereby, both bottom-up image classification which depends on features and top-down recognition which depends on category models are corroborating another by running in an interleaved manner during belief propagation.

The following section gives a detailed account of the different stages and describes the learning of models which are underlying the inference procedure used for recognition with the network illustrated in Figure 2. Due to the independence properties represented by this Bayesian network, the categorization probabilities factorize and their computation is split into separate parts [19]. This factorization is significant to dividing up the procedure into the different stages and making inference feasible. In [19] a message-passing algorithm is introduced that propa-

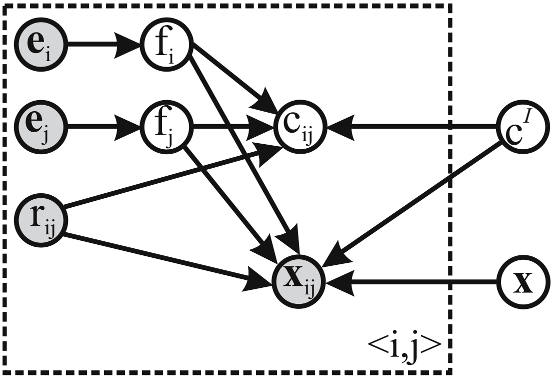


Fig. 2. Illustration of the Bayesian network. The evidence nodes (shaded variables) are $E = \{\mathbf{e}_i\}_i \cup \{r_{ij}, \mathbf{x}_{ij}\}_{<i,j>}$. Where $< i, j >$ denotes pairs of parts that are agglomerated into compositions—the dotted structure is therefore replicated for all these tuples, see text for details.

gates evidence through polytrees to estimate the *belief* of unobserved variables, that is their posterior given the evidence. For some random variable Y it is

$$BEL(y) := P(Y = y|E) \ , \tag{1}$$

where E denotes the observed evidence. Moreover, it has been widely advocated that this so called *sum-product algorithm* [10] yields good approximations even for Bayesian networks with loops (cf. [17]).

3.1 Localized Feature Histograms for Compositionality

As outlined above, representations of object categories have to deal with large intra-class variations. However, the local appearance patches that have been widely used in the field of object categorization are basically subsampled image patches. The clustering that is then performed to obtain patch representatives is usually based on normalized grayscale correlation whereby invariance to illumination changes of patches as a whole is obtained. However, since the resulting invariances are just established by a global subsampling and intensity normalization, translations or local alterations still have an overproportional influence on the complete feature. Moreover, due to the low-pass filtering, only information on the strongest edges is preserved while the remaining patch content is blurred.

To overcome these problems we follow the concept of compositionality [9] where models for complex objects are decomposed into more unspecific, robustly detectable and thus widely usable components. This strategy results in fairly short representations for components and facilitates robust estimation of the statistics that model the grouping of parts. This section starts by outlining the part representation, while later sections continue to present relations and compositions.

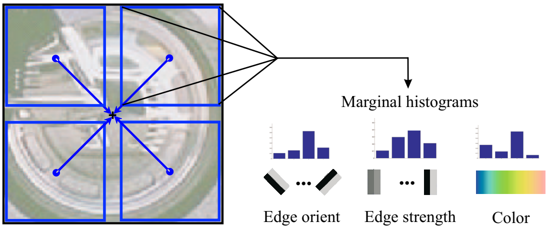


Fig. 3. Sketch of localized feature histograms

A crucial problem of forming a scene representation is the trade-off between its invariance properties, e.g. to varying influences of the imaging process, and its specificity for a certain task, e.g. distinguishing object categories. As delineated above, current approaches to categorization base their class recognition mainly on highly distinctive local appearance patches (e.g. [1, 12, 6]) and incorporate only limited invariance with respect to alteration of patch contents. An alternative approach at the other end of the modeling spectrum is that of using histograms over complete images (cf. [20]). Thereby, utmost invariances with respect to changes of individual pixels can be obtained. In conclusion, the former approach facilitates almost perfect localization while the latter one offers maximal invariance with respect to local distortions. We therefore aim at a representation whose invariance properties are transparently adjusted between these two classical extremes and add the specificity lost by invariance through the relations that are used for forming the compositions.

To process an image, we start by applying the interest point detector to obtain some 10^2 to 10^3 interest points together with rough local scale estimates. Although our implementation incorporates multiple scales, our current approach is based on a single estimated scale selected by the interest point detector. Therefore we extract quadratic image patches (with a side length of 10 to 20 pixel depending on scale) and subdivide them into a number of subpatches with fixed location relative to the patch center (see Figure 3). In each of these subwindows three types of marginal histograms are computed (four bins allocated for each), measuring edge strengths, edge orientations, and color. The statistics of each feature channel are estimated independently from another to make the estimates robust by having enough data support. In the following, the vector of measurements $\mathbf{e}_i$ denotes the combination of the features extracted in all the subpatches at interest point i. These vectors serve as evidence in our Bayesian network. The trade-off between invariance and localization is represented by the number of subpatches—in the current implementation a patch is divided up into four of these subwindows.

The proposed representation differs from the SIFT features [15] not only in that color is used. Whereas SIFT features aim at distinguishing different instances of the same object from another, we seek a representation that is invariant to the specificities of individual object instances and environment configurations. To obtain a small codebook of atomic representatives for compositionality, we reduce the complexity of the very features whereas the other approach would have to perform this indirectly by clustering in a high-dimensional space with few prototypes.

3.2 Codebook Representation of Atomic Compositional Parts

To facilitate a robust estimation of statistics in subsequent stages of the architecture, a small codebook for features is generated during learning. This set forms a representation of atomic compositional parts. The codebook is generated by clustering the features detected in training images of all the different categories with k-means—resulting in a set of 300 centroids in our current implementation. It should be emphasized that this representation is shared by all the different categories. During recognition the squared euclidean distance $d_\nu(\mathbf{e}_i)$ of a measured feature $\mathbf{e}_i$ to all the centroids $\mathbf{a}_\nu$ from the codebook is computed. The objective is to represent measurements not merely by their nearest prototype but by a distribution over the codebook, thereby leading to increased robustness. Now, for each measurement $\mathbf{e}_i$, a new random variable F_i is introduced that takes on cluster indices ν as its values. Each of these variables is coupled with the corresponding measurement using the same Gibbs distribution [22]

$$P(F_i = \nu | \mathbf{e}_i) := Z(\mathbf{e}_i)^{-1} \exp\left(-d_\nu(\mathbf{e}_i)\right) \quad , \tag{2}$$

$$Z(\mathbf{e}_i) := \sum_\nu \exp\left(-d_\nu(\mathbf{e}_i)\right) \quad . \tag{3}$$

Subsequently, $P(F_i = \nu)$ is abbreviated using its realization $f_i = \nu$ and simply writing $P(f_i)$ which models the first stage of the Bayesian network in Figure 2.

3.3 Forming Compositions

The part representations have to be augmented by additional evidence. Therefore, relations between image regions are taken into account. From the various principles of perceptual organization, investigated in [18] for grouping processes, we apply *good continuation*. This concept basically groups those entities together that form a common smooth contour. Hence we consider pairs of patches which lie on a common edge curve and measure their distance. To facilitate a later robust statistical estimation, this distance is discretized into three ranges (i.e. close/medium/far) which depend on a histogram over all these distances measured in the training data. The edge curves are obtained by performing Canny edge detection twice, once with a scale parameter that is the mean of the lower half of all scales detected at the interest points, and once with scale being equal to the mean of the upper half. The edge images are then added up. Now consider two patches at interest points i and j. If they are observed to lie on the same contour and have a discretized gap of r_{ij} they establish the relation $R_{ij} = r_{ij}$. The two parts are then forming a composition which we denote by $< i, j >$, i.e.,

$$< i, j > \Leftrightarrow \text{part } i \text{ \& part } j \text{ form a composition} \quad . \tag{4}$$

Since the relations between image regions are observed, all the random variables R_{ij} enter as evidence into the Bayesian network in Figure 2. It should be emphasized that this is a sparse set of nodes—iff both patches lie on a common contour, such a random variable is introduced. In conclusion, a grouping based on such relations incorporates additional edge information that takes compositions beyond a mere proximity or co-occurrence grouping.

Based on the detected relations the following modelling heuristic describes how compositions of parts are formed. Let the random variable C_{ij} represent a composition of the two image regions i, j. Each such composition is of a certain category. That is, it has a certain state $c_{ij} \in \mathcal{C}_C$, where this state space of compositions is a superset of the set $\mathcal{C}_I = \{\text{face}, \text{airplane}, \dots\}$ of all categories for images, i.e. $\mathcal{C}_I \subset \mathcal{C}_C$. Consider the illustrating example of an image that is recognized to contain a motorbike. Then compositions representing subparts such as tires might be added to the set of allowed image categories. In our current implementation both sets differ by an additional category for *background* that we have incorporated for compositions, i.e. $\mathcal{C}_C = \mathcal{C}_I \cup \{\text{background}\}$.

The distribution of a composition of the two parts i, j depends only on the representations of the involved parts, their relation, as well as on the categorization of the image, denoted by C^I, where $c^I \in \mathcal{C}_I$. Thereby, the invariances represented in the Bayesian network from Figure 2 are reflected,

$$P(C_{ij} = c_{ij} | F_i = f_i, F_j = f_j, R_{ij} = r_{ij}, C^I = c^I) \ . \tag{5}$$

All C_{ij} are assumed to be identically distributed and this distribution is split into

$$P(c_{ij} | f_i, f_j, r_{ij}, c^I) \propto P(c^I | c_{ij}) P(c_{ij} | f_i, f_j, r_{ij}) \tag{6}$$

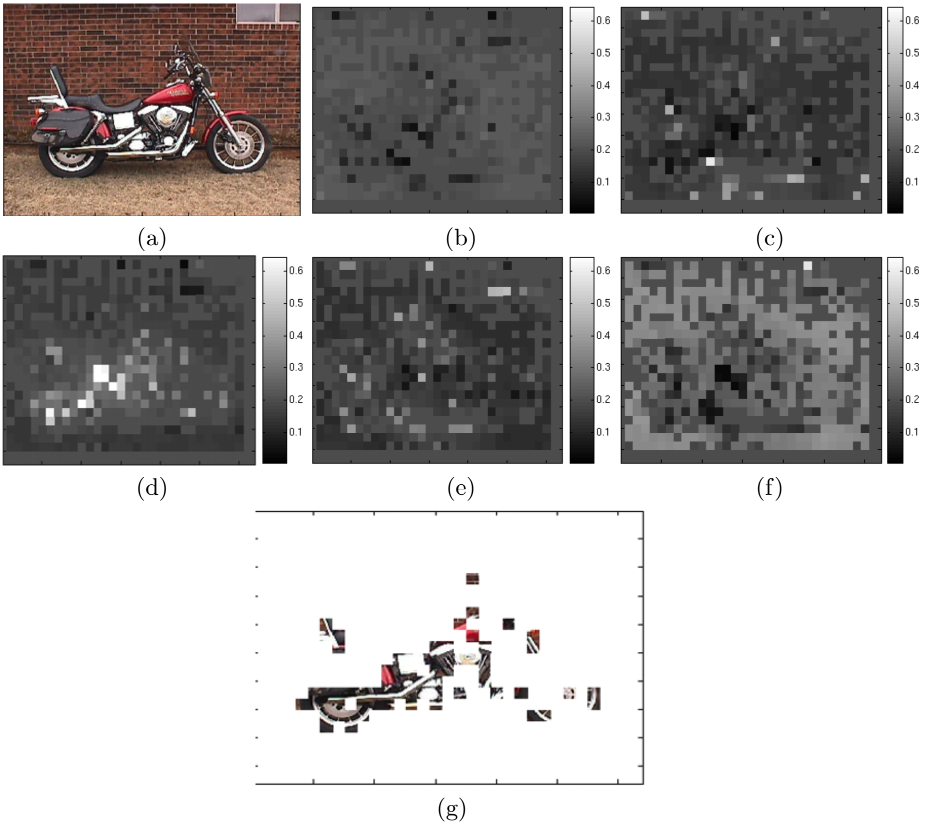
using Bayes formula and dropping a normalization constant. The first factor models category confusion probabilities which are assumed to be independent from features and relations $(P(c^I | c_{ij}) = P(c^I | c_{ij}, f_i, f_j, r_{ij}))$ when compositions c_{ij} are given. With no further assumptions on categories, we have made the following choice,

$$P(c^I | c_{ij}) = \begin{cases} |\mathcal{C}_I|^{-1}, & \text{if } c_{ij} = \text{background} \\ \eta, & \text{if } c^I = c_{ij} \\ 1 - \eta, & \text{otherwise} \ . \end{cases} \tag{7}$$

In our current implementation we simply set $\eta = 1$. The second distribution in Eq. (6) is the categorization probability of compositions: The underlying nonparametric model is obtained in the learning stage by processing all the training images as follows: for each detected grouping, the category label of the whole image is taken as c_{ij} and the distribution is estimated from the empirical histogram of all observed compositions. Figure 4 and 9 visualize the category beliefs of compositions for the different classes.

3.4 Modeling Object Shape

In the following, a model of the spatial configuration of object components is presented. This model is used to refine the image categorization by propagating



(a) (b) (c)

(d) (e) (f)

(g)

Fig. 4. Categorization belief of compositions. For each composition c_{ij}, (b) displays $P(C_{ij} = \text{car}|E)$ at the position of patches i and j. Image regions that are not used to form compositions are displayed with the uniform distribution over all categories. Regions that are involved in multiple compositions show the average belief of these compositions. (c) Displays the posterior for class *face*, (d) for *motorbike*, (e) for *airplane*, and (f) for *background*. Where the last category facilitates a figure-ground segregation of compositions. (g) Shows the regions selected by the algorithm to categorize the image as *motorbike*.

information on the estimated object location. The shape of an object of a given category is modeled by the displacement $\mathbf{s}_{ij}$ of all of its components from its center $\mathbf{x}$. Letting $\mathbf{x}_{ij}$ denote the location of a composition (the midpoint between its components) detected in the training data, its shift is computed as

$$\mathbf{s}_{ij} = \mathbf{x} - \mathbf{x}_{ij} . \tag{8}$$

During learning, the object centers are computed by

$$\mathbf{x} = \sum_{I \in \text{train data}} \sum_{<i,j> \in I} \mathbf{x}_{ij} \cdot P(C_{ij} = c^I | f_i, f_j, r_{ij}) \ . \tag{9}$$

To predict the location of the object center, a Parzen window density estimation is performed. The probability of a shift, given the features of a composition and the object category, is represented by the following non-parametric model

$$p(S = \mathbf{s}|f_i, f_j, r_{ij}, c^I) = \frac{1}{N}\sum_{l=1}^{N} \frac{K_{\sigma_N}\Big(\mathbf{s} - \mathbf{s}_{ij}^{(l)}\Big)}{\sigma_N}. \qquad (10)$$

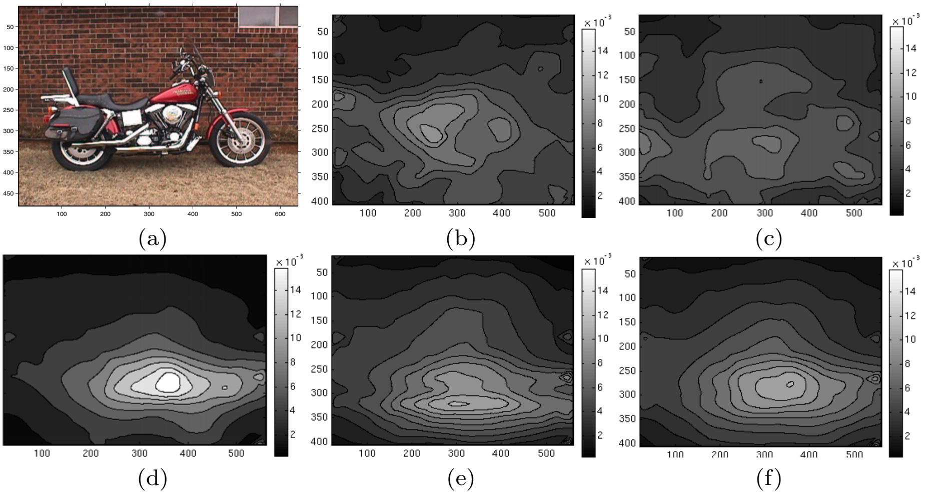
Here K_σ is a Gaussian kernel function with diagonal covariance matrix $\Sigma = \sigma \cdot \mathbf{I}$. Moreover, $\mathbf{s}_{ij}^{(l)}$ is the l-th shift vector found in the training data for a composition of parts represented by (f_i, f_j, r_{ij}). The number of shift vectors observed for such a composition in the training set is denoted $N = N(f_i, f_j, r_{ij})$. Therefore, the spatial density of the object center given one composition is

$$p(X = \mathbf{x}|f_i, f_j, r_{ij}, \mathbf{x}_{ij}, c^I) = p(S = \mathbf{x} - \mathbf{x}_{ij}|f_i, f_j, r_{ij}, c^I) \ . \qquad (11)$$

Using this equation the conditional probability to observe a composition at location $\mathbf{x}_{ij}$ can be written as

$$p(\mathbf{x}_{ij}|f_i, f_j, r_{ij}, c^I, \mathbf{x}) \propto p(S = \mathbf{x} - \mathbf{x}_{ij}|f_i, f_j, r_{ij}, c^I)\, p(\mathbf{x}_{ij}|f_i, f_j, r_{ij}, c^I) \ . \qquad (12)$$

To simplify the representation in the graphical model the locations $\mathbf{x}_{ij}$ are discretized on a regular 10×10 grid. The latter term is then approximated during learning by histogramming over the observed positions of compositions in the training data. Figure 5 gives an example for the estimation of object locations.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Spatial density of the object location, given some categorization C^I. (b) displays $p(\mathbf{x}|\{f_i, f_j, r_{ij}, \mathbf{x}_{ij}\}_{<i,j>}, C^I = car)$. In (c) the category C^I is *face*, in (d) *motorbike*, in (e) *airplane*. (f) shows the inferred final belief for the object center position, $p(\mathbf{x}|E)$. Note that the density for the true category in (d) and the final belief are both nicely peaked at the true center.

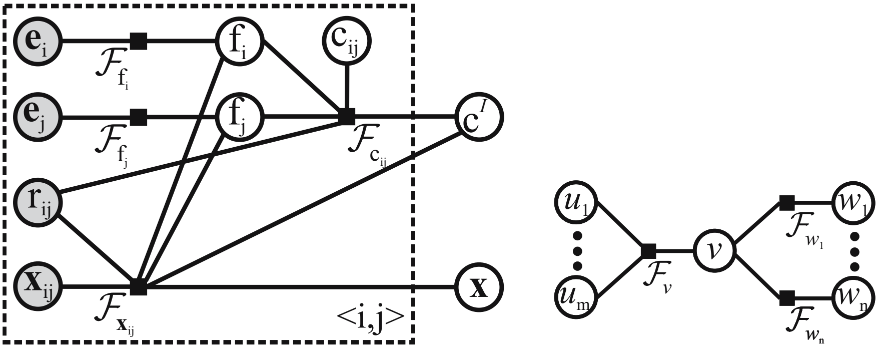
3.5 Inference of Image Categorization

During recognition, loopy belief propagation is performed using the evidence $E = \{\mathbf{e}_i\}_i \cup \{r_{ij}, \mathbf{x}_{ij}\}_{<i,j>}$ to categorize the scene as a whole, i.e. we are interested in the belief of the random variable C^I. Belief propagation simplifies the complex problem of optimizing a marginalization of the joint distribution over all model variables. This is rendered possible by using the independence properties represented in the graphical model and taking only the resulting local interactions into account. To simplify the computation scheme we transform the Bayesian network from Figure 2 into the factor graph (cf. [10]) displayed in Figure 6(a). Function nodes represent the conditional probability distributions, whereas the remaining variable nodes correspond to the random variables of the Bayes net. To propagate beliefs each vertex in this graph has to compute and send messages to its neighbors as follows: Consider some variable node v that has function node neighbors $\mathcal{F}_v$ and $\mathcal{F}_{w_1}, \dots, \mathcal{F}_{w_n}$ as depicted in Figure 6(b). Adjacent to each $\mathcal{F}_{w_i}$ is again some variable node w_i and $\mathcal{F}_v$ has variable node neighbors v and $u_1, \dots, u_m$. Now an unobserved variable sends messages to its function node neighbors by taking all the incoming messages into account [10],

$$\mu_{v \to \mathcal{F}_v}(v) := \prod_i \mu_{\mathcal{F}_{w_i} \to v}(v) \ . \tag{13}$$

If v is an evidence variable and observed to be in state v' then this message is just $\mu_{v \to \mathcal{F}_v}(v) = \mathbf{1}\{v = v'\}$, where $\mathbf{1}\{.\}$ denotes the characteristic function. Moreover a function node sends the following messages to its neighbors

$$\mu_{\mathcal{F}_v \to v}(v) := \sum_{u_1, \dots, u_m} \mathcal{F}_v(v, u_1, \dots, u_m) \prod_j \mu_{u_j \to \mathcal{F}_v}(u_j) \ . \tag{14}$$



(a) (b)

Fig. 6. (a) Conversion of the Bayesian network from Figure 2 into a factor graph representation. The function nodes $\mathcal{F}_\bullet$ represent the posterior of the corresponding random variable, e.g. $\mathcal{F}_{f_i} = P(f_i|\mathbf{e}_i)$, see text for details. (b) A simple factor graph used for illustrating belief propagation.

The belief of v given all the present evidence E is then the product of all of its incoming messages

$$P(v|E) \propto \mu_{\mathcal{F}_v \to v}(v) \prod_i \mu_{\mathcal{F}_{w_i} \to v}(v) \ . \tag{15}$$

In conclusion, our architecture propagates beliefs not only in a bottom-up manner from the observed image features to infer categorization and object location. The system also propagates information backwards in a top-down fashion from object localization and categorization to composition and part hypotheses, c_{ij} and f_i respectively. While the bottom-up feature-driven and the top-down category model driven updates are performed concurrently, hypotheses get improved by finding those which have optimal mutual agreement.

4 Evaluation

In the following, the proposed architecture is evaluated on the CalTech image database. During learning, some 700 images are presented to the system together

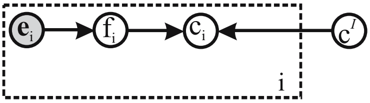


Fig. 7. A simple Bayesian network for categorization used to evaluate the gain of compositionality

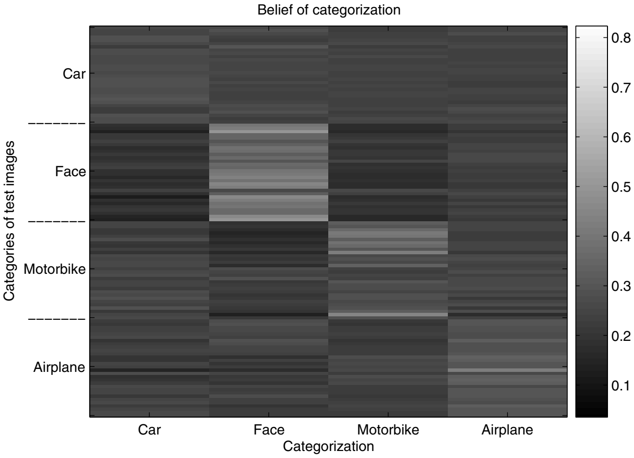


Fig. 8. Category confusion matrix for categorization without using compositionality. The predicted class is shown on the x-axis, whereas the true class is on the y-axis. This model achieves an overall classification rate of 82.5%.

with the image category labels as the only additional information. The test scenario is then to classify previously unknown images as belonging to one of the categories. Moreover, a confidence in this categorization is returned.

In order to evaluate the gain of compositionality in categorization, we first investigate a simpler model. It is based on the same image representation but with neither compositions nor a shape model (see Figure 7). Therefore the categorizations c_{ij} of compositions are replaced by the classification c_i of single parts, where $P(c_i|f_i, c^I)$ is empirically estimated from the training data in the same way as the c_{ij} in Section 3.3. Figure 8 displays the resulting category confusion matrix. The confidence in a categorization of class c^I (shown on the x-axis) of images with a given ground truth category (shown on the y-axis) is visualized in this figure. Therefore a row represents the beliefs of the different categories

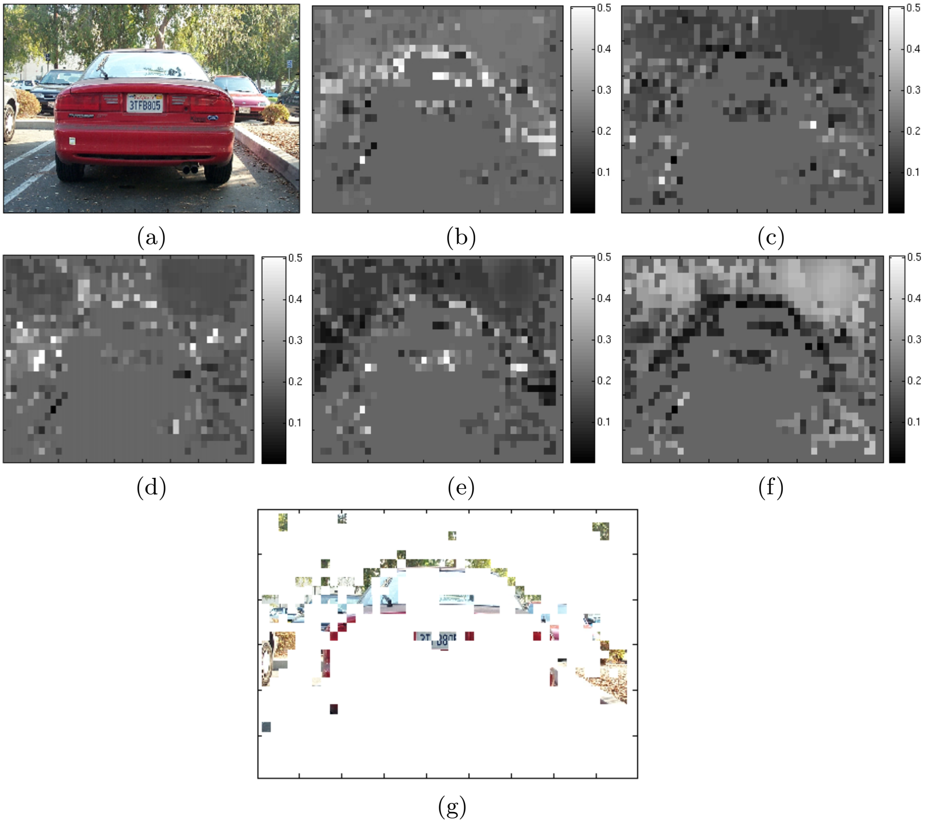


Fig. 9. Categorization belief of compositions. For each composition c_{ij}, (b) displays $P(C_{ij} = \text{car}|E)$ at the position of patches i and j. See Figure 4 for details. (c) Shows the posterior for class *face*, (d) for *motorbike*, (e) for *airplane*, and (f) for *background*. Where the last category facilitates a figure-ground segregation of compositions. (g) Shows the regions that support categorizing the image as *car*.

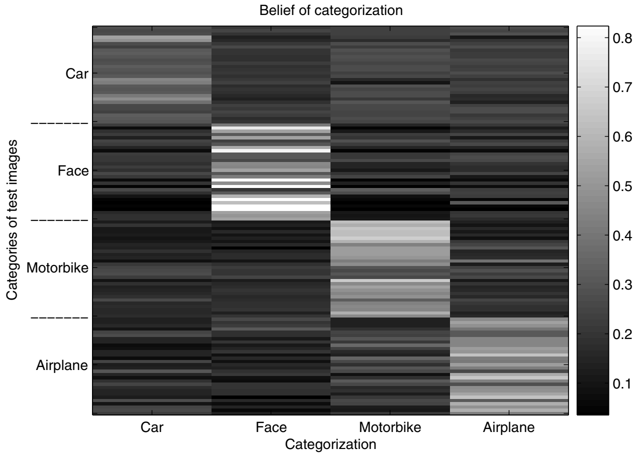


Fig. 10. Category confusion matrix for categorization based on the full model with compositionality and shape. This approach achieves an overall classification rate of 91.7% and has a significantly higher confidence than the previous one. Considering only the last three categories a recognition rate of 94.4% is achieved (see text for details). Compare this with the 93.0% reported in [6] for the same three categories.

for one test image. This model achieves an overall correct classification rate of 82.5%. However, the categorization confidence is quite low.

Subsequently, this simple model is to be compared with the full approach outlined in Section 3. Figure 10 displays the category confusion matrix of the model that is based on compositions and their spatial arrangement. When comparing the two plots it becomes evident that the system with compositionality and shape achieves a significantly increased confidence in the correct categorizations. Moreover, the recognition rate has increased to 91.7%. This illustrates that the relations used for compositions add crucial information to that already present in the individual parts. As one can see, most of the error results from falsely classified car images. This is due to the fact that the interest point detector returns only very few votes for the large homogeneous parts of the body of a car. Most detections are in the background or at the outline of the vehicle. This is also apparent in the illustration of compositions in Figure 9. Although for this specific dataset the background features would provide good indications for the presence of cars, we do not want to introduce such dependencies as they are to a great deal database specific and would very likely lead to an overfitting to this image collection. In a future extension of the approach we therefore plan to revise the interest point detection stage to also incorporate homogeneous regions. When leaving out the car category and considering only the remaining three ones, the present approach achieves an overall recognition rate of 94.4%. Compare this with the average recognition rate of 93.0% reported by Fergus et al. in [6] for the same three categories.

5 Conclusion and Future Work

Inspired by Geman's compositionality approach [9], we have devised a novel model for object categorization on the basis of a Bayesian network. The underlying image representation emphasizes keypoint relations and it accounts for large intra-class variations as they are usually encountered in general object categorization. Models for compositions and global object shape have been introduced and tightly combined in a single graphical model to infer image categorizations based on the underlying statistical framework. As a result, the system not only propagates information on an image category in a feature-driven, bottom-up manner. It also uses category models to corroborate such hypotheses by a model-driven, top-down inference, thereby reconciling different locally proposed categorization hypotheses by belief propagation. The system achieves competitive recognition performance on a standard image database used for categorization.

The approach shows significant potential for future extensions at several stages. First, feature representation should incorporate multiple scales and segmentation or other prior information to deal with homogeneous regions. Moreover, compositions could be formed in a recursive manner to yield a representation that is semantically closer to the final image categorization. Also, additional types of relations should add significant information on the objects present in a scene. All these refinements are expected to be necessary for large-scale experiments with hundreds of classes and a diverse nature of rigid, articulate and flexible objects.

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. I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.
3. E. Bienenstock, S. Geman, and D. Potter. Compositionality, mdl priors, and object recognition. In *NIPS*, volume 9, 1997.
4. E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *CVPR Workshop on Perceptual Organization in Computer Vision*, 2004.
5. E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, 2002.
6. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.
7. R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV*, 2004.
8. M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1), 1973.
9. S. Geman, D. F. Potter, and Z. Chi. *Composition Systems.* Technical report, Division of Applied Mathematics, Brown University, Providence, RI, 1998.
10. F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47(2), 2001.

11. M. Lades, J. C. Vorbrüggen, J. M. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput.*, 42, 1993.
12. B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on Stat. Learning in Computer Vision*, 2004.
13. B. Leibe and B. Schiele. Scale-invariant object categorization using a scale-adaptive mean-shift search. In *Pattern Recognition, DAGM*, 2004.
14. D. G. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Norwell, MA, 1985.
15. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2), 2004.
16. K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Computer Vision*, 60(1), 2004.
17. K. Murphy, Y. Weiss, and M. Jordan. Loopy-belief propagation for approximate inference: An empirical study. In *UAI*, 1999.
18. B. Ommer and J. M. Buhmann. A compositionality architecture for perceptual feature grouping. In *EMMCVPR*, 2003.
19. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.
20. R. C. Veltkamp and M Tanase. Content-based image and video retrieval. In O. Marques and B. Furht, editors, *A Survey of Content-Based Image Retrieval Systems.* Kluwer, 2002.
21. M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV*, 2000.
22. G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods—A Mathematical Introduction.* Springer, 2nd edition, 2003.