A Compositionality Architecture for Perceptual Feature Grouping

Björn Ommer and Joachim M. Buhmann

Rheinische Friedrich-Wilhelms-Universität
Institut für Informatik III, Römerstr. 164
D-53117 Bonn, Germany
{ommer, jb}@cs.uni-bonn.de
http://www-dbv.informatik.uni-bonn.de

Abstract. We propose a compositionality architecture for perceptual organization which establishes a novel, generic, algorithmic framework for feature binding and condensation of semantic information in images. The underlying algorithmic ideas require a hierarchical structure for various types of objects and their groupings, which are guided by gestalt laws from psychology. A rich set of predefined feature detectors with uncertainty that perform real-valued measurements of relationships between objects can be combined in this flexible Bayesian framework. Compositions are inferred by minimizing the negative posterior grouping probability. The model structure is founded on the fundamental perceptual law of *Prägnanz*. The grouping algorithm performs hierarchical agglomerative clustering and it is rendered computationally feasible by visual pop-out. Evaluation on the edgel grouping task confirms the robustness of the architecture and its applicability to grouping in various visual scenarios.

1 Introduction

Simple entities like points, isolated edgels or even small image patches provide only local, relatively unreliable information about objects in the image. This fact and the large number of such simple objects call for a processing step that concentrates information in few descriptive objects. The procedure focuses on the relevant entities in an image and increases the information content of single entities by forming complex *groupings* of simple ones. As a result, a hierarchy of compositions is created that contains objects of increasing robustness and growing relevance with respect to the whole image. This goal is exactly the primary objective of a composition system, as presented by Geman et al. [7]. Although our system resembles in its structure more neural nets and graphical models than stochastic grammars as used in [7, 4], the principle of compositionality constitutes a central part of our approach which follows from the philosophy of Geman. Psychological and neurophysiological concepts related to perceptual organization are ported and integrated into a uniform, algorithmic framework using methods common to computer science such as energy minimization and

Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'03),
Lisbon, LNCS 2683, pp. 275–290, Springer, 2003.

compositionality. The goal is to obtain a universally applicable, robust grouping algorithm that can be used as a preprocessing step for subsequent applications such as search and retrieval of objects.

The above described processing has the underlying psychological motivation that a representation of the original image which is based on such groupings is *perceptually more salient* and *simpler* than unstructured sets of pixels or edgels. To create simple, stable, and perceptually salient groupings, Gestalt psychology proposes the fundamental principle of *Prägnanz* [5, 8]. Moreover, numerous simple (compared to Prägnanz) *Gestalt laws* [5, 8] exist that approximately entail Prägnanz. These laws (e.g. proximity of entities, their closure, or their similarity in orientation) form the basis for the implemented feature detectors, called *relations* in the following. These relations act as *sensors* that perceive a certain mutual relatedness of objects in a perceptually meaningful way.

In this paper we introduce a highly flexible algorithmic framework that provides a uniform embedding of arbitrary, uncertain feature detectors. Consequently we also have to control the interaction of a large number of these components. To unify and ease the design of different kinds of feature detectors, we define a uniform basic structure for these detectors which spans the probability space for feature groupings. The negative posterior grouping probability serves as a cost function which measures the quality of different perceptual organizations. The grouping algorithm performs hierarchical agglomerative clustering [6] of these objects to minimize the grouping costs. To speed up the process, psychological concepts like visual pop-out efficiently prune the solution space.

As a result, our approach achieves the goal to reduce the dependence on single relations and, thereby, it is robustified against cues that are accidentally erroneous in certain situations. The analysis and processing of the numerous feature detectors produces a *sensor fusion* scheme that allows us to extract the various hints or cues for a grouping out of the scene.

The basic ideas on perceptual organization in computer vision relevant to the presented architecture are motivated by Lowe in [11]. However, our approach differs from Lowe's and other related work [9, 14, 17] in the fundamental way the feature detectors are used: Our framework handles numerous features in a flexible way compared to the alternative approaches which base their grouping decisions on a very limited number of fixed cues. Williams and Thornber [17] use a random walk process to formulate a new saliency measurement. By defining saliency measures for edgels, Shashua and Ullman [14] separate edgels into highly salient figure elements and background elements of lower saliency. Jacobs [9] computes the likelihood that a group of edgels is produced by a single object to direct a recognition system and improve its accuracy.

The strategy to implement a large number of uncertain feature detectors allows us to simplify the specification of concrete relations as far as possible. Related work by Amir and Lindenbaum [1] on the integration of large numbers of sensors differs exactly in this system design step. While the relations in their approach are mainly treated as binary valued random variables (presence or absence of relationships between objects), that have to be specified by the user, we offer a Bayesian concept for real-valued, uncertain feature detectors. Thereby different features of wide ranging strength can be detected and compared.

The next section presents a discussion of Prägnanz to provide the foundation for the cost function. Thereafter our algorithmic framework for the integration of arbitrary relations into a common cost function will be described. Section 4 will present the perceptually optimized grouping algorithms. Finally the performance of the presented architecture will be evaluated in section 5.

2 Prägnanz, Redundancy, and MDL

Fred Attneave [2] has stated that the received visual input is highly redundant and that large parts of a scene are predictable, given only a small fraction of the overall stimulus. This is due to some restrictions in the visually perceivable world he attributes to *lawfulness* of nature.

Given (an estimate of) the complexity of the visually perceivable world (e.g. by measuring the relative frequency of certain involved phenomena), it is possible to acquire a set of rules that predict large parts of a scene on the basis of only a small number of detected features. These rules, which exploit redundancy, are just the Gestalt laws. Moreover, a grouping of maximal Prägnanz can be understood as an optimal exploitation of this redundancy [2]. Consequently, this repetition of information within clusters can be avoided in the joint, probably even lossy encoding of the objects. The resulting strategy represents a realization of the minimum description length principle.

These ideas are illustrated by a brief example: Consider two straight edgels that are situated adjacent to each other, each represented by its respective endpoints. A joint encoding of the grouping can be represented by three points, or in case of collinearity by only two points. More complex entities such as squares or templates for natural objects such as cars need even fewer parameters in proportion to the number of their elementary constituents. Generalizing this idea, Biederman [3] proposed *geons*, a set of image primitives that can be used as components of complex visual scenarios. In conclusion, the goal of a grouping that obeys maximal Prägnanz can be reached by following the MDL principle and minimizing a corresponding cost function.

Splitting Up Prägnanz: A direct specification of a cost function for Prägnanz is too complicated and we, therefore, replace the Prägnanz concept by directly measurable relations: Gestalt psychology proposes a number of Gestalt laws, which analyze similarity of objects in certain image features (e.g. color or orientation). Thereby, the relations detect the redundancies of the resulting grouping which favors the simplicity in image coding.

3 An Energy Function for Gestalt Principles

Outline of the Algorithm: The input to the grouping algorithm are *edgels*, short straight edge elements gained from a Canny edge detector. Furthermore color histograms in a small neighborhood on both sides of each edgel are computed. Thereafter the grouping algorithm uses various feature detectors to obtain information about the mutual relationship of objects in order to create a hierarchy of groupings. This procedure is mainly bottom-up, thereby not requiring additional knowledge about the scene.

3.1 Objects

To support the idea of minimizing the overall description length, this approach defines a hierarchy of objects of different complexity. Composite types of objects are defined via inclusion of simpler ones, e.g. curves are defined as compositions of edgels or other curves. The basic objects are *edgels* that result from a Canny edge detector. An edgel E is described by its two endpoints in Euclidian space, $\mathtt{Ep}_i(E) \in \mathbb{R}^2, i \in \{-1,1\}$. The goal is to group these objects to perceptually meaningful curves which are the second type of objects we use. These entities are applicable to widely differing scenarios and are therefore chosen to exemplify the general architecture subsequently. Other types of objects are grouped in a similar fashion using appropriate features. The endpoints of a curve C are recursively defined via its components, which are edgels or other curves. Assuming a grouping of two objects O_{-1}, O_1 that can be edgels or curves, the endpoint of entity O_α that is closest to the endpoints of the other object has the index

$$\mathtt{cEp}(O_\alpha, O_{-\alpha}) := \underset{i\in\{-1,1\}}{\operatorname{argmin}} \min_{j\in\{-1,1\}} \left\{ \left\| \mathtt{Ep}_i(O_\alpha) - \mathtt{Ep}_j(O_{-\alpha}) \right\|_2 \right\} \ . \tag{1}$$

The endpoints of C are then

$$\mathtt{Ep}_i(C) := \mathtt{Ep}_{-\mathtt{cEp}(O_i, O_{-i})}(O_i), i \in \{-1,1\} \ . \tag{2}$$

Moreover the additional feature of orientation is added to the endpoints of an edgel (the curves take over this parameter during the grouping):

$$\mathtt{Ep}_i^{\mathrm{orient}}(E) := \measuredangle \left\{ \mathtt{Ep}_i(E) - \mathtt{Ep}_{-i}(E); \ \ (1,0)^T \right\}, i \in \{-1,1\} \ . \tag{3}$$

Each object, except for the singleton edgels, is a grouping of other objects. The goal is to group objects to perceptually meaningful compositions that constitute new objects. The hierarchy of the created compositions is logged in a rootless tree, the *dependency graph* of the participating objects. Each vertex corresponds to an object (the edgels form the leafs of this structure), while the arcs represent the grouping relationships between compositions and their subparts.

3.2 Relations

In the following, a generic and flexible framework for all different kinds of grouping cues will be presented. As a result a modular and adjustable cost function is obtained that measures the gain in Prägnanz given arbitrary features. Therefore each of the underlying relations corresponds to a technical *neuron* that perceives a specific mutual relatedness of objects from a previous level of the hierarchy mentioned above as its input. The outputs of sensors with the same entities as input are combined to obtain the overall grouping cost function. These measures are in turn used recursively for subsequent groupings via a feed-forward architecture of successively applied relations. Given this structure, perceptual organization can be understood literally as applying relations to their respective sensory input in order to *organize* these *percepts* in a dependency graph of objects.

The final output of each relation, the *vote*, corresponds to the probability that the considered objects belong to a joint object, given the information on their mutual relationship that is gathered by the relation. Letting r denote the output of the relation and using only two objects O_1 and O_2 for simplicity (each of these is a singleton or a grouping of objects), the probability of a perceptually favorable grouping is

$$P\left(O_1 \odot O_2 | r\left(O_1, O_2\right)\right) \ . \tag{4}$$

Here the symbol $\odot$ indicates that the objects join to form a new grouped object (e.g. a new composition $O := O_1 \odot O_2$).

In order to compute a relation it is splitted into two separate parts. At first a feature extration function is used to obtain information on the relationship between objects. This sensor data is then normalized and interpreted so that all different relations provide equation (4) with an unified input. The output of the relation is the composition

$$r := \mathtt{sI} \circ \mathtt{sD} \ . \tag{5}$$

The first part computes the strength of a specified relationship of the objects (e.g. their relative orientation or similarity of color) given the data. This part corresponds to the input nodes of a perceptron and its output resembles the internal activity level, or the *action potential* of their neural counterparts. Letting $\mathfrak{O}$ denote the set of all objects and assuming for simplicity a compact codomain $[sd_{min}, sd_{max}]$, this feature extraction can be summarized as

$$\mathtt{sD} : \mathfrak{O} \times \mathfrak{O} \rightarrow [sd_{min}, sd_{max}] \subset \mathbb{R} \ . \tag{6}$$

This function comprises both the perception of relevant features of the objects and the computation of a relation specific distance measure. In order to combine different measures they have to be normalized, i.e., this normalization corresponds to the output function of a neuron,

$$\mathtt{sI} : [sd_{min}, sd_{max}] \rightarrow [0, 1] \ . \tag{7}$$

Subsequently, the grouping probability can be computed using the output of the relation and Bayesian decision theory [6]

$$P(O_1 \odot O_2 | r) = \frac{p(r|O_1 \odot O_2) \cdot P(O_1 \odot O_2)}{p(r)} \ . \tag{8}$$

Using marginalization, the evidence can be written as

$$p(r) = \underbrace{p(r|O_1 \odot O_2)}_{\rightsquigarrow \text{ causal reason}} \cdot P(O_1 \odot O_2) + \underbrace{p(r|O_1 \not\odot O_2)}_{\rightsquigarrow \text{ accidentalness}} \cdot P(O_1 \not\odot O_2) \ . \tag{9}$$

In the following a Gaussian probability density function

$$\varphi(r) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r-\mu)^2}{2\sigma^2}} , \ \ \mu \in [0,1], \ \ \sigma \in \mathbb{R}_+ \tag{10}$$

is used as a parametric model of the likelihood of a relation, Φ being the cumulative distribution. The range of mean μ is restricted to the range of r. The likelihood is then a rectified Gaussian

$$p(r|O_1 \odot O_2) = \frac{\varphi(r)}{\Phi(1) - \Phi(0)} \ . \tag{11}$$

Furthermore, the mean μ is the expected value of the relation

$$\mu = E_{p(r|O_1 \odot O_2)}[r] \ , \tag{12}$$

given that the objects form a grouping. The variance σ^2 corresponds to the inverse significance of the relation and reflects its uncertainty. Having a distribution with a sharp peak, a response close to the optimum, $r(O_1, O_2) \approx \mu$, is a good indication for a reliable grouping.

In contrast to the likelihood, $p(r|O_1 \not\odot O_2)$ indicates how likely it is that r takes on a certain value, given that the objects do not form a perceptually meaningful composition. In absence of further information, most relations permit the approximation of this *accidentalness* term by assuming that all outputs are equally likely in this case, c.f. [13].

3.3 The Energy Function

The approach uses numerous relations that provide knowledge about different features in order to find out whether objects form a grouping. Therefore, sensors that are uncertain about a specific clustering of entities can be compensated by others. Given a *sensor fusion* of n different relations, a voting scheme (see figure 1) is proposed that pools them in the overall energy function

$$\mathcal{H}(O_1 \odot O_2 | r_1, \ldots, r_n) = -\log \left[\prod_{i=1}^{n} P(O_1 \odot O_2 | r_i)^{\frac{\mathtt{w}_i}{\sum_{j=1}^{n} \mathtt{w}_j}} \right] \ . \tag{13}$$

The weights $\mathtt{w}_i$ correct slight statistical dependencies between the features that are detected by different relations. If they are all independent, i.e., $p(r_i|r_{i+1}, \ldots, r_n, O_1 \odot O_2) = p(r_i|O_1 \odot O_2)$, all weights are equal and $\mathcal{H}$ leads to the same grouping decisions as $-\log P(O_1 \odot O_2 | r_1, \ldots, r_n)$, c.f. [13]. An example of such a correction is our model for the relation that detects the parallelism of the ends of two curves. Since it relies on similar features as the relation for collinearity, the weight is lower than the chosen standard value of 2. The management of the voting information is carried out in a knowledge base. With this framework each relation can be designed without much effort by specifying the two sensor functions and the parameters μ, σ, and the weight. Furthermore, these variables offer a uniform point of access to adjust the relations and their interaction.

The overall cost function for the current grouping of the complete image is formed by the set $\mathfrak{O}_l$ of all groupings in the latest level of the dependency graph,

$$\widehat{\mathcal{H}}(\mathfrak{O}_l|r_1,\dots,r_n) = \sum_{O\in\mathfrak{O}_l} \mathcal{H}(O|r_1,\dots,r_n) \ . \tag{14}$$

3.4 Implemented Relations

So far a number of relations have been specified to illustrate the potential of the presented architecture. In the following the design of some of these sensors will be described in detail. The implemented Gestalt laws are *proximity*, *similarity* of *orientation* and *color*, *good continuation*, and *closure* (see [5, 8]).

Let $O_{-1}, O_1 \in \mathfrak{O}$ be two curves or edgels with endpoints $\texttt{Ep}_i(O_j), i, j \in \{-1, 1\}$ and the corresponding grouping $O_{-1,1} := O_{-1} \odot O_2$. Furthermore, $\texttt{maxDist} \in \mathbb{R}$ is set to the length of the image diagonal and the abbreviation

$$\widehat{\texttt{cEp}}(O_i, O_{-i}) := \texttt{Ep}_{\texttt{cEp}(O_i,O_{-i})}(O_i), \ \ i \in -1, 1 \tag{15}$$

is used. The relation for *proximity* of the endpoints of two curves or edgels is

$$\texttt{sD}_{\texttt{prox}}(O_{-1}, O_1) := \left\|\widehat{\texttt{cEp}}(O_{-1}, O_1) - \widehat{\texttt{cEp}}(O_1, O_{-1})\right\|_2 \in [0, \texttt{maxDist}] \ , \tag{16}$$

$$\texttt{sI}_{\texttt{prox}}(d) := d/\texttt{maxDist} \ , \tag{17}$$

$$\mu_{\texttt{prox}} := \texttt{sI}(0), \ \sigma_{\texttt{prox}} := \texttt{sI}(\max\{t_l, \min\{t_h, \texttt{avgDist}\}\}), \ \texttt{w}_{\texttt{prox}} := 2.3 \ . \tag{18}$$

To estimate avgDist, the distances of a number of randomly located objects to their nearest neighbors (distance of endpoints) are averaged with threshold constants $t_l := 4, t_h := 40$.

Relative orientation (or parallelism) of the ends of two curves or edgels is modeled by

$$\texttt{sD}_{\texttt{orient}}(O_{-1}, O_1) := \left|\texttt{Ep}^{\text{orient}}_{\texttt{cEp}(O_1,O_{-1})}(O_1) - \texttt{Ep}^{\text{orient}}_{\texttt{cEp}(O_{-1},O_1)}(O_{-1})\right| \mod 2\pi \ , \tag{19}$$

$$\texttt{sI}_{\texttt{orient}}(\alpha) := 1 - \frac{|\pi - \alpha|}{\pi} \ , \tag{20}$$

$$\mu_{\texttt{orient}} := 1, \ \sigma_{\texttt{orient}} := \texttt{sI}_{\texttt{orient}}(10^\circ/180^\circ \cdot \pi), \ \texttt{w}_{\texttt{orient}} := 1.0 \ . \tag{21}$$

A simple, computationally feasible way to measure the gain in *closure* of object O_{-1} by grouping it with O_1 is modeled as follows:

$$\texttt{gap}_{\texttt{s}}(O_{-1}, O_1) := \left\|\widehat{\texttt{cEp}}(O_1, O_{-1}) - \widehat{\texttt{cEp}}(O_{-1}, O_1)\right\|_2 \ , \tag{22}$$

$$\texttt{gap}_{\texttt{o}}(O_{-1}, O_1) := \left\|\texttt{Ep}_{-\texttt{cEp}(O_1,O_{-1})}(O_1) - \texttt{Ep}_{-\texttt{cEp}(O_{-1},O_1)}(O_{-1})\right\|_2 \ , \tag{23}$$

$$\texttt{sD}_{\texttt{closure}}(O_{-1}, O_1) := \max\left\{\frac{\texttt{gap}_{\texttt{s}}(O_{-1}, O_1) + \texttt{gap}_{\texttt{o}}(O_{-1}, O_1)}{\left\|\texttt{Ep}_{-1}(O_{-1}) - \texttt{Ep}_1(O_{-1})\right\|_2}, l\right\} \ . \tag{24}$$

Of course equation (24) is only applicable to objects that are not perfectly closed and a parameter $l := 2$ penalizes losses in closure above this bound equally. Therefore short curves, whose closure can fluctuate a lot when they are grouped, can still receive a fixed lower probability mass from the rectified gaussian introduced above. The remaining components of this relation are

$$\mathtt{sI}_{\mathtt{closure}}(g) := g/l \ , \tag{25}$$

$$\mu_{\mathtt{closure}} := \mathtt{sI}(0), \ \ \sigma_{\mathtt{closure}} := \mathtt{sI}(1/2), \ \ \mathtt{w}_{\mathtt{closure}} := 2.0 \ . \tag{26}$$

Furthermore, our system computes color histograms for small areas (depending on the length of an edgel) on both sides of an edgel E. The means $\bar{c}^E_{-1}, \bar{c}^E_1 \in [0, 255]^3$ of these two histograms are added to the endpoint section of each edgel so that the curves inherit these features,

$$\mathtt{Ep}_i^{\mathrm{color}_j}(E) := \bar{c}^E_{i \cdot j}, i, j \in \{-1, 1\} \ . \tag{27}$$

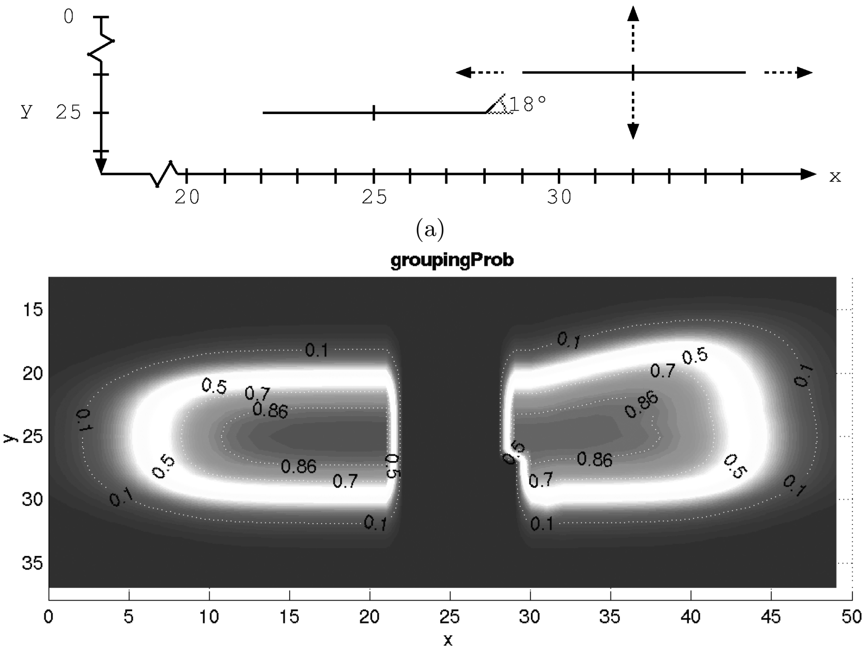
Relying on the color information of the edgels, *similarity in color* of the local surroundings of two curves and *color contrast* of both sides of an edgel are used. Another implemented feature detector is the *collinearity* of the ends of two curves. In combination with relative orientation and proximity, an implementation of the Gestalt law of good continuation is formed by these three relations. Finally another relation performs the recursive *propagation* of grouping probabilities from the component objects in the dependency graph to their composition.

4 Grouping Algorithms

The grouping algorithm perceives a certain mutual relatedness of some objects by applying the relations. The resulting energy function provides the necessary information on which entities to combine.

The solution space in which the grouping algorithm searches, consists of groupings of the whole image. Each of these is represented by a level in the dependency graph and consists of all grouped objects that are necessary to describe the relevant aspects of the image. The grouping algorithm produces a new level in the graph by grouping entities from a previous level. The resulting hierarchical agglomerative clustering starts on the initial level that contains only edgels and returns the last level of the graph that represents the final grouping of the image. The algorithm performs a search for a perceptually favorable grouping by minimizing the energy function. Successive states of the clustering procedure in solution space correspond to successive levels in the dependency graph.

On the one hand a perceptually optimal grouping of the whole image is to be found eventually. On the other hand the search for such a solution has to be speeded up in order to obtain computationally feasible algorithms that find these groupings in the enormously large solution space in reasonable time. This poses a trade-off between optimality and feasibility. It is possible to come closer to both objectives simultaneously by placing perceptually motivated restrictions



(a)

(b)

Fig. 1. (a) Two short curve segments, one moving about the other. The fixed one has a bend of 18° at one end. (b) Plot of the cost function $\exp -\mathcal{H}$ (labeled `groupingProb` in the figure) resulting from the application of the relations *proximity*, *relative orientation* and *collinearity* to these segments

(c.f. section 2) on the continuations of paths through solution space so that these lead to reasonably good solutions. In the following we will discuss the grouping of curves. The grouping of more complex objects proceeds in the same manner.

Greedy Approach: A significant decrease in complexity can be achieved by making grouping decisions in a *greedy* fashion. Provided the set of relations conveys enough information, groupings will persist once they are formed. In this case psychology indicates that the resulting speed up is *not* at the expense of optimality. The energy function $\mathcal{H}$ measures the gain in Prägnanz resulting from performing one additional grouping that is discriminating two successive states in solution space. In the above depicted case of persisting groupings this local optimization leads to a global optimum. Furthermore, the idea is to correct erroneous compositions in later stages by producing more complex objects than curves that are less prone to accidental influences. One cause for such a robustness is additional information about missing components that is available once a great fraction of a complex object has been detected. Therefore, complex compositions integrate information over large parts of the image. However, work on these concepts is still in progress.

To further improve the speed of the grouping algorithm, we reduce the branching factor by early commiting to grouping decisions: The procedure searches for a grouping partner of the curve with best $\mathcal{H}$, so that the composition yields improved Prägnanz, as measured by the energy function. If such a grouping has been found, the two components are not considered in further groupings. Starting with N curves each grouping reduces the number of curves by one, leading to an overall complexity of $\mathcal{O}(N^2)$.

Attention Control: The grouping algorithm applies the cost function $\mathcal{H}$ to pairs of curves. This process can be accelerated significantly by restricting the set of grouping partners and thereby limiting the possible successor states. This is motivated by observations on the way the (human) brain allots attention to the various inputs, thereby speeding up their perception enormously. Psychological and neurophysiological research analyzed the way these processes influence visual search [10, 16]. One key point is that targets that have features which differ from the surrounding distractors can *pop-out* pre-attentively. In contrast to ideas mentioned in [15], the algorithms developed for this contribution do not search for a known target in an image. Our goal is to speed up the grouping by developing processes similar to their pre-attentive neural counterparts that can group objects directly *without serially checking* all possible partners of an object as in regular pairwise grouping.

A uniform framework has been designed that uses some relations to compute a possibly multidimensional feature vector for each object. The features should constitute a relevant measure for both, *within-group attraction* and *between-group segregation*, so that groupable objects are similar in feature space. Therefore, the algorithm induces a partition of this space. The clusters are determined from an inverting of the relations so that groupable objects have at least one cell in common. The possible branchings of the decision tree of the previously presented grouping algorithm are significantly limited by performing this central clustering prematurely that assigns entities to the according cells. In contrast to the similarity measure used above, this preprocessing step computes the relevant features for each object on its own and not pairwise. Thereby only those groupings have to be reviewed by the pairwise procedure that have joint cells. Since only those groupings are neglected that would not have a chance to be grouped anyway, this acceleration still preserves the optimality and leaves the clustering resulting from the greedy approach unchanged.

The speed up is significantly higher than the acceleration that arises out of regular *shielding*. Shielding-effects occur when long range interactions are eclipsed by short, intermediate ones. The acceleration is due to the fact that a small set of grouping partners pops out immediately and no serial scanning is needed. Therefore the total complexity of the grouping algorithm reduces to $\mathcal{O}(N^{3/2})$. Images of size 400×400 pixels can thereby be processed in only a few seconds on a Pentium®II-400 with 128 Mb of RAM. Scenes of size 1000×1000

take only a few minutes to be grouped (depending on the choice of parameters and on the image this takes about one to three minutes).

5 Evaluation of the Framework

In the following, exactly the same grouping algorithm is applied to numerous different visual scenarios which emphasizes the flexibility of the architecture. Only the scale parameter and edgel density of the preprocessing edge detector are changed for the first image to illustrate the resulting phenomena. All other scenes are processed with a scale parameter $\sigma = 2$ of the edge detector. Distances are measured in pixel throughout this paper. To improve the legibility of the illustration we only visualize a certain fraction of the longest curves so that the output is not too densely filled with groupings. Again we are only using objects of type curve for these tests in order to present the architecture in its most general form.

Figures 2 a), c), and 3 a), c) show the original images, while figures 2 b), d), and 3 b), d) display the respective groupings. The first image illustrates the capabilities of the algorithm to group loosely coupled edgels coming from the edge detector. Moreover the scale parameter was set to 4 [pixel] for this image.

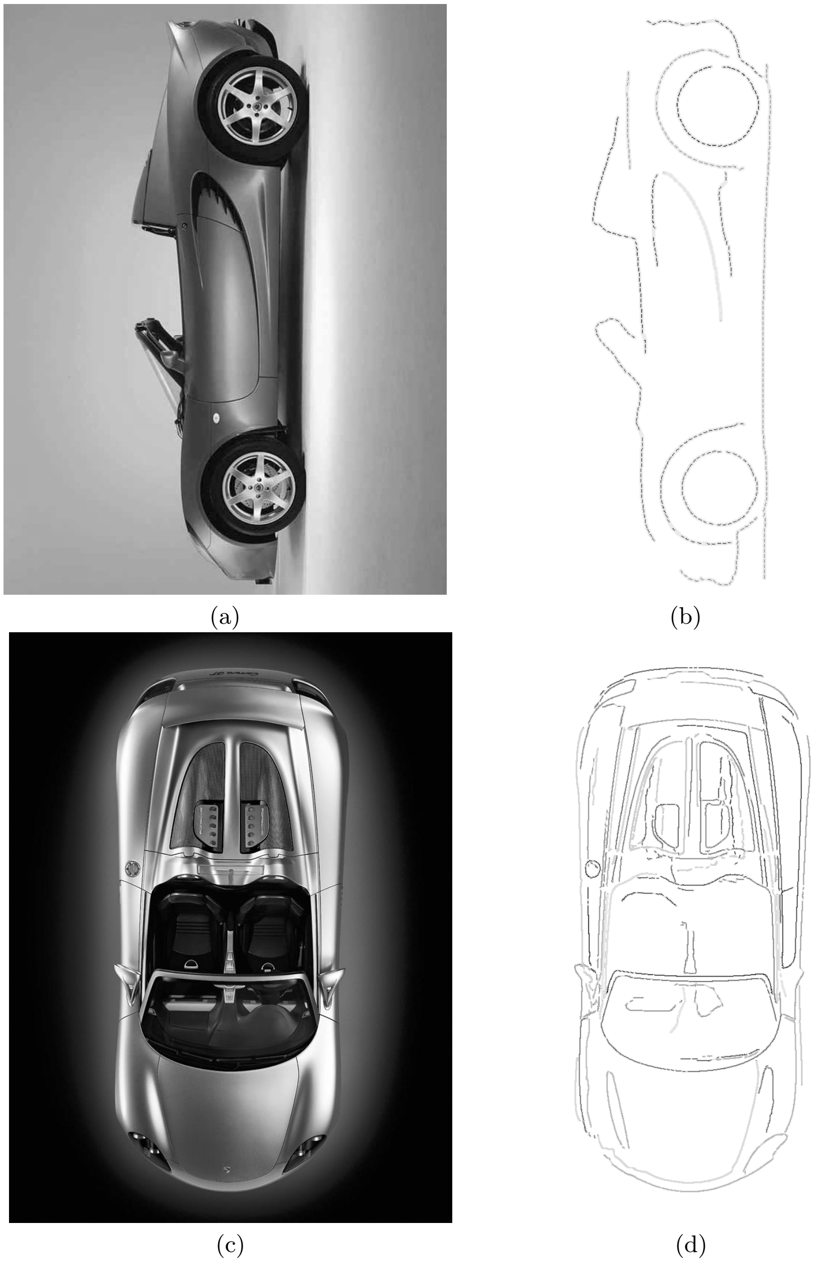
Figure 4 shows an image and the only minimally differing groupings of this original and of a version with added white noise. The signal to noise ratio is approximately 17 dB. Furthermore two images of similar cars viewed under different environmental conditions and perspectives were grouped (see figure 5) to illustrate the degree of invariance of certain curves against these effects.

Finally the grouping algorithm is tested using the human segmented images presented in [12]. Fifteen original images with about seven hand segmentations available for each image are used to select only those edgels coming from the Canny detector that lie on the respective human produced grouping. These remaining edgels are grouped using our system. The number of resulting curves and their length are then compared with those that are generated by using an edge point linking strategy as is common in implementations of the Canny detector.

On average, our algorithm generates only a fourth ($28\% \pm 3\%$) of the number of curves which are produced by the linking method while both methods explain the same number of edgels in the hand segmentations. Similarly, our curves are 4.2 ± 0.17 times longer. It remains to mention that these results probably would further improve, if the used images would have been less textured, since the currently implemented relations do not exploit region information.

6 Conclusion and Further Work

In this contribution a novel, flexible algorithmic framework for perceptual feature grouping has been presented based on fundamental principles of perceptual organization. The groupings are integrated into a hierarchical architecture that has been designed on the philosophy of compositionality. Moreover, a sensor fusion scheme has been presented that is flexible, problem independent, and easily



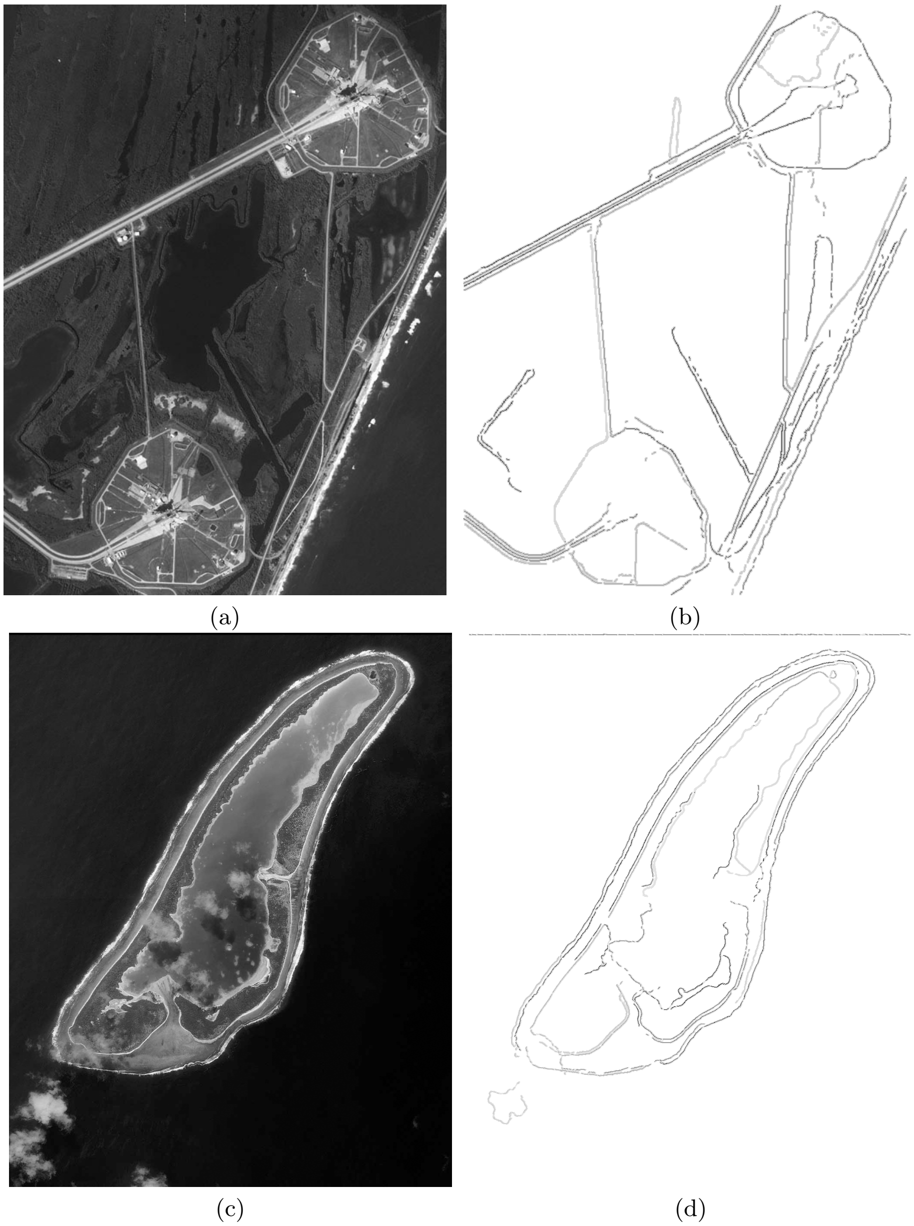
(a)

(b)

(c)

(d)

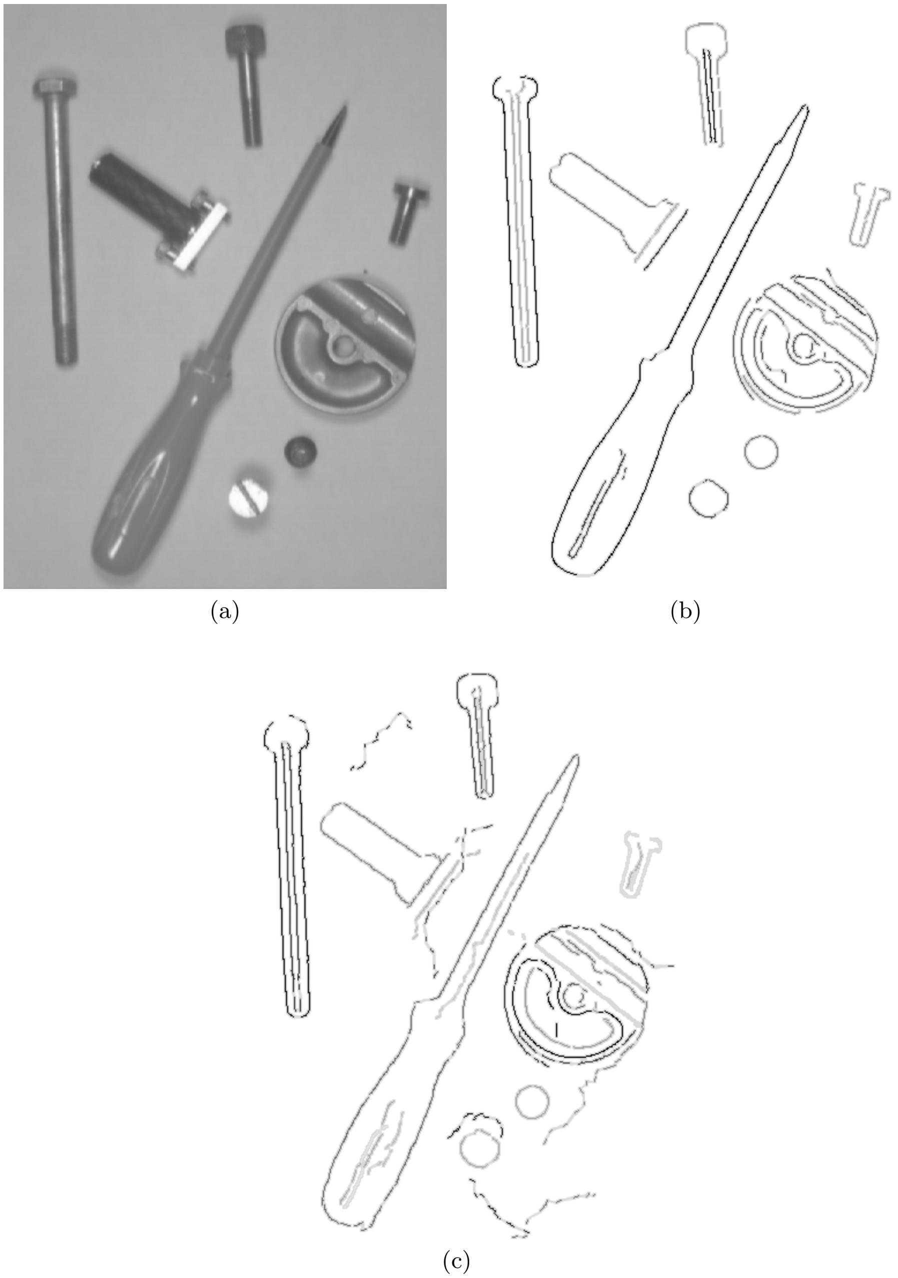
Fig. 2. (a) Original image. (b) Grouping after running Canny with $\sigma = 4$. Processing time is about 2 seconds on a PII-400 with 128Mb of RAM. (c) Original image. (d) Grouping after running Canny with $\sigma = 2$. Processing time is about 95 seconds



(a) (b)

(c) (d)

Fig. 3. (a) Original image from *www.spaceimaging.com* (b) Grouping after running Canny with $\sigma = 2$. Processing time is about 2.5 minutes. (c) Original image from *www.spaceimaging.com* (d) Grouping after running Canny with $\sigma = 2$. Processing time is about 3.5 minutes

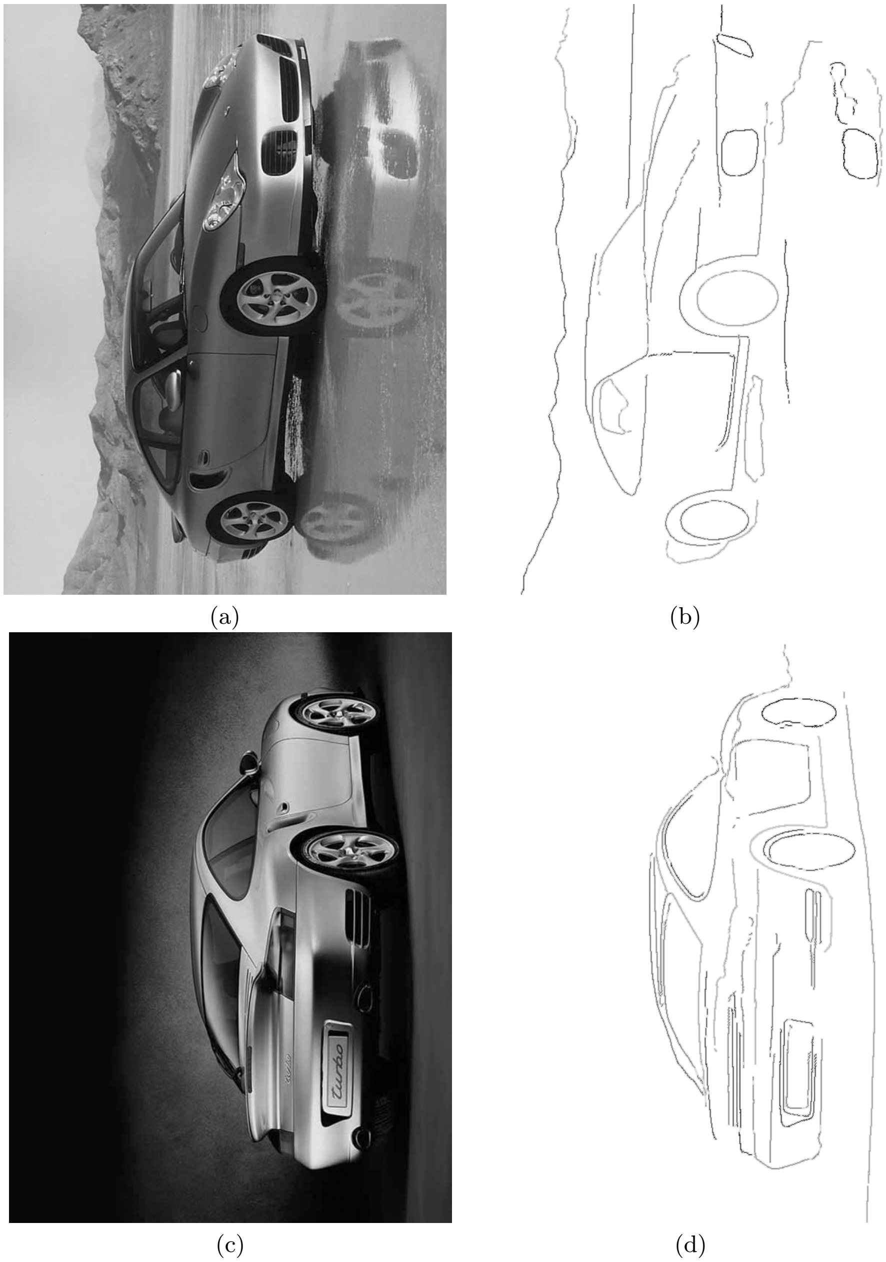


(a)

(b)

(c)

Fig. 4. (a) Original image. (b) Grouping after running Canny with $\sigma = 2$. Processing time is about 15 seconds. (c) Grouping of the noisy image after running Canny with $\sigma = 2$. Processing time is about 35 seconds



(a)

(b)

(c)

(d)

Fig. 5. (a) Original image. (b) Grouping after running Canny with $\sigma = 2$. Processing time is about 2 minutes. (c) Original image. (d) Grouping after running Canny with $\sigma = 2$. Processing time is about 70 seconds extendable with respect to the feature detectors that are used. As a result an open design of an energy function has been obtained that is directly based on fundamental perceptual principles. Using concepts from psychology the related grouping algorithm has been designed and brought to a computationally feasible form. Finally real world experiments have demonstrated that the architecture meets the aspired properties. Moreover, the framework has been embedded into a number of commonly used techniques including neural networks and graphical models (c.f. [13]) in order to gain new insights into perceptual organization and into the presented architecture.

References

1. Arnon Amir and Michael Lindenbaum. A generic grouping algorithm and its quantitative analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(2):186–192, 1998.
2. Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.
3. Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.
4. Elie Bienenstock. Notes on the growth of a composition machine. In *Proceedings of the Royaumont Interdisciplinary Workshop on Compositionality in Cognition and Neural Networks*, 1991.
5. Vicki Bruce, Patrick R. Green, and Mark A. Georgeson. *Visual Perception: Physiology, Psychology, and Ecology.* Psychology Press, 3rd edition, 1996.
6. Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification.* John Wiley & Sons, New York, NY, 2nd edition, 2001.
7. Stuart Geman, Daniel F. Potter, and Zhiyi Chi. *Composition Systems.* Technical report, Division of Applied Mathematics, Brown University, Providence, RI, 1998.
8. E. Bruce Goldstein. *Sensation and Perception.* Wadsworth, 3rd edition, 1989.
9. David W. Jacobs. Grouping for recognition. MIT AI Lab Memo 1177, MIT, Cambridge, MA, 1989.
10. Zhaoping Li. A V1 model of pop out and asymmetry in visual search. In *Advances in Neural Information Processing Systems*, volume 11, 1999.
11. David G. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Norwell, MA, 1985.
12. David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, volume 8, July 2001.
13. Björn Ommer. An algorithmic framework for visual grouping by perceptual organization. Diploma thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2003.
14. Amnon Shashua and Shimon Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *ICCV*, 1988.
15. Hemant D. Tagare, Kentaro Toyama, and Jonathan G. Wang. A maximum-likelihood strategy for directing attention during visual search. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(5):490–500, 2001.
16. Anne Treisman. Features and objects in visual processing. *Scientific American*, 254(11):114–125, 1986.
17. Lance R. Williams and Karvel K. Thornber. A comparison of measures for detecting natural shapes in cluttered backgrounds. *Int. J. Computer Vision*, 34:81–96, 1999.