Mathematical Biosciences and Engineering

MBE, 20(5): 9364–9384.
DOI: 10.3934/mbe.2023411
Received: 20 December 2022
Revised: 24 January 2023
Accepted: 06 February 2023
Published: 17 March 2023

http://www.aimspress.com/journal/MBE

Opinion paper

Few-parameter learning for a hierarchical perceptual grouping system

Eckart Michaelsen*

Department of Object Recognition, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Gutleuthausstrasse 1, 76275, Ettlingen, Germany

* **Correspondence:** Email: exemichaelsen58@gmail.com; Tel: +4915111818476.

Abstract: Perceptual grouping along well-established Gestalt laws provides one set of traditional methods that provide a tiny set of meaningful parameters to be adjusted for each application field. More complex and challenging tasks require a hierarchical setting, where the results aggregated by a first grouping process are later subject to further processing on a larger scale and with more abstract objects. This can be several steps deep. An example from the domain of forestry provides insight into the search for suitable parameter settings providing sufficient performance for the machine-vision module to be of practical use within a larger robotic control setting in this application domain. This sets a stark contrast in comparison to the state-of-the-art deep-learning neural nets, where many millions of obscure parameters must be adjusted properly before the performance suffices. It is the opinion of the author that the huge freedom for possible settings in such a high-dimensional inscrutable parameter space poses an unnecessary risk. Moreover, few-parameter learning is getting along with less training material. Whereas the state-of-the-art networks require millions of images with expert labels, a single image can already provide good insight into the nature of the parameter domain of the Gestalt laws, and a domain expert labeling just a handful of salient contours in said image yields already a proper goal function, so that a well working sweet spot in the parameter domain can be found in a few steps. As compared to the state-of-the-art neural nets, a reduction of six orders of magnitude in the number of parameters results. Almost parameter-free statistical test methods can reduce the number of parameters to be trained further by one order of magnitude, but they are less flexible and currently lack the advantages of hierarchical feature processing.

Keywords: traditional machine vision; perceptual grouping; parameter adjustment; forestry; Gestalt laws

1. Introduction

Ubiquitous widespread conviction has it that we are witnessing huge progress in automation by use of recent methods from artificial intelligence, in particular from deep-learning artificial neural nets. Some even see dangers of self-abolishment of the human species. For the falsification of such claims, one example capability of human subjects suffices, where the proposed automata do not really compete with human subjects. Let us pick vision—most humans can see. So, what progress has been made in machine vision?

There has been a major breakthrough in the performance of object recognition by machines due to the introduction of deep-learning convolutional neural nets—Section 1.1 of this manuscript reviews such work. However, this came at the cost of introducing very many parameters, which are tuned utilizing very large labeled sample sets. At least since the advent of machine-learning theory, we have been aware that large numbers of such parameters are risky, jeopardizing robustness of the machines. More traditional machine-vision methods—Section 1.2 refers to some of them—need also some such parameters. These can be adjusted on much smaller sample sets properly. Generally, their recognition performance is inferior, so such work has not been pursued with the same efforts and computing power anymore.

Often, vision has been sub-divided into several steps building upon each other. An early step—often referred to as feature recognition—produces intermediate results, upon which a following object recognition or localization step operates. There is an old branch of vision research that postulates a specific intermediate step, i.e., perceptual grouping, between feature recognition and the final decision steps. Section 1.3 reviews some of the proposals that were made along these lines. These include a parameter-free statistical approach claiming to work without any learning or adjustment on any imagery.

The paper at hand proposes a perceptual grouping method that still retains a handful of parameters that allow adjustment on a given environment represented by a small sample set, which may contain as few as one image. It is the opinion of the author that such method should be preferred, even if it performs a little worse. It is much less risky, does not depend so much on the presence or absence of specific samples in the training data and consumes much less training effort.

1.1. State of the art in machine vision

There has been a major change in paradigm in this field with the advent of *convolutional* deep-learning nets. The basic model, how a scientific machine-vision paper would look today, was given by Yann LeCun et al. more than three decades ago [1]: The method is a deep-learning convolutional neural net trained by back-propagation of the classification error. The technical part describes the architecture in detail and the loss function. A publicly available data benchmark is used to demonstrate the superior classification performance (the Modified National Institute of Standards and Technology (MNIST) handwritten digits). Details are given on how the data are split into test and training parts and how the then-available machinery was run for days on the training. Today, most papers accepted for the main conferences and journals follow exactly this pattern. However, three decades ago, something like that used to be rejected—except for special neural-net conferences and journals. The vision scientists reviewing the work would have had doubts about the scientific value of such an end-to-end method, concentrating on the performance only and yielding no new insights or theory of the process of vision at all. It was not before the very impressive object-recognition performance of massive GPU-boosted deep-learning convolutional neural nets was demonstrated, by Alex Krizhevsky et al. about a decade ago [2], that the community accepted such machines as state-of-the-art. The object-recognition benchmarks used by then were quite academic—in contrast to the MNIST—distinguishing objects such as cats, dogs, airplanes, cars, etc. in big numbers of arbitrarily collected images. It has been explored how deep convolutional networks can become—i.e., how many layers can be trained still yielding good recognition performance and acceptable training efforts [3]. In fact, with rising depth, a consistent increase in recognition performance was reported (up to 19 layers). In the most recent years, object detection and localization has focused on YOLO architectures, where the acronym translates as "You Only Look Once," emphasizing that multiple objects can be detected and located from one image in video in real time on a standard computer [4]. Microsoft's COCO has become a standard data repository, and it contains very many annotated images of all sorts. Figure 2 in [4] shows the established standard architecture, consisting of the *backbone*, the *neck* and the *head*, respectively. The backbone is traversed bottom-up and made of several convolutional layers for feature detection, just like already proposed in [1], only bigger. The neck implements a corresponding deconvolution, traverses down the scales and takes inputs from corresponding layers of the backbone. The head implements the decisions and regressions needed. There, fully connected neural-net architectures are used. Much of the gain in performance has been achieved through what is now called *bag of freebies*, i.e., manipulations that do not alter the architecture at all and instead only make the training smarter. The most important are probably data-augmentation methods inventing non-measured, but plausible, imagery from the images included. This somehow replaces the investigations of invariances that used to be important before the CNN revolution.

1.2. Traditional approaches to machine vision

In contemporary papers, terms like *traditional methods*, or *hand-crafted features* in a rather pejorative tone, are used for the sparring partner bound to lose in the empirical comparison. I am most familiar with the traditional machine-vision methods proposed for remote sensing. A good overview of how elaborated those already were about three decades ago can be found in the Ascona Workshop of ETH Zürich [5]. The community was interested in extracting objects from imagery and analyzing their mutual relations. In particular, the road-extraction section consists of papers very related to the work at hand. Often, such methods wander around in the image, constructing the road network on the way. Such work has been continued—there are two follow-up Ascona Workshops that are also recommendable, in 1997 and 2001—and sometimes such work is still published recently [6]. Of course, the medical application has a similar tradition of object-centered machine-vision proposals, e.g., for the extraction of blood vessels. However, the communities are somewhat separated, and I am not familiar enough with the literature to recommend specific papers from the vast body published there in the last decades.

The classic traditional object recognition paper was contributed by Viola & Jones in 2001 [7]. Using face detection as an example task, it presents a cascade of weak classifiers, each checking one Haar feature. Such features are very efficiently computed using integral images. They exhibit whether certain rectangular sub-regions are brighter than others and can be specified with few parameters. Initial rejections can be made on one or two most important features, but for sufficient recognition performance, up to two hundred features are required successively, yielding a strong classifier committee. Training commenced on 24 × 24 pixel positive and negative sample tiles. Scale invariance was achieved by scaling the Haar features when shifting the operator over the input image during the operational phase. There is neither rotational invariance nor hierarchical composition of features. Lighting invariance is in question.

Histogram of Gradients (HoG) features have been a second important proposal in traditional object detection. Ludwig et al. used them in [8], with person detection being the example task. Such features are more sophisticated and require more parameters to be adjusted in the learning phase. They also use a classifier cascade, but they seem more aware of the risk that comes with a rising number of parameters, and they balance the number of parameters with the recognition performance in their cost function. No part-of hierarchy can be seen in this or similar proposals.

Rule-based production systems were proposed to implement visual knowledge application in a data-driven way [9], highlighting the part-of hierarchies in the content of complex aerial imagery of suburban terrain. Such attempts to utilize the conscious visual knowledge and reasoning of interpreters analyzing the images were popular two or three decades ago. However, it turned out that probably much of their expertise remains unconscious and hard to model. Thus, the much more primitive end-to-end approaches of today prevail.

1.3. Perceptual grouping and Gestalt laws in machine vision

This topic has its own community within psychology. The classic source is Wertheimer [10], now quite exactly a hundred years old. Still, scientific approaches to the topic are older, and often one reads the names of Mach and Helmholtz. When computers became mature enough to actually consider machine vision in practice, the main textbooks showed strong awareness of the Gestalt laws and the work and knowledge that had been accumulated in psychology for decades—see Marr [11] and Lowe [12]. Meanwhile, the interdisciplinary research between machine vision and psychology of perception was continued, leading, e.g., to the very recommendable textbook of Pizlo and his group [13]. Pizlo emphasizes that human perception turns out 3-dimensional without intermediate steps—and accordingly, related machine vision should do so also. I do not agree with him on that point, as becomes evident from the paper at hand. However, on all other points I see agreement. The most important late contribution to the field of Gestalt law grouping has been given by Agnès Desolneux and her group, presenting the a contrario estimation method in [14]. She claims that all free parameters of such processes can be eliminated—with only the test-significance threshold remaining. The paper at hand can be regarded as an argument to retain some of the parameters still, in order to have some flexibility adjusting the vision machine to the task and data given. We introduced hierarchical cooperation between the Gestalt laws recently in [15].

2. Materials and methods

2.1. A traditional method—perceptual grouping

Traditional machine-vision literature, such as [11,12,14], proposes cooperation and interdisciplinary mutual learning with psychology and biology, an idea well suited to the journal at hand. In particular, the utilization of Gestalt laws was popular. Figure 1 illustrates what is meant: The visual scene is regarded as something that contains objects (so called *primitives*), and such objects may come arranged in aggregates—for which the German word *Gestalt* is used. This highlights the nonaccidental geometric compositions distinguishing such organized patterns from simple chaotic clusters which may be accidental. In [14], A. Desolneux gives a justification (first being used a hundred and fifty years or more ago by Helmholtz):

"If the arrangement of the parts of a Gestalt is highly unlikely as product of random independent positioning of each part, a common underlying cause must be inferred, and thus their aggregation has more meaning than just the sum of the meaning of the parts alone."

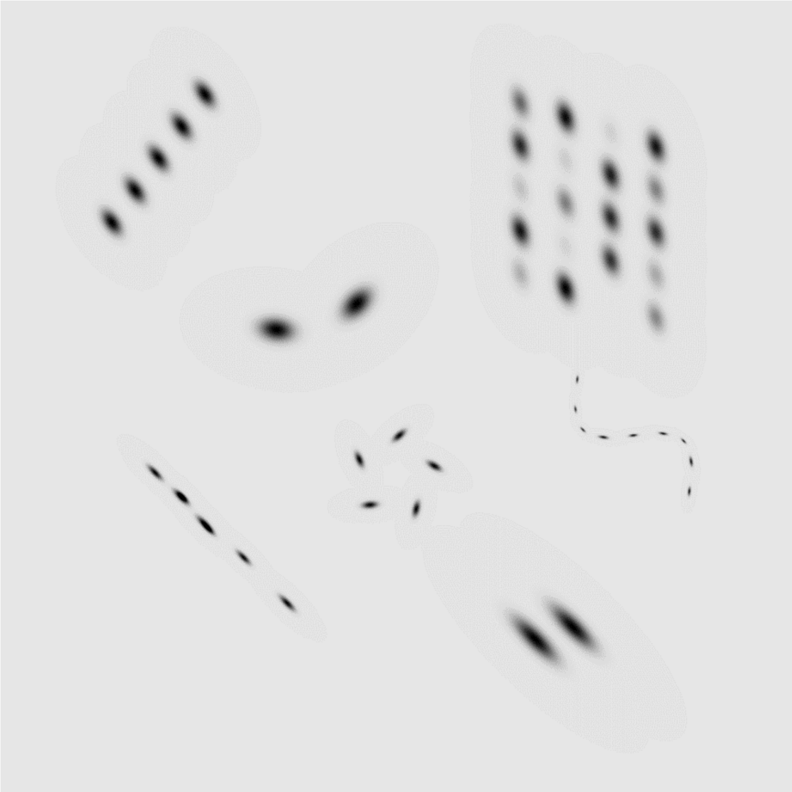

Figure 1. Seven laws for perceptual grouping.

The Gestalt literature shows no agreement on the number of corresponding grouping laws or their detailed geometric formulation. Thus, the seven laws presented in Figure 1 are a somewhat arbitrary choice following [15]. It is the result of practical application of Gestalt grouping to various applications of machine-vision. From top-left to bottom-right, there are the following:

1) Frieze law: A frieze or row enumerates its parts along a straight line in good continuation, i.e., equal spacing, using the same generator vector over and over again. This can have many applications, including the recognition of columns of military vehicles, of suburban row housing, of parking lots, etc. We encountered this law many times in machine vision for remote sensing. It is a strong law producing very salient Gestalts [16].

2) Reflection symmetry: The two parts of such mirror Gestalt must not be too far apart from each other. Their orientations and other features such as scale, eccentricities, etc. must be mapped on each other by a mirror mapping. This law can separate foreground from clutter in natural environments containing various living things, in particular, animals, humans, faces, etc. It is usually not very strong, but it is useful as an additional independent cue next to other evidence [16], and it has also been the main focus

of the symmetry-recognition competitions organized by Y. Liu and her group [17].
3) Lattices: For such Gestalt, the parts have two enumeration indices—row and column—and accordingly, there are two generator vectors. As such, one might see it as the hierarchical version of the frieze law: First, group the primitives into rows and subsequently, on a higher hierarchy, group the rows into a matrix. However, in practice, that has not worked well. Other methods of aggregation are more successful, and the whole lattice should be regarded as one Gestalt. Again, there are examples from remote sensing: in particular, façade recognition on synthetic aperture radar imagery [18].
4) Straight continuation: This is the usual standard Gestalt law, for which, e.g., A. Desolneux formulated her *a contrario* method first [14]. It will also be used as a main grouping principle in the paper at hand. Here, the primitives must be located on a common straight line. In contrast to the frieze law (#1), equal spacing is not an issue here. Instead, the orientation of the primitives must be consistent with the aggregate. This law does not require the other properties of the primitives—such as scale or eccentricity—to be mutually equal or even similar. Thus, it turns out quite distinct from law #1, while this difference is rarely explicit in the literature. Applications of this law are widespread. In particular, man-made objects tend to yield straight contours. Very long such Gestalten are very salient—such as the contours of runways in satellite images. However, the paper at hand will demonstrate that it can also be very useful in natural scenes.
5) Rotational-symmetry: This is rarely listed among the Gestalt laws in the literature, although such patterns, generated by a finite rotational mapping group (instead of the simple two-element reflection group of law #2), are very salient. It is a strong law but computationally not easy. Immediately, as a first application, flower recognition comes to mind. However, many religious and heraldic symbols also use the inherent psychological power of such Gestalt, and many man-made technical objects contain such patterns [15,17].
6) Curved continuation: This may be regarded as a generalization of straight continuation (law #4), the latter then being spurious in the list. However, a spline is a much more general curve in the plane as compared to the straight line. It poses a much weaker constraint, so that following Helmholtz, it may result from random positioning much more easily. Accordingly, it is less strong, and the search for such aggregates may branch wildly in the presence of clutter, yielding excessive and meaningless combinatoric enumerations. The prevention of such catastrophic failure necessitates complex coding and testing when utilizing this law for machine vision. Yet, it has very important applications, not only in remote-sensing (extracting roads and rivers [5]) but also in medicine and biology (extracting blood vessels, dendrites, plant roots, etc.). Many of the methods in use or proposed in these fields use Gestalt grouping law #6 without being aware of it.
7) Parallelism: Here, the evidence depends more than with the other laws on *proximity*. Grouping only pairs of primitives into one Gestalt (like in the reflection case #2), and the evidence provided is fairly weak. Yet, there are important applications—such as, again, recognition of roads, etc. Also, the samples provided in the paper at hand below are quite salient.

Of course, one might train deep-learning convolutional neural nets in order to implement these laws and recognize such Gestalten automatically. However, that is not necessary. The laws can easily be coded in an object-oriented way, defining the properties of such Gestalten and coding the functions that provide their saliency measures from the configurations and features of their parts. Corresponding search procedures can also be constructed, utilizing said saliencies for the smart administration of computational resources [19].

2.2. Parameters for Gestalts

The theory of a contrario tests [14], implementing the Helmholtz principle, eliminates all parameters from the perceptual-grouping process, with one exception: the statistic test-level. The user has to specify something like 95% or 98%—like in all statistic tests. However, this is done on the basis of certain assumptions on the distribution of the background clutter. This can, in practice, be violated frequently, so leaving some of the parameters in the Gestalt-construction functions adjustable turns out wise. Each law has its own set of parameters. Here, they are discussed in detail for laws #4 and #7.

2.2.1. Parameters for good continuation along straight lines

This law—#4 in the list above—combines four properties in one common saliency assessment: positioning along a line, angular consistency of the orientation of the parts with the orientation of the aggregate, sparsity of gaps in the covering of the line from start point to end point, inheritance of the saliency of the parts to the aggregate. So, the final saliency results as a function of a conjunctive evidence fusion—i.e., a product of values between zero and one. Such product is not directly accepted, because it tends to become smaller with every factor, and Gestalts with different numbers of saliency factors compete in the vision module for computational resources. Accordingly, I prefer a fourth root here for compensation in the final saliency:

$$a = \sqrt[4]{a_1 a_2 a_3 a_4} \tag{1}$$

where a is the overall saliency, and the a_i stand for the four sub-saliencies. The individual sub-laws work as follows:

1) Straight continuation: The straight-line orientation and location results from the first two moments of the locations of the parts. The eigenvector $\boldsymbol{v}_2$ corresponding to the larger eigenvalue sets the straight orientation. The eigenvector $\boldsymbol{v}_1$ corresponding to the smaller eigenvalue sets the equation for the residuals r_j for the individual parts. These give the sum of squared residuals R:

$$R = \frac{p_1^2}{s_a^2} \sum_{j=1}^{m} r_j^2 \qquad \text{and} \quad a_1 = \max(1 - R, 0). \tag{2}$$

However, there are necessary corrections here: The whole thing is scaled by the scale of the aggregate s_a, i.e., its length from start point to end point. That is also where the first parameter p_1 sits, controlling the process. The minus in the second equation in (2) provides low saliency a_1 for large cost R, and we must assure that no negative saliencies can be produced.

2) Angular consistency: Angular residuals are not that easy, because they do not sit in a vector space. Due to the rotational self-similarity of the short line-primitives, their orientations are given as $0 \le \rho_j < \pi$. For such angular residuals, the saliency is set as

$$a_2 = \prod_{j=1}^{m} (½ - ½\cos(2\rho_a - 2\rho_j))^{p_2}. \tag{3}$$

Here, ρ_a is the orientation of the aggregated Gestalt. Since the cosine evaluates between -1 and +1, the factors of this product will be limited by zero and one, the latter being reckoned when the orientations are completely consistent. The second control parameter p_2 acts as an exponent on each factor. This will make the cost curve more pointed or blunt.

3) Sparsity of gaps: Along the line from the start point to the end point, the aggregated line-segment is checked for overlap with the primitives. Where one or more are overlapping, everything will be fine, but gaps must cause costs in the saliency. Let s_g be the length of all gaps encountered, and then

$$a_3 = (1 - \frac{s_g}{s_a})^{p_3}. \tag{4}$$

This will be one for a gap-free Gestalt and zero for gaps-only, and it can be distorted in both directions by the control parameter p_3 as desired.

4) Saliency inheritance: Let the saliencies of the primitives be o_j and m be their number. Then,

$$a_4 = \left(\prod_{j=1}^{m} o_j\right)^{p_4/m} . \tag{5}$$

This can again be distorted in both directions by the control parameter p_4 as desired.

Of course, the saliency functions above are somewhat arbitrary and could be replaced with other choices—certain choices for cost functions and non-linearity in the deep-learning net science are similarly arbitrary, as well! However, here, along the lines of the a contrario theory [14], statistic justifications for them may well be found one day. The important point here is that we have a small number of very powerful and mutual independent controls in hand. Initially, all these parameters can be set to one, and the grouping will work somewhat. However, if the performance is not satisfactory, the user can start searching for better adjustments. Moreover, automatic learning procedures are also possible, if the automaton has input data which are annotated with the desired perceptual grouping outcome. In the case of hierarchical grouping, the inheritance (5) provides nested analytic functions. Thus, the chain rule applies, and learning adjustments can be propagated back through the hierarchy from the aggregates to the parts, all the way to the primitives—in analogy to backpropagation in deep neural nets.

2.2.2. Parameters for parallelism

This is the law #7 in the list above. Similar to #4, it combines four properties in one common saliency assessment. However, these are a little different: proximity, parallelism of the orientations of the two parts, overlap and inheritance of the saliency of the parts to the aggregate. So, the final saliency results as conjunctive evidence fusion using again function (1).

1) Proximity: Two roughly parallel lines give four locations: The heads $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ and the tails $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$. We construct the mid-head and the mid-tail. The corresponding connecting line from mid-tail to mid-head provides the major geometric attributes (position, orientation and scale s) of the aggregate. It also gives a normal form along the line $\boldsymbol{l}\cdot\boldsymbol{x}-c_a=0$ and perpendicular to the line $\boldsymbol{n}\cdot\boldsymbol{x}-c_n = 0$. For the width attribute w, the positions of the parts are entered as $\boldsymbol{x}$ into the perpendicular form. In order to be salient as parallel Gestalt, the ratio between width and scale w/s should be small but not zero (then, they are colinear, not parallel). We set

$$a_1 = e^{(2.0 - \frac{wp_1}{s} - \frac{s}{wp_1})} . \tag{6}$$

So, there we have the proximity-control parameter p_1.

2) Parallelism: This re-uses Eq (3), but only one factor is used, in which the two orientations of the parts are compared. The parameter p_2 must be chosen quite strict here.
3) Overlap: This is measured using the normal form $\boldsymbol{l}\cdot\boldsymbol{x}-\boldsymbol{c}_a = 0$ along the Gestalt and inserts the locations $\boldsymbol{h}_1$, $\boldsymbol{h}_2$, $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, respectively, as $\boldsymbol{x}$ in it. Also, this length must be set in ratio to the scale of the aggregate. The parameter p_3 acts like in (4).
4) Saliency inheritance: This is the same as above and uses Eq (5) with its parameter p_4—where the product has only two entries now.

2.2.3. Comparison of the numbers of parameters

As can be seen from the lists above, each Gestalt grouping law contributes about four parameters to be trained. Since there are seven such laws, the overall number of parameters of the proposed hierarchical grouping machinery may be estimated at about thirty. For comparison, the approximate numbers of parameters are given in Table 1 for four competing approaches to mid-level machine vision: several convolutional layers in the backbone of a state-of-the-art machine-vision neural net, a Haar feature cascade like in Viola & Jones [7], the proposed hierarchical Gestalt grouping and the statistical a contrario test approach, respectively.

Table 1. Estimated numbers of parameters (orders of magnitude).

Method	# of Parameters
a contrario test	10^0
hierarchical grouping	10^1
Haar cascade	10^4
YOLO-CNN	10^7

The numbers of parameters of the YOLO backbones are given in Table 1 of [4] as between 12.0 million and 27.6 million. For the a contrario test method, zero parameters are claimed [14]. However, there is a statistical significance to be specified in advance. This can be regarded as one parameter—though it is not trained from the data but instead chosen by the user at will. Also, the density of the irrelevant background clutter edge primitives must be known. In practice, this means that at least one sample image is needed, so that the density can be estimated (or bounded) as number of primitives divided by image area. In my view, this constitutes also a form of parameter learning. Some Haar features have six parameters, while some have ten. Basically, those specify the positions and sizes of the rectangles in which the intensities are considered—plus a sign setting darker or lighter. Several hundreds or thousands of such features sum up to ten thousand parameters or so. HoG features may be one order of magnitude higher in their demand of parameters to be trained.

2.3. Data

The work at hand intends to stay interdisciplinary between engineering and biology, the biological model being human vision. Humans are primates, and thus their vision developed in the forest. Therefore, we pick the example image displayed in Figure 2a) from the application of forestry.

Imagine it was your task to build the vision component for a smart—maybe autonomous—forestry robot assisting the rangers in their work in such environment. It may have legs to move freely in such scene and strong arms, maybe a saw, etc., to manipulate stems and logs. The task of the vision component, then, is the localization of said objects in the frame of reference of the robot, in order to avoid obstacles and give the motor control directions where to saw, grasp and lift. In this paper, the focus is set on training the parameters of the feature detection and grouping part of the vision system—roughly corresponding to the backbone in the neural net machines. The Gestalt vision system also needs a decision head, which can be a small shallow fully connected net working on features of the established Gestalts—such as colors, eccentricity, location, orientation, etc. It may as well be a support vector machine or some statistical classifier, and it may also be an inference machine as proposed by [9]. Such machine was examined in [19], where quite similar visual reasoning was used to implement landmark-based navigation for autonomous aerial vehicles.

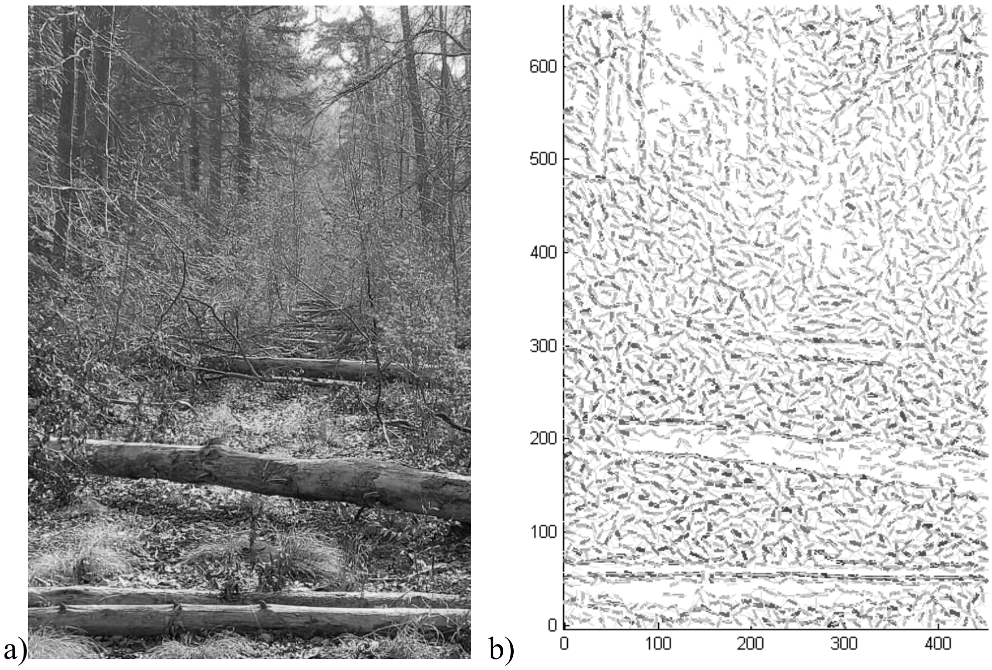

Figure 2. A scene in the forest: a) input intensity image; b) primitives extracted from it.

Obviously, possible application domains for perceptual grouping are very broad, almost ubiquitous. The reason why a forest example is chosen for this paper is twofold: On the one hand, it emphasizes the necessity and ability of such machinery to produce useful contributions in an open world, or “in the wild” as [17] calls it, as opposed to controlled man-made environments, such as in industrial applications. On the other hand, the human visual system—being a primate’s visual sense—has evolved in the forest, and the primary model for Gestalt-law research has been the human visual system for more than a hundred years. This does not imply that such method would not also be useful for completely controlled industrial applications, such as inspection of electronic circuits.

An intensity image as displayed in Figure 2a) might be a suitable input format for deep-learning convolutional neural nets, but it does not suit the needs of perceptual grouping as outlined above in Sections 2.1 and 2.2. To that end, a set of primitive objects must be extracted from the intensity matrix. There are many possible methods to do so. Here, a traditional standard method has been chosen: Canny edge filtering [20] with subsequent choice for local maxima. All parameters of these methods are set to the default values, as recommended by [20]. Of course, such parameters can also be subject to similar optimization procedures like the work presented below. The process results in a list of roughly five thousand short line-like primitives, attributed with location, orientation, length and saliency. They are depicted in Figure 2b) with their saliency given as grey tone. Such format fits into the needs for Gestalt grouping.

One might argue that only one image will not suffice as basis for the empirical evaluation of a machine vision proposal, stating that such a mistake was common practice in the early decades of the field and is generally not accepted anymore by the research-community, demanding empirical evaluation on a commonly available benchmark consisting of many images. However, as can be seen below in Sections 3.1 and 3.2, a lot can be learned from a single image and the behavior of the grouping processes on it. A single image already contains several hundred cases of Gestalt groups representative for the task at hand. Also, with only a handful of parameters to be adjusted, a hundred or so cases are already sufficient (as opposed to neural-net machine learning, where often millions of parameters are trained with only a few thousand image instances).

Often, a clear distinction between training data and test data is also demanded. That is justified for blind high-risk machines in the sense of the theory of machine learning [21], but in the case of robust perceptual grouping, it is less important. Nonetheless, a second image has been obtained after parametrizing the method and is used as a test datum in Section 3.3. It is from a different place and season of the year, has a different lighting situation, etc. Yet, it represents a similar application domain and scenery.

2.4. Searching the parameter domain

Parameter setting is a classical optimization task, for which the traditional mathematical methods apply that have proven successful. Most such optimization methods consist of an initialization method yielding a parameter setting to start from, followed by iterative stepwise improvement of said solution, until no further progress occurs.

For each parameter, a sophisticated a contrario reckoning can be done along the lines of the Desolneux theory [14]. Then, statistics of the density of clutter, the distribution of orientations, etc. are compiled from the sample imagery. Together with a test threshold—such as 95%—the a contrario estimations will yield an already quite useful parameter value.

In the absence of any smart parameter-initialization method, one can use random Monte Carlo type sampling of the parameter domain or smart non-random sampling methods, such as Latin Squares [22]. All the parameters presented in Section 2.3 above must remain positive, and the default setting is the value one. Their nature is logarithmic, like the frets on a guitar or the f-stops on the lens of an old camera. Accordingly, twenty random quadruples $(r_{1j}, r_{2j}, r_{3j}, r_{4j})$ were drawn uniformly from the range $-4.0 \leq r_{ij} \leq 4.0$ (j = 1...20 being the sample number). Then, for each such setting j, the parameters were set to

$$p_i = 2^{r_{ij}} \ . \tag{7}$$

Iterative improvement of parameter settings commences from good values to better values. Standard machine learning for artificial neural nets uses gradient decent with some arbitrary learning rate decaying during the process. In low-dimensional cases—such as the parameter optimization at hand here—a traditional Gauss/Newton-type optimization is often more appropriate, since it does not introduce a new meta-parameter (the learning rate) and converges faster. For the applicability of such methods, the goal function of the optimization must be differentiable.

3. Results

Results of the parameter adjustment on the training image are displayed in Sections 3.1 and 3.2. After the process of training the parameters was finished, a test was made on a new image, which was not at hand during development and tuning of the proposal. The corresponding results are given in Section 3.3.

3.1. Monte Carlo search on the training image

For law #4, the default parameter-setting $r_i = 0$ or $p_i = 1$ for all four parameters yields a rather unsatisfactory result, as is displayed in Figure 3a). Obviously, illusions run wild with such setting on densely cluttered input sets, such as the one displayed in Figure 2b). Such setting may, however, work on less crowded scenes, obtained by more sophisticated extraction methods—having better recognition performance than simple Canny edge filtering—or on less demanding scenery.

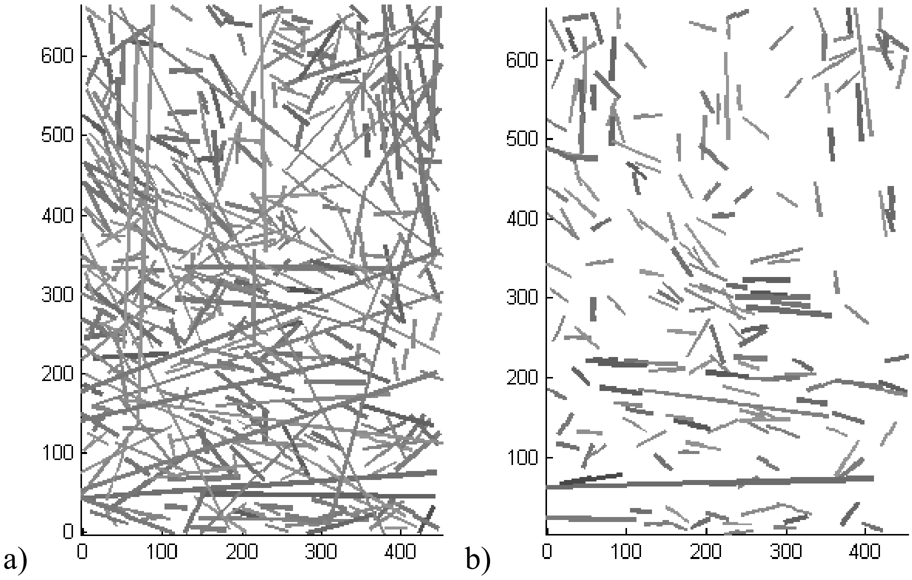

Figure 3. Results with law #4 for initial parameter setting: a) default setting *(1, 1, 1, 1)*; b) parameter setting *(1.46, 3.96, 5.28, 0.24)*.

Among the twenty results obtained by random Monte Carlo sampling, the one presented in Figure 3b) was most appealing. Some of the most salient trunks are outlined, while the number of obvious illusions is low. The corresponding setting is given in the figure caption. Apparently, as compared to the default setting, we have to be stricter with the deviations from good continuation, very much stricter with angular consistency, even more strict than that with the gaps but quite liberal with inheritance of saliency from the primitives to the aggregate.

3.2. Stepping around in the parameter domain for improvement on the training image

Equations (1)–(5) are analytic, so for the saliency attribute of the *k*-th prolonged Gestalt in Figure 3, we can reckon partial derivatives describing the dependency on the *i*-th parameter:

$$\frac{\partial a_k}{\partial p_i} \ . \tag{8}$$

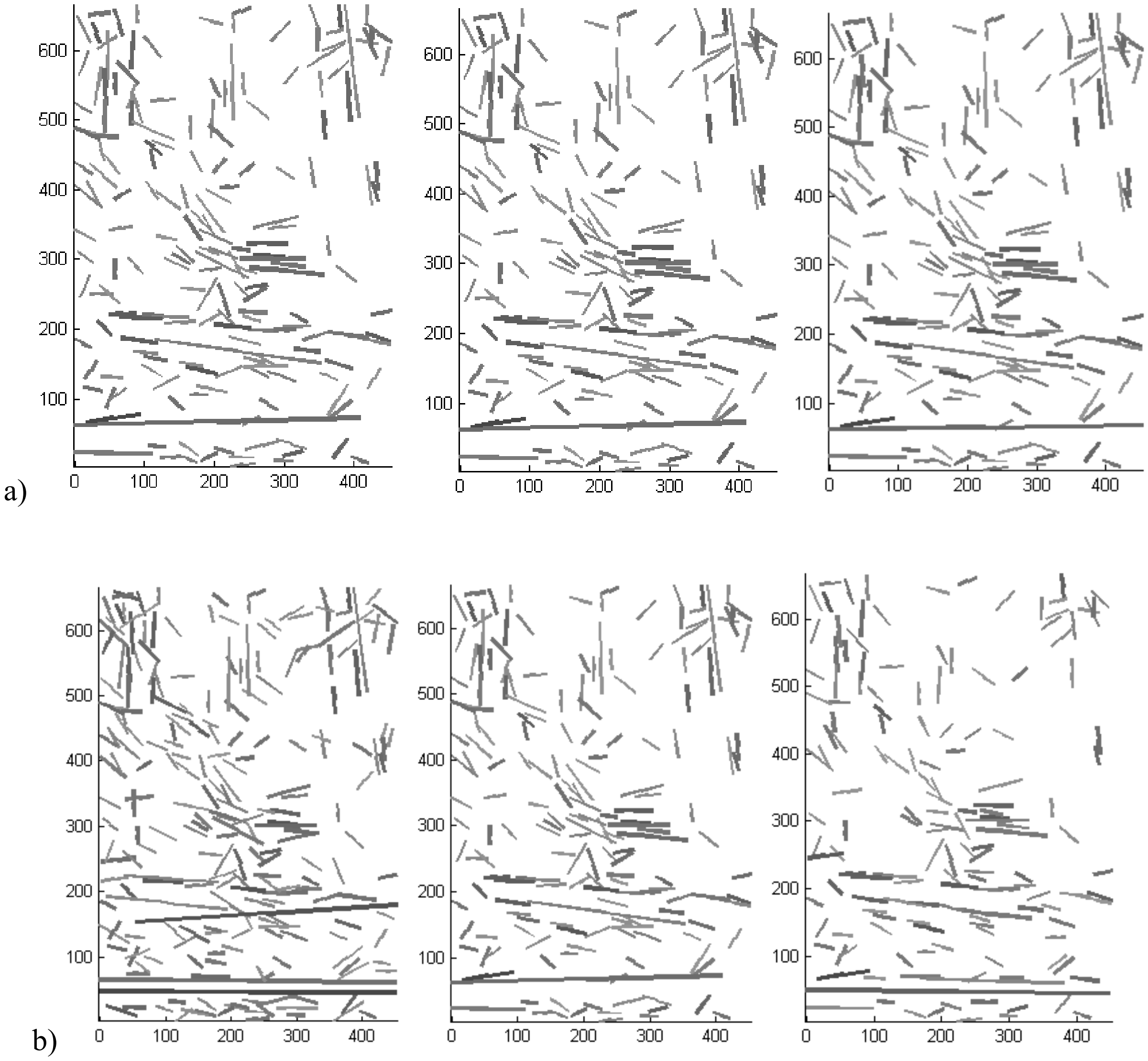

Figure 4. Steps in the parameter domain: a) $p_1 = 1.02, 1.46, 2.06$; b) $p_2 = 2.80, 3.96, 5.60$.

If we find a domain-expert labeling of the aggregated long line segments into the classes *illusory* and *consistent*, we could define a proper cost function as

$$c = \sum_{k \in illusory} a_k - \sum_{k \in consistent} a_k \tag{9}$$

yielding the perfect situation for Gauss/Newton minimization with the matrix (8) as Jacobian and using pseudo-inverse in each step. However, the set of Gestalten found is not fixed. Even small changes of parameters may lead to non-analytic events, such as new Gestalts appearing, existing Gestalts disappearing and aggregated Gestalts becoming substantially longer or shorter due to changes in their sets of parts. Figures 4 and 5 give an overview of what happens around the sweet spot shown in Figure 3b) when we disturb each parameter by a moderate adjustment, dividing or multiplying it by *√2*. Accordingly, the domain expert will be called up again in order to label new instances or to re-assess old ones that have become longer or shorter.

On the other hand, it may well happen that a particular change rate (8) at a particular current parameter setting may be comparably small. Figure 4a) shows the changes when the first parameter is altered considerably—nothing seems to change, with no new instances, no prolongation or shortening, no deletions and hardly any change in saliency (displayed again in grey tones). The change is not zero but very small. Accordingly, the inversion in a Gauss/Newton optimization would result in very large steps in the direction of this parameter. In contrast to that, at that sweet spot parameter setting, a change in the second parameter has salient consequences, as can be seen in Figure 4b).

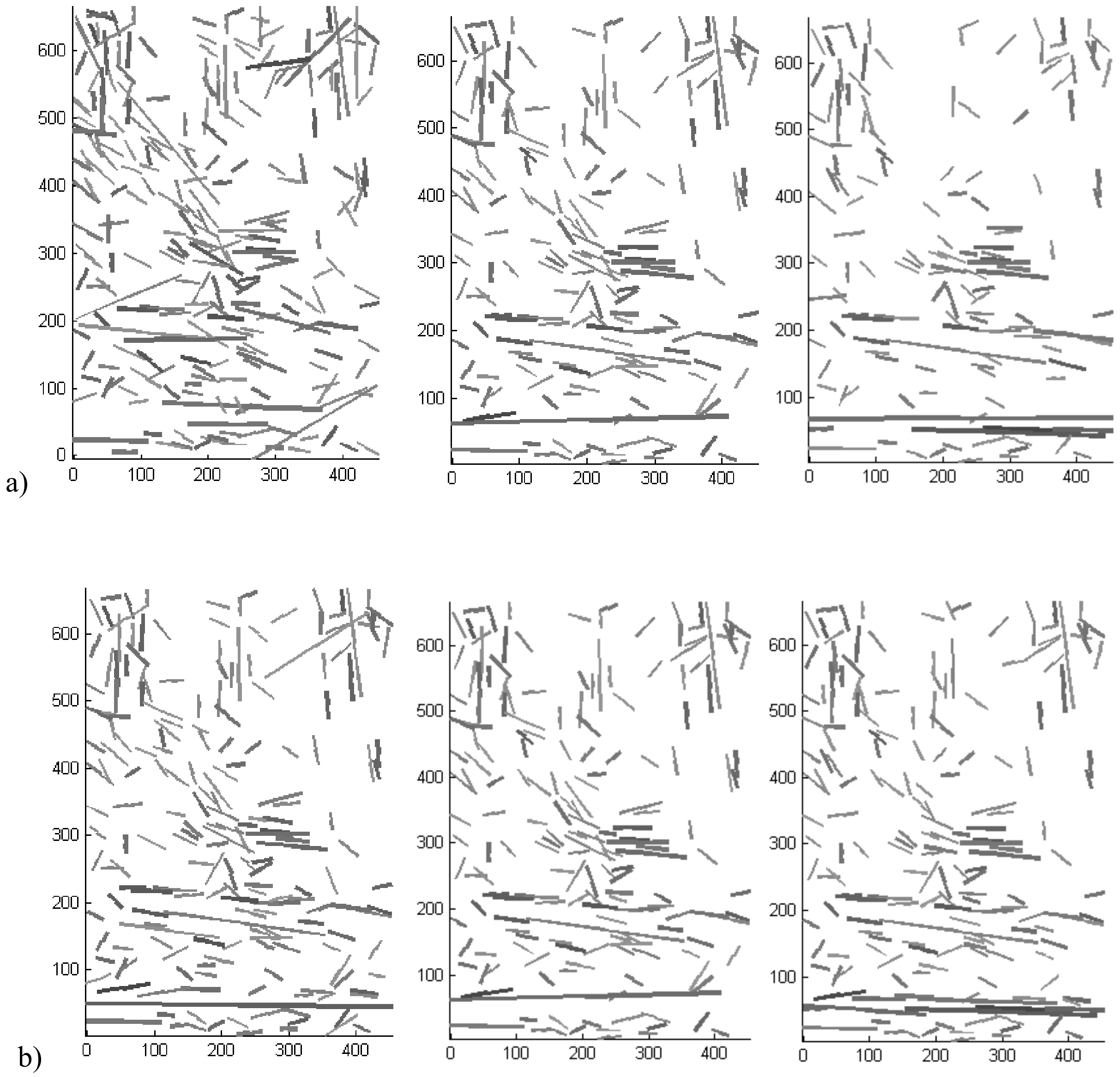

Figure 5. Steps in the parameter domain: a) $p_3 = 3.73, 5.28, 7.47$; b) $p_4 = 0.17, 0.24, 0.33$.

Probably the best result—with respect to merit/cost balance (9), i.e., balance between consistency and illusion—is found in Figure 5b) on the right: that is, setting *(1.46, 3.96, 5.28, 0.33)*. The full potential of perceptual grouping will be unleashed when the Gestalt laws are utilized in a hierarchical manner, so that such result is just regarded as an intermediate result on which further grouping should be performed—just like how artificial neural nets become much more potent when they are deepened from a single layer to deep-learning nets which use intermediate signals corresponding to rising scales and abstraction. The obvious next step in our case will be utilizing law #7, parallelism.

To this end—just like with a deep-learning-net—a new set of parameters must be adjusted, those discussed in Section 2.2.2. At this point, prior knowledge can give helpful advice. We know that proximity must be very close for parallelism to be salient—both from theoretical consideration (a contrario tests) and from practical reasoning (tree trunks are much narrower than long). We also suspect that orientation similarity must be strict for #7, along with high demands on overlap. Therefore, we do not start from parameter setting (1, 1, 1, 1) but from (10, 2, 2, 1) in our twenty Monte Carlo type search runs. Figure 6a) displays the default result, and b) shows the sweet spot found in this search run.

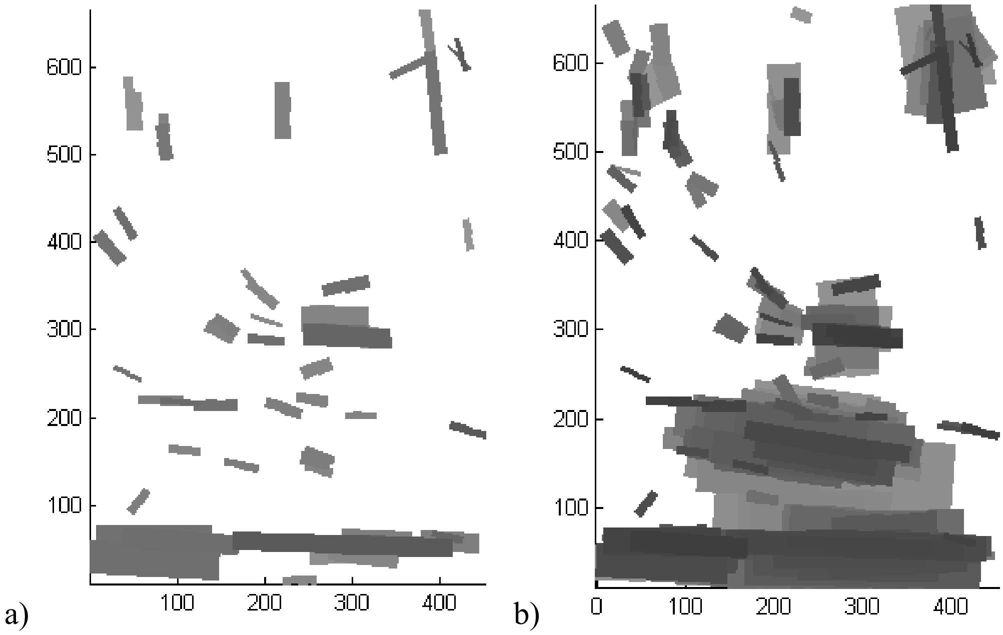

Figure 6. Results with law #7 for initial parameter setting: a) default setting *(10, 2, 2, 1)*; b) parameter setting *(9.22, 0.16, 0.31, 0.24)*.

We see that now large regions in the picture space are void. This is analogous to the situation with deep neural nets, where in deeper layers most activations tend to be zero. The full arithmetic power of the high-performance GPUs is usually wasted on summing up thousands of zeros in parallel billions of times in a second. In contrast, the object-centered method proposed here wastes no calculation and no storage on void regions. With rising depth of grouping, the objects will always cluster in isolated salient islands in the 2D-image domain (augmented by more features such as orientation, scale, width, etc., which make the true void regions much more dominating in the overall volume than they seem in 2D only).

This is a strong argument against the use of the a contrario method, which assumes uniformly distributed objects (like in Figure 2b)). Actually, the list of about a hundred Gestalts of the parallel-pair-of-long-contours type does not really correspond directly to objects in the scene. It corresponds to possible consistent combinations of features that may result from a common object in a scene. Accordingly, each object in the scene may well be found in such list in multiple combinatoric versions. Instead of wasting computational resources on enumerating zeros in void areas, the approach at hand tends to waste computational resources on enumerating sub-sets of features in a combinatorial way—but only in regions where some salient objects are present.

Therefore, the objects on such a higher level are sorted according to their saliency (displayed as grey tone in the figures), and among the many possibilities in one location, orientation and scale, only the best are kept for further processing. This is indicated in the figures by displaying the more salient (darker) stripes laid over the less salient.

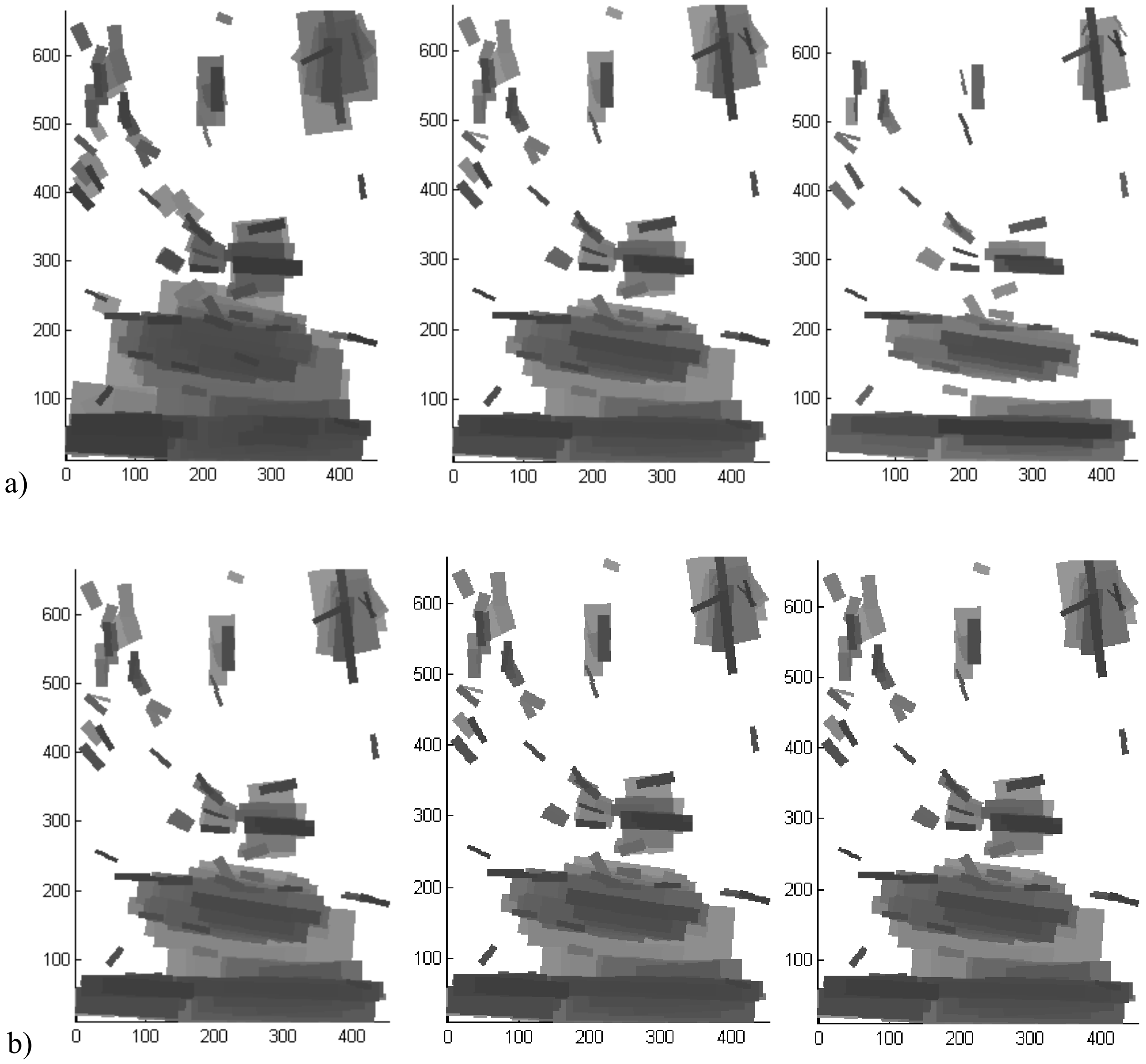

Figure 7. Steps in the parameter domain: a) $p_1 = 6.52, 9.22, 13.04$; b) $p_2 = 0.11, 0.16, 0.23$.

In Figure 6a) the salient almost horizontal trunk in the center at vertical coordinate 200 is missing, which would be an important object for the task at hand. Therefore, the decision must be in favor for some sweet spot setting like presented in Figure 6b), where all important objects are present.

In Figure 7, the first two parameters—for proximity and parallelism—are explored around the sweet spot found by Monte Carlo search and displayed in Figure 6b). Obviously, parameter p_1 should be adjusted a little stricter in order to suppress unnecessary combinatorial enumerations and illusions, while still listing the important objects. The result at this setting is quite robust with variations of parameter p_2. In Figure 8 parameters p_3 and p_4 are varied around said setting, displaying again great robustness against such disturbances. Actually, nothing substantial happens at all when applying drastic changes such as doubling or halving these parameters around this setting on these data.

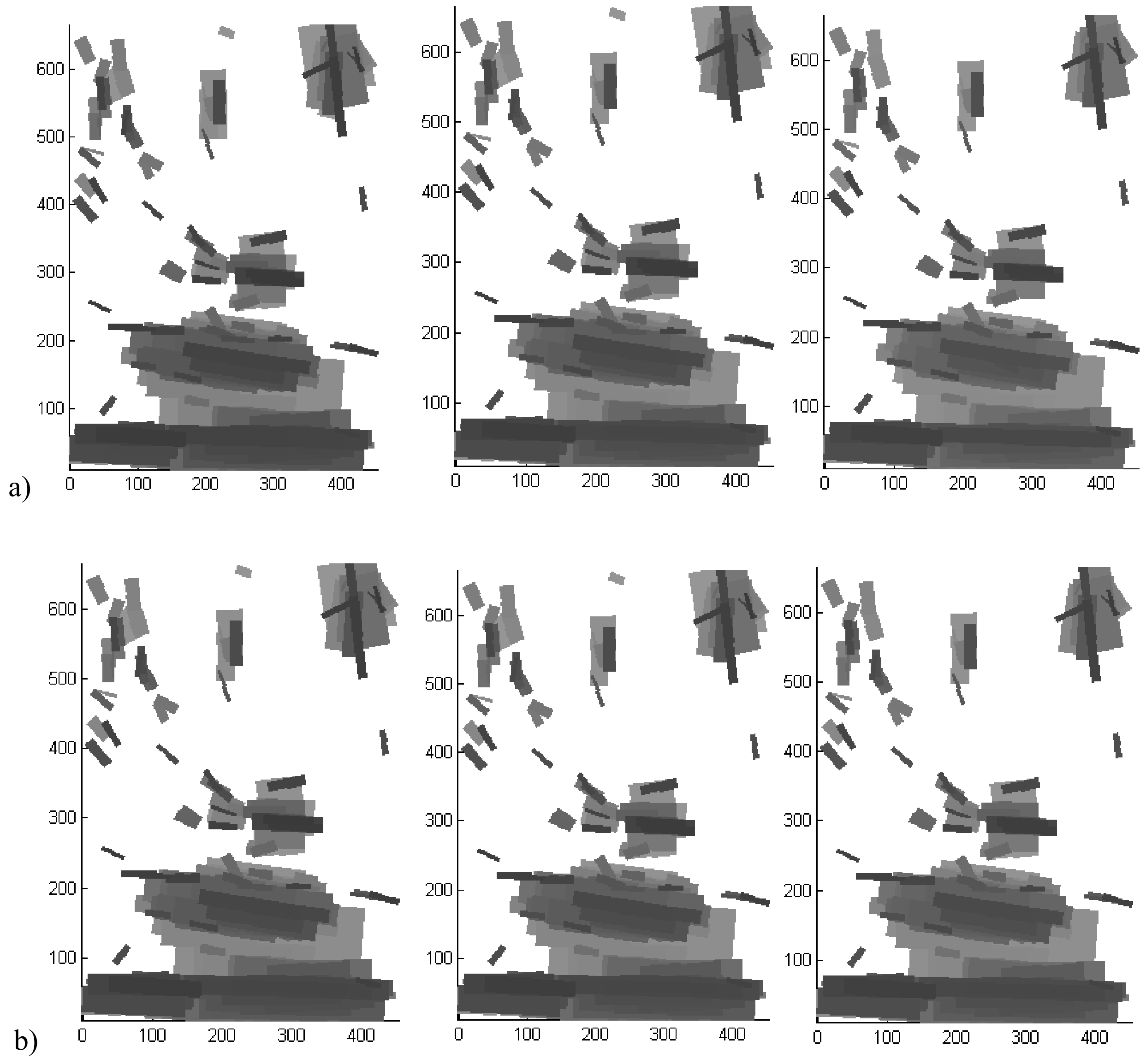

Figure 8. Steps in the parameter domain: a) $p_3 = 0{,}22\ 0.31, 0.44$; b) $p_4 = 0.17, 0.24, 0.34.$

The results can obviously be a useful input for obstacle avoidance and object manipulation for a potential forestry robot. Of course, the depth of the objects is still missing and required for such tasks. Additional sensors, such as an active laser rangefinder directed at the objects, may be a solution to that. Stereo vision can also be an option, which would be a separate module in the vison machine, cooperating with the perceptual grouping as well as with the motor control. The point here is that perceptual grouping can extract the relevant objects even from such a cluttered outdoor scene. It makes machines see.

3.3. Test image

In order to assess the dependency of the achieved parameter setting on the training image and thus the robustness of the method in an open world, a new image was obtained subsequently. It is displayed in Figure 9a).

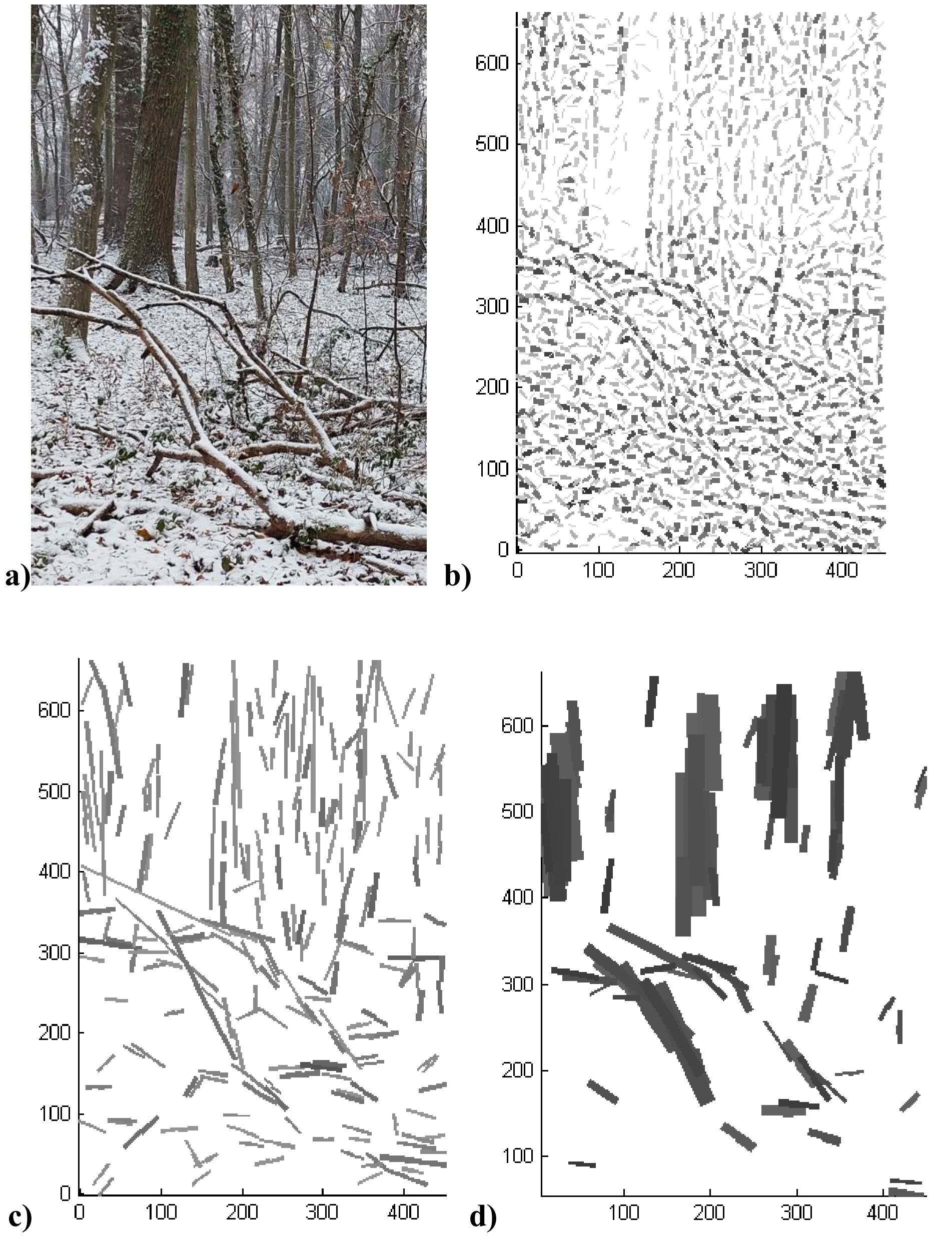

Figure 9. The test image: a) input image; b) Canny edge primitives; c) grouping law #4 parameters as found in Section 3.2; d) subsequent grouping law #7 parameters as found in 3.3.

This image was shot in a different season of the year and at a different location—though also in a similar forest. There are different lighting conditions, etc. Thus, it comes as no surprise that the resulting stripe Gestalts depicted in Figure 9d) are a bit less satisfying than those in the 7a) leftmost frame, which were obtained with the same parameter setting. However, the branches in the foreground are instantiated successfully. Closer objects are more important in robotics. Thus, this result, too, may still well contribute valuable input for obstacle avoidance or object manipulation.

4. Discussion

The grouping result on the test image depicted in Figure 9d) points to an important issue of further consideration: The salient foreground branches are instantiated in a still fragmented way with one object in the scene corresponding to several stripe Gestalts given by the vision system. It would probably be desirable to achieve a more one-to-one relation between what is seen and what is there. That calls for another grouping step, this time using law #6, curved continuation, adding one step more to the hierarchy. Were we aware of the desirability of such augmentation in advance, before seeing this sample image and the corresponding result?

I would say yes. The sequence in which the laws are applied and the depth of reasoning should not be fixed in advance by the architecture of the machine. The whole approach should be implemented in a data-driven fashion with the priorities given by the assessment attributes of the Gestalts, just like in a production system used for visual reasoning ([9] or [19]).

Table 1 suggests that the YOLO backbone, the Haar cascade, the Gestalt laws and the a contrario test may be compared to each other. The reader may object, however, that those proposals were meant for different ends: almost ubiquitous object detection and localization, specific object detection (faces) and detection of salient configurations. Also, the performance in terms of avoiding recognition failures is different. However, all these proposals claim to contribute to one major endeavor: making machines see.

Thus, we can also compare them with respect to other aspects, such as, e.g., their similarities to the archetype: human vision. CNNs inherit their name from biological neural nets, but there is no evidence for anything like backpropagation going on in animals. The architecture resembles that of actual brains only very superficially, and we do not really know to what degree and on what laws synapses are trained. The preference for horizontal rectangles in systems like YOLO is surely a consequence of our contemporary square computer world. It all results from the view of the visual world as seen through square pixel grids. This is even more evident for the feature cascades. Nothing can be further away from the human visual perception than integral images. This is totally different for Gestalt grouping—whether explicitly parametrized or in the form of statistical tests. Neither pixel grids nor image margins are of any interest in this world—only aggregates and parts and mutual geometric relations. Every structure used in such approaches can be straightforwardly double-checked by psychological experiments using computer graphics and random generators.

The success of deep-learning neural nets teaches that, often, the detailed structure of the formalism reckoning the scores inside a recognition machine does not really matter. Any sigmoid or even piecewise linear activation function will work. What matters, though, is the goal function (or loss, as connectionists call it), along with fidelity to the ground truth and whether the learning data set is large enough, balanced and really representative for the issue at hand.

Vapnik and Chervonenkis presented the only valid theory of machine learning in 1974 [21], to which, to my knowledge, nothing substantial has been added since then. It teaches that, on the one hand, such a machine should not be too simple—containing too few parameters to be adjusted in the learning process. In that case, it cannot solve more challenging tasks satisfactorily. On the other hand—the deeper insight of said theory—too many parameters in the machine will be very risky. The size of the training set—and with it the computational efforts for the optimization—limits the number of parameters in the machine. Over-adjustable recognition machines will be prone to sudden catastrophic failure when confronted with real-world situations.

The proposals popular for machine vision today contain dozens of millions of parameters to be adjusted. The only items they incorporate from the century-old science of vision are convolution (implementing a bit of shift invariance) and scale hierarchy (to a limited and fixed degree). Other proposals for machine vision, which were most popular two decades ago, were even more remote from said vision science as it is practiced in psychology, albeit tuning only dozens of thousands of parameters. The paper at hand recalls some traditional methods based on that interdisciplinary vison science as an almost forgotten alternative with even less parameters to be adjusted. In fact, only a handful of parameters suffices to solve very complex vison tasks, and these parameters have quite intuitive meanings well intelligible even to application experts who are not familiar with vision science itself.

Moreover, this kind of traditional approach leaves the domain of raster matrices early and continues working on segmented objects which are seen in part-of hierarchies. This view of the world is much closer to human intuition and understanding than the hierarchies of scales in rigid raster matrices given by state-of-the-art machine vision. It also suits the needs for cooperation with robotic and control mechanisms better, where the resulting lists of perceived objects can be further processed, e.g., in obstacle avoidance or object grasping and handling.

5. Conclusions

Machine vision can get along with far less parameters to be adjusted by machine learning. Training efforts can be reduced by orders of magnitude. Dependency on huge image repositories accessible only to a few major companies and of unknown content and composition can be avoided, while still retaining the advantages of machine learning for adaption to a specific practical vision task. The awkward bounding-box and rigid vector-format outputs of the state-of-the-art architectures can be replaced by flexible lists of the seen objects with their measured properties and features. All that can be achieved by using more traditional methods combined with hierarchical perceptual grouping instead of the convolutional backbone of the now popular machinery. It is the opinion of the author that such advantages justify accepting minor performance disadvantages.

Conflict of interest

The author declares there is no conflict of interest.

References

1. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, et al., Backpropagation applied to handwritten zip code recognition, *Neural Comput.*, **1** (1989), 541–551. https://doi.org/10.1162/neco.1989.1.4.541
2. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. https://doi.org/10.1145/3065386
3. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.
4. A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, YOLOv4: Optimal speed and accuracy of object detection, preprint, arXiv:2004.1093.
5. A. Gruen, O. Kuebler, P. Agouris, *Automatic Extraction of Man-made Objects from Aerial and Space Images*, Birkhäuser Verlag, Basel, 1995.
6. J. Dai, R. Ma, L. Gong, Z. Shen, J. Wu, A model-driven-to-sample-driven method for rural road extraction, *Remote Sens.*, **13** (2021), 1417. https://doi.org/10.3390/rs13081417

7. P. Viola, M. Jone, Rapid object detection using a boosted cascade of simple features, in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *CVPR 2001*, **1** (2001). https://doi.org/10.1109/CVPR.2001.990517
8. O. Ludwig, U. Nunes, B. Ribeiro, C. Premebida, Improving the generalization capacity of cascade classifiers, *IEEE Trans. Cybern.*, **46** (2013), 2135–2146. https://doi.org/10.1109/TCYB.2013.2240678
9. T. Matsuyama, V. S. S. Hwang, *SIGMA A Knowledge-based Aerial Image Understanding System*, Plenum Press, 1990.
10. M. Wertheimer, Untersuchungen zur Lehre der Gestalt. II, *Psychologische Forschung*, **4** (1923), 301–250. https://doi.org/10.1007/BF00410640
11. D. Marr, *Vision*, Freeman & Co., San Francisco, 1982.
12. D. G. Lowe, *Perceptual Organization and Visual Recognition*, Kluwer, Boston, 1986.
13. Z. Pizlo, Y. Li, T. Sawada, R. M. Steinman, *Making a Machine that Sees like Us*, Oxford University Press, USA, 2014.
14. A. Desolneux, L. Moisan, J. M. Morel, *From Gestalt Theory to Image Analysis: A Probabilistic Approach*, Springer, Berlin, 2008.
15. E. Michaelsen, J. Meidow, *Hierarchical Perceptual Grouping for Object Recognition*, Springer, Berlin, 2019.
16. E. Michaelsen, Self-organizing maps and Gestalt organization as components of an advanced system for remotely sensed data: An example with thermal hyper-spectra, *Pattern Recogn. Lett.*, **83** (2016), 169–177. https://doi.org/10.1016/j.patrec.2016.06.004
17. C. Funk, S. Lee, M. R. Oswald, S. Tsogkas, W. Shen, A. Cohen, et al., 2017 ICCV challenge: detecting symmetry in the wild, in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (2017), 1692–1701.
18. E. Michaelsen, A. Schunert, U. Soergel, Utilizing phase for the grouping of PS in urban high-resolution in-SAR-images, in *2011 Joint Urban Remote Sensing Event*, IEEE, (2011), 189–192. https://doi.org/10.1109/JURSE.2011.5764752
19. E. Michaelsen, J. Meidow, Stochastic reasoning for structural pattern recognition: an example from image-based UAV navigation, *Pattern Recogn.*, **47** (2014), 2732–2744. https://doi.org/10.1016/j.patcog.2014.02.009
20. Amarjot, *Canny Edge Detector*, Available from: https://www.mathworks.com/matlabcentral/fileexchange/40737-canny-edge-detector.
21. V. N. Vapnik, A. Ya, *Chervonenkis: Theory of Pattern Recognition*, Nauka, 1974.
22. L. Gao, *Latin Squares in Experimental Design*, Michigan State University, 2005.