Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Physiology - Paris 97 (2003) 311–324
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

Computational gestalts and perception thresholds

Agnès Desolneux *, Lionel Moisan, Jean-Michel Morel

CMLA, ENS Cachan, 61 av. du président Wilson, 94235 Cachan cedex, France

Abstract

In 1923, Max Wertheimer proposed a research programme and method in visual perception. He conjectured the existence of a small set of geometric grouping laws governing the perceptual synthesis of phenomenal objects, or "gestalt" from the atomic retina input. In this paper, we review this set of geometric grouping laws, using the works of Metzger, Kanizsa and their schools. In continuation, we explain why the Gestalt theory research programme can be translated into a Computer Vision programme. This translation is not straightforward, since Gestalt theory never addressed two fundamental matters: image sampling and image information measurements. Using these advances, we shall show that gestalt grouping laws can be translated into *quantitative laws* allowing the automatic computation of gestalts in digital images.

From the psychophysical viewpoint, a main issue is raised: the computer vision gestalt detection methods deliver *predictable perception thresholds*. Thus, we are set in a position where we can build artificial images and check whether some kind of agreement can be found between the computationally predicted thresholds and the psychophysical ones. We describe and discuss two preliminary sets of experiments, where we compared the gestalt detection performance of several subjects with the predictable detection curve. In our opinion, the results of this experimental comparison support the idea of a much more systematic interaction between computational predictions in Computer Vision and psychophysical experiments.

Keywords: Gestalt theory; Visual perception; Detection threshold; Computer vision

1. Introduction, Gestalt theory

The geometric Gestalt theory started in 1921 with the Max Wertheimer founding paper [33]. The Gestalt Bible *Gesetze des Sehens* by Wolfgang Metzger gave in its last edition in 1975 a broad overview of the extension and the results of the research. At about the same date, Computer Vision was an emerging new discipline, at the crossing point between Artificial Intelligence and Robotics. The foundation of signal sampling theory by Claude Shannon [30] was actually already 20 years old, but computers were able to deal with images with some efficiency only at the beginning of the seventies. Two things are noticeable:

- Computer Vision used very little and almost nothing of the Gestalt theory results: the founding book of David Marr [24] involves much more neurophysiology than phenomenology. Also, its programme and the robotics programme [13] based Computer Vision on binocular stereo vision. This was in total contradiction to, or ignorance of, the results explained at length in Metzger's chapters on *Tiefensehen*. Indeed, these chapters demonstrate that binocular stereo vision is a *parent pauvre* in human volume perception. The only bright exception is Attneave's attempt [3] to give a quantitative sampling theory adapted to shape perception. His paper had some influence in shape analysis algorithms.
- Conversely, the Shannon's information theory does not seem to have influenced at all gestalt research, as far as we can judge from Kanizsa's and Metzger's books.

Both facts are surprising. Indeed, both disciplines have attempted to answer the following question: how to arrive at global percepts from the local, atomic information contained in an image? Both groups were aware that the retina information is atomic and that synthesis laws are necessary to build up visual objects, let them be gestalts [1] or shapes (see Fig. 1).

* Corresponding author.

[1] We shall write gestalt and treat it as an English word when we talk about gestalts as groups and maintain the uppercase in Gestalt theory.

0928-4257/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2003.09.006

Fig. 1. Illustration of gestalt principles. From left to right and top to bottom: color constancy + proximity, similarity of shape and similarity of texture; continuity of direction; closure (of a curve); convexity; parallelism; amodal completion (a disk seen behind the square); color constancy; good continuation (dots building a curve); closure (of a curve made of dots); modal completion and pregnancy: by pregnancy of squares, we tend to see a square in the last figure and its sides are also seen in a modal way (subjective contour). Notice that the similarity of texture in the first and last figure. Most of the figures involve constant width and similarity of size of the objects.

In this paper, we shall summarize a computational theory which permits to find automatically gestalts in digital images. This theory essentially predicts *perception thresholds* which can be computed on every image and give a usually clear cut decision between what is seeable as a geometric structure (gestalt) in the image and what is not. Those thresholds are computable thanks to the discrete nature of images.

A physiological and psychophysical question arises: are the same kind of thresholds in game in our equally discrete retina images? We made two simple experiments to check whether a match between theoretically predicted thresholds and psychophysical ones can be made. In the first experiment, squares were displayed to subjects in more or less noisy images and the subjects were asked to decide whether they saw one or not. In the second experiment, aligned segments were displayed in a background of random segments. In both cases, subjects were asked to decide when they saw a square, or an alignment. Our theory predicted an a priori shape for the "pop-out" curve, depending in each case upon two parameters. In the first experiment, the pop-out depends upon relative contrast and size of the square; in the second one, the pop-out depends upon the number of aligned segments and the density of background distractors. Somewhat to our surprise, we found a significant agreement between the gestalt pop-out threshold curves predicted and the observed ones. This opens, to our opinion, a new way to develop a *quantitative Gestalt theory* and psychophysical devices. Indeed, this possibility of predicting *a priori* perception quantitative thresholds and checking them experimentally can be expanded to all gestalt qualities. The future computational theories about the perception of simple and complex gestalts should systematically entail psychophysical (and maybe neurophysiological) devices to decide between computational theories of human perception.

Our plan is as follows. We start in the next section with an account of Gestalt theory, centered however on the initial 1923 Wertheimer programme. We also address some slight changes in terminology necessitated by the computational developments, in particular the notion of *partial gestalt*. In continuation, we shall recall the very basics of Shannon theory and explain why, together with a probabilistic principle which we call Helmholtz principle, they permit a computation of gestalts in digital images. Section 4 is devoted to the description of two experimental devices permitting to check whether the psychophysical pop-out curves, depending on two parameters (roughly, size of the object and amount of noise) are in agreement with the computationally predicted ones.

2. Classification of gestalt laws from a computational viewpoint

According to Gestalt theory, "grouping" is the main process in our visual perception (see [15–18,21,25,33,34]). Whenever points (or previously formed visual objects) have one or several characteristics in common, they get grouped and form a new, larger visual object, a *gestalt*. The list given by Gaetano Kanisza in *Grammatica del Vedere* page 45 [15] is *vicinanza, somiglianza, continuita di direzione, chiusura, pregnanza, esperienza passata*, that is: vicinity, similarity, continuity of direction, closure, pregnancy, former experience. This is almost exactly the list stated in the founding paper of Wertheimer [33].

The above grouping laws have not at all the same status and will need some ordering. All of them, however, belong, according to Kanizsa to the so called *processo primario* (primary process), opposed to a more cognitive secondary process. Also, it may of course be asked *why and how* this list of geometric qualities has emerged in the course of biological evolution. Brunswick and Kamiya [5] were among the first to suggest that the gestalt grouping laws were directly related to the geometric statistics of the natural world. Since then, several works have addressed from different points of views these statistics and the building elements which should be conceptually considered in perception theory, and/or numerically used in Computer Vision [1,4,11,22,27].

2.1. The starting points or atomic data: Shannon theory and the discrete nature of images

Before proceeding to a classification of gestalt laws from the computational viewpoint, we must discuss the computational nature of images, let them be digital or biological. In order to define an image in the simplest possible way, we just need to fix a point of focus. Assume all photons converging towards this focus are intercepted by a surface which has been divided into regular cells, usually squares or hexagons. Each cell counts its number of photons hits during a fixed exposure time. This count gives a grey level image, that is, a rectangular, (roughly circular in biological vision) array of grey level values on a grid. In the case of digital images, CCD matrices give regular grids made of squares. In the biological case, the retina is divided into hexagonal cells with growing sizes from the fovea. Thus, in all cases, a digital or biological image contains a *finite* number of values on a grid. Shannon [30] made explicit the mathematical conditions under which, from this matrix of values, a continuous image can be reconstructed. By Shannon's theory, we can compute the grey level at *all* points, and not only the points of the grid. When we zoom in the interpolated image, however, it looks more and more blurry: the amount of information in a digital image is *bounded* and the *resolution* of the image is finite. The points of the grid together with their grey level values are called *pixels*, an abbreviation for *pict*ure *el*ements.

This raises a hope: aren't the pixels the atoms from which gestalt grouping can start? On the other hand, we see a paradox: *how can we infer sure events as lines, circles, squares, whatsoever gestalt from discrete data?* If the image is blurry, all of these structures cannot be inferred as completely sure ; their exact location must in fact remain uncertain. As we shall see, this is crucial: all basic geometric information in the image has an easy-to-guess *accuracy*. This accuracy parameter will be crucial in the computations of the next sections. Since all local information about a function u at a point (x,y) boils down to its Taylor expansion, we can assume that the atomic information from which gestalts can be built up are:

- the value $u(x,y)$ of the grey level at each point (x,y) of the image plane. Since the function u is blurry, this value is valid at points close to (x,y),
- the gradient of u at (x,y), the vector

$$Du(x,y) = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right)(x,y),$$

- the *orientation* at (x,y),

$$\mathrm{Orient}(x) = \frac{1}{\|Du(x,y)\|}\left(-\frac{\partial u}{\partial y}, \frac{\partial u}{\partial x}\right)(x,y).$$

This vector is visually intuitive, since it is tangent to the boundaries one can see in an image.

The orientation is invariant when the image contrast changes (which means robustness to illumination conditions). Attneave's and Julesz [14] constantly refer to it for shape recognition and texture discrimination theory.

2.2. From atomic data to partial gestalts

Before going on with the discussion, we must fix a bit the terminology to adapt it to the intermediate scopes Computer Vision has to deal with. There are three meanings for gestalt:

- the first one is the final global group seen, whose constitution it is the aim of Gestalt theory to explain;
- the second one is related to the geometric qualities involved, like good continuation or convexity; in that case, we shall talk about the "good continuation gestalt", which actually means "good continuation law" leading to detect smooth curves;
- finally we have partial gestalts, namely not the final gestalts, but the result of application of one of the grouping laws to the image.

This is better explained by an example. A simple object like a square, whose boundary has been drawn in black with a pencil on a white sheet, will be perceived by connectedness (the boundary is a black line), by constant width (of the stroke), convexity and closedness (of the black pencil stroke), parallelism (between opposite sides), orthogonality (between adjacent sides), finally equidistance (of both pairs of opposite sides). Thus, we must distinguish between what we shall call *global* gestalt and *partial* gestalt. The square is a global gestalt, but it is the result of a long list of concurring geometric qualities, leading to parts of it endowed with some gestalt quality. Such parts we shall call *partial gestalts*.

To be more precise, let us make the following definitions.

Definition 1. We call partial gestalt law any grouping process driven by a single grouping law.

As we shall see, the partial gestalt grouping laws are computationally treatable as they proceed from *local*, *atomic observations*. Also, all of them are *recursive*: they allow the grouping of already partially constituted groups into larger groups, and so on.

One can summarize the efforts of Computer Vision as a way to compute the (very diverse in nature) partial gestalts. To take an instance, the snakes method [19] attempts to capture the closed smooth curves, a combination of the "closure" and "good continuation" gestalts. In the same way, have been proposed in

Computer Vision: alignment detectors (e.g. Hough transforms), edge detectors, angle detectors, shape recognition methods (the "similarity of shape" gestalt), and texture segmenters, that is, a general way to group points according to common features which are, again, nothing but partial gestalts. The good continuation principle has been extensively addressed in Computer Vision, first in [26], more recently in [28] and still more recently in [12]. A recent example of computer vision paper implementing "good continuation", understood a "constant curvature", is [35].

3. Computing partial gestalts in digital images

3.1. General detection principles

In this section, we quickly review anterior work where we proposed a general principle for computing any partial gestalt and applied it to several gestalt qualities. This principle will be applied to several new examples of computable partial gestalts.

Helmholtz Principle (see Fig. 2): In [7], we outlined a computational method to decide whether a given partial gestalt (computed by any segmentation or grouping method) is reliable or not. We treated the detection of alignments, as one of the most basic gestalts (see [33]). As we shall recall, our method gives *absolute thresholds*, depending only on the image size, permitting to decide when a peak in the Hough transform is significant or not.

A geometrically meaningful event is an event that, according to probabilistic estimates, should not happen in an image and therefore makes sense. This informal definition immediately raises an objection: if we do probabilistic estimates in an image, this means that we have an a priori model. We are therefore loosing any generality in the approach, unless the probabilistic model could be proven to be "the right one" for the image under consideration. In fact, our proposition has been to do statistical estimates without any image model. Instead, we applied a general perception principle that we called Helmholtz principle. This principle yields computational grouping thresholds associated

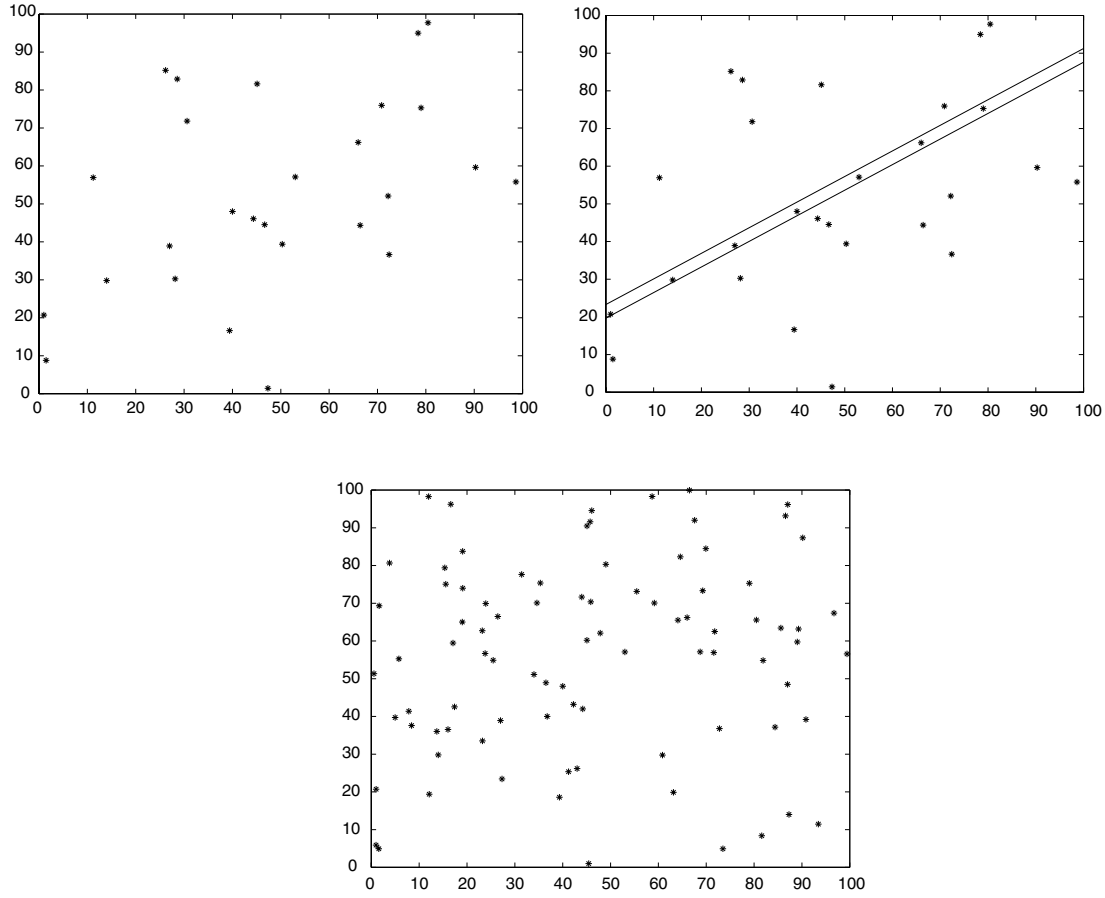

Fig. 2. An illustration of Helmholtz principle: non-casual alignments are automatically detected by Helmholtz principle as a large deviation from randomness. Top left: 20 uniformly randomly distributed dots, and 7 aligned added. Top right: this meaningful (and seeable) alignment is detected as a large deviation. Bottom: same alignment added to 80 random dots. The alignment is no more meaningful (and no more seeable). In order to be meaningful, it would need to contain at least 11 points.

with each gestalt quality. It can be stated in the following generic way. Assume that objects $O_1, O_2, \ldots, O_n$ are present in an image. Assume that k of them, say $O_1, \ldots, O_k$, have a common feature, say, same color, same orientation, etc. We are then facing the dilemma: is this common feature happening by chance or is it significant and enough to group $O_1, \ldots, O_k$? In order to answer this question, we make the following mental experiment: we assume a priori that the considered quality has been randomly and uniformly distributed on all objects, i.e. $O_1, \ldots, O_n$. Notice that this quality may also be spatial (like position, orientation). Then we (mentally) assume that the observed position of objects in the image is a random realization of this uniform process. We finally ask the question: is the observed repartition probable or not? If not, this proves *a contrario* that a grouping process (a gestalt) is at stake, since, according to Helmholtz principle, qualities of independent objects should be equally distributed. Mathematically, this can be formalized by

Definition 2 (ε-*meaningful event* [7]). We say that an event of type "such configuration of points has such property" is ε-meaningful if the expectation of the number of occurrences of this event is less than ε under the uniform random assumption.

As an example of generic computation we can do with this definition, let us assume that the probability that a given object O_i has the considered quality is equal to p. Then, under the independence assumption, the probability that at least k objects out of the observed n have this quality is

$$B(p,n,k) = \sum_{i=k}^{n} \binom{n}{k} p^i (1-p)^{n-i},$$

i.e. the tail of the binomial distribution. In order to get an upper bound for the number of false alarms, i.e. the expectation of the number of geometric events happening by pure chance, we can simply multiply the above probability by the number of tests we perform on the image. Let us call N_T the number of tests. Then in most cases we shall consider in the next subsections, a considered event will be defined as ε-meaningful if

$$N_T B(p,n,k) \leqslant \varepsilon.$$

We call in the following the left hand member of this inequality the "number of false alarms" (NFA).

When $\varepsilon \leqslant 1$, we talk about meaningful events. This seems to contradict the necessary notion of a parameterless theory. Now, it does not, since the ε-dependency of meaningfulness is low (it is in fact a $\log \varepsilon$-dependency [7]). The probability that a meaningful event is observed by accident will be very small. In such a case, our perception is liable to see the event, no matter whether it is "true" or not. We refer to [7] for a complete discussion of this definition.

The general method we have just outlined can be viewed as a systematization of Stewart's "MINPRAN" method [32]. The method was presented as a new paradigm, but was applied only to the 3D alignment problem.

The method we explain here has probably been proposed several times in Computer Vision (e.g. in the early Lowe work [23]), but, to the best of our knowledge, not systematically developed.

3.2. Meaningful alignments

Let us start by our first example, the detection of straight lines in an image. From the psychophysical and statistical viewpoint, alignment is considered as one of the main partial gestalts, already discussed at length in the founding Wertheimer paper [33]. A recent study on long range interactions in image perception due to alignments can be found in [31].

Since images are blurry, noisy and aliased, we cannot hope for a strong accuracy in direction measurement at each pixel, and we shall, without need for many explanations, fix the accuracy of a measured gradient direction at a point equal to a factor $p\pi$ radians. This means that a casual alignment of a direction with a prefixed one happens with probability p. In practice, $p = \frac{1}{16}$ is the best we can hope from digital images (and is even optimistic for aliased images). We consider the following event: "on a discrete segment of the image, joining two pixel centers, and with length l counted in points at Nyquist distance, at least k points have the same direction as the segment with precision p". The direction at each point is computed as the direction of the gradient rotated by $\frac{\pi}{2}$.

Definition 3 ([7]). Consider a segment S of length l containing k aligned points. We call number of false alarms of S,

$$\mathrm{NFA}(S) = N^4 \sum_{j=k}^{l} \binom{l}{j} p^j (1-p)^{l-j}.$$

We say that S is ε-meaningful if $\mathrm{NFA}(S) \leqslant \varepsilon$.

An example of alignment detection is given on Fig. 3 with $\varepsilon = 1$.

If on a straight line we have found a very meaningful segment S, then by enlarging slightly or reducing slightly S, we still find a meaningful segment. This means that meaningfulness cannot be a univoque criterion for detection, unless we can point out the "best meaningful" explanation of what is observed as meaningful. This is done by the following definition, which can be adapted as well to meaningful boundaries [8], meaningful edges [8], meaningful modes in a histogram [9] and clusters.

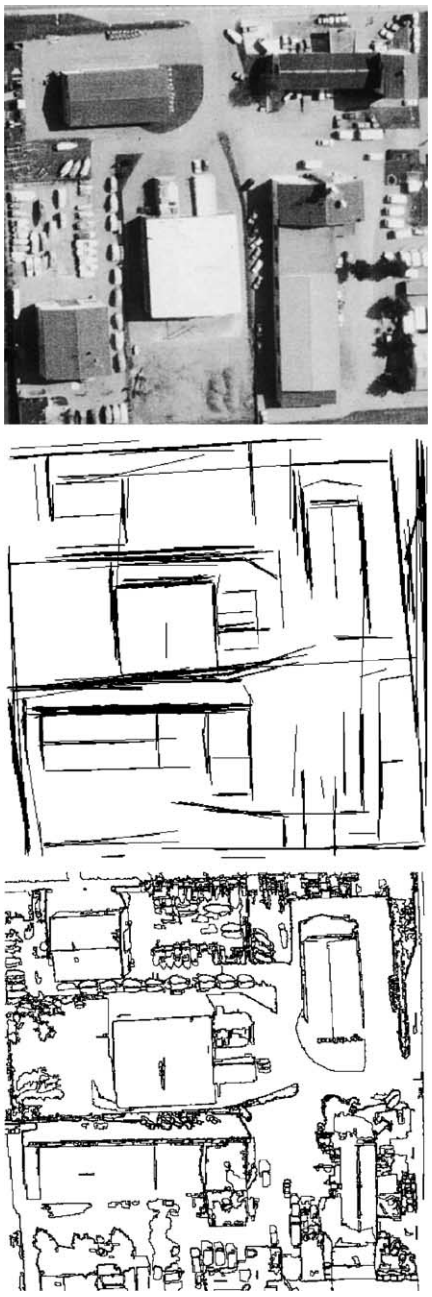

Fig. 3. Two partial gestalts, alignments and boundaries. Top: original aerial view (source: INRIA), middle: maximal meaningful alignments, bottom: maximal meaningful boundaries.

Definition 4 ([9]). We say that an ε-meaningful geometric structure A is maximal meaningful if

- it does not contain a strictly more meaningful structure: $\forall B \subset A$, $\mathrm{NFA}(B) \geqslant \mathrm{NFA}(A)$.
- it is not contained in a more meaningful structure: $\forall B \supset A$, $B \neq A$, $\mathrm{NFA}(B) > \mathrm{NFA}(A)$.

It is proved in [9] that maximal structures cannot overlap, which is one of the main theoretical outcomes validating the above definitions.

3.3. Edge and boundary detectors

We shall now review briefly our second example of partial gestalt, the boundaries: a classical example in Computer Vision! The scope here is to point out the existence of a parameterless boundary detector deduced from the Helmholtz principle and again to compare it with the other definitions of partial gestalts. A detailed treatment is given in [8]. Let u be a discrete image of size $N \times N$. We consider the level lines at quantized levels $\lambda_1, \ldots, \lambda_k$ (see [6]).

Level lines are simply curves along which $u(x,y)$ is constant. Since the image is only defined on a discrete grid, an interpolation close to Shannon's interpolation must be defined. In the experiments below we chose a less accurate interpolation, the bilinear interpolation, for a sake of simplicity.

Let L be a level line of the image u. We denote by l its length counted in independent points, and by x_1, $x_2, \ldots, x_l$ the l considered points of L. For a point $x \in L$, the contrast at x is defined by

$$c(x) = |\nabla u|(x), \tag{1}$$

where ∇u is computed by a standard finite difference scheme on a 2×2 neighborhood [7]. For $\mu > 0$, we consider the event: for all $1 \leqslant i \leqslant l$, $c(x_i) \geqslant \mu$, i.e. each point of L has a contrast larger than μ. From now on, all computations are performed in the Helmholtz framework: we make all computations as though the contrast observations at x_i were mutually independent. Since the l points are independent, the probability of this event is

$$\mathrm{Prob}[c(x_1) \geqslant \mu] \cdot \mathrm{Prob}[c(x_2) \geqslant \mu] \cdots \mathrm{Prob}[c(x_l) \geqslant \mu] = H(\mu)^l, \tag{2}$$

where $H(\mu)$ is the probability for a point on any level line to have a contrast larger than μ. An important question here is the choice of $H(\mu)$.

We decided to compute $H(\mu)$ from the empirical distribution given by the image itself, that is

$$H(\mu) = \frac{1}{M} \#\{x, |\nabla u|(x) \geqslant \mu\}, \tag{3}$$

where M is the number of pixels of the image where $\nabla u \neq 0$. In order to define a meaningful event, we have to compute the expectation of the number of occurrences of this event in the observed image. Thus, we first define the number of false alarms.

Definition 5 ([8]). Let L be a level line with length l, counted in independent points. Let μ be the minimal contrast of the points $x_1, \ldots, x_l$ of L. The number of false alarms of this event is defined by

$$\mathrm{NFA}(L) = N_{\mathrm{ll}} \cdot [H(\mu)]^l, \tag{4}$$

where N_{ll} is the number of level lines in the image.

Notice that the number N_{ll} of level lines is provided by the image itself. We now define ε-meaningful level

lines. The definition is analogous to the definition of ε-meaningful alignments.

Definition 6 (ε-*meaningful boundary* [8]). A level line L with length l and minimal contrast μ is ε-meaningful if

$$\mathrm{NFA}(L) \leqslant \varepsilon. \tag{5}$$

The above definition involves two variables: the length l of the level line, and its minimal contrast μ. The number of false alarms of an event measures the "meaningfulness" of this event: the smaller it is, the more meaningful the event is. An example of boundary detection is given on Fig. 3.

3.4. A general similarity grouping principle: histogram modes

As we mentioned in the introduction, the main gestaltic grouping principle is this: points or objects having one or several features in common are being grouped because they have this feature in common. We shall consider here only grouping by a single feature and we shall see that this single-feature grouping already yields relevant results. We face here a general problem: assume k objects $O_1, \ldots, O_k$, among a longer list $O_1, \ldots, O_n$, have some quality Q in common. Assume that this quality is actually measured as a real number. Then our decision of whether the grouping of $O_1, \ldots, O_k$ is relevant must be based on the fact that the values $Q(O_1), \ldots, Q(O_k)$ make a *meaningful mode* of the histogram of $Q(O_1), \ldots, Q(O_n)$. Thus, the single quality grouping is led back to the question of an automatic, parameterless, histogram mode detector. Of course, this mode detector depends upon the kind of feature under consideration. We shall consider two paradigmatic cases, namely the case of orientations, where the histogram can be assumed by Helmholtz principle to be flat, and the case of the objects sizes (areas) where the null assumption is that the size histogram is decreasing (see Fig. 4).

3.4.1. The similarity gestalt: objects grouped by orientation, or grey level

In the sequel, we quantize the possible orientations and grey levels in the usual way and we assume that the M values of orientation (or grey level) are independent and uniformly distributed on $\{1, 2, \ldots, L\}$. Consider an interval $[a, b] \subset [1, L]$ and let $k(a, b)$ denote the number of objects with gestalt value in $[a, b]$. We define $p(a, b) = (b - a + 1)/L$ as the a priori probability that the gestalt value of an object falls in $[a, b]$. With the same generic argument as in Section 1, we have

Definition 7 ([9]). An interval $[a, b]$ is ε-meaningful if

$$\mathrm{NFA}([a, b]) = N_i \cdot B(p(a, b), M, k(a, b)) \leqslant \varepsilon,$$

where N_i is the number of considered intervals ($Ni \simeq L(L + 1)/2$). An interval $[a, b]$ is said maximal meaningful if it is meaningful and if it does not contain, or is not contained in, a more meaningful interval (see Definition 4).

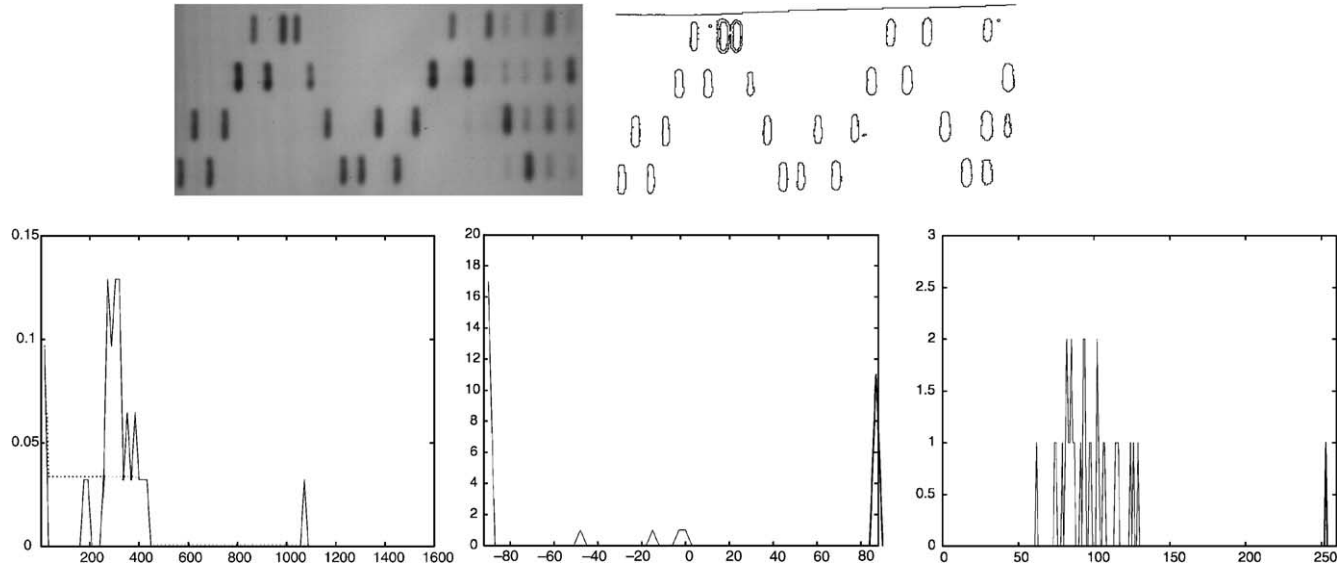

Fig. 4. Collaboration of gestalts: the objects tend to be grouped similarly by several different partial gestalts. First row: original DNA image (left) and its maximal meaningful boundaries (right). Second row, left: histogram of areas of the meaningful blobs. There is a unique maximal mode (256–416). The outliers are the double blob, the white background region and the three tiny blobs. Second row, middle: histogram of orientations of the meaningful blobs (computed as the principal axis of each blob). There is a single maximal meaningful mode (interval). This mode is the interval 85–95. It contains 28 objects out of 32. The outliers are the white background region and three tiny spots. Second row, right: histogram of the mean grey levels inside each block. There is a single maximal mode containing 30 objects out of 32, in the grey level interval 74–130. The outliers are the background white region and the darkest spot.

It can be proved in the same way as for alignments that maximal meaningful intervals do not intersect. Thus, we get an operational definition of meaningful modes as disjoint subintervals of $[1, L]$.

3.4.2. Size of objects

The preceding arguments are easily adapted to Helmholtz type assumptions on non-uniform histograms. A very generic way to group objects in an image is their similarity of size. This similarity lets groups perceptually pop out. Now, it would be a total nonsense to assume any uniform law on the objects sizes. There are several powerful arguments in favor of a statistical decreasing law for size. These arguments derive from perspective laws, or from the occlusion dead leaves model, or directly from statistical observations of natural images [2]. Our Helmholtz qualitative hypothesis is then: the prior distribution of the size of objects is *decreasing*.

Definition 8 ([10]). An interval $[a, b]$ is ε-meaningful (for the decreasing assumption) if

$$\mathrm{NFA}([a,b]) = N_i \cdot \max_{p \in \mathscr{D}} B(p(a,b), M, k(a,b)) \leqslant \varepsilon,$$

where $\mathscr{D}$ is the set of decreasing probability distributions on $\{1, 2, \ldots, L\}$, and $p(a,b) = \sum_{i=a}^{b} p_i$.

3.5. Alignments of objects (good continuation again)

The gestalt we now consider is not the same as the alignment gestalt considered at the beginning of Section 2, where the aligned points had their own orientation. Here, we consider the case of objects whose barycenters are aligned. Assume that we observe M objects of a certain kind in an image. Our null hypothesis for the application of Helmholtz principle will be that the M barycenters (x_i, y_i) are independent and uniformly distributed on a domain Ω. A meaningful alignment of points must be a meaningful peak in the Hough Transform (see [20,29] for a very similar approach). Now, the accuracy matter must be addressed. Points will be supposed to be aligned if they all fall into a strip thin enough, in sufficient number. Let S be a strip of width a. Let $p(S)$ denote the prior probability for a point to fall in S, and let $k(S)$ denote the number of points (among the M) which are in S. The following definition permits to compute all strips where a meaningful alignment is observed (see Fig. 5).

Definition 9 ([10]). A strip S is ε-meaningful if

$$\mathrm{NFA}(S) = N_s \cdot B(p(S), M, k(S)) \leqslant \varepsilon,$$

where N_s is the number of considered strips (one has $N_s \simeq 2\pi(R/a)^2$, where R is the half-diameter of Ω and a the minimal width of a strip).

In practice, we sample all possible strip widths in a logarithmic scale (about 8 widths) and we sample accordingly the angles between tested strips in order to get a good covering of all directions. Thus, the number of strips N_s only depends on the size of the image and this yields a parameterless detection method.

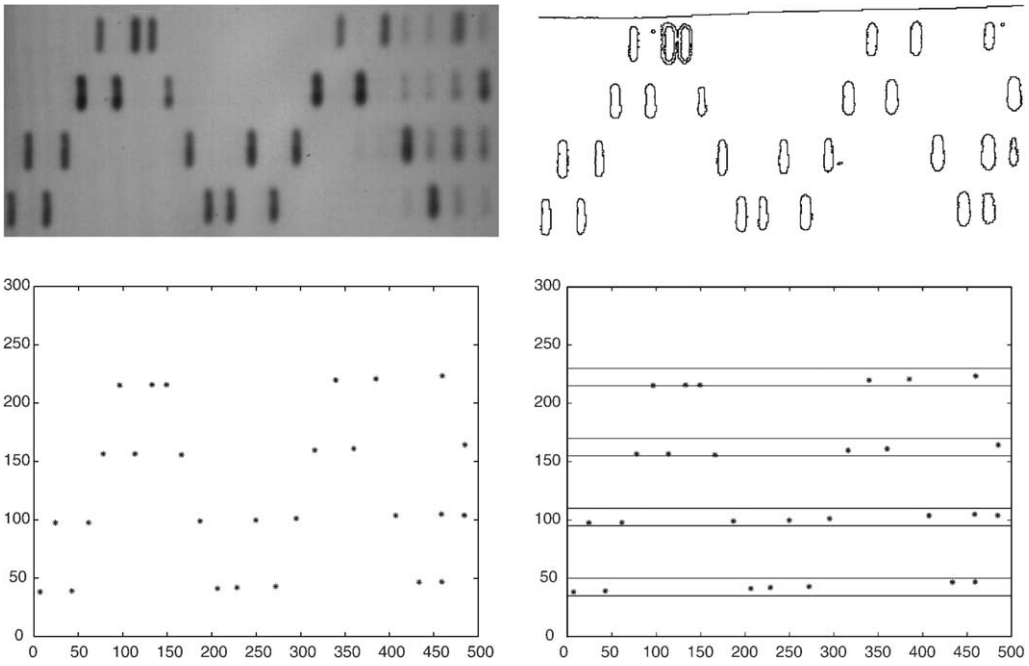

Fig. 5. Gestalt grouping principles at work for building an "order 3" gestalt (alignment of blobs of the same size). First row: original DNA image (left) and its maximal meaningful boundaries (right). Second row: left, barycenters of all meaningful regions whose area is inside the only maximal meaningful mode of the region areas histogram; right, meaningful alignments of these points.

4. Psycho-visual experiments

In this section, we describe two psycho-visual experiments that we built up in order to compare our approach of computational gestalts to the human gestalt perception. The question is: to what extent can one sustain that our perception is driven by Helmholtz principle? The experiments were designed to test the ability of a subject to detect the presence of a certain kind of structure in an image. For each experiment, we measured the subject response in function of two parameters, namely the size of the gestalt to be detected and the amount of noise, and we compared the average perception threshold to the computational one predicted by Helmholtz principle.

4.1. Detection of squares

4.1.1. Protocol

A sequence of images is presented to the subject. Each image appears on the screen during 1.5 s. For each image the subject has to answer to the question: "Voyez-vous un carré dans l'image?". [2] If his answer is yes, he has to press a key during the 1.5 s image display time. Between each image, a blank image is displayed during approximately 0.5 s to avoid interferences between successive images.

Each image is made of $N \times N$ black or white pixels. A square is randomly generated in the following way: its side length l and its position are chosen randomly (and uniformly), and a number $\bar{d} \in [0,1]$ (average density) is also randomly chosen. A random image is then generated as follows: each pixel is black with probability p and white with probability $1-p$ (Bernoulli process). All pixel values are chosen independently, except that we take $p = \bar{d}$ in the square domain and $p = 1/2$ outside. Such an image is shown on Fig. 6.

For each image, we record the answer of the subject ("can you see a square, yes or no"), along with the square side length (l) and its relative density δ, defined by $\delta = |d - 1/2|$, where d is the ratio of white pixels contained in the square (note that the expectation of d is $\bar{d}$, but the two numbers may differ a little). Each image is then represented as a point in the (δ, l) plane.

4.1.2. Prediction

If our perception is based on Helmholtz principle, we are supposed to detect the square against the hypothesis that the image is a white noise, in the present case a Bernouilli noise with parameter 1/2. The number of false alarms associated to a square with side length l and relative density δ is then

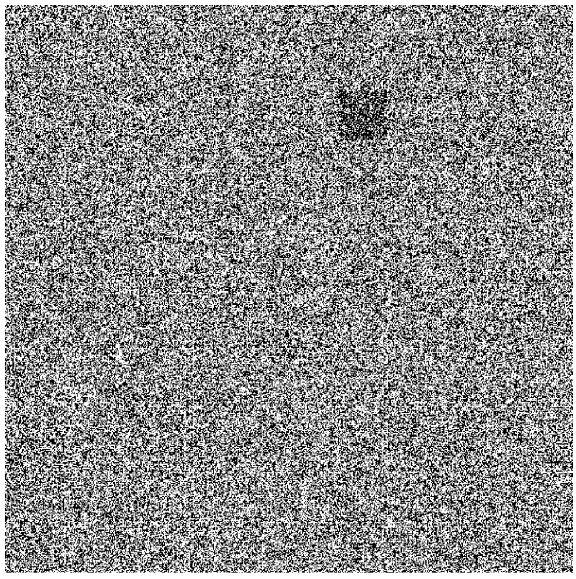

Fig. 6. Example of test image used for square detection.

$$F(l,d) = N^3 \mathrm{Prob}\left[\left|\frac{S_{l^2}}{l^2} - \frac{1}{2}\right| \geqslant \delta\right], \tag{6}$$

where S_n is the sum of n independent Bernoulli random variables with parameter 1/2, which means that

$$\mathrm{Prob}[S_n \geqslant k] = 2^{-n} \sum_{j=k}^{n} \binom{n}{k}.$$

The first factor of (6), N^3, counts the number of possible squares (N^2 locations and N side lengths). Using a large deviation estimate of (6) [9], we obtain

$$\log F(l,d) \simeq 3\log N - l^2((1+2\delta)\log(1+2\delta) + (1-2\delta)\log(1-2\delta)).$$

Each level line of F, defined by $F(l,d) = \varepsilon$, separates two regions in the (δ, l) plane: the squares that we are ε-meaningful and the other ones. These level lines are represented in Fig. 7 for several values of ε. Note that for small values of δ, we have

$$\log F(l,d) \simeq 3\log N - 8\delta^2 l^2 + O(\delta^4),$$

so that the level lines of F look like hyperbolae ($\delta \cdot l = cte$). If our visual perception has something to do with Helmholtz principle, the perception thresholds we measure in our experiments should correspond to a level line of F associated to some confidence level ε.

4.1.3. Results

We asked eight persons to realize the experiment described above. To each subject, we submitted a first (non-recorded) training set of 50 images, then a real set of 100 images. No other explanation were given to the subjects than just the written question "can you see ...".

[2] Can you see a square in this image?

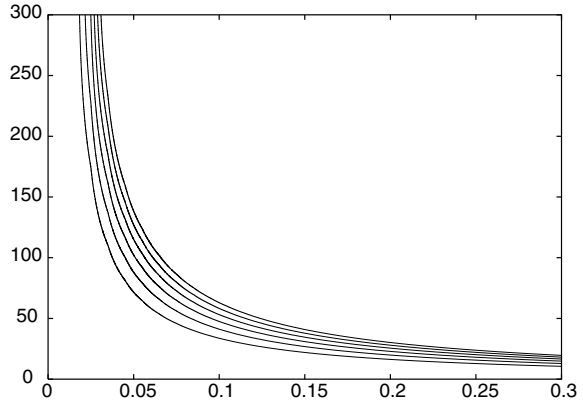

Fig. 7. Thresholds in the (δ, l) plane (square density and side length) predicted by Helmholtz principle for different values of ε ($\varepsilon = 10^{-0}$, $10^{-10}, 10^{-20}, \ldots, 10^{-50}$) when $N = 600$.

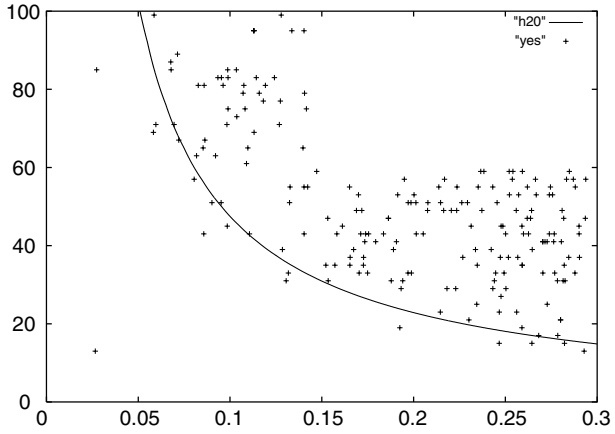

Fig. 9. Positive answers for $l \leqslant 100$ and the prediction curve ($\varepsilon = 10^{-20}$).

Each image had size 600×600 and was displayed on a $15''$ 1600×1200 LCD panel. The answers are reported on Fig. 8. Note that the values chosen randomly for d and l excluded the domain $\{(\delta, k), l > 60$ and $\delta > 0.15\}$, for which the detection of the square is too obvious.

The data we collected appeared to fit the Helmholtz model for $l \geqslant 100$, for ε around 10^{-20} (see Figs. 9 and 10). However, for larger values of l the perception threshold is worse than the one predicted by Helmholtz principle (see Fig. 11).

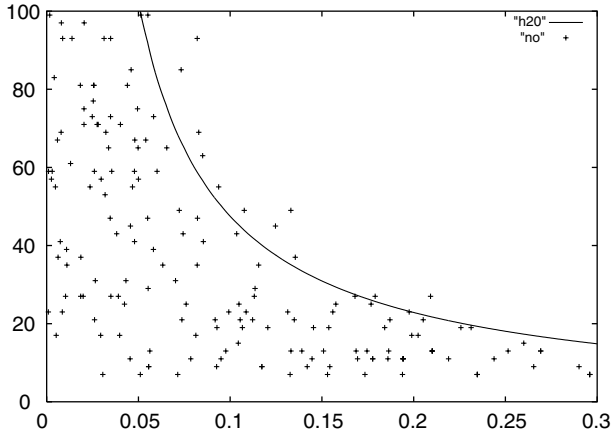

Fig. 10. Negative answers for $l \leqslant 100$ and the prediction curve ($\varepsilon = 10^{-20}$).

4.1.4. Discussion

Since the experiments involve two parameters (δ and l), whereas the model only has one (ε), the fit obtained between the measured data for $l \leqslant 100$ and the model seems to be relevant. How to explain the lack of fit for $l \geqslant 100$? For such a large square, our visual system has to "zoom out" the image, so that our model based on the fact that each point (black or white) is visible becomes questionable. Indeed, imagine that we keep a fixed physical image size but reduce the pixel size to zero (hence $N \to \infty$): then, our model predicts that any square with $\delta > 0$ will become visible for N large enough. This is clearly false, and can be explained by the

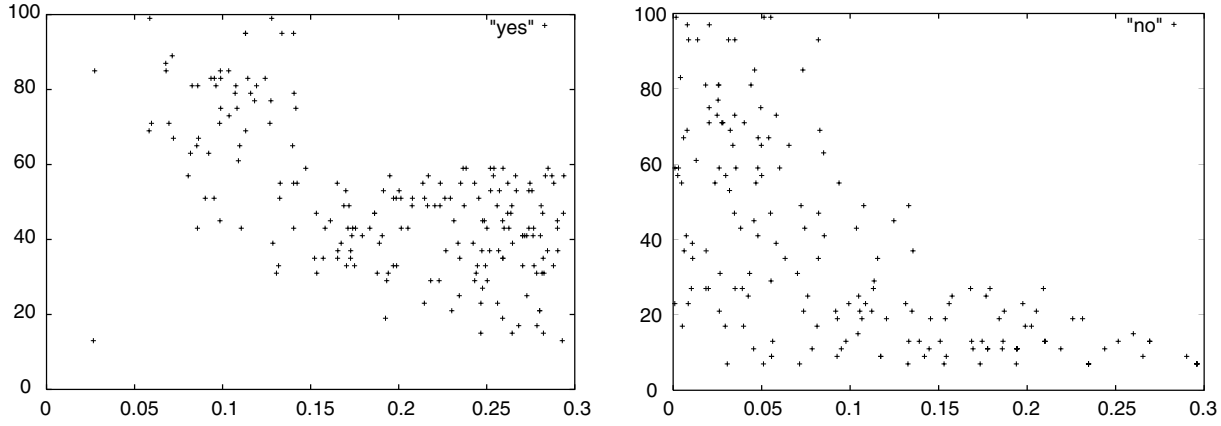

Fig. 8. Positive (left) and negative (right) answers for $l \leqslant 100$, in function of the relative density of the square (δ, horizontal axis) and its side length (l, vertical axis).

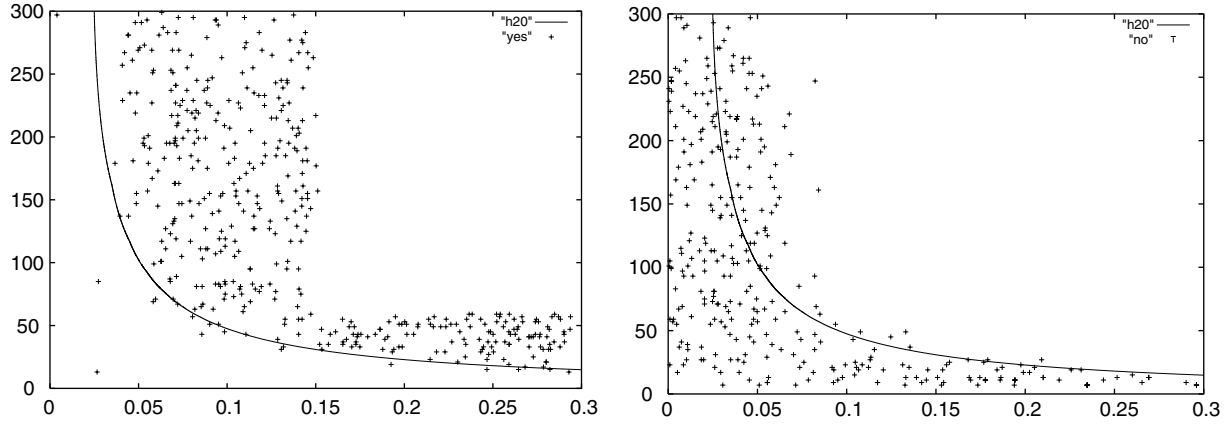

Fig. 11. Comparison between the prediction curve, positive (left) and negative (right) answers for $l \leqslant 300$.

fact that the resolution of our visual system is not infinite. Hence, we think that there is a maximum size for the squares beyond which the discrete model associated with Shannon sampling is no more valid.

The experiment we presented could be developed in several ways. First, one could try to perform more measurements, with one or several subjects. However, performing a lot of measurements with one single subject is difficult, because loss of concentration cannot be avoided beyond a certain number of experiments. When taking several subjects, there is a risk that the separation between "yes" and "no" answers becomes more fuzzy, since the threshold that each subject chooses may depend on the interpretation of the question (do I have to say yes when I am sure that I see a square, or when I think that there may be a square?). This question was actually asked to us by some subjects, but we gave no directions.

Another possibility to investigate further would be to change the image size (N) in order to check that the threshold value ε remains constant. However, our main concern here was not to predict exactly a perceptual threshold, but to predict its dependence upon the two parameters involved (the square density and side length).

4.2. Detection of alignments

4.2.1. Protocol

The protocol is essentially the same as for the square detection: images appear on the screen during 1.5 s, and the subject has to answer a question. For this experiment, the question is "Y a-t-il un alignement exceptionnel?". [3]

Each image is made of an hexagonal grid of size $N \times N$. At the center of each cell, there may be a little segment or not. This little segment has three possible orientations: 0°, 120° or 240°. To build each image, an alignment is generated in the following way: a position, a length l and a orientation are chosen, and then l segments are put in the adjacent cells defined by the initial position and the orientation. These segments are constrained to have the same direction as the alignment itself. Then, a density $\bar{d}$ is chosen randomly in [0,0.5], and the other cells are filled randomly and independently with respect to $\bar{d}$: each cell is empty with probability $1-d$ and contains one of the three possible segments with probability $d/3$. Such an image is shown on Fig. 12.

For each image, we record the answer of the subject (yes or no), along with the alignment length (l) and the density d of cell with segments (the expectation of d is $\bar{d}$). Each image is then represented as a point in the (d, l) plane.

4.2.2. Prediction

If we apply Helmholtz principle to this experiment, the number of false alarms associated to the encountering of an alignment of l consecutive segments is

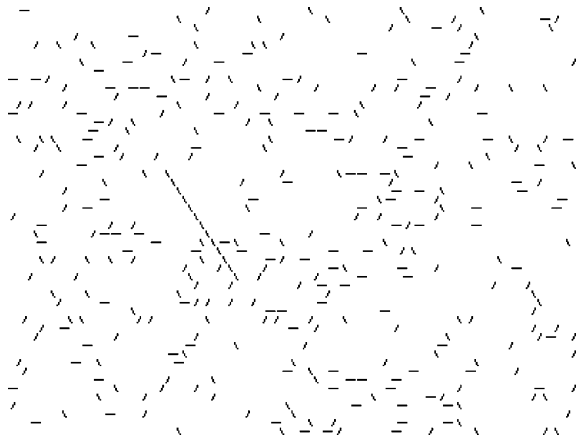

Fig. 12. Example of test image used for alignment detection ($N = 50$).

[3] Is there an exceptional alignment?

$$\mathrm{NFA} = \frac{3}{2}N^3\left(\frac{d}{3}\right)^l. \tag{7}$$

The number $3N^3/2$ approximately counts the number of possible alignments on the image (three orientations, N^2 positions for the center, $N/2$ possibilities for the segment length), and the second term is simply the probability that the l cells of the alignment have the proper orientation, knowing the empirical density d of non-empty cells in the image. From (7), we deduce that the threshold curve in the (d, l) plane corresponding to $\mathrm{NFA} \leqslant \varepsilon$ has equation

$$l = \frac{C}{\log(d/3)}, \quad \text{where } C = \log(\varepsilon) - \log\left(\frac{3}{2}N^3\right).$$

Some of these curves are displayed on Fig. 13.

4.2.3. Results

We collected 900 answers from seven different subjects. Like for the square detection, each subject was first submitted to a short (non-recorded) training set. This step was needed by the subject to understand well the question. The answers are shown on Fig. 14.

We found that for $\varepsilon \simeq 10^{-2}$, the threshold curve we predict separates not too badly the two sets of answers (see Figs. 15 and 16). The fitting could not have been

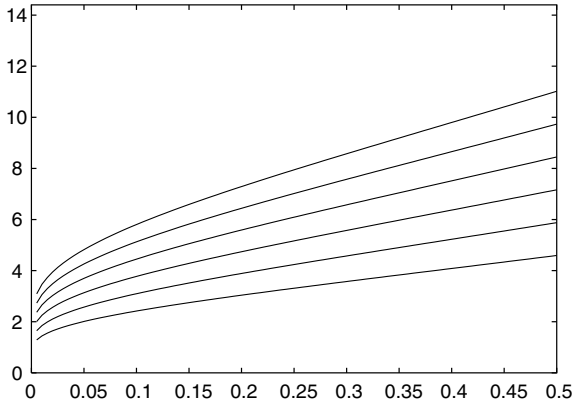

Fig. 13. Thresholds in the (d, l) plane (segment density and alignment length) predicted by Helmholtz principle for different values of ε ($\varepsilon = 10^{-0}, 10^{-1}, 10^{-2}, \ldots, 10^{-5}$) when $N = 50$.

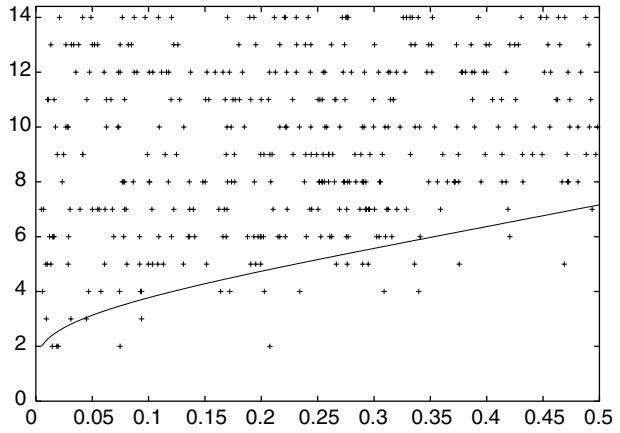

Fig. 15. Positive answers and the prediction curve ($\varepsilon = 10^{-2}$).

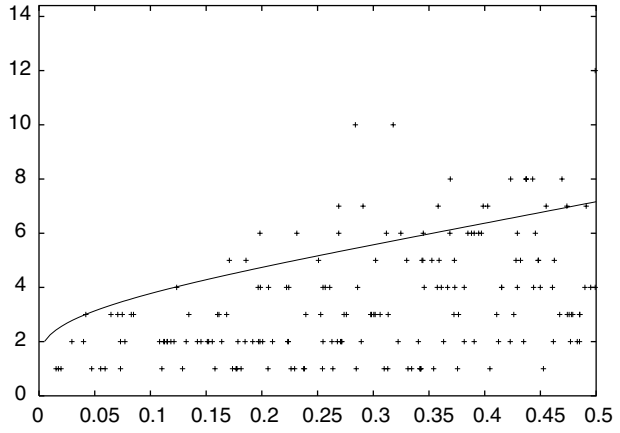

Fig. 16. Negative answers and the prediction curve ($\varepsilon = 10^{-2}$).

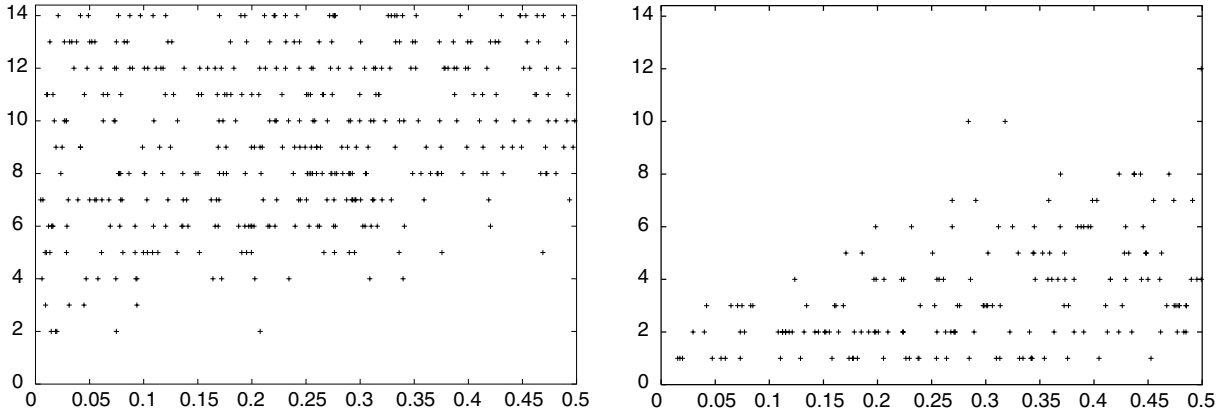

Fig. 14. Positive (left) and negative (right) answers.

much better because the answers are a little bit mixed up (the threshold is fuzzy).

4.2.4. Discussion

The predicted threshold curve seems to fit well the observed data, but one could argue that any straight line would do as well. This is probably true, but once again, since we only had one parameter (ε) to fit the data, we think that the fact that the position and the slope of our line-like threshold curve fits together the data for a convenient value of ε is significant. We do not claim that our prediction is satisfied very precisely (it would require much more experiments and probably a more constrained protocol to check it), but at least the predicted curve has a reasonable slope and offset.

4.3. Conclusion

What did we prove with these experiments? In some way, that the qualitative thresholds predicted by our computational approach of gestalt detection seem to fit the human perception, at least for the two examples above. One could argue that our experiments were designed in such a way that we answer "yes" when we perceive that the randomness of the image presented to the subject has been biased. Hence, there could be that we would fit as well the measured data by any threshold curve obtained from a criterion measuring the "randomness" of the image. In fact, this is exactly what we claim: we do not say that the model that allows us to compute the thresholds is reproduced in our brain, but simply that our perception is based on a detection of randomness that can be modeled by Helmholtz principle. To go further in the analogy between computational and psycho-visual gestalts, we would need to build experiments for which the Helmholtz model is visually achievable. This may be right for the second experiment (detection of aligned segments), but for the square detection it does not seem realistic to say that we can perceive each point of the 600×600 binary image presented during 1.5 s. This may explain why we find a fit for $\varepsilon = 10^{-20}$, which is in a sense far from being optimal. A significant improvement of our results would be realized by designing several experiments from which only one single threshold curve is predicted (corresponding to a kind of "universal" value of ε), instead of a family of curves from which we select the best fit as above.

Acknowledgements

Work partially supported by Office of Naval Research under grant N00014-97-1-0839, Centre National d'Etudes Spatiales, Centre National de la Recherche Scientifique et Ministère de la Recherche et de la Technologie. We thank the Fondation des Treilles for having hosted the authors during the redaction of this paper.

References

[1] J.-P. d'Alès, J. Froment, J.-M. Morel, Reconstruction visuelle et généricité, Intellectica 1 (28) (1999) 11–35.
[2] L. Alvarez, Y. Gousseau, J.-M. Morel, The size of objects in natural and artificial images, Adv. Imag. Electron Phys. 111 (1999) 167–242.
[3] F. Attneave, Some informational aspects of visual perception, Psych. Rev. 61 (1954) 183–193.
[4] A.J. Bell, T.J. Sejnowski, Edges are the 'independent components' of natural scenes, Adv. Neural Inform. Process. Syst. 9 (1996).
[5] E. Brunswik, J. Kamiya, Ecological cue-validity of 'proximity' and other Gestalt factors, Amer. J. Psychol. 66 (1953) 20–32.
[6] V. Caselles, B. Coll, J.-M. Morel, A Kanizsa programme, Prog. Nonlinear Different. Eqs. Applicat. 25 (1996) 35–55.
[7] A. Desolneux, L. Moisan, J.-M. Morel, Meaningful alignments, Int. J. Comp. Vision 40 (1) (2000) 7–23.
[8] A. Desolneux, L. Moisan, J.-M. Morel, Edge detection by helmholtz principle, J. Math. Imag. Vision 14 (3) (2001) 271–284.
[9] A. Desolneux, L. Moisan, J.-M. Morel, Maximal meaningful events and applications to image analysis, preprint CMLA no 2000–22, submitted. Available from <http://www.cmla.ens-cachan.fr/Cmla/Publications/>.
[10] A. Desolneux, L. Moisan, J.-M. Morel, Partial gestalts, preprint CMLA no 2001–22, submitted. Available from <http://www.cmla.ens-cachan.fr/Cmla/Publications/>.
[11] W.S. Geisler, J.S. Perry, B.J. Super, D.P. Gallogly, Edge co-occurrence in natural images predicts contour grouping performance, Vision Res. 41 (2001) 711–724.
[12] G. Guy, G. Medioni, Inferring global perceptual contours from local features, Int. J. Comp. Vision 20 (1) (1996) 113–133.
[13] B.K. Horn, Robot Vision, MIT Press, 1987.
[14] J.R. Bergen, B. Julesz, Textons, the fundamental elements of preattentive vision and perception of textures, Bell Syst. Tech. J. 62 (6) (1983) 1619–1645.
[15] G. Kanisza, Grammatica del Vedere, II Mulino, Bologna, 1980. Traduction française par Antonin Chambolle (ci-dessous).
[16] G. Kanizsa, La Grammaire du Voir, Editions Diderot, arts et sciences, 1997.
[17] G. Kanisza, Vedere e pensare, II Mulino, Bologna, 1991.
[18] G. Kanisza, Organization in Vision, Holt, Rinehart & Winston, 1979.
[19] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, in: 1st Intenational Computer Vision Conference IEEE 777, 1987.
[20] N. Kiryati, Y. Eldar, A.M. Bruckstein, A probabilistic Hough transform, Pattern Recog. 24 (4) (1991) 303–316.
[21] K. Koffka (Ed.), Principles of Gestalt Psychology, Harcourt and Brace, New York, 1935.
[22] N. Krüger, F. Wörgötter, Multi-modal estimation of collinearity and parallelism in natural image sequences, Network Comput. Neural Syst. 13 (4) (2002) 553–576.
[23] D. Lowe, Perceptual Organization and Visual Recognition, Kluwer Academic Publishers, 1985.
[24] D. Marr, Vision, Freeman and co., 1982.
[25] W. Metzger, Gesetze des Sehens, Waldemar Kramer, 1975.
[26] U. Montanari, On the optimal detection of curves in noisy pictures, CACM 14 (5) (1971) 335–345.
[27] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (6583) (1996) 607–609.

[28] A. Sha'Ashua, S. Ullman, Structural saliency: the detection of globally salient structures using a locally connected network, in: Proceedings of the 2nd International Conference on Computer Vision, 1988, pp. 321–327.
[29] D. Shaked, O. Yaron, N. Kiryati, Deriving stopping rules for the probabilistic hough transform by sequential analysis, Comp. Vision Image Understand. 63 (3) (1996) 512–526.
[30] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, and 623–656.
[31] M. Sigman, G.A. Cecchi, C.D. Gilbert, M.O. Magnasco, On a common circle: natural scenes and Gestalt rules, Proc. Nat. Acad. Sci. USA 98 (4) (2001) 1935–1940.
[32] C.V. Stewart, MINPRAN: a new robust estimator for computer vision, IEEE Trans. Patt. Anal. Mach. Intell. 17 (1995) 925–938.
[33] M. Wertheimer, Untersuchungen zur Lehre der Gestalt II, Psychol. Forsch. 4 (1923) 301–350, Translation published as Laws of Organization in Perceptual Forms, in: W. Ellis, A Source Book of Gestalt Psychology, Routledge and Kegan Paul, London, 1938, pp. 71–88.
[34] M. Wertheimer (Ed.), Laws of Organisation in Perceptual Forms, Harcourt & Brace & Javanowitch, London, 1935.
[35] D.M. Wuescher, K.L. Boyer, Robust contour decomposition using constant curvature criterion, IEEE Trans. Patt. Anal. Mach. Intell. 13 (1) (1991) 41–51.