When the a contrario approach becomes generative

Agnès Desolneux
CNRS and CMLA, ENS Cachan

April 9, 2014

Abstract

The *a contrario* approach is a statistical, hypothesis testing based approach to detect geometric meaningful events in images. The general methodology consists in computing the probability of an observed geometric event under a noise model (null hypothesis) H_0 and then declare the event meaningful when this probability is small enough. Generally, the noise model is taken to be the independent uniform distribution on the considered elements. Our aim in this paper will be to question the choice of the noise model: What happens if we "enrich" the noise model? How to characterize the noise models such that there are no meaningful event against them? Among them, what is the one that has maximal entropy? What does a sample of it look like? How is this noise model related to probability distributions on the elements that would produce, with high probability, the same detections? All these questions will be formalized and answered in two different frameworks: the detection of clusters in a set of points and the detection of line segments in an image. The general idea is to capture the perceptual information contained in an image, and then generate new images having the same visual content. We believe that such a generative approach can have applications for instance in image compression or for clutter removal.

Keywords: detection theory, non-accidentalness principle, maximum entropy distributions, clusters of points, line segments detection, image reconstruction, visual information theory.

1 Introduction

The *a contrario* approach is a statistical approach for detecting geometric events in an image that is inspired by the visual perception principle of *non-accidentalness*. This principle of non-accidentalness is also sometimes called *Helmholtz principle* and it is one of the fundamental ideas of visual recognition as developped for instance by D. Lowe in [21] and [22]. It is well summarized by S. C. Zhu in [35]: "Besides Gestalt psychology, there are two other theories for perceptual organization. One is the likelihood principle [30] which assigns a high probability for grouping two elements such as line segments, if the placement of these two elements as a low probability of resulting from *accidental arrangement* ([21] , [22])." The non-accidentalness principle is also stated by Witkin and Tenenbaum in [32], where they explain that "Because regular structural relationships are extremely unlikely to arise by the chance configuration of independent elements, such structure, when observed, almost certainly denotes some underlying unified cause or process." In [14], we have explained, developped and applied the *a contrario* approach in many different situations. It is a formalization of the non-accidentalness principle and it leads to a general methodology that can be summarized as follows:

1. Define a null-hypothesis H_0 that is a probability distribution on *elements*, and take a small number ε.

2. Observe a geometric event (that is a configuration of elements) in an image, and denote it by E.

3. Compute the probability of E under H_0.
4. If this probability is smaller than a threshold computed to ensure that on the average there are less than ε detections in an image where the elements would be distributed according to H_0, then declare E as ε-meaningful.

Under that form, the *a contrario* methodology has been applied to many detection problems, for instance: in [9], [13] and [15] for the detection of alignments in images; in [10] and [1] for the problem of edge detection in an image; in [12] and [8] for the problem of histogram segmentation (and the application to automatic color palette determination [7]); in [4] for the detection of good continuations and corners in images; in [5] for the detection of clusters of points; in [29] for motion detection; in [25] for shape recognition; in [23] for the detection of rigid point matches between two images; etc.

In most of these applications of the *a contrario* methodology the null-hypothesis H_0 (also sometimes called, in some papers, the *a contrario noise model*, the *background model* or the *naive model*) is taken as the independent uniform distribution on the elements. There are however noticeable exceptions to this choice of H_0. In particular, in [17], B. Grosjean and L. Moisan compute the detectability of spots in textured backgrounds and their null-hypothesis is a power-law Gaussian texture. They are able to obtain a formula that relates the detectability of a spot to its size, its contrast and the power exponent of the texture, and that matches the human visual perception. Another noticeable exception is [26] where A. Myaskouvskey, Y. Gousseau and M. Lindenbaum introduce in H_0 some correlation between the elements. The authors show in particular that such a noise model is more adapted to part-based object detection, and that it "enables reasonably accurate prediction of the false detection rate with no need for training data".

Our main goal in this paper will be to discuss the choice of the null hypothesis H_0. If the *a contrario* background noise model contains too many structures, i.e. if it is too rich, then the configurations are not unlikely anymore and therefore they are not meaningful against this rich background noise model. The extreme case is when the noise model is not random anymore but taken as being the image itself. In that case all geometric events have probability 1! Therefore, a natural question is: starting from the i.i.d. uniform noise model, then how far can we go in the "enrichment" still having some detections ? And what is the "first" distribution that does not lead to any detection ? Then, we can take a sample from this distribution and look what is like. It will of course contain, in some sense, the configurations of the original image. One can also ask what is the relationship between this distribution and the distributions on the elements that give, with high probability, the same detections as the original image. Such distributions are often looked for in texture synthesis problems. Indeed, the problem of texture synthesis is to be able to capture the "features" of a given original texture image in order to synthesize new sampled texture images that have the same "look and feel" as the original texture image. This question has been in particular formalized by Zhu, Wu and Mumford in [33] and [34], where they use exponential models, as these are the distributions that have at the same time the same histogram of filter responses as the original texture image and maximal entropy.

We will here also use the maximal entropy principle because we will look for background noise models that satisfy some constraints (such as : no meaningful configurations, or same detections as in an original given image) and are at the same time "as random as possible" (in the sense that they have maximal entropy). We will develop more precisely these questions in two different *a contrario* detection frameworks: the detection of clusters of points and the line segment detection problem in an image. In these two cases we will define distributions on elements (i.e. on points in the first case and on orientation fields in the second case) that generate new images having the same perceptual content (for the considered perception task, i.e. clusters of points in the first case and line segments in the second case) as the original given image. In that sense, the *a contrario* detection approach will become generative and will meet again visual perception modeling. We believe such an approach can have applications for image compression since we will capture and reproduce the meaningful configurations of an image while taking the rest "as random as possible".

From the point of view of information theory, we will have here to anwer questions of what could be called *visual information theory*, meaning that we will be interested in the amount of information contained in an image under visual perception constraints.

The paper is organized as follows: in Section 2 we will be interested in the detection of clusters of points in a planar domain. The clusters are defined as regions of the domain that contain "a lot of" points and they correspond to the Gestalt grouping law of *vicinity*. We will first, in Section 2.1, recall the *a contrario* framework in that case. Then, in Section 2.2, we will precisely describe the background noise distributions such that no regions are meaningful against them and will characterize the ones that have maximal entropy. In Section 2.3, we will be interested in distributions that have, most of the time, the same detections as the original set of points. Here again we will characterize the ones that have maximal entropy, and we will explore the link with the distributions of Section 2.2. In the second part of the paper (Section 3), we will consider a second *a contrario* detection framework, namely the one of the Line Segment Detector (LSD) of Grompone et al. [15]. We will recall precisely its framework in Section 3.1. Then in Section 3.2, we will be interested in distributions on orientation fields such that: else no rectangles are meaningful against them as a background noise model or they lead, most of the time, to the same detections as the original orientation field. We will answer simultaneously these two questions since here, unlike the case of clusters of points in Section 2, they are very closely related. In Section 3.3, we then address the question of the reconstruction of an image from an orientation field. Finally, we end the paper with Section 4 that contains a conclusion, a discussion and some views on future work.

2 Clusters of points: the vicinity grouping law

Assume we observe an original set of n points denoted by $s^0 = \{x_1^0, \dots, x_n^0\}$ in the domain $D = [0,1]^2$, the unit square of the plane. See Figure 1 for an example. Looking at these points, we are interested in the perception of the Gestalt grouping law of vicinity. According to the *a contrario* framework developed in [14], we detect meaningful groups of points for vicinity using the following methodology: start we a set of regions in D, then for each region, compute the number of points it contains, and if this number is significantly large (compared to what a noise model would give), then the region is said meaningful. We describe in the next subsection the whole method in more details, with precise definitions and properties.

2.1 A contrario framework and notations

Let $s^0 = \{x_1^0, \dots, x_n^0\}$ be a set of n points in the domain $D = [0,1]^2$. Let $\mathcal{R} = \{R_i\}_{1 \le i \le N_r}$ be a set of regions that cover D. It is not necessarily a partition of D (i.e. we may have $R_i \cap R_j \neq \emptyset$ for some $i \neq j$), but it can be. For a region $R \in \mathcal{R}$, we will denote by $k(R; s^0)$ the number of points, among $x_1^0, \dots, x_n^0$, it contains, that is:

$$k(R; s^0) = k(R; \{x_j^0\}_{1 \le j \le n}) = \sum_{j=1}^{n} \mathbb{1}_{x_j^0 \in R},$$

where $\mathbb{1}_E$ denotes the indicator function of an event E.

Definition 1 (Number of False Alarms)**.** *Let P be a probability distribution on the sets S of n points $S = \{X_1, \dots, X_n\}$ in D. For a region $R \in \mathcal{R}$, we define its Number of False Alarms under the law P by*

$$\mathrm{NFA}_P(R; s^0) = N_r \times \mathbb{P}_P[k(R; S) \ge k(R; s^0)], \tag{1}$$

where $k(R; S)$ is the random variable that counts the number of points of $S = \{X_1, \dots, X_n\}$ falling in the region R:

$$k(R; S) = k(R; \{X_j\}_{1 \le j \le n}) = \sum_{j=1}^{n} \mathbb{1}_{X_j \in R}.$$

Let $\varepsilon \in (0,1]$ be a small number. When $\mathrm{NFA}_P(R; s^0) < \varepsilon$, then we say that the region R is (ε, P)-meaningful for the set s^0.

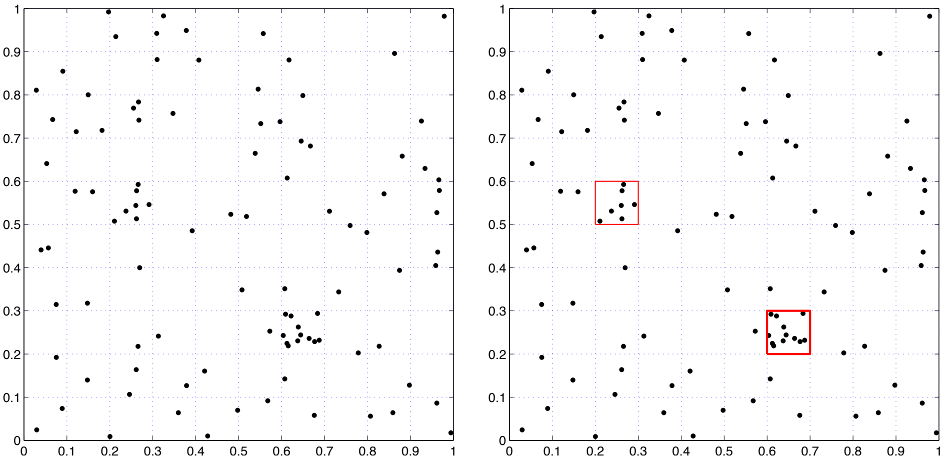


Figure 1: Example of points and regions. Here we have a set s^0 of $n = 100$ points in the domain $D = [0,1]^2$, and the tested regions are small squares of side length 0.1 and delimited by the dashed lines. On this image, one can clearly perceive that two regions are not like the others, in the sense that they contain an "unexpectedly high" number of points. On the right, we have delimited these two (ε, U)-meaningful regions (with here $\varepsilon = 1$) in red, with a line width proportional to $-\log_{10}(\mathrm{NFA}_U(R; s^0))$. The most meaningful region is the bottom right one.

The above defined Number of False Alarms measures how likely an observed event is. It is used as a measure of meaningfulness: the smaller the NFA is, the more meaningful the event is, in the sense that is has a low probability of having occurred just by chance.

The definition of the NFA and of ε-meaningful events is made in such a way that we have the following proposition ensuring that we control the number of "errors".

Proposition 1. *If a set S^0 of n points is randomly sampled from the probability distribution P, then the numbers of false alarms $\mathrm{NFA}_P(R_i; S^0)$, $1 \le i \le N_r$, become random variables and we have the fundamental property that*

$$\mathbb{E}_P\left(\sum_{i=1}^{N_r} \mathbb{1}_{\mathrm{NFA}_P(R_i;S^0)<\varepsilon}\right) < \varepsilon.$$

In other words, it means that the expected number (under the law P) of (ε, P)-meaningful regions is less than ε.

Proof. Let $R \in \mathcal{R}$ be a region and let us denote by F the tail distribution of the random variable $k(R; S)$ when S follows the probability distribution P, that is $F(k) = \mathbb{P}_P[k(R; S) \ge k]$ for all $k \in \mathbb{N}$. When the set S^0 of n points is randomly sampled from the probability distribution P, then $k(R; S^0)$ is a random variable and we have

$$\mathbb{P}_P[\mathrm{NFA}_P(R; S^0) < \varepsilon] = \mathbb{P}_P[F(k(R; S^0)) < \varepsilon/N_r] = \mathbb{P}_P[k(R; S^0) \ge F^{-1}(\frac{\varepsilon}{N_r})] = F(F^{-1}(\frac{\varepsilon}{N_r})) < \frac{\varepsilon}{N_r},$$

where the inverse F^{-1} is defined by $F^{-1}(\alpha) = \min\{k \in \mathbb{N}\,;\, F(k) < \alpha\}$ for all $\alpha \in [0,1]$. Now, to end the proof of the proposition we notice that, by linearity of the expectation,

$$\mathbb{E}_P\left(\sum_{i=1}^{N_r} \mathbb{1}_{\mathrm{NFA}_P(R_i;S^0)<\varepsilon}\right) = \sum_{i=1}^{N_r} \mathbb{P}_P[\mathrm{NFA}_P(R_i; S^0) < \varepsilon] < \sum_{i=1}^{N_r} \frac{\varepsilon}{N_r} = \varepsilon.$$

□

Generally in the use of the *a contrario* methodology [14], the noise distribution P (also sometimes called the *background noise model* or the *naive model*) is taken as U, the uniform distribution, making then the random variables $X_1, \dots, X_n$ independent and uniformly distributed on the domain D. The above proposition ensures that if the n points are sampled independently from the uniform distribution on D, then, on the average, the number of (ε, U)-meaningful regions will be less than ε. Notice that when $P = U$ is the uniform distribution, then the random variable $k(R; S)$ follows a binomial distribution and therefore the number of false alarms $\mathrm{NFA}_U(R; s^0)$ in Equation (1) is given by

$$\mathrm{NFA}_U(R; s^0) = N_r \times B(n, k(R; s^0), |R|),$$

where $|R|$ is the Lebesgue measure (*i.e.* the area) of the region R (we recall that $D = [0, 1]^2$ and therefore $|D| = 1$) and where $B(n, k, p)$ denotes the tail of the binomial distribution defined by

$$\forall p \in [0, 1],\ \forall 0 \leq k \leq n \text{ integers},\quad B(n, k, p) := \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j}. \tag{2}$$

A convenient way to make computations with the binomial tail is to rewrite it with the incomplete beta function defined by

$$\forall p \in [0, 1],\quad B_{n,k}(p) = \frac{\int_0^p t^{k-1}(1-t)^{n-k}\,dt}{\int_0^1 t^{k-1}(1-t)^{n-k}\,dt}. \tag{3}$$

The incomplete beta function is defined when n and k are real numbers, and it fits the binomial tail when n and k are integers, that is

$$B(n, k, p) = B_{n,k}(p) \text{ when } 0 \leq k \leq n \text{ are integers.}$$

This is why we use almost the same notation to denote them. Notice also that the function $p \mapsto B_{n,k}(p)$ defines, when $k > 0$, a continuous, strictly increasing mapping from $[0, 1]$ to itself.

An example of the (ε, U)-meaningful regions is given on Figure 1, with here $\varepsilon = 1$. These meaningful regions seem to correspond to the ones that we visually perceive as significant. Our aim here is not to discuss the pyschophysiological validity of the *a contrario* approach and the link between human visual perception thresholds and values of NFA_U. Some precise studies about these questions can be found in [3] and in [20]. What we want to discuss here, from a statistical point of view, is the choice of U as a background noise model. It is in some sense the only "natural" background choice since the i.i.d. uniform distribution U is the only stationary distribution with independent points, but in the hypothesis testing framework, it is worth questioning this specific choice. In particular, we can ask the following questions about the whole *a contrario* methodology:

- Question 1: What are the probability distributions P such that no regions are (ε, P)-meaningful for a given set of points s^0? What does a set of n points sampled from P look like?
- Question 2: Among all probability distributions P such that no regions are (ε, P)-meaningful, what is the one that has the maximal entropy?
- Question 3: What are the probability distributions Q such that if $S = \{X_1, \dots, X_n\}$ is a set of n points sampled from Q, then with high probability we have the same meaningful regions for S as for s^0?
- Question 4: Again: what is the probabilty distribution Q of the previous question that has the maximal entropy?
- Question 5: What is the link between the probability distributions of Questions 1 and 3?

These are very generic questions that can be addressed in any of the *a contrario* detection applications. We will here address them in the framework of clusters of points, starting with the first two questions.

2.2 Enrichment of the noise model

In all the following, the set of points $s^0 = \{x_1^0, \ldots, x_n^0\}$ is fixed. Let $\mathcal{P}$ denote the set of probability distributions P on $S = \{X_1, \ldots, X_n\}$ such that no regions are (ε, P)-meaningful for s^0. In fact this set $\mathcal{P}$ depends on both: the set of points $s^0 = \{x_1^0, \ldots, x_n^0\}$ and the set $\mathcal{R}$ of test regions. It is characterized by

$$\begin{aligned} P \in \mathcal{P} \quad &\Longleftrightarrow \quad \forall i = 1, \ldots, N_r, \ \ \mathrm{NFA}_P(R_i; s^0) \geq \varepsilon \\ &\Longleftrightarrow \quad \forall i = 1, \ldots, N_r, \ \ \mathbb{P}_P[k(R_i; S) \geq k(R_i; s^0)] \geq \varepsilon / N_r \end{aligned}$$

Notice that the set $\mathcal{P}$ is non-empty since it contains at least the Dirac distribution at $s^0 = \{x_1^0, \ldots, x_n^0\}$. Indeed, in that case we have $\mathbb{P}_P[k(R_i; S) \geq k(R_i; s^0)] = 1$ for all regions R_i. More generally $\mathcal{P}$ contains any distribution P that satisfies $k(R_i; S) = k_0(R_i; s^0)$ for all R_i.

Let $R \in \mathcal{R}$ be a region. We are first interested in probability distributions on $\Omega = D^n$ such that the region R is not (ε, P)-meaningful. We will then, in a second step, consider the case of all regions. Now, among the probability distributions P such that the region R is not (ε, P)-meaningful, we will be interested in the ones that are at the same time "as random as possible". A natural way to measure this is to use the entropy of the distribution. Therefore, we will consider only distributions P that admit a probability density denoted by f_P and we recall then that their differential (or continuous) entropy is defined by

$$H(P) = -\int_{\Omega} f_P(x_1, \ldots, x_n) \log f_P(x_1, \ldots, x_n) dx_1 \ldots dx_n,$$

with the usual convention that $0 \log 0 = 0$. Notice that by Jensen inequality, and the concavity of the log, we have

$$H(P) \leq H(U) = \log(|\Omega|) = 0,$$

because we recall that here $\Omega = D^n$ with $D = [0,1]^2$.

We will also be interested in the set denoted $\mathcal{I}$ of probability distributions that make the points $X_1, \ldots, X_n$ independent, which is equivalent to say that $P \in \mathcal{I}$ if and only if the density f_P of P is of the form $f_P(x_1, \ldots, x_n) = \tilde{f}_P(x_1)\tilde{f}_P(x_2) \ldots \tilde{f}_P(x_n)$. Notice that the uniform distribution U belongs to $\mathcal{I}$.

We now characterize the distributions P that make a region R not (ε, P)-meaningful and that have maximal entropy.

Proposition 2. *Let $R \in \mathcal{R}$ be a region. Let $\mathcal{P}_R$ denote the set of probability distributions P on Ω such that the region R is not (ε, P)-meaningful. We then have two cases:*

1. *If the region R is not (ε, U)-meaningful, that is if $\mathrm{NFA}_U(R) \geq \varepsilon$, then $U \in \mathcal{P}_R \cap \mathcal{I}$, and it is the maximal entropy distribution in $\mathcal{P}_R$.*

2. *If, on the contrary, the region R is (ε, U)-meaningful, that is if $\mathrm{NFA}_U(R) < \varepsilon$, then $U \notin \mathcal{P}_R$ and*

   (a) The distribution $P \in \mathcal{P}_R$ that has $H(P)$ maximal is given by

$$f_P(x_1, \ldots, x_n) = \begin{cases} \frac{\varepsilon}{N_r |\Omega_0|} & \text{if } (x_1, \ldots, x_n) \in \Omega_0 \\ \frac{N_r - \varepsilon}{N_r |\Omega \backslash \Omega_0|} & \text{if } (x_1, \ldots, x_n) \notin \Omega_0, \end{cases}$$

   where $\Omega_0 := \{(x_1, \ldots, x_n) \in \Omega; \sum_{j=1}^n \mathbb{1}_{x_j \in R} \geq k(R; s^0)\}$ is the set of configurations of n points such that at least $k(R; s^0)$ of them are in R. Notice that

$$|\Omega_0| = B(n, k(R; s^0), |R|).$$

(b) *The set $\mathcal{P}_R \cap \mathcal{I}$ is non-empty, and the probability distributions it contains are characterized by*

$$p_P(R) := \mathbb{P}_P[X_1 \in R] \geq B^{-1}_{n,k(R;s^0)}(\varepsilon/N_r),$$

where $\alpha \mapsto B^{-1}_{n,k}(\alpha)$ is the inverse of the incomplete beta function defined by (3)*. And the distribution $P \in \mathcal{P}_R \cap \mathcal{I}$ that has maximal entropy is given by*

$$\tilde{f}_P(x) = \begin{cases} \frac{1}{|R|} B^{-1}_{n,k(R;s^0)}(\varepsilon/N_r) & \textit{if } x \in R \\ \frac{1}{1-|R|}(1 - B^{-1}_{n,k(R;s^0)}(\varepsilon/N_r)) & \textit{if } x \notin R, \end{cases}$$

Proof. Let us start with the first point of the proposition: when the region R is not (ε, U)-meaningful, then by definition $U \in \mathcal{P}_R$. And the statement follows from the fact that U also belongs to $\mathcal{I}$, and that U is the maximal entropy distribution among all probability distributions on Ω.

In the second case, we now assume that the region R is (ε, U)-meaningful for s^0. This implies that $B(n, k(R; s^0), |R|) < \varepsilon/N_r$. Then, we first notice that

$$P \in \mathcal{P}_R \iff \mathbb{P}_P(k(R;S) \geq k(R;s^0)) \geq \frac{\varepsilon}{N_r} \iff P(\Omega_0) \geq \frac{\varepsilon}{N_r},$$

where $\Omega_0 = \{(x_1, \dots, x_n) \in \Omega; \sum_{j=1}^n \mathbb{1}_{x_j \in R} \geq k(R; s^0)\}$. Now, let P_a, $a \in (0,1)$, be the family of probability distributions on Ω defined by $f_{P_a}(x_1, \dots, x_n) = a/|\Omega_0|$ if $(x_1, \dots, x_n) \in \Omega_0$ and $f_{P_a}(x_1, \dots, x_n) = (1-a)/(1-|\Omega_0|)$ if $(x_1, \dots, x_n) \notin \Omega_0$. Notice that $P_a(\Omega_0) = a$ and therefore $P_a \in \mathcal{P}_R$ if and only if $a \geq \varepsilon/N_r$.

For any $P \in \mathcal{P}_R$, we have that the Kullback-Leibler divergence of P_{a_P} from P, where $a_P = P(\Omega_0)$ is always positive. And this divergence is also, by definition, equal to

$$\begin{aligned} D(P||P_{a_P}) &= \int_\Omega f_P(y_1, \dots, y_n) \log \frac{f_P(y_1, \dots, y_n)}{f_{P_{a_P}}(y_1, \dots, y_n)} \, dy_1 \dots dy_n \\ &= -H(P) - P(\Omega_0) \log \frac{a_P}{|\Omega_0|} - (1 - P(\Omega_0)) \log \frac{1 - a_P}{1 - |\Omega_0|} \\ &= -H(P) + H(P_{a_P}). \end{aligned}$$

This shows that $H(P) \leq H(P_{a_P})$. Moreover a simple study of the function

$$a \mapsto H(P_a) = -a \log \frac{a}{|\Omega_0|} - (1-a) \log \frac{1-a}{1-|\Omega_0|}$$

shows that it is increasing on $[0, |\Omega_0|]$ and decreasing on $[|\Omega_0|, 1]$. Therefore, under the constraint $a \geq \frac{\varepsilon}{N_r} > |\Omega_0| = B(n, k(R; s^0), |R|)$, it is maximal for $a = \frac{\varepsilon}{N_r}$, and this achieves the proof of *2.(a)*.

For the second point, if the probability distribution P makes the points $X_1, \dots, X_n$ independent then $k(R; S)$ follows a binomial distribution and more precisely we have

$$\mathbb{P}_P(k(R;S) \geq k(R;s^0)) = B(n, k(R;s^0), p_P(R)) = B_{n,k(R;s^0)}(p_P(R)),$$

where $p_P(R) := \mathbb{P}_P[X_1 \in R]$. Then, since $p \mapsto B_{n,k(R;s^0)}(p)$ is a continuous and strictly increasing function, we conclude that $P \in \mathcal{P}_R$ if and only if $p_P(R) \geq B^{-1}_{n,k(R;s^0)}(\varepsilon/N_r)$.

In the case of a probability distribution P making the points $X_1, \dots, X_n$ independent, its entropy can be easily computed since we have

$$H(P) = H(f_P) = nH(\tilde{f}_P).$$

Then by a computation analogous to the one with P_a above, we find that the distribution that has maximal entropy among all distributions in $\mathcal{P}_R \cap \mathcal{I}$ has a probability density $\tilde{f}_P$ that is constant on R, and on $D \setminus R$ and such that $\int_R \tilde{f}_P(x)\, dx = B^{-1}_{n,k(R;s^0)}(\varepsilon/N_r)$.

□

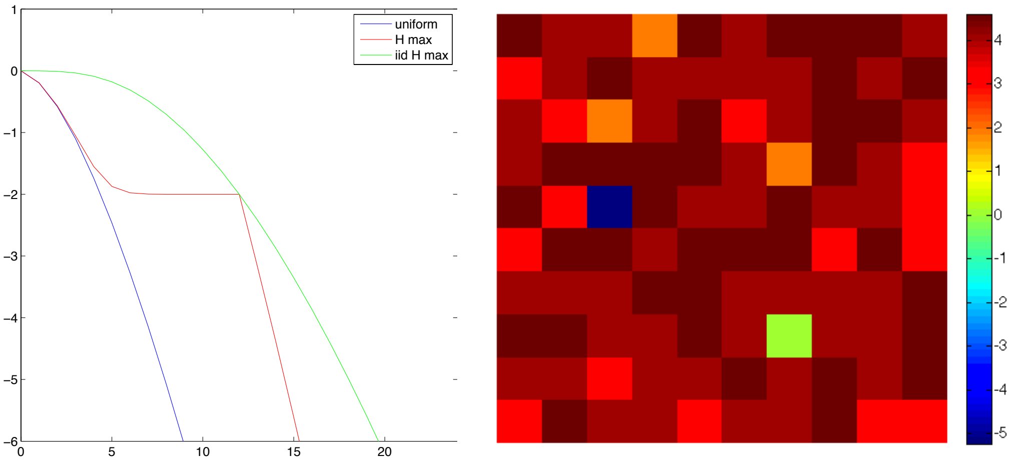


Figure 2: On the left, graphs of $k \mapsto \log_{10} \mathbb{P}_P(k(R;S) \geq k)$ for P being respectively the uniform distribution U (blue curve), the distribution in $\mathcal{P}_R$ that has maximal entropy (red curve) and the distribution in $\mathcal{P}_R \cap \mathcal{I}$ that has maximal entropy (green curve). On the right, image of the $\log(\mathrm{NFA}_P(R_i))$ for all regions R_i when P is the distribution in $\mathcal{P}_R \cap \mathcal{I}$ with maximal entropy. See the text for more details.

On Figure 2, we illustrate Proposition 2. The region R is here the most meaningful region of the points shown on Figure 1 (it is delimited by the bottom right red square). It contains $k(R;s^0) = 12$ points among a total of $n = 100$ points. We have here $|R| = 10^{-2}$, with $N_r = 100$ test regions, and therefore we compute $\log_{10} \mathrm{NFA}_U(R) = \log_{10}(100 \times B(100, 12, 0.01)) \simeq -7.3$. This region is obviously (ε, U)-meaningful (with $\varepsilon = 1$ in the following). On the left of Figure 2, we compare the graphs of $k \mapsto \log_{10} \mathbb{P}_P(k(R;S) \geq k)$ for P being respectively the uniform distribution U (blue curve), the distribution in $\mathcal{P}_R$ that has maximal entropy (red curve) and the distribution in $\mathcal{P}_R \cap \mathcal{I}$ that has maximal entropy (green curve). The value ε/N_r is here 0.01, and we see on the figure how the red and blue curves fit the value $-2 = \log_{10}(\varepsilon/N_r)$ at $k = 12$. On the right, we show the image of the $\log(\mathrm{NFA}_P(R_i;s^0))$ for all regions R_i when P is the distribution in $\mathcal{P}_R \cap \mathcal{I}$ with maximal entropy. Notice that the green value is 0, which means that the considered region R satisfies here $\mathrm{NFA}_P(R;s^0) = \varepsilon = 1$, all other regions are not (ε, P)-meaningful except the region in blue (denoted R_2), that was the other (ε, U)-meaningful region of Figure 1. It is even "more meaningful" here since we had $\log(\mathrm{NFA}_U(R_2;s^0)) = -4.9$ whereas we have now $\log(\mathrm{NFA}_P(R_2;s^0)) = -5.2$.

The distributions in $\mathcal{P}_R$ are able to "explain" or "erase" the region R, in the sense that the observed number of points in the region R is not a meaningful deviation of these distributions. We are now interested in distributions that can "explain" or "erase" all regions $R_i \in \mathcal{R}$.

Theorem 1. *Let $\mathcal{R} = \{R_i\}_{1 \leq i \leq N_r}$ be a set of regions that cover the domain D. And let $\mathcal{P}$ denote the set of probability distributions on $(X_1, \ldots, X_n)$ such that no regions are (ε, P)-meaningful. We then have the two following properties:*

1. *The distribution in $\mathcal{P}$ that has maximal entropy is of the form*

$$f_P(x_1, \ldots, x_n) = \frac{1}{Z_\lambda} \exp\left(\sum_{i=1}^{N_r} \lambda_i \mathbb{1}_{k(R_i;\{x_1,\ldots,x_n\}) \geq k(R_i;s^0)} \right),$$

   where the λ_i are real numbers and Z_λ is the normalizing constant that ensures $\int_\Omega f_P = 1$. If $U \in \mathcal{P}$, then U is the maximal entropy distribution, and it corresponds to $\lambda_i = 0$ for all i.

2. *We assume here that the regions make a partition of D. If $\varepsilon/N_r \leq 1/2$, then $\mathcal{P} \cap \mathcal{I} \neq \emptyset$ and in that case, if there is moreover at least one (ε, U)-meaningful region, then the maximum*

entropy distribution in $\mathcal{P} \cap \mathcal{I}$, denoted P_0, is given by

$$\tilde{f}_{P_0}(x) = \left\{ \begin{array}{l} \frac{\alpha_i}{|R_i|} \textit{ if } x \in R_i \textit{ with } 1 \leq i \leq m_r \\ \beta \textit{ otherwise }, \end{array} \right. \tag{4}$$

where $\alpha_i := B^{-1}_{n,k(R_i;s^0)}(\varepsilon/N_r)$, the regions are numbered in such a way that $\alpha_i/|R_i|$ is decreasing, i.e.

$$\frac{\alpha_1}{|R_1|} \geq \frac{\alpha_2}{|R_2|} \geq \dots \frac{\alpha_{N_r}}{|R_{N_r}|}.$$

And m_r is defined by

$$m_r := \min\{1 \leq j \leq N_r - 1 \textit{ such that } \frac{1 - \sum_{i=1}^{j} \alpha_i}{1 - \sum_{i=1}^{j} |R_i|} \geq \frac{\alpha_{j+1}}{|R_{j+1}|}\} \tag{5}$$

Finally, β is the constant such that $\int_D \tilde{f}_{P_0}(x)\, dx = 1$.

And when there are no (ε, U)-meaningful region, then $P_0 = U$ is the maximal entropy distribution in $\mathcal{P} \cap \mathcal{I}$.

Proof. We have the following characterization of $\mathcal{P}$: by definition, $P \in \mathcal{P}$ if and only if for all $R_i \in \mathcal{R}$, $\mathbb{P}_P[k(R_i;S) \geq k(R_i;s^0)] \geq \varepsilon/N_r$. The first point of the theorem is a consequence of a classical result on exponential distributions that we recall here (a proof can be found in [24] p. 221 or in [6] p. 410): If we look for a distribution f that has maximal entropy under the constraints $\int c_i(x)f(x)dx = a_i$ (that is $\mathbb{E}_f(c_i(X)) = a_i$) then f is of the form $f(x) = \frac{1}{Z}\exp(\sum_i \lambda_i c_i(x))$, where the constants Z and λ_i have to be determined so that the integral of f is 1 and the constraints on c_i are satisfied.
In our case, the constraints are $\mathbb{P}_P[k(R_i;S) \geq k(R_i;s^0)] \geq \varepsilon/N_r$. We then simply notice that $\mathbb{P}_P[k(R_i;S) \geq k(R_i;s^0)] = \mathbb{E}_P(\mathbb{1}_{k(R_i;S) \geq k(R_i;s^0)})$, and this gives the result. Now the main difficulty is that there is no closed formula to determine the λ_i. This has to be done numerically as it is done for instance by Zhu, Wu and Mumford in [34] in the framework of texture synthesis.

For the second point of the theorem, we fist notice that a distribution P in $\mathcal{P} \cap \mathcal{I}$ is characterized by

$$\forall R_i \in \mathcal{R}, \;\; p_P(R_i) := \mathbb{P}_P[X_1 \in R_i] \geq B^{-1}_{n,k(R_i;s^0)}(\frac{\varepsilon}{N_r}). \tag{6}$$

Therefore, since the regions $\{R_i\}_{1 \leq i \leq N_r}$ are a partition of the domain D, the set $\mathcal{P} \cap \mathcal{I}$ is non-empty if and only if $\sum_{i=1}^{N_r} B^{-1}_{n,k(R_i;s^0)}(\frac{\varepsilon}{N_r}) \leq 1$. Indeed otherwise if $\sum_{i=1}^{N_r} B^{-1}_{n,k(R_i;s^0)}(\frac{\varepsilon}{N_r}) > 1$, then by (6), we would have $\sum_{i=1}^{N_r} p_P(R_i) > 1$ which is impossible when the $\{R_i\}_{1 \leq i \leq N_r}$ are a partition of the domain D.

Now, we use the following lemma.

Lemma 1. *Let $0 \leq k \leq n$ be integers, then*

$$B_{n,k}\left(\frac{k}{n}\right) = B(n,k,\frac{k}{n}) \geq \frac{1}{2}.$$

Proof. The proof of this lemma can be found in [19], where the authors study the location of the median value of binomial distributions. It can also be found in [27] where the authors study the relative location of the median, the mean and the mode of a beta distribution. In particular they show (using our notations) that

$$\min\left(\frac{k-1}{n-1}, \frac{k}{n+1}\right) \leq B^{-1}_{n,k}\left(\frac{1}{2}\right) \leq \max\left(\frac{k-1}{n-1}, \frac{k}{n+1}\right). \tag{7}$$

And this implies Lemma 1 since $\max\left(\frac{k-1}{n-1}, \frac{k}{n+1}\right) \leq \frac{k}{n}$. □

Assume that $\varepsilon/N_r \leq 1/2$, then according to Lemma 1, we have $B^{-1}_{n,k(R_i;s^0)}(\frac{\varepsilon}{N_r}) \leq B^{-1}_{n,k(R_i;s^0)}(\frac{1}{2}) \leq \frac{k(R_i;s^0)}{n}$ for all R_i. Then, since the regions $\{R_i\}_{1\leq i\leq N_r}$ are a partition of D, we have $\sum_{i=1}^{N_r} k(R_i;s^0) = n$. Therefore $\sum_{i=1}^{N_r} B^{-1}_{n,k(R_i;s^0)}(\frac{\varepsilon}{N_r}) \leq 1$ and consequently $\mathcal{P}\cap\mathcal{I}\neq\emptyset$.

Let us now prove that the distribution P_0 defined by (4) is indeed the maximum entropy distribution of $\mathcal{P}\cap\mathcal{I}$. We first notice that, thanks to the concavity of the log function and using the fact that $\{R_i\}_{1\leq i\leq N_r}$ is a partition of D,

$$H(P) = nH(\tilde{f}_P) = -n\sum_{i=1}^{N_r}\int_{R_i}\tilde{f}_P(x)\log\tilde{f}_P(x)\,dx \leq -n\sum_{i=1}^{N_r} p_P(R_i)\log\frac{p_P(R_i)}{|R_i|},$$

where $p_P(R_i) := \int_{R_i}\tilde{f}_P(x)\,dx$, and where equality holds when $\tilde{f}_P$ is constant on each R_i. Therefore, to achieve maximum entropy, the density $\tilde{f}_P$ has to be constant on each region R_i. Using the notation $\alpha_i := B^{-1}_{n,k(R_i;s^0)}(\frac{\varepsilon}{N_r})$, the constraint of belonging to $\mathcal{P}\cap\mathcal{I}$ is equivalent to the constraint $p_P(R_i) \geq \alpha_i$ for all $1\leq i\leq N_r$. The proof of the optimality of the distribution P_0 of (4) then follows from the following lemma.

Lemma 2. *Let $p_1,\dots,p_N$ be a probability distribution on $\{1,2,\dots,N\}$. Let $r_1,\dots,r_N$ be another probability distribution on $\{1,2,\dots,N\}$, and let $\alpha_1,\dots,\alpha_N$ be positive real numbers such that $\sum_{i=1}^N\alpha_i\leq 1$. Assume moreover that the indices are ordered in such a way that*

$$\frac{\alpha_1}{r_1}\geq\frac{\alpha_2}{r_2}\geq\ldots\geq\frac{\alpha_N}{r_N}.$$

Then, under the constraints $p_i\geq\alpha_i$, for all $1\leq i\leq N$, the Kullback-Leibler distance

$$p=(p_1,\dots,p_N)\mapsto D(p||r)=\sum_{i=1}^N p_i\log\frac{p_i}{r_i}$$

is minimal for $p=p^*$ given by

$$p_i^*=\alpha_i \text{ for } 1\leq i\leq m \text{ and } p_i^*=\frac{1-\sum_{j=1}^m\alpha_j}{1-\sum_{j=1}^m r_j}r_i \text{ for } i>m \tag{8}$$

where $m:=\min\{0\leq j\leq N-1$ such that $\frac{1-\sum_{i=1}^j\alpha_i}{1-\sum_{i=1}^j r_i}\geq\frac{\alpha_{j+1}}{r_{j+1}}\}$.

Proof. We first notice that the set $\mathcal{D}_\alpha$ of discrete probability distributions $p=(p_1,\dots,p_N)$ on $\{1,2,\dots,N\}$ that satisfy the constraint $p_i\geq\alpha_i$ for all $1\leq i\leq N$, is a closed and convex subset of the set $\mathcal{D}$ of distributions on $\{1,2,\dots,N\}$. Then, thanks to the strict convexity of the function $p\mapsto D(p||r)$ (see [6] for instance), we have the existence and the unicity of p^* in $\mathcal{D}_\alpha$ that achieves the minimum of $D(p||r)$, $p\in\mathcal{D}_\alpha$. Let us now prove that p^* given by (8) is indeed the point achieving the minimum. In order to do this, we prove that any other $q\in\mathcal{D}_\alpha$, $q\neq p^*$ can not be a point of minimum. If $q\in\mathcal{D}_\alpha$, $q\neq p^*$, then there exist two indices i_0 and j_0 such that $q_{i_0}>p^*_{i_0}$ and $q_{j_0}<p^*_{j_0}$. Since $p^*_i=\alpha_i$ for $1\leq i\leq m$, we necessarily have $j_0>m$ (otherwise q is not in $\mathcal{D}_\alpha$). Now, we also have that $\frac{q_{j_0}}{r_{j_0}}<\frac{q_{i_0}}{r_{i_0}}$. Indeed, let us denote $\beta=\frac{1-\sum_{j=1}^m\alpha_j}{1-\sum_{j=1}^m r_j}$. Then we have two cases: $i_0>m$ and $i_0\leq m$. In the first case, when $i_0>m$, then $q_{i_0}>p^*_{i_0}=\beta r_{i_0}$. Since we have $q_{j_0}<p^*_{j_0}=\beta r_{j_0}$, this implies that $\frac{q_{j_0}}{r_{j_0}}<\frac{q_{i_0}}{r_{i_0}}$. In the second case, when $i_0\leq m$, then using the fact that $\sum_{i=1}^N\alpha_i\leq 1$ and that the ratios α_i/r_i are decreasing, we have $\beta\leq\frac{\sum_{j>m}\alpha_j}{\sum_{j>m}r_j}\leq\alpha_{i_0}/r_{i_0}$. And consequently: $q_{i_0}/r_{i_0}>p^*_{i_0}/r_{i_0}=\alpha_{i_0}/r_{i_0}\geq\beta=q_{j_0}/r_{j_0}$. Therefore we also have $\frac{q_{j_0}}{r_{j_0}}<\frac{q_{i_0}}{r_{i_0}}$.
Finally, considering the function $t\mapsto f(t)=(q_{j_0}+t)\log\frac{q_{j_0}+t}{r_{j_0}}+(q_{i_0}-t)\log\frac{q_{i_0}+t}{r_{i_0}}$ defined for $t\geq 0$ and small, we compute $f'(0)=\log\frac{q_{j_0}}{r_{j_0}}-\log\frac{q_{i_0}}{r_{i_0}}<0$ and that shows that q can not be a point of minimum of $D(\cdot||r)$ in $\mathcal{D}_\alpha$.

□

Notice that the result stated in Lemma 2 is very similar to the computations in [2] where the maximum likelihood discrete decreasing distribution is estimated from observed samples and yields to the so-called Grenander estimator. □

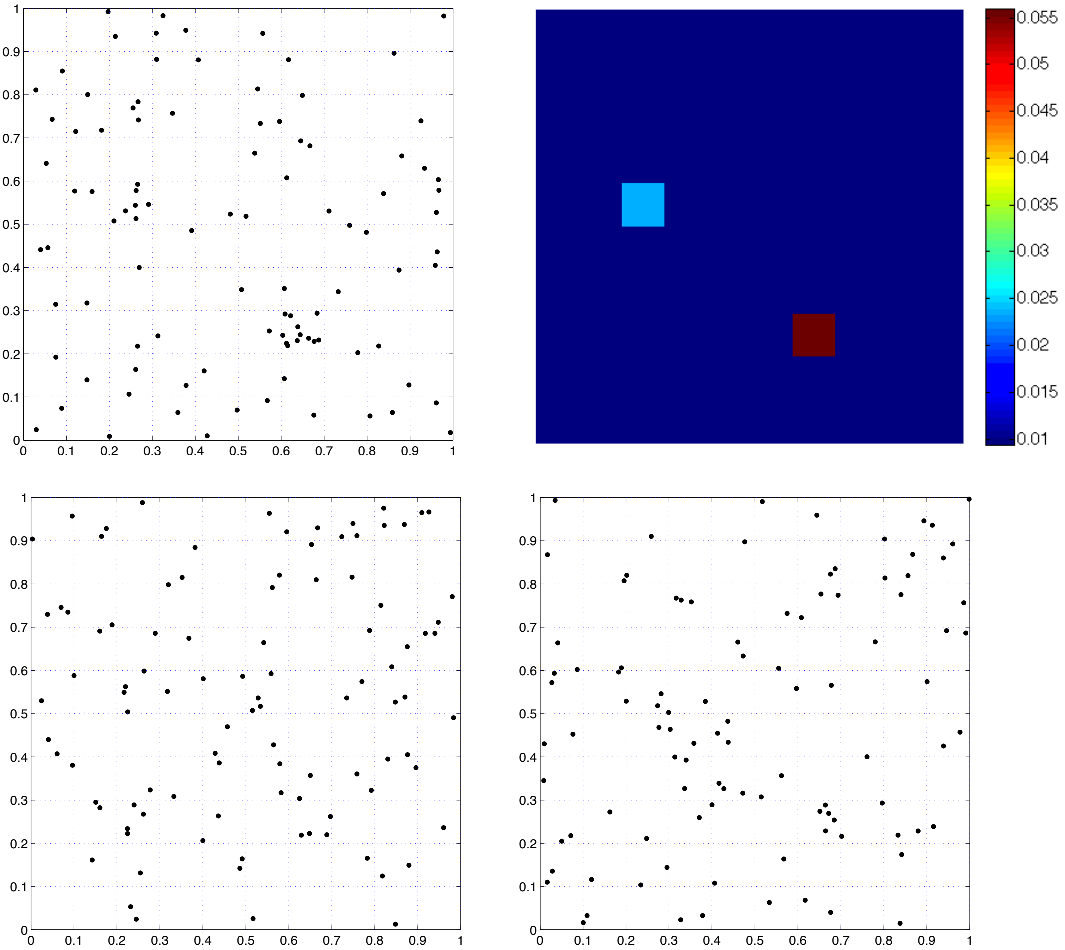


Figure 3: Top left: the initial set of points $s^0 = \{x_j^0\}_{1 \leq j \leq n}$ and the regions R_i (delimited by the dashed squares). Top right: value of $p_{P_0}(R_i)$ for each region R_i where P_0 is the maximal entropy distribution such that the points are independent and such that no regions are (ε, P_0)-meaningful for s^0, it is defined by (4). Bottom: two samples of $n = 100$ points with the law P_0. These samples look a bit more "structured" than uniform samples but they don't have the same "look and feel" as the initial set of points $s^0 = \{x_j^0\}_{1 \leq j \leq n}$. See the text for more detailed comments.

Let us also notice that the hypothesis ε / N_r in Theorem 1 is not a strong hypothesis since it is satisfied as soon as $\varepsilon \leq 1$ and there are at least two regions. In practice we are always in that case.

On Figure 3 we illustrate Theorem 1. On the top left of the figure, we first show again the initial set of points $s^0 = \{x_j^0\}_{1 \leq j \leq n}$ and the regions $\{R_i\}_{1 \leq i \leq N_r}$ (delimited by the dashed squares). On the top right, we show for each region the value of $p_{P_0}(R_i)$, where P_0 is the distribution defined in (4). Notice that under the uniform distribution U, all regions have a probability equal to 0.01. Now, we have here that under P_0, the most (ε, U)-meaningful region R_1 (i.e. the one with NFA_U minimal) has a probability 5 times larger, and the second region R_2 has a probability 2.5 larger. This implies that if we sample $n = 100$ points with the law P_0, the expectation of the number of points in R_1 is approximately 5 (whereas we have $k_0(R_1; s^0) = 12$) and the expectation of the number of points in R_2 is approximately 2.5 (whereas we have $k_0(R_2; s^0) = 7$). On the bottom of the figure we show two sets of $n = 100$ points sampled from P_0. These samples may look a bit more "structured" than uniform samples but they don't have the same "look and feel" as the initial set of points $s^0 = \{x_j^0\}_{1 \le j \le n}$. Yet, by definition, no regions are (ε, P_0)-meaningful, which means that there is no rejection of P_0 with the set s^0. This is in fact a kind of paradox of goodness of fit tests: it is not because you don't reject a distribution that this distribution fits your data.

In conclusion of this part, as it is illustrated by Figure 3, the samples from P_0 are more "structured" than uniform samples, but we don't perceive in them the same perceptual groups as in the original set of points. Therefore, we can ask: what distribution of points would produce, "most of the time", the same perception as $\{x_j^0\}_{1 \le j \le n}$? This is the aim of the next section.

2.3 Distributions generating the same perceptual groups

We want to formalize here in this section the notion of *"samples, that would, most of the time, show the same perceptual groups as the initial set of points $s^0 = \{x_j^0\}_{1 \le j \le n}$"*. In order to do this, we first need to define more precisely what "perceptual groups" are. According to several psychophysiological experiments ([11], [3] and [20]), it seems that there is a strong link between the general *a contrario* approach using a uniform i.i.d. background noise model and human visual perception. Therefore, in the following we define the perceptual groups of points in an original set of points $s^0 = \{x_j^0\}_{1 \le j \le n}$ as being the (ε, U)-meaningful regions for s^0. We recall that the U-meaningfulness of a region R for a set $s^0 = \{x_j^0\}_{1 \le j \le n}$ of points in $D = [0, 1]^2$ is measured by the Number of False Alarms given by

$$\mathrm{NFA}_U(R; s^0) := N_r \times B(n, k(R; s^0), |R|), \quad \text{where} \quad k(R; s^0) = \sum_{j=1}^{n} \mathbb{1}_{x_j^0 \in R}.$$

When $\mathrm{NFA}_U(R; s^0) < \varepsilon$, the region R is (ε, U)-meaningful and it is a perceptual group for s^0. The smaller the value of $\mathrm{NFA}_U(R; s^0)$ is, the more meaningful and the more perceptually significant the region is.

We can now mathematically define the set of distributions on points that generate, most of the time, the same perceptual groups as the initial set of points $s_0 = \{x_j^0\}_{1 \le j \le n}$

Definition 2. *Let $s^0 = \{x_j^0\}_{1 \le j \le n}$ be a set of n points in the domain $D = [0, 1]^2$. Let $\mathcal{R} = \{R_i\}_{1 \le i \le N_r}$ be a set of test regions. We then denote by $\mathcal{Q}$ the set of probability distributions Q on sets $S = \{X_1, \ldots, X_n\}$ of n points in D such that for any region R_i, $1 \le i \le N_r$,*

- *If $\mathrm{NFA}_U(R_i; s^0) < \varepsilon$ (i.e. if the region is (ε, U)-meaningful for the initial set s^0), then*

$$\mathrm{Med}_Q(\mathrm{NFA}_U(R_i; S)) \le \mathrm{NFA}_U(R_i; s^0),$$

  where Med_Q denotes the median value under the distribution Q on the random variable S.

- *Otherwise, if $\mathrm{NFA}_U(R_i; s^0) \ge \varepsilon$ (i.e. if the region is not (ε, U)-meaningful for the initial set s0), then*

$$\mathrm{Med}_Q(\mathrm{NFA}_U(R_i; S)) \ge \varepsilon.$$

This means that when the points $\{X_1, \ldots, X_n\}$ are sampled from a distribution in $\mathcal{Q}$, then in the majority of cases, the (ε, U)-meaningful regions R_i for s^0 are at least as meaningful, whereas non-meaningful regions remain non-meaningful.

Let us comment on the fact that in the above definition we use the median value instead of the expectation. The reason for this is that if Y is a real valued random variable and if f is any increasing function then we have $\mathrm{Med}(f(Y)) = f(\mathrm{Med}(Y))$, whereas this is not true for the expectation. This implies in particular that, in our case, the above definition of $\mathcal{Q}$ is invariant to the fact that the meaningfulness is measured by NFA_U, or $\log(\mathrm{NFA}_U)$, or any other increasing function of NFA_U.

We now characterize the maximum entropy distribution in $\mathcal{Q} \cap \mathcal{I}$, in a way similar to what we did in the previous section with $\mathcal{P} \cap \mathcal{I}$.

Theorem 2. *Assume that $\varepsilon/N_r \leq 1/2$ and that the regions $\{R_i\}_{1\leq i\leq N_r}$ are a partition of the domain D. Let $R_1, \dots R_d$ denote the (ε, U)-meaningful regions for s^0. Then, we have that $\mathcal{Q}\cap\mathcal{I} \neq \emptyset$ and the maximum entropy distribution in $\mathcal{Q}\cap\mathcal{I}$, denoted by Q_0, is given by: its probability density $\tilde{f}_{Q_0}$ is constant on each R_i, and*

$$p_{Q_0}(R_i) := \int_{R_i} \tilde{f}_{Q_0}(x)\,dx = \begin{cases} B^{-1}_{n,k(R_i;s^0)}(\frac{1}{2}) & \text{if } 1 \leq i \leq d \\ \frac{1-\sum_{j=1}^{d} B^{-1}_{n,k(R_j;s^0)}(\frac{1}{2})}{1-\sum_{j=1}^{d}|R_j|}\cdot|R_i| & \text{if } i > d. \end{cases} \tag{9}$$

In particular when $d = 0$, i.e. when there are no (ε, U)-meaningful regions for s^0, then the maximum entropy distribution in $\mathcal{Q}$ is the uniform distribution U.

Proof. By definition of $Q \in \mathcal{Q}$ and of the median value, we have that, for $1 \leq i \leq d$,

$$\begin{aligned} Q \in \mathcal{Q} \quad &\iff \quad \begin{cases} \mathbb{P}_Q[\mathrm{NFA}_U(R_i;S) \leq \mathrm{NFA}_U(R_i;s^0)] \geq \frac{1}{2} & \text{for } 1 \leq i \leq d \\ \mathbb{P}_Q[\mathrm{NFA}_U(R_i;S) \leq \varepsilon] \leq \frac{1}{2} & \text{for } i > d \end{cases} \\ &\iff \quad \begin{cases} \mathbb{P}_Q[k(R_i;S) \geq k(R_i;s^0)] \geq \frac{1}{2} & \text{for } 1 \leq i \leq d \\ \mathbb{P}_Q[k(R_i;S) \geq k_{min}(R_i)] \leq \frac{1}{2} & \text{for } i > d \end{cases} \end{aligned}$$

where $k_{min}(R_i) := \min\{0 \leq k \leq n; B(n,k,|R_i|) \leq \varepsilon/N_r\}$ is the minimal number of points that have to be in R_i in order to make it (ε, U)-meaningful.
Now, when Q also belongs to $\mathcal{I}$, the points are independent identically distributed and therefore $k(R_i;S)$ follows a binomial distribution of parameters n and $p_Q(R_i)$. This implies that for any k integer, $\mathbb{P}_Q[k(R_i;S) \geq k] = B(n,k,p_Q(R_i))$. Using the fact that $p \mapsto B(n,k,p)$ is an increasing function, we have

$$Q \in \mathcal{Q}\cap\mathcal{I} \iff \begin{cases} p_Q(R_i) \geq B^{-1}_{n,k(R_i;s^0)}(\frac{1}{2}) & \text{for } 1 \leq i \leq d \\ p_Q(R_i) \leq B^{-1}_{n,k_{min}(R_i)}(\frac{1}{2}) & \text{for } i > d. \end{cases}$$

Finally, thanks to Lemma 2 applied with the constraints $\alpha_i = B^{-1}_{n,k(R_i;s^0)}(\frac{1}{2})$ for $1 \leq i \leq d$ and $\alpha_i = 0$ for $i > d$, we have the announced formula for the maximum entropy distribution Q_0 in $\mathcal{Q}\cap\mathcal{I}$ satisfying the first set of constraints (the ones on meaningful regions).

Now, notice that since for $1 \leq i \leq d$, we have $B(n,k(R_i;s^0),|R_i|) \leq \frac{\varepsilon}{N_r} \leq \frac{1}{2}$, this implies $B^{-1}_{n,k(R_j;s^0)}(\frac{1}{2}) \geq |R_i|$ for $1 \leq i \leq d$, and therefore $\gamma := \frac{1-\sum_{j=1}^{d} B^{-1}_{n,k(R_j;s^0)}(\frac{1}{2})}{1-\sum_{j=1}^{d}|R_j|} \leq 1$. Consequently, for $i > d$, we have $\gamma|R_i| \leq |R_i| \leq B^{-1}_{n,k_{min}(R_i)}(\frac{1}{2})$ (because by definition $B(n,k_{min}(R_i),|R_i|) \leq \varepsilon/Nr \leq \frac{1}{2}$), which means that the second set of constraints (on non-meaningful regions) is automatically satisfied. Therefore the probability distribution Q_0 defined by (9) is the maximum entropy distribution in $\mathcal{Q}\cap\mathcal{I}$. □

On Figure 4, we illustrate Theorem 2. On the top left, we show again the original set of points s^0 together with the test regions $\{R_i\}_{1\leq i\leq N_r}$, and on the top right we show the image of the values of $p_{Q_0}(R_i)$. In particular we have $p_{Q_0}(R_1) = B^{-1}_{100,12}(\frac{1}{2}) \simeq 0.116$, $p_{Q_0}(R_2) = B^{-1}_{100,7}(\frac{1}{2}) \simeq 0.066$ and $p_{Q_0}(R_i) = 0.01 \times (1 - B^{-1}_{100,12}(\frac{1}{2}) - B^{-1}_{100,7}(\frac{1}{2}))/(1-0.02) \simeq 0.0083$ for $i > 2$ (all these values have to be compared to the value $0.01 = |R_i| = p_U(R_i)$). Then on the bottom we show two samples from Q_0. These two samples of $n = 100$ points are respectively denoted by s^1 and s^2. In the left sample, we have two (ε, U)-meaningful regions: the region R_1 with 11 points has $\log_{10}(\mathrm{NFA}_U(R_1,s^1) \simeq -6.2$ (we had $\log_{10}(\mathrm{NFA}_U(R_1,s^0)) \simeq -7.3$ in the original set of points s^0) and the second region R_2 with 7 points has $\log_{10}(\mathrm{NFA}_U(R_2,s^1) \simeq -2.1$ (we had also exactly the same value $\log_{10}(\mathrm{NFA}_U(R_2,s^0))$ in the original set of points s^0). For the right sample, we also have two (ε, U)-meaningful regions: the region R_1 with 19 points has now $\log_{10}(\mathrm{NFA}_U(R_1,s^2) \simeq -16.2$ and the second region R_2 with 8 points has here $\log_{10}(\mathrm{NFA}_U(R_2,s^2) \simeq -3.1$.

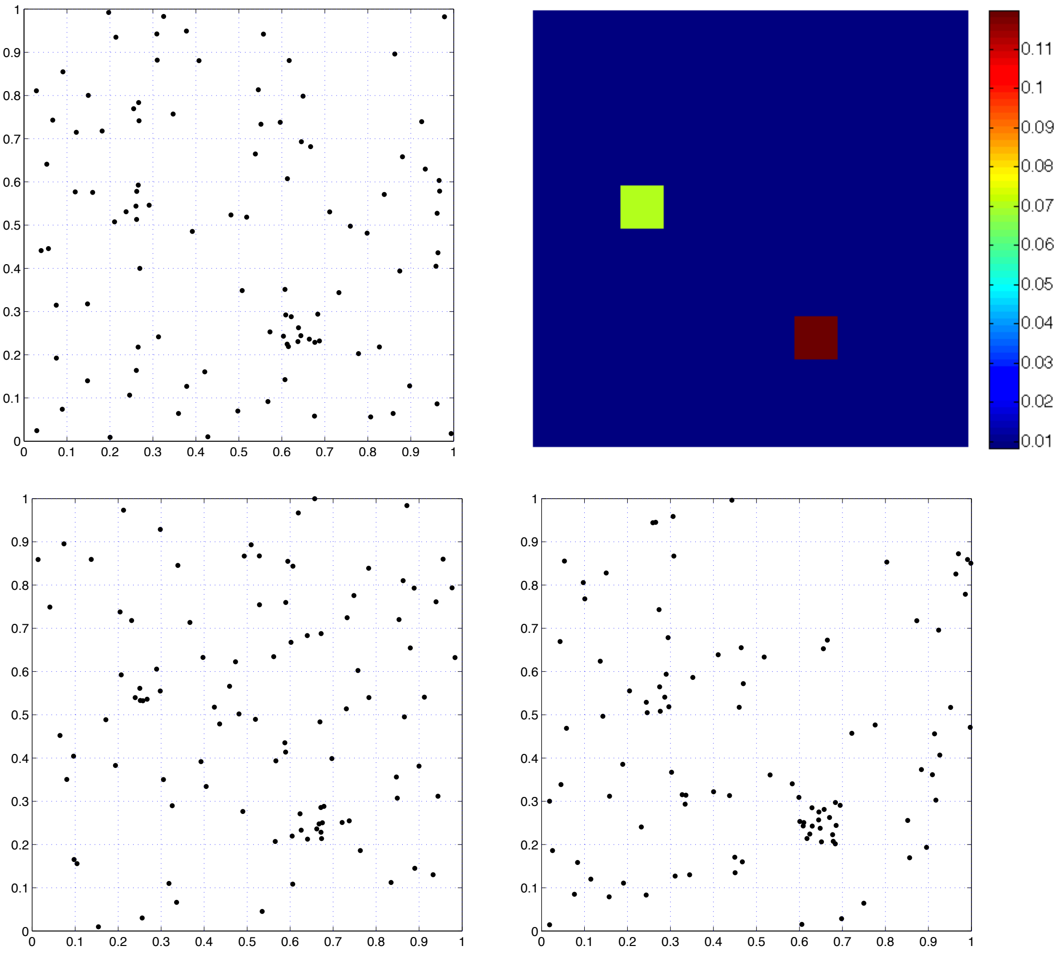


Figure 4: Top left: the original set of points s^0 with the regions R_i as squares delimited by the dashed lines. Top right: values of $p_{Q_0}(R_i)$ where Q_0 is the maximum entropy distribution in $\mathcal{Q} \cap \mathcal{I}$ given by Equation (9). Bottom : two samples of $n = 100$ points from the probability distribution Q_0. These two samples have the same (ε, U)-meaningful regions as the original set of points S_0. See the text for more details about the NFA_U exact values for these regions. One can clearly perceive in the two bottom figures perceptual groups that are similar to the ones of the original set of points.

Now, a natural question is: what is the link between the sets $\mathcal{P}$ and $\mathcal{Q}$? When $\varepsilon/N_r \leq 1/2$, we have that if $P \in \mathcal{Q}$, then for a (ε, U)-meaningul region R_i of S_0, $1 \leq i \leq d$,

$$\mathbb{P}_P[k(R_i; S) \geq k(R_i; s^0)] \geq \frac{1}{2} \geq \frac{\varepsilon}{N_r},$$

and therefore $P \in \mathcal{P}_{R_i}$. This is quite logical: if a distribution P is such that, in the majority of cases, samples from it have the same meaningful regions as the original set of points s^0, then these regions are not P-meaningful (they are not rare events under the law P).
Now, for regions R_i, $i > d$, that are not (ε, U)-meaningful for s^0, the situation is quite different. Indeed it may happen that $P \in \mathcal{Q}$ and at the same time $\mathrm{NFA}_P(R_i; s^0) < \varepsilon$. It is for instance the case when P makes the points independent and such that $p_P(R_i)$ is very small (i.e. smaller than $B^{-1}_{n,k_{min}(R_i)}(1/2)$ and than $B^{-1}_{n,k(R_i;s^0)}(\varepsilon/N_r)$).

The conclusion of this section is that we have been able to turn the *a contrario* detection approach (that is originally a hypothesis testing method with multiple tests) into a generative approach. To do this, we have defined a "canonical" way to associate to any set of points $s^0 = \{x^0_j\}_{1 \leq j \leq n}$ a probability distribution Q_0 on sets of n points that

- makes the points independent;
- belongs to $\mathcal{Q}$ (i.e. points sampled from Q_0 have, most of the time, the same perceptual groups as the initial set s^0);
- and is at the same time as "random" as possible, in the sense that it has maximum entropy.

The approach developed in this section is quite general and can be applied to other detection problems that are formalized in terms of grouping laws. We will in particular, in the next section, see how to extend it to the framework of line segments detection in an image.

3 Line segments in an image

In this second part of the paper, we will be interested in another application of the *a contrario* approach, that is line segment detection in an image, as performed by the LSD algorithm of Grompone, Jakubowicz, Morel and Randall in [15] and [16]. This algorithm is a great generalization of the alignment detection method of [9]. It can be easily tested on any image thanks to the online demo on the Image Processing Online Journal (IPOL) at `http://demo.ipol.im/demo/gjmr_line_segment_detector/`.

We first recall the general framework of the LSD algorithm (with slight modifications and simplifications) and we introduce some notations.

3.1 Framework and notations

Given a grey level image I^0 defined on a discrete domain $\Omega = \{1, \ldots, M\} \times \{1, \ldots, N\}$, we compute its orientation field $\theta^0 : \Omega \to S^1 = [0, 2\pi)$ as the orientation field of the level lines of I^0, or, that is equivalent, by

$$\forall x \in \Omega, \quad \theta^0(x) = \frac{\pi}{2} + \mathrm{Arg}\frac{\nabla I^0(x)}{\|\nabla I^0(x)\|},$$

where ∇I^0 is the gtradient of I^0.

For a rectangle r in Ω (that is a set of pixels in Ω that belong to an underlying "true" rectangle in the continuous domain, see Figure 5 for an illustration), we define its orientation $\varphi(r)$ as the orientation of one of its axis. The number of aligned points in θ^0 it contains, according to a given precision p, is defined by

$$k(r; \theta^0) := \sum_{x \in r} \mathbb{1}_{|\theta^0(x) - \varphi(r)| \leq p\pi}.$$

We will also denote by $N_p = \#\Omega = M \cdot N$ the total number of pixels in Ω and by $n(r) = \#r$ the number of pixels in a rectangle r. For an orientation field $\theta : \Omega \to S^1$, we will denote its value at a pixel x by $\theta(x)$ or by θ_x.

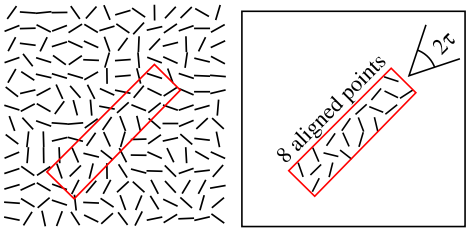


Figure 5: Example of an orientation field (elementary small segments), a rectangle r (delimited in red), and its aligned points (for a precision p, hence here on the figure $\tau = p\pi$). Figure courtesy of R. Grompone von Gioi.

Given a probability distribution P on a random orientation field $\Theta : \Omega \to S^1$, the Number of False Alarms of r in θ^0 under the noise model P is defined by

$$\mathrm{NFA}_P(r; \theta^0) = N_{tests} \cdot \mathbb{P}_P[k(r; \Theta) \geq k(r; \theta^0)], \tag{10}$$

where N_{tests} is the number of tests, that is the number of rectangles in an image of size $M \times N$. It is of the order of $(MN)^{5/2}$ (see [16]). In the definition (10), Θ is a random orientation field that follows the distribution P and $k(r; \Theta)$ is therefore a random variable defined by

$$k(r; \Theta) := \sum_{x \in r} \mathbb{1}_{|\Theta(x) - \varphi(r)| \leq p\pi}.$$

In the LSD algorithm, the Number of False Alarms is computed with the uniform independent distribution as a noise model (*i.e.* assuming the $\Theta(x)$ are independent and uniformly distributed on S^1). The LSD algorithm performs a region growing process, with a validation step, resulting in a list of U-meaningful rectangles $r_1, \ldots, r_m$. This is a short and incomplete description of the full algorithm (there is in particular a scaling step that helps to cope with aliasing and quantization artifacts), but we don't need here to include all the improvement steps of the full algorithm.

Let us notice that in the original LSD algorithm, the precision p of the alignments is not fixed, several precisions are tested. However since we have observed in experiments on several images that most of the rectangles (about 95% of them) are obtained for the precision value $p = 1/8$, and in order to make the ideas and computations clearer and simpler, we will in all the following assume that there is only one tested precision p that is fixed to $p = 1/8$.

Let us also emphasize that the resulting meaningful rectangles of the LSD algorithm are "almost" disjoint. Indeed, we have noticed on several experiments that the percentage of pixels that belong to two meaningful rectangles at the same time is less than 1%. This is a very nice feature of the LSD algorithm: because it works like a growing process, it does not need a "maximality" or "optimality" step as it is the case of many *a contrario* detection methods (the original meaningful alignments of [9], or the meaningful boundaries [10], etc.). In all the following we will assume that the resulting rectangles of the LSD are disjoint. This assumption will help to simplify the mathematical results, while not affecting the practical results.

3.2 Enriching the a contrario noise model

In a way similar to what we did for clusters of points in the first part of the paper, we here again question the choice of the noise model P on the orientation field Θ. And in particular we introduce the following sets of probability distributions.

Definition 3. *Let $\theta^0 : \Omega \to S^1$ be an orientation field. Let $r_1, \dots, r_m$ be the resulting disjoint U-meaningful rectangles of the LSD algorithm on θ^0. We then define the three following sets of distributions:*

- *Let $\mathcal{P}$ denote the set of probability distributions P on Θ such that none of the rectangles $r_1, \dots, r_m$ are (ε, P)-meaningful in θ^0, i.e.*

$$P \in \mathcal{P} \iff \forall 1 \leq j \leq m, \ \ \mathrm{NFA}_P(r_j; \theta^0) \geq \varepsilon.$$

- *Let $\mathcal{Q}$ denote the set of probability distributions P on Θ such that if the orientation field is sampled from P then, in most cases, the meaningul rectangles will be at least as meaningful as in θ^0.*

$$Q \in \mathcal{Q} \iff \forall 1 \leq j \leq m, \ \ \mathrm{Med}_Q(\mathrm{NFA}_U(r_j; \Theta)) \leq \mathrm{NFA}_U(r_j; \theta^0).$$

- *Finally, let $\mathcal{I}$ denote the set of probability distributions on Θ such that the $\Theta(x)$ are independent (but not necessarily identically distributed).*

We will also here again be interested in distributions in $\mathcal{P}$ (or $\mathcal{Q}$) that have maximal entropy. Notice however that we are not exactly in the same framework as in the case of clusters of points. To see this, let us look at the formula of $\mathrm{NFA}_U(r; \theta^0)$. When the orientations are independent and uniformly distributed on the circle S^1, the law of $k(r; \Theta)$ is the binomial distribution of parameters $n(r)$ and p. And this is a very different situation from the clusters of points where the binomial distribution involved under the uniform independent noise model was of parameters n (total number of points, fixed) and $|R|$ (size of the region R). Therefore we can not directly apply the propositions and theorems of the first part of the paper. Now, even if the situations are different, we have “analogous” results.

Proposition 3. *1. The distribution P in $\mathcal{P}$ that has maximal entropy admits a probability density f_P given by*

$$f_P(\theta) = \frac{1}{(2\pi)^{N_p}} \prod_{j=1}^{m} a_j^{h_j(\theta)} \tilde{a}_j^{1-h_j(\theta)}, \tag{11}$$

where h_j is the function defined by $h_j(\theta) = 1$ if $k(r_j; \theta) \geq k(r_j; \theta^0)$ and 0 otherwise (notice that in fact h_j is only a function of $\theta(x)$, $x \in r_j$) and where

$$a_j = \frac{\varepsilon}{N_{tests} \cdot B(n(r_j), k(r_j; \theta^0), p)} \quad \text{and} \quad \tilde{a}_j = \frac{1 - \varepsilon / N_{tests}}{1 - B(n(r_j), k(r_j; \theta^0), p)}.$$

2. The distribution P_0 in $\mathcal{P} \cap \mathcal{I}$ defined by $f_{P_0}(\theta) = \prod_{x=1}^{N_p} f_{P_0}^{(x)}(\theta_x)$ with

$$f_{P_0}^{(x)}(\theta_x) = \begin{cases} \frac{1}{2\pi} & \text{if } x \notin \cup_{j=1}^m r_j \\ \frac{1}{2p\pi} B^{-1}_{n(r_j), k(r_j;\theta^0)}\left(\frac{\varepsilon}{N_{tests}}\right) & \text{if } x \in r_j \text{ and } |\theta_x - \varphi(r_j)| \leq p\pi \\ \frac{1}{2(1-p)\pi}\left(1 - B^{-1}_{n(r_j), k(r_j;\theta^0)}\left(\frac{\varepsilon}{N_{tests}}\right)\right) & \text{if } x \in r_j \text{ and } |\theta_x - \varphi(r_j)| > p\pi \end{cases}$$

is a local maximum of the entropy in $\mathcal{P} \cap \mathcal{I}$.

Proof. A probability distribution P on Θ belongs to $\mathcal{P}$ if, by definition, for all regions r_j, $1 \leq j \leq m$, we have $\mathbb{P}_P[k(r_j; \Theta) \geq k(r_j; \theta^0)] \geq \frac{\varepsilon}{N_{tests}}$. To such a probability distribution P we associate the probability distribution P_a defined by

$$f_{P_a}(\theta) = \frac{1}{(2\pi)^{N_p}} \prod_{j=1}^{m} a_j^{h_j(\theta)} \tilde{a}_j^{1-h_j(\theta)},$$

where h_j is the function defined by $h_j(\theta) = 1$ if $k(r_j; \theta) \geq k(r_j; \theta^0)$ and 0 otherwise; and with a_j and $\tilde{a}_j$ respectively given by $a_j B(n(r_j), k(r_j; \theta^0), p) = \mathbb{P}_P[k(r_j; \Theta) \geq k(r_j; \theta^0)]$ and $\tilde{a}_j(1 -$

$B(n(r_j), k(r_j;\theta^0), p)) = 1 - a_j B(n(r_j), k(r_j;\theta^0), p)$. The condition $P \in \mathcal{P}$ then simply becomes $a_j B(n(r_j), k(r_j;\theta^0), p) \geq \varepsilon / N_{tests}$. In a way similar to the proof of Proposition 2, we first prove that $H(P_a) \geq H(P)$. To see this we compute the Kullback-Leibler divergence $D(P||P_a)$:

$$
\begin{aligned}
D(P||P_a) &= \int_{[0,2\pi)^{N_p}} f_P(\theta) \log \frac{f_P(\theta)}{f_{P_a}(\theta)} d\theta_1 \dots d\theta_{N_p} = -H(P) - \int_{[0,2\pi)^{N_p}} f_P(\theta) \log f_{P_a}(\theta) d\theta_1 \dots d\theta_{N_p} \\
&= -H(P) + N_p \log(2\pi) - \sum_{j=1}^{m} \mathbb{P}_P[k(r_j;\Theta) \geq k(r_j;\theta^0)] \log a_j + (1 - \mathbb{P}_P[k(r_j;\Theta) \geq k(r_j;\theta^0)]) \log \tilde{a}_j \\
&= -H(P) + N_p \log(2\pi) - \sum_{j=1}^{m} a_j B(n(r_j), k(r_j;\theta^0), p) \log a_j + \tilde{a}_j (1 - B(n(r_j), k(r_j;\theta^0), p)) \log \tilde{a}_j \\
&= -H(P) + H(P_a),
\end{aligned}
$$

because we have that

$$
\int_{[0,2\pi)^{N_p}} \mathbb{1}_{h_j(\theta)=1} \, d\theta_1 \dots d\theta_{N_p} = (2\pi)^{N_p} B(n(r_j), k(r_j;\theta^0), p).
$$

Then, since $D(P||P_a) \geq 0$, this proves that $H(P_a) \geq H(P)$. Finally, a simple study of the function $t \mapsto -t \log \frac{t}{B} - (1-t) \log \frac{1-t}{1-B}$ shows that it is decreasing when $t \geq \varepsilon / N_{tests}$, where here $B := B(n(r_j), k(r_j;\theta^0), p)$ is such that $B < \varepsilon / N_{tests}$ (because the rectangle r_j is (ε, U)-meaningful in θ^0). Therefore, the maximum entropy is achieved for P_a when $a_j B(n(r_j), k(r_j;\theta^0), p) = \varepsilon / N_{tests}$ for all $1 \leq j \leq m$.

For the second part of the proposition, let us look for the probability distribution $P_0 \in \mathcal{P} \cap \mathcal{I}$ that has maximal entropy. Since $P_0 \in \mathcal{I}$, its probability density f_{P_0} is a product and therefore

$$
H(P_0) = -\sum_{x=1}^{N_p} \int_0^{2\pi} f_{P_0}^{(x)}(\theta_x) \log f_{P_0}^{(x)}(\theta_x) \, d\theta_x.
$$

When $x \notin \cup_{j=1}^m r_j$, there is no constraint on $f_{P_0}^{(x)}$ and therefore the entropy of $f_{P_0}^{(x)}$ is maximal when it is the uniform distribution on $[0, 2\pi)$.
Now, let r_j be a rectangle and consider all $x \in r_j$. Then since $P_0 \in \mathcal{P}$, we have the constraint

$$
\mathbb{P}_{P_0}[k(r_j;\Theta) \geq k(r_j;\theta^0)] \geq \frac{\varepsilon}{N_{tests}}, \quad \text{where } k(r_j;\Theta) = \sum_{x \in r_j} \mathbb{1}_{|\Theta_x - \varphi(r_j)| \leq p\pi}.
$$

Under the law P_0, since the Θ_x are independent (but nor necessarily identically distributed), the random variable $k(r_j;\Theta)$ follows a so-called Poisson binomial distribution of parameters $\{q_x\}_{x \in r_j}$ where

$$
q_x := \mathbb{P}_{P_0}[|\Theta_x - \varphi(r_j)| \leq p\pi] = \int_{\varphi(r_j)-p\pi}^{\varphi(r_j)+p\pi} f_{P_0}^{(x)}(\theta_x) \, d\theta_x.
$$

Then by considering the Kullack-Leibler divergence between $f_{P_0}^{(x)}$ and the distribution on θ_x that is constant equal to $q_x / 2p\pi$ on $[\varphi(r_j) - p\pi, \varphi(r_j) + p\pi]$ and equal to $(1 - q_x)/2(1-p)\pi$ on the remainder part of $[0, 2\pi)$, we have that

$$
H(f_{P_0}^{(x)}) \leq \log(2\pi) - q_x \log \frac{q_x}{p} - (1 - q_x) \log \frac{1 - q_x}{1 - p},
$$

with equality when the two distributions are equal. Then the statement of the second part of the proposition follows from the following lemma.

Lemma 3. *Let $n \geq 1$ be an integer, let $p \in (0,1)$ and let H_p be the function defined on $[0,1]^n$ by $H_p(q_1, \dots, q_n) = -\sum_{i=1}^n q_i \log \frac{q_i}{p} + (1 - q_i) \log \frac{1-q_i}{1-p}$. Let $1 < k_0 \leq n$ be an integer, and let C be the function defined on $[0,1]^n$ by $C(q_1, \dots, q_n) = \mathbb{P}[K \geq k_0]$, where K follows a Poisson binomial* *distribution of parameters $q_1,\dots,q_n$. We also make the assumption that n is large enough, and p not too small, in such a way that $1+\frac{1}{n+1}\log\frac{p}{1-p}>0$. Finally let $\eta\in(0,1/2]$ be such that $B(n,k_0,p)<\eta$. Then, under the constraint $C(q_1,\dots,q_n)\geq\eta$, a local maximum of H_p is achieved when all the q_i are equal to $\overline{q}:=B^{-1}_{n,k_0}(\eta)$.*

Let us first comment on the assumption that $1+\frac{1}{n+1}\log\frac{p}{1-p}>0$. This assumption is equivalent to $p>1/(1+e^{n+1})$. It will be always satisfied in our case since we take in practice $p=1/8$ and n is always larger than 1, having thus $1+e^{n+1}\geq 1+e^2\simeq 8.4$.

The proof of the lemma is postponed to the Appendix. It is based on the computations of the gradient and the Hessian of a symmetric function (symmetric means here invariant under any permutation of the variables) under a symmetric constraint. Actually, we conjecture that the point $(\overline{q},\dots,\overline{q})$ is in fact the point of global maximum of H_p under the constraint $C\geq\eta$. Now, this is not obvious since symmetric functions under symmetric constraint do not necessarily have a global maximum at a symmetric point. This is for instance discussed in the paper of Waterhouse [31] where he shows that symmetric points are local extrema but where he also gives examples where the global maximum of a symmetric function is not a symmetric point. However, we believe that here in our case the point $(\overline{q},\dots,\overline{q})$ is indeed a global maximum. We have in particular checked this when $n=2$. We also think that it is related to the result of Harremoës in [18] where he proves that the binomial distribution is the maximum entropy distribution on suitably defined sets.

□

We show examples of samples from P_0 on Figures 6, 7 and 8. As in the case of clusters, it seems that these samples are not visually very "close" to θ^0 in terms of line segment detection. And this is again explained by the same phenomenon: in hypothesis testing, it is not because you don't reject a distribution that this distribution fits your data.

Now, for the set $\mathcal{Q}$, the situation will be different since this set of probability distributions is build on purpose to contain, in most cases, the same meaningful line segments (rectangles) as in the original orientation field. We first state the proposition.

Proposition 4. *1. The distribution Q in $\mathcal{Q}$ that has maximal entropy admits a probability density f_Q given by*

$$f_Q(\theta)=\frac{1}{(2\pi)^{N_p}}\prod_{j=1}^{m}b_j^{h_j(\theta)}\tilde{b}_j^{1-h_j(\theta)},\qquad(12)$$

where h_j is the function defined by $h_j(\theta)=1$ if $k(r_j;\theta)\geq k(r_j;\theta^0)$ and 0 otherwise (notice that in fact h_j is only a function of θ_x, $x\in r_j$) and where

$$b_j=\frac{1}{2B(n(r_j),k(r_j;\theta^0),p)}\quad\text{and}\quad\tilde{b}_j=\frac{1}{2(1-B(n(r_j),k(r_j;\theta^0),p))}.$$

2. The probability distribution Q_0 in $\mathcal{Q}\cap\mathcal{I}$ given by $f_{Q_0}(\theta)=\prod_{x=1}^{N_p}f_{Q_0}^{(x)}(\theta_x)$ with

$$f_{Q_0}^{(x)}(\theta_x)=\begin{cases}\frac{1}{2\pi} & \text{if } x\notin\cup_{j=1}^m r_j\\ \frac{1}{2p\pi}B^{-1}_{n(r_j),k(r_j;\theta^0)}(\frac{1}{2}) & \text{if } x\in r_j \text{ and } |\theta_x-\varphi(r_j)|\leq p\pi\\ \frac{1}{2(1-p)\pi}(1-B^{-1}_{n(r_j),k(r_j;\theta^0)}(\frac{1}{2})) & \text{if } x\in r_j \text{ and } |\theta_x-\varphi(r_j)|>p\pi\end{cases}$$

is a local maximum of the entropy in $\mathcal{Q}\cap\mathcal{I}$.

Proof. The proof of the proposition is analogous to the proof of the previous one. Indeed, we notice that, starting from the definition of $\mathcal{Q}$ and of the median, we get

$$\begin{aligned}Q\in\mathcal{Q}&\iff\forall 1\leq j\leq m,\ \ \mathrm{Med}_Q(\mathrm{NFA}_U(r_j;\Theta))\leq\mathrm{NFA}_U(r_j;\theta^0)\\&\iff\forall 1\leq j\leq m,\ \ \mathbb{P}_Q[\mathrm{NFA}_U(r_j;\Theta)\leq\mathrm{NFA}_U(r_j;\theta^0)]\geq\frac{1}{2}\\&\iff\forall 1\leq j\leq m,\ \ \mathbb{P}_Q[k(r_j;\Theta)\geq k(r_j;\theta^0)]\geq\frac{1}{2}.\end{aligned}$$

This last equivalence has to be compared with the definition of $\mathcal{P}$, where we had: $P \in \mathcal{P}$ if and only if $\mathbb{P}_P[k(r_j;\Theta) \geq k(r_j;\theta^0)] \geq \frac{\varepsilon}{N_{tests}}$. Thanks to this remark, it shows that the proof of the proposition is exactly the same as in the case of $\mathcal{P}$, except that we simply need to replace $B^{-1}_{n(r_j),k(r_j;\theta^0)}(\frac{\varepsilon}{N_{tests}})$ by $B^{-1}_{n(r_j),k(r_j;\theta^0)}(\frac{1}{2})$. Notice that this remark also implies that

$$\mathcal{Q} \subset \mathcal{P}.$$

□

On Figures 6, 7 and 8, we show for each image *Pirée*, *Chairs* and *Valbonne*: the original image I^0, the orientation field θ^0, the rectangles output of the LSD algorithm, a sample from P_0 and a sample from Q_0. Notice on these figures that, as expected, the samples from Q_0 are more structured than the one of P_0, and resemble the original orientation field θ^0 in terms of straight structures perception.

3.3 Image reconstruction from an orientation field

Now that we have samples of orientation fields, the next natural question is : how can we reconstruct an image from an orientation field? We first notice that not all orientation fields are the orientation field of an image. Indeed, only the ones that satisfy a kind of null circulation condition will correspond to conservative fields (and therefore to an image gradient). Notice also that even in the case of an orientation field that comes from an image, there is no unicity since any change of contrast on the image does not modify its orientation field.

We will solve here the question of reconstructing an image from a given orientation field θ by looking for an image that has a gradient orientation field as close as possible to $\theta - \frac{\pi}{2}$ in a sense that we will define. This problem is highly related to the Poisson image editing algorithm of Pérez et al. [28], where they copy the gradients of an image inside a domain in another image and then recover an image without color or boundary artefacts.

Consider an orientation field θ defined on a discrete domain Ω of size $M \times N$ pixels. We then choose an arbitrary amplitude function $R : \Omega \to \mathbb{R}_+$ (the choice of this amplitude function will be discussed later at the end of the section). We define the vector field V on Ω by $V(x) = (v_1(x), v_2(x)) = (R(x)\sin\theta(x), -R(x)\cos\theta(x))$. And we look for an image $u : \Omega \to \mathbb{R}$ such that

$$\sum_{x\in\Omega} \|\nabla u(x) - V(x)\|_2^2 = \sum_{x\in\Omega} \left(\frac{\partial u}{\partial x_1}(x) - v_1(x)\right)^2 + \left(\frac{\partial u}{\partial x_2}(x) - v_2(x)\right)^2 \quad \text{is minimal.}$$

Thanks to Parseval's theorem and to the property that the discrete Fourier transform (DFT) of the partial derivatives of u are related to the DFT of u by

$$\forall \xi = (\xi_1, \xi_2) \in \Omega, \quad \widehat{\frac{\partial u}{\partial x_1}}(\xi) = \frac{2i\pi\xi_1}{M}\widehat{u}(\xi) \quad \text{and} \quad \widehat{\frac{\partial u}{\partial x_1}}(\xi) = \frac{2i\pi\xi_2}{N}\widehat{u}(\xi),$$

the above minimization problem is equivalent to find $\widehat{u}$ such that

$$\sum_{\xi\in\Omega} \left|\frac{2i\pi\xi_1}{M}\widehat{u}(\xi) - \widehat{v_1}(\xi)\right|^2 + \left|\frac{2i\pi\xi_2}{N}\widehat{u}(\xi) - \widehat{v_2}(\xi)\right|^2 \quad \text{is minimal.}$$

This is a very simple quadratic minimization problem, and it is solved by taking

$$\forall \xi \in \Omega \setminus \{0\}, \quad \widehat{u}(\xi) = \frac{\frac{2i\pi\xi_1}{M}\widehat{v_1}(\xi) + \frac{2i\pi\xi_2}{N}\widehat{v_2}(\xi)}{\left(\frac{2i\pi\xi_1}{M}\right)^2 + \left(\frac{2i\pi\xi_2}{N}\right)^2}$$

and $\widehat{u}(0)$ equal to any arbitrary constant (this is the mean value of the reconstructed image u).

Samples of P_0 or Q_0 provide an orientation field θ, and then we have to choose an amplitude field R. We have investigated three possible choices:

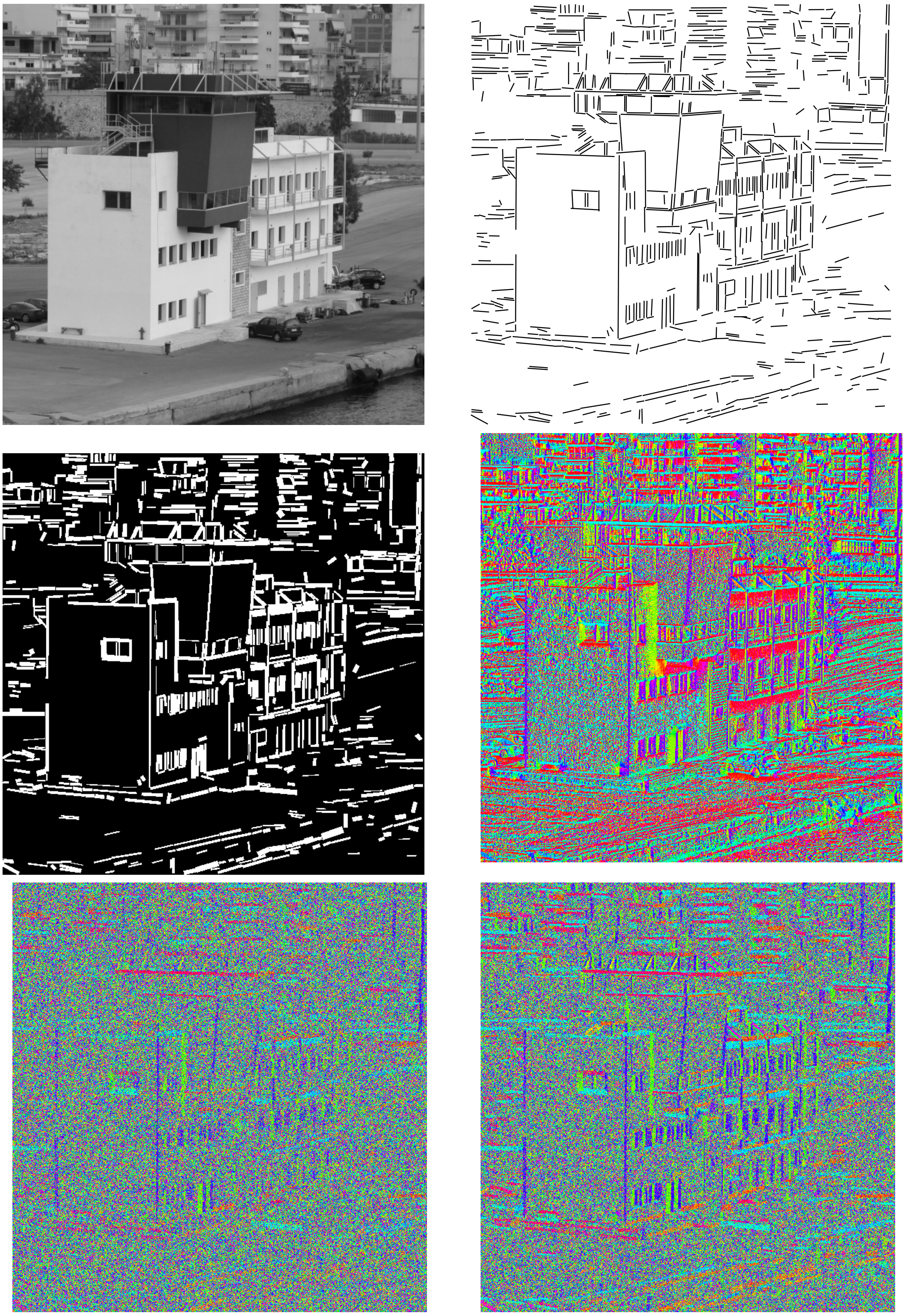


Figure 6: From left to right, top to bottom: the original *Pirée* image I^0 of size 600×600 pixels, the output of the LSD, the whole output rectangles of the LSD, the orientation field θ^0, a sample from P_0 and a sample from Q_0.

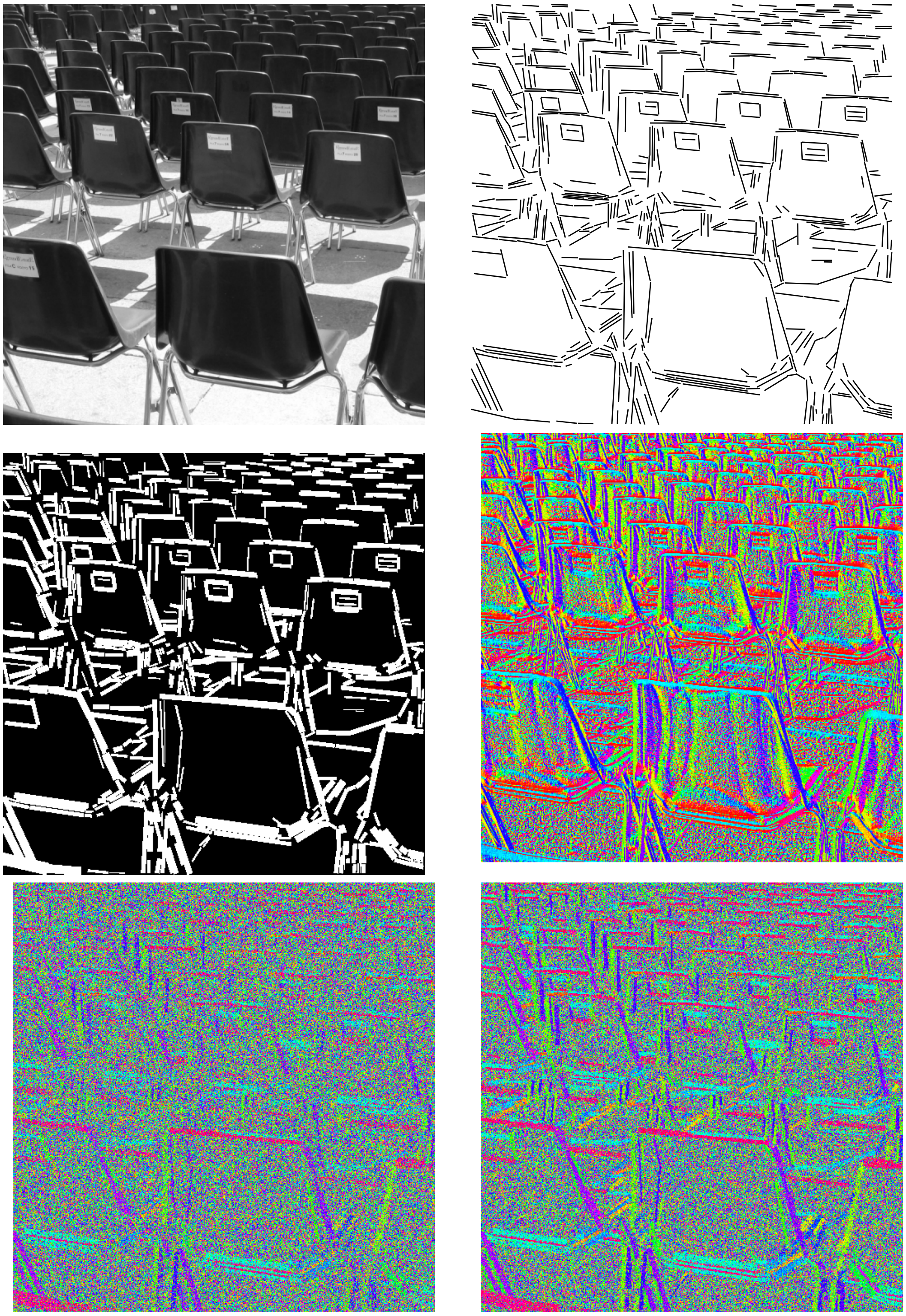


Figure 7: From left to right, top to bottom: the original *Chairs* image I^0 of size 512×512 pixels, the output of the LSD, the whole output rectangles of the LSD, the orientation field θ^0, a sample from P_0 and a sample from Q_0.

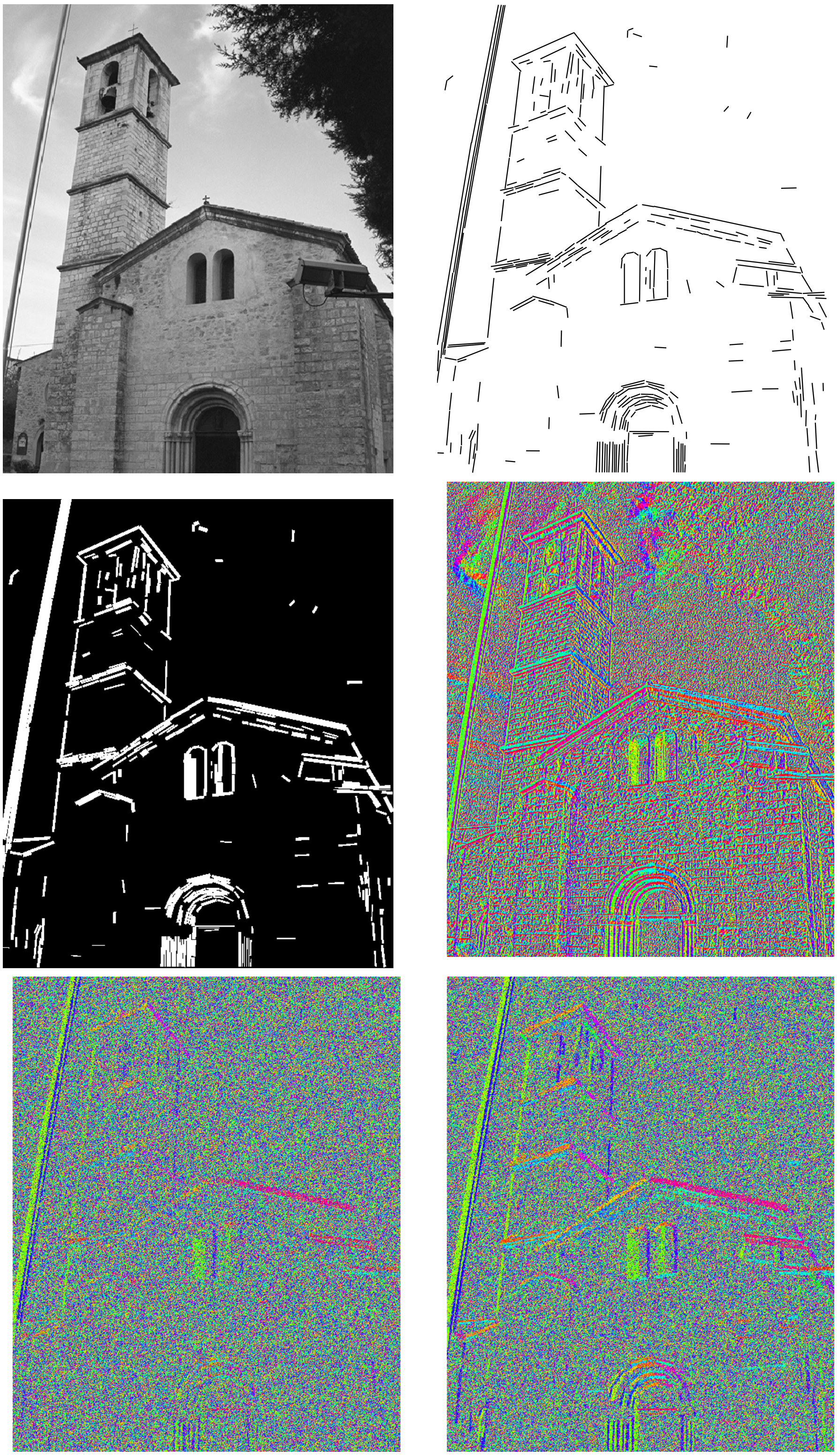


Figure 8: From left to right, top to bottom: the original *Valbonne* image I^0 of size 539×648 pixels, the output of the LSD, the whole output rectangles of the LSD, the orientation field θ^0, a sample from P_0 and a sample from Q_0.

1. R_{rand} that is obtained by taking the $R(x), x \in \Omega$, independent identically distributed uniformly on $[0, 1]$.

2. R_{cst} that is simply constant, *i.e.* $R(x) = 1$ for all $x \in \Omega$.

3. R_{100} that is defined by $R(x) = 100$ for $x \in \cup_{j=1}^{m} r_j$ (where the r_j, $1 \leq j \leq m$, are the output rectangles of the LSD) and $R(x) = 1$ otherwise.

The idea underlying the third choice R_{100} is that the output rectangles of the LSD are generally also edges in the image and the contrast (gradient amplitude) on these rectangles is large. However, we believe that the choice of R is an important question that is beyond the scope of the present paper and that should be studied in some future works.

Examples of images u reconstructed from a sample of P_0 or Q_0 with the three different choices for R are shown on Figures 9, 10 and 11. Notice how these reconstructed images are "close" to the original ones in the sense that they contain the same perceptual straight structures. Notice also how the tree in the *Valbonne* image disappears in the reconstructed images. This is easily explained by the fact that the tree was not detected by the LSD algorithm. This also shows that the proposed reconstruction method could be used for clutter removal.

Now, for any reconstructed image we can apply the LSD algorithm on it and look at what the ouput is. This is illustrated on Figure 12 where the LSD is applied respectively to the image reconstructed from the sample of P_0 and R_{100}, to the image reconstructed from the sample of Q_0 and R_{100} (these two images are the ones of the last row of Figure 9, and we denote them respectively u_{P_0} and u_{Q_0}) and to the original image I^0. As expected, the LSD on u_{P_0} does not produce many segments whereas the ouput segments on u_{Q_0} are almost identical to the ones of the original image I^0. This is a sanity check for the proposed method since the distribution Q_0 is build on purpose to ensure that, in most cases, the LSD output will be the same as in the original image.

4 Discussion, conclusion and future work

In this paper, we have shown how the *a contrario* detection approach can be changed into a generative approach enabling the generation of new images that have the same perceptual content as an original given image. We have extensively used the maximum entropy principle as a way to get probability distributions that satisfy some visual detection constraints while being at the same time as random as possible. In that sense this paper belongs to the field of what could be called *visual information theory.*

This paper answered some questions but it raises also many ones in different directions:

- It raises theoretical question: for instance in the first part, some results were stated making the assumption that the regions are disjoint, but what happens if they are not disjoint? Is is still possible to have explicit formulas for P_0 and Q_0? In the second part, we also have open theoretical questions mainly about P_0 and Q_0 as global maximum entropy distributions (and not only local ones), and also about the image reconstruction process (like for instance: what is the "best" choice for the amplitude R?)

- It will lead to extensions: we have shown how to get generative models from clusters of points and from line segment detection in an image. But we believe other *a contrario* detection frameworks could be explored, for instance edge detection or histogram modes. Then we will have to face the problem of combining several detectors in an common generative model.

- It can have applications: as illustrated by Figure 11, the proposed generative model leads to a clutter removal method. More generally, the generative model is able to capture the perceptual structures (line segments for instance) while making the rest of the image "random". This could lead to applications in image compression.

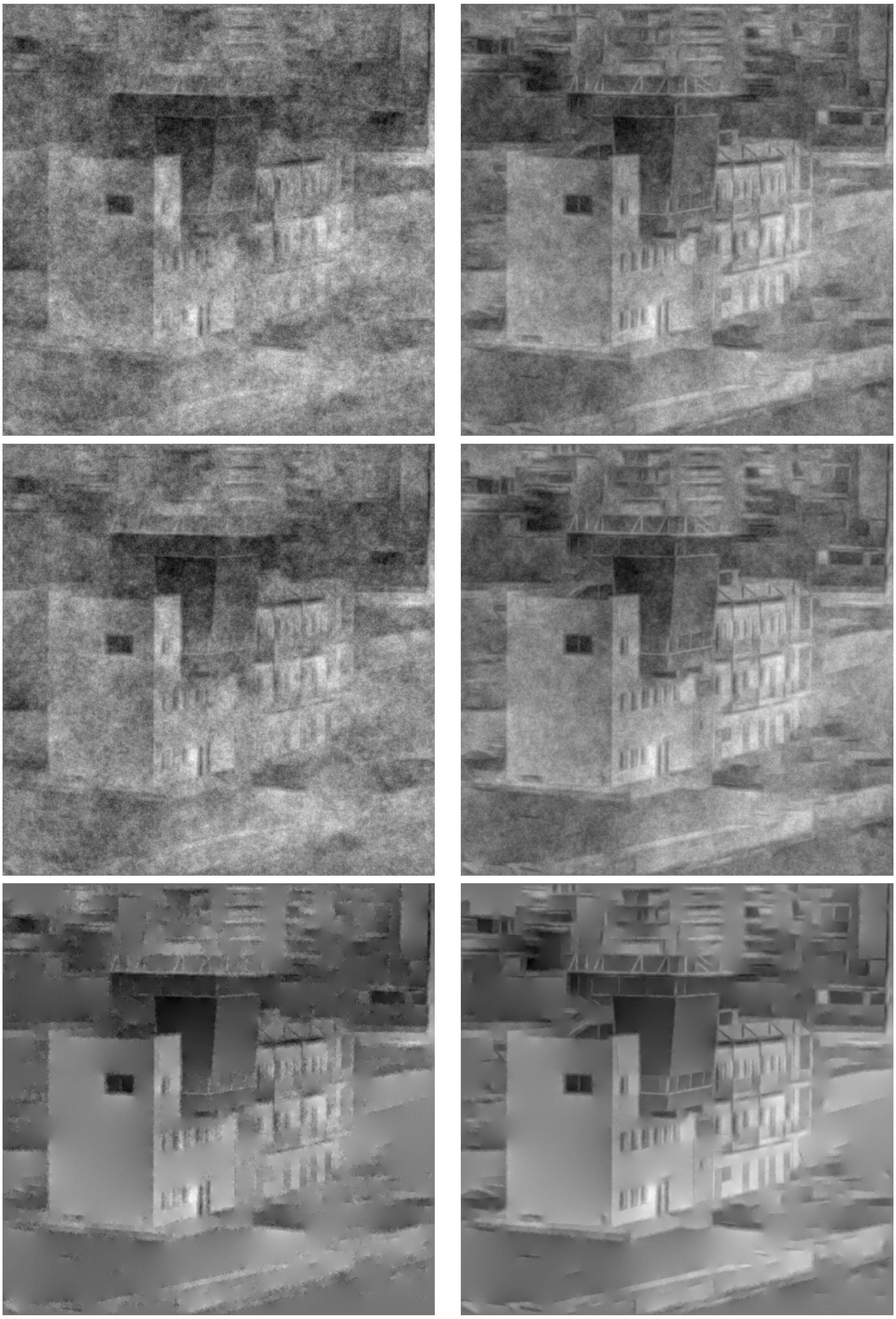
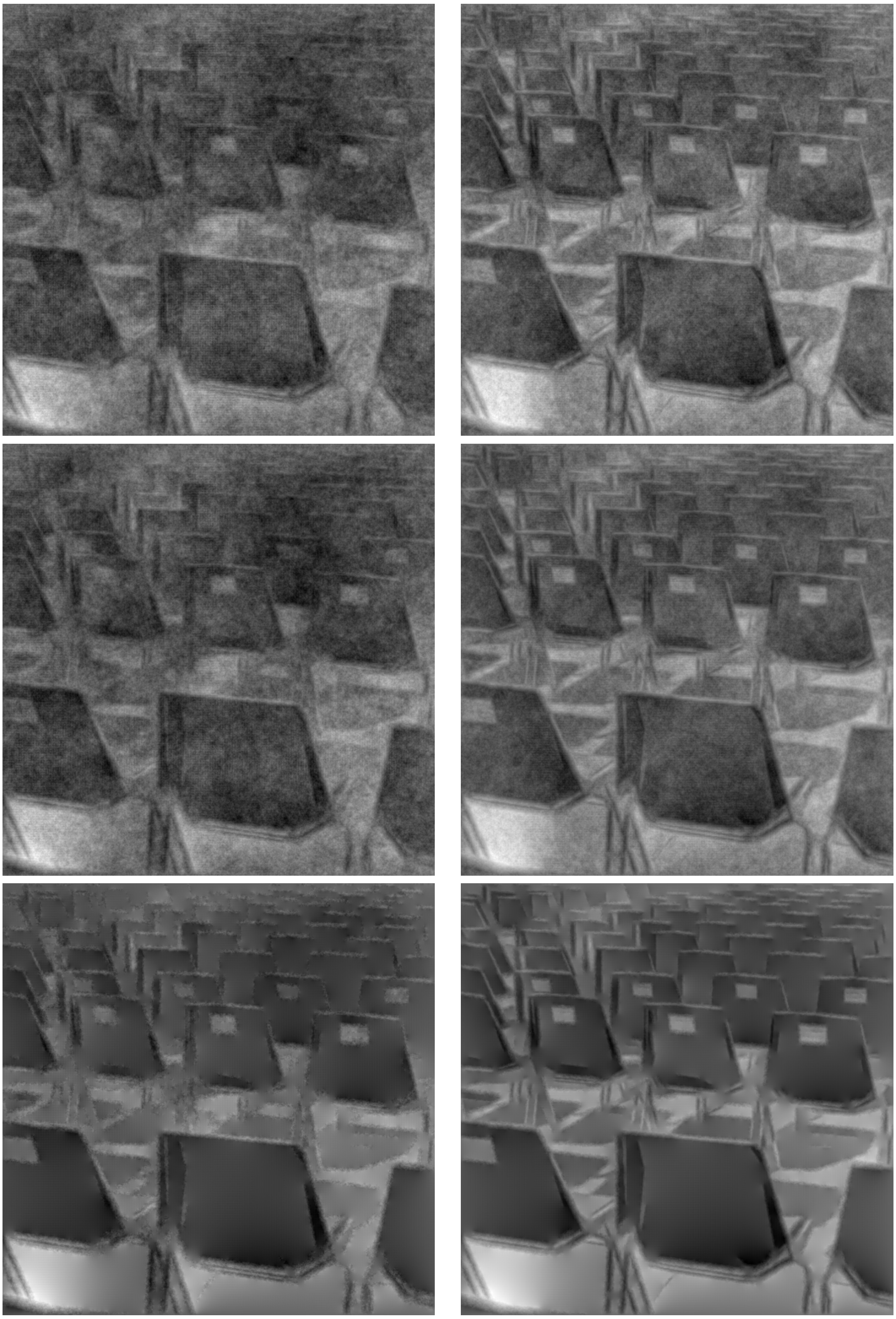


Figure 9: The original image I^0 was here the *Pirée* image of Figure 6. On the left column, we show the images reconstructed from a sample of P_0 (the one shown on Figure 6) and with respectively from top to bottom R_{rand}, R_{cst} and R_{100}. On the right column, same experiment but with the sample from Q_0.

Figure 10: The original image I^0 was here the *Chairs* image of Figure 7. On the left column, we show the images reconstructed from a sample of P_0 (the one shown on Figure 7) and with respectively from top to bottom R_{rand}, R_{cst} and R_{100}. On the right column, same experiment but with the sample from Q_0.



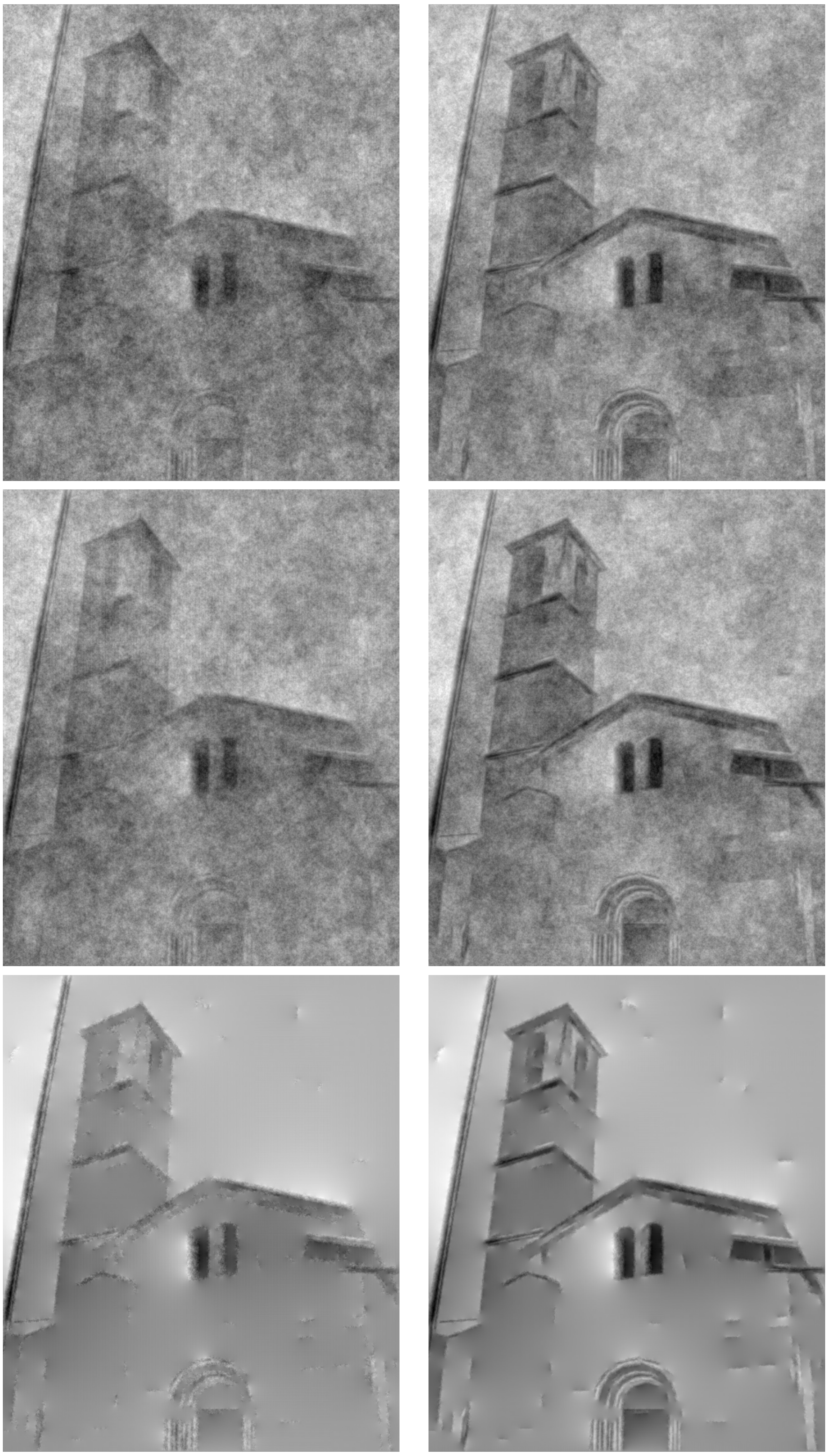


Figure 11: The original image I^0 was here the *Valbonne* image of Figure 8. On the left column, we show the images reconstructed from a sample of P_0 (the one shown on Figure 8) and with respectively from top to bottom R_{rand}, R_{cst} and R_{100}. On the right column, same experiment but with the sample from Q_0. Notice how the tree, present in the top right corner of the original image I^0, has here disappeared in the reconstructed images.

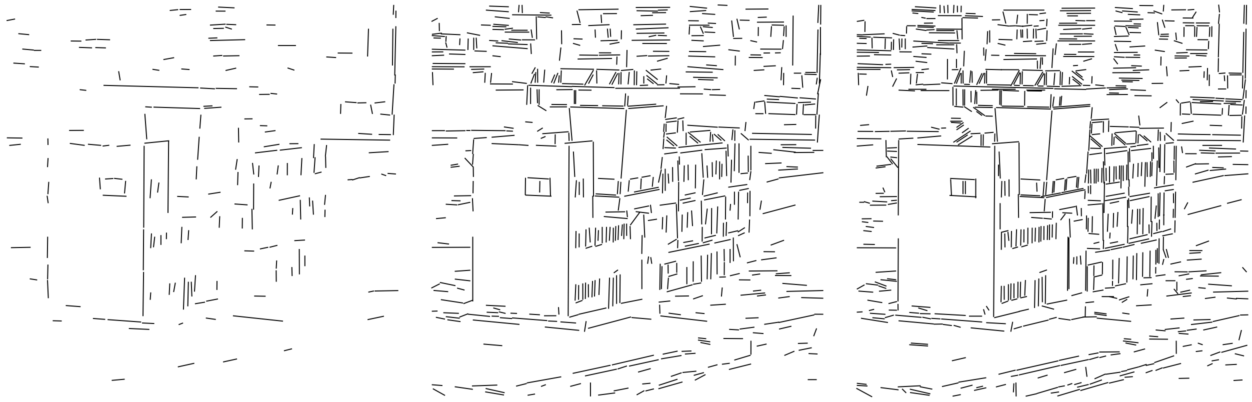


Figure 12: Output of the LSD algorithm on respectively, from left to right : the image reconstructed from the sample of P_0 and R_{100}, the image reconstructed from the sample of Q_0 and R_{100} (these two images are the ones of the last row of Figure 9), and the original *Pirée* image I^0. Notice haw the last two images are very similar.

Appendix A

Proof of Lemma 3

Proof. It is based on the fact that we consider here a symmetric function H_p (*i.e.* symmetric in the sense that it is invariant under any permutation of $q_1, \dots, q_n$) under a constraint that is also symmetric. We first notice that a point of local maximum of $H_p(q_1, \dots, q_n) = -\sum_{i=1}^n q_i \log \frac{q_i}{p} + (1-q_i) \log \frac{1-q_i}{1-p}$ under the constraint $C(q_1, \dots, q_n) \geq \eta$ (where $\eta \in (0, 1/2]$ and $C(q_1, \dots, q_n) = \mathbb{P}[K \geq k_0]$, with K following a Poisson binomial distribution of parameters $q_1, \dots, q_n$) is necessarily achieved when $C(q_1, \dots, q_n) = \eta$. Indeed if it is not the case then we will have a point $(q_1, \dots, q_n)$ of local maximum of H_p with $C(q_1, \dots, q_n) > \eta$. Now, at least one the q_i is not equal to p (because $B(n, k_0, p) < \eta$ by hypothesis) and therefore we can slightly modify it, still satisfying the constraint $C(q_1, \dots, q_n) \geq \eta$ and increasing H_p.

Let us now prove that $(\overline{q}, \dots, \overline{q})$ with $\overline{q} := B^{-1}_{n,k_0}(\eta)$ is a point of local maximum of H_p under the constraint $C(q_1, \dots, q_n) = \eta$. We consider smooth (at least C^2) curves $t \mapsto q_i(t)$ defined for t real in a neighbourhood I of 0, and such that $\forall t \in I$, $C(q_1(t), \dots, q_n(t)) = \eta$ and $q_1(0) = \ldots = q_n(0) = \overline{q}$. Then we define for all $t \in I$, $h(t) = H_p(q_1(t), \dots, q_n(t))$ and we will compute $h'(0)$ and $h''(0)$. A simple computation leads to

$$h'(0) = -\left(\log \frac{\overline{q}}{1-\overline{q}} - \log \frac{p}{1-p}\right) \sum_{i=1}^n q_i'(0) \tag{13}$$

$$\text{and} \quad h''(0) = -\left(\log \frac{\overline{q}}{1-\overline{q}} - \log \frac{p}{1-p}\right) \sum_{i=1}^n q_i''(0) - \frac{1}{\overline{q}(1-\overline{q})} \sum_{i=1}^n q_i'(0)^2. \tag{14}$$

Now, since $C(q_1(t), \dots, q_n(t)) = \eta$ for all t, and since C is symmetric, we get

$$\frac{\partial C}{\partial q_1}(\overline{q}, \dots, \overline{q}) \sum_{i=1}^n q_i'(0) = 0 \quad \text{and} \tag{15}$$

$$\frac{\partial C}{\partial q_1}(\overline{q}, \dots, \overline{q}) \sum_{i=1}^n q_i''(0) + \frac{\partial^2 C}{\partial q_1 \partial q_2}(\overline{q}, \dots, \overline{q}) \sum_{i,j=1, i \neq j}^n q_i'(0) q_j'(0) + \frac{\partial^2 C}{\partial q_1^2}(\overline{q}, \dots, \overline{q}) \sum_{i=1}^n q_i'(0)^2 = 0. \tag{16}$$

To compute the partial derivatives of C, we go back to its definition. Indeed C is defined by: $C(q_1,\ldots,q_2) = \mathbb{P}[Y_1+\ldots+Y_n \geq k_0]$ where the Y_i are independent Bernoulli random variables with respective parameter q_i. We can then develop the probability term and rewrite C as

$$C(q_1,\ldots,q_2) = q_1(\mathbb{P}[Y_2+\ldots+Y_n \geq k_0-1] - \mathbb{P}[Y_2+\ldots+Y_n \geq k_0]) + \mathbb{P}[Y_2+\ldots+Y_n \geq k_0]$$

$$\text{and also} \quad C(q_1,\ldots,q_2) = q_1q_2(P_2-2P_1+P_0)+(q_1+q_2)(P_1-P_0)+P_0,$$

where $P_2 := \mathbb{P}[Y_3+\ldots+Y_n \geq k_0-2]$; $P_1 := \mathbb{P}[Y_3+\ldots+Y_n \geq k_0-1]$ and $P_0 := \mathbb{P}[Y_3+\ldots+Y_n \geq k_0]$. These functions depend only on $q_3,\ldots,q_n$. This allows us to easily compute the partial derivatives of C at the point $(\overline{q},\ldots,\overline{q})$ and get

$$\frac{\partial C}{\partial q_1}(\overline{q},\ldots,\overline{q}) = \frac{(n-1)!}{(k_0-1)!(n-k_0)!}\overline{q}^{k_0-1}(1-\overline{q})^{n-k_0}\ , \quad \frac{\partial^2 C}{\partial q_1^2} = 0$$

$$\text{and} \quad \frac{\partial^2 C}{\partial q_1 \partial q_2} = \frac{(n-2)!}{(k_0-1)!(n-k_0)!}\overline{q}^{k_0-2}(1-\overline{q})^{n-k_0-1}(k_0-1-(n-1)\overline{q}).$$

Then from (15), we deduce that

$$\sum_{i=1}^n q_i'(0) = 0,$$

and as a first consequence we thus have, from (13), that $h'(0)=0$. Then using (16), the values of the partial derivatives of C and the fact that $\sum_{j\neq i} q_j'(0) = -q_i'(0)$, we get that (14) gives:

$$h''(0) = -\frac{\overline{q}}{1-\overline{q}}\left(1+(\frac{k_0-1}{n-1}-\overline{q})(\log\frac{\overline{q}}{1-\overline{q}}-\log\frac{p}{1-p})\right)\sum_{i=1}^n q_i'(0)^2.$$

By definition of $\overline{q}$, we have $B(n,k_0,\overline{q}) = \eta \leq \frac{1}{2}$ (by hypothesis). Then by Lemma 1, and more precisely by Equation (7) in its proof, we get

$$1+(\frac{k_0-1}{n-1}-\overline{q})(\log\frac{\overline{q}}{1-\overline{q}}-\log\frac{p}{1-p}) \geq 1+\frac{1}{n+1}\log\frac{p}{1-p} > 0$$

by hypothesis. Finally, we have obtained that

$$h'(0) = 0 \quad \text{and} \quad h''(0) < 0,$$

showing that the point $(\overline{q},\ldots,\overline{q})$ is a local maximum of H_p under the constraint $C \geq \eta$.

□

References

[1] I. Abraham, R. Abraham, A. Desolneux, and S. Li-Thiao-Té. Significant edges in the case of non-stationary Gaussian noise. *Pattern Recognition*, 40(11):3277 – 3291, 2007.

[2] M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26(4):641–647, 1955.

[3] S. Blusseau, J. Lezama, R. Grompone von Gioi, J.-M. Morel, and G. Randall. Comparing human and machine detection thresholds: An a-contrario model for non accidentalness. In *European Conference on Visual Perception*, 2012.

[4] F. Cao. Application of the Gestalt principles to the detection of good continuations and corners in image level lines. *Computing and Visualisation in Science. Special Issue, Proceeding of the Algoritmy 2002 Conference*, 7:3–13, 2004.

[5] F. Cao, J. Delon, A. Desolneux, P. Musé, and F. Sur. A unified framework for detecting groups and application to shape recognition. *Journal of Mathematical Imaging and Vision*, 27(2):91–119, 2007.

[6] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[7] J. Delon, A. Desolneux, J.L. Lisani, and A.B. Petro. Automatic color palette. *Inverse Problems and Imaging*, 1(2):265–287, 2007.

[8] J. Delon, A. Desolneux, J.L. Lisani, and A.B. Petro. A non parametric approach for histogram segmentation. *IEEE Transactions on Image Processing*, 16(1):253–261, 2007.

[9] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.

[10] A. Desolneux, L. Moisan, and J.-M. Morel. Edge detection by Helmholtz principle. *Journal of Mathematical Imaging and Vision*, 14(3):271–284, 2001.

[11] A. Desolneux, L. Moisan, and J.-M. Morel. Computational Gestalts and perception thresholds. *Journal of Physiology*, 97(2-3):311–324, 2003.

[12] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003.

[13] A. Desolneux, L. Moisan, and J.-M. Morel. Maximal meaningful events and applications to image analysis. *Annals of Statistics*, 31(6):1822–1851, 2003.

[14] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis: A Probabilistic Approach.* Springer-Verlag, 2008.

[15] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: A Fast Line Segment Detector with a False Detection Control. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(4):722–732, 2010.

[16] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: a Line Segment Detector. *Image Processing On Line*, 2:35–55, 2012.

[17] B. Grosjean and L. Moisan. A-contrario detectability of spots in textured backgrounds. *Journal of Mathematical Imaging and Vision*, 33(3):313–337, 2009.

[18] P. Harremoës. Binomial and Poisson distributions as maximum entropy distributions. *IEEE Transactions on Information Theory*, 47(5):2039–2041, 2001.

[19] R. Kaas and J.M. Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34:13–18, 1980.

[20] J. Lezama, S. Blusseau, J.-M. Morel, G. Randall, and R. Grompone von Gioi. Psychophysics, Gestalts and Games. In G. Citti and A.Sarti, editors, *Neuromathematics of Vision*, Lecture Notes in Morphogenesis, pages 217–242. Springer Berlin Heidelberg, 2014.

[21] D. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Amsterdam, 1985.

[22] D. Lowe. Visual recognition as probabilistic inference from spatial relations. In *AI and the Eye*, pages 261–2793. A. Blake and T. Troscianko (John Wiley), 1990.

[23] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3):201–218, 2004.

[24] D. Mumford and A. Desolneux. *Pattern Theory : the stochastic analysis of real-world signals.* AK Peters - CRC Press, 2010.

[25] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel. An a contrario decision method for shape element recognition. *International Journal of Computer Vision*, 69(3):295–315, 2006.

[26] A. Myaskouvskey, Y. Gousseau, and M. Lindenbaum. Beyond independence: An extension of the a contrario decision procedure. *International Journal of Computer Vision*, 101(1):22–44, 2013.

[27] M. Payton, L. Young, and J. Young. Bounds for the difference between median and mean of beta and negative binomial distributions. *Metrika*, 36:347–354, 1989.

[28] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics (SIGGRAPH'03)*, 22(3):313–318, 2003.

[29] T. Veit, F. Cao, and P. Bouthemy. An a contrario decision framework for region-based motion detection. *International Journal on Computer Vision*, 68(2):163–178, 2006.

[30] H. von Helmholtz. *Treatise on Physiological Optics.* Thoemmes Press, 1999.

[31] W. C. Waterhouse. Do symmetric problems have symmetric solutions? *The American Mathematical Monthly*, 90(6):378–387, 1983.

[32] A.P. Witkin and J. Tenenbaum. On the role of structure in vision. In *Human and Machine Vision*, pages 481–543. A. Rosenfeld ed., Academic Press, New York, 1983.

[33] S. C. Zhu, Y. N. Wu, and D. Mumford. Minimax Entropy Principle and Its Application to Texture Modeling. *Neural Computation*, 9(8):1627–1660, 1997.

[34] S. C. Zhu, Y. N. Wu, and D. Mumford. Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.

[35] S.C. Zhu. Embedding Gestalt Laws in Markov Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1170–1187, 1999.