# A unified framework for detecting groups and application to shape recognition

Frédéric Cao, Julie Delon, Agnès Desolneux, Pablo Musé, Frédéric Sur

I R I S A
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION
INTERNE
N° 1746

# A UNIFIED FRAMEWORK FOR DETECTING GROUPS AND APPLICATION TO SHAPE RECOGNITION

F. CAO , J. DELON , A. DESOLNEUX , P. MUSÉ , F. SUR

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

# A unified framework for detecting groups and application to shape recognition

F. Cao[*] , J. Delon[**] , A. Desolneux[***] , P. Musé[****] , F. Sur[*****]

Systèmes cognitifs
Projets Vista

Publication interne n ° 1746 — Septembre 2005 — 36 pages

**Abstract:** A unified *a contrario* detection method is proposed to solve three classical problems in clustering analysis. The first one is to evaluate the *validity* of a cluster candidate. The second problem is that meaningful clusters can contain or be contained in other meaningful clusters. A rule is needed to define locally optimal clusters by inclusion. The third problem is the definition of a correct merging rule between meaningful clusters, permitting to decide whether they should stay separate or unit. The motivation of this theory is shape recognition. Matching algorithms usually compute correspondences between more or less local features (called shape elements) between images to be compared. This paper intends to form spatially coherent groups between matching shape elements into a shape. Each pair of matching shape elements indeed leads to a unique transformation (similarity or affine map.) As an application, the present theory on the choice of the right clusters is used to group these shape elements into shapes by detecting clusters in the transformation space.

**Key-words:** Cluster validity, merging criterion, number of false alarms, shape recognition

*(Résumé : tsvp)*

[*] IRISA/INRIA, fcao@irisa.fr
[**] CMLA, ENS-Cachan, delon@cmla.ens-cachan.fr
[***] MAP5/CNRS, Agnes.Desolneux@math-info.univ-paris5.fr
[****] CMLA, ENS-Cachan, muse@cmla.ens-cachan.fr
[*****] Loria/CNRS, sur@loria.fr

Centre National de la Recherche Scientifique
(UMR 6074) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

# Une méthode de détection de groupes et application à la reconnaissance de formes

**Résumé :** Une méthode de détection *a contrario* est proposée pour résoudre trois problèmes classiques de clustering. Le premier est d'évaluer la validité d'un cluster candidat. Le second problème est qu'un cluster peut en contenir un autre, ou lui même être contenu dans un cluster. Une règle est nécessaire pour choisir les clusters optimaux lors de telles inclusions. Le troisième problème est la définition d'une règle de fusion correcte entre deux clusters, permettant de décider si ceux-ci doivent rester séparés ou au contraire être fusionnés. Cette étude est motivée par une application de reconnaissance des formes. Les algorithmes de mise en correspondance calculent en général des caractéristiques locales (appelées éléments de forme dans le présent article), qui sont ensuite comparées. Notre but est de montrer qu'on peut former des groupes spatialement cohérents d'éléments de formes. Chaque correspondance entre élément de forme définit en effet une unique transformation (similitude ou transformation affine dans le cas présent). Le groupement de ces transformations permet de détecter des formes à partir des éléments de formes.

**Mots clés :** Validité d'un cluster, critère de fusion, nombre de fausses alarmes, reconnaissance de formes

# Contents



# 1 Introduction

## 1.1 Problem statement

Clustering aims at discovering structure in a point data set, by dividing it into its "natural" groups. There are three classical problems related to the construction of the right clusters. (See Fig. 1.)

1. The first one is to evaluate the *validity* of a cluster candidate. In other words, is a group of points really a cluster, *i.e.* a group with a large enough density?

2. The second problem is that meaningful clusters can contain or be contained in other meaningful clusters. A rule is needed to define locally optimal clusters by inclusion. This rule, however, is not enough to interpret correctly the data.

3. The third problem is the definition of a correct merging rule between meaningful clusters, permitting to decide whether they should stay separate or unit.

A unified *a contrario* method will be proposed. It consists in detecting regions of the space with an unexpectedly high concentration of points, relatively to a statistical background model. In continuation, some complexity issues and heuristics to find sound candidate clusters will be considered.

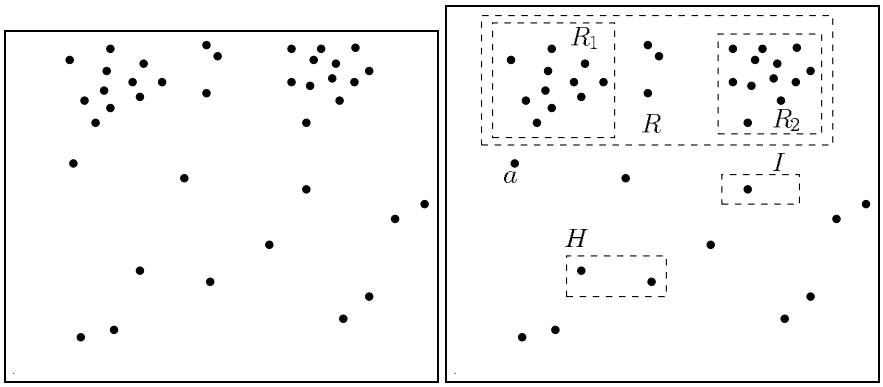



Figure 1: This figure illustrates three aspects of the grouping problem. The figure presents a set of data points in the plane and some test regions where an exceptional density may be observed, or not. Intuitively, the regions $H$ and $I$ do not contain clusters. So a first question is to rule out such non meaningful clusters. A second question is the choice of sound candidate regions: for instance, should not $R_1$ be enlarged to include the point $a$? As a last question, what is the best description of the observed clusters? The region $R$ is a possible good candidate, but it also contains the points of regions $R_1$ and $R_2$ which also are sound candidates. Thus, the question arises of whether $R$ should be chosen as cluster region, rather than the pair $(R_1, R_2)$.

This theory is then used to address a shape recognition problem. Given two images, how to anwer the question "do these two images have shapes in common?". This question only makes sense if a set of invariance properties is also given. For instance, it is sound to assume that the perception of a shape is widely independent from the viewpoint. Hence, the recognition procedure should be projective invariant, or, at least for remote planar shapes, affine invariant. It should also be quite independent from illumination conditions. And finally, it should resist to partial occlusions. This last requirement implies that, unless in specific applications, recognition cannot be the mere research of global templates. Instead, more simple and local parts of shapes have to be analyzed and identified in each image of the considered pair. Such local parts, or *shape elements* can be defined in several ways. In [30], shape elements are pieces of level lines that have been encoded in an affine invariant way. This representation will be adopted in the following, but is definitively not the main scope of the paper. Moreover, the theory that follows can be applied exactly in the same way to other types of descriptors. For instance, SIFT descriptors [24] are local histograms of directions, in similarity or affine invariant neighborhoods of keypoints. The main point is that shape elements define local, invariant frames. The next recognition step is to match similar shape elements. When shape elements are pieces of level lines, such a matching procedure, with an analysis of the detection threshold has been decribed in [30] and will not be detailed further in this paper. The result of this procedure is a set of pairs of matching shape elements.
Now, recognition is obviously not terminated at this point, and this is where the results of this paper come into action. Indeed, the local matching does not detect that two shape elements belong to the same single shape. For this purpose, shape elements have to be grouped together, whenever they form coherent wholes. It is then natural to define groups, as sets of shape elements that are transformed from the first image to the second one, by the same transformation. (In the present setting, a similarity or an affine map.) Thus, the problem of finding

groups of shape elements can be formulated as the detection of groups of transformations, *i.e.* a clustering problem. These groups of shape elements are more proper to define shapes.

The plan of this paper is as follows. Sect. 1.2 gives a short overview of the related problems in clustering analysis and grouping in shape recognition. Sect. 2 is the theoretical core of the paper and proposes an answer to the three questions of validity, stopping rule and merging. In Sect. 3, this theory is applied to perceptual grouping, illustrated by simple experiments. In Sect. 4, the application is to group points that are geometric transformations, corresponding to matches between parts of images. Sect. 5 contains numerical experiments, showing the validity of the proposed approach.

## 1.2 Related work

The problem of finding groups in a data set is an active research field. It is involved in data-mining, pattern recognition and pattern classification. The main clustering techniques are presented in [7, 9, 16, 17, 18, 20, 34]. All these methods face the three general problems above. Dubes [8] and Milligan and Cooper [28] proposed solutions to the choice of the number of clusters, which are related to the stopping rule in hierarchical methods. Bock [2] and Gordon [11, 12] are particularly interested in the validity assessment. Their approach is close to what we call an *a contrario method*: they define a background model in which they measure the likelihood of the concentration of points. A uniform model may not be the most adapted method, and it may be useful to define a data-dependent background model as shall be done in the next section. The method of the present paper is directly inspired by Desolneux *et al.*'s method for detecting groups of dots in an image [6]. In this method, a hierarchical classification of the set of dots is considered, and meaningful clusters are detected as large deviations from a standard Poisson null model. A maximality criterion was also defined but had several flaws that are taken in consideration in the approach proposed in this paper.

Grouping phenomena are also probably essential in human perception. In vision, the grouping phenomenon was thoroughly explored by the Gestalt school, from the founding paper of Wertheimer [36]. In Computer Vision, the first attempts to model a computational perceptual organization date back to Marr [26]. More recently Lowe [25] proposed a detection framework based on the computation of accidental occurrences. Even though the relation with perceptual organization was not highlighted, Computer Vision also used spatial coherence for shape or object detection. One of the first and best example is Ballard's work on the generalized Hough transform [1]. In his paper, Ballard proposed a method extending the Hough transform to any kind of planar shape, not necessarily described by an analytic formula. Stockman [33] presented another early work based on the same principle (recognize a target shape by finding clusters in the transformation space), where he introduced a coarse to fine technique allowing to reduce the search complexity. Other voting schemes, like Geometric Hashing [21, 37], the Alignment method [15], or tensor-voting [27], are frequently used in detection or recognition problems. They are computationally more expansive and can be less accurate. An advantage of these voting procedures is that they are systematic, and can in principle be generalized in any dimension (although the computational burden often becomes too heavy). However, they do not solve the decision problem. In [13, 14], Grimson and Huttenlocher presented a study on the likelihood of false peaks in the Hough parameter space. Their work inspired the detection method adopted in this paper. They indeed proposed a detection framework where recognition thresholds are derived from a null model ("*the conspiracy of random*"). Previous recognition methods generally associated a single threshold with each target image, independently of the scene complexity. In contrast to these methods, the grouping thresholds derived in this paper satisfy an important property: they are functions of the scene complexity and of the uncertainty in feature extraction. The method of the present article shares these fundamental ideas with Grimson and Huttenlocher's work. The computational swiftness is obtained by a hierarchical representation of the transformation points. The definition of a data-dependent background model is crucial for avoiding false clusters: Grimson and Huttenlocher's method assumes that matched features are uniformly distributed in the image. This assumption is usually not valid [31]. One of the observation of this paper is that an empirical distribution can be used to detect groups in arbitrary data points.

# 2 Hierarchical clustering and validity assessment

## 2.1 *A contrario* cluster validity

The first contribution is to define a quantitative measure of validity of a group of points. A group will be considered as meaningful whenever it is contained in a region in which only few points are expected if the data were drawn at random. Hence, a probability model has to be defined, as well as the precise event that will be sought.

### 2.1.1 The background model

In all what follows, $E$ is a given subset of $\mathbb{R}^D$, endowed with a probability measure $\pi$ (which will be also called *background law*.) By definition, $\pi(R)$ is the probability that a random point belongs to $R$. We do not mention measurability issues here. They are straightforward in this context.
The definition of $\pi$ is problem specific. In general, it is given *a priori*, or can be empirically estimated over the data. (See next section.)

**Definition 2.1** *A* background process *is a finite point process $(X_i)_{i=1,\ \dots\ M}$ in $E$ made of $M$ mutually independent variables, identically distributed with law $\pi$.*

Let us now consider an observed data set of $M$ points $(x_1,\ \dots,\ x_M)$ in $E^M$. A subset of the data set will form a meaningful group if an important part of its points belong to a "small" given region, whenever the probability of this event is small. In other words, it could not be explained by the background model. Therefore, the cornerstone of the *a contrario* method is to contradict the following assumption:
**(A)** *The observed M-tuple $(x_i)_{i\in\{1\dots M\}}$ is a realization of the background process.*

Before going further let us make the following crucial remark. Let us assume for instance that $E = (0,1)^2$, and $\pi$ is the uniform law on $E$. Then, given $M$ points in $E = (0,1)^2$, it is always possible to find a connected set $R$ of arbitrary small probability $\pi(R)$ containing all the datapoints. Of course, it would be a non sense to conclude that any set of points form a group. It means that the regions that are to be considered cannot be completely given a posteriori to the observation. Special care must be brought to the type of event that are to be studied. In particular, the definition of a meaningful group will involve the total number of candidate regions, and this number has to be finite. This limits the complexity of the family of regions.

### 2.1.2 Meaningful groups

Consider a region $R \subset E$ containing the origin, typically a hyperrectangle centered at the origin. Assume that $k$ points among $(x_1,\ \dots,\ x_M)$ belong to a region of the type $x_j + R$, for some $j$, $1 \leqslant j \leqslant M$. If $k$ is large enough, and $\pi(x_j + R)$ small enough, one will observe a cluster of points in $x_j + R$ which can hardly have been generated by the background model. This group of points will then be detected in $x_j + R$, by contradicting the hypothesis that the points are due to chance. Clusters can be grouped around any of the $x_j$ and can have any shape. A generic shape for the tested regions must, however, be fixed *a priori*. The region $R$ will have to belong to a finite family $\mathcal{R}$ of regions, which will be detailed further. For the time being, let us simply assume that $\mathcal{R}$ has finite cardinality $\#\mathcal{R}$ and that for all $R \in \mathcal{R}$, $0 \in R$.

In the following, for $k \leqslant M \in \mathbb{N}$ and $0 \leqslant p \leqslant 1$, let us denote by

$$\mathcal{B}(M,k,p) = \sum_{j\geqslant k} \binom{M}{j} p^j (1-p)^{M-j}$$

the tail of the binomial law. Given a background process $X_1,\ \dots,\ X_M$ and a region $R$ of $E$ with probability $\pi(R)$, one can interpret $\mathcal{B}(M,k,\pi(R))$ as the probability that *at least $k$ out of the $M$ points of the process fall*

*into* $R$. A thorough study of the binomial tail and its use in the detection of geometric structures can be found in [4].

Fix $1 \leqslant j \leqslant M$ and $R' \in \mathcal{R}$. We note:

- $X = (X_1, \ldots, X_M)$, the background process.
- $X^j = (X_1, \ldots, X_M)$ with $X_j$ omitted in the list
- $K(X^j, X_j, R')$, number of points in the list $X^j$ belonging to $X_j + R'$.

**Definition 2.2** *Let $R$ be a region of the type $R = X_j + R'$ for some $j \in \{1, \ldots, M\}$ and $R' \in \mathcal{R}$. We call number of false alarms of $R = X_j + R'$,*

$$NFA_g(X, j, R') \equiv \#\mathcal{R} \cdot M \cdot \mathcal{B}(M-1, K(X^j, X_j, R'), \pi(X_j + R')). \tag{2.1}$$

*We say that $R = X_j + R'$ is an $\varepsilon$-meaningful region if $NFA_g(X, j, R') \leqslant \varepsilon$.*
*By a slight abuse of notation, $NFA_g(X, j, R')$ will also be denoted by $NFA_g(R)$.*

As a sanity check of the above definition, our aim is to prove that the expected number of $\varepsilon$-meaningful regions is less than $\varepsilon$.

**Proposition 2.1** *If $X_1$, ... $X_M$ is a background process, the expected number of $\varepsilon$-meaningful regions is less than $\varepsilon$.*

The proof is given in appendix.
*Remark.* The key point is that the *expectation* of the number $S$ of meaningful regions is easily controlled. The probability law of $S$ would instead be extremely difficult to compute because of the interactions between regions.

Let us summarize: the number of false alarms is a measure of how likely it is that a group contained in a region $R$ centered on a data point, containing at least $k$ of the other data points, was generated "by chance", as a realization of the background process. The lower $NFA_g(R)$, the less likely the observed cluster in the background process. By Prop. 2.1, the only parameter controlling the detection is $\varepsilon$. This provides a handy way to control false detections. If, on the average, one is ready to tolerate one "non relevant region" among all regions, then $\varepsilon$ can be simply set to 1.

The following proposition shows that the influence of the parameter $\#\mathcal{R}$ and of the decision parameter $\varepsilon$ on the detection results is very weak.

**Proposition 2.2 ( [4])** *Let $R$ be a region in $\mathcal{R}$ and let*

$$k^*(\varepsilon) = \min\{k : \#\mathcal{R} \cdot \mathcal{B}(M-1, k, \pi(R)) \leqslant \varepsilon\}.$$

*Then*

$$\alpha(M, \varepsilon)\sqrt{2\pi(R)(1-\pi(R))} \leqslant k^*(\varepsilon) - \pi(R)(M-1) \leqslant \frac{\alpha(M, \varepsilon)}{\sqrt{2}}, \tag{2.2}$$

*where $\alpha(M, \varepsilon) = \sqrt{(M-1)\ln(\#\mathcal{R}/\varepsilon)}$.*

Notice that $k^*(\varepsilon)$ is the minimal number of points in a $\varepsilon$-meaningful group. By the preceding result, this decision threshold only has a logarithmic dependance upon $\#\mathcal{R}$ and $\varepsilon$.

Figure 2 shows an example of clustering. The data consists of 950 points uniformly distributed in the unit square, and 50 points manually added around the positions $(0.4, 0.4)$ and $(0.7, 0.7)$. The figure shows the result of a numerical method involving the above NFA. The background distribution $\pi$ is taken uniform in $(0, 1)^2$. Both visible clusters are found and happen to belong to regions whose NFA's are respectively $10^{-7}$ and $10^{-8}$. Such low numbers can barely be the result of chance. How to obtain *exactly* these two clusters and no other larger or smaller ones which would also be meaningful? This will be the object of the next two sections.

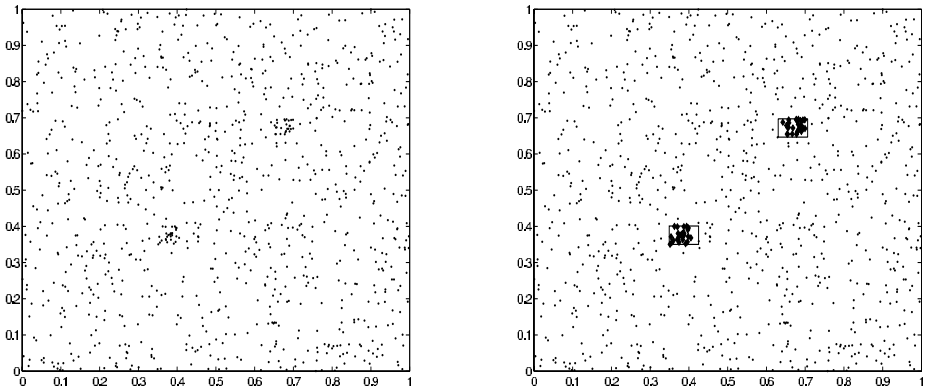


Figure 2: Clustering of twice 25 points around $(0.4, 0.4)$ and $(0.7, 0.7)$ surrounded by 950 i.i.d. points, uniformly distributed in the unit square. The regions of $\mathcal{R}$ are rectangles as described in Sect. 2.3.1. In this example $\#\mathcal{R} = 2500$ (50 different sizes in each direction). Exactly two maximal meaningful clusters are detected. (See Sect. 2.2 for the definition of maximality.) The NFA of the lower left one is $10^{-8}$ while the upper-right one has a NFA equal to $10^{-7}$.

## 2.2 Optimal merging criterion

In Sect. 2.1.2, it was proposed to restrict the space of tests to regions of the form $x_i + R$, where $x_i$ is an observed data point and $R \in \mathcal{R}$, a fixed finite set of regions containing the origin in $\mathbb{R}^D$. While each meaningful region is relevant by itself, the whole set of meaningful regions exhibits, in general, a high redundancy. Indeed, a very meaningful region $R$ usually remains meaningful when it is slightly enlarged or shrunk into a region $R'$. (See Fig. 1.)

If, *e.g.* $R \subset R'$, this question is easily answered by comparing $NFA_g(R)$ and $NFA_g(R')$. The region with the smallest number of false alarms must of course be preferred. Another more subtle question arises when three or more regions interact. Let $R_1$ and $R_2$ be two tested regions and $R$ another tested region containing all the points of $R_1$ and $R_2$. Let us call $R$ a "merged region" of $R_1$ and $R_2$. We then face two conflicting interpretations of the data: two clusters or just one? The merged region $R$ is not necessarily a better data representation than the two separate cluster regions $R_1$ and $R_2$. A first possibility is that $R$ is less meaningful than each one of the merging ones. In such a case, $R_1$ and $R_2$ should be kept, rather than $R$. The situation is less obvious when $R$ is more meaningful than both $R_1$ and $R_2$. In that case, keeping $R_1$ and $R_2$ apart may still be opportune. So a quantitative merging criterion is required. We shall first define a *number of false alarms for a pair of regions*. This new value will be compared to the NFA of the merged region. Let us introduce the trinomial coefficient

$$\binom{M}{i,j} = \binom{M}{i}\binom{M-i}{j}.$$

We note

$$\mathcal{M}(M, k_1, k_2, \pi_1, \pi_2) = \sum_{i=k_1}^{M} \sum_{j=k_2}^{M-k_1} \binom{M}{i,j} \pi_1^i \pi_2^j (1 - \pi_1 - \pi_2)^{M-i-j}. \tag{2.3}$$

This number can be interpreted as follows. Let $R_1$ and $R_2$ be two disjoint regions of $E$ and $\pi_1 = \pi(R_1)$, $\pi_2 = \pi(R_2)$ their probabilities. Then $\mathcal{M}(M, k_1, k_2, \pi_1, \pi_2)$ is the probability that at least $k_1$ among the $M$, and then at least $k_2$ points among the remaining $M - k_1$, fall into $R_1$ and $R_2$ respectively. Thus, this probability measures how exceptional a pair of concentrated clusters can be in the background model. We aim at defining a new NFA for such events. As in the study of $\varepsilon$-meaningful regions, some care must be taken of notation and abbreviations. Let $1 \leqslant i \neq j \leqslant M$ and $R', R'' \in \mathcal{R}$. Now, two tested regions $x_i + R'$ and $x_j + R''$ may intersect and we have to deal with this possibility. We note

- $X = (X_1, \ldots, X_M)$, background process
- $X^{ij} = (X_1, \ldots, X_M)$ with $X_i,\ X_j$ omitted in the list
- $R_i = X_i + R'$, $R_j = X_j + R''$
- $K(X, i, j, R', R'') =$ number of points among $X^{ij}$ belonging to $R_i \setminus R_j = (X_i + R') \setminus (X_j + R'')$.
- $K_i = K(X, i, j, R', R'')$, and $K_j = K(X, j, i, R'', R')$ obtained by reversing the role of $i$ and $j$.
- $\Pi_i = \pi((X_i + R') \setminus (X_j + R''))$, $\Pi_j = \pi((X_j + R'') \setminus (X_i + R'))$

**Definition 2.3** *We call number of false alarms of the random pair of regions $(R_i, R_j) = (X_i + R', X_j + R'')$*

$$NFA_{gg}(X, i, j, R', R'') = \frac{M(M-1)^2(\#\mathcal{R})^2}{2}\mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j). \tag{2.4}$$

*We say that a random pair of regions $(R_i, R_j)$ is $\varepsilon$-meaningful if $NFA_{gg}(X, i, j, R', R'') < \varepsilon$. Without ambiguity, $NFA_{gg}(X, i, j, R', R'')$ will also be denoted by $NFA_{gg}(R_i, R_j)$.*

Again, the aim is to prove that the expected number of $\varepsilon$-meaningful pairs of regions is less than $\varepsilon$.

**Proposition 2.3** *The expected number of $\varepsilon$-meaningful pairs of regions is less than $\varepsilon$.*

This proposition leads to the following heuristic. Two measures of meaningfulness are available: the $NFA$ of a region and the $NFA$ of a pair of regions. Since the number of $\varepsilon$-meaningful regions or pairs of regions is about $\varepsilon$ in the background model, we consider that they have the same order of magnitude and they can be compared to define a mergion criterion.

**Definition 2.4 (Indivisible region)** *Let $R_1$ and $R_2$ be two regions and $R$ a region containing all the data points of $R_1$ and $R_2$. We say that $R$ is indivisible relatively to $R_1$ and $R_2$ if*

$$NFA_g(R) \leqslant NFA_{gg}(R_1, R_2). \tag{2.5}$$

*Given a set $\mathcal{R}$ of test regions and $R$ an element of $\mathcal{R}$, $R$ is said to be indivisible in $\mathcal{R}$ if it is indivisible relatively to all pairs $(R_1, R_2)$ of regions in $\mathcal{R}$ such that $R$ contains the data points of $R_1$ and $R_2$.*

Equation (2.5) represents a crucial test for the coherence of a cluster region. If it is not fulfilled, $R$ will not be considered as a valid region, as it can be divided into a more meaningful pair of cluster regions. The next lemma will prove useful in speeding up the merging decision.

**Lemma 2.1** *For every $k_1$ and $k_2$ in $\{0, \ldots, M\}$, such that $k_1 + k_2 \leqslant M$ and for every $\pi_1$ and $\pi_2$ in $[0, 1]$ such that $\pi_1 + \pi_2 \leqslant 1$,*

$$\mathcal{M}(M, k_1, k_2, \pi_1, \pi_2) \leqslant \mathcal{B}(M, k_1, \pi_1) \cdot \mathcal{B}(M, k_2, \pi_2). \tag{2.6}$$

This result of "negative dependence" of the binomial distribution is not obvious and has been proved in [19] by Joag-Dev and Proschan. We are actually interested in its consequence to follow.

**Proposition 2.4** *If $R$ is indivisible with respect to $R_1$ and $R_2$, then*

$$NFA_g(R) < \frac{M}{2} \cdot NFA_g(R_1) \cdot NFA_g(R_2)$$

*Proof.* From (2.4) and Def. 2.4, one has

$$NFA_{gg}(R_1, R_2) = \frac{M(M-1)^2(\#\mathcal{R})^2}{2}\mathcal{M}(M-2, k_1, k_2, \pi_1, \pi_2)$$

and

$$NFA_g(R_i) = M(\#\mathcal{R})\mathcal{B}(M-1, k_i, \pi_i),\ i = 1, 2.$$

By Lem. 2.1, it follows

$$NFA_{gg}(R_1, R_2) \leqslant \frac{1}{2}M^3(\#\mathcal{R})^2\mathcal{B}(M-2, k_1, \pi_1)\mathcal{B}(M-2, k_2, \pi_2).$$

Since $\mathcal{B}(M-1, k, p) \leqslant \mathcal{B}(M, k, p)$ for all $M$, $k$ and $p$, the result follows. □

Proposition 2.4 is useful from the computational viewpoint, since in many cases one can avoid computing the tail of the trinomial distribution by "filtering" those clusters that do not pass the necessary condition.

## 2.3 Computational issues

### 2.3.1 The choice of test regions

What is the right set of test regions $\mathcal{R}$? This question is obviously application driven. To fix ideas, let us just indicate a possible choice. For some reasonably fixed $a > 0$, $r > 1$ and $n \in \mathbb{N}$, let us consider all hyperrectangles whose edge lengths belong to the set $\{a,\ ar,\ ar^2,\ \ldots,\ ar^n\}$. This allows one to consider a tractable number of test regions with very different sizes and shapes. The choice of the hyperrectangles is particularly opportune when the probability distribution $\pi$, defined on a hyperrectangle $E$ of $\mathbb{R}^D$, is a tensor product of one-dimensional densities $\pi_1,\ \ldots,\ \pi_D$. We address the question with more details in the next section.

Definition 2.2 permits to compute the NFA of any test region centered at a data point. Since the number of scales is $n$ in each dimension, there are $Mn^D$ regions centered at a data point. In the next section, $D = 4$ or 6. From the numerical feasibility viewpoint, $Mn^D$ becomes too large when $n$ grows. This explains why the testing cannot generally be performed this way. It is better to involve a tree structure of the point data set obtained by a hierarchical clustering algorithm. A hierarchical organization of the data can be used to limit the number of tested regions, by proceeding as follows.

One starts by applying to the data point set a so-called *hierarchical clustering* method. The hierarchical clustering methods provide a family of nested partitions of the point data set. They yield a tree structure in which each node is a part of the set and a candidate cluster. This tree is sometimes called *dendrogram* [17].

Many of the most common aggregation procedure proceed by a recursive binary merging procedure. Thus, they directly yield binary trees. In such methods, the initial set of nodes is the set of data singletons, $\{x_1\},\ \ldots,\ \{x_N\}$. At each stage of the construction, the two closest nodes are united to form their parent node. The inter-cluster distance must be chosen *ad hoc*. In the case of sparse data, one can take the minimal distance $d(x_i, x_j)$ where $x_i$ belongs to the first cluster and $x_j$ to the second one. The nodes of the tree are all merged parts at all levels and the daughters of a node are the two parts it was merged from.

Let it be clear why such a construction can become necessary. The set of all possible partitions of a data point set is huge. A tree structure permits to reduce the exploration to the search of an optimal subtree of the initial tree structure. This reduction makes sense if the set of nodes of the initial tree structure contains roughly all groups of interest. The choices of the right metric on the data point set and of the right inter-cluster distances must be carefully specified for the problem of interest.

Given a dendrogram of the data point set, the following algorithm permits to explore all regions centered at data points and containing a node of the dendrogram.

**Grouping algorithm**
For each node $G$ (candidate group) in the clustering tree or dendrogram,

1. for each point $x$ of the node:
   (a) Find the smallest region $x + R$ centered at this point, and containing the other data points of the node. Call $k + 1$ the number of data points it contains.
   (b) Compute the NFA of the region as $M \cdot \#\mathcal{R} \cdot \mathcal{B}(M - 1, k, \pi(x + R))$.
2. Associate with the node $G$ the region $R(G)$ with lowest NFA thus computed; it contains the points of the node $G$, but may also contain other data points.

Once this algorithm has been performed, a candidate cluster region is associated with each node. By a harmless abuse of notation, let us note $NFA_g(G) = NFA_g(R(G))$ . In the same way, if $G_1$ and $G_2$ are a pair of nodes and $R(G_1)$ and $R(G_2)$ their regions, note $NFA_{gg}(G_1, G_2) = NFA_{gg}(R(G_1), R(G_2))$. In that way, the clustering tree is endowed with NFA's for nodes and for pairs of nodes. Conversely, the set $\mathcal{R}$ of regions of the form $R(G)$ inherits the tree structure.

### 2.3.2 Indivisibility and maximality

We are now faced with Questions 2 and 3 mentioned at the beginning of the present article: we can get many meaningful clusters by the preceding method. Their NFA is known. One can also compute the NFA of a pair of clusters, and compare it roughly to the NFA of their union. The next definition proposes a way to select the right clusters, by using the cluster dendrogram.

**Definition 2.5 (Maximal $\varepsilon$-meaningful group)** *A node region $R = R(G)$ in $\mathcal{R}$ is maximal $\varepsilon$-meaningful if and only if*

1. $NFA_g(R) \leqslant \varepsilon$,
2. *$R$ is indivisible with respect to all its pairs of descent.*
3. *for all indivisible descent $R'$, $NFA_g(R') \geqslant NFA_g(R)$,*
4. *for all indivisible ascent $R'$, either $NFA_g(R') > NFA_g(R)$ or there exists an indivisible descent $R''$ of $R'$ such that $NFA_g(R'') < NFA_g(R')$.*

*We say that $G$ is a maximal $\varepsilon$-meaningful if $R(G)$ is.*

Condition 4 implies that $R$ can be abandoned for a larger region only if this region has not been beaten by one of its descents. Imposing conditions 3 and 4 ensures that two different maximal meaningful groups are disjoint. Let us also remark that the indivisibility is required only with respect to pairs of descent. Indeed, Def. 2.4 is theoretically satisfying but not pratically tractable. Hence, a slightly weaker condition is imposed.

Let us illustrate the critical importance of the merging condition with two simple examples. Figure 3 shows a configuration of 100 points, distributed on $[0, 1]^2$, and naturally grouped in two clusters $G_1$ and $G_2$, for a background model which is uniform in $[0, 1]^2$. In the hierarchical structure, $G_1$ and $G_2$ are the children of $G = G_1 \cup G_2$. All three nodes are obviously meaningful, since their $NFA_g$ is much lower than 1. Their $NFA_g$ also is lower than the $NFA_g$ of the other groups in the dendrogram. It has been checked that for this particular configuration,

$$NFA_g(G_2) < NFA_g(G) < NFA_g(G_1).$$

It is clear that $G_1$ represents an informative part of the data that should be kept. This will be the case. Notice that $G_2$ is more meaningful than $G$ and is contained in $G$. Thus, $G$ would be eliminated if only the most meaningful groups by inclusion were kept. On the other hand, $G$ is more meaningful than $G_1$, so that $G_1$ is not a local maximum of meaningfulness, with respect to inclusion. So, without the notion of indivisibility and maximality, trouble would arise: $G$ would eliminate $G_1$ and $G_2$ would eliminate $G$. One would get the solution indicated in the middle column of Fig. 3. In fact, $G$ is not indivisible since it is less meaningful than the pair

$(G_1, G_2)$. Thus, the result of the grouping procedure yields, in accordance with the rule of Def. 2.5, the pair $(G_1, G_2)$.

In [6], the above mentioned maximality definition was proposed: it consists of taking the lowest NFA in all the branches of the tree. As has been just seen, this definition is not suitable here. By this definition, $G_2$ would have been considered as the only maximal meaningful cluster of the tree.

Fig. 4 illustrates another situation where the indivisibility check yields the intuitively right solution. In this example, the union $G$ of two clusters $G_1$ and $G_2$ is more meaningful than each separate cluster. Without the indivisibility requirement, $G$ would be the only maximal meaningful group. This would have been coherent, had $G_1$ and $G_2$ been intricate enough. In the presented case, the indivisibility condition yields two clusters $G_1$ and $G_2$, since $NFA_{gg}(G_1, G_2) < NFA_g(G)$.

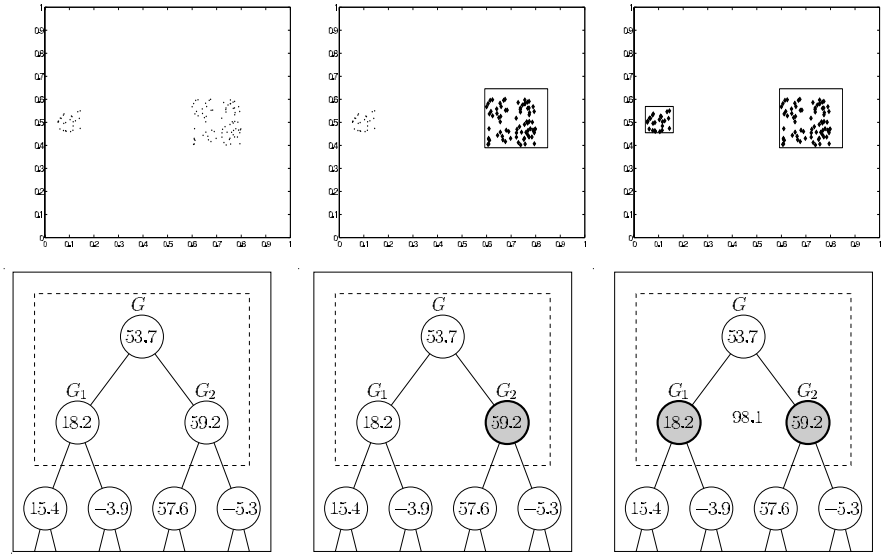



Figure 3: Indivisibility prevents collateral elimination. Each subfigure shows a configuration of points, and a piece of the corresponding dendrogram, with the selection of maximal meaningful groups, depicted in grey. The numbers in each node corresponds to $-\log_{10}(NFA_g)$ of its associated cluster, so that the cluster is meaningful when this number is large. The number placed between two nodes is the $NFA_{gg}$ of the corresponding pair. Left: original configuration. Middle: the node selected by taking only the most meaningful group in each branch. The right-most group $G_2$ is eliminated. It is, however, very meaningful since $NFA_g(G_2) = 10^{-18}$. Right: by combining indivisibility and maximality criteria, both clusters $G_1$ and $G_2$ are selected.

## 3 Experimental validation: object grouping based on elementary features

Grouping phenomena are essential in human perception, since they are responsible for the organization of information. In vision, grouping has been especially studied by Gestalt psychologists like Wertheimer [36]. The aim of these experiments is to extract the groups of objects in an image, that share some elementary geometrical properties. The objects boundaries are extracted as some contrasted level lines in the image, called *meaningful level lines* (see [5] for a full description of this extraction process). Once these objects are detected, say $O_1, ...O_M$, we can compute for each of them a list of $D$ features (grey level, position , orientation, etc...). If $k$ objects among $M$ have one or several features in common, we wonder if it is happening by chance or if it is enough to group them. Each data point is a point in a bounded subset of $\mathbb{R}^D$ and the method described above is applied. (Actually, some coordinates, as angles, belong to the unit circle, since periodicity must be taken into account. This can be done all the same.)

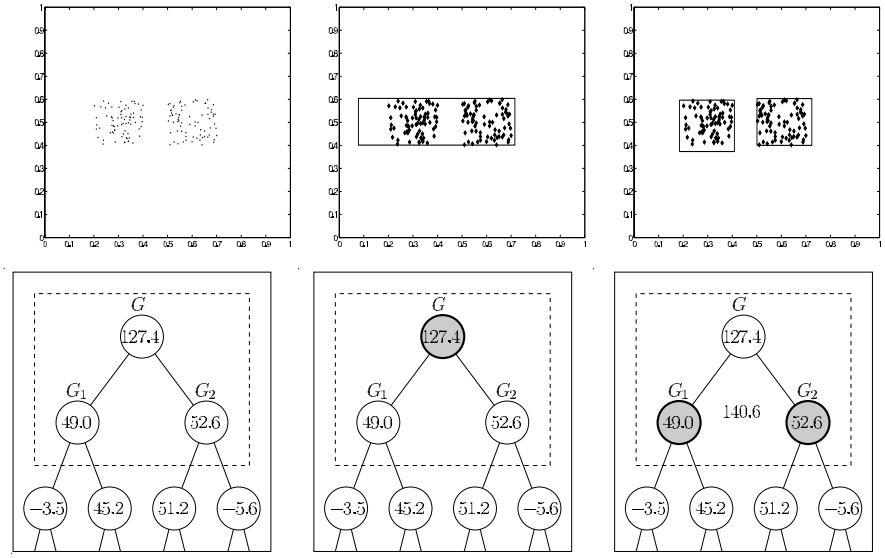



Figure 4: Indivisibility prevents faulty union. Each sub-figure shows a configuration of points, and a piece of the corresponding dendrogram, with the selection of maximal meaningful groups, depicted in grey. The number in each node corresponds to the $NFA_g$ of its associated cluster. The number between two nodes is the $NFA_{gg}$ of the corresponding pair. Left: original configuration. Middle: the node selected if one only checks maximality by inclusion and not indivisibility. The largest group $G$ has the lowest $NFA_g$ and would be the only one kept. Note that the optimal region is not symmetric, since it must be centered on a datapoint. Right: selected nodes obtained by combining the indivisibility and maximality criteria. Since $NFA_{gg}(G_1, G_2) = 10^{-140} < 10^{-127} = NFA_g(G)$, the pair $(G_1, G_2)$ is preferred to $G$.

### 3.1 Dots in noise

The first experiment is Fig. 2, which contains two groups of 25 points in addition to 950 i.i.d uniformly in the unit square. Two groups and two groups only are detected with very good $NFA_g$ (less than $10^{-7}$). The experiment on Fig. 5 shows the importance of the a priori distributions on the data points. Two different distributions lead to two different maximal meaningful groups. Both interpretations are correct but depend on the context.

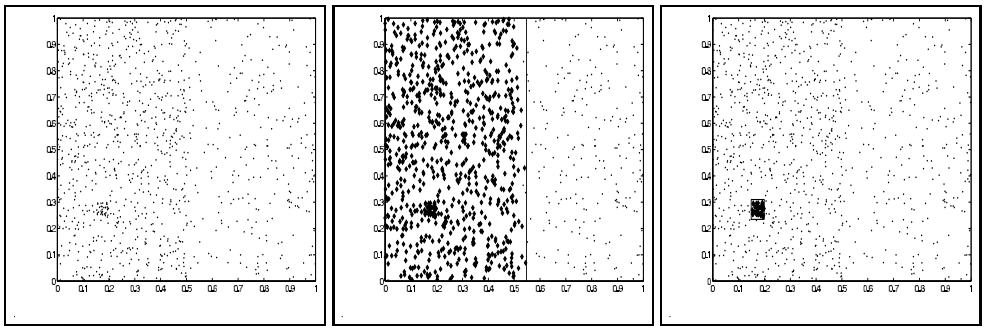


Figure 5: Importance of the distribution of the background model. The original data is the left-most figure. It is the superposition of 500 i.i.d points in $(0,1)^2$, of 500 i.i.d points in $(0, 0.5) \times (0, 1)$ and 25 points around $(0.2, 0.3)$. In the middle plot, the *a priori* distribution in the background model is taken uniform. Then, a single large maximal meaningful group is detected, containing 793 points, and $-\log_{10}(NFA_g) = 44.9$. On the right-most plot, the distribution is defined as the product of the marginal empirical distributions in the horizontal and vertical directions. There is still a single maximal meaningful group ($-\log_{10}(NFA_g) = 1.6$), but it now corresponds to the smallest group.

### 3.2 Segments

In the second example, groups are perceived as a result of the collaboration between two different features. Figure 6 shows 71 straight segments with different orientations, almost uniformly distributed in position. As expected, no meaningful cluster is detected in the space of position coordinates of the barycenters. In all the experiments, the number of rectangle sizes in each direction is 50. Thus $\#\mathcal{R} = 50^D$.

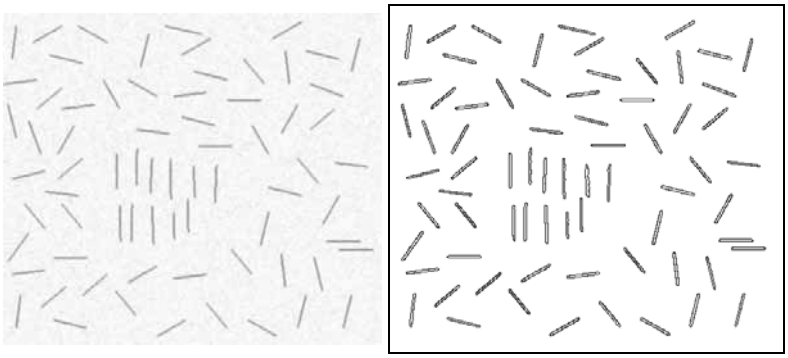


Figure 6: An image of a scanned drawing of segments, and its 71 maximal meaningful level lines [5].

If orientation is chosen as the only feature ($D = 1$), 8 maximal meaningful groups are detected, corresponding to the most represented orientations. None of these clusters exhibits a very low $NFA_g$. Only one of those group is conspicuous (the central one), but orientation is obviously not the only factor. Note that this group does not contain all the central segments. Indeed, their orientation slightly differ, and the group of 11 segments is not maximal. All the other groups are actually not perceived, because they are masked by the clutter made of all the other objects. However, one cannot object that they have a coherent direction.

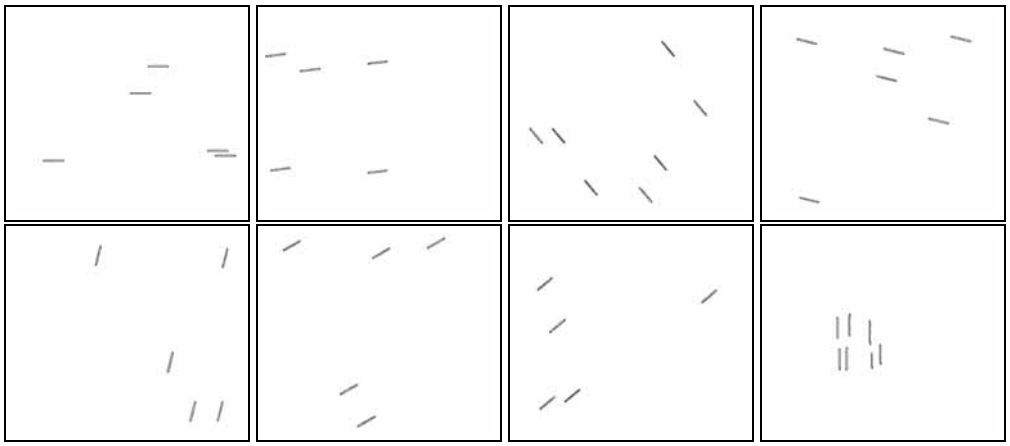


Figure 7: Grouping with respect to orientation: there are 8 maximal meaningful groups. $NFA_g$ range is between $10^{-1}$ and $10^{-5}$. The central group does not contain all the vertical segments, because the orientation is not very accurate. Hence, the maximal group containing these vertical segments does not include all the central objects. This means that orientation alone is not sufficient to detect this group. On the contrary, it allows to detect good groups, but their position is not coherent enough to make them conspicuous.

Now, let us see what happens when considering two features ($D = 2$, $\#\mathcal{R} = 2500$). In the space ($x$-coordinate, orientation), two maximal meaningful clusters are found (Fig. 8). As expected, the most meaningful is the group $G$ of 11 central vertical segments. Its $NFA_g$ is equal to $10^{-1.5}$, which is not that low. The second one is correct, but hardly meaningful $NFA_g = 0.3$. In the space ($y$-coordinate, orientation), the central group $G$ is splitted into two maximal meaningful clusters. They correspond to the two rows of segments composing $G$. The role of the merging criterion is decisive here. In the space ($y$-coordinate, orientation), the combination

of the maximality and the merging criterion yields that it is more meaningful to observe at the same time the two rows of segments than the whole $G$. This is coherent with the visual perception, since we actually see two lines of segments here. On the contrary, in the ($x$-coordinate, orientation) space, the merging criterion indicates that observing $G$ is more meaningful than observing simultaneously its children in the dendrogram. This decision is still conform with observation: no particular group within $G$ can be distinguished with regards to the $x$-coordinate. The same group is obtained in the space ($x$-coordinate, $y$-coordinate, orientation), with a lower $NFA_g = 10^{-3.4}$.

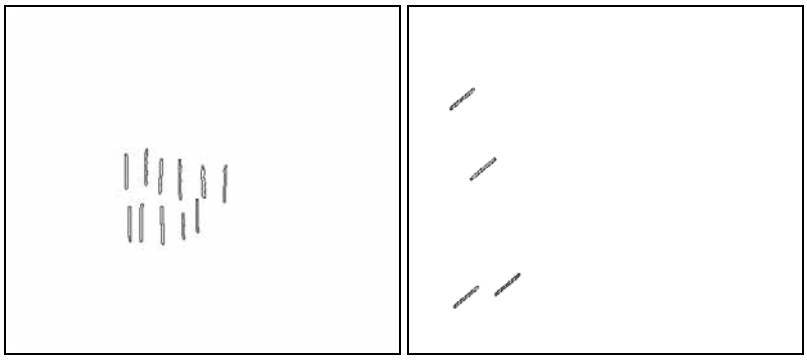


Figure 8: Grouping in the space ($x$-coordinate,orientation). There are two maximal meaningful groups. This time, the whole central group is detected ($NFA_g = 10^{-1.5}$), but there is still another group (which is a part of the 7th group in the orientation grouping (see Fig. 7)). However, its $NFA_g = 0.3$, which means that it is hardly meaningful. If grouping is done with respect to full 2D-position and orientation, only the central group is detected with $NFA_g = 10^{-3.4}$.

### 3.3 DNA image

The 80 objects in Fig. 9 are more complex, in the sense that more features are needed in order to represent them (diameter, elongation, orientation, *etc.*). It is clear that a projection on a single feature is not really enough to differentiate the objects. Globally, we see three groups of objects: the DNA marks, which share the same form, size and orientation; the numbers, all on the same line, almost of the same size; finally the elements of the ruler, also on the same line and of similar diameters. The position appears to be decisive in the perceptive formation of these groups.

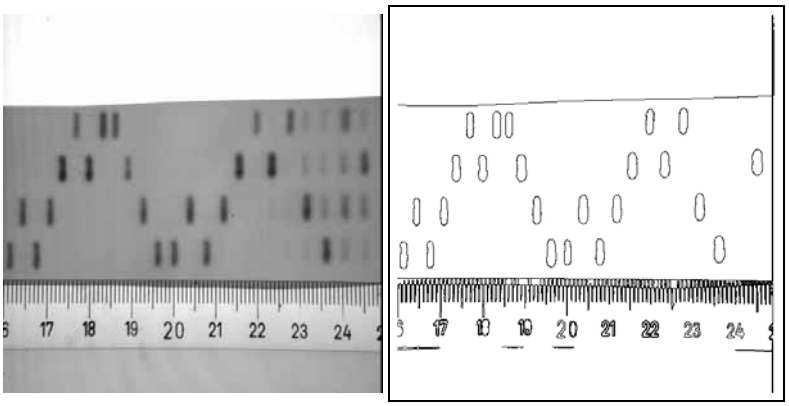



Figure 9: An image of DNA and its 80 maximal meaningful level lines [5].

In the space (diameter, $y$-coordinate), 6 maximal meaningful groups are detected (Fig. 10). Four of them correspond to the lines of DNA marks (from left to right and top-down), $-\log_{10}(NFA_g) = 2.6, 7.6, 6.4, 5.6$. The group of numbers contains 23 objects (a group of two digits sometimes contains three objects: the two

digits and a level line surrounding both of them) and $-\log_{10}(NFA_g) = 43$. The last group, composed of the vertical graduation of the ruler contains 31 objects and is even more meaningful, $-\log_{10}(NFA_g) = 54$.

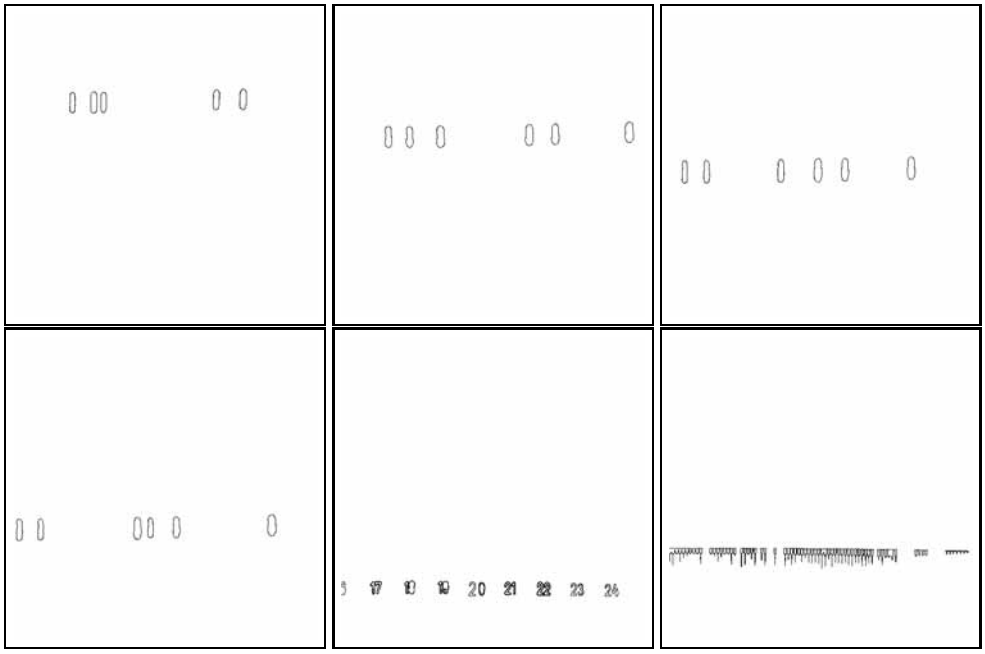


Figure 10: Grouping with respect to diameter and $y$ coordinate. Six groups are detected, 4 of which are rows of DNA marks. The last two ones correspond to the ruler. $-\log_{10}(NFA_g)$ range from 2.6 to 7.6 for the DNA. The last two groups are larger and are obviously more meaningful: $-\log_{10}(NFA_g) = 43$ and 54.

Now, let us give up considering the position information. Do we still see the DNA marks as a group? By taking several other features into account (see Fig. 11), the DNA marks form an isolated and very meaningful group: the combination of features (orientation, diameter, elongation, convexity coefficient) reveals the DNA marks as a very good maximal meaningful cluster ($NFA_g = 10^{-10}$). However, to our surprise, two other groups are also detected (though not very meaningful since their $NFA_g$ is about $10^{-1}$): the 1's and the 2's of the ruler. Let us detail how $\pi$, the law of the background model was estimated on the data itself: the marginal distribution of each characteristic is approximated by the empirical histogram. Then all the characteristics are assumed to be independent. Let us point out that the obtained distribution is not uniform at all.

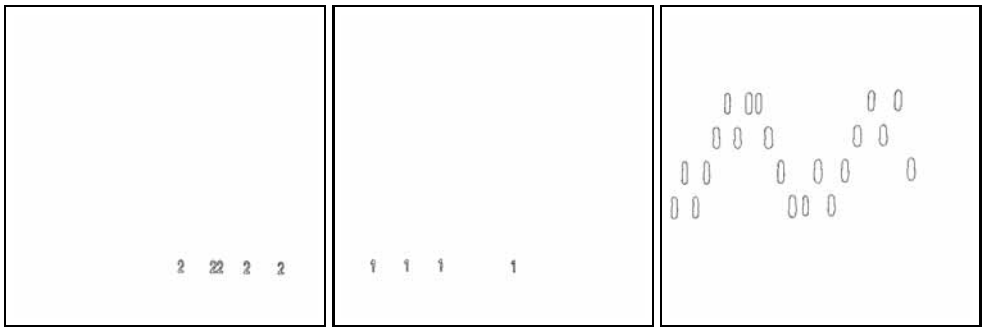


Figure 11: Grouping with respect to orientation, elongation, diameter, and a convexity coefficient. The DNA marks are the most meaningful group $NFA_g = 10^{-10}$, but the 1 and 2's also form groups, with $NFA_g$ close to 1.

# 4 Grouping spatially coherent matches for planar shape recognition

## 4.1 Why spatial coherence detection?

Looking at Fig. 12, everybody can obviously recognize on the bottom left image a detail of Picasso's painting *Guernica* shown on the top left image. However, the painting is incomplete and partially occluded in the bottom image. It is also deformed by the perspective view. Moreover, the compression rates are also different. Recognizing shapes which are observed from different viewpoints and are partially occluded requires shape descriptors to be discriminative enough, local or semi-local, and invariant to subgroups of the projective group [23, 24, 32]. Shape descriptors having this properties will be called *shape elements* in the sequel.

Assume now that instances of a query shape are present in a scene, and that a method to identify similar shape elements is available. It will certainly provide several correct pairings, but also some false ones; indeed, since shape elements only provide local information, two different objects having similar parts may present some shape elements that match. Thus, recognition requires finding a consistent set of pairings, that is, a set of pairing in a particular geometrical configuration.

In this framework, one possible strategy consists in associating with each pairing between shape elements the underlying transformation, and then detecting sets of pairings for which the underlying transformations are "close" in a certain sense.

## 4.2 Matching shape elements

In a sake of completeness, we briefly review the main steps of the shape elements extraction and matching algorithms described in [30] and that feed the grouping procedure described below. However, let us point out that the grouping procedure is applied independently from this particular procedure. A first observation is that the contours of objects in grey level images very well coincide, at least locally, with pieces of level lines (or isophotes). The converse is not always true: indeed, level lines provide a complete representation of a grey level image [29], and there are many of them in textures. Thus, a first step is to select a small subset of all the level lines of an image. In [5], an *a contrario* method is proposed, and the selected level lines are called *meaningful boundaries*. It allows to select about $1\%$ of the level lines of an image, without perceptual loss of shape content. These level lines are simple curves that are closed or meet the image border at their endpoints.

Shape recognition should be robust to partial occlusion. Hence, meaningful boundaries should be cut in smaller pieces, called *shape elements* that are to be recognized. Since geometric invariance is also required, the encoding of shape elements also has to be invariant. In [23, 30], an affine invariant encoding method is proposed. Let us remark that, in some cases, a similarity invariant method may be accurate enough. Along each meaningful line, local affine invariant frames are computed, based on affine invariant robust directions, as bitangent lines. Each local frame uniquely defines a system of coordinates. The coordinates of the points of a curve in this system of coordinates is affine invariant. In other terms, two curves differing from an affine transformation define different local frames. However, when described in their respective system of coordinates, they are located at the same position. Hence they define a piece of normalized curve, an *affine invariant shape element*. A single meaningful boundary usually contains several shape elements.

Now, given two images and the sets of their shape elements, how to find shape elements in common? Since shape elements are normalized, this recognition is naturally affine invariant. In [30], an *a contrario* dedicated method is proposed to match shape elements. A number of false alarms of a match is defined, and the matches with a low number of false alarms are kept.

Let $I$ and $I'$ be two images, referred to as the *target* image and the *scene* image. For each match between a shape element $\mathcal{S}$ in $I$ and a shape element $\mathcal{S}'$ in $I'$, a geometric transformation (a similarity or an affine transform) can be computed. In what follows, the parameters involved in these transformations are described, as well as the way they can be estimated, both for the similarity and the affine transformation cases.

The objective of this part is twofold: first, to prove that shape elements corresponding to a single shape can be accurately grouped together. Second, that this grouping procedure is robust enough to discard all false matches. The group NFAs' are usually very small. This makes the detection very reliable.

The overall strategy is as follows. In Section 4.3, the parameterization of similarities or general affine transformations is described. Section 4.4 applies the general clustering ideas presented in Sect. 2, first by defining a dissimilarity measure between transformations, then by defining a suitable background model on the sets of transformations.

Fig. 13 displays the shape elements common to these two images. Since no restriction is made on the affine distortion, a lot of normalized convex shape elements look quite the same. A unique affine transformation corresponds to each match between shape elements.

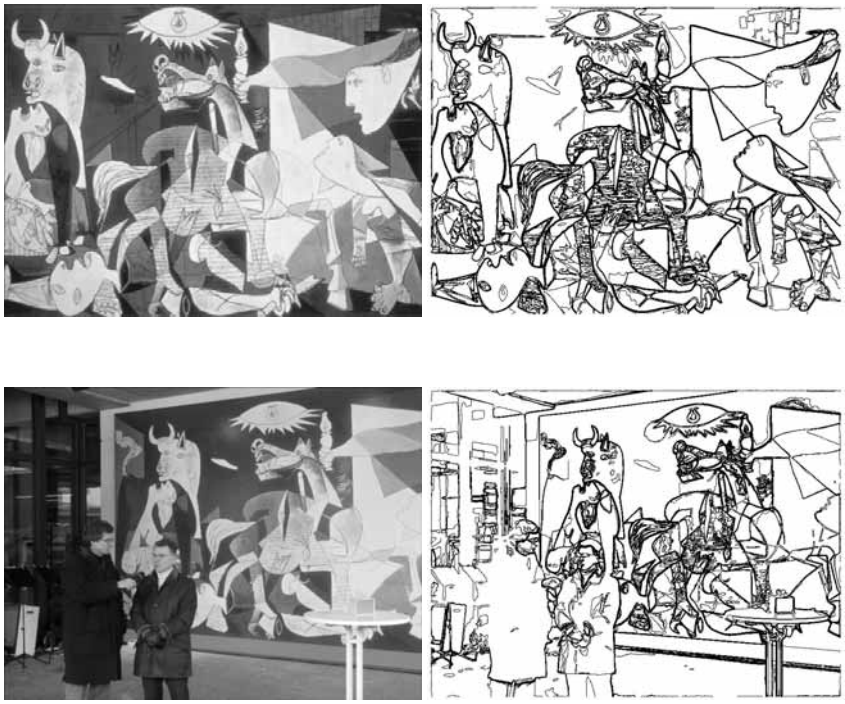


Figure 12: "Guernica" experiment. Original images and maximal meaningful level lines [5]. All these level lines are encoded into normalized affine invariant shape elements [30], based on robust directions as bitangent and flat parts. Top: target image, bottom: scene image.

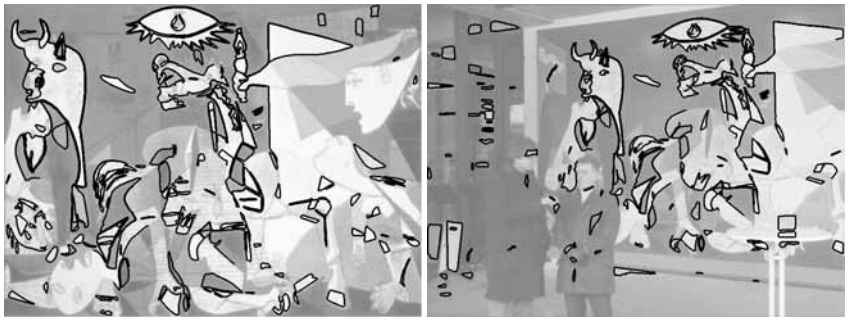


Figure 13: "Guernica" experiment: affine invariant meaningful matches [30]. Since all parallelograms differ from an affine transformation (*idem* for triangles or ellipses), there are many casual matches.

## 4.3 Describing transformations

### 4.3.1 The similarity case

Let $\mathcal{S}$ and $\mathcal{S}'$ be two matching shape elements. Recall that a shape element is a normalized piece of level line described in a local frame (See Fig. 14). A similarity invariant frame is completely determined by two points, or equivalently a point and a vector. This last representation will be chosen. A local frame is then given by a couple $(p, v)$ where $p$ gives the origin of the frame and $v$ gives its scale and orientation. Let us assume that $\mathcal{S}$ is related to $(p, v)$ and $\mathcal{S}'$ to $(p', v')$. Since $\mathcal{S}$ and $\mathcal{S}'$ match, they differ by a similarity transformation. Now, there exists a unique similarity mapping the local frame $(p, v)$ onto $(p', v')$. By using complex numbers notations, this similarity can be uniquely expressed as

$$\forall z \in \mathbb{C},\ \mathbf{T}(z) = az + b,\ \text{with } a = \frac{v'}{v} \text{ and } b = p' - ap, \tag{4.1}$$

with $(a, b) \in \mathbb{C}^2$.

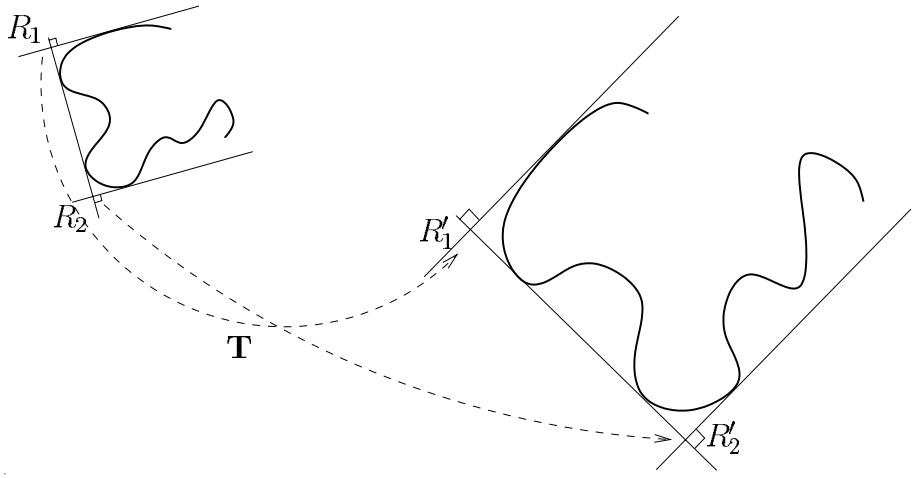



Figure 14: Two pieces of level lines and their corresponding local similarity frames. The similarity $\mathbf{T}$ maps $R_1$ into $R'_1$ and $R_2$ into $R'_2$. Equivalently the local frame, $(R_1, R_2)$ may be represented by $(p, v) = \left(\frac{R_1+R_2}{2}, R_2 - R_1\right)$.

### 4.3.2 The affine transformation case

Let us now consider the case of affine invariant normalization. Three non-aligned points are now necessary to define a local frame. Affine normalization of a piece of curve is performed by mapping these three points $\{R_1, R_2, R_3\}$ onto the triplet $\{(0, 0), (1, 0), (0, 1)\}$. Given another triplet $\{R'_1, R'_2, R'_3\}$ of non aligned points, there is a unique affine transform mapping $\{R_1, R_2, R_3\}$ on $\{R'_1, R'_2, R'_3\}$, again denoted by $\mathbf{T}$. There exists a unique $2 \times 2$ matrix $\mathbf{M}$ and a unique $(t_x, t_y) \in \mathbb{R}^2$ such that

$$\mathbf{T}(x, y) = \mathbf{M} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

Calculating $\mathbf{M}$ boils down to the solution of a $2 \times 2$ linear system. By the $QR$ decomposition [10], $\mathbf{M}$ can be written

$$\mathbf{M} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & \varphi \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}. \tag{4.2}$$

This decomposition is unique and completely determines $(\theta, \varphi, s_x, s_y)$ in $[0, 2\pi) \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$. Let us denote by $(x_{R_1}, y_{R_1})$ and by $(x'_{R_1}, y'_{R_1})$ the pair of coordinates of $R_1$ and $R'_1$ respectively. The transformation

parameters $T = (\theta, \varphi, s_x, s_y, t_x, t_y)$ are determined by elementary algebraic calculations. Again, the vector $T$ characterizes the transformation $\mathbf{T}$.

Without risk of ambiguity, one can adopt the same notation for similarities or affine transformations. In addition, since $T$ characterizes $\mathbf{T}$, both of them can be identified. Thus we write, for $X \in \mathbb{R}^2$, $T(X)$ instead of $\mathbf{T}(X)$.

Figure 15 shows three 2-D projections of the transformation points $T_k$ corresponding to the "Guernica" affine invariant meaningful matches of Fig. 13).

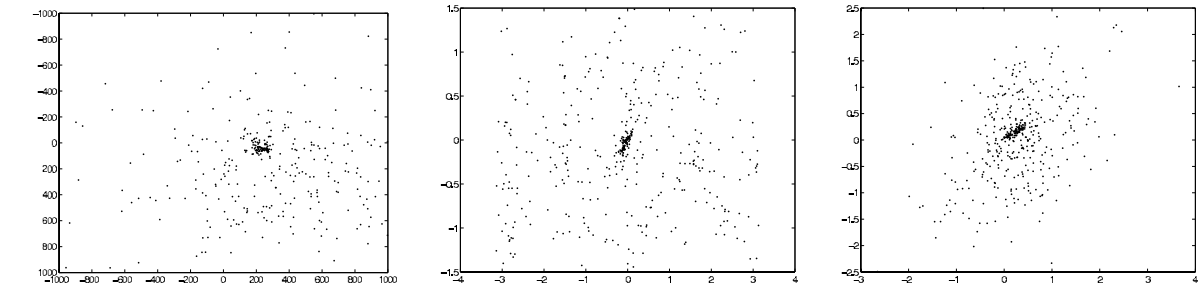


Figure 15: "Guernica experiment: Each point represents a transformation associated with an affine invariant meaningful match, described by 6 parameters. Each figure represents a two-dimenstional projection of the points, respectively $t_x$ vs. $t_y$ (translation coordinates), $\theta$ (rotation) vs. $\varphi$ (shear), and $\ln(s_x)$ vs. $\ln(s_y)$ (zooms in the $x$ and $y$) directions. The noise is mainly due to global shape elements that are very much alike up to affine transformations, and which do not belong to the same real shape. The main cluster is also spread because of the effect of perspective.

## 4.4 Meaningful clusters of transformations

The problem of planar shape detection is by now reduced to a clustering problem in the transformation space. According to Sect. 2, it is necessary to define

1. a dissimilarity measure between points in the transformation space,
2. a probability on the space of transformations,
3. a grouping strategy.

### 4.4.1 A dissimilarity measure between transformations

Defining a distance between transformations is not trivial, for two reasons. First, the magnitudes of the parameters of a transformation are not directly comparable. This problem is not specific to transformation clustering but general to clustering of any kind of data. Second, our representation of similarities or affine transformations does not behave well in a vector space. A sound distance is not necessarily derived from a norm.

**Definition 4.1 (similarity case)** *Let $(P_1, Q_1)$ (resp. $(P_1', Q_1')$) be the points determining the local frame of $\mathcal{S}_1$ in image $I$ (resp. $\mathcal{S}_1'$ in image $I'$). Let $T_1$ the unique similarity determined by $(P_1, Q_1)$ and $(P_1', Q_1')$. In the same way, let $T_2$ be the similarity determined from a match between the shape elements with frames $(P_2, Q_2)$ and $(P_2', Q_2')$ in $I$ and $I'$. We call dissimilarity measure between $T_1$ and $T_2$,*

$$d_S(T_1, T_2) = \max\left\{\|T_1(P_i) - T_2(P_i)\|,\ \|T_1(Q_i) - T_2(Q_i)\|,\ i \in \{1, 2\}\right\}. \tag{4.3}$$

For completeness, let us define a dissimilarity between affine transforms.

**Definition 4.2 (affine case)** *Let $T_1$ (resp. $T_2$) be an affine transform determined by two shapes elements $(\mathcal{S}_1, \mathcal{S}'_1)$ (resp. $(\mathcal{S}_2, \mathcal{S}'_2)$) matching from $I$ to $I'$. Let also $(P_1, Q_1, R_1)$ and $(P'_1, Q'_1, R'_1)$ (resp. $(P_2, Q_2, R_2)$ and $(P'_2, Q'_2, R'_2)$) the points determining the local frame of $\mathcal{S}_1$ and $\mathcal{S}'_1$ (resp. $\mathcal{S}_2$ and $\mathcal{S}'_2$). We set*

$$d_A(T_1, T_2) = \max\left\{\|T_1(P_i) - T_2(P_i)\|,\ \|T_1(Q_i) - T_2(Q_i)\|,\ \|T_1(R_i) - T_2(R_i)\|,\ i \in \{1,2,3\}\right\}. \quad (4.4)$$

### 4.4.2 Background model: the similarity case

In order to apply the detection framework of Sect. 2, a background law is first needed. A data point here is a similarity transformation represented by a pair of complex numbers $(a, b) \in \mathbb{C}^2$. The purpose of this section is to devise a sound background law $\pi$ on the set of similarity transformations. To this aim, recall that $(a, b)$ is determined by two local frames in the images to be matched, respectively $(p, v)$ and $(p', v')$. Let us now assume that these observations are the realization of a random variable $(P, V, P', V') \in \mathbb{C}^4$. It is natural to assume that the position, the size and the orientation of an object are independent. This is certainly sound, up to some border effects. In addition, two images which do not contain common shapes also can be assumed independent. This leads us to the following independence assumption for the background model.

**(A')** *Consider a random model image $\mathcal{I}$ and a random scene image $\mathcal{I}'$. Then the random variables $P$, $|V|$, $\arg V$, $P'$, $|V'|$, $\arg V'$ associated with matches between both images are mutually independent.*

The marginal laws of the six previous random variables can easily be learned from the two images. Hence, the law of $(P, V, P', V')$ is assumed to be known. By (4.1), such a 4-tuple uniquely defines a random similarity pattern denoted by $(A, B)$, where $A$ represents the rotation and zoom, and $B$ the translation. The background law $\pi$ is nothing but the distribution of $(A, B)$. The expression of $(A, B)$ as a function of $(P, V, P', V')$ is explicit and given by

$$(A, B) : (P, V, P', V') \mapsto \left(\frac{V'}{V}, P' - \frac{V'}{V}P\right).$$

The background law $\pi$ is the image of the law $(P, V, P', V')$ by this application. It is also clear that $A$ and $B$ are not independent. Nevertheless, by definition of the conditional law,

$$d\pi(a, b) = d\pi^B(b\,|A = a)\, d\pi^A(a), \quad (4.5)$$

where $\pi^A$ is the marginal of $A$ and $\pi^B(\cdot\,|A = a)$ is the law of $B$ knowing $A = a$. Since $|A| = |V'|/|V|$ and $\arg A = \arg V' - \arg V \mod (2\pi)$, these two variables are independent under Assumption **(A')**. Thus, the distribution $\pi^A$ can easily be computed. Moreover, it turns out that $A$ is independent from $P$ and $P'$. Hence, the law of $B = P' - AP$, conditionally to $A = a$ is the law of $P' - aP$, which can also be easily computed under **(A')**. The background law $\pi$ follows from (4.5).

In practice, the computation of $\pi$ between two images is as follows:

1. Compute all the shape elements of model and target images.
2. Compute the empirical laws of $P, V, P', V'$ giving the position, the scale and the orientation of the local frames related to shape elements in the two images. Under the independence assumption **(A')**, this yields the law of the background model $(P, V, P', V')$.
3. Under the same assumption, compute the empirical laws of $|A| = \frac{|V'|}{|V|}$ and $\arg A = \arg V' - \arg V \mod (2\pi)$.
4. For each value $a$ of $A$ with non null frequency, compute the empirical distribution of $P' - aP$.

The probability of a region $R$ is then given by approximating the integral

$$\pi(R) = \int_R d\pi^B(b|A = a)\, d\pi^A(a).$$

A few words about the estimation of the background model: one would expect $\arg A$ to be uniformly distributed in $[-\pi, \pi)$, and this belief was experimentally confirmed, although the horizontal and vertical directions may sometimes be privileged. (See Fig. 17 and experiments.) The distribution of the zoom factor $|A|$ is instead far from being uniform. There is no way to figure out a realistic *a priori* distribution for $|A|$, or for $B$ given $A$. The background model distributions must be learned from the scene and target images.
*Remark.* The ideas presented here also hold for the affine transformation clustering. For this case, $\theta$, $\varphi$, $s_x$ and $s_y$ are considered to be mutually independent. Their distributions can be learned empirically, as well as the joint probability of $(t_x, t_y)$ given $(\theta, \varphi, s_x, s_y)$. This construction, experimentally satisfying though it is (see the experiments), has no righteous theoretical justification. The problem of finding the right independent marginal variables in the affine case is left open.

### 4.4.3 Grouping strategy

There are several methods to build a binary tree from a dataset and a dissimilarity measure. In this paper, the minimal spanning tree is used. Its construction uses a classical *single linkage algorithm* working as follows. The dissimilarity $d$ between two datapoints is extended to any pair of disjoint sets of datapoints $A$ and $B$ by setting

$$d(A, B) = \min_{(a,b)\in(A,B)} d(a, b).$$

A binary tree is constructed by the following iterative process: each datapoint is taken as a leaf-node. Then merge the closest pair of nodes into a single node. Repeat this until all nodes have been merged in the whole dataset. By replacing the "$\min$" by a "$\max$" in the above formula, a maximal spanning tree is obtained instead. Choosing one tree or the other may be very application dependent but none is universally better than the other [17].

## 5 Experimental results

The consistency of the previous definitions is now empirically checked. All the experiments will be performed with a pair of images. It is worth summarizing the steps leading a complete experimental setting for shape recognition.

1. Extraction of all the images level lines. An efficient algorithm due to Monasse and Guichard is used [29]. There are typically $10^5$ level lines in a $512 \times 512$ image.

2. Selection of the most meaningful level lines [3, 5]. This step can be viewed as a compression of the shape information of the image. Only a small set of level lines (between 100 and 1000) is selected by this fully automatic procedure.

3. Encoding of shape elements: robust directions (bitangent or flat parts) are computed on the level lines. Based on *all* those directions, local frames are computed, and pieces of level lines are described in normalized frames, typically a few thousands per image [22, 30].

4. The method of [30] is then applied and yields a set of $M$ pairs of matching shape elements, one in the target image and one in the scene image. A fundamental hypothesis for the *a contrario* detection of groups is that, under the *background model*, transformation points are mutually independent. In order to comply with this hypothesis, a greedy algorithm that eliminates matched shape elements which share a large piece of curve with other pairs of matching shape elements.

5. A background model $\pi$ on the set of similarities or on the set of affine transforms $E$ is built according to Sect. 4.4.2.

6. The transforms $T_1, \ldots, T_M$ associated with the matching pairs form a point data set in $E$. From this set, a clustering tree is built according the dissimilarity measures of Definitions 4.1 or 4.2.

7. Maximal meaningful groups are computed by Def. 2.5.

The final outcome of the shape identification method of this paper is, for each pair of images, a set of maximal meaningful clusters. Each cluster is likely to correspond to an identified shape. One can display for each cluster its associated shape elements. If the grouping is correct, this set of shape elements must correspond to a *matching shape* in both the target image and the scene image. In practice, the identified shapes have dramatically low NFA's. Thus, they yield an overwhelming certainty about identification. This certainty is, however, not fully unambiguous because of the Strobe effect. Indeed, shapes often have self-similar parts: windows, or rows of windows in a building are a good example. Other examples are given by symmetries. For instance, the letter N is self-similar by a $\pi$ rotation. In these cases, two or more very meaningful groups can be found, each one corresponding to a shape self-similarity. Such self-similarities can, however, easily be anticipated by a previous comparison of the target image with itself. This comparison can be performed by the above algorithm. The main group will then correspond to the global match of the shape with itself and the other groups to Strobe effects between parts of the shape.

## 5.1 A single group

Figure 16 depicts the maximal meaningful groups for the "Guernica" experiment. There is one single maximal meaningful group, with $-\log_{10}(NFA_g) = 196.23$. Hence grouping gives a dramatic confidence in detections, while all the false matches are eliminated. Figure 17 shows the learned distribution of the zoom factors in the $x$ and $y$ directions as well as the shear and rotation angle. This last one is not perfectly uniform in this case, because the vertical and horizontal directions are privileged in these geometrical images. Figure 18 shows the meaningful cluster.

Figure 16: "Guernica" experiment: a single maximal meaningful group was detected. Zoom on the matches of the group for the target image (left) and the scene image (right). The group is composed by 117 good matches, and its $-\log_{10}(NFA_g)$ is 196.23.

## 5.2 Two different groups

The similarity invariant procedure is applied in the same way to the images of Fig. 19. Two maximal meaningful groups are detected: the faces and the title. The corresponding points in the similarity space are displayed on Fig. 20. The two groups with their different translation and their different scaling are clearly visible this time.

The indivisibility criterion (2.4) decides that two separate groups (the actors' faces on the one hand and the word "Casablanca" on the other hand) are a better representation than a single large group containing both groups. Indeed, while the large group in Fig. 21 has a lower $NFA_g$ than one of its children ($10^{-7}$), it is not indivisible. Indeed, the $NFA_g$ of its two children are $10^{-7.6}$ and $10^{-6.6}$. By Prop.2.4, the largest group is not indivisible, and thus cannot be maximal.

The examination of the transformation histograms (Fig. 22) shows that the rotation angle is nearly uniformly distributed. The zooming factor, on the other hand, does not have an intuitive distribution. The translation has to be learned conditionnally to the rotation and the zoom. The last two plots are the two-dimensional distribution

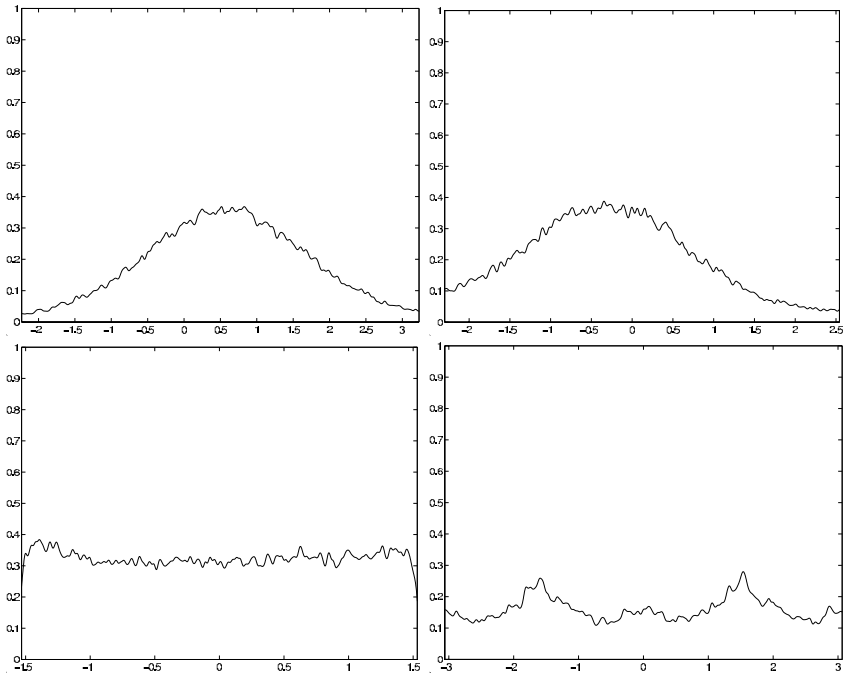


Figure 17: Empirical histograms for affine invariant matching for the experiment of Fig. 12. On the first row, the empirical zoom factors in the $x$ and $y$ direction (logscale), which are image dependent. On the second row, the distribution of the shear and the rotation angle. The shear is basically uniform, but the rotation exhibits some peaks around $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ because of the numerous lines in the image.

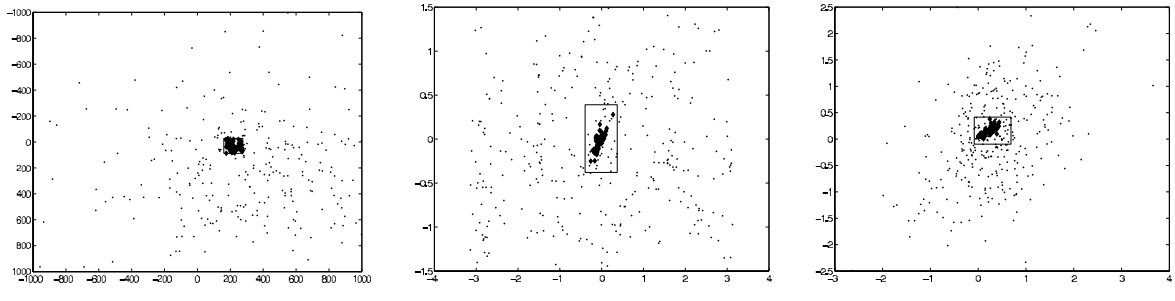
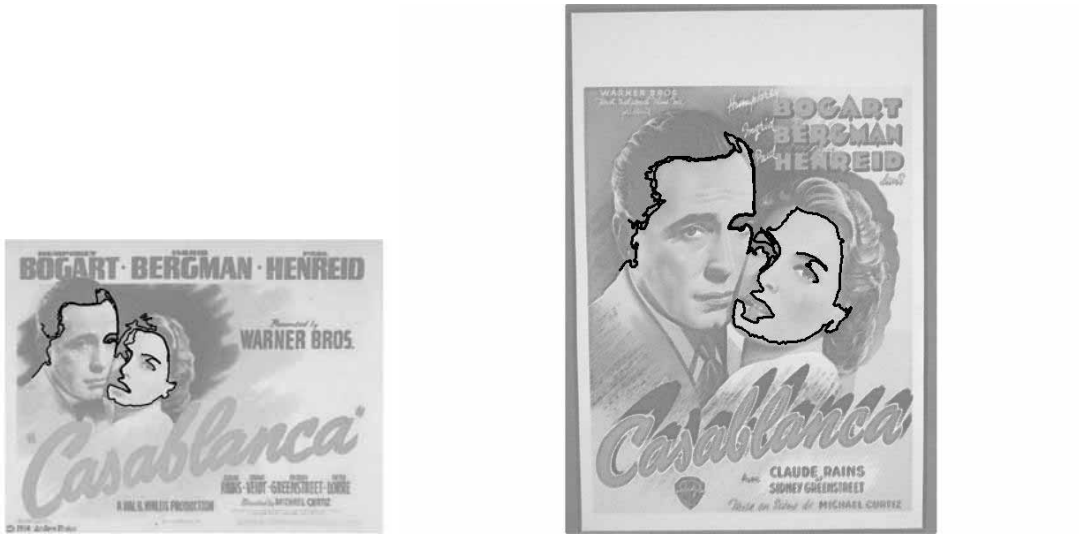


Figure 18: "Guernica experiment: data points of Fig. 15, where the points corresponding to the only affine invariant group are represented with diamonds. The boundaries of the corresponding hyperrectangle are drawn.


(a) First maximal meaningful group: 12 meaningful matches, $-\log_{10}(NFA_g) = 7.6$




(b) Second maximal meaningful group: 7 meaningful matches, $-\log_{10}(NFA_g) = 6.62$

Figure 19: "Casablanca" experiment: there are exactly two maximal meaningful groups, corresponding to the faces and the title. The relative scale of the images presented above is the same as the original one. One should note that the faces and the title actually lie in different relative positions and scales.

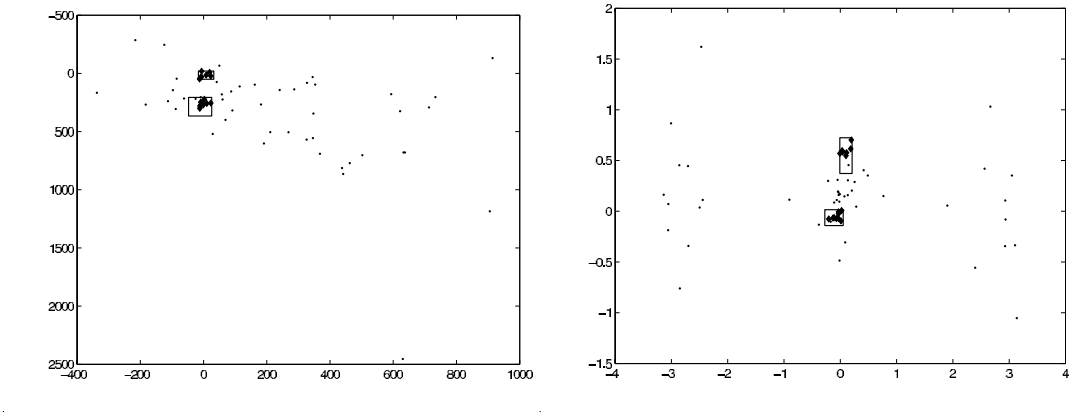



Figure 20: Casablanca experiment. Meaningful clusters in the similarity space. Left: projection in the translation dimensions. Right: projection on the rotation and zoom (log scale) axes. In this case, two clusters are clearly visible. Their position but also their scale is different.


Figure 21: "Casablanca" experiment. Meaningful group corresponding to the merging of groups in Fig. 19. This group contains 23 meaningful matches, and its $-\log_{10}(NFA_g)$ is 7.0. It is more meaningful than the faces group, but it is not maximal. Note the "Strobe" effect of the lower part of "cASablanca" in the first image that matches with "casABlanca" in the second one.

of the translation, conditioned by the rotation and zoom of the two detected maximal meaningful groups. As can be seen, these distributions are not simple and cannot be deduced from one another by a single scaling.

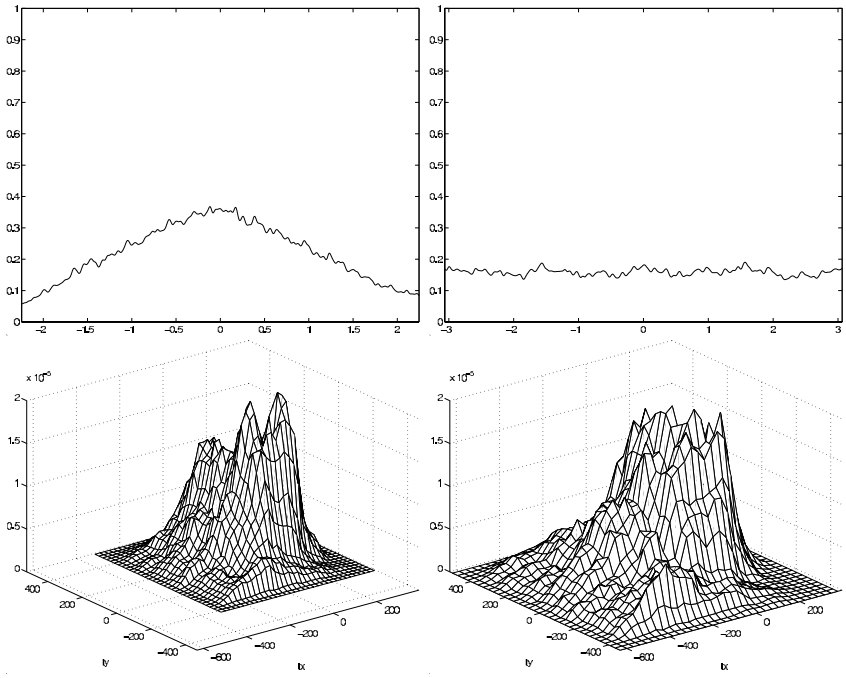


Figure 22: Empirical histograms for similarity invariant matching for the experiment of Fig. 19. On the first row, the log-empirical zoom factor $\ln(|a|)$ and the rotation angle $\arg a$. This last one is nearly uniform in this case. On the bottom row, the distribution of the translation vector, conditioned by two different values of the couple $(\ln(|a|), \arg a)$. These values correspond to the two maximal groups that are depicted on Fig. 19. Since the scales are different, so are the distributions.

### 5.3 Detecting multiple groups

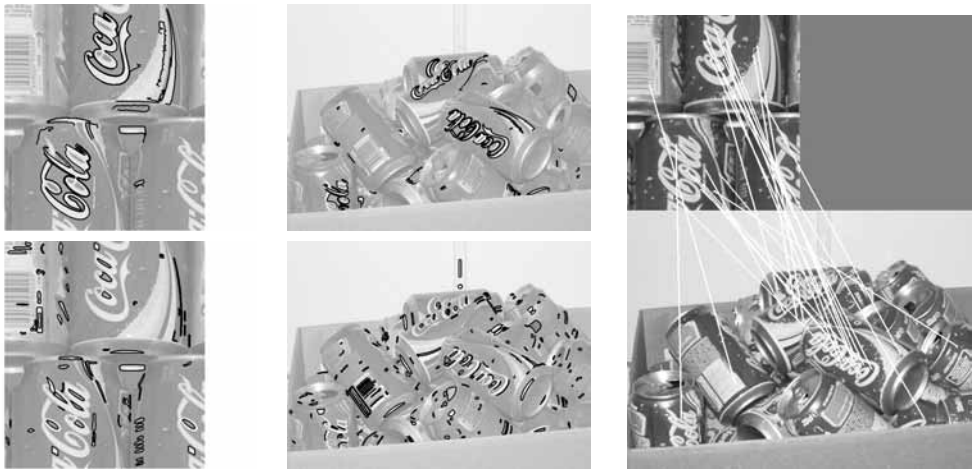
The next example illustrates the performance of the proposed methodology in detecting multiple groups in an image. Two images containing multiple occurrences of parts of the Coca-Cola logo are compared (Fig. 23). Figure 24 shows the affine invariant meaningful matches. Five groups are detected. The corresponding shape elements are displayed for each group in Fig. (25) and (26). The $NFA_g$ of maximal meaningful groups are reported in Tab. 1. The three first groups are very meaningful, while the two other $NFA_g$ are much closer to 1: about $10^{-4}$.

| Group nb. | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| nb. of matches | 15 | 7 | 5 | 6 | 4 |
| $-\log_{10}(NFA_g)$ | 20.6 | 16.7 | 5.8 | 4.0 | 3.0 |

Table 1: "Coca-Cola" experiment: $NFA_g$ for the maximal meaningful groups in Fig. 25 and 26.

Maximal meaningful groups can be used for registration. Since a group contains several points (*i.e.* several affine transforms), a standard least squares procedure allows to compute the best plane projective transform describing the group. As can be seen on the left parts of Fig. 27 and 28, this registration is very accurate since no blur is visible when the two registered images are superposed. Another way to check the accuracy of the registration is to find all the pieces of level lines in common in the two images, as made as follows. The two images are first registered. All pieces of meaningful level lines with a length $l$ are parameterized by their arc-length. If for two pieces $C_1$ and $C_2$, belonging to the first and second image satisfy $|C_1(s) - C_2(s)| < \delta$ for all

Figure 23: "Coca-Cola" experiment: original images and maximal meaningful level lines. Top: images, bottom: meaningful level lines [5].



(a) (b)

Figure 24: "Coca-Cola" experiment. (a) meaningful matches, with local encoding (top) and globally encoding (down). Number of tests: $1.57\ 10^7$ ($591$ shape elements in the target image, $26,621$ in the scene image). (b) As a comparison, matches with Lowe's SIFT features [24], code courtesy of D. Lowe; 23 matches, represented by white segments joining the matching locations between the two images. The algorithm is less accurate in this case, but much faster. Since both matching algorithms work on parts of the images (which is mandatory if robustness to occlusion and ability to detect multiple groups are required), casual matches are inevitable. The grouping phase attempts to build a more global context, and to discard those false matches. Contrary to Lowe's grouping algorithm (not presented in this figure) which is based on a Hough Transform clustering, the method proposed in this paper does not depend on bins quantization and has automatic detection thresholds.

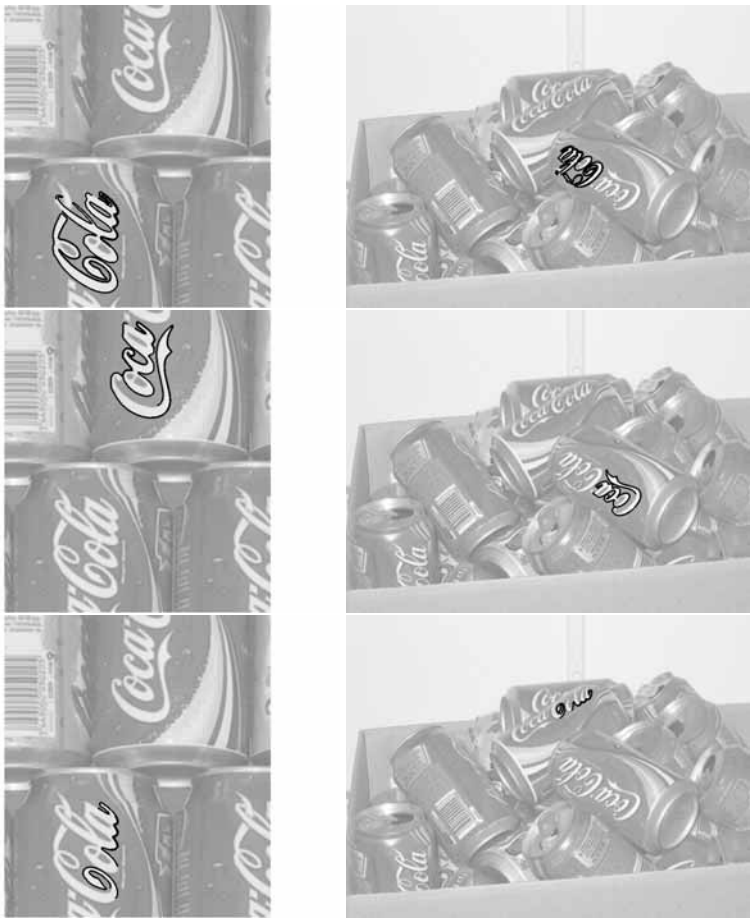


Figure 25: "Coca-Cola" experiment: first three maximal meaningful groups (among 5). Their $-\log_{10}(NFA)$ are respectively 20.6, 16.7, 5.8, showing that they are indeed very meaningful.

Figure 26: "Coca-Cola" experiment: maximal meaningful groups (last two among five). Their $-\log_{10}(NFA)$ are respectively $4.0$ and $3.0$.

$s \in (0, l)$ then, keep $C_1$ and $C_2$. In the experiments, $l = 40$ and $\delta = 4$. All these pieces of level lines that are close to each other are plotted on the right part of Fig. 27 and 28.

Figure 27: "Coca-Cola" experiment: registration with respect to the meaningful groups. Because there are several affine matches per group, one can compute the best projective mapping by a standard least squares method. The projective transformation is used to superpose the two images (on the left). On the right side, pieces of level lines that are close to each other in the registered images (see text).

## 6 Conclusion

This paper presents a general setting of detection and selection of groups in a collection of data points. The meaningful groups are those that cannot be generated by chance. As such, they can be defined as large deviations from an independence hypothesis of the points they contain. This allows to define a measure of meaningfulness, the number of false alarms. Among all the meaningful groups, only those which cannot be split into two smaller groups are relevant. The same kind of methodology can lead to the selection of these maximal meaningful groups. This framework is then applied to the grouping of transformations resulting from a preliminary local matching algorithm. The method is less sensitive to quantization than Hough Transform type algorithms, because the size of the region leading to the most meaningful event is automatically chosen. Let us point out that the present method intends to detect cluster with "no shape", *i.e.* groups of points that should be equal but differ because of noise. In particular, it needs further work to deal with clusters with holes, or nested. Because of the preliminary clustering step, it also much depends on the used distance, but there does not seem to be a choice which is completely independent of the application. However, the NFA calculation should be adapted to more general types of groups.

Figure 28: "Coca-Cola" experiment: maximal meaningful groups (last two among five). Their $-\log_{10}(NFA)$ is respectively 3.39 and 4.60. Both of them correspond to a Strobe effect, since the lower part of "oca" is identical to the lower part of "ola". Left: the registered images. Right: registered pieces of level lines.

The method could be used to find out the characteristics that are really relevant to form perceptual groups in a set of objects. How to select the characteristics to obtain the most meaningful groups? Another application where these clustering procedures are proposed is the analysis of visual motion [38], where the purpose is to detect spatio-temporal coherence. Elementary types of motions (ideal zooming, pure rotation, rectilinear motion) are parameterized, and local observations are grouped with respect to these criterions. Works in progress are exposed in [35].

*Acknowledgments*. This work was partially financed by the Centre National d'Etudes Spatiales, the Centre National de la Recherche Scientifique, the Office of Naval research under grant N00014-97-1-0839 and the Ministère de la Recherche (project ISII-RNRT).

# A Proofs

## A.1 Proof of Prop. 2.1

A careful notation is needed. Fix $1 \leqslant j \leqslant M$ and $R' \in \mathcal{R}$. We note:

- $X = (X_1, \ldots, X_M)$, the background process and $d\Pr(x_1, \ldots, x_M) = d\pi(x_1) \cdots d\pi(x_M)$ its distribution.
- $x = (x_1, \ldots x_M)$ a set of $M$ points in $E$
- $X^j = (X_1, \ldots, X_M)$ with $X_j$ omitted in the list
- $x^j = (x_1, \ldots, x_M)$ with $x_j$ omitted in the list
- $d\pi^j(x^j) = d\pi(x_1) \ldots d\pi(x_M)$ with $d\pi(x_j)$ omitted in the product
- $\Pr^j$ the joint marginal of $\Pr$ with respect to $X^j$
- $K(X^j, X_j, R')$, number of points in the list $X^j$ belonging to $X_j + R'$.

**Lemma A.1** *For every $x_j \in E$,*

$$\Pr{}^j\left(\mathcal{B}(M-1, K(X^j, x_j, R'), \pi(x_j + R')) < \frac{\varepsilon}{\#\mathcal{R} \cdot M}\right) < \frac{\varepsilon}{\#\mathcal{R} \cdot M}.$$

*Proof.* Since $x_j$ and $R'$ are fixed, $K(X^j, x_j, R')$ is a random variable whose survival function is exactly $k \mapsto F(k) = \mathcal{B}(M-1, k, \pi(x_j + R'))$. Hence $\Pr{}^j(F(K(X^j, x_j, R') < t) < t$, proving the result. □

*Proof.* [of Prop. 2.1] Let us note

- The Bernoulli variable

$$Y_{j,R'} = \begin{cases} 1 & \text{if } \#\mathcal{R} \cdot M \cdot \mathcal{B}(M-1, K(X^j, X_j, R'), \pi(X_j + R')) < \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

- $S = \sum_{i,R'} Y_{j,R'}$ the number of $\varepsilon$-meaningful regions.

We then have by Fubini theorem

$$\begin{aligned} \Pr(Y_{j,R'} = 1) &= \mathbb{E}(Y_{j,R'}) \\ &= \int_E d\pi(x_j) \int_{E^{M-1}} \mathbb{1}_{\{\#\mathcal{R} \cdot M \cdot \mathcal{B}(M-1, K(x^j, x_j, R'), \pi(x_j + R')) < \varepsilon\}} d\pi^j(x^j) \\ &= \int \Pr{}^j\left(\mathcal{B}(M-1, K(X^j, x_j, R'), \pi(x_j + R')) < \frac{\varepsilon}{\#\mathcal{R} \cdot M}\right) d\pi(x_j) \\ &\leqslant \frac{\varepsilon}{\#\mathcal{R} \cdot M}, \end{aligned}$$

where Lemma A.1 has been used in the last inequality. A region $R$ is $\varepsilon$-meaningful if and only if $R = X_j + R'$ for some point $X_j$ and some region $R' \in \mathcal{R}$ and if $Y_{j,R'} = 1$. Hence,

$$\mathbb{E}(S) = \sum_{j,R'} \mathbb{E}(Y_{j,R'}) < \sum_{j,R'} \frac{\varepsilon}{\#\mathcal{R} \cdot M} = \varepsilon. \qquad \square$$

## A.2 Proof of Prop. 2.3

Let $1 \leqslant i \neq j \leqslant M$ and $R',\ R'' \in \mathcal{R}$. Now, two tested regions $x_i + R'$ and $x_j + R''$ may intersect and we have to deal with this possibility. We note

- $X = (X_1, \ldots, X_M)$, background process
- $x = (x_1, \ldots x_M)$ a set of $M$ dots in $E$
- $X^{ij} = (X_1, \ldots, X_M)$ with $X_i,\ X_j$ omitted in the list
- $x^{ij} = (x_1, \ldots, x_M)$ with $x_i,\ x_j$ omitted in the list
- $X_{ij} = (X_1, \ldots, X_M)$ with $X_j$ and $X_j$ replaced by $x_i$ and $x_j$
- $d\pi^{ij}(x^{ij}) = d\pi(x_1) \ldots d\pi(x_M)$ with $d\pi(x_i)$ and $d\pi(x_j)$ omitted in the product
- $\Pr^{ij}$ the joint marginal of $\Pr$ with respect to $x^{ij}$
- $R_i = X_i + R'$, $R_j = X_j + R''$
- $K(X, i, j, R', R'') =$ number of points among $X^{ij}$ belonging to $R_i \setminus R_j = (X_i + R') \setminus (X_j + R'')$.

- $K_i = K(X, i, j, R', R'')$, $K_j = K(X, j, i, R'', R')$
- $\tilde{K}_i = K(X_{ij}, i, j, R', R'')$, $\tilde{K}_j = K(X_{ij}, j, i, R'', R')$.
- $k_i = K(x, i, j, R', R'')$, $k_j = K(x, j, i, R'', R')$
- $\pi_i = \pi((x_i + R') \setminus (x_j + R''))$, $\pi_j = \pi((x_j + R'') \setminus (x_i + R'))$
- $\Pi_i = \pi((X_i + R') \setminus (X_j + R''))$, $\Pi_j = \pi((X_j + R'') \setminus (X_i + R'))$
- $\epsilon = \frac{2\varepsilon}{M(M-1)^2(\#\mathcal{R})^2}$.

**Lemma A.2** *For every $x_i, x_j \in E$,*

$$\Pr{}^{ij}\left[\mathcal{M}(M-2, \tilde{K}_i, \tilde{K}_j, \pi_i, \pi_j) < \epsilon\right] < (M-1)\epsilon.$$

*Proof.* The proof extends the arguments used for Lemma A.1. We have

$$\begin{aligned}
&\Pr{}^{ij}\left[\mathcal{M}(M-2, \tilde{K}_i, \tilde{K}_j, \pi_i, \pi_j) < \epsilon\right] \\
&\quad = \sum_{(k_i,k_j)|\mathcal{M}(M-2,k_i,k_j,\pi_i,\pi_j)<\epsilon} \Pr{}^{ij}(\tilde{K}_i = k_i, \tilde{K}_j = k_j) \\
&\quad = \sum_{(k_i,k_j)|\mathcal{M}(M-2,k_i,k_j,\pi_i,\pi_j)<\epsilon} \binom{M-2}{k_i, k_j} \pi_i^{k_i} \pi_j^{k_j} (1-\pi_i-\pi_j)^{M-2-k_i-k_j}.
\end{aligned}$$

Let

$$k_i(\epsilon, k_j) = \inf\{0 \leqslant k \leqslant M | \mathcal{M}(M-2, k, k_j, \pi_i, \pi_j) < \epsilon\},$$

with the useful conventions $\mathcal{M}(M-2, k, k_j, \pi_i, \pi_j) = 0$ and $\binom{M-2}{k,k_j} = 0$ if $k \geqslant M-1-k_j$. The map $k \mapsto \mathcal{M}(M-2, k, k_j, \pi_i, \pi_j)$ being monotone, one has

$$\mathcal{M}(M-2, k, k_j, \pi_i, \pi_j) < \epsilon \Leftrightarrow k \geqslant k_i(\epsilon, k_j). \tag{A.1}$$

Summarizing and using the definition of $k_i(\epsilon, k_j)$,

$$\begin{aligned}
&\Pr{}^{ij}\left[\mathcal{M}(M-2, \tilde{K}_i, \tilde{K}_j, \pi_i, \pi_j) < \epsilon\right] \\
&\quad = \sum_{k_j=0}^{M-2} \sum_{k=k_i(\epsilon,k_j)}^{M-2} \binom{M-2}{k, k_j} \pi_i^k \pi_j^{k_j} (1-\pi_i-\pi_j)^{M-2-k-k_j} \\
&\quad \leqslant \sum_{k_j=0}^{M-2} \sum_{k=k_i(\epsilon,k_j)}^{M-2} \sum_{l=k_j}^{M-2} \binom{M-2}{k, l} \pi_i^k \pi_j^l (1-\pi_i-\pi_j)^{M-2-k-l} \\
&\quad = \sum_{k_j=0}^{M-2} \mathcal{M}(M-2, k_i(\epsilon, k_j), k_j, \pi_i, \pi_j) < (M-1)\epsilon. \qquad \square
\end{aligned}$$

*Proof.* Let us note for $R' \neq R''$,

- The Bernoulli variable

$$Y_{i,j,R',R''} = \begin{cases} 1 & \text{if } \frac{M(M-1)^2(\#\mathcal{R})^2}{2} \cdot \mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j) < \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

- $S = \sum_{i,j,\{R',R''\}} Y_{i,j,R',R''}$ the number of $\varepsilon$-meaningful pairs of different regions.

By Fubini theorem,

$$\begin{aligned}
\mathbb{E}(Y_{i,j,R',R''}) &= \Pr(Y_{i,j,R',R''} = 1) \\
&= \int_{E^2} d\pi(x_i) d\pi(x_j) \int_{E^{M-2}} \mathbb{1}_{\{\mathcal{M}(M-2,k_i,k_j,\pi_i,\pi_j)<\epsilon\}} d\pi^{ij}(x^{ij}) \\
&= \int_{E^2} d\pi(x_i) d\pi(x_j) \Pr^{ij}(\mathcal{M}(M-2,\tilde{K}_i,\tilde{K}_j,\pi_i,\pi_j) < \epsilon) \\
&< (M-1)\epsilon,
\end{aligned}$$

where Lemma A.2 has been used in the last inequality and $\epsilon = \frac{2\varepsilon}{(\#\mathcal{R})^2 M(M-1)^2}$. Finally,

$$\begin{aligned}
\mathbb{E}(S) &= \sum_{i,j,\{R',R''\}} \mathbb{E}(Y_{i,j,R',R''}) \\
&< \sum_{i,j,\{R',R''\}} (M-1)\epsilon \\
&= \sum_{i,j,R'\neq R''} \frac{2\varepsilon}{(\#\mathcal{R})^2 \cdot M(M-1)} = \varepsilon. \qquad \square
\end{aligned}$$

# References


[1] D.H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[2] H.H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2:77–108, 1985.

[3] F. Cao, P. Musé, and F. Sur. Extracting meaningful curves from images. *Journal of Mathematical Imaging and Vision*, 22(2-3):159–181, 2005.

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

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

[6] 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.

[7] P.A. Devijver and J. Kittler. *Pattern recognition - A statistical approach*. Prentice Hall, 1982.

[8] R . C. Dubes. How many clusters are best? – an experiment. *Pattern Recognition*, 20(6):645–663, 1987.

[9] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[10] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1989.

[11] A.D. Gordon. Null models in cluster validation. In W. Gaul and D. Pfeifer, editors, *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis, and Knowledge Organization*, pages 32–44. Springer Verlag, 1996.

[12] A.D. Gordon. *Classification*. Monographs on Statistics and Applied Probability 82, Chapman & Hall, 1999.

[13] W.E.L. Grimson and D.P. Huttenlocher. On the sensitivity of the Hough transform for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3):255–274, 1990.

[14] W.E.L. Grimson and D.P. Huttenlocher. On the verification of hypothesized matches in model-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(12):1201–1213, 1991.

[15] D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *International Conference of Computer Vision*, pages 267–291, London, UK, 1987.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[17] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Advanced Reference Series. Prentice-Hall, 1988.

[18] A.K. Jain, R.P.W Duin, and M. Jiachang. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–36, 2000.

[19] K. Joag-Dev and F. Proschan. Negative association of random variables, with applications. *Annals of Statistics*, 11(1):286–295, 1983.

[20] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, 1990.

[21] Y. Lamdan and H.J. Wolfson. Geometric hashing: a general and efficient model-based recognition scheme. In *Proceedings of IEEE International Conference on Computer Vision*, pages 238–249, Tampa, Florida, USA, 1988.

[22] J.L. Lisani. *Shape Based Automatic Images Comparison*. PhD thesis, Université Paris 9 Dauphine, France, 2001.

[23] J.L. Lisani, L. Moisan, P. Monasse, and J.-M. Morel. On the theory of planar shape. *SIAM Multiscale Modeling and Simulation*, 1(1):1–24, 2003.

[24] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[25] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publisher, 1985.

[26] D. Marr. *Vision*. Freeman Publishers, 1982.

[27] G. Medioni, M. Lee, and C. Tang. *A Computational Framework for Segmentation and Grouping*. Elsevier, 2000.

[28] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[29] P. Monasse and F. Guichard. Fast computation of a contrast invariant image representation. *IEEE Transactions on Image Processing*, 9(5):860–872, 2000.

[30] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.M. Morel. An a contrario decision method for shape element recognition. *submitted to IJCV*, 2004.

[31] X. Pennec. Toward a generic framework for recognition based on uncertain geometric features. *Videre: Journal of Computer Vision Research*, 1(2):58–87, 1998.

[32] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[33] G. Stockman, S. Kopstein, and S. Benett. Matching images to models for registration and object detection via clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(3):229–241, 1982.

[34] P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[35] T. Veit, F. Cao, and P. Bouthemy. A grouping algorithm for early motion detection and the analysis of visual motion. Technical report, 2005.

[36] M. Wertheimer. Untersuchungen zur Lehre der Gestalt, II. *Psychologische Forschung*, 4:301–350, 1923. Translation published as Laws of Organization in Perceptual Forms, in Ellis, W. (1938). A source book of Gestalt psychology (pp. 71-88). Routledge & Kegan Paul.

[37] H.J. Wolfson and I. Rigoutsos. Geometric hashing: an overview. *IEEE Computational Science & Engineering*, 4(4):10–21, 1997.

[38] A.L. Yuille and N.M. Grzywacz. A theoretical framework for visual motion. In T. Watanabe, editor, *High-Level Motion Processing*. MIT Press, 1998.