J Math Imaging Vis 27: 91–119, 2007

DOI: 10.1007/s10851-006-9176-0

A Unified Framework for Detecting Groups and Application to Shape Recognition

FRÉDÉRIC CAO
IRISA/INRIA
fcao@irisa.fr

JULIE DELON
LTCI, Télécom Paris (CNRS UMR 5141), 46, rue Barrault, F-75634 Paris cedex 13, France
julie.delon@enst.fr

AGNÈS DESOLNEUX
MAP5/CNRS
Agnes.Desolneux@math-info.univ-paris5.fr

PABLO MUSÉ
CMLA, ENS-Cachan
muse@cmla.ens-cachan.fr

FRÉDÉRIC SUR
Loria, Bâtiment C - projet Magrit, Campus Scientifique - BP 239, 54506 Vandoeuvre-lès-Nancy cedex, France
sur@loria.fr

Published online: 21 September 2006

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 unite. 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. Each pair of matching shape elements leads to a unique transformation (similarity or affine map.) The present theory is used to group these shape elements into shapes by detecting clusters in the transformation space.

Keywords: clustering, a contrario detection, perceptual grouping, shape recognition

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?

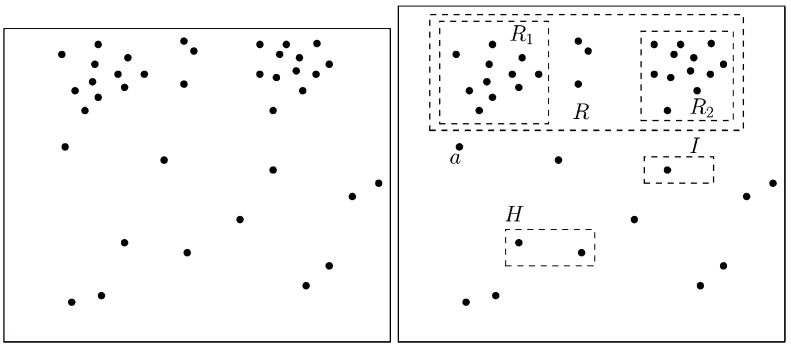

Figure 1. This figure illustrates three aspects of the grouping problem. The figure presents a set of data points in the plane and some test regions where an exceptional density may be observed, or not. Intuitively, the regions H and I do not contain clusters. So a first question is to rule out such 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).

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.

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. The representation that will be used in this paper has been introduced in [30] (see Section 4.2), 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. The first recognition step is to match similar 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. Section 1.2 gives a short overview of the related problems in clustering analysis and grouping in shape recognition. Section 2 is the theoretical core of the paper and proposes an answer to the three questions of validity, stopping rule and merging. In Section 3, this theory is applied to perceptual grouping, illustrated by simple experiments. In Section 4, the application is to group points that are geometric transformations, corresponding to matches between parts of images. Section 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, 16–18, 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 examples 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. An advantage of these voting procedures is that they are systematic, and can in principle be generalized to 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. They 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 observations 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 π (which will be also called *background law*.) By definition, for any $R \subset E$, $\pi(R)$ 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 π 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 π.

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.*

Let us remark that the clusters that are sought in this paper do not have any particular shape: they are merely a high concentration of points. Techniques aiming at finding a submanifold of any codimension containing the data points (up to error measurements) is usually called dimensionality reduction, and is not the aim of this work. As a consequence, the set of regions will be very simple: hyperrectangles with sides parallel to the axes of coordinates. This choice will prove rich enough to yield a very robust detection, but still allowing quite easy computations. Moreover, the background process will also often assume that all the coordinates are independent. In this case, the probability of a rectangle is a product of one dimensional probabilities. The complexity of computation of the probability of a rectangle then linearly increases with the dimension space D. Moreover, while it is impossible to accurately learn empirical probabilities as soon as the dimension is more than, say, 3, it is easy to compute one dimensional histograms.

However, the theory is quite independent of the choice of the regions. For the time being, let us simply assume that $\mathcal{R}$ is a set of parts of E, with finite cardinality $\#\mathcal{R}$ and such that $0 \in R$ for all $R \in \mathcal{R}$.

Another requirement is the knowledge of an agglomeration algorithm. This is defined as a function

$$\begin{gathered}\mathcal{A} : E^M \to (\mathcal{P}(E))^P \\ (x_1, \dots, x_M) \to \mathcal{A}(x_1, \dots, x_M) = (G_1, \dots, G_P)\end{gathered} \tag{2.1}$$

which to any M-tuple of data points associates a P-tuple of sets, $G_1, \dots, G_P$, such that each G_k is a part of $\{x_1, \dots, x_M\}$. The algorithm $\mathcal{A}$ is designed from any clustering algorithm and proposes a set of groups candidates from a set of data points. The number of group candidates P only depends on the number of data points M and not on the particular values of $x_1, \dots, x_M$. Actually, some of the groups can even be empty, so that P is only an a priori upper bound of the number of group candidates. In this paper, $\mathcal{A}$ is chosen as a standard single linkage hierarchical clustering method, but all the theory below does not depend on this choice. Remark that knowing $\mathcal{A}$ does not solve the problem of cluster validity. It only aims at selecting a few group candidates among the 2^M subsets of $\{x_1, \dots, x_M\}$ which is of course an untractable number in any realistic application. So the question is: among $G_1, \dots, G_P$, are there any valid groups, and how to define a quantitative measure of validity?

2.1.2. Meaningful Groups. 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].

Definition 2.2. Let $G \subset \{x_1, \dots x_M\}$ be a subset of k points out of the M data points. We call number of false alarms (NFA) of G,

$$\begin{aligned}&NFA_g(G) \\ &\equiv \#\mathcal{R} \cdot M \cdot P \cdot \min_{\substack{x_j \in G,\ R \in \mathcal{R} \\ G \subset x_j + R}} \mathcal{B}(M-1, k-1, \pi(x_j + R)).\end{aligned} \tag{2.2}$$

We say that G is an ε-meaningful group if $NFA_g(G) < \varepsilon$.

Let us see how this quantity is computed. Among all the regions of the type $x_j + R$ containing G, centered at $x_j \in G$, with $R \in \mathcal{R}$, the one with the smallest probability is selected. Then, $NFA_g(G)$ is, up to a multiplicative constant, the tail of the binomial law with parameters $M-1, k-1$ and $\pi(x_j + R)$. Let us remark that, since each group contains at most M points, and there are at most P group candidates, the total number of possible rectangles is $\#\mathcal{R} \cdot M \cdot P$ which is exactly the multiplicative constant in (2.2). This quantity is deterministic. It has of course a probabilistic interpretation which is as follows.

Proposition 2.1. *Let* $X_1, \dots X_M$ *be a background process, and* $(\Gamma_1, \dots, \Gamma_P) = \mathcal{A}(X_1, \dots, X_M)$ *the associated group candidates. Then, the expected number of* ε*-meaningful groups is less than* ε.

The proof is given in appendix.

Remark. The key point is that the *expectation* of the number of meaningful groups is easily controlled. The probability distribution of this number would instead be extremely difficult to compute, since groups may interact.

Let us interpret Definition 2.2 with Proposition 2.1. If the data points are random points following the background process, the NFA of a (random) group Γ is a random number proportional to the probability that Γ is contained in a region of $\mathcal{R}$ centered at a point of Γ. If there is such a small region, or if the cardinality of Γ is large, this probability is small. In other terms, the NFA measures how likely it is to observe a random group in a region, *by chance*. In the background model, data points are assumed independent, that is to say the data has no particular structure. Under this assumption, any candidate group has no other explanation than chance, and any detection has to be considered as a false alarm, hence the denomination. Definition 2.2 and Proposition 2.1 ensure that there are at most ε detections in the background model.

Another interpretation can be made in terms of classical hypothesis testing, in the case of multiple tests. The most conservative threshold is given by Bonferroni's method: if at most N tests are to be performed, requiring a p-value less than $\frac{\varepsilon}{N}$ for each test, implies that there are less than ε positive answers among all the tests. The definition of the NFA only consists in finding a suitable set of tests.

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

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

Proposition 2.1 ([4]). *Let R be a region in $\mathcal{R}$ and let*

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

Then

$$\alpha(M,\varepsilon)\sqrt{2\pi(R)(1-\pi(R))} \leqslant k^*(\varepsilon) - \pi(R)(M-1) \leqslant \frac{\alpha(M,\varepsilon)}{\sqrt{2}}, \qquad (2.3)$$

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

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

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 π is taken uniform in $(0,1)^2$. Both visible clusters are found and their 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.

2.2. *Optimal Merging Criterion*

While each meaningful group is relevant by itself, the whole set of meaningful groups exhibits, in general, a high redundancy. Indeed, a very meaningful group G usually remains meaningful when it is slightly enlarged or shrunk into a group G'. (See Fig. 1.)

If, e.g. $G \subset G'$, this question is easily answered by comparing $NFA_g(G)$ and $NFA_g(G')$. The group with the smallest number of false alarms must of course be preferred. Another more subtle question arises when three or more groups interact. Let G_1 and G_2 be two groups and G another group containing $G_1 \cup G_2$. We then face two conflicting interpretations of the data: two clusters or just one? The merged group G is not necessarily a better data representation than the two separate clusters G_1 and G_2. A first possibility is that G is less meaningful than each one of the merging ones. In such a case, G_1 and G_2 should be kept, rather than G. The situation is less obvious when G is more meaningful than both G_1 and G_2. In that case, keeping G_1 and G_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 groups*. This new value will be compared to the NFA of the merged group. Let us

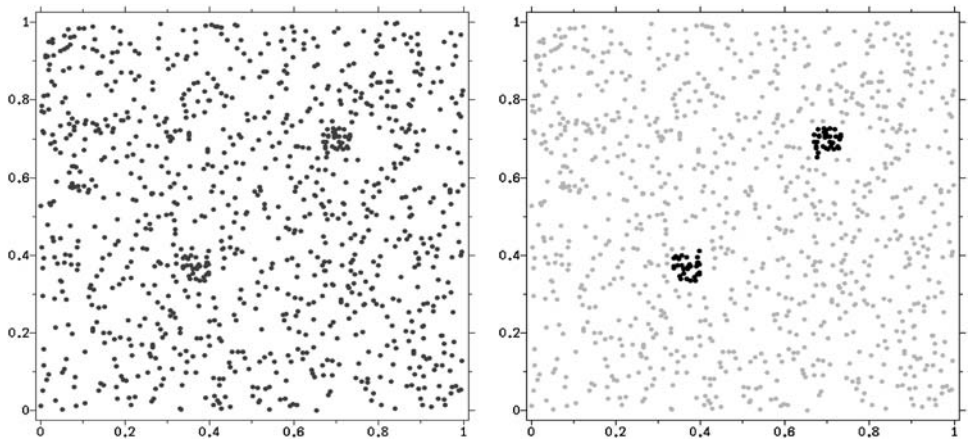

Figure 2. Clustering of twice 25 points around (0.4, 0.4) and (0.7, 0.7) surrounded by 950 i.i.d. points, uniformly distributed in the unit square. The regions of $\mathcal{R}$ are rectangles as described in Section 2.3.1. In this example $\#\mathcal{R} = 2500$ (50 different sizes in each direction). Exactly two maximal meaningful clusters are detected. (See Section 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}.

introduce the trinomial coefficient

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

We note

$$\mathcal{M}(M, k_1, k_2, \pi_1, \pi_2) = \sum_{i=k_1}^{M} \sum_{j=k_2}^{M-k_1} \binom{M}{i,\, j} \pi_1^i \pi_2^j (1-\pi_1-\pi_2)^{M-i-j}. \quad (2.4)$$

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 $M - k_1$, belong to R_1 and R_2 respectively. Thus, this probability measures how exceptional a pair of concentrated clusters can be in the background model.

As for meaningful groups, a NFA for pairs of groups is to be defined. It is assumed that a set of P *pairs* of group candidates are obtained by an operator $\mathcal{A}_2$

$$\begin{aligned} &\mathcal{A}_2 : E^M \to (\mathcal{P}(E) \times \mathcal{P}(E))^P \\ &(x_1, \ldots, x_M) \to \mathcal{A}_2(x_1, \ldots, x_M) \\ &= \left(\left(G_1^1, G_1^2\right), \ldots, \left(G_P^1, G_P^2\right)\right), \end{aligned} \quad (2.5)$$

where it is assumed that $G_i^k \subset \{x_1, \ldots x_M\}$, for $k = 1, 2$ and $1 \leqslant i \leqslant P$. Actually, the number of candidate pairs P does not need to equal the number of candidate groups. However, since some of the groups may be empty, this does not make any difference.

Definition 2.3. Consider two candidate groups of data points (G^1 and G^2). Let $(z_1, z_2) \in G^1 \times G^2$ be two data points, and R_1 and R_2 in $\mathcal{R}$. Let us denote by

- k_1 (resp. k_2) the cardinality of $G^1 \backslash (z_2 + R_2)$ (resp. $G^2 \backslash (z_1 + R_1)$), i.e. the number of points of G^1 (resp. G^2) that are not in $z_2 + R_2$ (resp. $z_1 + R_1$).
- $\pi_1 = \pi((z_1 + R_1) \backslash (z_2 + R_2))$ and $\pi_2 = \pi((z_2 + R_2) \backslash (z_1 + R_1))$.

Let us define the number of false alarms of the pair (G^1, G^2) by

$$\begin{aligned} &NFA_{gg}(G^1, G^2) \\ &= M^3 \cdot P \cdot (\#\mathcal{R})^2 \min_{\substack{(z_1,z_2)\in G^1\times G^2,\\ R_1,R_2\in\mathcal{R}\\ G^1\subset z_1+R_1, G^2\subset z_2+R_2}} \mathcal{M}(M-2, k_1-1, \\ &\quad k_2 - 1, \pi_1, \pi_2). \end{aligned} \quad (2.6)$$

We say that a pair of groups (G^1, G^2) is ε-meaningful if $NFA_{gg}(G^1, G^2) < \varepsilon$.

Let us sum up how to compute this quantity: choose a region centered at one point of G^1 (resp. G^2) and containing G^1 (resp. G^2). Those two regions may intersect, so remove their intersection and the points it may contain. Then, k_1 and k_2 points are left in each group, and the trinomial tail can be computed. Now, take the minimal value, by varying the regions and their center. Again, this is a deterministic quantity that only has a probabilistic interpretation once the background model has been introduced.

Proposition 2.3. *Let $(X_1, \ldots, X_M)$ be a background process. Let $((\Gamma_1^1, \Gamma_1^2), \ldots, (\Gamma_P^1, \Gamma_P^2)) = \mathcal{A}_2(X_1, \ldots, X_M)$, the P candidate pairs. Then, the number of ε-meaningful pairs of regions among them is less than ε.*

See Appendix 6 for the proof.

The NFA of a pair (Γ_1, Γ_2) measures how (un)likely it is to observe a large concentration of points in both of its elements. Removing the intersection is a mere technicality so that the probability of this event is the tail of the trinomial law. An alternate definition that keeps the points in the intersection has also been considered, leading to equivalent experimental results. (It turns out that the most meaningful groups usually belong to disjoint rectangles.) What really matters is that the expected number of meaningful groups in the background model (false alarms) is under control.

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

Definition 2.4 (Merging condition). Let G_1 and G_2 be two groups and G containing $G_1 \cup G_2$. We say that G is indivisible relatively to G_1 and G_2 if

$$NFA_g(G) \leqslant NFA_{gg}(G_1, G_2). \tag{2.7}$$

Equation (2.7) represents a crucial test for the coherence of a cluster region. If it is not fulfilled, G will not be considered as a valid group, as it can be divided into a more meaningful pair of groups.

2.3. Computational Issues

2.3.1. The Choice of Test Regions. What is the right set of test regions $\mathcal{R}$? All quantities previously defined can be theoretically computed, but complexity depends on $\mathcal{R}$. The choice of this paper is the following. 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 π, defined on a hyperrectangle E of $\mathbb{R}^D$, is a tensor product of one-dimensional densities $\pi_1, \ldots \pi_D$. Indeed, the probability of a rectangle is the product of independent marginal probabilities. Hence, the algorithmic complexity is a linear function of the dimensionality.

2.3.2. Agglomeration Algorithms. In this paper, the algorithms $\mathcal{A}$ and $\mathcal{A}_2$ are actually derived from a single algorithm. Indeed, a hierarchical single linkage algorithm is used. It provides a binary tree. Each level of the tree is a partition of the data set, and each node is a group candidate. Non leaf nodes are the union of their exactly two children. This tree is sometimes called *dendrogram* [17]. Hence, the total number of non leaf nodes in the tree is $P = M - 1$, which is also the number of pairs.

Many of the most common aggregation procedures 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, $\{x_1\}, \ldots \{x_M\}$. It is assumed that between two data points x_i and x_j, a dissimilarity measure $d(x_i, x_j)$ is given. (It does not need to be a distance.) 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. Pairs of sibling nodes are the candidate pairs, whose NFA is computed.

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 validity of each node is computed as in Definition 2.2.

Grouping algorithm
For each node G (candidate group) with cardinality k in the clustering tree or dendrogram,

1. find the smallest region $x + R$, with $x \in G$ and $R \in \mathcal{R}$ centered at x, and containing the other data points of the node.

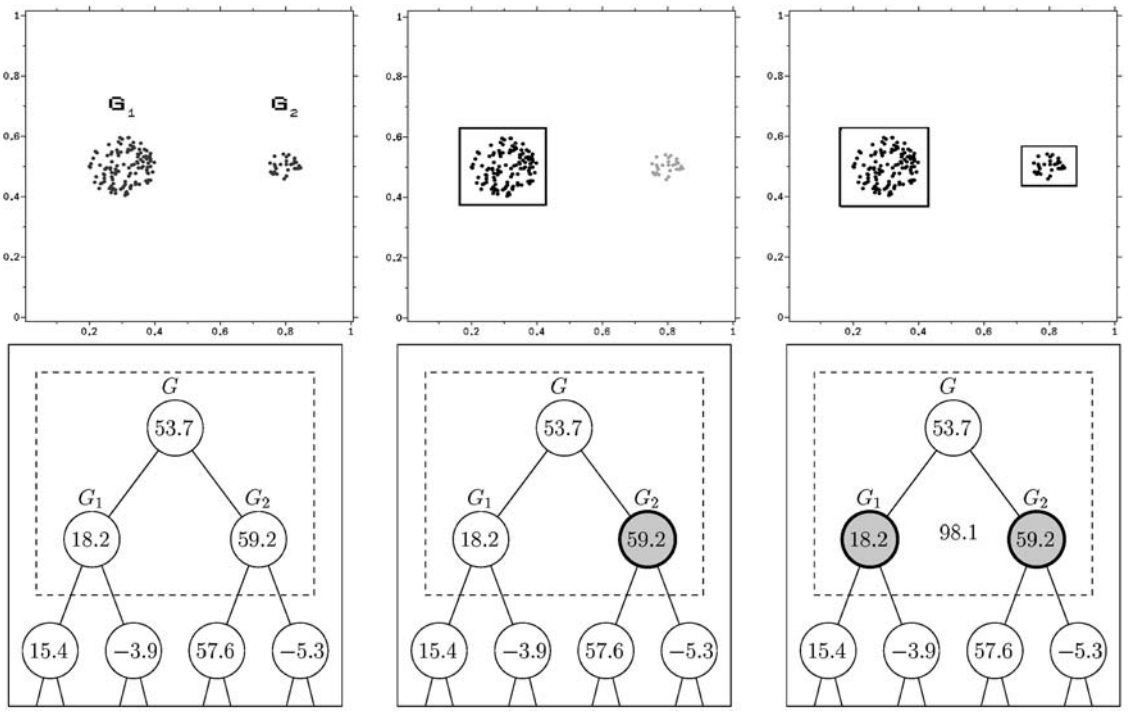

Figure 3. Indivisibility prevents collateral elimination. Each subfigure shows a configuration of points, and a piece of the corresponding dendrogram, with the selection of maximal meaningful groups, depicted in grey. The numbers in each node corresponds to $-\log_{10}(NFA_g)$ of its associated cluster, 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 left-most group G_1 is eliminated. It is, however, very meaningful since $NFA_g(G_1) = 10^{-18}$. Right: by combining indivisibility and maximality criteria, both clusters G_1 and G_2 are selected.

2. Compute the NFA of G as the minimum of $M(M-1)\cdot\#\mathcal{R}\cdot\mathcal{B}(M-1, k-1, \pi(x+R))$ when x describes all G and R is any element of $\mathcal{R}$.

2.3.3. 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 ε-meaningful group). A node G is maximal ε-meaningful if and only if

1. $NFA_g(G) \leqslant \varepsilon$,
2. G is indivisible with respect to any pair of sibling descendents,
3. for all indivisible descendent G', $NFA_g(G') \geqslant NFA_g(G)$,
4. for all indivisible ascendent G', either $NFA_g(G') > NFA_g(G)$ or there exists an indivisible descendent G'' of G' such that $NFA_g(G'') < NFA_g(G')$.

Condition 4 implies that G can be abandoned for a larger group only if this group has not been beaten by one of its descendents. Imposing conditions 3 and 4 ensures that two different maximal meaningful groups are disjoint.

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,

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

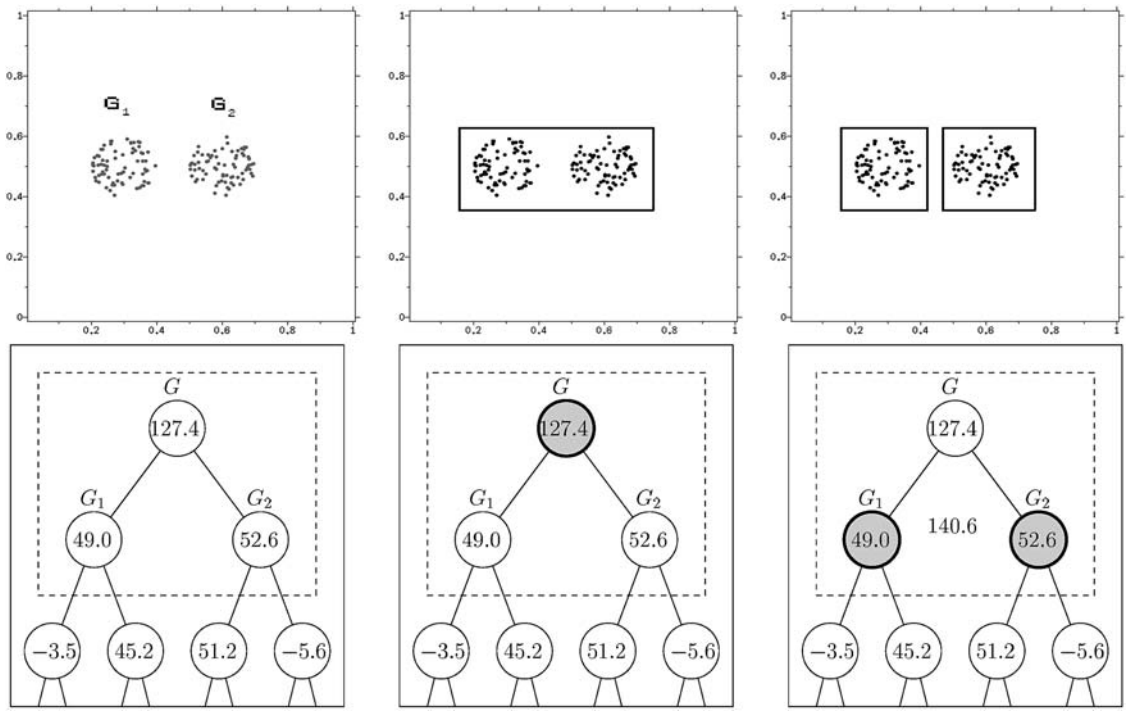

Figure 4. Indivisibility prevents faulty union. Each sub-figure shows a configuration of points, and a piece of the corresponding dendrogram, with the selection of maximal meaningful groups, depicted in grey. The number in each node corresponds to the NFA_g of its associated cluster. The number between 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.

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 Definition 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.

Figure 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 $NFA_{gg}(G_1, G_2) < NFA_g(G)$.

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,

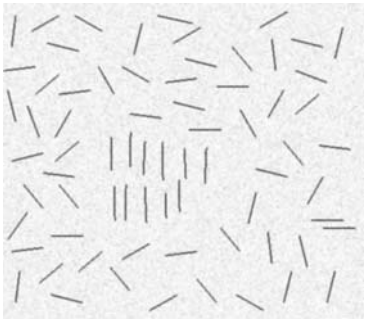

Figure 5. An image of a scanned drawing of 71 segments.

called *meaningful level lines* (see [5] for a full description of this extraction process). Once these objects are detected, say $O_1, \ldots O_M$, 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.) In all the experiments, the number of rectangle sizes in each direction is 50. Thus $\#\mathcal{R} = 50^D$. Let us also give a few words on the dissimilarity measures in this section. Up to an affine change of variables, all observations are assumed to belong to the interval (0, 1) (possibly with periodic boundaries). The Euclidean metric is then used as a dissimilarity measure.

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. These points are grouped with respect to their x and y coordinates in the square, so that $D = 2$. In the background model, x and y are assumed independent. Two groups and two groups only are detected with very good NFA_g (less than 10^{-7}).

3.2. Segments

In the second example, groups are perceived as a result of the collaboration between two different features. Figure 5 shows 71 straight segments with different orientations, almost uniformly distributed in position. The position of the barycenter and the orientation of the principal axis of the strokes are computed. As expected, no meaningful cluster is detected in the space of 2D position coordinates of the barycenters.

If orientation is chosen as the only feature ($D = 1$), 8 maximal meaningful groups are detected, corresponding to the most represented orientations, see Fig. 6. None of these clusters exhibits a very low NFA_g. Only one of those groups 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 orientations 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.

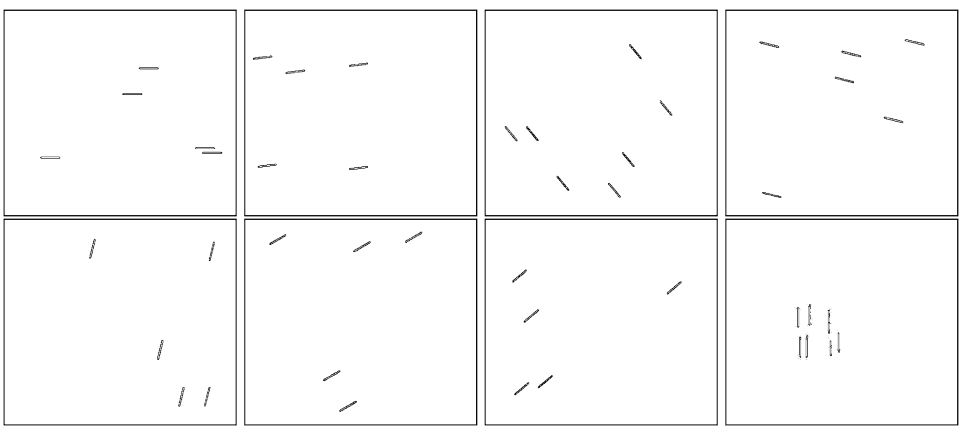

Figure 6. Grouping with respect to orientation: there are 8 maximal meaningful groups. NFA_g range is between 10^{-1} and 10^{-5}. The central group does not contain all the vertical segments, because their orientation are actually slightly different. Hence, the maximal group containing these vertical 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.

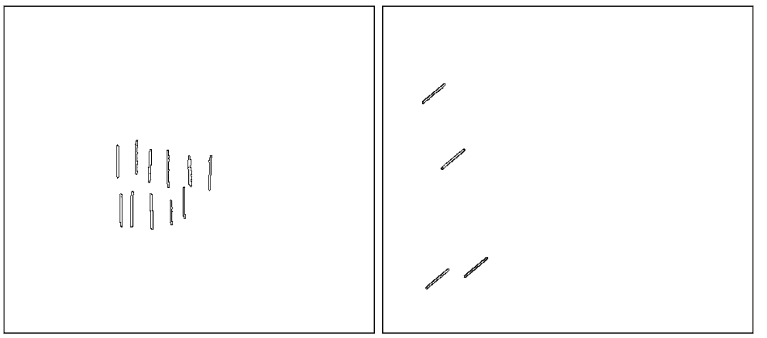

Figure 7. Grouping in the space (x-coordinate, orientation). There are two maximal meaningful groups. This time, the whole central group is detected ($NFA_g = 10^{-1.5}$), but there is still another group (which is a part of the 7th group in the orientation grouping (see Fig. 6)). However, its $NFA_g = 0.3$, which means that it is hardly meaningful. This group is not perceived because it is masked by all the other segments. If grouping is done with respect to full 2D-position and orientation, only the central group is detected with $NFA_g = 10^{-3.4}$.

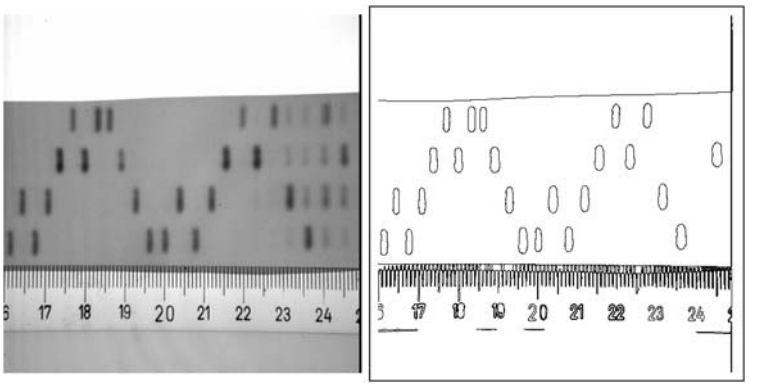

Figure 8. An image of DNA and its 80 maximal meaningful level lines [5].

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. 7). 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 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 $NFA_g = 10^{-3.4}$.

3.3. DNA Image

The 80 objects in Fig. 8 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 digits, 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.

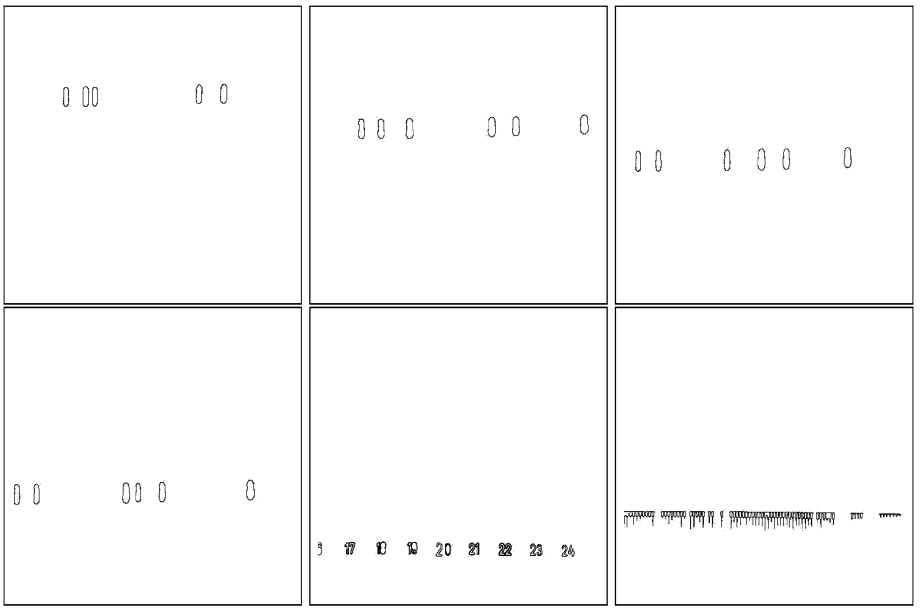

Figure 9. Grouping with respect to diameter and y coordinate. Six groups are detected, 4 of which are rows of DNA marks. The last two ones correspond to the ruler. $-\log_{10}(NFA_g)$ range from 2.6 to 7.6 for the DNA. The last two groups are larger and are obviously more meaningful: $-\log_{10}(NFA_g) = 43$ and 54.

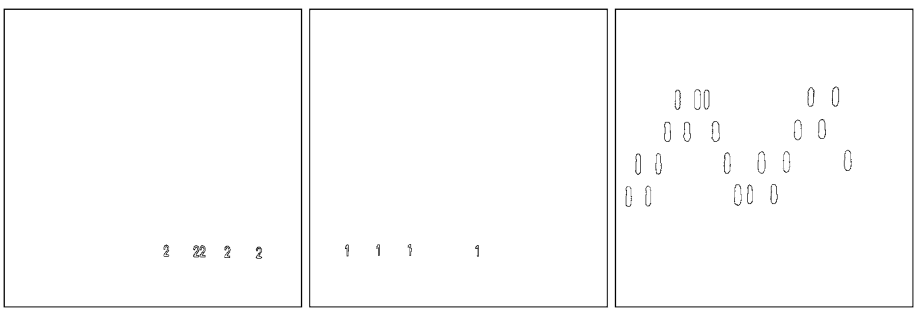

Figure 10. Grouping with respect to orientation, elongation, diameter, and a convexity coefficient. The DNA marks are the most meaningful group $NFA_g = 10^{-10}$, but the 1 and 2's also form groups, with NFA_g close to 1.

In the space (diameter, y-coordinate), 6 maximal meaningful groups are detected (Fig. 9). 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 digits 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, $-\log_{10}(NFA_g) = 54$.

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. 10), 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 π, 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.

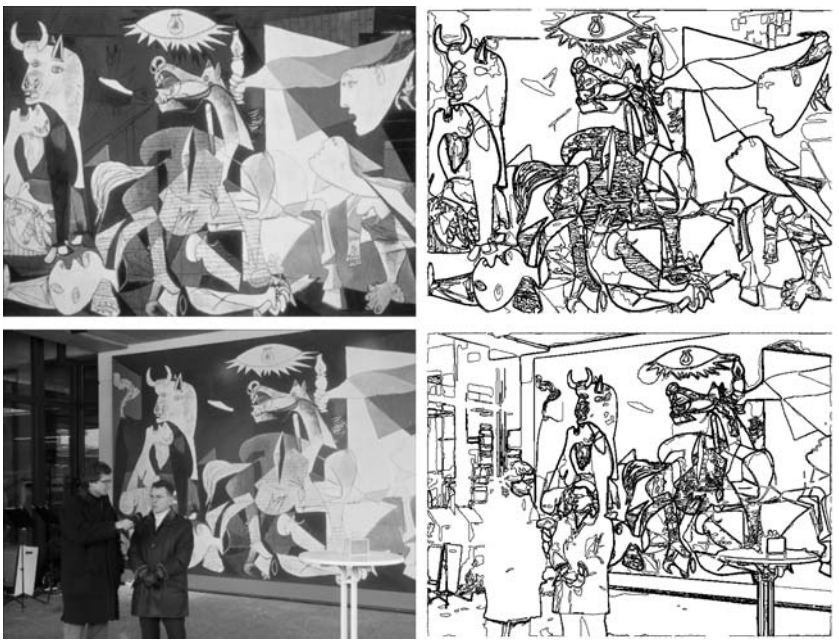

Figure 11. "Guernica" experiment. Original images and maximal meaningful level lines [5]. All these level lines are encoded into normalized affine invariant shape elements [30], based on robust directions as bitangent and flat parts. Top: target image, bottom: scene image.

4. Grouping Spatially Coherent Matches for Planar Shape Recognition

4.1. Why Spatial Coherence Detection?

Looking at Fig. 11, 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 pairings 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

For the 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 are 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.

Figure 12 displays the shape elements common to the two images of Fig.11. 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.

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 Section 2, first by defining a dissimilarity measure between transformations, then by defining a suitable background model on the sets of transformations.

Let us remark that any representation that allows to match local features with a given group of invariance can be used as the input of the grouping procedure. Instead of matching level lines, we also tested Lowe's SIFT descriptors [24], which are similarity invariant. The results are equivalent for images differing by a similarity transformation. For general rigid deformation, the local matching should be at least affine invariant.

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. 13.) 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

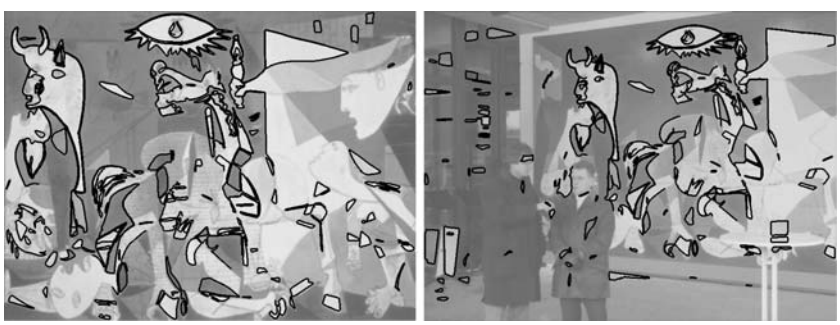

Figure 12. "Guernica" experiment: affine invariant meaningful matches [30]. Since all parallelograms differ from an affine transformation (*idem* for triangles or ellipses), there are many casual matches.

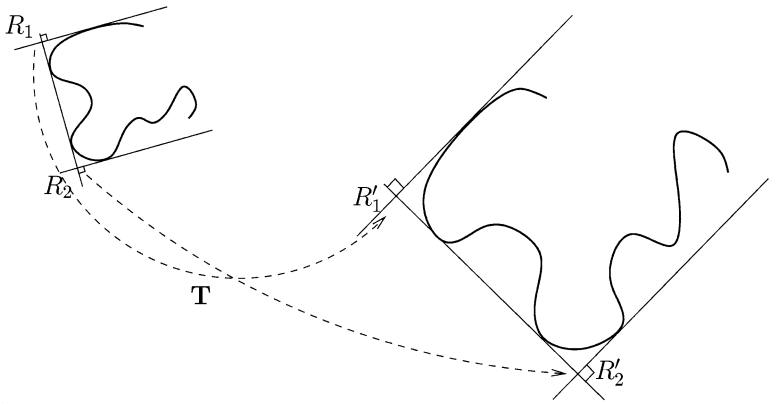

Figure 13. Two pieces of level lines and their corresponding local similarity frames. The similarity **T** maps R_1 into R'_1 and R_2 into R'_2. Equivalently the local frame, (R_1, R_2) may be represented by $(p, v) = (\frac{R_1+R_2}{2}, R_2 - R_1)$.

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

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

with $(a, b) \in \mathbb{C}^2$. The transformation **T** is uniquely determined by the 4-tuple

$$T = (Re(b), Im(b), \arg a, |a|),$$

and T and **T** will be identified.

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 **T**. There exists a unique 2×2 matrix **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 **M** boils down to the solution of a 2×2 linear system. By the QR decomposition [10], **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 $[0, 2\pi) \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$. 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 **T**.

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

Figure 14 shows three 2-D projections of the transformation points T_k corresponding to the "Guernica" affine invariant meaningful matches of Fig. 12.

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 Section 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

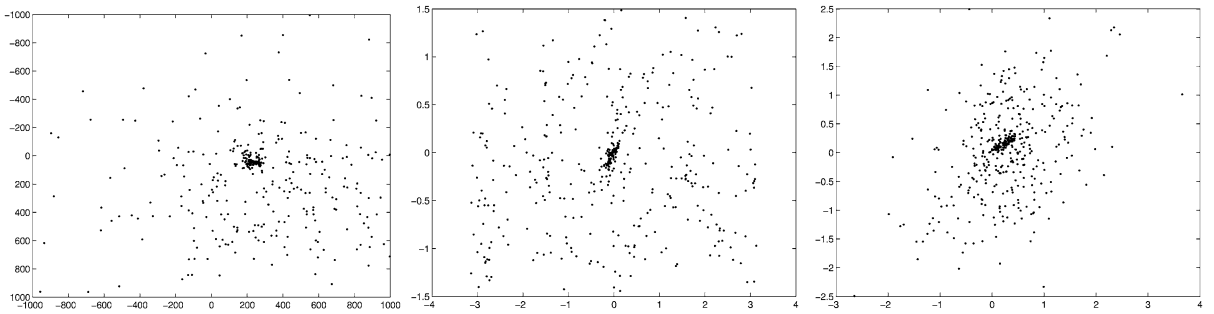

Figure 14. "Guernica experiment: Each point represents a transformation associated with an affine invariant meaningful match, described by 6 parameters. Each figure represents a two-dimensional projection of the points, respectively t_x vs. t_y (translation coordinates), θ (rotation) vs. φ (shear), and $\ln(s_x)$ 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.

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,

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

Let us remark that this dissimilarity not only depends on the transformation themselves, but also on the location and size of the shape elements in the image.

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 shape 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

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

4.4.2. Background Model: The Similarity Case. In order to apply the detection framework of Section 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 π 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 π 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 π 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,

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

where π^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 π^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 π follows from (4.5).

In practice, the computation of π 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. 16 and experiments.) The distribution of the zoom factor $|A|$ is instead far from being uniform and even showing a constant shape in the different experiments we have made. 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, θ, φ, 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 solid 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 to 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 512×512 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 is applied.
5. A background model π on the set of similarities or on the set of affine transforms E is built according to Section 4.4.2.
6. The transforms $T_1, \dots T_M$ associated with the matching pairs form a point data set in E. From this set, a clustering tree is built according to the dissimilarity measures of Definitions 4.1 or 4.2.
7. Maximal meaningful groups are computed by Definition 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 π 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.

Some experiments are displayed both in the similarity invariant case, and affine invariant case. In theory, affine invariance is always better. First because it is a better approximation of projective transformation. Second because the probability of regions are usually smaller, because they are the product of 6 marginal probabilities instead of 4 in the similarity invariant case. In practice, affine invariant encoding is more demanding, and there are usually less affine invariant shape elements in an image than similarity invariant shape elements. Therefore groups usually contain less points in the affine invariant setting, which counterbalances the smaller probabilities of the regions.

5.1. A Single Group

Figure 15 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 16 shows the learned distribution of the zoom factors in the x and y directions as well as the shear and rotation angle. The latter is not perfectly uniform in this case, because the vertical and horizontal directions are privileged in these geometrical images. Figure 17 shows the meaningful cluster.

Figure 15. "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.

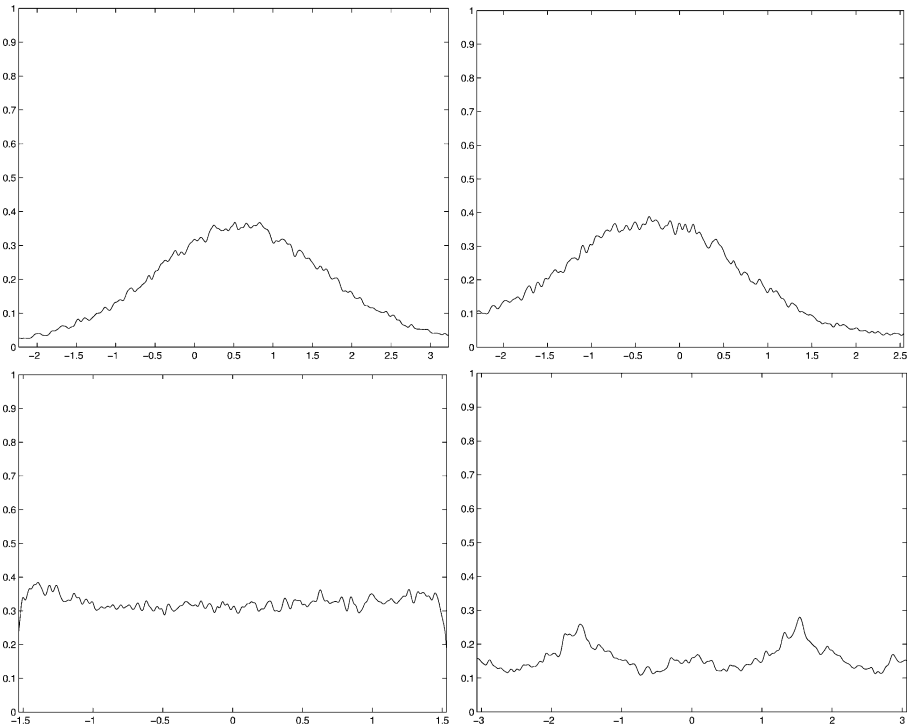

Figure 16. Empirical histograms for affine invariant matching for the experiment of Fig. 11. On the first row, the empirical zoom factors in the x and y direction (logscale), which are image dependent. On the second row, the distribution of the shear and the rotation angle. The shear is basically uniform, but the rotation exhibits some peaks around $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ because of the numerous horizontal and vertical lines in the image.

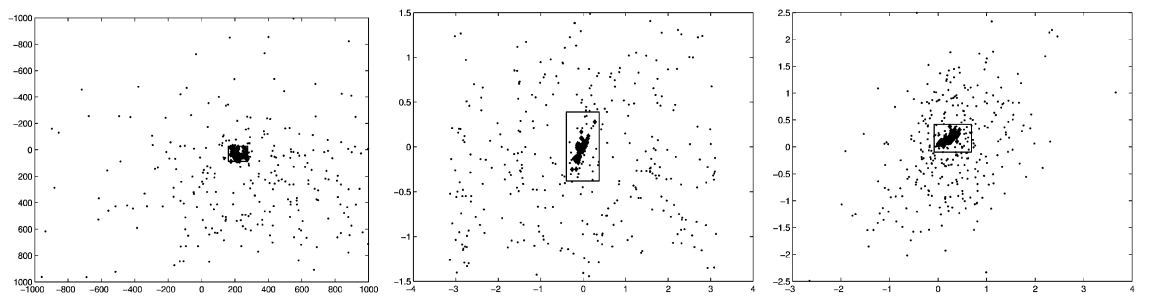

Figure 17. "Guernica experiment: data points of Fig. 14, where the points corresponding to the only affine invariant group are represented with larger dots. The boundaries of the corresponding hyperrectangle are drawn.

5.2. Two Different Groups

The similarity invariant procedure is applied in the same way to the images of Fig. 18. Two maximal meaningful groups are detected: the faces and the title. The corresponding points in the similarity space are displayed on Fig. 19. 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. 20 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 $10^{-6.6}$. The largest group is not indivisible, and thus cannot be maximal.

The examination of the transformation histograms (Fig. 21) shows that the rotation angle is nearly uniformly distributed. The zooming factor, on the other hand, does not have an intuitive distribution. The

(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 18. "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.

translation has to be learned conditionnally to the rotation and the zoom. The last two plots are the two-dimensional distribution 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.

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. 22). Figure 23 shows the affine invariant meaningful matches. Five groups are detected. The corresponding shape elements are displayed for each group in Figs. 24 and 25. The NFA_g of maximal meaningful groups are reported in Table 1. The three first groups are very meaningful, while the two other NFA_g are much closer to 1: about 10^{-4}.

Maximal meaningful groups can be used for registration. Since a group contains several points (i.e. several affine transforms), a standard least squares procedure

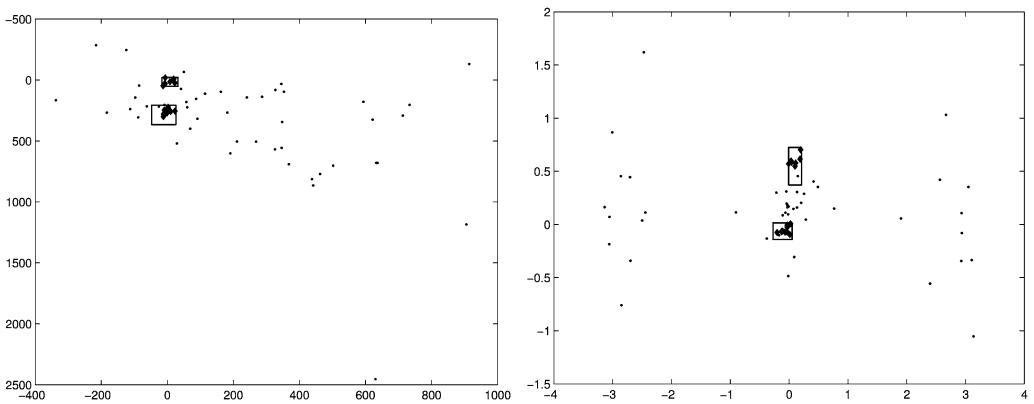

Figure 19. Casablanca experiment. Meaningful clusters in the similarity space. Left: projection in the translation dimensions. Right: projection on the rotation and zoom (log scale) axes. In this case, two clusters are clearly visible. Their position but also their scale is different.

Figure 20. "Casablanca" experiment. Meaningful group corresponding to the merging of groups in Fig. 18. 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.

Table 1. "Coca-Cola" experiment: NFA_g for the maximal meaningful groups in Figs. 24 and 25.

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

allows to compute the best planar projective transform describing the group. As can be seen on the left parts of Figs. 26 and 27, 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 two pieces C_1 and C_2, belonging to the first and second image satisfy $|C_1(s) - C_2(s)| < \delta$ for all $s \in (0, l)$ 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 Figs. 26 and 27.

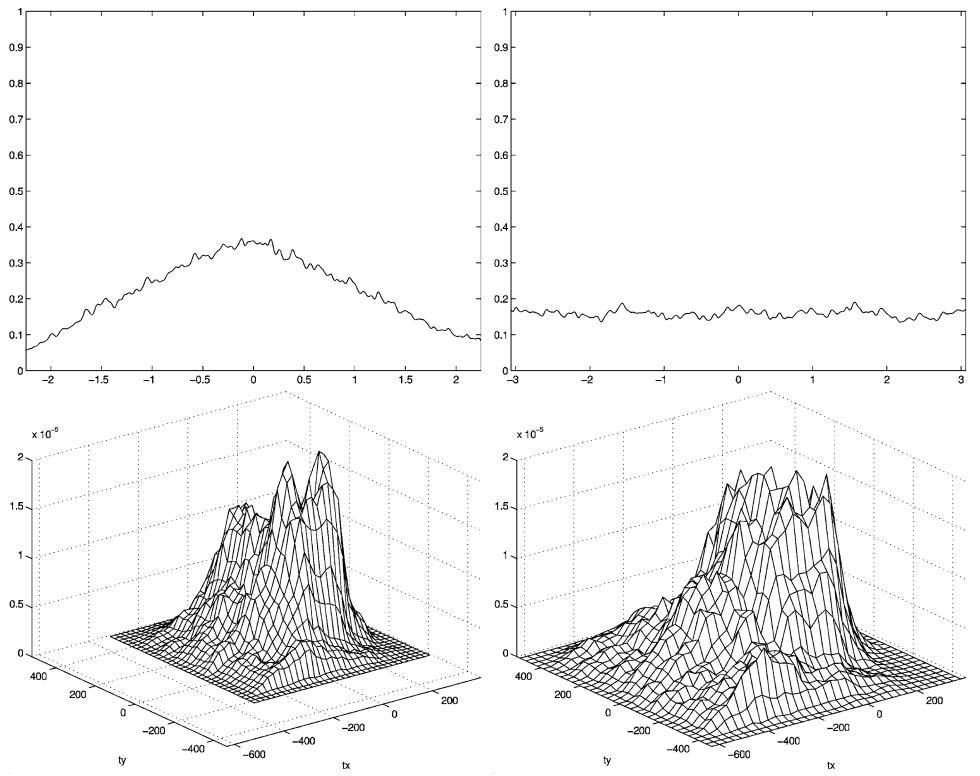

Figure 21. Empirical histograms for similarity invariant matching for the experiment of Fig. 18. On the first row, the log-empirical zoom factor $\ln(|a|)$ and the rotation angle $\arg a$. This last one is nearly uniform in this case. On the bottom row, the distribution of the translation vector, conditioned by two different values of the couple $(\ln(|a|), \arg a)$. These values correspond to the two maximal groups that are depicted on Fig. 18. Since the scales are different, so are the distributions.

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 defining 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 clusters with "no shape". 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.

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].

Perceptual grouping laws works recursively, as observed by Gestaltists. Applying the same kind of iteration, with the presented algorithm would lead to a

Figure 22. "Coca-Cola" experiment: original images and maximal meaningful level lines. Top: images, bottom: meaningful level lines [5].

Figure 23. "Coca-Cola" experiment. Meaningful matches. Number of tests: $1.57\ 10^7$ (591 shape elements in the target image, $26,621$ in the scene image). Since the matching algorithm works 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 Hough Transform clustering, the method proposed in this paper does not depend on bins quantization and has automatic detection thresholds.

Figure 24. "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.

further automatic analysis of images. However, some questions need to be examined. First, after an iteration, meaningful groups with respect to some properties are detected. Each one of these groups have a NFA. Is it possible to integrate this NFA as a weight in further grouping iterations? Another major problem is the masking phenomenon: grouping laws often compete with one another. As a consequence, a structure that is conspicuous out of any context may not be visible when surrounded by other percepts. The simplest case of masking is corruption by noise, which is more or less handled by the method, because the background model precisely measures how much the observation differs from a noise model. A much more complicated case is masking by another organized structure. In this case, how can NFA be used to decide which structure has to be kept? The influence of the choice of the dissimilarity should also be studied further. Other questions are about the limitations of the method, due to high dimensionality. How hard is it to design a background model in high dimension? How sensitive are the results with respect to this model? This needs to be investigated further.

Appendix. A. Proofs

A.1. Proof of Proposition 2.1

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

- $X = (X_1, \ldots X_M)$, the background process,
- $x = (x_1, \ldots x_m)$ a set of M points in E,

Figure 25. "Coca-Cola" experiment: maximal meaningful groups (last two among five). Their $-\log_{10}(NFA)$ are respectively 4.0 and 3.0.

- $X^j = (X_1, \dots X_M)$ with X_j omitted in the list,
- $x^j = (x_1, \dots x_M)$ with x_j omitted in the list,
- $d\pi^j(x^j) = d\pi(x_1) \dots d\pi(x_M)$ with $d\pi(x_j)$ omitted in the product,
- $\Pr^j$ the 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. *Let us fix $x_j \in E$. Consider a random process $X_1, \dots, X_M$. Then*

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

Proof: The cumulative distribution of the random variable $K(X^j, x_j, R)$ is $k \to \mathcal{B}(M-1, k, \pi(x_j+R))$. If X has a cumulative distribution F, it is a classical result that the probability that $F(X) < t$ is less or equal than t. The results follows. □

Proof of Proposition 2.1: Let us note

- For $1 \leqslant i \leqslant P$, the Bernoulli variable

$$Y_i = \begin{cases} 1 & \text{if } \Gamma_i \text{ is } \varepsilon\text{-meaningful,} \\ 0 & \text{otherwise.} \end{cases}$$

- $S = \sum_i Y_i$ the number of ε-meaningful groups.

Let us also denote by K_i the (random) cardinality of Γ_i and $\epsilon = \frac{\varepsilon}{MP\#\mathcal{R}}$.

$$\Pr(Y_i = 1)$$

$$= \Pr\left(\min_{\substack{X_j \in \Gamma_i,\, R\in\mathcal{R} \\ \Gamma_i \subset X_j + R}} \mathcal{B}(M-1, K_i - 1, \pi(X_j+R)) < \epsilon\right). \tag{A.1}$$

$$= \Pr(\exists j, R \text{ s.t. } X_j \in \Gamma_i, \Gamma_i \subset X_j + R, \mathcal{B}(M-1, K_i-1, \pi(X_j+R)) < \epsilon) \tag{A.2}$$

$$\leqslant \Pr(\exists j, R \text{ s.t. } \mathcal{B}(M-1, K(X^j, X_j, R), \pi(X_j+R)) < \epsilon) \tag{A.3}$$

$$\leqslant \sum_{\substack{1\leqslant j\leqslant M \\ R\in\mathcal{R}}} \Pr(\mathcal{B}(M-1, K(X^j, X_j, R), \pi(X_j+R)) < \epsilon). \tag{A.4}$$

Figure 26. "Coca-Cola" experiment: registration with respect to the meaningful groups. Because there are several affine matches per group, one can compute the best planar 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).

The first inequality results from $\Gamma_i \subset X_j + R \Rightarrow K_i - 1 \leqslant K(X^j, X_j, R)$ and the monotonicity of the map $k \mapsto \mathcal{B}(M-1, k, p)$. Now, Lemma A.1 cannot be directly applied. Indeed, the considered region is centered at a random point X_j and thus has a random probability. However, by Fubini Theorem

$$\begin{aligned}
&\Pr(\mathcal{B}(M-1, K(X^j, X_j, R), \pi(X_j + R)) < \epsilon),\\
&= \int d\pi(x_j) \Pr{}^j(\mathcal{B}(M-1, K(X^j, x_j, R),\\
&\quad \pi(x_j + R)) < \epsilon),\\
&\leqslant \int d\pi(x_j)\epsilon \quad \text{by Lemma A.1,}\\
&= \epsilon.
\end{aligned}$$

Thus

$$P(Y_i = 1) \leqslant M\#\mathcal{R}\epsilon = \frac{\varepsilon}{P}.$$

Figure 27. "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.

Finally,

$$\mathbb{E}(S) = \sum_{i=1}^{P} \mathbb{E}(Y_i) \leqslant \sum_{i=1}^{P} \frac{\varepsilon}{P} = \varepsilon,$$

where $\mathbb{E}$ is the expectation under the background model. □

A.2 Proof of Proposition 2.3

Let $1 \leqslant i \neq j \leqslant M$. We note

- $X = (X_1, \ldots X_M)$, the 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_i 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 marginal of Pr with respect to x^{ij},
- $K(X, i, j, R_i, R_j)$ = the number of points among X^{ij} that are in $X_i + R_i$ but not in $X_j + R_j$, i.e. belonging to $(X_i + R_i) \setminus (X_j + R_j)$,
- $K_i = K(X, i, j, R_i, R_j)$, $K_j = K(X, j, i, R_j, R_i)$,

- $\tilde{K}_i = K(X_{ij}, i, j, R_i, R_j)$, $\tilde{K}_j = K(X_{ij}, j, i, R_j, R_i)$,
- $k_i = K(x_i, i, j, R_i, R_j)$, $k_j = K(x_j, j, i, R_j, R_i)$,
- $\pi_i = \pi((x_i + R_i) \setminus (x_j + R_j))$, $\pi_j = \pi((x_j + R_j) \setminus (x_i + R_i))$,
- $\Pi_i = \pi((X_i + R_i) \setminus (X_j + R_j))$, $\Pi_j = \pi((X_j + R_j) \setminus (X_i + R_i))$,
- $\epsilon = \frac{2\varepsilon}{M^3 P(\#\mathcal{R})^2}$.

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 to the case of two variables. Remark that this proof is true for discrete variables, since it uses the fact that $\tilde{K}_j$ and $\tilde{K}_i$ can only take $M-1$ different values. Indeed,

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

Let

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

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 $k \to \mathcal{M}(M-2, k, k_j, \pi_i, \pi_j)$ being monotone,

$$\mathcal{M}(M-2, k, k_j, \pi_i, \pi_j) < \epsilon \Leftrightarrow k \geqslant k_i(\epsilon, k_j). \tag{A.5}$$

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

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

□

Proof: Let us note for $k = 1, \ldots, P$

- The Bernoulli variable

$$Y_k = \begin{cases} 1 & \text{if } (\Gamma_k^1, \Gamma_k^2) \text{ is } \varepsilon\text{-meaningful,} \\ 0 & \text{otherwise.} \end{cases}$$

- $S = \sum_{k=1}^{P} Y_k$ the number of ε-meaningful pairs of regions.

Let us fix k. Let X_i and X_j be two points in the process, belonging to Γ_k^1 and Γ_k^2. Let R_i and R_j be two regions in $\mathcal{R}$, such that $\Gamma_k^1 \subset X_i + R_i$ and $\Gamma_k^2 \subset X_j + R_j$. Let also $\hat{K}_i$ the number of points of Γ_k^1 that are not in $X_j + R_j$ and $\hat{K}_j$ the number of points of Γ_k^2 that are not in $X_i + R_i$. Remark that with the notations above, $\hat{K}_i \leqslant K_i$ and $\hat{K}_j \leqslant K_j$. Then,

$$\begin{aligned}&\Pr(Y_k = 1) \\ &= \Pr(\exists i, j, R_i, R_j \quad \text{s.t.} \quad X_i \in \Gamma_k^1,\ X_j \in \Gamma_k^2, \\ &\quad \Gamma_k^1 \subset X_i + R_i, \Gamma_k^2 \subset X_j + R_j, \\ &\quad \mathcal{M}(M-2, \hat{K}_i - 1, \hat{K}_j - 1, \Pi_i, \Pi_j) < \epsilon). \\ &\leqslant \Pr(\exists i, j, R_i, R_j \quad \text{s.t.} \\ &\quad \mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j) < \epsilon) \\ &\leqslant \sum_{i,j=1}^{M} \sum_{R_i, R_j} \Pr(\mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j) < \epsilon)\end{aligned}$$

The first inequality results from $\hat{K}_i - 1 \leqslant K_i$ and $\hat{K}_j - 1 \leqslant K_j$ and the monotonicity of the map $(k, l) \mapsto \mathcal{M}(M-2, k, l, p, q)$ with respect to each

of its variables. By Fubini theorem,

$$\begin{aligned}
&\Pr(\mathcal{M}(M-2, K_i, K_j, \Pi_i, \Pi_j) < \epsilon) \\
&\quad = \int_{E^2} d\pi(x_i) d\pi(x_j) \\
&\qquad \times \int_{E^{M-2}} \mathbb{1}_{\{\mathcal{M}(M-2,k_i,k_j,\pi_i,\pi_j)<\epsilon\}} d\pi^{ij}(x^{ij}) \\
&\quad = \int_{E^2} d\pi(x_i) d\pi(x_j) \\
&\qquad \times \Pr{}^{ij}(\mathcal{M}(M-2, \tilde{K}_i, \tilde{K}_j, \pi_i, \pi_j) < \epsilon) \\
&\quad < (M-1)\epsilon,
\end{aligned}$$

where Lemma 6.2 has been used in the last inequality. Finally,

$$\begin{aligned}
\mathbb{E}(S) &= \sum_{k=1}^{P} \mathbb{E}(Y_k) \\
&< \sum_{k=1}^{P} (M-1) P M (\#\mathcal{R})^2 \epsilon \\
&\leqslant \varepsilon.
\end{aligned}$$

□

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).

References

1. D.H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognition*, Vol. 13, No. 2, pp. 111–122, 1981.
2. H.H. Bock, "On some significance tests in cluster analysis," *Journal of Classification*, Vol. 2, pp. 77–108, 1985.
3. F. Cao, P. Musé, and F. Sur, "Extracting meaningful curves from images," *Journal of Mathematical Imaging and Vision*, Vol. 22, No. 2–3, pp. 159–181, 2005.
4. A. Desolneux, L. Moisan, and J.-M. Morel, "Meaningful alignments," *International Journal of Computer Vision*, Vol. 40, No. 1, pp. 7–23, 2000.
5. A. Desolneux, L. Moisan, and J.-M. Morel, "Edge detection by Helmholtz principle," *Journal of Mathematical Imaging and Vision*, Vol. 14, No. 3, pp. 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*, Vol. 25, No. 4, pp. 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*, Vol. 20, No. 6, pp. 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 *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis, and Knowledge Organization*, W. Gaul and D. Pfeifer (Eds.), Springer Verlag, 1996, pp. 32–44.
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*, Vol. 12, No. 3, pp. 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*, Vol. 13, No. 12, pp. 1201–1213, 1991.
15. D.P. Huttenlocher, and S. Ullman, "Object recognition using alignment," In *International Conference of Computer Vision*, London, UK, 1987, pp. 267–291.
16. A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, Vol. 31, No. 3, pp. 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*, Vol. 22, No. 1, pp. 4–36, 2000.
19. K. Joag-Dev and F. Proschan, "Negative association of random variables, with applications," *Annals of Statistics*, Vol. 11, No. 1, pp. 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*, Tampa, Florida, USA, 1988, pp. 238–249.
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*, Vol. 1, No. 1, pp. 1–24, 2003.
24. D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 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*, Vol. 50, No. 2, pp. 159–179, 1985.
29. P. Monasse, and F. Guichard, "Fast computation of a contrast invariant image representation," *IEEE Transactions on Image Processing*, Vol. 9, No. 5, pp. 860–872, 2000.

30. P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.M. Morel, "An a contrario decision method for shape elements recognition," *International Journal of Computer Vision*, Vol. 69, No. 3, pp. 295–315, 2006.
31. X. Pennec, "Toward a generic framework for recognition based on uncertain geometric features," *Videre: Journal of Computer Vision Research*, Vol. 1, No. 2, pp. 58–87, 1998.
32. C. Schmid, and R. Mohr, "Local greyvalue invariants for image retrieval," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 5, pp. 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*, Vol. 4, No. 3, pp. 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, INRIA.
36. M. Wertheimer, "Untersuchungen zur Lehre der Gestalt, II," *Psychologische Forschung*, Vol. 4, pp. 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*, Vol. 4, No. 4, pp. 10–21, 1997.
38. A.L. Yuille, and N.M. Grzywacz, "A theoretical framework for visual motion," in *High-Level Motion Processing* Watanabe, T. (Eds.), MIT Press, 1998.

Frédéric Cao graduated from the Ecole Polytechnique in 1995 and obtained a PhD in applied mathematics in Ecole Normale Supérieure de Cachan in 2000. He defended his "Habilitation à diriger des recherches" in 2004. His research interests include partial differential equations for image and shape filtering, but also statistical approaches to shape recognition and data analysis, or motion analysis.

Julie Delon was born in France in 1978. During the period 1997–2001, she has studied applied mathematics at the Ecole Normale Supérieure de Cachan. From 2001 to 2004, she prepared a Ph.D. thesis in image analysis at the CMLA (Cachan, France), and defended it in December 2004. She is currently a research scientist with CNRS at Télécom Paris.

Agnès Desolneux was born in France in 1974. She defended her PhD thesis in applied mathematics in 2000 under the direction of Jean-Michel Morel at the ENS Cachan. She is currently CNRS researcher at the MAP5, University Paris 5. She is working on statistical methods in image analysis. Web page: http://www.math-info.univ-paris5.fr/~desolneux/

Pablo Musé was born in Montevideo, Uruguay, in 1975. He received the Electrical Engineer degree from the Universidad de la República, Uruguay, in 1999, and the DEA in Mathematics, Vision and Learning from the Ecole Normale Supérieure de Cachan, France, in 2001. In 2004 he obtained his Ph.D. in Applied Mathematics, also from ENS Cachan, where he had a researcher position until May 2005. Since then he has been with Cognitech Inc., Pasadena, CA, USA.

Frédéric Sur was born in 1976. He is a former student of École Normale Supérieure de Cachan, France. In 2004, he received his PhD degree in applied mathematics and image analysis from Université Paris Dauphine. He is now an assistant professor at École des Mines de Nancy (France) and with Loria laboratory.