Lecture Notes in Mathematics

Frédéric Cao
José-Luis Lisani
Jean-Michel Morel
Pablo Musé
Frédéric Sur

A Theory of Shape Identification

1948

Springer

Frédéric Cao · José-Luis Lisani
Jean-Michel Morel · Pablo Musé
Frédéric Sur

A Theory of Shape Identification

Springer

Frédéric Cao
DxO Labs
3 rue Nationale
92100 Boulogne Billancourt, France
fcao@dxo.com

José-Luis Lisani
Dep. Matemàtiques i Informàtica
University Balearic Islands
ctra. Valldemossa km.7,5
07122 Palma de Mallorca, Balears
Spain
joseluis.lisani@uib.es

Jean-Michel Morel
CMLA, Ecole Normale
Supérieure de Cachan
61 av. du Président Wilson
94235 Cachan Cédex, France
morel@cmla.ens-cachan.fr

Pablo Musé
Instituto de Ingeniería Eléctrica
Facultad de Ingeniería
Julio Herrera y Reissig 565
11300 Montevideo, Uruguay
pmuse@fing.edu.uy

Frédéric Sur
Loria Bat. C - projet Magrit
Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cédex
France
sur@loria.fr

ISBN 978-3-540-68480-0 e-ISBN 978-3-540-68481-7
DOI 10.1007/978-3-540-68481-7

Lecture Notes in Mathematics ISSN print edition: 0075-8434
ISSN electronic edition: 1617-9692

Library of Congress Control Number: 2008927359

Mathematics Subject Classification (2000): 62C05, 62G10, 62G32, 62H11, 62H15, 62H30, 62H35, 68T10, 68T45, 68U10, 91E30, 94A08, 94A13, 94B70

Cover design: WMX Design Bender

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

Recent years have seen dramatic progress in shape recognition algorithms applied to ever-growing image databases. They have been applied to image stitching, stereo vision, image retrieval, image mosaics, solid object recognition and video and web shape retrieval. More fundamentally, the ability of humans and animals to detect and recognize shapes is one of the enigmas of perception. Digital images and computer vision methods open new ways to address this enigma.

Given a dictionary of digitized shapes and a previously unobserved digital image, the aim of shape recognition algorithms is to know whether some of the shapes in the dictionary are present in the image. This book describes a complete method that starts from a query image and an image database and yields a list of the images in the database containing the query shapes.

Technically speaking there are two main issues. The first is extracting invariant shape descriptors from digital images. Indeed, a shape can be seen from various angles and distances and in various lights. A shape can even be partially occluded by other shapes and still be identifiable. Because the extraction step is so crucial, three acknowledged shape descriptors, SIFT (Scale-Invariant Feature Transform), MSER (Maximally Stable Extremal Regions) and LLD (Level Line Descriptor) will be introduced.[1]

The second issue is deciding whether two shape descriptors are identifiable as the *same shape* or not. This decision process will derive from a unique paradigm, called the Helmholtz principle. For each decision a background model is introduced. Then one decides whether an event of interest (such as the presence of a shape in the image) has occurred if it has a very low probability of occurring by chance in the background model. Thus from the statistical viewpoint shape identification goes back to *multiple hypothesis testing.*

A shape descriptor is recognized if it is not likely to appear by chance in the background model. At a higher complexity level, a group of shape descriptors is recognized if its spatial arrangement could not occur just by chance. These two decisions

[1] In a recent review paper on affine invariant recognition written by a pool of experts, SIFT and MSER were actually acclaimed as the best shape descriptors [122].

rely on simple stochastic geometry and eventually compute a false alarm number for each shape descriptor. The lower this number, the more secure the identification. In that way most familiar simple shapes or images can be reliably identified. Many realistic experiments show false alarm rates ranging from 10^{-5} to less than 10^{-300}.

All in all these lecture notes prove that many shapes can indeed be identified. For these shapes one needs no *a priori* model and no training, just one sample of the shape and what statisticians call a *background model*, or *a null model*. In the case of shape recognition, the term background is to be taken to the letter. By the Helmholtz principle a shape is conspicuous if and only if it cannot be generated by the image background on which it is perceived. The background model is therefore easily learnt from the image database itself.

The above description should not be taken to suggest that the shape recognition problem is solved. The methods described only apply to solid shapes and not to deformable shapes. They only deal with individual shapes and images such as logos or paintings, and not with wide classes of objects such as all humans, all cats or all cars. This latter problem is known as *categorization* and is still widely open to research.

The authors are indebted to their collaborators for many important comments and corrections, particularly to Andrés Almansa, Yann Gousseau and Guoshen Yu. David Mumford and another anonymous referee made valuable comments which reshaped the book. All experiments were done using the public software MegaWave (authors: Jacques Froment and Lionel Moisan). The SIFT method is also public and downloadable.

The present theory was mainly developed at the Centre de Mathématiques et Leurs Applications, at ENS Cachan, at the Universitat de les Illes Balears and at IRISA, Rennes. It was partially financed for the past eight years by the Centre National d'Etudes Spatiales, the Centre National de la Recherche Scientifique, the Office of Naval research (grant N00014-97-1-0839) and the Ministère de la Recherche (project ISII-RNRT), and the Ministerio de Educación y Cultura (project MTM2005-08567). Special thanks to Bernard Rougé and Wen Masters for their great interest and support. We are indebted to Nick Chriss for numerous stylistic corrections.

Frédéric Cao
José Luis Lisani
Jean-Michel Morel
Pablo Musé
Frédéric Sur

Contents

Part II Level Line Invariant Descriptors

Part III Recognizing Level Lines

Chapter 1
Introduction

1.1 A Single Principle

Digital images became an object of scientific interest in the seventies of the last century. The emerging science dealing with digital images is called *Computer Vision*. Computer Vision aims to give wherever possible a mathematical definition of visual perception. It can be therefore viewed in the realm of perception theory. Images are, however, a much more affordable object than percepts. Indeed, digital images are sampled real or vectorial functions defined on a part of the plane, usually a rectangle. They are accessible to all kinds of numerical, geometric, and statistical measurements. In addition, the results of artificial perception algorithms can be confronted to human perception. This confrontation is both advantageous and dangerous. Experimental results may easily be misinterpreted during visual inspection. The results look disappointing when matched with our perception. Obvious objects are often very hard to find in digital images by an algorithm.

In a recent book by Desolneux et al. [54], a general mathematical principle, the so called Helmholtz principle, was extensively explored as a way to define all visual percepts (gestalts) as large deviations from randomness. According to the main thesis of these authors one can compute detection thresholds deciding whether a given geometric structure is present or not in each digital image. Several applications of this principle have been developed by these authors and others for the detection of alignments [50], boundaries [51, 35], clusters [53, 33], smooth curves and good continuations [30, 31], vanishing points [2] and robust point matches through epipolar constraint [128].

These works make extensive use of a computed function, the so called *number of false alarms* (NFA) of a perception. The NFA of a perception is the expected number of times this perception could have arisen just by chance in the background. An observed configuration in an image can be numerically defined as a perception if and only if its NFA is smaller than 1. Experimental evidence has confirmed that the NFA of many human percepts of geometric figures is actually much smaller than 1,

F. Cao et al., *A Theory of Shape Identification*. Lecture Notes in Mathematics 1948.

typically less than 10^{-n} where n ranges from 10 to 100 and more. See [52] and the textbook [54].

Thus theory and experiments give a mathematical and experimental basis to the existence of sure percepts. Their existence had been stated for a long time by phenomenology, in particular the Gestalt theory [95], but without quantitative evidence.

The idea that perceptions are objects unlikely to form just by chance in a background goes back to Helmholtz [83]. This principle could not be tested until images became digital and therefore accessible to computational experiments. Before the above-mentioned works, there had been several attempts to define percepts as exceptional events. Stewart [168] proposed to detect planes in a cloud of points by what he called the MINPRAN method. He computed the probability that MINPRAN will "hallucinate a fit where there is none". This probability was computed under the *a contrario* assumption that the points were randomly distributed. Lowe [112] proposed a detection framework based on the computation of accidental occurrence.

> *In other words, we can shift our attention from finding properties with high prior expectations to those that are sufficiently constrained to be detectable among a realistic distribution of accidentals.[...] Even when we do not know the ultimate interpretation for some grouping and therefore its particular* a priori *expectation, we can judge it to be significant based on the non-accidentalness criteria.*

Later Barlow [17] interpreted the perceptual goals of the neocortex as a search for *suspicious coincidences*. In the same spirit, Grimson and Huttenlocher [75, 76] proposed to compute shape recognition thresholds from a null model viewed as *the conspiracy of random*.

There is a common sense objection to applying the Helmholtz principle to shape recognition. Watching the sky, one often sees castles, cats and dogs in the clouds. Humans have a high capacity for hallucinating familiar generic shapes such as faces in rich visual environments. Thus the Helmholtz principle is not suited for all sorts of shapes.

The situation is, however, quite different regarding more specific iconic shapes such as letters, logos and in general solid shapes. One sees faces in the sky, but certainly not this or that particular face. It is to be expected that any complex enough solid shape will be recognizable in the Helmholtz sense: no random arrangement would be able to reproduce it accurately.

The mathematics to prove this are quite simple. Let $\mathcal{S}$ and $\mathcal{S}'$ be two shapes observed in two different images and which happen to be similar. Denote their (small) Hausdorff distance after registration by $\delta = d(\mathcal{S}, \mathcal{S}')$. Assume we know enough of the background model to compute the probability $\Pr(\mathcal{S}, \delta) = \Pr(d(\mathcal{S}, \Sigma) \leqslant \delta)$ that some shape in the background, Σ be as similar to $\mathcal{S}$ as $\mathcal{S}'$ is. If this probability is very small one can deduce that $\mathcal{S}'$ does not look like $\mathcal{S}$ just by chance. Then $\mathcal{S}$ and $\mathcal{S}'$ will be *identified as the same shape*.

Digital images contain thousands of significant *shape elements* that constitute their shape contents. (Several kinds of *shape elements, or descriptors* will be considered in this book.) Controlling the number of wrong matches involves the computation of the probability of a casual match with the background. This probability should be very small. But also the number of tests, which can be huge, must be under control.

Definition 1. Let $\mathcal{I}$ and $\mathcal{I}'$ be two images and N, N' the number of shape elements in each. Let $\mathcal{S}$ and $\mathcal{S}'$ be two shape elements extracted from I and I' respectively, lying at distance δ. We call number of false alarms of the match between $\mathcal{S}$ and $\mathcal{S}'$ the number

$$\mathrm{NFA}(\mathcal{S}, \mathcal{S}') = N \cdot N' \cdot \Pr(d(\mathcal{S}, \Sigma) \leqslant \delta).$$

If $\mathrm{NFA}(\mathcal{S}, \mathcal{S}')$ is much smaller than 1, one deduces that $\mathcal{S}'$ could not look like $\mathcal{S}$ just by chance and concludes that $\mathcal{S}$ and $\mathcal{S}'$ have *the same shape*.
There is an important phenomenological consequence: Shapes can be defined without learning, that is without empirical knowledge. By definition a shape is any part of an image which has been identified (in the sense of low NFA) at least once in another image.

From an empirical point of view, there are two kinds of shapes. First, any solid physical object can be photographed under many views and illuminations. If, by using the above definition, two snapshots of the same physical object happen to contain recognizable shape elements, one may say that the object itself is identifiable. These shape elements will constitute the object signature.

Second, humans build all kinds of standardized objects. Likewise, two different standard objects can be identified if they stem from the same industrial process. This also applies to the very numerous iconic planar shapes generated by human visual communication, in particular characters and logos. In the experimental parts of this book, we shall study the identifiability of several such iconic shapes: the lower the NFAs they generate at a given Hausdorff distance, the more recognizable they are.

As a consequence of the present study, one can define solid shapes as equivalence classes of recognized pairs without reference to empirical knowledge or *ground truth*. Thus, one should demonstrate the existence of, say, the Coca-Cola logo just by the fact that a certain group of shape elements appears in several images with very low NFA for all pairwise comparisons. Experiments will compare several snapshots of the same painting or poster, various images extracted from the same movie, or various logos of the same firm. The aim in all cases is to single out and group in clusters all shape elements common to both images. Conversely, the same method gives a negative answer when two images have no shape in common. In that case the NFA is above 1, which means that the shape is likely to occur casually in the background.

From the mathematical and numerical point of view, the main challenge in the whole study is the accurate computation of numbers of false alarms (NFA). This requires the computation of very small probabilities. Small probabilities cannot be directly measured from a shape database as frequencies. Thus, a probabilistic model of the set of all possible shapes should be built. Such a realistic experimental *background model* should be made of a large and representative set of all kinds of digital images. Unfortunately, there is no available probabilistic model for a large set of images. It is as hopeless as building a global model of the world. Even if such a model were available, one would still face the challenge of computing accurately the probability of very rare events in this world model.

Fortunately enough it is possible to overcome or rather to circumvent these two obstacles. The only information needed is the probability for a background shape to be very close to a given query shape. By a geometric independence argument, this probability will be made into a product of much larger probabilities. These probabilities instead become observable as frequencies in a small image database.

1.2 Shape Invariants and Consequences

1.2.1 Shape Distortions

In order to find the shape invariance classes, it suffices to give a rough typology of the transformations that affect images but not our recognition of the shapes they contain. Following Lisani *et al.* [109], the main classes of perturbations which do not affect recognition are:

1. **Changes in the color and luminance scales (contrast changes).** According to Gestaltists Attneave [13] and Wertheimer [179], shape perception is independent of the gray level scale or of measured colors.

   > *The concentration of information in contours is illustrated by the remarkable similar appearance of objects alike in contour and different otherwise. The "same" triangle, for example, may be either white on black or green on white. Even more impressive is the familiar fact that an artist's sketch, in which lines are substituted for sharp color gradients, may constitute a readily identifiable representation of a person or thing.* Attneave, 1954.

   > *I stand at the window and see a house, trees, sky.*
   > *Theoretically I might say there were 327 brightnesses and nuances of color. Do I have "327"? No. I have sky, house, and trees. It is impossible to achieve "327" as such. And yet even though such droll calculation were possible and implied, say, for the house 120, the trees 90, the sky 117 – I should at least have this arrangement and division of the total, and not, say, 127 and 100 and 100; or 150 and 177.* Wertheimer, 1923.

   Refer to Fig. 1.1 designed by E. H. Adelson for a striking illustration of illumination invariance.

2. **Occlusions and background modification.** Shape recognition can also be performed in spite of occlusion and varying background, as shown in Fig. 1.2. The phenomenology of occlusion was thoroughly studied by Kanizsa [95] who argues that occlusion is always present in every day vision: most objects are partially hidden by others. Human perception must therefore be able to recognize partial shapes. Conversely, if a shape occludes a background, its recognition is invariant to changes in the background. This independence of shape

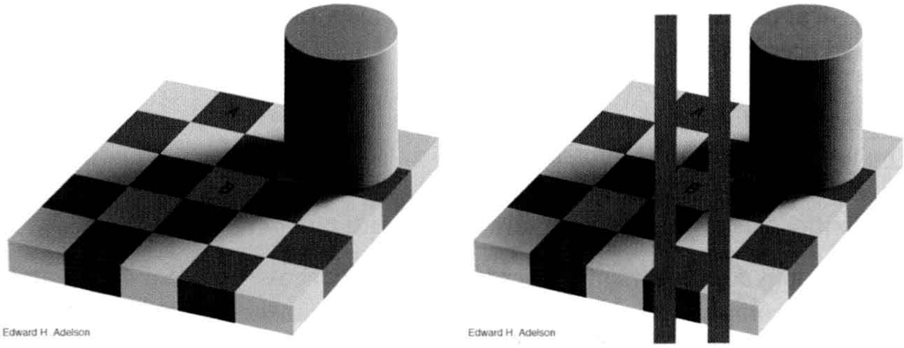

Fig. 1.1 Contrast change invariance. In the left hand image, the A and B squares have exactly the same gray level. This incredible fact is easily checked in the right hand image where A and B are linked by two rectangles with the same gray level. This experiment by E.H. Adelson illustrates the unreliability of brightness perception and the invariance of shape recognition with respect to illumination changes. (Courtesy E.H. Adelson, http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html)

recognition from its background is known in perception psychology as the *figure-background problem* (Rubin [153]). The figure-background problem is part of the occlusion problem. A shape is superimposed on a background, which can be made of various objects. How can the shape be singled out from that clutter? This poses a dilemma. Extract the shape and then recognize it or extract it *because* it has been recognized?

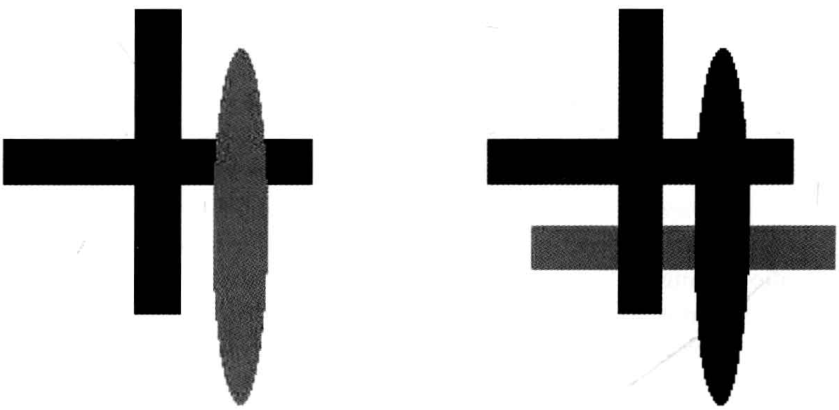

Fig. 1.2 Left: According to Kanizsa and his school, shapes can be recognized even when they undergo several occlusions. Our perception is trained to recognize shapes which are only seeable in part. Here the occluded cross can be easily recovered. Right: the *figure-background* problem. Our perception is adapted to recover a figure on the foreground, independently from the background

3. **Classical noise and blur**, inherent to any perception task and to any image generated according to Shannon's theory.

4. **Geometrical distortions or deformations.** Perspective is deeply incorporated in human perception. Humans can recognize objects and shapes under perspective distortion as long as perspective is not too strong. Recognition is also invariant to elastic deformations, always within some limits.

The previous four invariant properties fix requirements a good image representation should comply with. It will be necessary to formulate a mathematical model for each of them and to derive a well adapted image representation.

1. a. **The local contrast invariance requirement.** A digital image is usually defined as a function $u(x)$, where $u(x)$ represents the gray level or luminance at x. The first task is to extract from the image topological information independent from the varying and unknown contrast change function of the optical or biological apparatus. One can model such a contrast change function as any continuous increasing function g from $\mathbb{R}^+$ to $\mathbb{R}^+$. The real datum corresponding to the observed u could be as well any image $g(u)$. This simple argument can lead to select the level sets of the image [161] or its set of level lines as a complete contrast invariant image description [37]. If u is of class say C^1, then the level lines are the connected components of $u^{-1}(\lambda)$, which are C^1 curves for almost every $\lambda \in \mathbb{R}$. This is the choice adopted by the LLD (level line descriptor) and MSER (maximally stable extremal regions) methods. Another way to handle the contrast invariance requirement is to encode only the direction of the gradient of u, $\frac{Du}{|Du|}$ and not the gradient Du itself. Indeed, the direction of gradient is normal to the level line and is not altered by any increasing contrast change. This is the way adopted by the SIFT method to cope with contrast changes.
   b. **The concentration of information requirement.** Somewhat in contradiction to this contrast invariance principle, the Gestaltist Attneave [13] asserted that *"Information is concentrated along contours (i.e., regions where color changes abruptly)"*. Indeed not all the level lines are needed to have a complete shape description. Most of them are due to noise or to tiny illumination changes. Thus, it makes sense to select only the most contrasted level lines. That is to say, those along which the gradient of u is large enough. Such a selection is not invariant to all contrast changes, since it explicitly uses the gradient value. However, it is still invariant to affine global contrast changes. Figure 1.3 shows an example of such level lines selection. The selection of the most contrasted level lines will be the subject of Chap. 2. It will be applied to the LLD and the MSER methods. The SIFT methods actually weights its gradient orientation histograms by the gradient magnitude (see Chap. 10).
2. **The occlusion and figure-background requirements.** Even the best adapted choice of level lines is not totally suited to describing image parts. Indeed, when a shape A partially occludes a shape B, the level lines of the resulting image

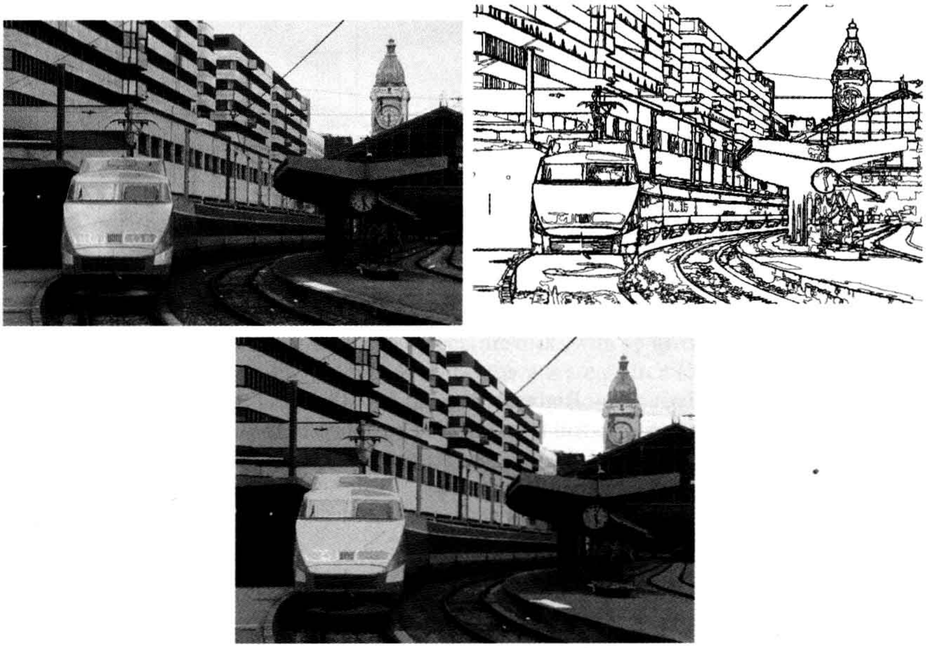

Fig. 1.3 Top left: original image, 83,759 level lines. Top right: meaningful boundaries (883 level lines). Bottom: reconstruction from the meaningful boundaries. Only 883 boundaries remain. The structure of the image is preserved and perceptual loss is very weak. The LLD and MSER methods use these boundaries for building up normalized shape descriptors. See Chap. 4

are a concatenation of pieces of the level lines belonging to A and to B. This is shown with a very simple example in Fig. 1.4. Even if a shape is not occluded, but simply occludes its own background, there may be no level line surrounding the whole shape, as shown in Fig. 1.5. These remarks show that whole level lines are too big and too sensitive to occlusion. In order to overcome this obstacle the general idea is to build shape recognition on shape elements as local as possible. The SIFT method takes small image patches. The LLD method splits level lines into small pieces.

3. **The smoothing requirement.** It is an easy experiment to check that shapes are easily recognized in images subject to noise. This means that shape information is not affected by noise. Noise introduces details which are too fine (in relation to the essential shape information) to be perceptually relevant in terms of recognition. Quoting Attneave (ibid., 1954):

> *It appears, then, that when some portion of the visual field contains a quantity of information grossly in excess of the observer's perceptual capacity, he treats those components of information which do not have redundant representation somewhat as a statistician treats "error variance", averaging out particulars and abstracting certain statistical homogeneities.*