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

Chapter 11
Securing SIFT with *A Contrario* Techniques

Abstract In the previous chapter two shortcomings of Lowe's SIFT algorithm have been pointed out, namely its low matching efficiency (ratio between the number of correct matches and the total number of matches) and its inability to match several instances of the same object. The grouping stage of the method also is widely empirical and requires some fix.
In this chapter we shall examine three easy improvements of the SIFT method, all based on the *a contrario* techniques developed in the present book. They permit to treat all raised issues. The first one (Sect. 11.1) is the direct application of the theory for *a contrario* grouping of transformations developed in Chap. 8. The second one (Sect. 11.2) is the use of a background model for SIFT matches which prevents the elimination of multiple matches. Finally Sect. 11.4 yields an efficient *a contrario* technique computing a NFA for each SIFT match. In summary, the aim is to demonstrate that the whole SIFT algorithm can be secured and associated realistic NFAs, as we did in Chap. 5 and 8 for the LLD method.

11.1 *A Contrario* Clustering of SIFT Matches

The problem of matching efficiency of the SIFT algorithm was already remarked in [114] by D. Lowe. He proposed to address this problem by a generalized Hough transform, in order to identify subsets of matching key points that also agree on location, scale, and orientation. Quoting Lowe:

> *The correct matches can be filtered from the full set of matches by identifying subsets of keypoints that agree on the object and its location, scale, and orientation in the new image. The probability that several features will agree on these parameters by chance is much lower than the probability that any individual feature match will be in error. The determination of these consistent clusters can be performed rapidly by using an efficient hash table implementation of the generalized Hough transform. Each cluster of 3 or more features that agree on an object and its pose is then subject to further detailed verification. First, a least-squared estimate is made for an affine approximation to the object pose. Any other image features consistent with this pose are identified, and outliers are discarded. Finally,*

F. Cao et al., *A Theory of Shape Identification*. Lecture Notes in Mathematics 1948.
© Springer-Verlag Berlin Heidelberg 2008

> *a detailed computation is made of the probability that a particular set of features indicates the presence of an object, given the accuracy of fit and number of probable false matches. Object matches that pass all these tests can be identified as correct with high confidence.*

In this section we propose to develop this technique and to replace the Hough transform by a clustering step identical to the one described in Chap. 8 and 9.

The location, scale and orientation of each one of the SIFT matching pairs can be represented as a point in the space of similarity transformations. These points can be grouped together using the technique of Chap. 8. The resulting meaningful clusters correspond to sets of spatially coherent matches. Matches not belonging to any meaningful cluster are rejected as wrong.

The following images illustrate the use of this clustering technique. In Fig. 11.2 the clustering method has been applied to a pair already shown in the previous chapter. Each circle represents a SIFT key point that got a match. As usual, the radius represents the scale of the key point. White circles represent correct matches, black circles represent wrong matches. The results for the original SIFT method are shown on the left part of the image. On the right side, the new results using the clustering technique are shown. The most meaningful cluster is displayed and the key points of the SIFT descriptors contributing to the cluster are shown in white. The efficiency of the matching procedure increases from 37.9% to 67.8%. Notice that efficiency did not reach 100% because some imprecise (though not completely incoherent) matches were included in the maximal meaningful cluster. For all of them, the location and scale are quite close to the one expected according to the underlying homography, this explains why these matches were included in the cluster.

The second figure (Fig. 11.1) compares the results of the original SIFT algorithm (left) and those obtained after clustering (right). As can be observed, the final number of matches decreases but the efficiency increases significantly (from 82.1% to 100%).

In the next figures we show an example with several clusters of matches. Figure 11.3 displays the result of the comparison of two of the images displayed in the previous chapter (Fig. 10.8). Observe that some of the matches are wrong. After applying the clustering step two clusters of matches are found (see Fig. 11.4). All of the matches in these clusters are correct.

11.2 Using a Background Model for SIFT

The matching algorithm for SIFT used in the previous chapter consists in selecting the matches by thresholding the ratio between the distance from the query match to the first closest database match and the distance from the query match to the second closest one. This ensures the selected matches to be clearly separated from the clutter. The best threshold was determined empirically by Lowe by making a statistics on 40,000 key points:

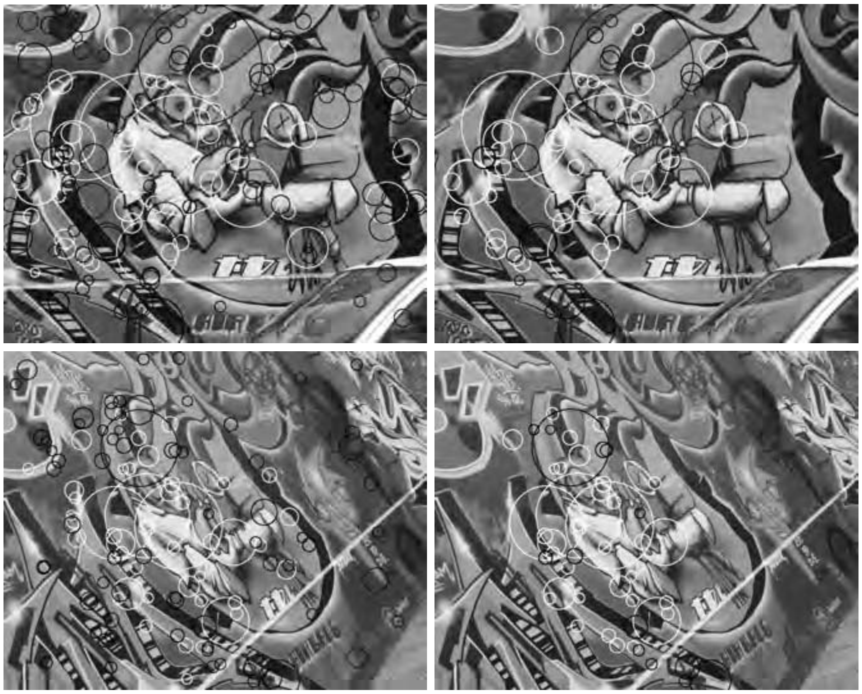

Fig. 11.1 Left: SIFT matching (the radius represents the scale of the key point; white circles represent correct matches, black circles represent wrong matches). Right: SIFT matching followed by *a contrario* clustering (white circles mark the key points belonging to the cluster, all of them are correct matches). Top: image 1; bottom, image 4

> *The probability that a match is correct can be determined by taking the ratio of distance from the closest neighbor to the distance of the second closest. Using a database of 40,000 keypoints (...).*

Actually it is clear that this empirical probability is not given by a model. Another obvious drawback is that no repeated match can be detected.

A very simple way to overcome this last drawback consists in using a third image as *background model* for learning the rejection thresholds. Here we are applying the very same SIFT method, but we just notice that any image can be used as background model (the one we have used in our tests is shown in Fig. 11.5). The matching procedure simply consists in selecting the matches by thresholding the ratio between the distance from the query match to the database match and the distance from the query match to the closest match in the background model. As a consequence of this modification several new matches have been detected and, in particular, several instances of the same object have now been detected (see Figs. 11.6 and 11.7).

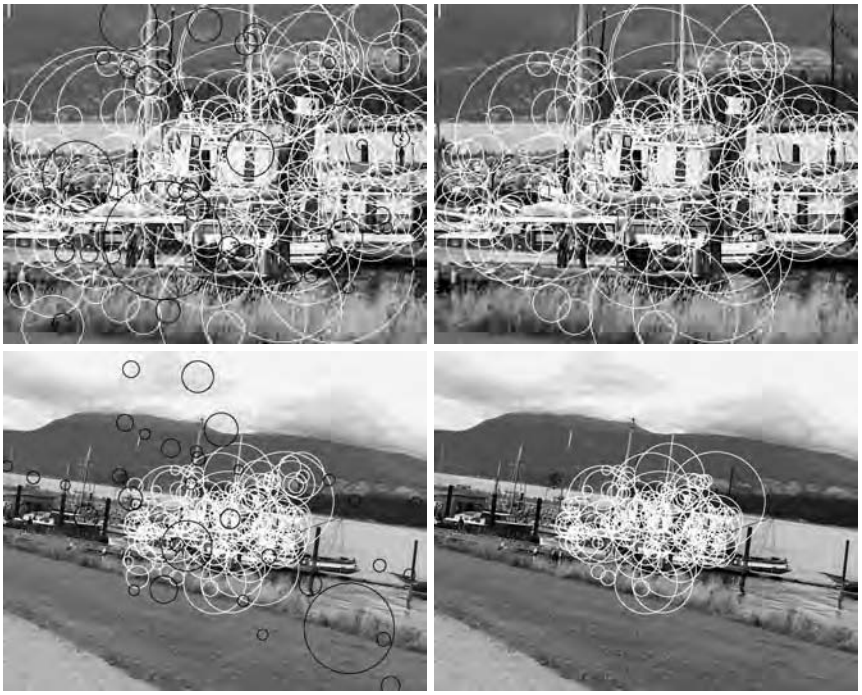

Fig. 11.2 Left: SIFT matching (the radius represents the scale of the key point; white circles represent correct matches, black circles represent wrong matches). Right: SIFT matching followed by *a contrario* clustering. Top: image 1; bottom: image 5

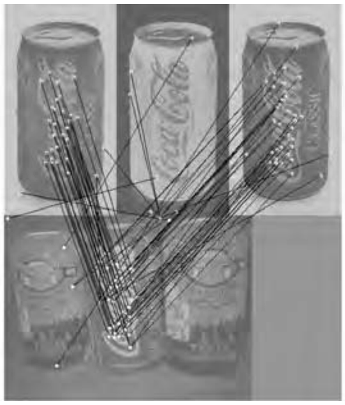

Fig. 11.3 SIFT matching between top and bottom images

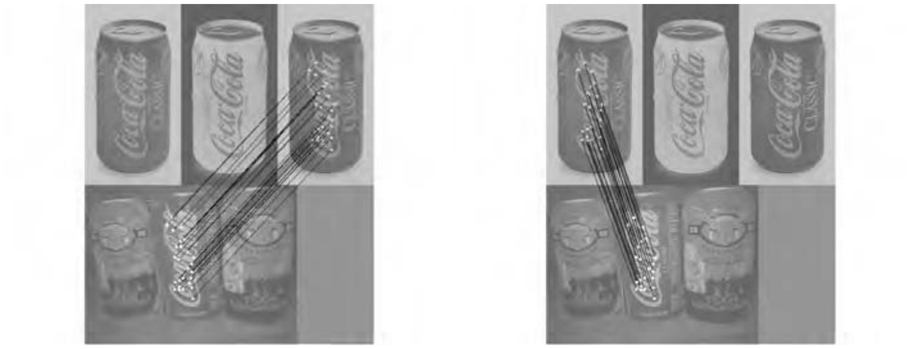

Fig. 11.4 SIFT matching between top image (query) and bottom image (database), followed by *a contrario* clustering. Two maximal meaningful clusters were detected

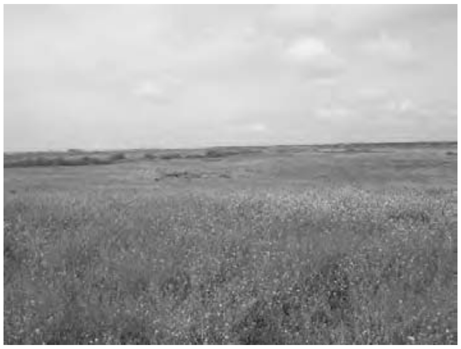

Fig. 11.5 Image used as background model for the rejection thresholds of matches

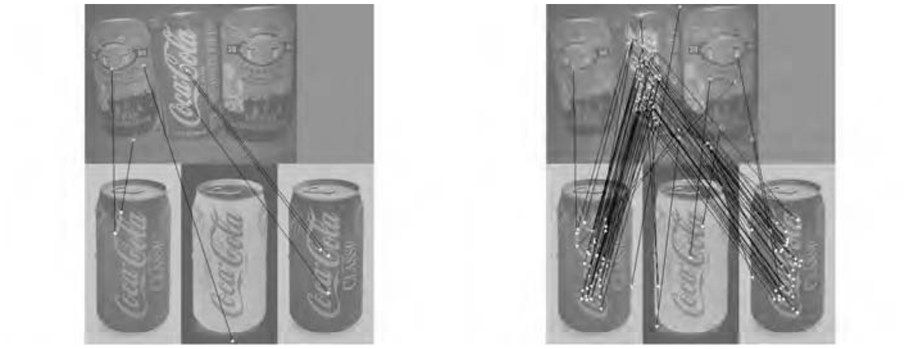

Fig. 11.6 Left: original SIFT matching result. Since the same shapes occur twice in the bottom image, the standard SIFT threshold procedure eliminates almost all matches. Right: result when using a background model. Top: query image; bottom: database image. The wrong matches can be eliminated by clustering in the transformations' space, as illustrated in the next figure 11.7

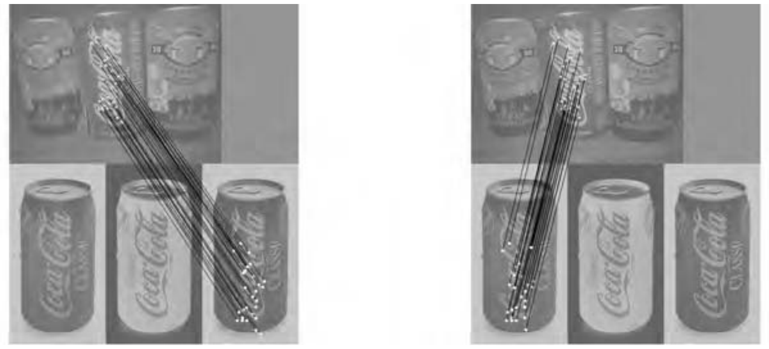

Fig. 11.7 SIFT matching between the top image (query) and the bottom image (database), with background model thresholding, followed by *a contrario* clustering. Two maximal meaningful groups were detected

11.3 Meaningful SIFT Matching

The SIFT algorithm is based on the use of a set of descriptors for each key point detected in the image scale-space. Descriptors proposed in [114] are based on local histograms of gradient directions. Matching these descriptors involves a threshold computed empirically from the images themselves. In order to make the method more robust, this threshold should be derived from statistical arguments. In this section, we intend to do so, by deriving the matching threshold following an *a contrario* approach. We shall propose a new SIFT descriptor for which an automatic matching strategy similar to the one presented in Chap. 5 can be applied. The new SIFT algorithm exhibits high matching efficiency and is able to detect several instances of the same object. And again, the clustering step proposed in Sect. 11.1 can be applied to further improve the results.

11.3.1 Normalization

Exactly the same key points as in [114] will be used in the proposed algorithm. We gave some details on the procedure to compute them in Sect. 10.1. Each key point comes with a position, but also with its scale and an orientation (which is one of the dominant gradient directions in the vicinity of the point). Hence it is characterized by an element $(x, s, \varphi) \in \mathbb{R}^2 \times \mathbb{R}_+ \times (-\pi, \pi)$. There are usually a few hundreds key points in a 512×512 image. Even though a very simple interpolation procedure attempts to refine the position of the key points, Lowe estimates that the position error is of the magnitude of the scale of the interest point, the error on the orientation is $\pm 15°$, and the scale is determined up to a $\sqrt{2}$ factor. It turns out that the accuracy is often much better than that, thus permitting a fair enough registration of the images. A pair of interest points (x, s, φ) and (y, t, ψ) in two images defines a unique similarity (four scalar parameters). This similarity writes

$$F(\xi) = zR\xi + b,$$

where $z = \frac{t}{s}$, R is the plane $\psi - \varphi$ rotation, and $b = y - zRx$. Let us assume that u is a gray level image and that v is obtained by applying the similarity F to u followed by a contrast change g (we recall that contrast changes are modeled by an increasing function $g : \mathbb{R} \to \mathbb{R}$). Thus, $v(x) = g(u(zRx + b)) = g(u(F(x)))$. An elementary calculation shows that

$$Dv(F(x)) = g'(u)zDu(F(x))R, \tag{11.1}$$

meaning that the gradient of u has simply been rotated by the rotation R^{-1} and multiplied by a positive number. Therefore the direction of the gradient of v at point $F(x)$ is obtained by rotating the direction of the gradient of u at x.

11.3.2 Matching

Let us assume that u and v are two images, or pieces of images, of the same size belonging respectively to a set of images $\mathcal{Q}$ (for query) and $\mathcal{B}$ (for base). Let us denote by $N_{\mathcal{Q}}$ and $N_{\mathcal{B}}$ the cardinality of $\mathcal{Q}$ and $\mathcal{B}$. How to compare u and v and to come to the robust decision that they are similar? A similarity measure is often defined as a distance between descriptors. The comparison relies on the following fact: two images differing by a contrast change have the same gradient direction. On the contrary, if u and v are not related, then so are the directions of their gradient. In this case, it is sound to assume *a contrario* that the difference of these directions (in absolute value) is a uniform random variable in $(0, \pi)$, and independent of the values taken at remote enough points.

The *a contrario* approach consists in deciding that two images are actually similar when their observed similarity could not occur just by chance. More precisely, we have to check if the gradients of u and v are much more often aligned than the *a contrario* model can allow. For any point x, let us denote by $D(x)$ the difference of the directions of $Du(x)$ and $Dv(x)$. It is a number in the interval $(0, \pi)$, defined if both $Du(x)$ and $Dv(x)$ are nonzero. In order to avoid quantization effects on the gradient direction, only points where the gradient norm is larger than $\tau > 0$ can be considered. In practice, $\tau = 5$.

Let x_1, ..., x_M, be M points in the image domain of u. The way they are chosen will be the object of a further careful analysis. Let us consider the following *a contrario* hypothesis.

$\mathcal{H}_0$: the M values $(D(x_i))_{1 \leqslant i \leqslant M}$ are i.i.d., uniform in $(0, \pi)$.

Again, this hypothesis is clearly false if the images are similar. Our purpose is precisely to adequately reject this hypothesis. Let $\alpha \in (0, \pi)$, and $q_\alpha = \frac{\alpha}{\pi}$. The probability, under $\mathcal{H}_0$, that at least k among the M values $\{D(x_1), \dots D(x_M)\}$ are less than α is given by the tail of the binomial law

$$B(M, k, q_\alpha) = \sum_{j=k}^{M} \binom{M}{j} q_\alpha^j (1 - q_\alpha)^{M-j}. \tag{11.2}$$

Otherwise said, $B(M, k, q_\alpha)$ is the probability that the directions of Du and Dv coincide at (at least) k points out of M, *by chance*. If for two images u and v, we indeed observe k such points and if $B(M, k, q_\alpha)$ happens to be very small, then chance is certainly not a good explanation. Let us note, however, that if the image sets $\mathcal{Q}$ and $\mathcal{B}$ are very large, then such an observation may indeed be casual. The fact that an observation should be considered as surprising (or not) depends on the size of the database. This leads us to the following definition, which follows the Desolneux et al. method [50]. For a detailed account, we refer to the book [54].

Definition 19. Let $0 \leqslant \alpha_1 \leqslant \dots \leqslant \alpha_L \leqslant \pi$ be L values in $[0, \pi]$. For any $(u, v) \in \mathcal{Q} \times \mathcal{B}$, we call number of false alarms of (u, v) the quantity

$$\mathrm{NFA}(u, v) = N_{\mathcal{Q}} \cdot N_{\mathcal{B}} \cdot L \cdot \min_{1 \leqslant i \leqslant L} B(M, k_i, q_{\alpha_i}), \tag{11.3}$$

where k_i is the cardinality of

$$\{j,\ 1 \leqslant j \leqslant M, D(x_j) \leqslant \alpha_i\}.$$

We say that (u, v) is ε-meaningful, or that u and v are ε-similar if $\mathrm{NFA}(u, v) \leqslant \varepsilon$.

Since numbers of false alarms can be very small, the logarithmic scale is more intuitive and we call meaningfulness of (u, v) the value $\mathcal{M}(u, v) = -\log_{10}(\mathrm{NFA}(u, v))$.

The interpretation of this definition will be made clear after stating the following proposition. We put its proof for a sake of completeness, but it it just a variant of the other meaningfulness propositions in the present book, in particular Props. 8 and 10.

Proposition 12. *For two image sets $\mathcal{Q}$ and $\mathcal{B}$ such that $\mathcal{H}_0$ holds, the expected number of ε-meaningful pairs is less than or equal to ε.*

Proof. For all i, let us denote by K_i the random number of points among the x_j such that $D(x_j)$ is less than α_i. For any v, (u, v) is ε-meaningful, if there is at least $1 \leqslant i \leqslant L$ such that $N_{\mathcal{Q}} \cdot N_{\mathcal{B}} \cdot L \cdot B(M, K_i, q_{\alpha_i}) < \varepsilon$. Let us denote by $E(u, v, i)$ this event. Its probability $P_{\mathcal{H}_0}(E(u, v, i))$ satisfies

$$P_{\mathcal{H}_0}(E(u, v, i)) \leqslant \frac{\varepsilon}{L \cdot N_{\mathcal{Q}} N_{\mathcal{B}}}.$$

Indeed, for any real random variable X with survival function $H(x) = \Pr(X > x)$, it is a classical fact that $\Pr(H(X) < x) \leqslant x$. By applying this result to K_i, we get the upper bound on $P_{\mathcal{H}_0}(E(u, v, i))$. The event $E(u, v)$ defined by "(u, v) is ε-meaningful" is $E(u, v) = \cup_{1 \leqslant i \leqslant L} E(u, v, i)$. Let us denote by $\mathbb{E}_{\mathcal{H}_0}$ the mathematical expectation under $\mathcal{H}_0$. Then

$$\begin{aligned}\mathbb{E}_{\mathcal{H}_0}\left(\sum_{u\in\mathcal{Q},\, v\in\mathcal{B}} \mathbf{1}_{E(u,v)}\right) &= \sum_{u\in\mathcal{Q},\, v\in\mathcal{B}} \mathbb{E}_{\mathcal{H}_0}(\mathbf{1}_{E(u,v)}) \\ &\leqslant \sum_{\substack{u\in Q,\, v\in\mathcal{B}\\ 1\leqslant i\leqslant L}} P_{\mathcal{H}_0}(E(u,v,i)) \\ &\leqslant \sum_{\substack{u\in\mathcal{Q},\, v\in\mathcal{B}\\ 1\leqslant i\leqslant L}} \frac{\varepsilon}{LN_{\mathcal{Q}}N_{\mathcal{B}}} = \varepsilon. \qquad \square\end{aligned}$$

Remark that if $\mathcal{Q}$ and $\mathcal{B}$ are white noise images, then $\mathcal{H}_0$ trivially holds. For two such bases, any detection is a false alarm. Indeed, it is *a priori* known that the images are unrelated, which does not mean that they have nothing in common. The number of false alarm quantifies what has to be accepted as a casual similarity. Hence, by setting $\varepsilon = 1$, one (false) detection may be observed on average in databases of noise images. Obviously, this still holds if only one of the images u and v is made of noise. It actually turns out that $\mathcal{H}_0$ is reasonable if the sample points x_1, ..., x_M are carefully chosen, as discussed in Sect. 11.3.3.

Thus, Def. 19 together with Prop. 12 mean that there are, on average, less than ε pairs of images (u, v) in $\mathcal{Q} \times \mathcal{B}$ that match by chance, that is to say, when $\mathcal{H}_0$ holds. Under this hypothesis, any detection must be considered as a false alarm (hence the denomination of NFA). Thus, it is chosen to eliminate any observation having a frequency of the order of ε in the *a contrario model*.

The values q_α are simply quantization steps and are known *a priori*. Hence, it is possible to tabulate the values of the binomial law once and for all, and to rapidly compute the number of false alarms. It is possible to figure out the behavior of the NFA with respect to the parameters, thanks to the following asymptotic expansion, first proved by Hoeffding (for more details, see the original article [84] and the textbook [54]).

Proposition 13. *Let* $H(r,p) = r\ln\frac{r}{p} + (1-r)\ln\frac{1-r}{1-p}$, *be the relative entropy of two Bernoulli laws with parameters* r *and* p. *Then, for* $k \geqslant Mp$,

$$B(M,k,p) \leqslant \exp\left(-M\cdot H\left(\frac{k}{M},p\right)\right). \tag{11.4}$$

This inequality leads to the following sufficient condition of meaningfulness.

Corollary 2. *If*

$$\max_{\substack{1\leqslant i\leqslant L\\ k_i\geqslant Mq_{\alpha_i}}} H\left(\frac{k_i}{M}, q_{\alpha_i}\right) > \frac{1}{M}\ln\frac{LN_{\mathcal{Q}}N_{\mathcal{B}}}{\varepsilon}, \tag{11.5}$$

the pair (u,v) *is* ε*-meaningful.*

In this corollary, it appears clearly that the value of k such that (u, v) is ε-meaningful only depends on the logarithm of L, $N_{\mathcal{Q}}$, $N_{\mathcal{B}}$ and ε. In practice, we choose L about 10 which is compatible with our perceptual accuracy of directions. We also take $\varepsilon = 1$ since it means that we have on average less than 1 false detection. But, as we shall see, really similar images have much smaller NFA and the choice of ε is not really important. Thus, in all experiments, we always set $\varepsilon = 1$, and we can therefore claim that the decision threshold is automatically derived. The asymptotic estimate given by Hoeffding's inequality [84] also shows the conditions to obtain a small number of false alarms. If M is fixed, then the NFA is a decreasing function of the proportion of coincidental direction $\frac{k}{M}$. If the proportion is assumed to be fixed, then the number of false alarms is exponentially (thus very fast) decreasing with respect to the number of samples M.

11.3.3 Choosing Sample Points

The computation of the number of false alarms is made under the assumption $\mathcal{H}_0$. Thus, we assume u and v to be independent randomly chosen images. But we also assume that the fact that $Du(x)$ and $Dv(x)$ are collinear is independent from the fact that $Du(y)$ and $Dv(y)$ are collinear at some other pixel y.

We must make this assumption realistic. There are two reasons for being careful. The first one is that if x and y are too close to each other, then $Du(x)$ and $Du(y)$ can be correlated. Thus, we must take pixels at a critical minimal distance to ensure independence of their gradients. Since the gradient is computed by a 2×2 finite difference scheme, sample points must be at least two pixels afar, in the original images. This has to be corrected by the scaling introduced by the normalization. If for instance, u has to be zoomed in by a factor 4 before comparison with v, then the minimal distance between two samples in the resulting image is $2 \times 4 = 8$.

The second issue for ensuring independence of observations is what we shall call the *alignment problem.* Images contain shapes, whose boundary are often piecewise smooth curves, that can locally be approximated by straight lines. The orientations of the gradient at two points on an alignment are the same, and cannot be assumed independent, even though the points may be far from each other. Hence sample points must be chosen sparse enough to minimize the probability that they fall on an edge which is common to both u and v. This obviously puts strong limits on the number of samples that may be drawn. In [32], calculations showed that the number of samples that are necessary to attain very low numbers of false alarms (yielding detection) is about 100, with reasonable noise conditions. When globally comparing two images, drawing about 100 samples yields very good results with a number of false detections which is conform to the prediction.

When dealing with image patches, the results may be less satisfactory. Indeed, patches result from a normalization procedure. Two parts of images basically containing an edge lead to two normalized patches that are very similar, and aligned with the edge direction. Moreover, if points are uniformly sampled over the sets of pixels with a large enough gradient, the gradient direction difference will be small with a high probability, since all the samples are very likely to belong to the (single) edge. Hence, the number of false alarms will be small. The patches are very similar indeed, and the detection is not a false alarm to this respect. However, it is not very informative. This problem is analog to the aperture problem in the estimation of optical flow: edges are locally indistinguishable.

This implies that the comparison of gradient directions is sensible only if the images are complex enough. An easy way to check this is to impose to draw samples in u and v such that the gradient direction in u (for instance) is close to uniform over the sample points. This way, samples are constrained not to lie on the same alignment and to restore some of the complexity of the images. These considerations lead us to the sampling algorithm described in the next section.

11.4 The Detection Algorithm

For two images I_1 and I_2 compute their interest points (x_i, s_i, φ_i), (y_j, t_j, ψ_j). With no loss of generality, assume that $t_j \geqslant s_i$ (else reverse the role of I_1 and I_2). Let F be the similarity $F(x) = zRx + b$ mapping (x_i, s_i, φ_i) on (y_j, t_j, ψ_j), where $z > 0$, R is a plane rotation, and $b \in \mathbb{R}^2$, are uniquely determined. Define two normalized images u and v. The image v is obtained by cropping I_2 in a patch $\mathcal{P}$ around y_j. The image u is simply $I_1 \circ F^{-1}$, also restricted to $\mathcal{P}$.

Let us quantize the direction of Dv in $(0, 2\pi)$ on $2N$ values. Sample points are then chosen by the following recursion. Let us consider $x_1 \in \mathbb{R}^2$ such that $|Du(x_1)| > \tau$ and $|Dv(x_1)| > \tau$. Assume that n points have been sampled. Then, a $n+1$th point x_{n+1} is chosen such that

- $|Du(x_{n+1})| > \tau$ and $|Dv(x_{n+1})| > \tau$.
- For each k, $0 \leqslant k < 2N$, the number of points of $\{x_1, ..., x_{n+1}\}$ such that the direction of Dv belongs to the interval $\left(\frac{k\pi}{N}, \frac{(k+1)\pi}{N}\right)$ is less than $1 + \frac{n}{2N}$.

This simply means that the repartition of the sample points is required to stay essentially uniform. This is not always possible. If the histogram of directions in v is unimodal, then no long sequence will fit the condition. As a consequence, if K points are tested, then they may lead to a set of sample points containing much less than K points. Let M be this number of points. The number of false alarms between u and v is then computed by using Def. 19. Hence, the final number of samples depends on the complexity of v. If v is essentially an edge, the sample sequence will be short and the number of false alarms will be large (since computed on very few points). On the contrary, for complex images (for instance pure texture), sample sequences will be long, thus leading to very small NFAs.

Let us remark that this peculiar sampling introduces a slight asymmetry in the algorithm between u and v and that the algorithm would be strictly contrast invariant only if $\tau = 0$. In practice, letting $\tau = 5$ removes gray level quantization effects which can entail false detections [49].

Numerically, experiments were led with $K = 1000$, and M is restricted to be less than 300. This means that points are drawn according to the two conditions above. The loop ends either when the total number of drawn points reaches $K = 1000$

or when the number of admissible points reaches 300. In practice the number of admissible points ranges from 50 to 300. When the patches present the alignment problem, M is obviously small. In this case, numbers of false alarms are large, as can be seen on the asymptotic development (11.4).

11.4.1 Experiments: Securing SIFT Detections

In the experiments, $\mathcal{P}$ is a square patch whose size is proportional to the key point scale (a factor 25 is used).

In all the experiments, we observed the following facts:

- Right matches can be very meaningful ($-\log_{10}(\mathrm{NFA}) \simeq 30$);
- on the contrary the wrong matches meaningfulness is close to 0. When it is not the case, there is actually a strong geometrical similarity between the compared patches.

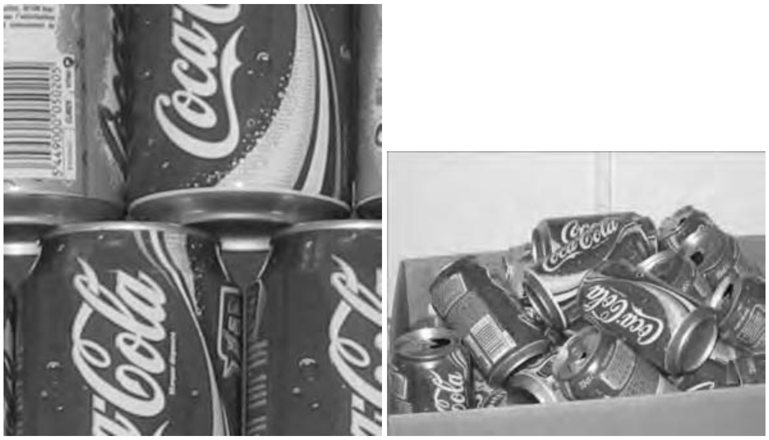

Fig. 11.8 Multiple matches experiment

In the experiment of Fig. 11.8 and 11.9, there are several partial occurrences of a logo. The usual procedure for matching SIFT descriptors (nearest neighbor, then comparison with second nearest neighbor) is inefficient in this case. In this experiment, 72 matches with an NFA less than 1 are detected. The best match has a meaningfulness (*i.e.* $-\log_{10}(\mathrm{NFA})$) equal to 25.4. Half the matches have NFA less than 10^{-3}, and are of course correct.

That there can be meaningful but wrong matches is not only inevitable but semantically sound. Fig. 11.12 shows such an example. Clearly the two matched SIFT patches do not correspond to the same motorcycle. All the same both present a wheel with the same orientation, taken from the same perspective and under the same light.

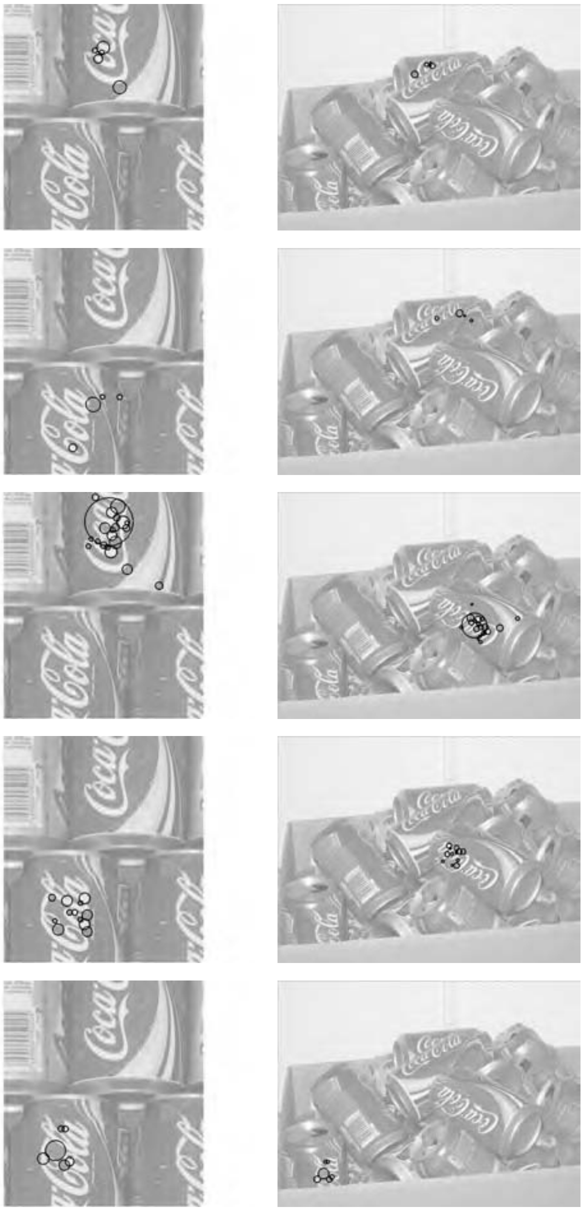

Fig. 11.9 Matching groups in the images of Fig. 11.8. Each circle represents a SIFT key point which got a match. The radius is twice the scale of the key point. The grouping procedure is the same as in in Chap. 8

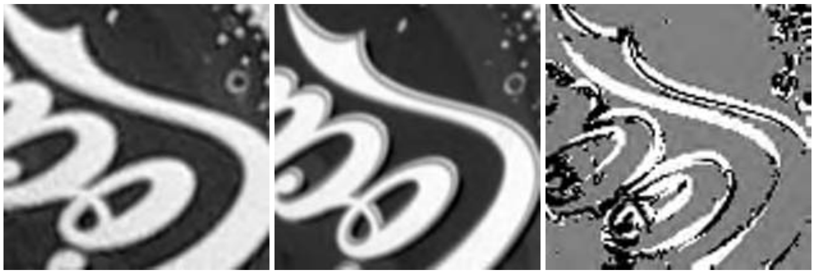

Fig. 11.10 Best SIFT match between the images of Fig. 11.8. Its meaningfulness is 25.4. On the left, the two registered patches. On the right figure, points where the gradient is not larger than 5 in both patches are displayed in gray. When the gradient is large in both images and when directions of the gradient coincide up to 40°, the pixel is plotted in white. It is plotted in black otherwise. Even though the images are different, the registration provided by the key points is accurate enough and yields a very meaningful match

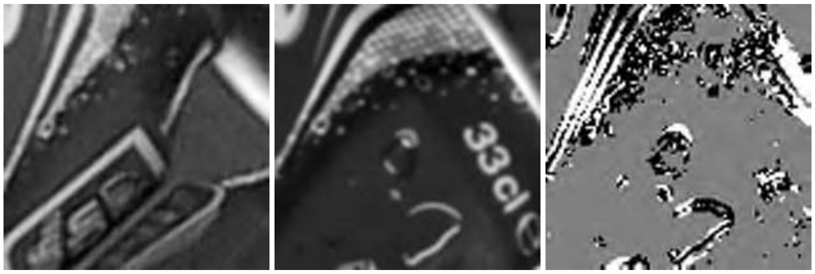

Fig. 11.11 A false and meaningful SIFT match between the two patches displayed on the left. The meaningfulness is 2.4. As can be seen, there are many white dots. Moreover, the orientation of the gradient at these locations is clearly not unimodal and the detection is not due to the presence of a single alignment

Under such circumstances, it is not only acceptable, but even desirable to make the detection. This is clearly illustrated in Fig. 11.8 which displays many physically different but identical Coca Cola cans. Fig. 11.10 shows an excellent match between two of them, taken from the same image. As we commented in Sect. 10.3.2.4 the SIFT original threshold procedure discards such matches. However, they are obviously of high interest. The detection and grouping of similar shapes in the same image is actually a fundamental gestalt [95]. However, some meaningful (but usually not very meaningful) casual matches can occur, which do not correspond to any recurrence of the same shape. Fig. 11.11 shows such a match between two SIFT patches, with 2.4 meaningfulness. Clearly the gradient orientations are very similar in both patches at many pixels, and the patches are complex enough to make such a coincidence an unlikely event. The overall explanation of such coincidence is given by the Gestalt Theory, which points out the recurrence of standard shapes in most images. Such shapes are called *gestalts* and include among others convex curves and bars with constant width [95]. The fact that two different patches show a similar arrangement of frequent shapes is therefore not unlikely. But this experiment proves that the *a contrario* model for patches similarity developed in this chapter is still a bit too primitive.

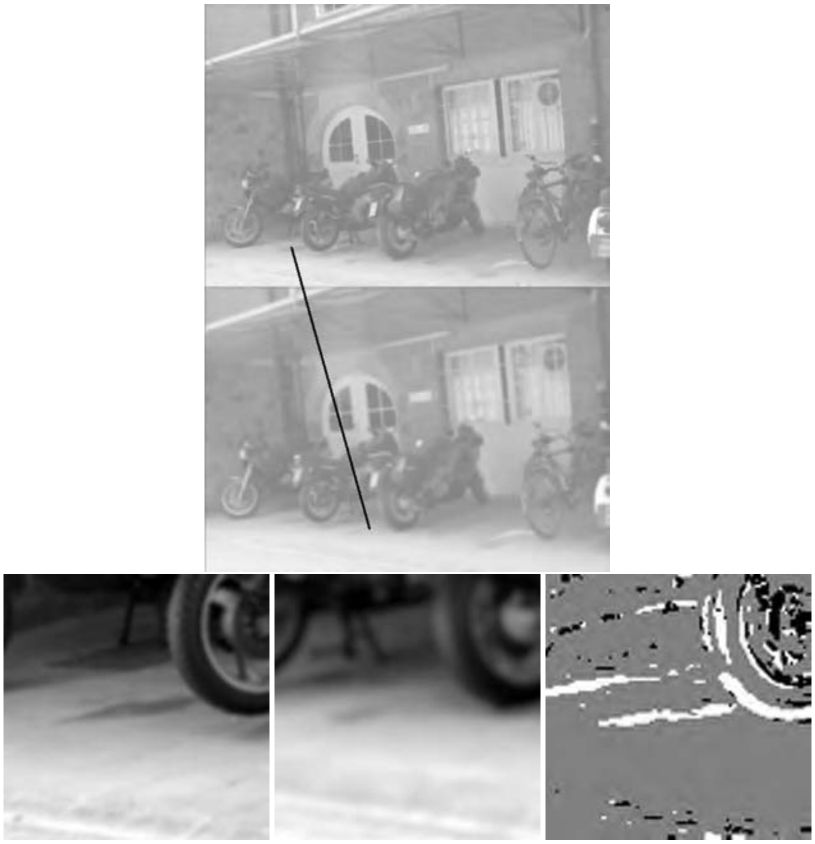

Fig. 11.12 Stroboscopic effect due to similarity within a class of objects. The second image has been obtained by a real defocus of the camera. On the bottom left, two patches that are very alike. The meaningfulness of their SIFT match is 2.2. The gradient directions difference is less than 40° at many locations (dots in white in the bottom patch). These patches do not come from exactly the same object, but retrieving them cannot be considered as a false alarm. Original images courtesy of the LEAR Team, INRIA

The next experiment shows matching and grouping results between two different views of the church of Valbonne (Fig. 11.13). Because of parallax, two different groups are detected, which is correct.

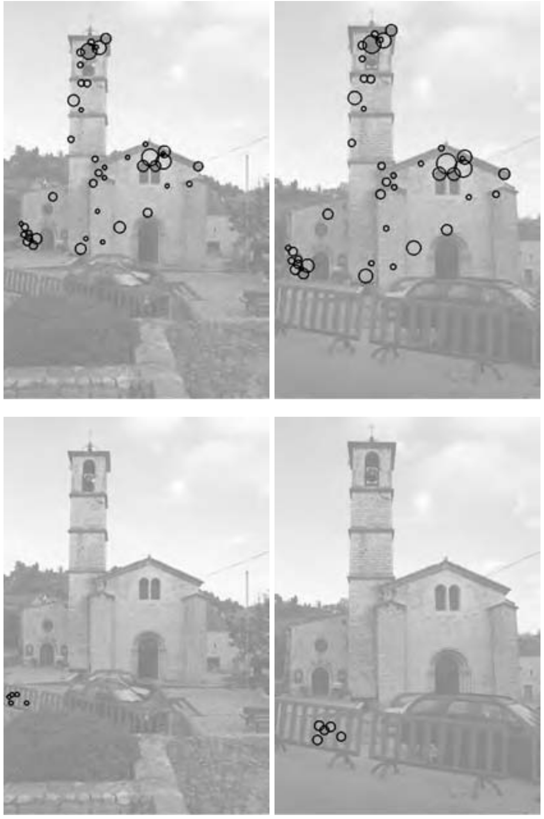

Fig. 11.13 Two different views of Valbonne church. There are two groups, corresponding to two different parts of the scene with different depths. In the first group, the largest meaningfulness is 30.8, in the second group, 2.7 because the resolution is lower in this part

11.5 Bibliographic Notes

Following [109, 139] the fact that some complex enough element recurs in two different images or even in the same can be taken as a basic definition of shape. Shapes simply are parts of an image which can be recognized in another one. There have been of course several attempts to measure the certainty of detected shape similarities. Schmid et al. [158, 157] use statistics of the distance between descriptors to recognize parts of objects of the same type, in a semi-supervised way. The grouping of SIFT matches for attaining certainty was pointed out in Lowe [113] and more recently in Cao et al. [33]. The grouping phase in [114, 157] is used *a posteriori*, both to eliminate possible casual matches, and to reinforce the detection of right matches. The novelty in [33] is the accurate computation of the number of false alarms assigned to a group. An alternative to the method described in Sect. 11.3 and 11.4 for meaningful SIFT matching has been recently proposed in [150].