Master MVA, IPOL #4 – Lecture Notes

A contrario matching of SIFT-like descriptors

Julien Rabin, with Julie Delon, Yann Gousseau and Jean-Michel Morel
julien.rabin@cmla.ens-cachan.fr

October 21, 2011

∴

In this course, we take interest in the matching of local feature between two images. After the description of the major issues related to this task in the introduction, the first section gives an overview of the generic a contrario *methodology introduced by A. Desolneux, L. Moisan and J.-M. Morel in [DMM00]. This method makes it possible to design robust decision criteria that prevent false detections. In the next section is proposed a brief description of the SIFT [Low99] method (Scale Invariant Feature Transform), which is nowadays at the heart of many algorithms in the computer vision literature. In this approach, images are coded from a few interest points* (i.e. *with loss), for which the neighborhood is described from similarity-invariant feature. These interest points and local descriptors are obtained in practice from a multitude of thresholds, the optimality of which is the key of the success of the SIFT method. In the case of orientation assignment of interest points, an alternative based on the* a contrario *histogram mode selection [DMM03] is proposed. The last section is devoted to the "safe" matching of SIFT-like features between images, in which several matching criteria from [Low04, MSC+06, RDG09] are discussed.*

∵

Contents

1 Introduction

Compared to global approaches, local representation of images has proved to be more efficient in many computer vision areas. Indeed, nowadays most algorithms are based on local features for various applications, such as image registr

ation, image stitching, object detection, pose estimation, object recognition, 3-D reconstruction, image classification, etc. However, dealing with local features requires some specific processing tasks which are discussed in the next paragraphs, particularly regarding the robust matching of those features.

Processing with local representation of images Let consider the object recognition task illustrated in Figure 1: given two images, it consists in **deciding** if a common object is present or not, and if so, in **estimating** the pose of this object in each image. In the most general case (without any prior knowledge on the object), the accomplishment of such task requires numerous invariance and robustness to various phenomena, such as:

- ▷ partial occlusion;
- ▷ change of point of view (perspective effects, self occlusion, geometric distortions due to camera);
- ▷ change of illumination source (white balance, shadows, contrast change, specularity);
- ▷ elastic and articulated deformations;
- ▷ multiple occurrences;
- ▷ self-similarity;
- ▷ change of context (background/foreground).

Early approaches proposed in the literature were based on template matching (*e.g.* global fitting using correlation), which does not fulfill many of those requirements. In contrast, local representation of images has been shown to yield naturally more invariance and robustness (among the most famous approaches see *e.g.* the Shape Context [BMP02], the MSER [MCUP02], the *a contrario* shape recognition framework introduced in [MSC+06]) and the Scale Invariant Feature Transform [Low04]). In Section 3 is given an overview of the SIFT method which is now extensively used in Computer Vision to extract invariant features.

Nevertheless, dealing with local features also requires extra processing. In the following is a sketch of the processing chain used for object recognition (see Figure 2 for an overview):

1. Interest point or region detection *Goal:* Find and select a small but sufficient number of salient structures (*e.g.* edges, corners, T-junctions, blobs, contrasted level line, etc) in image. *Examples:* Harris corner detector, Canny edge detector, normalized Laplacian, SIFT key-points (See Section 2), MSER, *etc.*

(a) Where is this building ...

(b) ... in this picture ?

(c) Is there any book in common ...

(d) ... in these two book shelves ?

Figure 1: Illustration of the challenging object recognition task.

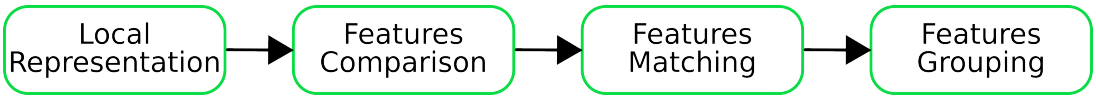


Figure 2: Processing with local representation of images.

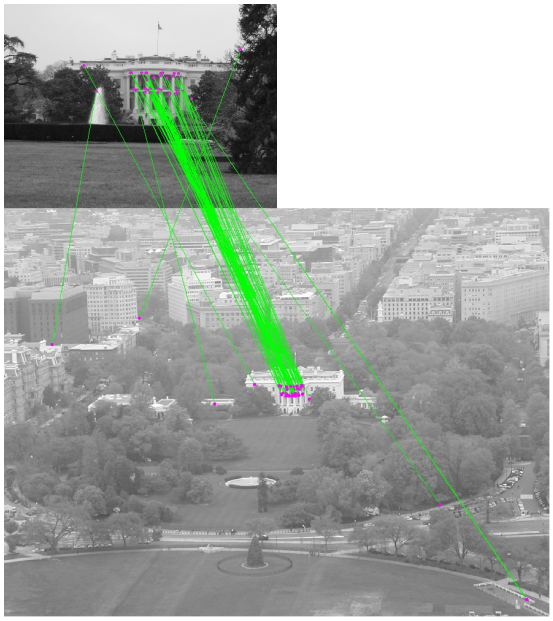
2. Feature description *Goal:* Compact and Invariant (robust) representation of local information. *Examples:* Color histograms or histograms, Shape context, SIFT (See Section 3).

3. Feature comparison *Goal:* Define a similarity measure to reject spurious correspondences, invariant (robust) to geometric distortions, noise, *etc.* *Examples:* Euclidean metric, χ^2-distance, Mahalanobis distance, Earth Mover's Distance, *etc.*

4. Feature matching *Goal:* Select relevant correspondences between two images. *Examples:* Thresholding techniques on similarity measures (See Section 3).

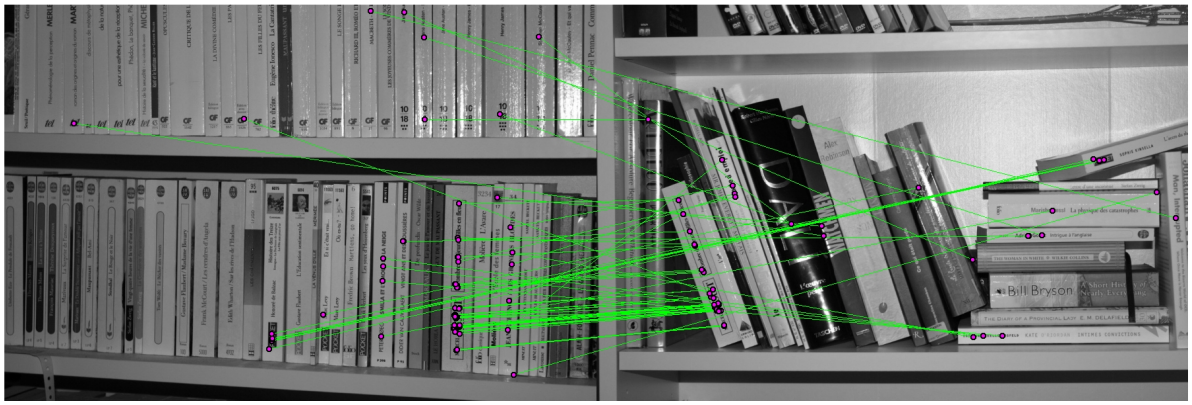
5. Grouping matches *Goal:* Group matches that are consistent under geometric constraints (hypotheses merging). *Examples:* RANSAC algorithm, Hough transform.

Each of these steps relies on decision criteria, which aim at reduce computations in further steps by rejecting false hypotheses while preserving the likely ones. Some examples of this processing chain are given in Figure 3. In this course, we will focus on the matching step.



(a) Local correspondences after the matching step.

(b) Pose estimation from correspondences.



(c) Local correspondences after the matching step.

(d) Pose estimation from correspondences.

Figure 3: **Object recognition from matches between local features.**

Why the feature matching step is critical? Without entering into details about grouping methods (step 5 in the aforementioned processing chain), one should know that such algorithms explore hypothesis based on interest point correspondences to merge consistent clues on the object pose (combinatorial problem). Therefore, it is very sensitive to the number of selected correspondences between the two images but also to the presence of irrelevant data ("outliers"), which corresponds to false matches between the two images. Let now define more precisely what are those false matches that the matching step should be able to discard.

When deciding if a match between two features should be validated or rejected, four scenarios can happen depending on if the match is correct or not. This is a classical hypothesis testing classification given by table 1.

Table 1: Taxonomy of hypothesis testing

Ground truth \ Test	Positive	Negative
Correct	True-Positive (tp)	False-Negative (fn) [Type II error]
Incorrect	False-Positive (fp) [Type I error]	True-Negative (tn)

We can see that two different types of matches are false: false positive and false negative (respectively type I and type II error). Depending on the application, one may especially want to avoid the former or the later (for instance, in vaccine testing, a good test should avoid False-Negative). In our case of study, the only type of false match we have to take into account is the False-Positive. Our goal in this course is to show how to select correct matches between two images (*i.e.* maximize the number of true-positive) while discarding incorrect matches (*i.e.* minimize type I error). In the section 4, several matching criteria devoted to this task are presented. However, two ingredients are first required: a **robust and invariant feature selection** method to describe local neighborhood of interest points (Section 3) and a the a contrario methodology (Section 2).

2 A short overview of the *a contrario* methodology

Most computer vision algorithms rely on decision criteria to decide if an hypothesis valid or not, which often make use of the Bayesian inference or statistical tests. Such methods generally require some prior knowledge on the specific object one want to detect.

In a study of perceptual grouping in the Gestalt theory, Desolneux, Moisan and Morel [DMM00, DMM08] introduced a new and generic decision method which exclusively relies on the rejection of a background model, *i.e.* without any prior on the object of interest itself. This "*a contrario*" framework – already seen in *IPOL#1 : "LSD: a Line Segment Detector"*– is today at the heart of numerous algorithms in the literature, and particularly in this course for local orientation selection and feature matching. In what follows, we give a brief overview of this methodology which estimate automatically adaptive decision thresholds by controlling the number of false alarms (type I error).

The Helmholtz principle The *a contrario* detection philosophy is based on the Helmholtz principle, which can be stated as follows:

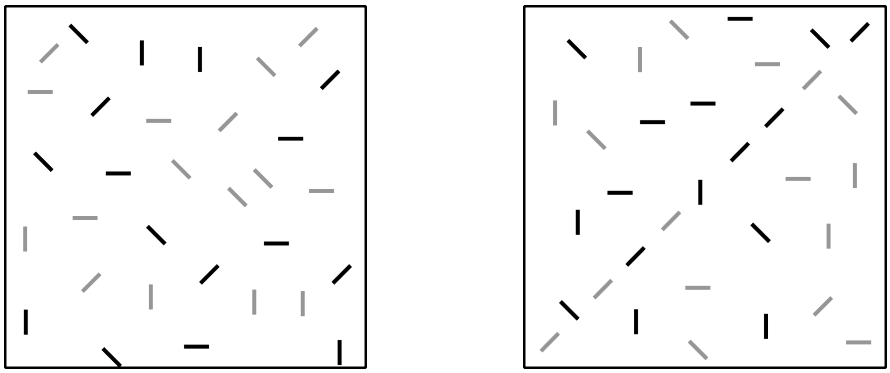
≪no structure should be detected in a noise image.≫

It comes from the assumption in Gestalt theory that we not perceive a structure when it is likely to happen by chance.

To illustrate this principle, an example is displayed on figure 4, were two different configurations of segments are shown. In the first one (figure 4(a)), the position and the orientation of segments are randomly –say uniformly and independently– drawn, so that no structure is perceptible. On the contrary, when considering the second configuration in figure 4(b), we can not help but group some segments which

are aligned. Indeed, such structure is very unlikely to happen under the assumption that the segments have been drawn **independently**.

The underlying idea behind the *a contrario* theory is to formalize this grouping principle to define adaptive decision criterion.



(a) (b)

Figure 4: Illustration of perceptual grouping (Helmholtz principle).

Formalization The background model rejection in the *a contrario* approach rely on a **null hypothesis** testing, denoted $\mathcal{H}_0$, that states that the observed features (the segments' orientation and position in our previous example) are **mutually independent**.

Now, given a group of features G, a similarity measure S has to be defined to evaluate the adequacy of this group with the type of structure we want to detect (for instance, the consistency of orientation to detect alignment).

Using this similarity measure, we can estimate the "p-value" corresponding to $\mathcal{H}_0$, that is the probability of observing a random group of features $\mathbf{G}$ following the null hypothesis with a better similarity measure than G:

$$\mathbb{P}_{\mathcal{H}_0}(S(\mathbf{G}) \geq S(G)) \,. \tag{1}$$

Clearly, the higher this probability, the higher the chance that the observed structure G can happen by chance; hence, that it should be rejected.

A classical approach to decide if the considered structure G should be validated is to threshold this probability. Groups of features that are validated while following the null hypothesis are called "false alarms" (*i.e.* false positive). The remaining question, which is at the heart of the *a contrario* methodology, is now: How to threshold such a probability to avoid false alarms?

False alarm rejection It must be noticed at this point that such a decision criterion ("Does structure G have to be validated ?") is meant to be used several times (multiple testing) in practice since many structures may be of interest in a single image (*e.g.* shape recognition). Let $\{G_i\}_i$ be the set of tested structures and $\mathcal{N} = |\{G_i\}_i|$ be the total number of tests. It is obvious that the **number of false alarms** (type I error) depends on $\mathcal{N}$

The *a contrario* methodology consists in taking into account this number of tests $\mathcal{N}$ to design adaptive thresholds to the detection of structure of interest.

To do so, it is usual to define the a quantity called "NFA", standing for Number of False Alarms, which is defined as follows (general formalization):

$$\mathrm{NFA}(G_i, S(G_i)) := \mathcal{N} \times \mathbb{P}_{\mathcal{H}_0}(S(\mathbf{G}) \geq S(G_i)) \,. \tag{2}$$

Indeed, this quantity represents the expected number of false alarms (*i.e.* an "E-value") for $\mathcal{N}$ tests when validating every random groups of features following the background model with quality measure greater than $S(G_i)$.

The smaller the NFA quantity, the more significant is the tested group. A group for which the NFA is smaller than ε is called ε-meaningful group. The fundamental interest of the *a contrario* approach lies in the fact that it ensures the following property:

> *≪The expected number of ε-meaningful groups following the null hypothesis is smaller than ε.≫*

This means that by selecting only groups G_i that are ε-meaningful, we control the expected number of false alarms. In practice, the "meaningfulness threshold" ε is generally set equal to one, meaning that at least one false alarm should arise when testing all the structures in the image.

Automatic thresholds We have seen that only groups that are ε-meaningful are validated. This boils down to threshold the probability under the null hypothesis with the constant $\frac{\varepsilon}{\mathcal{N}}$. This could be seen as a Bonferroni correction to control the type I error (*i.e.* the expected number of false alarms). The thresholding of the probability term is in fact equivalent to define adaptive thresholds on the similarity measure S:

$$t_i(\varepsilon) = \min\{t, \mathrm{NFA}(G_i, t) \leq \varepsilon\} \tag{3}$$

Then, a group is ε-meaningful *if* $S(G_i) > t_i(\varepsilon)$.

3 SIFT method

The SIFT method has been already described in details in *IPOL#4 : "Is the 'Scale Invariant Feature Transform' (SIFT) really Scale Invariant?"*. In this section is given a short overview of this method, for which some of its aspects are developed in more details.

Recall that the SIFT approach consists, in the first part, in detecting point in the scale space of the image of interest (§ 3.1), and then filtering those interest points (§ 3.2), and assigning them an orientation (§ 3.3). The second part is then devoted to the compact description of these points (§ 3.4), based on local histograms of the gradient orientation.

3.1 Salient point detection in the scale-space

Scale-space analysis Let $I : (x, y) \in \Omega \subset \mathbb{R}^2 \mapsto I(x, y) \in \mathbb{R}$ be a scalar image from which we want to extract some interest points in location, scale and orientation. The (unique) linear scale-space representation of I at point $(x, y, \sigma) \in \mathbb{R}^2 \times \mathbb{R}^+$ is given by the computation of the convolution with a centered Gaussian kernel with variance σ^2:

$$I_\sigma(x, y) := G_\sigma * I(x, y) = \frac{1}{2\pi\sigma^2} \iint_{(u,v)\in\Omega} I(u, v)\, e^{-\frac{(x-u)^2+(y-v)^2}{2\sigma^2}}\, dudv\ . \tag{4}$$

which corresponds to the solution to the heat diffusion PDE $\partial_t I_t(x, y) = \sigma \Delta I_t(x, y)$ at time $t = \sigma$, where at $t = 0$, $I_t := I$.

Normalized Laplacian In order to detect from this tri-dimensional representation of I some salient features in a robust and repeatable fashion, T. Lindeberg [Lin94] proposed to use the normalized Laplacian:

$$\Delta_\sigma I(x, y) = \sigma^2 \cdot \Delta I_\sigma(x, y) := \left\{ \left(\frac{x^2 + y^2}{\sigma^2} - 2 \right) G_\sigma(x, y) \right\} * I(x, y)\ . \tag{5}$$

The local extrema of such an operator corresponds to edges, junctions or "blobs". For instance,

- for a disk $D(C,\rho)$, the extremum is located in the center C of the disk, at scale $\sigma = \rho/\sqrt{2}$;
- for a Gaussian kernel $\mathcal{N}(\mu, s)$, the extremum is located in μ, at scale $\sigma = s$;
- for a corner (intersection of two edges), the extrema of the normalized Laplacian operator (in space, but not in scale) lay on the bisector. Theoretically, there is no extremum in scale because the Laplacian response is decreasing in scale. Nevertheless, because of the noise or the discretization in scale, several local extrema may be found in practice.

SIFT approximations The computation of the local extrema of the Laplacian on the scale space representation of the image is time consuming. In SIFT [Low99, Low04], D. Lowe proposed numerous approximations to speed-up the process. Some other have been later proposed (see *e.g.* [BTG06]).

First of all, the scale space representation is discretized in scale with very few samples $\{I_{\sigma_k}\}_k$ to avoid spurious computation of convolutions. In [Low99], the scale of index k is defined geometrically: $\sigma_k = \sigma_0 \cdot r^k$, where $\sigma_0 = 0.8$ is the initial scale and $r = 2^{\frac{1}{3}}$ (three samples for an octave). Each time a new octave is obtained, the image I_{σ_k} is down-sampled by a factor of two (to obtain a so-called pyramid).

The normalized Laplacian operator is then approximated by *difference of Gaussian* (DoG). Indeed, at scale of index k, we can write the difference of two successive images as:

$$I_{\sigma_{k+1}}(x,y) - I_{\sigma_k}(x,y) = \{G_{\sigma_{k+1}} - G_{\sigma_k}\} * I(x,y) \approx \sigma_k (r-1) \left.\frac{\partial G_\sigma(x,y)}{\partial \sigma}\right|_{\sigma_k} * I(x,y) \ . \tag{6}$$

Now, from the heat diffusion equation $\frac{\partial G_\sigma(x,y)}{\partial \sigma} = \sigma \Delta G_\sigma(x,y)$, so we get:

$$I_{\sigma_{k+1}}(x,y) - I_{\sigma_k}(x,y) \approx (r-1)\Delta_{\sigma_k} I_{\sigma_k}(x,y) \ . \tag{7}$$

Local extrema of the DoG operator are then found by checking for every pixel the values of its 26 neighbors in the scale space.

Observe that, in order to compensate those approximations, a second-order interpolation of the extremum coordinate $m_i : (x_i, y_i, \sigma_i)^T$ is then performed in [Low04]:

$$\hat{m}_i = (x_i, y_i, \sigma_i)^T + (\nabla^2 I_{\sigma_i}(x_i, y_i))^{-1} \nabla I_{\sigma_i}(x_i, y_i) \ , \tag{8}$$

where ∇ and ∇^2 respectively stands for first and second order derivative operator in $\mathbb{R}^3$.

3.2 Interest-point filtering

Detection of local extrema is very sensitive to noise, yielding some irrelevant detections on edges. To avoid the detection of such uninformative structure (location of point along edges is not reliable), a decision criterion based on geometrical purpose is driven.

In [Low04], this criterion is based on the comparison of the eigenvalues of the 2-D Hessian matrix (related to the curvature of the surface) computed at the scale of the point of interest. This Hessian matrix writes as follows:

$$H_\sigma(x,y) = \nabla^2 I_\sigma(x,y) = \begin{bmatrix} \frac{\partial^2 I_\sigma(x,y)}{\partial x^2} & \frac{\partial^2 I_\sigma(x,y)}{\partial y \partial x} \\ \frac{\partial^2 I_\sigma(x,y)}{\partial x \partial y} & \frac{\partial^2 I_\sigma(x,y)}{\partial y^2} \end{bmatrix} \ .$$

To avoid eigenvalues computation, Lowe propose the following test inspired from the Harris corner detector [HS88]:

$$Tr(H_\sigma(x,y))^2 > t \cdot Det(H_\sigma(x,y)) \ , \tag{9}$$

where Det and Tr respectively stands for the determinant and the trace of the matrix H_σ. The threshold t (fixed to 12.1) ensures that the ratio between the eigenvalues do not exceed 10.

Note that a similar technique has been proposed previously by Mikolajczyk and Schmid in [MS01], generalizing the Harris corner detector [HS88] for scale-space representation. The Harris detector, based on previous work of Moravec [Mor80], consists in the analysis of the eigenvalues of the structure tensor $T(x, y)$, an average correlation matrix of first-order derivative defined as follows:

$$T(x,y) = G_s * \begin{bmatrix} (\frac{\partial I}{\partial x})^2 & \frac{\partial I}{\partial x}\frac{\partial I}{\partial y} \\ \frac{\partial I}{\partial x}\frac{\partial I}{\partial y} & (\frac{\partial I}{\partial y})^2 \end{bmatrix} , \tag{10}$$

where G_s is a Gaussian kernel with standard deviation s. It is well known that the eigenvectors of the structure tensor are strongly related to the main directions of the gradient (T could be seen as a moment of inertia matrix). The original *cornerness* measure of Harris and Stephen is

$$\mathcal{C}(x,y) = Det\left(T(x,y)\right) - k \cdot Tr(T(x,y))^2 > t \ , \tag{11}$$

where t and k are two parameters. In [MS01], this criterion is adapted to the scale space analysis by considering the following matrix $T_\sigma(x, y)$, define by taking into account the scale parameter σ:

$$T_\sigma(x,y) = \sigma^2 \cdot G_s * \begin{bmatrix} (\frac{\partial I_\sigma}{\partial x})^2 & \frac{\partial I_\sigma}{\partial x}\frac{\partial I_\sigma}{\partial y} \\ \frac{\partial I_\sigma}{\partial x}\frac{\partial I_\sigma}{\partial y} & (\frac{\partial I_\sigma}{\partial y})^2 \end{bmatrix} . \tag{12}$$

Observe the presence of the normalizing term σ^2, as for the Laplacian operator. The scale-space measure of cornerness is now:

$$\mathcal{C}_\sigma(x,y) = Det\left(T_\sigma(x,y)\right) - k \cdot Tr(T_\sigma(x,y))^2 > t \ . \tag{13}$$

where $k = 4\,10^{-2}$ and $t = 2\,10^3$.

3.3 Interest-point orientation assignment

The previous detection steps provides interest points $\{m_i : (x_i, y_i, \sigma_i)\}$ in scale-space with similarity invariant operator (linear scale space and normalized Laplacian operator). A last step is now required to compute descriptor that are fully invariant to similarity: indeed, in addition to the space and scale localization, a orientation assignment is also required to achieve complete rotation invariance.

3.3.1 Original SIFT approach

The key idea of D. Lowe in [Low99] is to estimate the main orientation of a given structure by looking at the statistical distribution of the gradient orientation in the vicinity of the interest point. To do so, an histogram is first built in which peaks are then detected.

Notation Let consider a circular histogram H defined from M samples $\{\theta_1, \ldots, \theta_M\}$ (orientations) quantized uniformly on L values for the interval $[-\pi, \pi[$:

$$\forall\, i \in \{1, \ldots, L\}\, H[i] = \left| \left\{ \theta_m \in \left[\frac{2\pi}{L}(i-1) - \pi, \frac{2\pi}{L} i - \pi \right[, m = 1, \ldots, M \right\} \right| . \tag{14}$$

We will denote by $\mathcal{I}$ the set of indexes $\{1, 2, \ldots, L\}$ and by $\mathcal{L}$ the set of quantized values $\{\frac{\pi}{L} i - \pi\}_{i \in \mathcal{I}}$ corresponding to the center of the i-th bin. Recall that the histogram being circular, it means that the bins with respective index 1 and L are adjacent. A discrete circular interval $[a, b]$ is hence the set of index values in $\mathcal{I}$ starting from bin a towards bin b, *i.e.* $[a, b] := [a, L] \cup [1, b]$ when $b < a$. For commodity, we will refer from now to bin "$i-1$" as the bin with index $i-1$ if $i \leq 2$, and as the bin with index L if $i = 1$. The length of an interval is defined as its number of bins.

Peack detection In [Low04], the histogram of the local gradient orientation is quantized on $L = 36$ accumulators and is built from gradient samples extracted in the neighborhood of the interest point $mi : (x_i, y_i, \sigma_i)$. Each entry in the histogram is weighted by the gradient norm and by a Gaussian kernel $G_{3\sigma_i}(. - x_i, . - y_i)$ with variance $9\sigma_i^2$ and mean (x_i, y_i).

The peaks in the histogram are then simply defined as local maxima, where local maxima lower than 80% to the global maximum are pruned, that is

$$P = \left\{ i \in \mathcal{I}\,,\ H[i-1] \leq H[i] \text{ and } H[i] \geq H[i+1] \text{ and } H[i] \geq 0.8 \max_{j \in I} H[j] \right\} \tag{15}$$

The location of each peak is then refined as the vertex α of the parabola fitted to the 3 histogram values around the peak, that is

$$\forall i \in P\,,\ \alpha_i = -\frac{\pi\, b}{L\, a}\,[2\pi] \text{ with } \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} (i-1)^2 & i-1 & 1 \\ i^2 & i & 1 \\ (i+1)^2 & i+1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} H[i-1] \\ H[i] \\ H[i+1] \end{bmatrix}, \tag{16}$$

where $H[0] = H[L]$ and $H[L+1] = H[1]$.

Finally, the orientation α_i is assign to the interest point m_i so that $m_i : (x_i, y_i, \sigma_i, \alpha_i)$. Note that the interest point are duplicated so that we obtain as many interest points as detected peaks.

3.3.2 *A contrario* detection of orientation

It is obvious from the previous method that the definition of orientation is extremely sensitive to the quantization level L. There is a trade-off between the precision of the location (which increases with L) and the detection of the local maxima (which is only robust for small values of L). Ideally, the detection of dominant mode in the histogram of gradient orientation should not dependent on the quantization and without any threshold setting.

A good alternative is to use the *a contrario* algorithm proposed by Desolneux *et al.* [Des00, DMM03] to detect modes and gaps in histograms. We will now introduced this approach based on the "*a contrario* recipe" described in Section 2.

Null hypothesis $\mathcal{H}_0$ Recall that the histogram H is built from samples $\theta_m \in [-\pi, \pi[$.

Definition 1 (Null Hypothesis (background model)). *the samples $\{\theta_m\}_{\{m=1,\dots,M\}}$ are mutually independent random variables, identically and uniformly distributed in $[-\pi, \pi($.*

Observe that this null hypotheses corresponds to a background model of white noise image, since the orientation of the discrete gradient is uniformly distributed [DLMM02].

Then, the probability that one sample θ_m fall into $[a, b]$ under the null hypothesis $\mathcal{H}_0$ (*i.e.* assuming that the sample has been drawn using the background model) is hence:

$$p(a,b) = \frac{|[a,b]|}{L} = \frac{1}{L} \begin{cases} b-a+1 & \text{if } b \geq a \\ a-b+1+L & \text{otherwise} \end{cases}. \tag{17}$$

Consider now several samples: we want to be able to decide if a given interval of the histogram is meaningful, *i.e.* with a number of samples being very unlikely under the null hypothesis. Let denote by $k(a, b)$ the number of samples $\{\theta_m\}_m$ which belong to the discrete circular interval $[a, b]$. Then, the probability that at least $k(a, b)$ samples of points among M fall into interval $[a, b]$ under the null hypothesis is given by the tail of the binomial distribution $\mathcal{B}(M, k(a, b), p(a, b))$ defined as:

$$\mathbb{P}_{\mathcal{H}_0}(\mathbf{k} \geq k(a,b)) = \mathcal{B}(M, k(a,b), p(a,b)) = \sum_{i=k(a,b)}^{M} \binom{M}{i} p(a,b)^i (1-p(a,b))^{M-i}\,. \tag{18}$$

The lower this probability, the more unlikely the interval $[a,b]$ is under the background model. Deciding if this interval can be selected boils down to the thresholding of this probability.

Meaningful interval We have previously seen that such a threshold must take into account the number of tests in order to avoid spurious false alarms (Bonferroni correction). In our setting, this number of tests is the number of interval that will be considered, that is the total number of distinctive circular intervals

$$\mathcal{N} := |\{(a,b)\,,\ a \in \mathcal{I}, b \in \mathcal{I} \setminus \{a-1\}\} \cup \{[1,L]\}| = \sum_{i \in \mathcal{I}, j \in \mathcal{I} \setminus \{i-1\}} 1 + 1 = L(L-1) + 1\ . \tag{19}$$

Then, we define the following quality measure, called NFA, for a given interval $[a,b]$ with $k(a,b)$ samples as:

$$\mathrm{NFA}([a,b]) = \mathcal{N}\,\mathcal{B}(M, k(a,b), p(a,b))\ . \tag{20}$$

Definition 2 (Meaningful interval). *An interval I is a $\varepsilon - meaningful$ interval when*

$$NFA(I) = \mathcal{N}\,\mathcal{B}(M, k(a,b), p(a,b)) \leq \varepsilon\ .$$

Recall that a false alarm is a validated interval that follows the null-hypothesis.

Proposition 1. *Let H be an histogram built from M random samples following the null hypothesis $\mathcal{H}_0$. The expected number of false alarms when validating $\varepsilon - meaningful$ intervals is smaller than ε.*

Computation time consideration: Hoeffding's inequality The estimation of the NFA for each possible interval involves the computing of the binomial tail, which can be time consuming. Of course, since binomial parameters take discrete values (depending on constants L and M), the probability under the null hypothesis can be precomputed into a look-up table. Nevertheless, it remains fastidious for large values of M and L. Another interesting alternative is to use the Hoeffding's inequality to define a sufficient condition of meaningfulness.

Let first introduce the relative entropy definition. Let $r(a,b) = \frac{k(a,b)}{M}$ be the posterior probability for a sample to belongs to the interval $[a,b]$, and $p(a,b)$ the prior (uniform).

Definition 3 (Relative entropy of an interval). *The relative entropy of probabilities r and p is defined as:*

$$\forall\,(r,p) \in [0,1]\ , H(r,p) = r \ln \frac{r}{p} + (1-r) \ln \frac{1-r}{1-p}\ . \tag{21}$$

This quantity is the non-symetric Kullback-Leibler distance between two Bernoulli distributions with parameters r and p respectively.

Lemma 1 (Hoeffding's inequality). *If $\frac{k(a,b)}{M} \geq p(a,b)$, then*

$$\mathcal{B}(M, k(a,b), p(a,b)) \leq e^{-M \cdot H(\frac{k(a,b)}{M}, p(a,b))}\ . \tag{22}$$

Proposition 2 (Sufficient condition of meaningfulness). *Sufficient conditions for a given interval I=[a,b] to be ε-meaningful is:*

$$r(a,b) = \frac{k(a,b)}{M} \quad > \quad p(a,b) \tag{23}$$

$$H(r(a,b), p(a,b)) \quad > \quad \frac{1}{M} \ln \frac{\mathcal{N}}{\varepsilon} \tag{24}$$

This simple test is faster to evaluate in practice than the original one in Proposition (1).

Automatic threshold The thresholding of the quantity NFA with a fixed parameter ε makes it possible to define automatically thresholds on the number k of samples required for a discrete interval to be validated. Such a threshold k_ε, which only depends on the length l (*i.e.* the number of bins) of the interval, is defined as

Definition 4 (Meaningful sample threshold)**.**

$$k_\varepsilon(l) = arg\min_k \{k\,,\ \mathcal{B}(M,k,\frac{l}{L}) \leq \frac{\varepsilon}{\mathcal{N}}\}\ . \tag{25}$$

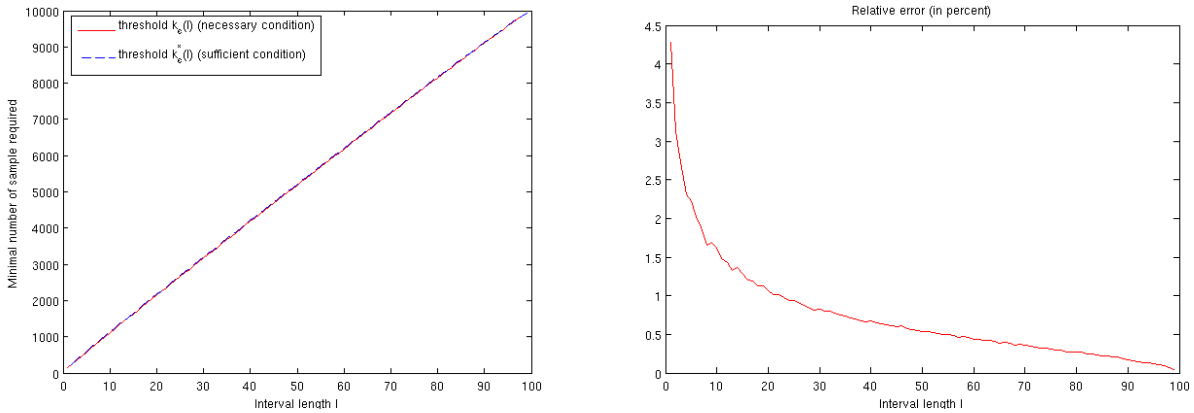
Then, a interval of length l with k samples is ε-meaningful if $k \geq k_\varepsilon(l)$

Using the sufficient condition on meaningfulness (Proposition (2)), we can define equivalently a sufficient threshold $\hat{k}_\varepsilon(l)$ on the number of samples such that the considered interval is ε-meaningful.

Definition 5 (Sufficient sample threshold)**.**

$$\hat{k}_\varepsilon(l) = arg\min_k \{k\,, k > M\frac{l}{L}\ \ and\ H(\frac{k}{M},p) > \frac{1}{M}\ln\frac{\mathcal{N}}{\varepsilon}\}\ . \tag{26}$$

A comparison of these two thresholds is given in Figure 5, with the following experimental setting: $L = 10^2$, $M = 10^4$. Observe that the sufficient sample threshold $\hat{k}_\varepsilon(l)$ do better approximate the meaningful sample threshold $k_\varepsilon(l)$ for large values of l.



(a) Thresholds for $\varepsilon = 1$

(b) Relative error

Figure 5: Comparison of thresholds.

Meaningful gaps and modes The previous definitions enables us to define a meaningful interval as an interval containing sufficiently more samples than it is expected from the background model. Nevertheless, it is insufficient in practice to select main modes of an histogram, since a meaningful interval can contains some "gaps". To take this into account, we first define meaningful gaps and then meaningful modes.

The definition of a meaningful gap is straight-forward: it is an interval that contains sufficiently *few* samples than expected from the background model. More precisely, the probability that at most $k(a,b)$ samples of points among M fall into interval $[a,b]$ under the null hypothesis is given by $\mathbb{P}_{\mathcal{H}_0}(\mathbf{k} \leq k(a,b)) = \sum_{i=0}^{k(a,b)} \binom{M}{i} p(a,b)^i (1-p(a,b))^{M-i}$, which can be expressed from again from the tail of the binomial distribution

$$\mathbb{P}_{\mathcal{H}_0}(\mathbf{k} \leq k(a,b)) = \mathcal{B}(M, M-k(a,b), 1-p(a,b))\ . \tag{27}$$

Then, we define

Definition 6 (Meaningful gap). *An interval $[a,b]$ is a $\varepsilon-$meaningful gap when*

$$NFA'(a,b) = \mathcal{N}\mathcal{B}(M, M-k(a,b), 1-p(a,b)) \leq \varepsilon\ . \tag{28}$$

Definition 7 (Meaningful mode). *An interval I is a ε-meaningful mode if it is a ε-meaningful interval that does not contain any ε-meaningful gap.*

Maximality From the previous paragraph we can guess that several meaningful modes with overlaps could be detected (redundant detections). Therefore we need a maximality criterion to select only non-overlapping modes.

Definition 8 (Maximal meaningful mode). *An interval I is a maximal-ε-meaningful mode if it is a ε-meaningful mode and if for all ε-meaningful mode $J \subset I$, $NFA(J) \geq NFA(I)$ and for all ε-meaningful mode $J \not\supseteq I$, $NFA(J) > NFA(I)$.*

Algorithm The *a contrario* algorithm for the detection of dominant mode in circular histogram is described in Table 2.

Table 2: A Contrario mode detection.

Algorithm 2 Automatic mode selection.
Input: Histogram H with M samples and L bins.
parameter $\varepsilon = 1$.
1) Find ε-meaningful intervals (definition 2);
2) Find ε-meaningful gaps (definition 6);
3) Find ε-meaningful modes (definition 7);
4) Find maximal ε-meaningful modes (definition 8).
Output: List of intervals.

Interest point filtering and orientation assignment Now that we have defined an algorithm to detect meaningful mode in histograms, the definition of orientation is straight-forward. For a given 1-meaningful mode $[a,b]$, the corresponding orientation $\alpha_{[a,b]}$ is simply defined as the *circular barycenter* of the histogram values in this interval, *i.e.*

$$\alpha_{[a,b]} = \frac{2\pi}{L} \begin{cases} \sum_{i=a}^{b} i \cdot H[i] & \text{if } a \leq b \\ \sum_{i=a}^{L} i \cdot H[i] + \sum_{i=1}^{b} (i+L) \cdot H[i] & \text{if } a > b \end{cases} [2\pi]\ . \tag{29}$$

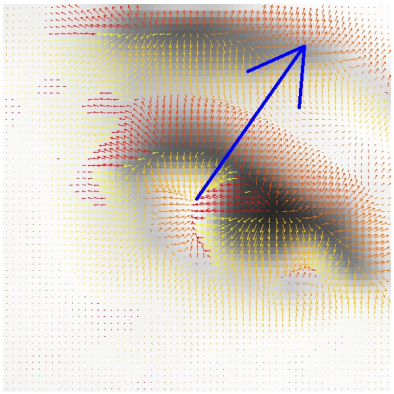
Figure 6 shows an example of the detection of the dominant orientation of an interest point from its histogram of gradient orientation.

A comparison the original SIFT orientation assignment technique with the a contrario mode detection is provided on a toy example in Figure 7.

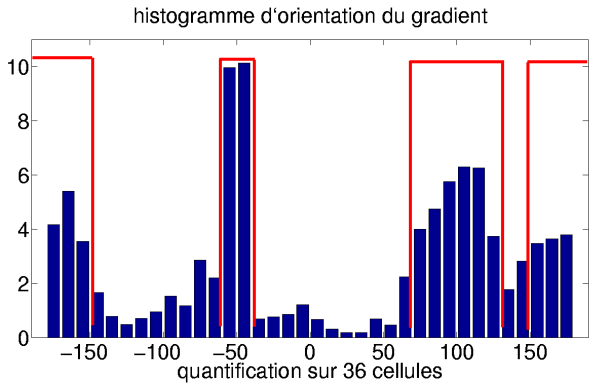
Eventually, observe that the *a contrario* mode selection makes it also possible to discard interest points that lies on edges structures. Indeed, only interest points with at least two orientation assignments should be kept.

3.4 Descriptor construction

Once interest points $\{m_i : (x_i, y_i, \sigma_i, \alpha_i)\}_i$ have been identified (typically a few thousand for a 1 megapixel digital image), a descriptor is then build for each of them. The SIFT descriptor is compact vector made of local histograms of the gradient orientation, extracted from the vicinity of the interest point.

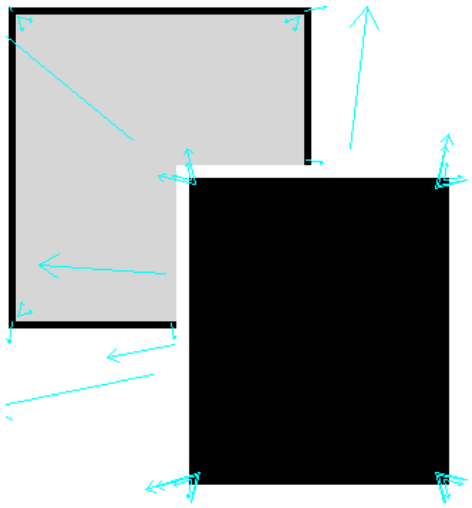


(a) An interest point and its orientation assignment

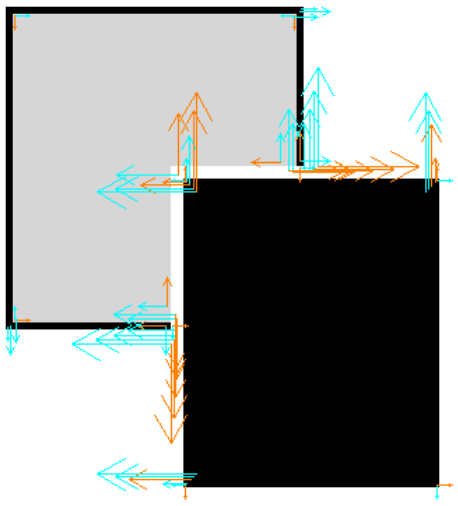


(b) *a contrario* selection of histogram mode [DMM03]

Figure 6: Illustration of the detection of the dominant orientation of an interest point based on mode detection.



(a) Orientation assignment with the original SIFT method



(b) Orientations assignment with the a contrario framework

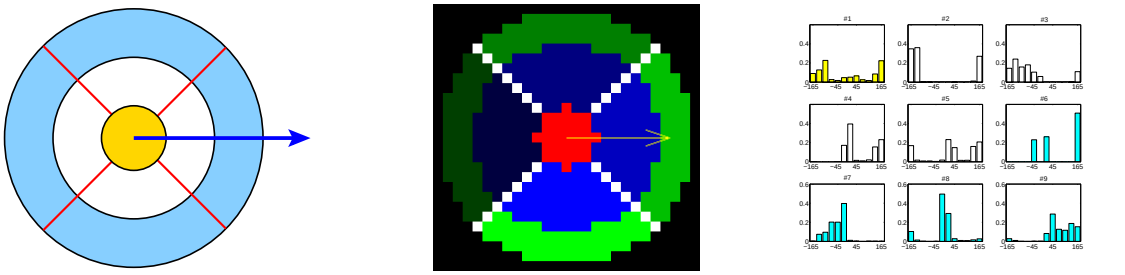
Figure 7: Comparison of orientation assignment techniques.

Local histograms First, a localization grid (see Figure 8(a)) is centered in the interest-point location $(x_i, y_i, \sigma_i, \alpha_i)$ in the scale-space representation of the image. Its size is proportional to the interest-point scale σ_i, and its orientation is defined accordingly to the point orientation α_i.

Tough D. Lowe makes use of a Cartesian grid in [Low04], one can use radial grid to achieve complete rotation invariance [MS05, Rab09]. This grid is divided into M radial regions (here $M = 9$), for each of which a local 1-D histogram of the gradient orientation is built from the image I_{σ_i}.

Orientation histograms are composed of $N = 8$ accumulators (or bins). In order to be robust to noise (the estimation of gradient orientation is not reliable for gradients with small magnitude), to the quantization and also to boundary effects, the entry of a sample (x_m, y_m, θ_m) into an histogram is weighted by

- the gradient magnitude in (x_m, y_m);
- a weight defined from the distance of the sample (x_i, y_i, α_i) to the nearest bin, using linear inter-



(a) Radial localization grid with 9 regions.

(b) Discrete mask with a radius of 12 pixels.

(c) Corresponding histograms of the gradient orientation

Figure 8: Radial localization grid used to extract local histograms.

polation;

- a weight defined from the distance of the gradient sample to the center (x_i, y_i) of the grid, using a Gaussian kernel.

Normalization The SIFT descriptor, which is composed of M 1-D histograms $\{h_m\}_{m=1,\dots,M}$, can be seen as a compact vector with $N \times M$ dimensions. The last step of the SIFT method consists in scaling the values of this vector in such a way that it has a unit Euclidean-norm. This normalization achieves the invariance of the descriptor to affine contrast change.

4 SIFT-like features matching

We have seen in the previous section how to get a local representation of images from SIFT descriptors. Within this framework, the comparison of several images boils down to the matching of such local features.

In the following, we consider a situation where one seeks for correspondences between N_Q *query descriptors* $\{a^i\}$ and a database of N_C *candidate descriptors* $\{b^j\}$, as illustrated by Figure 9.

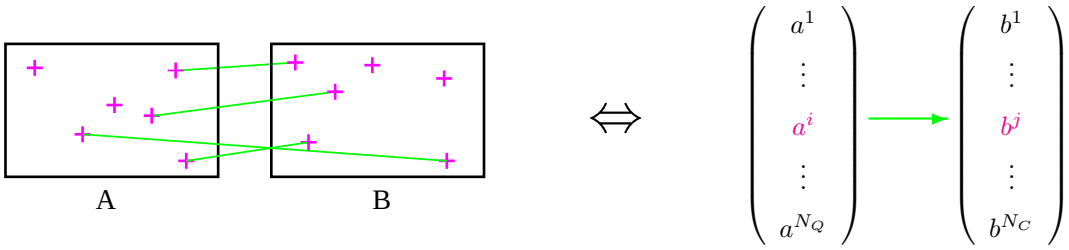


Figure 9: **Image matching *via* local representation** A is a query image described with N_Q query descriptors, and B is an image database composed of N_C candidates. The comparison of the query image with the database boils down to the problem of local feature matching.

The matching criterion should adapt itself to the complexity and diversity of the features.

4.1 Preliminary step: similarity measure

As it has been previously stressed in the introduction, a preliminary step to the feature matching consists in defining a similarity measure. From now, we denote by $D(a, b)$ the similarity measure descriptors a and b. We assume that the lower the quantity $D(a, b)$, the better the similarity. The definition of such a measure is very important in practice: for a given query descriptor a^i, we can sort the candidates

in the dataset $\{b^j\}_j$ by increasing order of similarity. The matching criterion then simply consists in thresholding this similarity measure to select relevant matches.

The dissimilarity measure should provide relevant comparisons between features and should be robust enough to cope with small variations of these features (quantization effect, geometrical distortions, *etc*).

The most classical way to compare SIFT-like descriptors is simply to use the L^p distance as in Formula (30), usually with $p = 2$ (Euclidean distance) [Low04]. Other distances that are used to compare local features include the χ^2 distance, as in [FL07] or the Mahalanobis distance [MM07]. The definitions of these distances in the framework of SIFT-like descriptors are recalled in Formula (31) and (32) respectively.

$$\mathrm{D}_{L^p}(a,b) \;:=\; \left(\sum_{m=1}^{M} \sum_{i=1}^{N} \mid a_m[i] - b_m[i] \mid^p \right)^{\frac{1}{p}} \tag{30}$$

$$\mathrm{D}_{\chi^2}(a,b) \;:=\; \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{(a_m[i] - b_m[i])^2}{a_m[i] + b_m[i]} \tag{31}$$

$$\mathrm{D}_{M}(a,b) \;:=\; \left(\sum_{m,n=1}^{M} \sum_{i,j=1}^{N} (a_m[i] - b_m[i]) \cdot w_{i,j,m,n} \cdot (a_n[j] - b_n[j]) \right)^{\frac{1}{2}} , \tag{32}$$

where $w_{i,j,m,n}$ is a weighting term (generally learn from the covariance of the data) enabling the comparison of bins with different index.

In what follows, we assume that distances have been computed between each a^i and each b^j. Note that this step can sometimes be replaced by approximate allocation algorithms, as in [BL97, ML09].

Two different criteria are often used in practice to validate matches, as detailed in [MS05, MP07], both relying on user-selected thresholds on the similarity measure. These criteria are discussed in Sections 4.2 and 4.3. Ideally, these thresholds should be set automatically and should depend on both the query and candidate descriptors. In sections 4.4 and 4.5 are shown two alternative matching criteria relying on such adaptive thresholds. Matches between the query and candidate descriptors are validated by rejecting casual matches, that are matches that can be produced by chance. Similar ideas are present in works dealing with the statistical analysis of object recognition processes [Lin97, Lin95]. Specifically, we resort to an *a contrario* methodology, first introduced in [DMM00] and then applied, among other things, to shape matching [MSC+06]. This approach provides thresholds on the dissimilarity measure that adapt to the query and candidate descriptors. This matching procedure also allows multiple detections over a database, while controlling the total number of matches, in particular in cases where the structure of interest is not present.

4.2 Fixed distance thresholding

The simplest matching criterion, that we call DT (Distance Threshold), relies on a global threshold on distances. That is, each query a^i is simply matched with candidates $\{b^j\}$ that are at a distance $d(a^i, b^j)$ smaller than the threshold.

Definition 9 (DT Criterion)**.** *For each query descriptor a^i, every correspondence with a candidate descriptor for which the similarity measure is smaller than a global threshold t is validated. The set of selected matches is therefore:*

$$\mathcal{C}_{DT} := \left\{ (a^i, b^j)\,,\; i \in \{1,\dots,N_Q\} \;\; and \;\; j \in \{1,\dots,N_C\} \;:\; D(a^i,b^j) \leq t \right\}$$

Usually, matches are restricted to the nearest neighbor [Bau00, JT08] for each query descriptor, in order to limit multiple false detections that often affect some query descriptors. We will refer to this criterion as NN-DT (Nearest Neighbor Distance Threshold).

Definition 10 (NN-DT Criterion). *For each query descriptor a^i, a unique correspondence with the nearest candidate descriptor is validated if the similarity measure is smaller than a global threshold t. The set of selected matches is therefore:*

$$\mathcal{C}_{NN\text{-}DT} := \left\{ (a^i, b^{J(i)}) \ , \ i \in \{1, \dots, N_Q\} \ : \ D(a^i, b^{J(i)}) \leq t \ \ s.t. \ \ J(i) = \underset{j \in \{1,\dots,N_C\}}{\arg\min} \ D(a^i, b^j) \right\}$$

Three main drawbacks inherent to this approach restrict its use in practice. First, the nearest neighbor restriction limits the number of correct matches so that, in some applications, one prefers to select the K nearest neighbors: $K = 3$ in [FTG06], $K = 4$ in [BL07] for image stitching, and K between 5 and 10 in [RLSP07]. The price to pay is then a higher proportion of false matches. Secondly, the nearest neighbor restriction is also problematic in cases where there are multiple occurrences of the structure of interest, for instance when the target object is present more than once in the database (see for instance [DMA07]), when dealing with objects having repetitive parts, such as buildings (this issue is studied in [ZK06]), or when the interest point detector yields spurious repetitions of the structure to be coded. Lastly, the great variability of distances between descriptors from images to images (as shown in Section 4.6) makes it particularly difficult to set the right threshold for a particular application.

Note that in [DZLF94], a variant of NN-DT consists in keeping only matches (a, b) for which a is also the nearest neighbor of b.

4.3 Distance ratio thresholding (Lowe criterion)

In order to reduce the variability of the chosen threshold, Lowe [Low04] introduces another criterion by comparing the distances between a^i and its closest and second-closest neighbors respectively. If the ratio between the two distances is below a threshold r, the match with the closest neighbor is validated.

Definition 11 (NN-DR Criterion). *For each query descriptor a^i, compute the nearest and the second nearest candidate descriptor in the database, respectively referred to as b^{J1} and b^{J2}. A unique correspondence with the nearest candidate descriptor is validated if the ratio of similarity measures $D(a^i, b^{J1(i)})/D(a^i, b^{J2(i)})$ is smaller than a global threshold r. The set of selected matches is therefore:*

$$\mathcal{C}_{NN\text{-}DR} := \left\{ (a^i, b^{J1(i)}) \ , \ i \in \{1, \dots, N_Q\} \ : \ \frac{D(a^i, b^{J1(i)})}{D(a^i, b^{J2(i)})} \leq r \ , \ s.t. \right.$$
$$\left. J1(i) = \underset{j \in \{1,\dots,N_C\}}{\arg\min} \ D(a^i, b^j) \ \ et \ J2(i) = \underset{j \in \{1,\dots,N_C\} \setminus J1(i)}{\arg\min} \ D(a^i, b^j) \right\}$$

This popular criterion, that we call NN-DR (Nearest Neighbor Distance Ratio), benefits from its simplicity and the fact that it is by far more robust than a simple threshold on distances. However, the choice of the "optimal" threshold r is strongly dependent on both the application and the database: $r = 0.8$ in [Low04], $r = 0.6$ in [SSS08], $r = 0.95$ in [CM05], or r between 0.56 and 0.7 in [MP07] for instance. In practice, the NN-DR criterion behaves very well (and in particular significantly better than the NN-DT criterion as shown in [MP07]) when the target to be matched is present *exactly once* in the candidate database. Indeed, in this case, it makes sense to assume that the distance to the nearest neighbor is small compared to distances to other candidates and in particular to the second nearest neighbor.

Now, the reason why this criterion should work when the structure of interest is not present is less clear. Such a situation is shown in the experimental section 4.6. It is of great practical importance, because in real situations a computer vision system relying on the matching of local features has to deal with situations when the target is present as well as with situations when the target is missing. Moreover, this criterion is by nature limited to the nearest neighbor, and, as NN-DT, may fail in the case where the structures of interest appear more than once, as already mentioned.

4.4 Adaptive distance thresholding (Lisani criterion)

Several variants of the NN-DR matching criterion have been proposed. In [BSW05], it is suggested to adapt the NN-DR criterion by averaging the distance to the second neighbor over several images for panorama stitching.

Another interesting alternative matching criterion has been proposed by J-L. Lisani in [CLM+08]. His idea is to adapt the distance ratio test in such a way that it tolerates multiple detections. To do so, one consider an alternative database B' composed of N'_C descriptors in which we look for the nearest neighbor for each query descriptor. Now, a match is validated if the ratio between the distance from the query to a given candidate in B and the distance from the query to its most similar candidate in B' is below a fixed threshold r. This method can be seen as an *a contrario* approach in which one estimates *adaptive thresholds* on the similarity measure depending on the query and a background model (the database B').

Definition 12 (ADT Criterion). *For each query descriptor a^i, define an adaptive threshold $t(a^i)$ as the similarity measure with the nearest descriptor in an alternative database. Then, every correspondence with a candidate descriptor b^j for which the similarity measure is smaller than the threshold $t(a^i)$ is validated. The set of selected matches is therefore:*

$$\mathcal{C}_{ADT} := \left\{ (a^i, b^j)\,,\, i \in \{1,\dots,N_Q\} \;\; and \;\; j \in \{1,\dots,N_C\} \;:\; D(a^i,b^j) \leq r \cdot \min_{j \in \{1,\dots,N'_C\}} D(a^i, b'^j) \right\}$$

This criterion, that we call ADT(Adaptive distance thresholding), has two major advantages over the NN-DR criterion: the threshold automatically adapt to the query (without being affected by self-similarity and multiple occurrences) to the query and the criterion enables multiple matches.

Nevertheless, it requires some extra computations (the database B' should have at least the same size as B'). Moreover, contrary to the *a contrario* methodology introduced in Section 2, it does not make it possible to set the threshold r in such way that we control the expected number of false alarms.

4.5 A contrario matching criterion

In this section, we introduce a generic way to compute matching thresholds in the framework of local, SIFT-like descriptors. Recall that we consider N_Q query descriptors $\{a^i\}$ and N_C candidate descriptors $\{b^j\}$. The question is then: for each a^i, to which b^j (if any) should it be matched? To answer this question, we rely on the general principle of *a contrario* methodology presented in Section 2 and fix matching thresholds that ensure the rejection of casual matches.

Note that an alternative *a contrario* approach for the matching of SIFT descriptors has been proposed in [CLM+08]. Other *a contrario* matching criteria as also been proposed for shape recognition [MSC+06], image retrieval [HGS08] and stereoscopic vision [SAM08].

4.5.1 The null hypothesis

Recall that each descriptor a^i is made of M orientation histograms, $a^i = (a^i_1, \dots, a^i_M)$. In order to define the null hypothesis, we assume that the distance between two descriptors a^i and b^j can be written as $D(a^i, b^j) = \sum_{m=1}^{M} d(a^i_m, b^j_m)$. This means that the total distance between two descriptors is merely a sum of distances between histograms. Observe that this is a very mild assumption, satisfied for classical bin-to-bin distances (see *e.g.* L^2, L^1 or χ^2 in formulas (30 and (31)), as well as for the Circular Earth Mover's Distance (CEMD) proposed in [RDG09].

Now, the idea is to fix matching thresholds in such a way that a given descriptor a^i would scarcely be matched with a generic random descriptor. In what follows, we will write $\mathbf{b}$ for a random descriptor, that is a collection $(\mathbf{b}_1, \dots, \mathbf{b}_M)$ of M random histograms. In order to specify what we mean by a "generic" random descriptor, we now define a hypothesis on $\mathbf{b}$ relying on the independence of the M real random variables $d(a^i_m, \mathbf{b}_m)$.

Definition 13. *For $i \in \{1, \dots, N_Q\}$, a random descriptor $\mathbf{b}$ is said to satisfy the i^{th} null hypothesis, $\mathcal{H}_0^i$, if*
$\left\{d(a_m^i, \mathbf{b}_m)\right\}_{m \in \{1, \dots M\}}$ *are mutually independent random variables.*

Under hypothesis $\mathcal{H}_0^i$, the law of the random variable $D(a^i, \mathbf{b})$ has density $\underset{m=1}{\overset{M}{*}} p_m^i$, where $*$ denotes the convolution product and p_m^i the probability density function of the random variable $d(a_m^i, \mathbf{b}_m)$. Thus,

Proposition 3. *Under the hypothesis $\mathcal{H}_0^i$, the probability that the distance between a^i and $\mathbf{b}$ is smaller than a given threshold δ is*

$$f_i(\delta) := \mathbb{P}\left(D(a^i, \mathbf{b}) \leq \delta \,|\, \mathcal{H}_0^i\right) = \int_0^\delta \underset{m=1}{\overset{M}{*}} p_m^i(x)\, dx\,. \tag{33}$$

The validity of a match will then be decided by thresholding this probability, as explained in the next section. This in turn yields thresholds on distances that depend on both a^i and the observed distribution of candidate descriptors.

In order to numerically compute the probability given by Equation (33), we need to estimate the probability density functions p_m^i, for each $i \in \{1, \dots N_Q\}$ and each $m \in \{1, \dots M\}$. These laws are empirically estimated over the database $\{b^1, \dots, b^{N_C}\}$. For this, we simply use histograms of realizations of the distances over the database.

4.5.2 Meaningful matches

Let us consider two given descriptors a^i and b^j at distance $\delta = D(a^i, b^j)$. We decide to match these descriptors as soon as $f_i(\delta) = \mathbb{P}(D(a^i, \mathbf{b}) \leq \delta \,|\, \mathcal{H}_0^i)$ is small enough. In other words, we match these descriptors if it is unlikely that a generic random descriptor $\mathbf{b}$ (that is, a descriptor satisfying $\mathcal{H}_0^i$) is closer from a^i than b^j is. In this case, we conclude that the proximity between a^i and b^j cannot be due to chance. It therefore remains to define what we mean by "small", thus to automatically fix a threshold on this probability. Following the general approach of *a contrario* methods, we choose the threshold in order to control the average number of false detections.

Definition 14. *For a given $\varepsilon > 0$, a match between $\{a^i\}$ and a candidate descriptor $\{b^j\}$ is said to be ε-meaningful if*

$$f_i\left(D(a^i, b^j)\right) \leq \frac{\varepsilon}{N_Q N_C}, \tag{34}$$

where the function f_i is defined by Formula (33), N_Q is the number of query descriptors and N_C the number of candidate descriptors.

Observe that for a given $\varepsilon > 0$, the unique threshold $\frac{\varepsilon}{N_Q N_C}$ yields N_Q adaptive thresholds on distances, defined for $i = 1, \dots, N_Q$ by

$$\delta_i(\varepsilon) = \arg\max_\delta \left\{ f_i(\delta) \leq \frac{\varepsilon}{N_Q N_C} \right\}. \tag{35}$$

For each query descriptor a^i, a match between a^i and some b^j will be ε-meaningful if $D(a^i, b^j) \leq \delta_i(\varepsilon)$. The reason behind the choice of the threshold $\frac{\varepsilon}{N_Q N_C}$ is the following:

Proposition 4. *When testing N_Q queries against N_C candidates satisfying all the null hypotheses,* the expected number *of ε-meaningful matches is smaller than ε.*

Proof. The function $f_i : \delta \mapsto \mathbb{P}_{\mathcal{H}_0^i}(D(a^i, \mathbf{b}) \leq \delta)$ is the repartition function of the random variable $D(a^i, \mathbf{b})$ when the random descriptor $\mathbf{b}$ follows the null hypothesis $\mathcal{H}_0^i$. Recall that if X is a random variable with

a repartition function F_X, then $\mathbb{P}(F_X(X) \leq \alpha) \leq \alpha \ \forall \alpha \leq 0$. Hence, $\mathbb{P}_{\mathcal{H}_0^i}(F_i(D(a^i, \mathbf{b})) \leq \alpha) \leq \alpha$. We denote by $\mathbb{1}(\chi)$ the indicator of event χ, and $\mathbb{E}_{\mathcal{H}_0}$ the expectation under the global null hypothesis $\mathcal{H}_0$.

$$\begin{aligned}\mathbb{E}_{\mathcal{H}_0}\Big[\sum_{i=1}^{N_C}\sum_{j=1}^{N_Q} \mathbb{1}\big\{(a^i, \mathbf{b}^j) \text{ is } \varepsilon\text{-significatif }\big\}\Big] &= \sum_{i=1}^{N_C}\sum_{j=1}^{N_Q}\mathbb{P}_{\mathcal{H}_0^i}\Big[\, F_i\left(D(a^i, \mathbf{b}^j)\right) \leq \frac{\varepsilon}{N_Q N_C}\,\Big] \\ &\leq \sum_{i=1}^{N_C}\sum_{j=1}^{N_Q}\frac{\varepsilon}{N_Q N_C} = \varepsilon\,.\end{aligned}$$

□

This result is a simple consequence of the linearity of the mathematical expectation. Observe that it would have been much more difficult to bound the *probability* of false detections, since distances between different descriptors are not necessarily independent. A more in-depth analysis of this interesting aspect can be found in [DMM08]. Let us also remark that this choice of δ is actually one of the simplest approaches to multiple testing, and is known in the statistical community as a Bonferroni correction [Mil91].

Definition 15 (AC criterion)**.** *A correspondence (a^i, b^j) is validated if the quantity $NFA(a^i, D(a^i, b^j))$ is lower or equal to the threshold ε. The set of matches is therefore:*

$$\mathcal{C}_{AC} := \Big\{(a^i, b^j)\ ,\ NFA\big(a^i, D(a^i, b^j)\big) \leq \varepsilon \ \ \forall i \in \{1, \dots, N_Q\} \ \text{ and } \ \forall j \in \{1, \dots, N_C\}\Big\}$$

In practice, for a fixed ε and for each descriptor a^i we perform the several steps summed up in the table 3 to achieve the matching procedure.

Table 3: A Contrario (AC) matching procedure.

Algorithm 3 Automatic distance threshold setting.

Input: N_Q query descriptors $\{a^i\}$ and N_C candidate descriptors $\{b^j\}$, parameter $\varepsilon > 0$.

For each query descriptor a^i, $i = 1, \dots, N_Q$:

1) computation of distances $d_m(a^i, b^j)$ for all $m = 1, \dots, M$ and $j = 1, \dots, N_C$;
2) estimation of probability density functions: for each m, p_m^i computed as the empirical distribution of $d_m(a^i, b^j)$, when b^j spans the database;
3) computation of $f_i : \delta \mapsto \mathbb{P}\left(D(a^i, \mathbf{b}) \leq \delta \,|\, \mathcal{H}_0^i\right)$ using Formula (33);
4) computation of threshold $\delta_i(\varepsilon)$ using Formula (35);
5) matching of a^i with each descriptor b^j such that $D(a^i, b^j) \leq \delta_i(\varepsilon)$;

Output: List of correspondences.

From now on, we will refer to this matching criterion as AC. Let us now comment on this criterion. First, one needs to fix the value of ε, that in turn yields a threshold on distances. Since this value corresponds to an expected number of false detections, we claim that it is much simpler to set than a threshold on distances. Indeed, it is well known that distances between descriptors vary very much from one descriptor to another or one image to another, as will be illustrated in the experimental section. Now, the threshold on distances computed thanks to step 3) above depends on both the particular descriptor at hand, a^i, and the database (e.g. an image, or a set of images) against which it is matched. This is due both to the learning of marginals p_m^i and to the fact that the number of descriptors is taken into account by Formula (35). In particular, one can hope that the proposed matching criterion works well

over a relatively large image database and in the presence of distractors, as will be confirmed by the experimental section.

Last, observe also that the number of matches is not restricted to the nearest neighbor, even though one has the possibility to add such a restriction depending on the application. As already mentioned, this is in contrast with classical approaches to the matching of local features. In the experimental section, we will see that removing the nearest neighbor restriction on the *a contrario* criterion does not yield an explosion of the number of wrong matches, contrarily to what is observed when simply thresholding distances between local features.

Observe now that the only parameter to be set is the matching threshold ε in Formula (35). Following the classical framework of *a contrario* methods [DMM08], ε can be safely set to the value $\varepsilon = 1$ in all cases. However, because of some dependencies introduced in particular by the Gaussian blur involved in the scale space computation, histograms from neighboring regions in the localization grid are not fully independent for a given SIFT descriptor. As a consequence, values of ε equal to 10^{-1} or 10^{-2} often yield better results, as will be shown in the experiments. It is important at this point to notice that, as it is common with *a contrario* methods[DMM08], the influence of ε is better expressed on a logarithmic scale.

Proposition 5. *Meaningfulness in the Gaussian case Let a be a given descriptor and b a random descriptor, such that the distances $d = d(a_m, b_m)$ are independent and identically distributed, and assume that their probability density function can be well approximated by a Gaussian distribution. The convolution of those M marginals is hence a Gaussian distribution, from which the mean and the standard deviation are respectively noted μ and σ.*

Then, if $\varepsilon \leq \dfrac{N_Q N_C}{2e\sqrt{\pi}}$ and if $D(a,b) < \mu - \sigma\sqrt{2}\sqrt{\log \dfrac{N_Q N_C}{2\sqrt{\pi}\varepsilon}}$, then the matching of a and b is ε-meaningful.

Proof. Following Definition 14, the matching between a and b is ε-meaningful if $f(D(a,b)) \leq \dfrac{\varepsilon}{N_Q N_C}$. Under the previous hypotheses, f can be written as the Gauss error function:

$$f(\delta) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\delta - \mu}{\sqrt{2}\sigma}\right)\right),$$

where the error function is defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \int_{-\infty}^x e^{-t^2} dt - 1 .$$

Now, for $x < 0$,

$$erf(x) \leq \frac{2}{\sqrt{\pi}} \int_{-\infty}^x e^{-t^2} \left(1 + \frac{1}{2t^2}\right) dt - 1 = \frac{1}{\sqrt{\pi}} \frac{e^{-x^2}}{|x|} - 1 .$$

Consequently, if $\delta < \mu$,

$$f(\delta) \leq \frac{1}{2\sqrt{\pi}} \frac{e^{-\left(\frac{\delta - \mu}{\sqrt{2}\sigma}\right)^2}}{\left|\frac{\delta - \mu}{\sqrt{2}\sigma}\right|} . \tag{36}$$

Thus, if $D(a,b) < \mu - \sigma\sqrt{2}\sqrt{\log \dfrac{N_Q N_C}{2\sqrt{\pi}\varepsilon}}$, then

$$\frac{D(a,b) - \mu}{\sigma\sqrt{2}} < -\sqrt{\log \frac{N_Q N_C}{2\sqrt{\pi}\varepsilon}} \text{ , and } \left(\frac{D(a,b) - \mu}{\sigma\sqrt{2}}\right)^2 > \log \frac{N_Q N_C}{2\sqrt{\pi}\varepsilon} .$$

Moreover, if we assume that $\varepsilon \leq \frac{N_Q N_C}{2e\sqrt{\pi}}$, then $\log \frac{N_Q N_C}{2\sqrt{\pi}\varepsilon} > 1$, and it follows that $\frac{|D(a,b)-\mu|}{\sigma\sqrt{2}} > 1$. As a consequence,

$$\left(\frac{D(a,b)-\mu}{\sigma\sqrt{2}}\right)^2 + \log\left|\frac{D(a,b)-\mu}{\sigma\sqrt{2}}\right| > \log\frac{N_Q N_C}{2\sqrt{\pi}\varepsilon} .$$

Finally, using Eq.(36), we obtain that $f(D(a,b)) \leq \frac{\varepsilon}{N_Q N_C}$. □

4.6 Experiments

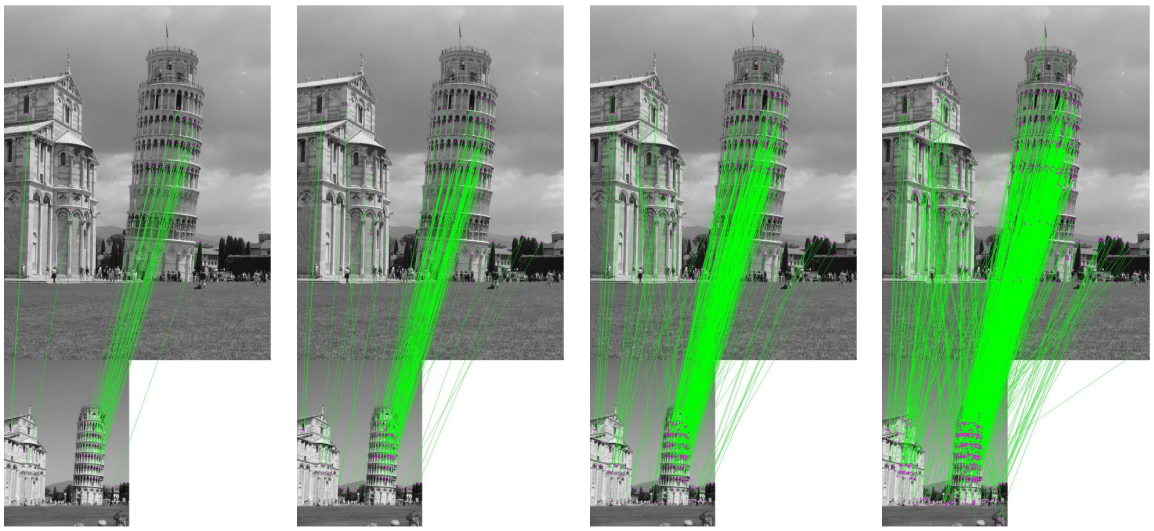
In this section are shown some experiments illustrating the aforementioned approaches.

Firstly, we show the behavior of the proposed matching procedure using different thresholds in the case of a scene with repetitive structures. Such a situation is common in the case of, e.g., images of buildings. As pointed out in [ZK06], these are difficult correspondence problems. Classical approaches could fail to provide enough relevant correspondences between images of the same scene. We compare two different views of the tower of Pisa shown in Fig. 10. Criterion NN-DR (used in Figs. 10(e), 10(f), 10(g) and 10(h) with respectively $r = 0.7$, $r = 0.8$, $r = 0.85$ and $r = 0.9$) can only correctly match a relatively low number of points while controlling the number of false matches. Indeed, the presence of repetitive structures can foul the NN-DR criterion because of several candidate descriptors at a similar distance to the query. On the contrary, using the AC matching criterion -which is not restricted to the nearest neighbor-, results in multiple matches between columns and arches (Figs. 10(a), 10(b), 10(c) and 10(d) with respectively $\varepsilon = 10^{-2}$, $\varepsilon = 10^{-1}$, $\varepsilon = 1$, and $\varepsilon = 10$).

Next, a single image (blue-framed) is matched separately with 8 different images (Fig. 11(a)). Four of them contain (one or several times) a common object with the query image (a can). The four other images do not contain the can. The AC matching procedure presented in section 4.5 is shown in Fig. 11(b)). It is compared to two classical matching procedures: NN-DR criterion in Fig. 12(a) or NN-DT criterion in Fig. 12(b). For each method, all images are matched with the same threshold ($\varepsilon = 10^{-2}$ for AC, $r = 0.8$ for NN-DR, and $t = 0.45$ for NN-DT). These thresholds are set in such a way as to obtain roughly the same number of correct matches between the query image and the image at the center of the leftmost column.

This matching experiment shows that the AC criterion yields much fewer false matches on images where the object is not present and better detection of multiple occurrences. It is also interesting to notice that there are less false matches even in images where the object is present.

The last figure shows the interest of the AC CRITERION for multiple occurrences. Even in such an extremal situation where the structure of interest (the logo) is present 28 times in the database, the corresponding matches are still meaningful and thus validated.



(a) AC, $\varepsilon = 10^{-2}$ (b) AC, $\varepsilon = 10^{-1}$ (c) AC, $\varepsilon = 1$ (d) AC, $\varepsilon = 10$

(e) NN-DR, $r = 0.7$ (f) NN-DR, $r = 0.8$ (g) NN-DR, $r = 0.85$ (h) NN-DR, $r = 0.9$

Figure 10: Matching an object with repetitive structures: the tower of Pisa. Two different matching procedures are used with different thresholds: the first row corresponds to the AC criterion and the second row corresponds to the NN-DR criterion. The first criterion permits to match the repeated elements of the tower.

(a) query image and 8 images dataset

(b) AC matching criterion with $\varepsilon = 10^{-2}$

Figure 11: Comparison of different matching procedures. One query image (blue framed) containing a can is matched separately against 8 images (Fig. 11(a)). Only half of these images contain the can, present one or several times. For each matching procedure, the query image is compared with all 8 images using the same threshold.

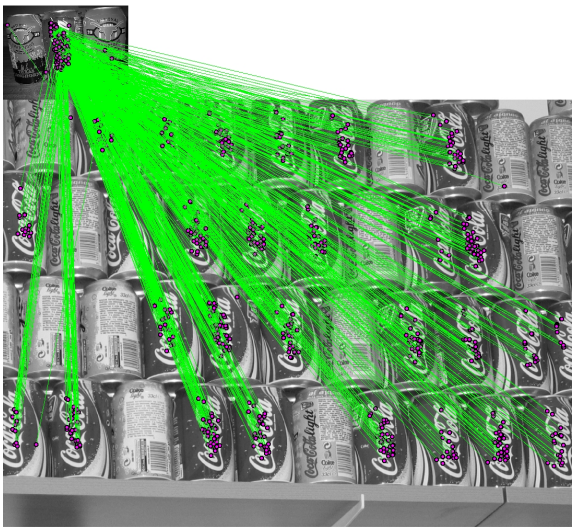
(a) NN-DR matching criterion with $r = 0.8$

(b) NN-DT matching criterion with $t < .45$

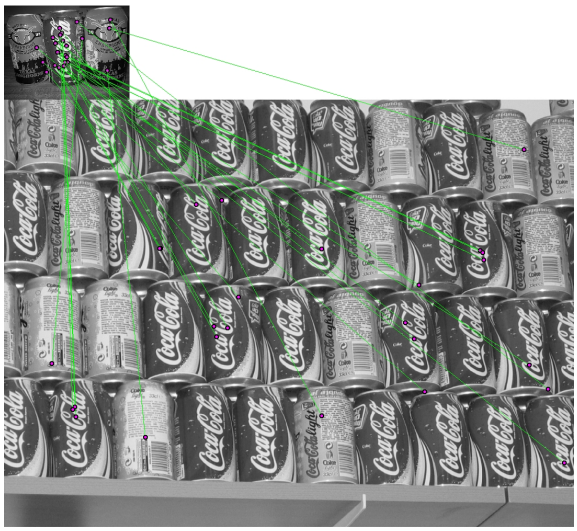
Figure 12: Observe that for a given number of correct matches with this left-centered image, the AC matching procedure introduced in FIgure 11(b) (§ 4.5) yields more correct matches in other images while providing a better control of the number of false detections than classical procedures (NN-DT and NN-DR criteria).

(a) Query image.

(b) A bunch of soda cans.



(c) AC with $\varepsilon = 10^{-1}$



(d) NN-DR with $r = 0.8$

Figure 13: **Multiple object matching.** (Photography courtesy of Frédéric Sur)

5 Projects

IPOL project reports are due to January; it should be composed of a detailed description of the algorithm and an experimental study (exhibiting both its properties and its limitations).

An IPOL publication may be eventually proposed, with a short description of the studied algorithm and a clean implementation.

Project 1 - A SIFT descriptor for MSER (Maximally Stable Extremal Regions)

This project is devoted to the study of the SIFT description technique introduced in [FL07] for Maximally Stable Extremal Regions (MSER) detector proposed in [MCUP02] for wide baseline stereo imaging. Paper is freely available here: http://www.cs.ubc.ca/~perfo/papers/forssen_iccv07.pdf.

First, it is asked to study the MSER algorithm which –similarly to SIFT– aims at detecting robust and repeatable local features for image comparison (codes are available for experiments). This algorithm starts from the exhaustive level-set description of an image to select only contrasted level-lines. The major interest of the detected MSER regions is that it is very easy to normalize such features to obtain **affine-invariant** descriptors (up to a scale change).

The second goal of the project is to study and implement the multi-scale variant of the MSER proposed by Forssen and Lowe.

Project 2 - Fast SIFT matching

This project is devoted to the study of the *M-SIFT* descriptors [MM07] (paper is available here: http://cmp.felk.cvut.cz/~matas/papers/mikolajczyk-msift-iccv07.pdf).

In this course, only exhaustive local feature comparison techniques for image matching have been studied. For on-line image search with large database and for real-time applications, using such approaches are generally too much time consuming. In practice, *Approximate Nearest-Neighbor* (ANN) search approaches are often used to speed-up the matching step.

The goal of this project is to analyze the SIFT variant proposed in [MM07], which is claimed to improve SIFT matching accuracy when using ANN techniques.

Project 3 - A contrario SIFT matching

This project is devoted to the study of the *a contrario* matching criterion presented in Section 4.5 (see [RDG09] for further details).

It is first required to compare the AC matching criterion to other matching criteria described in Sections § 4.2, § 4.3 and § 4.4. Matlab and C/C++ codes are available for experimental study.

The second part of this project is the study of the experimental limitations of the the *a contrario* matching criterion. The major drawback of this technique is that it greatly depends on the accuracy of the null hypothesis to reject mismatches. For SIFT matching, the null hypothesis rely on the assumption that the distances between histograms are independent, which is not completely verified in practice for random pairs of descriptors. The goal of this project is to study the practical interest of Principal Component Analysis (PCA) to decorrelate histograms and to improve matching results.

References

[Bau00] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. CVPR*, 2000. 16

[BL97] J. Beis and D. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proc. CVPR*, pages 1000–1006, 1997. 16

[BL07] Matthew Brown and David G. Lowe. Automatic panoramic image stitching using invariant features. *Int. J. Comput. Vision*, 74(1):59–73, 2007. 17

[BMP02] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002. 2

[BSW05] M. Brown, R. Szeliski, and S. Windner. Multi-image matching using multi-scale oriented patches. In *Proc. CVPR*, pages 510–517, 2005. 18

[BTG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *ECCV*, pages 404–417, 2006. 8

[CLM+08] F. Cao, J.L. Lisani, J.-M. Morel, P. Musé, and F. Sur. *A theory of shape identification*, volume 1948 of *Lecture Notes in Mathematics*. Springer, 2008. 18

[CM05] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *Proc. CVPR*, pages 220–226, 2005. 17

[Des00] A. Desolneux. *Maximal Meaningful Events and Applications to Image Analysis*. PhD thesis, ENS de Cachan, 2000. 10

[DLMM02] A. Desolneux, S. Ladjal, L. Moisan, and J.-M. Morel. Dequantizing image orientation. 11(10):1129–1140, 2002. 10

[DMA07] Francois Destrempes, Max Mignotte, and Jean-Francois Angers. Localization of shapes using statistical models and stochastic optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1603–1615, 2007. 17

[DMM00] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. Meaningful alignments. *Int. J. Comput. Vision*, 40(1):7–23, 2000. 1, 5, 16

[DMM03] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003. 1, 10, 14

[DMM08] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis: A Probabilistic Approach*. Springer Verlag, 2008. 5, 20, 21

[DZLF94] R. Deriche, Z. Zhang, Q.-T. Luong, and O. Faugeras. Robust recovery of the epipolar geometry for an uncalibrated stereo rig. In *ECCV '94: Proceedings of the third European conference on Computer vision (vol. 1)*, pages 567–576, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc. 17

[FL07] P.E. Forssén and D.G. Lowe. Shape descriptors for maximally stable extremal regions. In *IEEE ICCV*, 2007. 16, 27

[FTG06] Vittorio Ferrari, Tinne Tuytelaars, and Luc Gool. Simultaneous object recognition and segmentation from single or multiple model views. *Int. J. Comput. Vision*, 67(2):159–188, 2006. 17

[HGS08] T. Hurtut, Y. Gousseau, and F. Schmitt. Adaptive image retrieval based on the spatial organization of colors. *Computer Vision and Image Understanding, in Press*, 2008. 18

[HS88] Chris Harris and Mike Stephens. A combined corner and edge detector. *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988. 8, 9

[JT08] Jiaya Jia and Chi-Keung Tang. Image stitching using structure deformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):617–631, 2008. 16

[Lin94] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Norwell, MA, USA. Kluwer Academic Publishers, 1994. 7

[Lin95] M. Lindenbaum. Bounds on shape recognition performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(7):666–680, 1995. 16

[Lin97] M. Lindenbaum. An integrated model for evaluating the amount of data required for reliable recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11):1251–1264, 1997. 16

[Low99] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society. 1, 8, 9

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 1, 2, 8, 10, 14, 16, 17

[MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002. 2, 27

[Mil91] R. G. Miller. *Simultaneous Statistical Inference*. Springer-Verlag, New York, 1991. 20

[ML09] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP'09 : Proceedings of the International Conference on Computer Vision Theory and Applications*, 2009. 16

[MM07] Krystian Mikolajczyk and Jiri Matas. Improving SIFT for fast tree matching by optimal linear projection. page 8, Los Alamitos, CA, USA, 2007. IEEE Computer Society. 16, 27

[Mor80] Hans Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. In *tech. report CMU-RI-TR-80-03, Robotics Institute, Carnegie Mellon University & doctoral dissertation, Stanford University*, number CMU-RI-TR-80-03. September 1980. 9

[MP07] Pierre Moreels and Pietro Perona. Evaluation of features detectors and descriptors based on 3D objects. *Int. J. Comput. Vision*, 73(3):263–284, 2007. 16, 17

[MS01] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. *International Conference on Computer Vision*, pages 525–531, 2001. 9

[MS05] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. 14, 16

[MSC+06] Pablo Musé, Frédéric Sur, Frédéric Cao, Yann Gousseau, and Jean-Michel Morel. An a contrario decision method for shape element recognition. *Int. J. Comput. Vision*, 69(3):295–315, 2006. 1, 2, 16, 18

[Rab09] J. Rabin. *Approches robustes pour la comparaison d'images et la reconnaissance d'objets, disponible à:* `http://tel.archives-ouvertes.fr/tel-00472442/en/`. PhD thesis, Télécom ParisTech, 2009. 14

[RDG09] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3):931–958, 2009. 1, 18, 27

[RLSP07] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Segmenting, modeling, and matching video clips containing multiple moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):477–491, 2007. 17

[SAM08] Neus Sabater, Andres Almansa, and Jean-Michel Morel. Rejecting wrong matches in stereovision. *Preprint CMLA 2008-28*, 2008. 18

[SSS08] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 2008. 17

[ZK06] Wei Zhang and Jana Kosecka. Generalized RANSAC Framework for Relaxed Correspondence Problems. In *3DPVT '06: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pages 854–860, Washington, DC, USA, 2006. IEEE Computer Society. 17, 22