A contrario matching of local descriptors

Julien Rabin, Yann Gousseau, Julie Delon

A contrario matching of local descriptors

Julien Rabin, Yann Gousseau, Julie Delon
GET/TELECOM-PARIS, LTCI (UMR CNRS 5141)
46 rue Barrault, 75013 Paris, France
{rabin,gousseau,delon}@enst.fr

Abstract

This contribution focuses on the matching of local features between images. Given a set of query descriptors and a database of candidate descriptors, the goal is to decide which ones should be matched. This is a crucial issue, since the matching procedure is often a preliminary step for object detection, scene identification or image matching. In practice, this matching step is often reduced to a specific threshold on the Euclidean distance to the nearest neighbor.

We first introduce a robust distance between descriptors, making use of the Earth Mover's Distance (EMD). We then propose an a contrario *framework for the matching procedure, which enables us to control the number of false alarms. This approach yields validation thresholds automatically adapted to the complexity of the descriptor to be matched and to the diversity and size of the database. The method makes it possible to detect multiple occurrences and to rate the validated matches according to their meaningfulness.*

1. Introduction

Matching local features is a very convenient way of comparing several pictures. Many applications -such as object detection, stereo correspondence, image stitching, 3D reconstruction- are based on such procedures. An exhaustive list of the applications of the matching of local descriptors is beyond the scope of this paper. Illustrating examples can be found in [13, 2, 9].

Such methods require two preliminary steps. First, a few interest points are selected to reduce the coding of information. A descriptor is then built for each detected interest point. Many studies have proposed different interest points and geometric descriptors. In a comparative study [15], the SIFT descriptor [13] has proven to be the most robust and invariant representation method.

The second step consists in matching some of the *query descriptors* $\{a^i\}_{i=1...N_A}$ (*e.g.* extracted from a query image) with *candidate descriptors* $\{b^j\}_{j=1...N_B}$ from a database (*e.g.* another image or a set of images), using a dissimilarity measure and a selection criterion. For each query descriptor a^i, elements b^j from the database are ranked according to their similarity with a^i. Then, a criterion is used to validate the matches, that is to decide which candidates should be matched with the query.

In many applications, the matching procedure is followed by a validating step based on the global coherence of matches and making use of the Hough transform, RANSAC or alternatives, see *e.g.* [13, 4,

3]. In more specific applications, it is possible to get rid of false matches using geometrical constraints, see *e.g.* [5]. The quality of the results of such methods strongly depends on the proportion of false matches and it is crucial to have a high true matching rate, especially in the case of multiple or complex transformations.

Whereas the extraction and representation of descriptors has been thoroughly studied (see *e.g.* the references in [15]), there are few studies about their matching. In the literature, most matching processes start by computing distances between the N_a query descriptors a^i and the database $\{b^j\}_{j=1...N_B}$. Then three different criteria are used in practice to validate matches, as detailed in [15]. The simplest one uses a global threshold on distances $d(a^i, b^j)$. A refinement is to restrict such matches to only the closest neighbor for each a^i, in order to avoid multiple false detections that often occur. Such simple approaches are not satisfactory, essentially because optimal thresholds vary greatly depending on the query and candidate descriptors. For that purpose, Lowe [13] introduces another criterion by comparing the distance between a^i and its closest and second-closest neighbors. Only matches with the closest neighbor are validated if the ratio between the two distances is below a threshold. This method often performs well in image matching, but it has several drawbacks. There is at most one match per query and the optimal threshold varies greatly from one query to the other. Moreover, the diversity of the whole database is not taken into consideration in the matching process, since only the first and second neighbors are considered. A variant on this criteria has been proposed in [2] and consists in averaging the distance to the second neighbor over several images when performing multi-image matching, for instance in the context of panorama stitching. Another possibility -when thresholding the distance to the nearest neighbor- is to keep only matches (a, b) for which a is also the nearest neighbor of b, see [5].

In different settings, the control of false matches has been taken into account, see [17, 11, 16]. But to the best of our knowledge, no generic procedure for the matching of local, SIFT-like features has been proposed beyond the already mentioned thresholds applied to the nearest neighbor.

In this contribution, we propose to validate matches between the query and candidate descriptors by rejecting casual matches, that is matches that can be produced by chance. Specifically, we make use of an *a contrario* methodology, first introduced in [6] and then applied, among other things, to grouping [7] and shape matching [16]. The principle of such approaches is to detect or match features when a certain *null hypothesis* is rejected.

The plan of the paper is as follows. In Section 2, we detail the keypoints and descriptors to be used, in a very similar way to [13]. In Section 3 a robust dissimilarity measure between features is introduced, based on the Earth Mover's Distance described in [18]. Section 4 is the main contribution of this communication, where we present a new matching criterion that is inspired from the *a contrario* methodology. It provides an adaptive threshold on the dissimilarity measure that allows multiple detections over a database. Experimental validations are performed in Section 5.

2. Features extraction

This section briefly presents our version of SIFT-like (see [13]) descriptors. Classically, a scale-space representation is used to detect and select interest points. Descriptors based on the distribution of gradient orientations are then built for each of these points.

Detection of interest points A "Laplace-Harris" detector is used to select high curvature structures, typically multi-scale corners and "blobs". First, the image I_0 is convolved with Gaussian kernels g_{σ_k} to

obtain its linear scale-space representation $\{I_{\sigma_k}\}$. Then, the local extrema in scale and space of $\{L_k\}$ - the normalized Laplacian operator response of $\{I_{\sigma_k}\}$ [10]- provide a set of possible interest points $\{(x_i, y_i, s_i)\}$ with their scale estimation $s_i = \sigma_k$. Finally, the multi-scale Harris [8] criterion is applied to eliminate edge structures which are redundant and not significant enough for the matching process.

Orientation assignment In order to achieve rotation invariance, up to two different orientations are given to each interest point. A circular histogram of gradient orientations is built from the neighborhood of each interest point. We then use an automatic histogram segmentation method proposed in [7] -that we adapted to circular histograms- to extract the modes and keep the two most significant ones (or only one if it is unique). For each mode the center of mass is computed, yielding oriented interest points. This orientation assignment procedure is more robust than selecting the extrema of the histogram (as in [13]) and is performed very quickly.

Descriptor design In the same manner as SIFT, the descriptor consists of histograms of gradient orientations, weighted by the gradient magnitude and computed for different subregions of a location grid. Each histogram is quantized to N bins ($N = 12$ by default) and normalized to have unit weight. Orientations are defined with respect to the reference direction (there is one descriptor per reference direction, thus one or two descriptors for each interest point).

We use a circular location grid divided into M sectors on a disk ($M = 9$ by default, see Figure 1). This is known to be more robust to rotations than square sectors, [15]. The size of the disk is proportional to the scale s_i to achieve scale invariance [13], and sectors are defined so that they contain the same number of pixels. Thanks to the central sector and angular splitting, the descriptor is robust to small angular or translation shifts. Nevertheless, it is important to define a dissimilarity measure robust to angular quantization, and also to local deformations that result in angular shifts in the histograms. This is the aim of the next section.

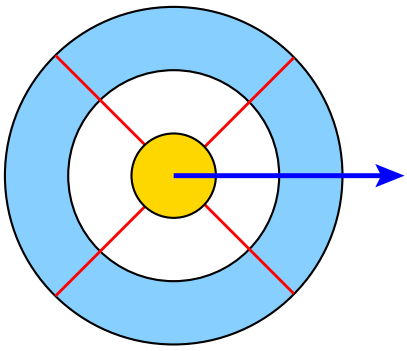


Figure 1. Location grid for the computation of orientation histograms

3. Dissimilarity measure

Bin-to-bin distances (such as the Euclidean, Mahalanobis or Manhattan distances) measure simply and quickly the dissimilarity between two vectors. However, these distances are obviously not robust to the orientation histogram quantization as shown in [13], where N is limited to 8 to make a compromise between angular quantization error and robustness to small angular shifts.

This quantization problem can be avoided by using a cross-bin distance, like the Earth Mover's Distance, proposed by Rubner [18] as a metric for color histograms. This distance can be seen as the solution to a "transportation" problem. In [12], Ling and Okada use an interesting variant of this measure (called "EMD-L_1") for SIFT descriptors as three dimensional histograms. However, this measure remains computationally too expensive to be applied efficiently to the matching problem when the number of descriptors increases (see Section 3.2). Moreover, using the same L^1 ground distance for the three dimensions of the histogram yields tricky parameter tuning. Indeed, mixing orientation and space makes transportation costs depend on the number of sectors M and on the angular quantization step N. We propose in the next paragraph a dissimilarity measure based on the Earth Mover's Distance for unidimensional and circular histograms which is specifically adapted to SIFT descriptors and has a low time complexity.

3.1. Earth Mover's Distance between normalized circular histograms

Consider two discrete circular (or periodic) histograms $f = (f[i])_{i=1\dots N}$ and $g = (g[i])_{i=1\dots N}$ with samples on N bins and normalized, in the sense that $\sum_{i=1}^{N} f[i] = \sum_{i=1}^{N} g[i] = 1$. In the non-circular case, it is well known [20] that the Earth Mover's Distance (EMD) between two unidimensional normalized histograms is equal to the L^1-distance between their cumulative histograms. In the periodic case, it can be shown that the EMD between f and g is the minimum in k of the L^1-distance between F_k and G_k, the cumulative histograms of f and g starting at the k^{th} quantization cell. That is, writing $d(f, g)$ for the EMD between f and g,

$$d(f,g) = \min_{k\in\{1,\dots,N\}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \mid F_k[i] - G_k[i] \mid \right\} , \tag{1}$$

where, $\forall\ k \in \{1, \dots, N\}$ (the definition is similar for G_k by replacing f by g),

$$F_k[i] = \begin{cases} \displaystyle\sum_{j=k}^{i} f[j] & \text{if} \quad i \geq k \\ \displaystyle\sum_{j=k}^{N} f[j] + \sum_{j=1}^{i} f[j] & \text{if} \quad i < k \end{cases} .$$

A descriptor a, as defined in Section 2, is made of M circular normalized histograms $(a_1, \dots, a_M)$. The dissimilarity measure between two descriptors a and b is then defined as the sum of the distances between a_m and b_m,

$$D(a,b) := \sum_{m=1}^{M} d(a_m, b_m). \tag{2}$$

We choose this dissimilarity measure because it is less sensitive to the context (change of background or occlusion) than using $\sum d(a_m, b_m)^2$ or $\max d(a_m, b_m)$.

3.2. Performance evaluation

The performances of this dissimilarity measure can be evaluated by comparing the descriptors of an image with the descriptors of the same image after an affine transform (an approximation for a limited

viewpoint change). For each descriptor of the original image, the best match among the descriptors of the transformed image is kept if the distance between these two descriptors is below a threshold. By varying this threshold value, we get a performance curve which shows the evolution of the number of correct matches according to the number of false matches. Four different curves are obtained (Figure 2) depending on the distance (EMD or Euclidean) and the quantization used for the histograms ($N = 12$ or 24). This experiment confirms the advantage of the EMD over the Euclidean distance in this context: the EMD yields a higher proportion of correct matches and is all the more efficient as the histogram quantization increases, which is obviously not the case with the Euclidean distance.

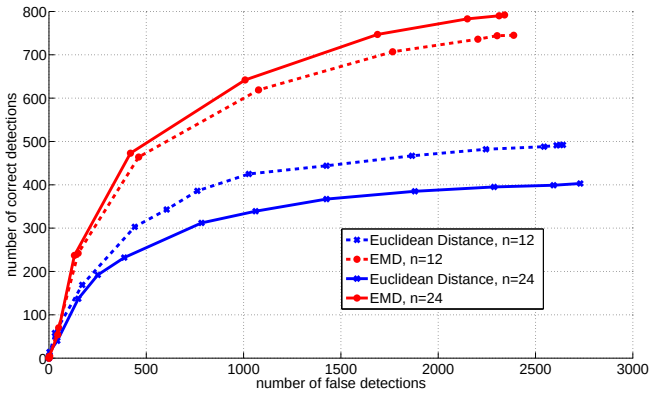


Figure 2. Comparison between the EMD and the Euclidean distances between descriptors, with two different histogram quantizations ($N = 12$ or $N = 24$).

Another interesting asset of our dissimilarity measure is that its time complexity is close to that of the Euclidean distance. Indeed, the time complexity of the naive EMD computation between M pairs of circular histograms with N bins is $O(MN^2)$, but a computation trick using the median reduces this time complexity to $O(MN\log(N))$. This is not far from the time complexity $O(MN)$ obtained with the Euclidean distance. In contrast, using the EMD-L^1 distance empirically [12] yields a time complexity of $O(M^2N^2)$.

4. Matching criterion

A dissimilarity measure between descriptors being defined, deciding whether a query descriptor a^i matches one or several descriptors from the database $\{b^1, \ldots, b^{N_B}\}$ boils down to the setting of a threshold on the distances. Ideally, this threshold should be set automatically and should depend on a^i and on the entire database. As explained in the introduction, one of the most popular matching criteria has been introduced by Lowe [13] and consists in thresholding the distance ratio between the first and second nearest neighbors of a^i in the database. If this ratio is below a threshold r, **the nearest neighbor** is matched with a^i, otherwise there is no match.

This criterion (that from now on we will refer to as NN-2) benefits from its simplicity and the fact that it is by far more robust than a simple threshold on distances. However, it has the following drawbacks:

– only the first and second nearest neighbors are considered to describe the complexity of the database;

– if a structure appears more than once in the database, it cannot be matched. This is a strong limitation whenever objects appear several times or have repetitive structures;

– the choice of an optimal r greatly depends on the experiment.

In the next section, we show how it is possible to overcome these difficulties by computing adaptive thresholds. Roughly speaking, the method rests on the rejection of matches that are due to chance.

4.1. *A contrario* methodology

The *a contrario* framework has been initially proposed by Desolneux *et al.* [6] in order to group low-level visual features. The basic principle is to detect groups of features that are very unlikely under the hypothesis that features are independent. In what follows, we call such a hypothesis a *null hypothesis*. The unlikeliness is ensured by controlling the average number of false detections. This generic approach has been applied with success to, among other things, the detection of alignments [6], contrasted edges, vanishing points, and grouping [7].

Recently, this methodology has been adapted to shape matching [16]. The main idea (also present in previous works such as [11, 17]) is again to reject matches that can happen "by chance". That is, relevant matches are detected *a contrario* as events contradicting the null hypothesis. Again, the null hypothesis is based on an independence assumption.

4.2. The background model

A candidate descriptor a^i being given, it is matched with b^j if $D(a^i, b^j)$ is small enough under the assumption that all b^js from the database follow a random model that is called a *background model*. This model should be seen as a model of generic descriptors. Remember that each descriptor a^i is made of M orientation histograms, $a^i = (a^i_1, \dots, a^i_M)$ and that the distance introduced in Section 3 is defined as $D(a^i, b^j) = \sum_{m=1}^{M} d(a^i_m, b^j_m)$. Two descriptors are all the more similar as distances between histograms are simultaneously small. The background model is defined through the independence of these distances, as in [16]. That is, the background model is any probabilistic model on a descriptor b such that, for all query descriptors a^i,

$\mathcal{H}_0$: *"$d(a^i_m, b_m)$ ($m \in \{1, \dots M\}$) are mutually independent random variables"*.

For a random descriptor following such a background model, the probability that the distance between a^i and b is smaller than δ can be written

$$\mathbb{P}\left(D(a^i, b) \leq \delta \,|\, \mathcal{H}_0\right) = \int_{-\infty}^{\delta} \underset{m=1}{\overset{M}{*}} p^i_m(x)\, dx \;, \tag{3}$$

where $*$ denotes the convolution product and p^i_m the density of the random variable $d(a^i_m, b_m)$. For each $i \in \{1, \dots N_A\}$ and each $m \in \{1, \dots M\}$, the laws p^i_m are empirically estimated over the database $\{b^1, \dots, b^{N_B}\}$. In other words, for each circular histogram a^i_m, one computes the distribution function of the distance $d(a^i_m, b_m)$ when b_m spans the m^{th} histogram of the descriptors in the database.

4.3. Number of false alarms

A match between a^i and an element b^j in the database is considered as meaningful and validated as soon as the distance $\delta = D(a^i, b^j)$ between them is much smaller than it can be expected to be under the hypothesis $\mathcal{H}_0$, *i.e.* as soon as the probability $\mathbb{P}(D(a^i, b) \leq \delta \,|\, \mathcal{H}_0)$ is small enough. Now, setting a threshold δ_i for each descriptor a_i is not an easy task. With the *a contrario* framework, the choice of these thresholds is replaced by a unique bound on the expectation of the global number of false alarms, which is more intuitive and handy. To this end, we introduce the following function of a^i and δ,

$$\mathrm{NFA}(a^i, \delta) = N_A\, N_B\, \mathbb{P}(D(a^i, b) \leq \delta \,|\, \mathcal{H}_0). \tag{4}$$

The value $\mathrm{NFA}(a^i, \delta)$ measures how likely it is that the distance between a^i and b is lower than δ under the hypothesis $\mathcal{H}_0$ on b. It also enables us to sort all the possible $N_A \times N_B$ matches and to evaluate their relevance. Thereby, a match between a^i and b^j is said to be ε-**meaningful** if $\mathrm{NFA}(a^i, D(a^i, b^j)) \leq \varepsilon$. With this definition, it is easy to prove that,
the expected number of ε-meaningful matches, when testing N_A queries against N_B candidates following the background model, is smaller than ε.

Along these lines, the threshold

$$\tilde{\delta}_i(\varepsilon) = \arg\max_{\delta} \left\{ \mathrm{NFA}(a^i, \delta) \leq \varepsilon \right\} \tag{5}$$

makes it possible to validate or reject the different correspondences between a^i and the elements of the database. A match is validated if $d(a^i, b^j) \leq \tilde{\delta}_i(\varepsilon)$. For each descriptor a^i, the threshold $\tilde{\delta}_i(\varepsilon)$ is automatically computed in function of the value ε.

Anticipating on the experimental section, let us underline the conceptual advantages of fixing ε to control the matches over other thresholds on distances. First, ε has the relatively intuitive meaning of a number of false alarms. Second, as said earlier, a single number yields thresholds that adapt to the query and the database. Last, the number of possible matches is not restricted.

5. Results

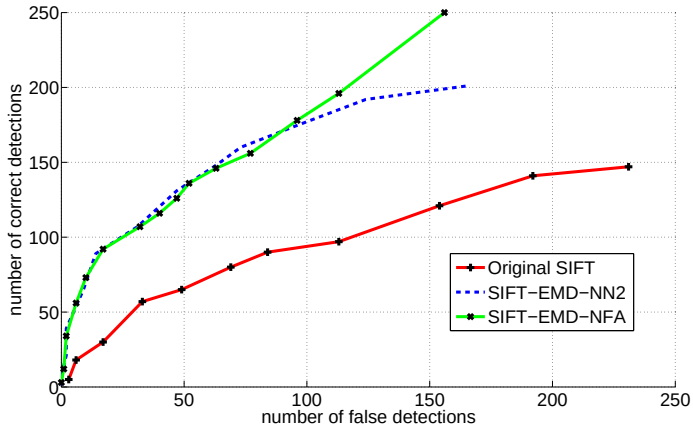
As defined in the previous section, the smaller the threshold ε, the more significant the selected matches. In practice, the threshold $\varepsilon = 10^{-1}$ appears to be satisfying for most experiments, since it limits the number of matches perceived as false detections -matches of descriptors belonging to different objects- even though these matches are not always false alarms, as they can represent really similar geometrical structures. In order to illustrate the efficiency of the proposed approach, we present various experiments comparing the following three matching procedures : "original SIFT" using D. Lowe's algorithm[1] (with the Euclidean distance between features and the NN-2 criterion), "SIFT-EMD-NN2" and "SIFT-EMD-NFA" using our SIFT-like descriptors (Section 2) and the EMD distance (Section 3), with the NN-2 and *a contrario* matching criteria respectively.

In the first experiment, two photographs of the same graffiti[2] taken with a very different viewpoint are matched. Figure 3(a) shows the matching result with the *a contrario* criterion, using $\varepsilon = 10^{-1}$. Figure 3(b) shows the number of correct detections against the number of false detections for the three different matching methods (3105x4123 original SIFT descriptors are computed and 3857x5193 with our scheme). The first observation concerns the superiority of both SIFT-EMD-NN2 and SIFT-EMD-NFA over the original SIFT. In this case the use of the EMD distance is efficient, similarly as in Section 3.2. The second observation is that the NN-2 and *a contrario* criteria perform quite similarly in this experiment. There are nor multiple occurrences neither repetitive structures and it seems sound to restrict matches to the nearest neighbor. This is a kind of sanity check for the matching criterion that we propose. Indeed, no restriction is made on the number of matches when using the *a contrario* approach. Nevertheless, the criterion adapts to the situation, and most matches are unique. Another observation is that when using high thresholds for both criteria (NN-2 and *a contrario*) to obtain a large number of matches, the *a contrario* criterion permits a better control of the number of false detections.

[1] The original Lowe's algorithm is kindly made available by its author on http://www.cs.ubc.ca/~lowe/keypoints/

[2] from the INRIA Graffiti dataset available at http://lear.inrialpes.fr/people/mikolajczyk/Database/index.html

(a) Result of matching with SIFT-EMD-NFA, at $\varepsilon = 10^{-1}$



(b) The number of correct detections is plotted against the number of false detections for the three methods: Original SIFT, SIFT-EMD-NN2, and SIFT-EMD-NFA.

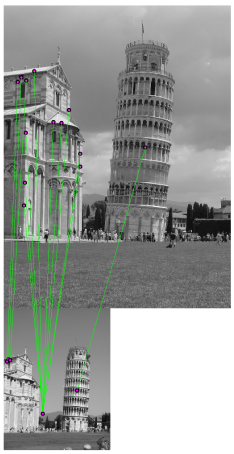
Figure 3. Matching two pictures from the INRIA Graffiti image dataset, with three different methods.

The two next experiments (Figure 4 and 5) illustrates one of the drawbacks of the NN-2 criterion described in the introduction, *i.e.* the difficulty of matching objects with repetitive structures.

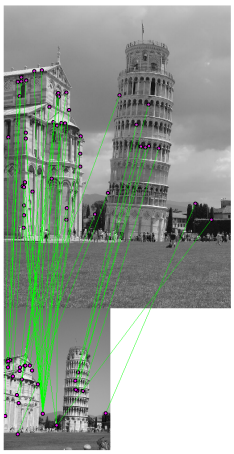
The first example with the White House front clearly shows that it is difficult to detect an object which has many repetitive structures with the NN-2 criterion (only 8 good detections among 41 at $r = 0.8$ using the original SIFT). Our approach avoids this problem by comparing each candidate to the whole database (68 good detections among 73 at $\varepsilon = 10^{-1}$ using the SIFT-EMD-NFA).

Two pictures of the leaning tower of Pisa and the front of the neighboring cathedral -with a little change in the viewpoint- are matched, using the same methods as in the previous experiment with several thresholds: the original SIFT results are shown at $r = 0.6$ (13 matches on 4(a)), 0.7 (45 matches on 4(b)), 0.8 (203 matches on 4(c)); the SIFT-EMD-NN2 results are shown at $r = 0.7$ (8 matches on 4(d)) and 0.8 (62 matches on 4(e)) -there is no match at $r = 0.6$; the SIFT-EMD-NFA results are shown at $\varepsilon = 10^{-2}$ (41 matches on 4(f)), 10^{-1} (104 matches on 4(g)), 1 (292 matches on 4(h)). The first observation is that, as expected, the NN-2 criterion fails to match the tower of Pisa, so that the matches obtained with a low threshold are mostly false detections (0 correct match with SIFT-EMD-NN2 at $r = 0.7$, and 3 correct matches with the original SIFT at $r = 0.6$). On the contrary, the *a contrario* criterion, with the same descriptors and dissimilarity measure, makes it possible to match these structures: all descriptors with $\varepsilon = 10^{-2}$ are matched with the correct object (the cathedral or the tower).

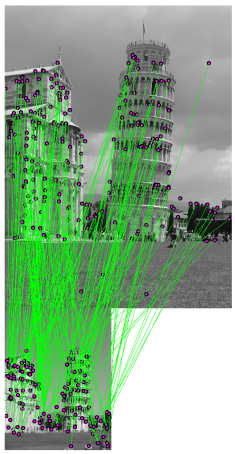
The second point highlighted by this experiment is the evolution of results as a function of the thresholds r and ε. On the one hand, it is difficult to choose an optimal result for the NN-2 criterion since the number of matches increases dramatically with r. Moreover, in order to obtain a few correct detections on the cathedral and on the tower, it is necessary to use a high threshold -$r = 0.8$ as recommended



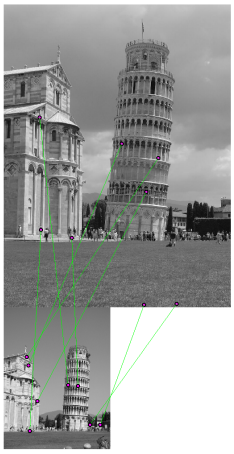
(a) Original SIFT, $r = 0.6$



(b) Original SIFT, $r = 0.7$

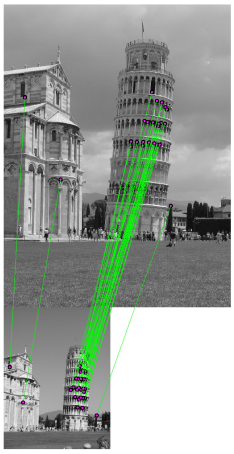


(c) Original SIFT, $r = 0.8$

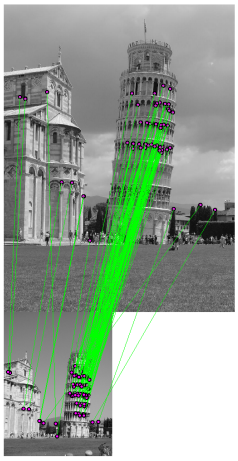


(d) SIFT-EMD-NN2, $r = 0.7$

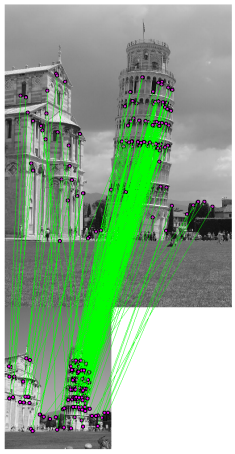
(e) SIFT-EMD-NN2, $r = 0.8$



(f) SIFT-EMD-NFA, $\varepsilon = 10^{-2}$



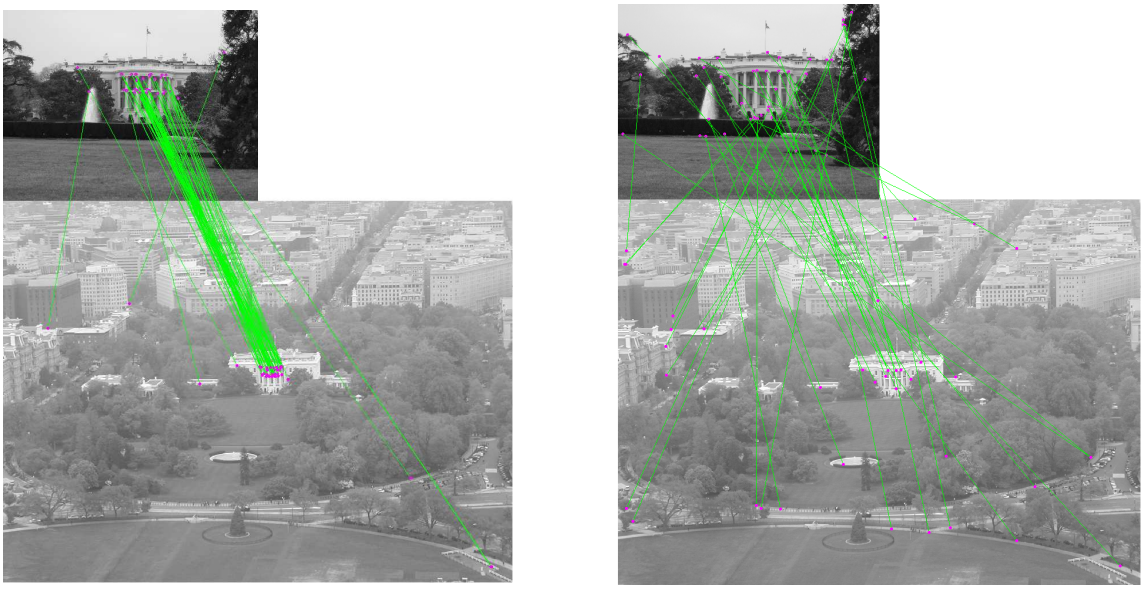
(g) SIFT-EMD-NFA, $\varepsilon = 10^{-1}$



(h) SIFT-EMD-NFA, $\varepsilon = 1$

Figure 4. Matching an object with repetitive structures: the tower of Pisa. Three different matching procedures are used: original SIFT, SIFT-EMD-NN2, and SIFT-EMD-NFA. The third method permits to match the tower even though it contains many repetitions.

in [13]- validating many false detections (nearly all the matches for SIFT-EMD-NN2 and roughly 75% of the matches for the original SIFT). On the other hand, when ε increases from 10^{-1} to 1 with the *a contrario* criterion, the number of matches increases from 104 to 292, but the number of false detections is limited to only a few points.



(a) SIFT-EMD-NFA at $\varepsilon = 10^{-1}$: 68 good detections among 73

(b) Original SIFT at $r = 0.8$: 8 good detections among 41

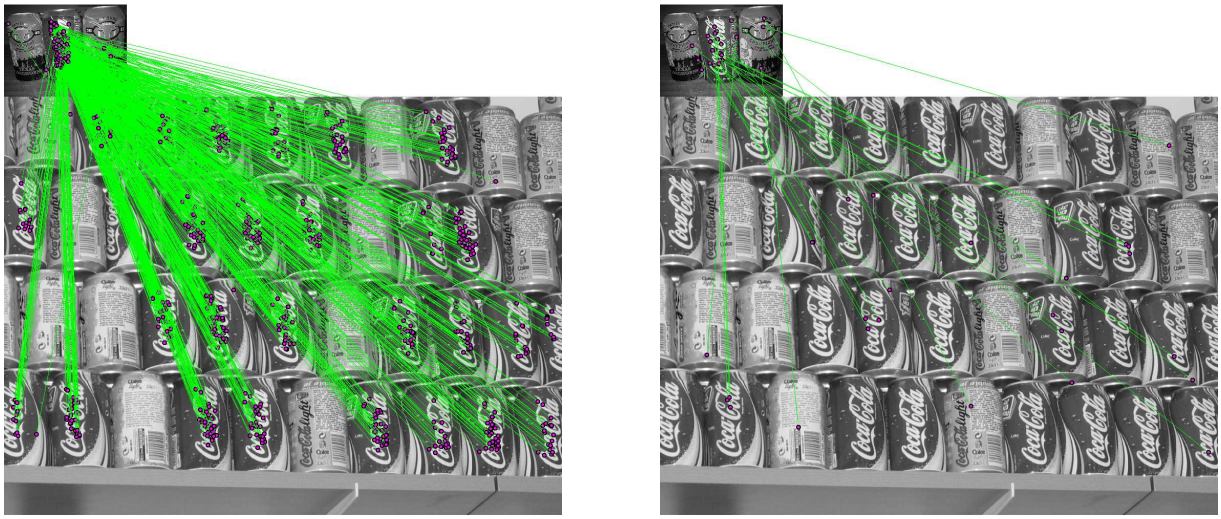
Figure 5. Matching an object with repetitive structures: the front of the White House. The *a contrario* criterion (Figure 5(a)) provides a higher number of good detections than NN-2 (fig. 5(b)).!

In the last three experiments, SIFT-EMD-NN2 and SIFT-EMD-NFA are compared in the case of multiple occurrences of an object, in order to illustrate the possibility offered by the *a contrario* matching.

The first one, Figure 6, shows the matching result between two pictures of cans. The logo of the central can in the query image appears several times in the second image. The result shown for $r = 0.8$ on Figure 6(b) confirms the difficulty of matching an object which appears several times when using the NN-2 criterion. Most matches are correct detections with $r = 0.8$ (20 matches out of 29) but it is not sufficient to detect the logo; when the threshold value r is increased, we obtain mostly false detections (only 33 out of 254 are correct between $r = 0.8$ and 0.9). In contrast, the *a contrario* criterion at $\varepsilon = 10^{-1}$ gives 1115 matches between the logos out of 1120 matches, using the same SIFT-EMD descriptors.

A similar experiment is shown on Figure 7 with a can of bean. Our SIFT-EMD-NFA approach selects automatically the correspondances between the cans in the second image with only few false detections (20 among 228 at $\varepsilon = 10^{-1}$), whereas the NN-2 criterion -using original SIFT or SIFT-EMD descriptors- hardly separates the good and the false detections (20 false detections among 55 with SIFT-EMD-NN2 at $r = 0.8$ and 301 among 378 at $r = 0.9$, 95 among 165 with original SIFT at $r = 0.8$).

In the last experiment, Figure 8, a query picture of a remote control is matched with a database of 6 pictures, the first two of which contain the same object. In order to detect this object in the database, the NN-2 criterion is applied when matching the query image with only one image at a time. By using



(a) SIFT-EMD-NFA, $\varepsilon = 10^{-1}$ (b) SIFT-EMD-NN2, $r = 0.8$

Figure 6. Multiple occurrences of a soda can in the database. The two matching criteria are used on the same SIFT-EMD descriptors: NN-2 matching with $r = 0.8$ and *a contrario* matching with $\varepsilon = 10^{-1}$.

$r = 0.6$ (Figure 8(a)), only one remote control is detected (5 matches in the second picture of the database) and a few false detections are obtained (6 matches with the "Rubik's cube" in the last picture). To obtain matches on the second remote control (in the first picture), $r = 0.8$ has to be used (Figure 8(b)) but the number of false detections is then very high (there are only 81 matches between the remote controls among 268). This also shows that the choice of the threshold r greatly depends on the similarity between the two objects and their context. Figures 8(c) (54 matches with $\varepsilon = 10^{-2}$) and 8(d) (116 matches with $\varepsilon = 10^{-1}$) shows the result obtained with the sames descriptors with the *a contrario* criterion. Mostly matches are between the remote controls (52 and 105 respectively).

6. Conclusion

In this contribution, we propose a procedure for the matching of local, SIFT-like descriptors. The procedure rests on a robust distance between descriptors and an automatic matching criterion. In contrast with most existing approaches, the criterion is not restricted to the nearest neighbor and allows multiple matches.

Several extensions of this work are foreseen. First, even though the computation of the proposed matching thresholds is not computationally demanding (it only requires to compute M convolutions for each query descriptor a^i), it cannot benefit in a straightforward way from fast nearest neighbor search schemes [13, 1]. We plan to adapt these by approximating the NFA using only a small number of candidate descriptors.

Another interesting point is that the matching methodology presented in Section 4 is completely generic and could be applied to other local descriptors, such as affine invariant descriptors described in [14]. We are currently working on the joint use of color and direction histograms as descriptors, within the same matching framework.

Next, we plan to take the global coherence of matches into account, as it is classical in object detection.

Here again, the same *a contrario* methodology can be used, in the same way as in [3]. In particular, this makes it possible to take the size and content of the database into account, as it is the case with the matching step presented in this paper. The adaptivity of the resulting object detection method will then be tested on very large databases, for instance through global image search over the Internet. Several recent applications such as [19] could benefit from searches for which the number of false detections remains controlled.

References

[1] J. Beis and D. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proc. CVPR*, pages 1000–1006, 1997. 11

[2] M. Brown, R. Szeliski, and S. Windner. Multi-image matching using multi-scale oriented patches. In *Proc. CVPR*, pages 510–517, 2005. 1, 2

[3] F. Cao, J. Delon, A. Desolneux, P. Musé, and F. Sur. A unified framework for detecting groups and application to shape recognition. *J. of Mathematical Imaging and Vision*, 27(2), 2007. 2, 12

[4] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *Proc. CVPR*, pages 220–226, 2005. 2

[5] R. Deriche, Z. Zhang, Q. Luong, and O. Faugeras. Robust recovery of the epipolar geometry for an uncalibrated stereo rig. In *Proc. ECCV*, pages 567–576, 1994. 2

[6] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *Int. J. Comput. Vision*, 40(1):7–23, 2000. 2, 6

[7] 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. 2, 3, 6

[8] C. Harris and M. Stephens. A combined corner and edge detector. *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988. 3

[9] A. Kushal and J. Ponce. Modeling 3D objects from stereo view and recognizing them in photographs. In *Proc. ECCV*, 2006. 1

[10] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Norwell, MA, USA. Kluwer Academic Publishers, 1994. 3

[11] 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. 2, 6

[12] H. Ling and K. Okada. An efficient Earth Mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):840–853, may 2007. 4, 5

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 1, 2, 3, 5, 10, 11

[14] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004. 11

[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. 1, 2, 3

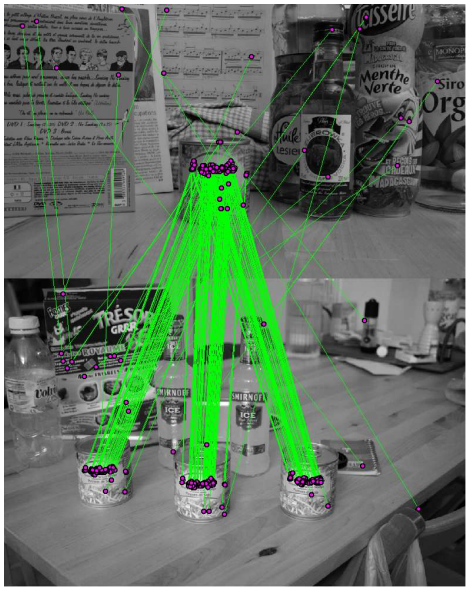
[16] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel. An a contrario decision method for shape element recognition. *Int. J. Comput. Vision*, 69(3):295–315, 2006. 2, 6

[17] C. Olson and D. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *IEEE Transactions on Image Processing*, 6(12):103–113, 1997. 2, 6

[18] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, 2000. 2, 4

[19] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism : exploring photo collection in 3D. In *ACM Transactions on Graphics (SIGGRAPH Proceedings)*, volume 25, pages 835–846, 2006. 12

[20] C. Villani. *Topics in optimal transportation*. American Math. Soc., 2003. 4



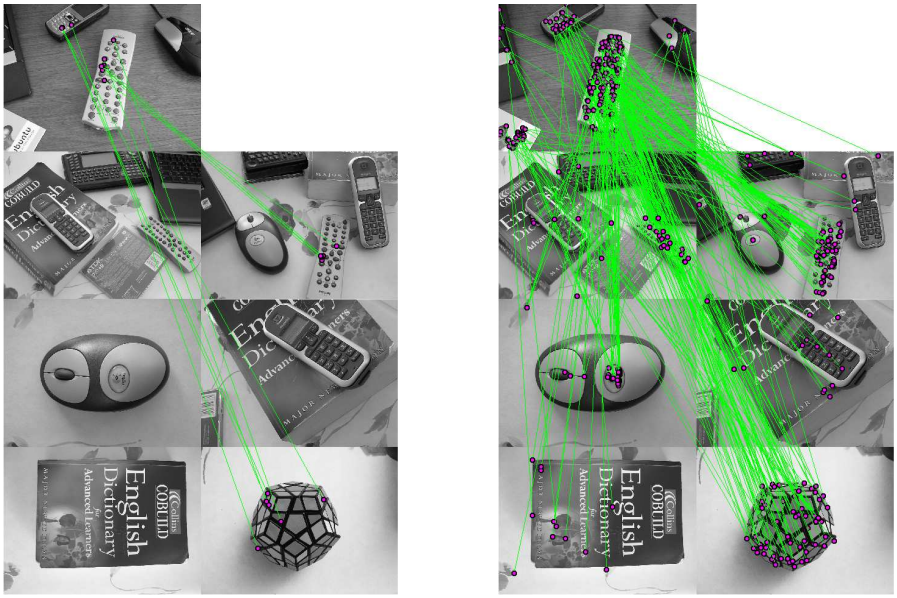
(a) SIFT-EMD-NFA à $\varepsilon = 10^{-1}$

(b) Original SIFT à $r = 0.8$

(c) SIFT-EMD-NN2 à $r = 0.8$

(d) SIFT-EMD-NN2 à $r = 0.9$

Figure 7. Multiple occurrences of a can of bean in the database. The a contrario matching criterion selects automatically the thresholds to match simultaneously descriptors of the query object with those of the 3 cans in the database (208 good detections among 228). The false detections rate is low contrary to NN-2 criterion (10% with SIFT-EMD-NFA at $\varepsilon = 10^{-1}$, 36% with SIFT-EMD-NN2 at $r = 0.8$, 80% with SIFT-EMD-NN2 at $r = 0.9$ and 59% with original SIFT at $r = 0.8$).



(a) NN2, $r = 0.6$ (b) NN2, $r = 0.8$

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

Figure 8. The remote control in the top picture is present in the first two pictures of the database. Figures (a) and (b) show the results of the NN-2 matching criterion (applied separately to each picture in the database, to allow multiple detections) for $r = 0.6$ (no detection in the first picture) and $r = 0.8$ (the remote control in the first picture is matched). Figures (c) and (d) show the results of the *a contrario* matching criterion for $\varepsilon = 10^{-2}$ and 10^{-1}, enabling us to match the two objects simultaneously with only a few false detections.