A STATISTICAL APPROACH TO THE MATCHING OF LOCAL FEATURES*

J. RABIN†, J. DELON†, AND Y. GOUSSEAU†

Abstract. This paper 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 or image matching. In practice, this matching step is often reduced to a specific threshold on the Euclidean distance to the nearest neighbor.

Our first contribution is a robust distance between descriptors, relying on the adaptation of the Earth Mover's Distance to circular histograms. It is shown that this distance outperforms classical distances for comparing SIFT-like descriptors, while its time complexity remains reasonable. Our second and main contribution is a statistical framework for the matching procedure, which yields validation thresholds automatically adapted to the complexity of each query descriptor and to the diversity and size of the database. The method makes it possible to detect multiple occurrences, as well as to deal with situations where the target is not present. Its performances are tested through various experiments on a large image database.

Key words. Statistical analysis of matching processes, local feature matching, dissimilarity measure, Earth Mover's Distance, a contrario.

AMS subject classifications. 62H35, 68T45, 68T10

1. Introduction. The matching of common structures between digital images is an important issue for a large number of computer vision applications: finding correspondences between images of the same scene [17], image classification [4], image and video retrieval [46, 45, 43], image stitching [5, 18], stereo vision [27, 11], object detection [49] and recognition [26, 16], and 3D object modeling [19]. One of the most classical approaches to this problem consists in using local features around interest points or regions. The locality of the features ensures robustness to occlusion or context change, while the coding of the features should be invariant or robust to various geometrical, photometric or radiometric changes. Numerous local approaches have been proposed in the literature, the exhaustive study of which is beyond the scope of the present paper. In two relatively recent comparative studies [30, 33], the SIFT descriptor [26] has been proven to be one of the most robust and invariant representation methods. As a result, the problem of finding correspondences between images often boils down to the matching of such local features. Nevertheless, whereas the extraction and representation of local descriptors has been thoroughly studied (see *e.g.* the references in [30]), their matching has not been the object of a systematic study. In practice, the matching step relies on simple but somehow limited procedures, as detailed further in the paper.

In many applications, this matching procedure is yet a crucial preliminary step. It can for instance be used as a pre-processing stage (before resorting to some geometric consistency algorithm like RANSAC [9, 43, 5] or some mean square error minimization [26]) for finding common objects between images. The matching step is at the core of many recent methods relying on image similarities, see e.g. [18, 16, 26, 9, 5, 11, 27, 43, 46, 1, 19, 32, 3, 50]. At this point, it is worth noticing that this matching step can serve to localize common structures between images, but

†LTCI CNRS, Telecom ParisTech, Paris, France.
{rabin,delon,gousseau}@telecom-paristech.fr

*Preliminary version of this work has appeared in proceedings of IEEE ICPR'08 [40, 41].

also to *decide whether a structure is present*. In fact, this is a crucial issue since a computer vision system has to deal with situations where the object of interest is not present. In such cases, it is of great interest to be able to limit the number of false matches, especially in the case of very large databases, see e.g. [46].

Now, as pointed out in [33],

> *Important aspects of matching are metrics and criteria to decide whether two features should be associated, and data structures and algorithms for matching efficiently.*

In the present paper we focus on the matching problem, and more precisely on two important steps:

- the choice of a dissimilarity measure between features;
- the choice of a matching criterion, used to decide which matches are valid.

Observe that we do not consider data structure optimization, such as approximate nearest neighbor approaches (see for instance [2]), or bag of features framework (*e.g.* [45]).

The dissimilarity measure should provide relevant comparisons between features and should be robust enough to cope with small variations of these features. The matching criterion should adapt itself to the complexity and diversity of the features. These two aspects (dissimilarity measure and matching criterion) are at the core of this paper. Our first contribution is a dissimilarity measure relying on the adaptation of the Earth Mover's Distance [44] to circular histograms. This measure is proven to behave well with respect to histogram quantization and to outperform classical bin-to-bin distances in the framework of local features comparison. Our second and principal contribution is a matching criterion relying on a statistical framework. This criterion provides thresholds which adapt to the complexity of the features and allow multiple detections over a database, while controlling the total number of matches. In particular, this criterion deals well with situations where we do not know whether the object of interest is present, as will be demonstrated by a specific experimental protocol. Conference proceedings versions of this work have appeared in [40] (dissimilarity measure) and [41] (matching criterion).

1.1. Related work.

Dissimilarity measure. As previously mentioned, the choice of a metric is fundamental for the matching of local features. Indeed, the matching criteria that are commonly used (as detailed in the next paragraph) directly rest on a thresholding of the similarity score.

The most classical local features, such as SIFT [26], reduce the geometrical information to one-dimensional circular histograms of local gradient orientations. Usually, "bin-to-bin" distances, such as the Euclidean distance [26, 30, 46, 43, 33] or the χ^2 distance [3, 49], are considered as the simplest way to quickly measure the dissimilarity between such histograms at a low computational cost. The term "bin-to-bin" refers to the fact that, to compare two histograms, each bin of the first histogram is compared exclusively to the bin of same rank of the second histogram. These distances are obviously not robust to histogram quantization. Therefore, one has to choose the number of bins to reach a good compromise between discriminative power and robustness of the comparison. For instance, the number of bins of gradient orientation histograms for original SIFTs [26] is limited to $N = 8$.

Bin-to-bin distances are intrinsically limited since they only compare the intensity of modes and not their relative positions. This limitation can be overcome by using cross-bin distances. A classical cross-bin distance, the Mahalanobis distance, requires

the computation of the covariance matrix of descriptors over a training database. This distance has been used in the context of local features comparison, but without meaningful gain: as pointed out in [33], *although the Mahalanobis distance is more general than the Euclidean distance, most relative performances were not modified.* Other cross-bin distances, such as the so-called quadratic distance [36] or the diffusion distance [24], rely on smoothings of the histograms. These methods necessitate non-trivial parameter adjustments, such as the choice of a kernel or the scale of smoothings.

The Earth Mover's Distance, proposed by Rubner *et al.* [44] and often used to compare image signatures, is probably one of the most elegant and robust ways of comparing histograms. However, it is computationally far more expensive than bin-to-bin distances when the dimension of histograms becomes strictly greater than one. A nice variant of this distance has been proposed by Ling *et al.* [25] as a way to speed up the comparison. This distance is applied in [25] to the comparison of local features. However, this measure remains too expensive to be applied to the matching problem when the number of features increases (as will be detailed in Section 2) and does not explicitly address the circularity of orientation histograms.

These limitations led us to propose a dissimilarity measure, called CEMD, specifically designed to compare one-dimensional circular histograms (see Section 2). This measure, based on the Earth Mover's Distance, is computationally efficient and deals with circular histograms, such as orientation histograms in SIFT descriptors [26]. In the experimental section, CEMD is used for the comparison of SIFT descriptors. This distance is shown to be more robust to quantization effects and small geometric perturbations than bin-to-bin distances. The first version of this distance was published as a research preprint in [39], and then as a conference proceeding in [40]. Independently, another interesting proposition of one-dimensional and circular distance was later published in [38].

Matching criterion. In order to introduce the most classical criteria that are used to match local descriptors, it is useful to give some vocabulary and notations that will be used throughout this paper. We consider a situation where one seeks for correspondences between N_Q *query descriptors* $\{a^i\}$ and a database of N_C *candidate descriptors* $\{b^j\}$. We assume that distances have been computed between each a^i and each b^j. This step can sometimes be replaced by approximate allocation algorithms, as in [2]. Two different criteria are used in practice to validate matches, as detailed in [30, 33], both relying on user-selected thresholds. Ideally, these thresholds should be set automatically and should depend on both the query and candidate descriptors.

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. Usually, matches are restricted to the nearest neighbor [1, 18] 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). 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 [16], $K = 4$ in [5] for image stitching, and K between 5 and 10 in [43]. 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 [15]), when dealing with objects having repetitive parts, such as buildings (this issue is

studied in [50]), 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) makes it particularly difficult to set the right threshold for a particular application.

In order to reduce the variability of the chosen threshold, Lowe [26] 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. 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 [26], $r = 0.6$ in [46], $r = 0.95$ in [9], or r between 0.56 and 0.7 in [33] for instance. In practice, the NN-DR criterion behaves very well (and in particular significantly better than the NN-DT criterion as shown in [33]) 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. This situation will be considered in the experimental section. 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.

Several variants of these matching criteria have been proposed. In [6], it is suggested to adapt the NN-DR criterion by averaging the distance to the second neighbor over several images for panorama stitching. In [11], a variant of NN-DT consists in keeping only matches (a, b) for which a is also the nearest neighbor of b. More specific matching criteria with geometric constraints have been proposed (see e.g. [8, 19, 32, 50]), but to the best of our knowledge, no generic procedure for the matching of local, SIFT-like features has been proposed beyond the aforementioned thresholds on distances.

In this paper, we propose in Section 3 an alternative matching criterion relying on 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 [23, 22]. Specifically, we resort to an *a contrario* methodology, first introduced in [12] and then applied, among other things, to shape matching [35]. 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. Observe that a different *a contrario* approach has been recently proposed in [7] in order to match local descriptors composed of a collection of gradient orientations on normalized patches.

1.2. Outline. In Section 2, we introduce the new transportation distance for comparing local descriptors, CEMD. Then in Section 3, the matching criterion relying on the *a contrario* methodology is introduced. In Section 4 the advantages of both contributions over classical approaches are demonstrated on an image database through the use of several experimental protocols.

2. Dissimilarity Measure. In this section, we introduce a dissimilarity measure designed to compare circular histograms (such as orientation histograms). This measure is a generalization of the classical Earth Mover's Distance to the circular case. It can also be seen as an application of the statistical Mallows distance to probability distributions on the unit circle. We then explain how to apply this measure to compare local, SIFT-like features.

2.1. Circular Earth Mover's Distance (CEMD). Consider two discrete circular[1] histograms $f = (f[i])_{i=1\dots N}$ and $g = (g[i])_{i=1\dots N}$, sampled on N bins. Both histograms are supposed to be normalized, that is, $\sum_{i=1}^{N} f[i] = \sum_{i=1}^{N} g[i] = 1$.

The Earth Mover's Distance between f and g is then defined in [44] as

$$\text{EMD}(f,g) := \min_{(\alpha_{i,j})\in\mathcal{M}} \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{i,j} c(i,j), \tag{2.1}$$

where

$$\mathcal{M} = \{(\alpha_{i,j});\ \alpha_{i,j} \geq 0,\ \sum_j \alpha_{i,j} = f[i],\ \sum_i \alpha_{i,j} = g[j]\}$$

and where $c(.,.)$ is a ground distance between bins. For circular histograms, this ground distance can naturally be chosen as

$$c(i,j) = \frac{1}{N}\min(|i-j|, N-|i-j|),\ \forall (i,j) \in \{1,\dots N\}^2. \tag{2.2}$$

The distance $\text{EMD}(f,g)$ can be understood as a transportation cost. The value $c(i,j)$ measures the cost of moving a unit mass from bin i to bin j, and $\alpha_{i,j}$ is the amount of mass carried from i to j. This definition can be used in any dimension. However the computation of the Earth Mover's Distance involves heavy computations when the dimension of histograms becomes larger than two. Note that this distance is known by statisticians as the Mallows distance between probability distributions [20], and is also one of the Monge-Kantorovich distances, defined in the mass transportation theory [47].

Now, for non-circular and one-dimensional histograms, when the ground distance is chosen as $c(i,j) = \frac{1}{N}|i-j|$, it is well known (see for instance chapter 2 in [47] for a proof) that $\text{EMD}(f,g)$ equals $\|F-G\|_1 = \frac{1}{N}\sum_{i=1}^{N}|F[i]-G[i]|$, where F and G are the cumulative histograms of f and g, defined as

$$F[i] = \sum_{j=1}^{i} f[j], \quad G[i] = \sum_{j=1}^{i} g[j]. \tag{2.3}$$

The generalization of this formula to circular histograms is not straightforward. Indeed, if f is a circular histogram, one can build as many cumulative histograms as there are bins in f, since any bin can be chosen as a starting point to cumulate the histogram. More precisely, for each k in $\{1,\dots,N\}$, we define F_k, the cumulative

[1] Circular means that the first and the last bins of the histogram are neighbors.

histogram of f starting at the k^{th} quantization bin, by

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

Cumulative histograms G_k are defined in a similar way by replacing f by g. Now, it can be shown that the (circular) Earth Mover's Distance between the circular histograms f and g equals

$$\textsc{cemd}(f,g) = \min_{k \in \{1,\dots N\}} \|F_k - G_k\|_1. \tag{2.4}$$

A proof of this result can be found in [42]. This means that the distance $\textsc{cemd}(f,g)$ 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 bin. This formula is equivalent to the one proposed in lemma 4 in the work of Werman *et al.* [48] on point matching on a circle.

2.2. Comparing SIFT-like features. In this section, we first briefly recall the classical way to compare SIFT-like features by using bin-to-bin distances, and then explain how to apply the CEMD to the comparison of such local features.

Let us recall [26, 30] that a SIFT-like descriptor a consists of M circular histograms a_m of gradient orientations, weighted by the gradient magnitude and computed for different subregions of a location grid around an interest point. Thus, the comparison of two descriptors a and b boils down to the comparison of circular histograms a_m and b_m. We suppose here that each histogram is quantized to N bins and that the whole descriptor $a = (a_1, \dots, a_M)$ is normalized to have unit weight [26].

2.2.1. Bin-to-bin distances. The most classical way to compare SIFT-like descriptors is simply to use the L^p distance as in Formula (2.5), usually with $p = 2$ (Euclidean distance) [26]. Applying this distance requires a global L^p normalization of descriptors a and b. Other bin-to-bin distances that are used to compare local features include the χ^2 distance, as in [49] or the Jeffrey divergence. The definitions of these distances in the framework of SIFT-like descriptors are recalled in Formula (2.6) and (2.7) respectively.

$$\mathrm{D}_{L^p}(a,b) := \left(\sum_{m=1}^{M} \sum_{i=1}^{N} \mid a_m[i] - b_m[i] \mid^p \right)^{\frac{1}{p}} \tag{2.5}$$

$$\mathrm{D}_{\chi^2}(a,b) := \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{(a_m[i] - b_m[i])^2}{a_m[i] + b_m[i]} \tag{2.6}$$

$$\begin{aligned} \mathrm{D}_J(a,b) := \quad & \sum_{m=1}^{M} \sum_{i=1}^{N} a_m[i] \log \left(\frac{2\, a_m[i]}{a_m[i] + b_m[i]} \right) \\ & + b_m[i] \log \left(\frac{2\, b_m[i]}{a_m[i] + b_m[i]} \right) \end{aligned} \tag{2.7}$$

2.2.2. Applying CEMD to local features. Two descriptors $a = (a_1, \ldots, a_M)$ and $b = (b_1, \ldots, b_M)$ can be compared by applying CEMD to each pair of histograms a_m and b_m. Now, observe that a_m and b_m can have different total masses, while Formulas (2.1) and (2.4) assume that both histograms have equal masses.

There is no obvious way to generalize the Earth Mover's Distance to histograms of different total masses. One possibility would be to normalize both histograms. In practice, however, it is by far more robust to globally normalize SIFT-like descriptors to unit weight (as shown in [26]) than to normalize each histogram a_m individually. Another possibility, often proposed in papers dealing with EMD [44], is to transport the histogram with the smallest mass onto the other one, and to discard the supplementary mass. This solution is obviously not satisfactory for comparing SIFT-like descriptors: empty histograms, which correspond to flat zones in the image, would be very close to all histograms, in particular histograms of very structured zones.

In this paper, we choose to simply apply Formula (2.4) to non-normalized histograms. The resulting distance is not a metric, but is easy and fast to compute, and yields excellent results on SIFT-like descriptors.

Let us mention that an interesting alternative has been independently proposed in [38], where the difference of masses between two histograms, weighted by the maximum ground distance between two bins, is added to the definition of the distance. This boils down to add mass to the smallest histogram so that masses of both histograms become equal, and to decide that the distance of this mass to all masses in the larger histogram is $max(c(i, j))$. In this paper, a fast scheme for the computation of the distance is proposed, based on the truncated circular L_1 ground distance. Contrary to Equation (2.2), the ground distance is considered as constant for any moves exceeding two bins (among N). This choice enables to speed up the computation time, but is somewhat arbitrary, the definition being not invariant to the quantization effects.

The last step in order to define a distance between descriptors is to combine CEMD-distances corresponding to different subregions (different values of m). Here we choose to use the following distance between two descriptors,

$$D_{\text{CEMD}}(a, b) := \sum_{m=1}^{M} \text{CEMD}(a_m, b_m). \tag{2.8}$$

Other dissimilarity measures could have been chosen (such as $\sum \text{CEMD}(a_m, b_m)^2$ or $\max \text{CEMD}(a_m, b_m)$). However, we observed experimentally that the distance (2.8) is more robust.

2.2.3. Implementation and computational cost. Let $X_k[i] = F_k[i] - G_k[i]$ be the difference of the cumulative histograms computed in Formula (2.4). X_k can be written as a function of X_1,

$$X_k[i] = \begin{cases} X_1[i] & \text{if } k = 1 \\ X_1[i] - X_1[k-1] & \text{if } i \geq k > 1 \\ X_1[i] - X_1[k-1] + X_1[N] & \text{if } i < k. \end{cases}$$

Consequently, computing CEMD does not require to compute the k different cumulative histograms F_k and G_k in the circular case. Note that $X_1[N]$ is equal to zero when the two histograms f and g have the same weight. Compared to the classical L^1 bin-to-bin distance, the only required extra computation is the minimization over k of $\|X_k\|_1$, the L^1 norm of X_k. It follows that the complexity of the CEMD computation

is approximately N times the complexity of the Euclidean distance computation, where N is the number of bins of each local histogram ($N = 8$ for classical SIFT).

Observe that in [25], Ling and Okada present an interesting variant of the Earth Mover's Distance, called EMD-L_1, designed to speed up the computation of EMD in the multidimensional case. Among their experiments, they show an application of their distance to SIFT descriptors, considered as three dimensional histograms (coding both orientation and localization). Nevertheless, this distance remains too expensive to be applied to large descriptors databases: EMD-L_1 is empirically 720 times slower than computing the Euclidean distance, according to Table VII in [25]. As an order of magnitude, performing the same evaluation as the one to be done in Section 4 with EMD-L_1 would require more than one year on a standard 2.5 GHz computer.

3. *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* methods and fix matching thresholds that ensure the rejection of casual matches.

3.1. *A contrario* methodology. The general principles of *a contrario* methods have first been proposed by Desolneux *et al.* [12] in order to detect alignments. The same principles have then been applied to a wide variety of computer vision tasks, such as the detection of contrasted edges, good continuation, vanishing points, rigid transforms or motion, see the recent monograph [14]. These approaches rely on a hypothesis testing framework. The main idea, presented in a generic manner in [13], is to detect groups of features that are very unlikely under the hypothesis that these features are *independent*. This hypothesis is called the *null hypothesis* in this paper. Loosely speaking, the detected groups are those that cannot result from chance. The second important point of *a contrario* methods is that to compute the degree of unlikeliness of a group, one predicts the expected number of groups under the null hypothesis, and not the (generally intractable) probability of existence of the group, see [14].

Recently, this methodology has been adapted to the problem of shape matching [35]. Again, the main idea is to reject matches that could have occurred by chance. Similar ideas are present in studies dealing with the statistical analysis of matching processes [23, 22, 37], in particular when predicting the number of false alarms. One difference is that these studies are more elaborated, but also less generic, because the analysis of the matching process relies on some shape model. When using *a contrario* approaches, one only needs a distance and an independence assumption (the null hypothesis) to validate matches. In the next two paragraphs, this methodology is adapted to the matching of SIFT-like features.

3.2. 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 (L^2, L^1 or χ^2), as well as for the CEMD-based distance proposed in formula (2.8).

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 $d(a_m^i, \mathbf{b}_m)$.

DEFINITION 3.1. *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 3.2. *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*

$$f_i(\delta) := \mathbb{P}\left(D(a^i, \mathbf{b}) \leq \delta \,|\, \mathcal{H}_0^i\right) = \int_0^\delta \underset{m=1}{\overset{M}{*}} p_m^i(x)\, dx\,. \tag{3.1}$$

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 (3.1), 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.

3.3. 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 3.3. *For a given $\varepsilon > 0$, a match between $\{a^i\}$ and a candidate descriptor $\{b^j\}$ is said to be ε-meaningful if*

$$f_i\left(D(a^i, b^j)\right) \leq \frac{\varepsilon}{N_Q N_C}, \tag{3.2}$$

where the function f_i is defined by Formula (3.1), 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

$$\delta_i(\varepsilon) = \arg\max_\delta \left\{ f_i(\delta) \leq \frac{\varepsilon}{N_Q N_C} \right\}. \tag{3.3}$$

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:

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.

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 [14]. 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 [31].

In practice, for a fixed ε and for each descriptor a^i we perform the several steps summed up in the table 3.1 to achieve the matching procedure.

TABLE 3.1
A Contrario (AC) matching procedure.

Algorithm 3.1 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 (3.1);
4) computation of threshold $\delta_i(\varepsilon)$ using Formula (3.3);
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 (3.3). 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 (3.3). Following the classical framework of a contrario methods [14], 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[14], the influence of ε is better expressed on a logarithmic scale. We provide a quick proof of this interesting property in the appendix.

4. Experimental results. In this section, several experiments are performed on an image database to illustrate the performances of both the dissimilarity measure and the matching criterion introduced in this paper. These experiments are performed on images modified by synthetic degradations (affine transformation and noise). We introduce several experimental protocols to illustrate the behavior of the proposed matching method in cases of single or multiple occurrences of the structure of interest, as well as in the presence of distractors.

4.1. Experimental setup.

4.1.1. Local features. In this paragraph, we briefly describe the local features that are used for the experiments. These are obtained in a very similar way to the classical SIFTs [26]. We did not use the original SIFT code in order to be able to vary the quantization step in the building of orientation histograms, as needed in Section 4.2.

The first step consists in detecting interest points in the image. These are first detected as extrema of the normalized Laplacian operator [21] over the linear scale-space of the image. Then a geometric criterion (scale-adapted Harris cost function [28]) is used to eliminate points lying on edges, and validate more complex structures like multi-scale corners. This results in a set of interest points with their corresponding scales.

The second step consists in building a local descriptor for each interest point detected, depending on its scale. Following the original methodology, we first build a histogram of gradient orientations in a neighborhood of each point. The neighborhood is centered on each point location and its size is proportional to the scale of the point. We extract the main orientations of this histogram with the non parametric analysis proposed in [10]. A polar localization grid as in [30] is then used to segment the neighborhood into M non overlapping regions. The grid is aligned with the reference orientation, as it is illustrated in Figure 4.1(a). We chose to use $M = 9$, providing a central region, 4 regions on a first ring and 4 more regions on a second ring. For each of these regions, a circular histogram a_m of gradient orientations with respect to the reference direction is built, the interval $[0, 2\pi[$ being quantized into n bins ($n = 8$ or 12 in the following experiments).

A SIFT descriptor is hence composed of the collection of these M circular histograms: $\{a_1, \ldots, a_M\}$.

4.1.2. Image database. Performances of both the dissimilarity measure and the matching criterion introduced in this paper are evaluated on approximately 3.10^6 descriptors, extracted from a set of 732 generic images[2]. We use this database because there is no large database of natural images that would be standard for evaluating

[2]Images available at: `http://www.tsi.enst.fr/~rabin/matching/`

(a) Localization grid (b) Affine transform

FIG. 4.1. *Left: Polar localization grid used to extract histograms of gradient orientation. Right: Affine transform applied to an image before adding noise.*

performances of feature matching. For instance, the classical INRIA database used in [30] only contains 8 different scenes. The size of the database we use is in the same order of magnitude as the one used in the evaluation paper [33], containing 100 query objects and 535 irrelevant images which constitute a 10^5 feature set. In this paper as in ours, an exhaustive feature comparison is performed. The use of such a dataset is of great importance, because performances can vary very much from an image to another. Observe also that using much bigger datasets to perform *exhaustive comparisons* would require quite heavy computing facilities. To the best of our knowledge, the present paper and [33] propose to date the largest scale experimental setup to evaluate a matching procedure.

4.1.3. Experimental protocols. We use several protocols to illustrate the versatility of the proposed matching criterion: ability to detect a structure when we know it is present exactly once, ability to decide whether the structure is present or not and ability to detect multiple occurrences.

The first protocol, called $A \to A'$, consists in matching keypoints between an image A and an image A' obtained by applying an affine transform (see Fig. 4.1(b)) and adding Gaussian noise to A (with a standard deviation $\sigma = 5$ for 8-bit images). Note that the authors of [34] showed that in the case of affine transform, the *tilt* parameter is the most critical for SIFT descriptors. More precisely, it is said that SIFTs can be considered roughly invariant to affine transform for which the tilt parameter is approximately less than 2. In order to illustrate the improvement provided by CEMD with perturbations for which SIFTs are not fully invariant, we hence use a tilt of 2.5. A match is declared false (*i.e.* a false positive) or correct (*i.e.* a true positive) depending on some spatial tolerance. More precisely, and following the protocol of [30], a match between a and b is considered as correct if the overlap error is below 50 percent. The overlap error between a and b is defined from the ratio between the area intersection and the area union of the corresponding SIFT regions in the image A, respectively R_a and R_b:

$$1 - |R_a \cap R_b| / |R_a \cup R_b| \,,$$

where $|R|$ is the area of the region R. This classical protocol, $A \to A'$, measures very simply the behavior of a matching procedure when two images containing exactly the same "objects" (before and after some transformations) are compared.

Now, many real computer vision systems involving a matching step have to deal with situations in which the target is not always present (*e.g.* the search of an object in an image database). In order to estimate matching procedures in such situations, we introduce another protocol called $A \to \{^{A'}_{B}$. In this protocol, the image A is

first compared with the modified image A' and then with an image B, independent of A (the next image in the database presented in Section 4.1.2). Of course, both comparisons are made using the same thresholds. Correct and false matches between A and A' are defined in the same way as in the protocol $A \to A'$. Meanwhile, all matches between A and B are considered as false matches. The total number of false matches is the addition of false matches in A' and B. A matching procedure should be able to match A and A' without finding too many correspondences between A and B.

In Section 4.3.2, protocol $A \to \{_B^{A'}$ will be extended by replacing B by the entire database introduced in Section 4.1.2. In Section 4.4, another protocol will be introduced to test the ability to detect multiple occurrences.

4.1.4. Performance evaluation. As is usually done, for each matching experiment between an image and its distorted version, a ROC curve shows the ratio of correct matches as a function of the ratio of false matches for different values of the matching threshold. More precisely, for a given threshold, the ratios of correct matches and false matches are defined as

$$\begin{cases} \text{correct matches ratio} = \dfrac{\#\text{correct matches}}{\#\text{possible matches}}, \\ \text{false matches ratio} = \dfrac{\#\text{false matches}}{\#\text{total number of matches}}. \end{cases}$$

Such a curve can be obtained for each image of the database. In order to evaluate the performances of different matching procedures (distances and criteria) on the whole database, thorough comparisons and analyses are made in the next sections, relying on these ROC curves.

4.2. Evaluation of the dissimilarity measures. We compare here the performances of the usual L^1 (Manhattan) distance, L^2 (Euclidean) distance, Jeffrey divergence, and χ^2 distance with the performances of the proposed Circular Earth Mover's Distance (CEMD). Since our purpose in this paragraph is not to evaluate matching criteria, we choose to use a simple threshold on distances restricted to the nearest neighbor (that is, criterion NN-DT) with the $A \to A'$ protocol. The comparison is performed for two quantization levels ($N = 8$ and $N = 12$) of the circular histograms.

Some images from the database and their associated ROC curves are shown in Fig. 4.2. For the sake of clarity, only CEMD, L^1 and L^2 distances are represented on these examples, respectively in red, blue and green continuous lines, for the value $N = 12$. We see on these curves that results can be quite different from one image to the other, even though CEMD shows better results than other distances.

Performances of the various distances are thus evaluated on the complete database. We follow the classical protocol used for image retrieval evaluation, see e.g. [44], and draw average performance curves to evaluate the ability of a given distance to retrieve correct information first. Average ROC curves show the average ratio of correct matches as a function of the ratio of false matches. The average correct matches ratio is defined (see (4.1)) as the average of correct matches ratio for the same given false matches ratio with each query image A_i, weighted by its number of descriptors $N_{Q,i}$, so that the larger the number of descriptors in an image, the greater

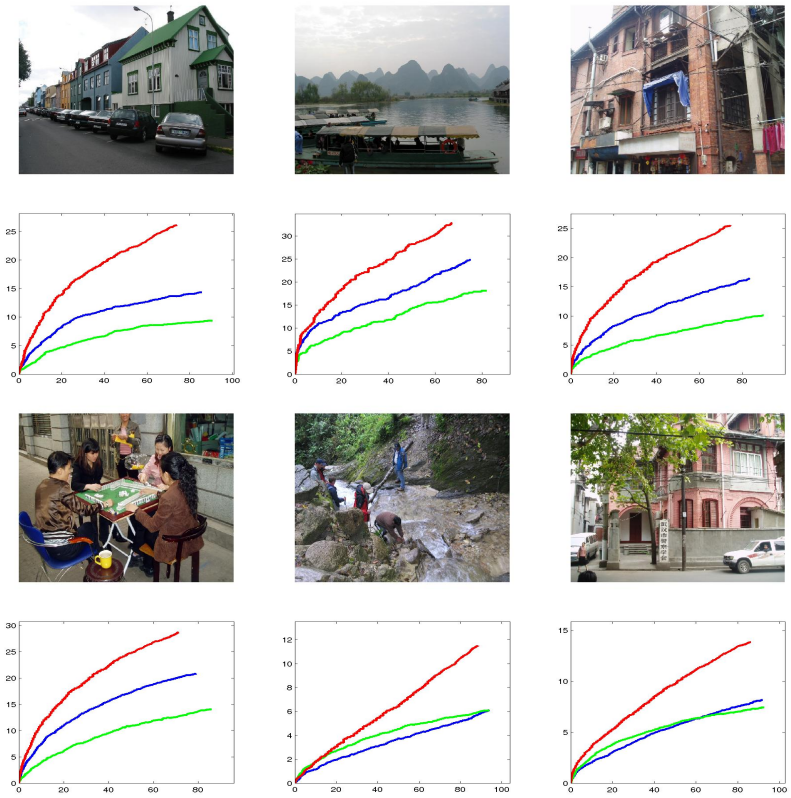


FIG. 4.2. *Six sample images from the database and the corresponding ROC curves. The red curve corresponds to* CEMD*, the blue one to the* L^1 *distance and the green one to the* L^2 *distance.*

its weight in the final average ROC curve.

$$\text{average correct matches ratio} = \frac{1}{\sum_{i=1}^{732} N_{Q,i}} \sum_{i=1}^{732} \left(N_{Q,i} \frac{\#\text{correct matches}(A_i)}{\#\text{possible matches}(A_i)} \right) \tag{4.1}$$

Consequently, for each distance defined in Section 2.2, performances are evaluated on the database (involving approximately 25.10^9 descriptor comparisons). Observe that curves are quite smooth because of this large number of comparisons.

Fig. 4.3 clearly shows the advantage of CEMD for all quantization choices. As one could expect, this measure deals well with the geometric deformations applied to each image which induce slight shifts in orientation histograms. Moreover, one observes that increasing N increases the quality of the matching when using CEMD. The number of bins is therefore only driven by computational complexity. In contrast, this is not the case for classical bin-to-bin distances, for which using too many bins yields inefficient comparisons between histograms. The average ROC curve in the case of the $A \to \{^{A'}_{B}$ protocol shows a similar behavior and is omitted for brevity. We will see in the next paragraphs that, in contrast, matching criteria behave differently depending on the matching protocol.

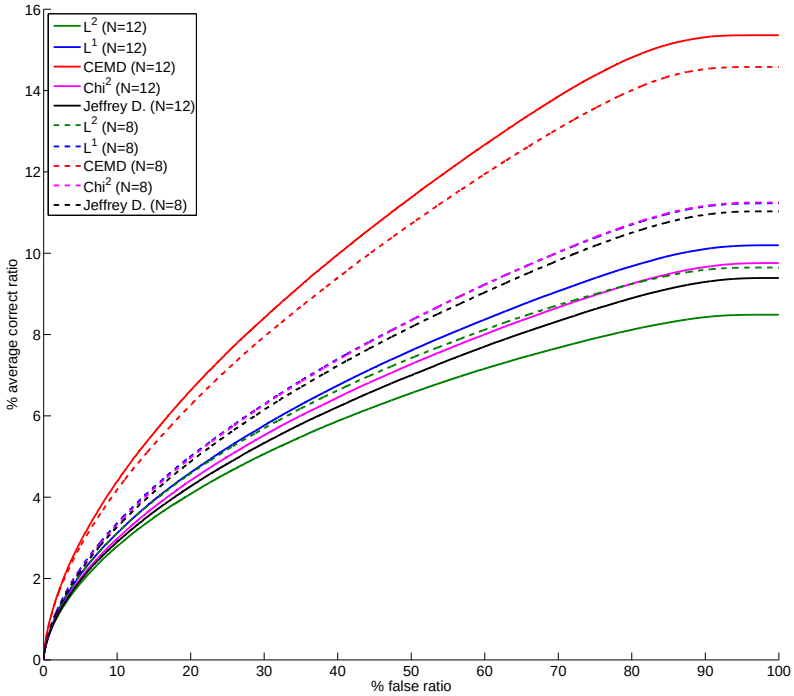


FIG. 4.3. *Average ROC curves (on* 732 *images) and* 3.1 *million descriptors for* CEMD *(red), L^1 (blue), L^2 (green), χ^2 distance (magenta) and Jeffrey divergence (black), with two different quantization levels ($N = 8$ for dashed lines and $N = 12$ for continuous lines).*

As previously mentioned in paragraph 2.2.2, EMD could be used to compare descriptors considered as three dimensional histograms (one dimension for the gradient orientations and two for the location of the region on the polar localization grid). We also saw that such a method involves intractable computation times, even when using the efficient implementation proposed by Ling *et al.*. [25]. Nevertheless, we performed a small scale experiment comparing such a use of EMD and the proposed CEMD on ten images from the database. The 3-dimensional EMD makes use of a ground distance obtained from a circular L^1 distance between orientations histograms and a L^1 ground distance on the position of regions of the descriptor. Since no indication is given in [25] on how to combine the two ground distances (respectively on position and orientation), we chose to simply add these two ground distances without trying to optimize their combination. We used the EMD code kindly provided by Y. Rubner [44]. Firstly, this implied computation times approximately 1000 times slower than when using CEMD. Secondly, the results displayed in Figure 4.4 suggest that 3-dimensional EMD, with this choice of ground distance, is less efficient than CEMD.

Effect of the tilt perturbation. In the previous experiments, the tilt parameter [34] was fixed to 2.5. As it has been suggested by the reviewers, we illustrate in the following experiment the matching performance for L_1 distance and CEMD in function of the amount of affine transform between images. This is achieved by computing the area under average ROC curves for different tilt values (the other parameters being set to 1 for scale, and 0 for both rotations). The figure 4.5 shows the resulting performance curves for CEMD and L_1 distance, with quantization set to $N = 12$ bins. As it could be expected, the performances of L_1 distances decrease faster than CEMD

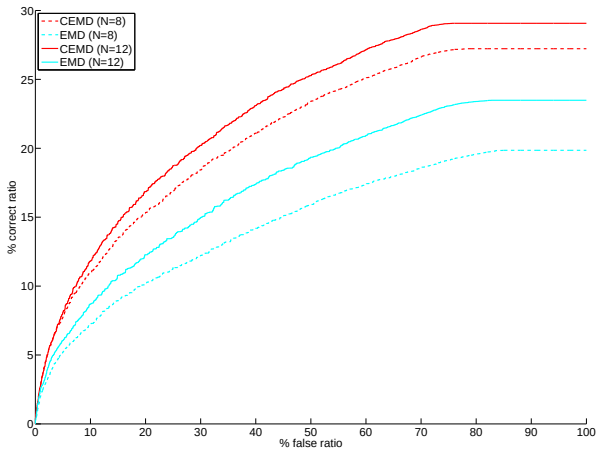


FIG. 4.4. *Average ROC curves (on* 10 *images) for* CEMD *(red) and 3-dimensional* EMD *(cyan), with two different quantization levels ($N = 8$ for dashed lines and $N = 12$ for continuous lines).*

when the tilt increases.

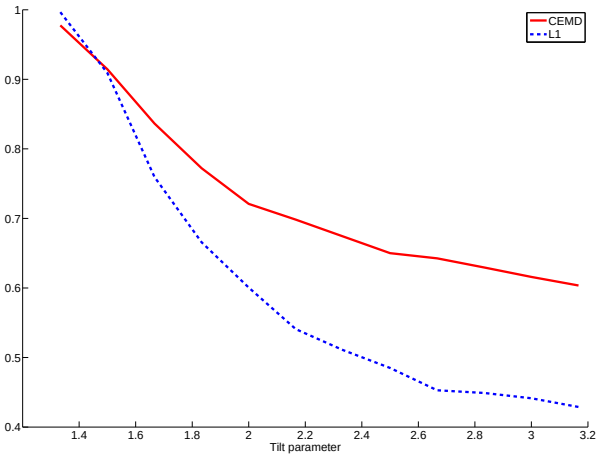


FIG. 4.5. *Area under average ROC curves (average on* 65 *images) for* CEMD *(red and continuous line) and L_1 (blue and dashed line), with $N = 12$ bins, for different tilt factors.*

4.3. Comparison of matching criteria - single match. Three matching criteria are compared in this section. All three criteria *limit matches to the nearest neighbor*, but make use of different thresholds. The first one is a threshold on distances, called NN-DT. The second threshold acts on the ratio between the distance to the nearest neighbor and the distance to the second nearest neighbor, as explained in Section 1.1. This criterion will be called NN-DR. The third criterion, called NN-AC, is the restriction to the nearest neighbor of the new matching criterion introduced in Section 3. Recall that a threshold on distances is obtained by thresholding a probability of false detections (see (3.3)). For the $A \to A'$ protocol, (3.3) is applied with $N_Q = N_C = N_A$, and for the $A \to \{^{A'}_{B}$ protocol with $N_Q = N_A$ and $N_C = N_A + N_B$. We use CEMD for all three matching methods.

Some images and associated ROC curves are shown in Fig. 4.6, both using the $A \to A'$ protocol (second and fifth rows) and the $A \to \{^{A'}_{B}$ protocol (third and sixth rows). In these curves, the NN-AC, NN-DT and NN-DR matching criteria are represented respectively in red, blue and green continuous lines. As in the previous paragraph, we can see that results can be quite different from one image pair to the other.

In order to compare the relative performances of different matching criteria, the same decision thresholds should be used for different query images, as is done in [33]. A *global ROC curve* is thus obtained by plotting the total number of correct matches on the whole database versus the total number of false matches, for different threshold values. Such a curve permits to evaluate how stable a given threshold is from one experiment to the other. The next two paragraphs present and interpret results on the whole database, relying on such curves, respectively for the $A \to A'$ and $A \to \{^{A'}_{B}$ protocols.

4.3.1. Single match. The target is present. Global ROC curves are displayed on Fig. 4.7 for the nearest neighbor criteria (namely NN-AC, NN-DT and NN-DR), using the $A \to A'$ protocol. We observe that both NN-AC and NN-DR have very similar global ROC curves, and that the NN-DT criterion is especially unstable. In this case, the NN-AC criterion proposed in this paper does not offer significant advantages in comparison with the classical NN-DR criterion. Indeed, as explained in Section 1.1, the NN-DR criterion is well adapted to the case where the target is present and yields excellent results in the special case of two images A and A' of the same scene, containing no distractors. Let us remark that we obtain results that are extremely close to the one shown in [33], where the authors obtain a flat global ROC curve for the NN-DT criterion and significant improvement with the NN-DR criterion. This analogy between our results and the ones in [33] also confirms the interest of using relatively large databases.

4.3.2. Single match. Is the target present ?. This section investigates the performances of the matching criteria on the whole database when using the $A \to \{^{A'}_{B}$ protocol. Fig. 4.8 shows the global ROC curve for this protocol. We can see that the performances of NN-DR clearly decrease in comparison to the ones of the proposed NN-AC criterion. For a given number of correct correspondences between A and A', NN-AC yields fewer false correspondences than NN-DR. As explained earlier, this shows the ability of the NN-AC criterion to discriminate between cases where the target is present and cases where it is not, which can be crucial for practical applications.

Next, we propose an extension of this last protocol where, for each query image A, the distractor image B is replaced by the entire database (deprived of A). Since this test involves much more computations than the previous one, it has been performed for only 100 images from the database (representing approximately $1.5\,10^{12}$ descriptor comparisons). Fig. 4.9 shows the corresponding global ROC curve. Again, one observes the substantial improvement provided by the NN-AC criterion. In fact, the improvement is greater than when only one image is used as a distractor, which suggests that the NN-AC criterion behaves well when the object of interest is seldom encountered.

4.4. Comparison of matching criteria - multiple matches. This section is a first attempt to compare matching criteria allowing multiple matching, thus *not restricted to the nearest neighbor*. First, observe that there is no obvious way to define such an extension for the NN-DR criterion. We therefore compare the following two criteria: a simple threshold on distances, that we call DT and the criterion introduced

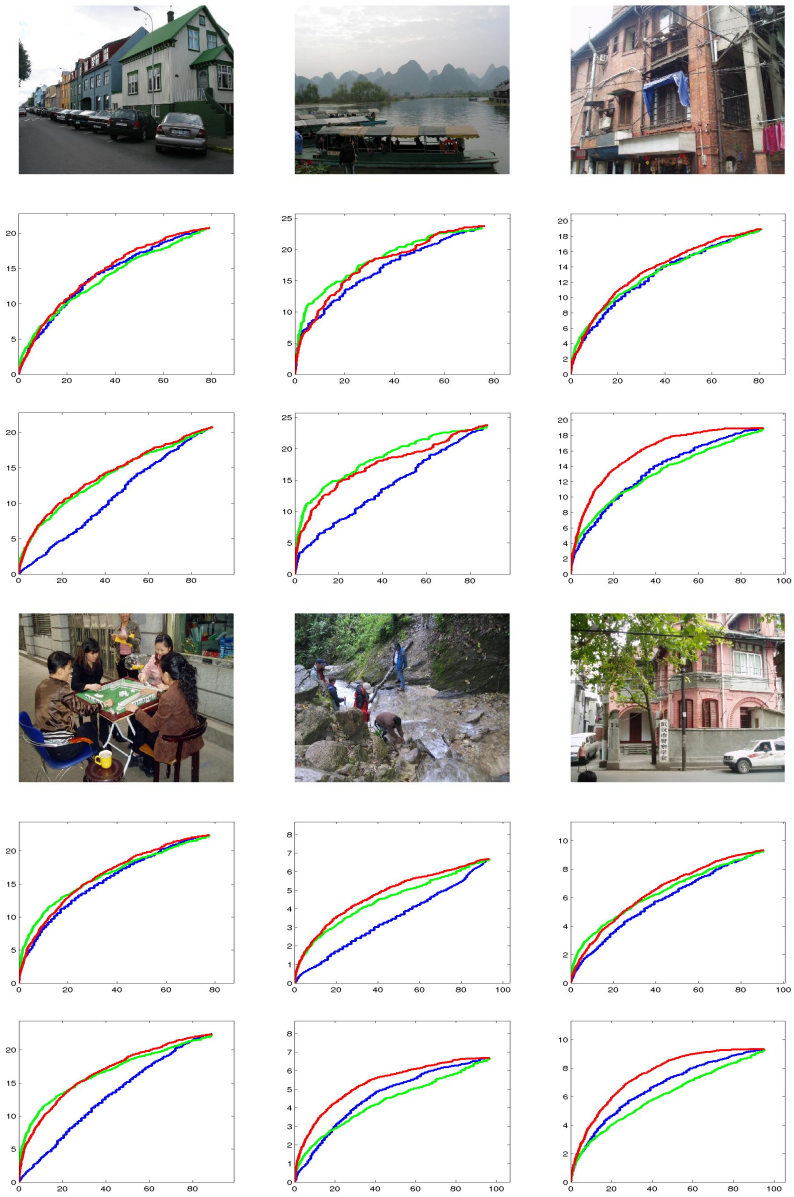


FIG. 4.6. *Six sample images from the database and the corresponding ROC curves. The red curves correspond to* NN-AC*, the blue ones to* NN-DT *and the green ones to* NN-DR*. The second and fifth rows show the curves obtained with the $A \to A'$ protocol, and the third and sixth rows show the results of the $A \to \{^{A'}_{B}$ protocol. Note that the relative performances of the three criteria depend strongly on the experiment.*

in this paper (without restricting matches to nearest neighbors), that we called AC. Both criteria allow multiple correspondences for each query descriptor.

For this comparison, we propose a protocol similar to $A \to \{^{A'}_{B}$, except that A' is replaced by a single image, called $A' + A''$, which is the concatenation of two different transformations of A. In this experiment, each structure of A appears twice

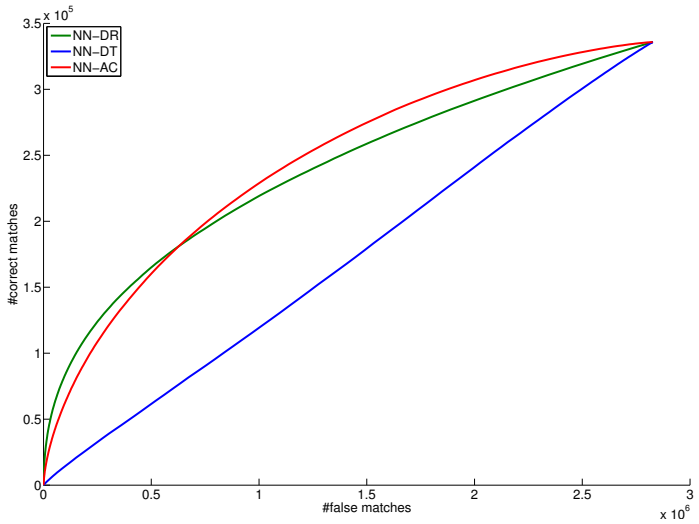


FIG. 4.7. *Global ROC curves (on the whole database) for different matching criteria:* NN-AC *(red),* NN-DT *(blue) and* NN-DR *(green). Experimental protocol is* $A \to A'$ *(an image* A *is matched against its transformed version* A'*).*

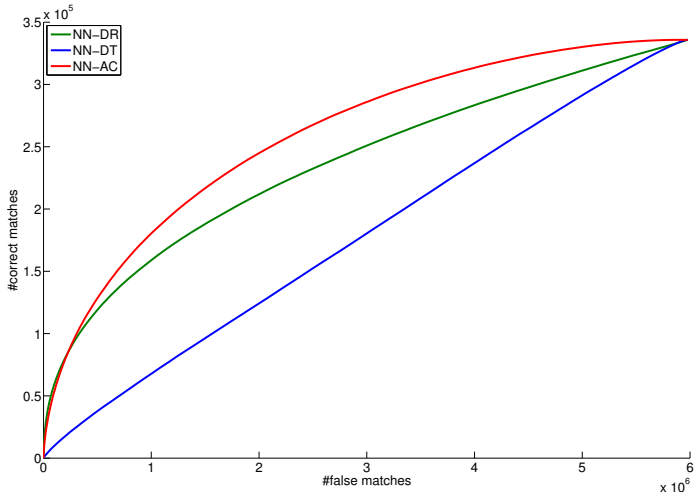


FIG. 4.8. *Global ROC curves (on the whole database) for different matching criteria:* NN-AC *(red),* NN-DT *(blue) and* NN-DR *(green). Experimental protocol is* $A \to \{^{A'}_{B}$ *(an image* A *is matched separately against* A' *and an independent image* B*).*

in $A' + A''$. Correct and false matches are counted exactly in the same way as in the protocol $A \to \{^{A'}_{B}$. This protocol is called $A \to \{^{A'+A''}_{B}$. Fig. 4.10 shows how the AC criterion clearly outperforms the DT criterion on the image database in this case of multiple matches.

4.5. Is the nearest neighbor restriction necessary ?. Following the previous section, it is quite natural to wonder whether not reducing the matches to the nearest neighbor yields a strong loss of performance *in the case where the target is present at most once.*

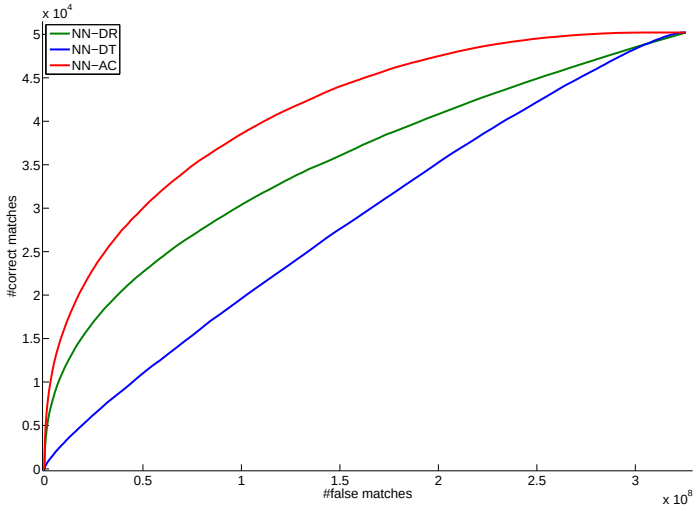


FIG. 4.9. *Global ROC curves (on* 100 *images) for different matching criteria:* NN-AC *(red),* NN-DT *(blue) and* NN-DR *(green). Experimental protocol is the same as* $A \to \{^{A'}_{B}$*, except that* B *is replaced by the complete database. That is, an image* A *is matched separately against* A' *and against each other image.*

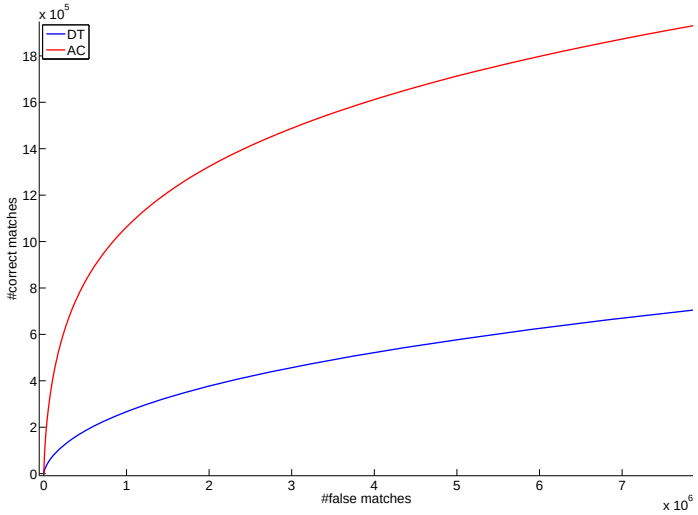


FIG. 4.10. *Global ROC curves (on the whole database) for the* $A \to \{^{A'+A''}_{B}$ *protocol (the target is present twice, see Section (4.4)). Criterion* AC *is shown in red and criterion* DT *is shown in blue.*

In Figure 4.11 we show, in continuous lines, global ROC curves for the two matching criteria AC and DT using the $A \to \{^{A'}_{B}$ protocol. Results for the matching criteria NN-AC and NN-DT, previously shown in Fig. 4.8, are represented in dashed lines. As could be expected, the performance of DT decreases significantly in comparison to NN-DT. Yet, we observe that AC and NN-AC criteria have similar results, even though AC does not have any restriction on the number of matches per query descriptor. This quite remarkable result indicates that the adaptive matching criterion introduced in this paper permits the rejection of false matches without any restriction on the number

of possible matches.

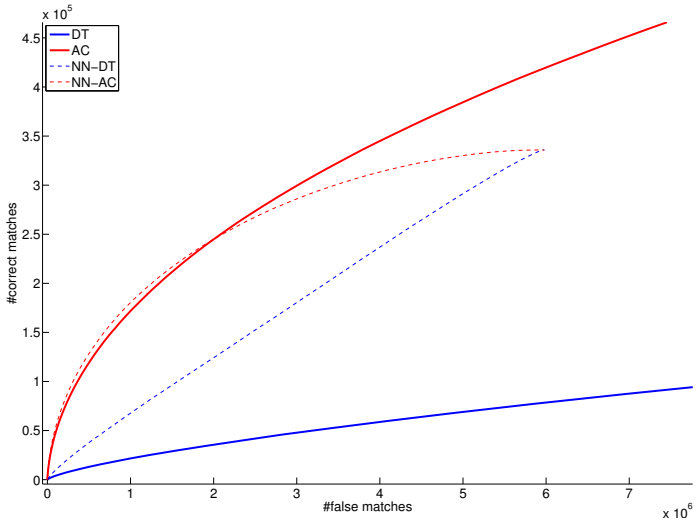
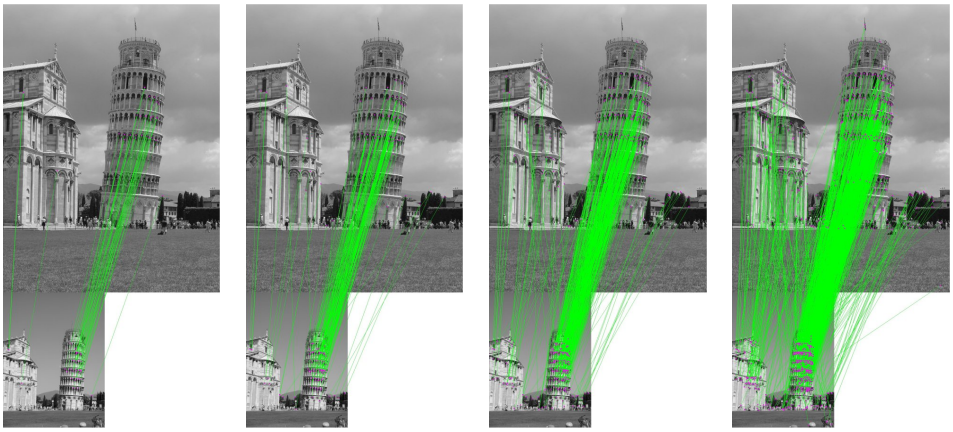


FIG. 4.11. *Global ROC curves (on the whole database) for different matching criteria and for the* $A \rightarrow \{^{A'}_{B}$ *protocol. Dashed lines: matches are restricted to the nearest neighbor (*NN-AC *in red and* NN-DT *in blue). Continuous lines: the number of matches per query is not restricted (*AC *in red and* DT *in blue).*

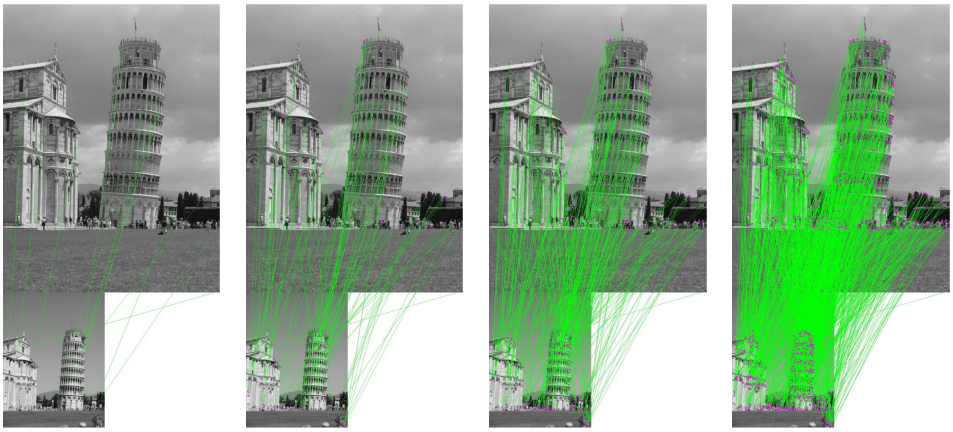
4.6. Some more experiments. In order to visually illustrate the behavior of the proposed matching procedure, this section presents some additional examples of matching between images.

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 [50], 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. 4.12. Criterion NN-DR (used in Figs. 4.12(e), 4.12(f), 4.12(g) and 4.12(h) with CEMD and 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. 4.12(a), 4.12(b), 4.12(c) and 4.12(d) with CEMD and 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. 4.13(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 complete matching procedure presented in this paper (CEMD for the distance and the AC criterion) is shown in Fig. 4.13(b)). It is compared to two classical matching procedures: Euclidean distance and NN-DR criterion in Fig. 4.13(c) or NN-DT criterion in Fig. 4.13(d). 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



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



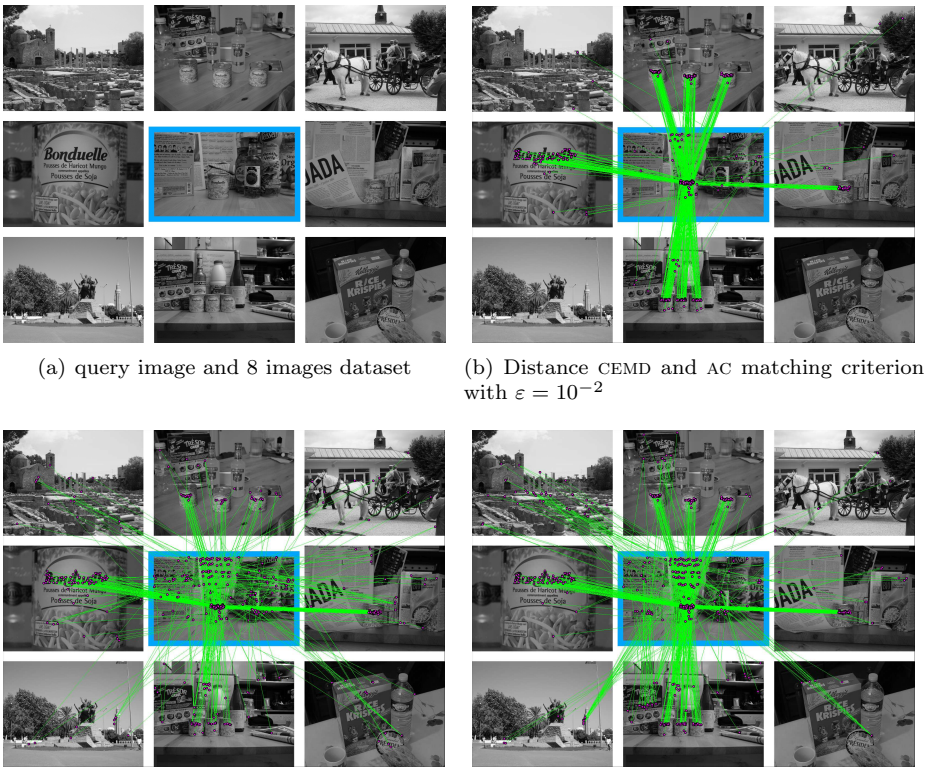
(e) NN-DR, $r = 0.7$ (f) NN-DR, $r = 0.8$ (g) NN-DR, $r = 0.85$ (h) NN-DR, $r = 0.9$

FIG. 4.12. *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.*

the query image and the image at the center of the leftmost column.

This matching experiment leads us to the same conclusions as the previous ROC curves. 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. This is not contradictory with the results of Section 4.3.1 (concluding to the equivalence of NN-DR and NN-AC when using the $A \to A'$ protocol), since many descriptors of either the query or the candidate image do not correspond to the object shared by the two images. This experiment shows (on a specific example) the versatility and adaptivity (all images are matched using the same threshold) of the proposed matching procedure.

5. Conclusion. In this paper, a new procedure for the matching of local, SIFT-like features has been proposed. First, a robust distance between circular histograms



(a) query image and 8 images dataset

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

(c) Euclidean distance and NN-DR matching criterion with $r = 0.8$

(d) Euclidean Distance and NN-DT matching criterion with $t < .45$

FIG. 4.13. *Comparison of different matching procedures (Distance + Matching Criterion). One query image (blue framed) containing a can is matched separately against* 8 *images (Fig. 4.13(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. These thresholds are chosen such that the number of correct matches with the image at the center of the left column is the same for all procedures. Observe that for a given number of correct matches with this left-centered image, the matching procedure introduced in this paper (*CEMD + AC *criterion) yields more correct matches in other images while providing a better control of the number of false detections than classical procedures.*

has been introduced and its advantages have been experimentally demonstrated on an image database. Second, a statistical matching criterion has been defined, relying on a threshold on a probability of false detections. The ability of this criterion to deal with situations where we do not know if the target is present has been demonstrated, as well as its ability to deal with 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 $(M-1)$ convolutions for each query descriptor a^i), it cannot benefit in a straightforward way from fast nearest neighbor search schemes [26, 2]. It is of interest to investigate the possibility to approximate the probability of false detections using only a small subset of candidate descriptors.

The distance introduced in Section 2 can also be applied to other descriptors made of circular histograms, such as color (hue) histograms. Observe also that the matching methodology presented in Section 3 is completely generic and could be

applied to other local descriptors, such as the affine invariant descriptors described in [29]. This matching methodology also enables us to simultaneously use different local features, by adapting the independence assumptions made in Section 3. Preliminary experiments on the joint use of color and direction histograms show promising results.

Acknowledgments. The authors thank Henri Maître for his useful comments. J. Delon acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant BLAN07-2_183172, Optimal transport: Theory and applications to cosmological reconstruction and image processing (OTARIE).

Appendix. Sufficient condition of meaningfulness.

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.

In the following, we show that 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 3.3, 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 $\delta < \mu$,

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

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 $\varepsilon \leq \dfrac{N_Q N_C}{2e\sqrt{\pi}}$, then $\log \dfrac{N_Q N_C}{2\sqrt{\pi}\varepsilon} > 1$, and it follows that $\dfrac{|D(a,b)-\mu|}{\sigma\sqrt{2}} > 1$. 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.(A.1), we obtain that $f(D(a,b)) \leq \dfrac{\varepsilon}{N_Q N_C}$. □

REFERENCES

[1] A. Baumberg, *Reliable feature matching across widely separated views*, in Proc. CVPR, 2000.

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

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

[4] A. Bosch, A. Zisserman, and X. Muñoz, *Scene classification using a hybrid generative/discriminative approach*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (2008), pp. 712–727.

[5] M. Brown and D. G. Lowe, *Automatic panoramic image stitching using invariant features*, Int. J. Comput. Vision, 74 (2007), pp. 59–73.

[6] M. Brown, R. Szeliski, and S. Windner, *Multi-image matching using multi-scale oriented patches*, in Proc. CVPR, 2005, pp. 510–517.

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

[8] G. Carneiro and A. D. Jepson, *Flexible spatial configuration of local image features*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (2007), pp. 2089–2104.

[9] O. Chum and J. Matas, *Matching with PROSAC - progressive sample consensus*, in Proc. CVPR, 2005, pp. 220–226.

[10] J. Delon, A. Desolneux, J. L. Lisani, and A. B. Petro, *A nonparametric approach for histogram segmentation*, IEEE Transactions on Image Processing, 16 (2007), pp. 253–261.

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

[12] A. Desolneux, L. Moisan, and J. Morel, *Meaningful alignments*, Int. J. Comput. Vision, 40 (2000), pp. 7–23.

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

[14] ———, *From Gestalt Theory to Image Analysis: A Probabilistic Approach*, Springer Verlag, 2008.

[15] F. Destrempes, M. Mignotte, and J. F. Angers, *Localization of shapes using statistical models and stochastic optimization*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (2007), pp. 1603–1615.

[16] V. Ferrari, T. Tuytelaars, and L. Gool, *Simultaneous object recognition and segmentation from single or multiple model views*, Int. J. Comput. Vision, 67 (2006), pp. 159–188.

[17] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision – 2nd Edition*, Cambridge University Press, 2004.

[18] J. Jia and C. K. Tang, *Image stitching using structure deformation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (2008), pp. 617–631.

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

[20] E. Levina and P. Bickel, *The Earth Mover's Distance is the Mallows distance: some insights from statistics*, in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 2, 2001, pp. 251–256 vol.2.

[21] T. Lindeberg, *Feature Detection with Automatic Scale Selection*, International Journal of Computer Vision, 30 (1998), pp. 79–116.

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

[23] ———, *An integrated model for evaluating the amount of data required for reliable recognition*, IEEE Trans. Pattern Anal. Mach. Intell., 19 (1997), pp. 1251–1264.
[24] H. Ling and K. Okada, *Diffusion distance for histogram comparison*, in CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2006, IEEE Computer Society, pp. 246–253.
[25] ———, *An efficient Earth Mover's distance algorithm for robust histogram comparison*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (2007), pp. 840–853.
[26] D. G. Lowe, *Distinctive image features from scale-invariant keypoints*, Int. J. Comput. Vision, 60 (2004), pp. 91–110.
[27] J. Matas, O. Chum, M. Urban, and T. Pajdla, *Robust wide baseline stereo from maximally stable extremal regions*, in BMVC, 2002, pp. 384–393.
[28] K. Mikolajczyk and C. Schmid, *Indexing based on scale invariant interest points*, International Conference on Computer Vision, (2001), pp. 525–531.
[29] ———, *Scale & affine invariant interest point detectors*, Int. J. Comput. Vision, 60 (2004), pp. 63–86.
[30] ———, *A performance evaluation of local descriptors*, IEEE Trans. Pattern Anal. Mach. Intell., 27 (2005), pp. 1615–1630.
[31] R. G. Miller, *Simultaneous Statistical Inference*, Springer-Verlag, New York, 1991.
[32] L. Moisan and B. Stival, *A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix*, International Journal of Computer Vision, 57 (2004), pp. 201–218.
[33] P. Moreels and P. Perona, *Evaluation of features detectors and descriptors based on 3d objects*, Int. J. Comput. Vision, 73 (2007), pp. 263–284.
[34] J.M. Morel and G. Yu, *ASIFT: A New Framework for Fully Affine Invariant Image Comparison*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 438–469.
[35] 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 (2006), pp. 295–315.
[36] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, *Qbic project: querying images by content, using color, texture, and shape*, vol. 1908, SPIE, 1993, pp. 173–187.
[37] C. Olson and D.P. Huttenlocher, *Automatic target recognition by matching oriented edge pixels*, IEEE Transactions on Image Processing, 6 (1997), pp. 103–113.
[38] O. Pele and M. Werman, *A linear time histogram metric for improved sift matching*, in ECCV08, 2008.
[39] J. Rabin, J. Delon, and Y. Gousseau, *A contrario matching of local descriptors*, tech. report, HAL, 2007.
[40] ———, *Circular Earth Mover's Distance for the comparison of local features*, in Proc. ICPR, IEEE Computer Society, 2008.
[41] ———, *A contrario matching of SIFT-like descriptors*, in Proc. ICPR, IEEE Computer Society, 2008.
[42] ———, *Transportation distances on the circle*, tech. report, HAL, 2009.
[43] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, *Segmenting, modeling, and matching video clips containing multiple moving objects*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (2007), pp. 477–491.
[44] Y. Rubner, C. Tomasi, and L. J. Guibas, *The Earth Mover's distance as a metric for image retrieval*, Int. J. Comput. Vision, 40 (2000), pp. 99–121.
[45] J. Sivic and A. Zisserman, *Video Google: Efficient visual search of videos*, in Toward Category-Level Object Recognition, vol. 4170 of LNCS, Springer, 2006, pp. 127–144.
[46] N. Snavely, S. M. Seitz, and R. Szeliski, *Modeling the world from internet photo collections*, To appear in Int. J. Comput. Vision, (2008).
[47] C. Villani, *Topics in optimal transportation*, American Math. Soc., 2003.
[48] M. Werman, S. Peleg, R. Melter, and TY Kong, *Bipartite graph matching for points on a line or a circle*, Journal of Algorithms, 7 (1986), pp. 277–284.
[49] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, *Local features and kernels for classification of texture and object categories: A comprehensive study*, Int. J. Comput. Vision, 73 (2007), pp. 213–238.
[50] W. Zhang and J. Kosecka, *Generalized ransac framework for relaxed correspondence problems*, in Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), Washington, DC, USA, 2006, IEEE Computer Society, pp. 854–860.