

Interpreting the Ratio Criterion for Matching SIFT Descriptors

Avi Kaplan, Tamar Avraham, Michael Lindenbaum

Computer science department,
Technion - I.I.T.
`kavi,tammya,mic@cs.technion.ac.il`

Abstract. Matching keypoints by minimizing the Euclidean distance between their SIFT descriptors is an effective and extremely popular technique. Using the ratio between distances, as suggested by Lowe, is even more effective and leads to excellent matching accuracy. Probabilistic approaches that model the distribution of the distances were found effective as well. This work focuses, for the first time, on analyzing Lowe's ratio criterion using a probabilistic approach. We provide two alternative interpretations of this criterion, which show that it is not only an effective heuristic but can also be formally justified. The first interpretation shows that Lowe's ratio corresponds to a conditional probability that the match is incorrect. The second shows that the ratio corresponds to the Markov bound on this probability. The interpretations make it possible to slightly increase the effectiveness of the ratio criterion, and to obtain matching performance that exceeds all previous (non-learning based) results.

Keywords: SIFT, matching, *a contrario*

1 Introduction

Matching objects in different images is a fundamental task in computer vision, with applications in object recognition, panorama stitching, and many more. The common practice is to extract a set of distinctive keypoints from each image, compute a descriptor for each keypoint, and then match the keypoints using a similarity measure between the descriptors and possibly also geometric constraints. Many methods for detecting the keypoints and computing their descriptors have been proposed. See the reviews [23, 11, 25].

The scale invariant feature transform (SIFT) suggested by Lowe [8, 9] is arguably the dominant algorithm for both keypoint detection and keypoint description. It specifies feature points and corresponding neighborhoods as maxima in the scale space of the DoG operator. The descriptor itself is a set of histograms of gradient directions calculated in several (16) regions in this neighborhood, concatenated into a 128-dimensional vector. Various normalizations and several filtering stages help to optimize the descriptor. The combination of scale space, gradient direction, and histograms makes the SIFT descriptor robust to scale, rotation, and illumination changes, and yet discriminative. Keypoints are matched

by minimizing the Euclidean distance between their SIFT descriptors. However, to rank the matches, it is much more effective to use the distance ratios:

$$\text{ratio}(\mathbf{a}_i, \mathbf{b}_{j(i)}) = \frac{\|\mathbf{a}_i - \mathbf{b}_{j(i)}\|_2}{\|\mathbf{a}_i - \mathbf{b}_{j'(i)}\|_2} \quad (1)$$

and not the distances themselves [8, 9]. Here, \mathbf{a}_i denotes a descriptor in one image, and $\mathbf{b}_{j(i)}, \mathbf{b}_{j'(i)}$ correspond to the closest and the second-closest descriptors in the other image.

SIFT has been challenged by many competing descriptors. The variations try to achieve faster runtime (e.g. SURF [2]), robustness to affine transformation (ASIFT [14]), compatibility with color images (CSIFT [1]) or simply represent the neighborhood in a different but related way (PCA-SIFT [7] and GLOH [11]). Performance evaluations [11, 13, 12, 22] conclude, however, that while various SIFT alternatives may be more accurate under some conditions, the original SIFT generally performs as accurately as the best competing algorithms, and better than the speeded-up versions.

SIFT descriptors are matched based of their dissimilarity, which makes the choice of dissimilarity measure important. The Euclidean distance (L_2) [9] is still the most commonly used. Being concatenations of orientation histograms, SIFT descriptors can naturally and effectively be compared using measures for comparing distributions, such as χ^2 distance [28] and circular variants of the Earth mover’s distance. Alternative, probabilistic approaches consider the dissimilarities as random variables. In [10], for instance, the dissimilarities are modeled as Gaussian random variables. The probabilistic *a contrario* theory, which we follow in this work, was effectively applied to matching SIFT-like descriptors [21] (as well as many other computer vision tasks [4]).

This work focuses, for the first time, on using a probabilistic approach for analyzing the ratio criterion. We show that this effective yet nonetheless heuristic criterion may be justified by two alternative interpretations. One shows that the ratio corresponds to a conditional probability that the match is incorrect. The second shows that the ratio corresponds to the Markov bound on this probability. These interpretations hold for every available distribution of dissimilarities between unrelated descriptors, and in particular, for all the distributions suggested later in this paper.

We also consider several dissimilarity measures, including, unusually, a multi-value (vector) one. The distributions of the dissimilarities, corresponding to incorrect matches, are constructed by partitioning the descriptor into parts (following [21]), estimating a rough distribution of the dissimilarities between the corresponding parts, and combining the distributions of the partial dissimilarities into a distribution of the full dissimilarity. These distributions, denoted as *background models* (as in [4]), are estimated, online, only from the matched images, requiring no training phase and making them image adaptive.

Combining these estimated distributions with the conditional probability (the first probabilistic interpretation of the ratio criterion) provides a new cri-

terion, or algorithm, for ranking matches. With this algorithm, we obtain state-of-the-art matching accuracy (for non-learning methods)¹.

In this paper, we make the following contributions:

1. A mathematical explanation of the ratio criterion as a conditional probability, which justifies this effective and popular criterion.
2. A second justification of the ratio criterion using the Markov inequality.
3. New measures of dissimilarity and methods for deriving their distribution.
4. A new matching criterion combining the estimated dissimilarity distribution with the conditional probability interpretation, which obtains excellent results.

Outline: Sec. 2 shows how to use dissimilarity distributions to rank matching hypotheses and, in particular, to justify the ratio criterion as a conditional distribution. Sec. 3 describes various dissimilarities and the corresponding partition-based background model distributions and summarizes the proposed matching process. Sec. 4 provides an additional justification of the ratio criterion. Experimental results are described in Sec. 5, and Sec. 6 concludes.

2 Using a background model for matching

2.1 Modeling with background models

We would like to generate a set of reliable matches between the feature points of two images A and B , using their corresponding sets of SIFT descriptors, $\mathcal{A} = \{\mathbf{a}_i\}$ and $\mathcal{B} = \{\mathbf{b}_j\}$. Let $a_i \in A$ be a specific feature point and let $\mathbf{a}_i \in \mathcal{A}$ be its corresponding descriptor. Most descriptors in \mathcal{B} (all except possibly one) are unrelated to \mathbf{a}_i in the sense of not corresponding to the same scene point. Our goal is to find the single feature point, if it exists, which matches \mathbf{a}_i .

The proposed matching process is based on statistical principles. We model the non-matching descriptors in \mathcal{B} as realizations of a random variable X drawn from some distribution. This distribution is high-dimensional and complex, and we refrain from working with it directly. Instead we consider associated dissimilarity values. Let $\delta(\mathbf{u}, \mathbf{v})$ be some dissimilarity measure between two descriptors \mathbf{u} and \mathbf{v} . We shall be interested in the dissimilarities $\delta(\mathbf{a}_i, \mathbf{b}_j)$ between the descriptors in \mathcal{A} and in \mathcal{B} . We consider \mathbf{a}_i to be a specific fixed descriptor, and \mathbf{b}_j to be drawn from the distribution of X . Then, the associated dissimilarity $\delta(\mathbf{a}_i, \mathbf{b}_j)$ is an instance of another random variable, which we denote $\mathbf{Y}^{\mathbf{a}_i}$. $\mathbf{Y}^{\mathbf{a}_i}$ is distributed by $F_{\mathbf{Y}^{\mathbf{a}_i}}(y)$, representing the dissimilarity distribution associated with \mathbf{a}_i and non-matches to it. Following the *a contrario* approach, we denote this distribution a *background model*. Note that the dissimilarities associated with different descriptors in \mathcal{A} follow different background models.

¹ Recently, learned patch descriptors were introduced as alternatives to SIFT (e.g. [6, 27]). We do not compete with their performance, as the focus in this work is to suggest and validate our new explanation to the ratio criterion

In this section we assume that this distribution is known. In Sec. 3 we consider several dissimilarity measures as well as methods for estimating the corresponding background models. As usual, we consider a scalar dissimilarity. Later, we show that an extension to multi-value (vector) dissimilarity is beneficial.

Given a matching hypothesis by which the two descriptors, $(\mathbf{a}_i, \mathbf{b}_j)$ correspond to the same 3D point, we contrast this hypothesis with the alternative, null hypothesis, by which the descriptor \mathbf{b}_j is drawn from a distribution of false matches. This is a false correspondence, or false alarm event, and we denote its probability (following *a contrario* notations [3]), as the probability of false alarm (PFA). The null background model hypothesis is rejected more strongly if the PFA is lower, and therefore matching hypotheses with lower PFA are preferred. This approach is further developed in the rest of this section.

The PFA replaces the commonly used distance between descriptors with a probabilistic measure. This new matching criterion is mathematically well defined, quantitatively and intuitively meaningful, and possibly image adaptive.

2.2 Ranking hypothesized matches by PFA

Let $(\mathbf{a}_i, \mathbf{b}_j)$ be a hypothesized match. To evaluate the null hypothesis we calculate the probability of drawing, from the background model, a value that is as extreme (small) as the dissimilarity of the hypothesized match, $\delta(\mathbf{a}_i, \mathbf{b}_j)$. Let $E_{1-1}^{\mathbf{a}_i}(d)$ denote the event that the value drawn from the distribution of $\mathbf{Y}^{\mathbf{a}_i}$ is smaller or equal to d . Then,

$$PFA_{1-1}(\mathbf{a}_i, \mathbf{b}_j) = \Pr(E_{1-1}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_j))) = F_{\mathbf{Y}^{\mathbf{a}_i}}(\delta(\mathbf{a}_i, \mathbf{b}_j)). \quad (2)$$

Thus, PFA_{1-1} is just the one-sided (lower) tail probability of the distribution. A hypothesized match is ranked higher if its PFA_{1-1} is lower, which enables us to reject the null hypothesis with higher confidence. We use this ranking to specify the feature point (and the descriptor) in the image B , corresponding to \mathbf{a}_i :

$$\mathbf{b}_{j(i)} = \arg \min_{\mathbf{b}_j \in \mathcal{B}} PFA_{1-1}(\mathbf{a}_i, \mathbf{b}_j). \quad (3)$$

Thus, for every $\mathbf{a}_i \in \mathcal{A}$ we get a single preferred match $(\mathbf{a}_i, \mathbf{b}_{j(i)})$, denoted selected match². Next, we would like to rank these selected matches. To that end, we may use, again, $PFA_{1-1}(\mathbf{a}_i, \mathbf{b}_{j(i)})$, as explained in the next section. Further below (Sec. 2.4), we propose another, more reliable PFA expression.

2.3 Ranking selected matches as complex events

The selected match, $(\mathbf{a}_i, \mathbf{b}_{j(i)})$, is a false alarm event when at least one of the $|\mathcal{B}|$ random, independently drawn, dissimilarity $\mathbf{Y}^{\mathbf{a}_i}$ values is as extreme as

² One might think that ranking by the PFA_{1-1} is equivalent to ranking by the dissimilarity measure itself. This is true in the simpler case when the dissimilarity is scalar and the distribution depends directly on this scalar, but not in more complex cases, as we shall see in Sec. 3.

$\delta(\mathbf{a}_i, \mathbf{b}_{j(i)})$. This is a complex event, denoted $E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)}))$. Its probability may be calculated as the binomial distribution

$$PFA_{1-|\mathcal{B}|}(\mathbf{a}_i, \mathbf{b}_{j(i)}) = 1 - (1 - PFA_{1-1}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)})))^{|\mathcal{B}|}. \quad (4)$$

As observed in our experiments, the typical values of $PFA_{1-1}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)}))$ are small, and usually much smaller than $1/|\mathcal{B}|$. Under this condition,

$$PFA_{1-|\mathcal{B}|}(\mathbf{a}_i, \mathbf{b}_{j(i)}) \approx |\mathcal{B}| \cdot F_{\mathbf{Y}^{\mathbf{a}_i}}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)})). \quad (5)$$

We may use this approximation for ranking the selected matches. Note that ranking by PFA_{1-1} is equivalent. We shall also use this approximation below for deriving a more effective, conditional, PFA.

2.4 Ranking selected matches by conditional PFA

Indirect evidence about the dissimilarity values drawn from the background model may improve the estimate of the PFA. Such evidence may be available from small dissimilarity values in $\{\delta(\mathbf{a}_i, \mathbf{b}_j) | j \neq j(i)\}$. These dissimilarities do not correspond to correct matches and come only from the background model. Consider the complex event associated with the second lowest dissimilarity, denoted $\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)})$. Knowing that the event $E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)}))$ occurred allows us to recalculate the probability that the event $E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)}))$ occurred as well, as a conditional probability:

$$\begin{aligned} PFA_C(\mathbf{a}_i, \mathbf{b}_{j(i)}) &= \Pr(E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)})) | E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)}))) \\ &= \frac{\Pr(E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)})) \cap E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)})))}{\Pr(E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)})))} \\ &\leq \frac{\Pr(E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)})))}{\Pr(E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)})))} \approx \frac{|\mathcal{B}| \cdot F_{\mathbf{Y}^{\mathbf{a}_i}}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)}))}{|\mathcal{B}| \cdot F_{\mathbf{Y}^{\mathbf{a}_i}}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)}))}. \end{aligned} \quad (6)$$

For scalar dissimilarities, the event $E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)}))$ is included in the event $E_{1-|\mathcal{B}|}^{\mathbf{a}_i}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)}))$, and the inequality above is actually an equality. For vector dissimilarities (Sec 3.2) the equality does not strictly hold. We shall use the approximation to the bound as an estimate for PFA_C ,

$$PFA_C(\mathbf{a}_i, \mathbf{b}_{j(i)}) \approx \frac{F_{\mathbf{Y}^{\mathbf{a}_i}}(\delta(\mathbf{a}_i, \mathbf{b}_{j(i)}))}{F_{\mathbf{Y}^{\mathbf{a}_i}}(\delta(\mathbf{a}_i, \mathbf{b}_{j'(i)}))}, \quad (7)$$

and prefer matches with lower PFA_C .

The expression (7) is reminiscent of the ratio of distances (Eq. 1) used by Lowe in the original SIFT paper [9]. We argue, moreover, that the derivation in Sec. 2 mathematically generalizes and justifies the two criteria suggested in [9]. First, (incorrectly) assuming a uniform distribution of the Euclidean distance dissimilarity makes the PFA_{1-1} expression proportional to the Euclidean distance.

Thus, minimizing PFA, as suggested here, is exactly equivalent to minimizing the Euclidean distance (as in [9]). Furthermore, with the uniform distribution assumption, the conditional PFA_C expression (Eq. 7) is simply the distance ratio (Eq. 1). As we show later, replacing the incorrect uniform distribution assumption with a more realistic distribution further improves the matching accuracy.

3 Partition-based dissimilarities and background models

3.1 Standard estimation of dissimilarity distribution

The various PFA expressions developed in Sec. 2 are valid for any available distribution. To convert these general expressions into a concrete algorithm, we consider several dissimilarity measures, and derive their distributions. We first consider the standard nonparametric distribution estimation method. Then, following the *a contrario* framework, we suggest partition-based methods that overcome some limitations of the standard method.

We may estimate the background distribution from different sources of data; each has pros and cons. One option uses an offline *external* database \mathcal{F} containing many non-matching descriptor pairs, (\mathbf{u}, \mathbf{v}) , unrelated to \mathcal{A}, \mathcal{B} . The distribution, $F(y) = \Pr(\delta(\mathbf{u}, \mathbf{v}) \leq y)$, may be estimated using the formal definition of the empirical distribution function: $\hat{F}(y) = |\{(\mathbf{u}, \mathbf{v}) \in \mathcal{F} : \delta(\mathbf{u}, \mathbf{v}) \leq y\}|/|\mathcal{F}|$ or more advanced methods, such as kernel density estimation; see [24]. The second option, which is image adapted and denoted *internal*, is to estimate the distribution from all pairs $(\mathbf{a}_i, \mathbf{b}_j)$ related to the specific image. The third option, which is also online, but *point adapted*, is to estimate the distribution using only $\mathcal{F} = \{(\mathbf{a}_i, \mathbf{b}_j) : \mathbf{b}_j \in \mathcal{B}\}$, separately for every descriptor \mathbf{a}_i in \mathcal{A} . While more data is available for the two first options, the last option is the one that fits our task. Since \mathcal{B} is relatively small and may contain the descriptor corresponding to the correct match to \mathbf{a}_i , more creative methods should be used to estimate the point-adapted distribution. Standard techniques would lead to a coarsely estimated distribution, which is a staircase function with steps of size $1/|\mathcal{B}|$, and $PFA_{1-1}(\mathbf{a}_i, \mathbf{b}_{j(i)})$ will always be $1/|\mathcal{B}|$, regardless of whether the match is correct, which would mean that it is useless for matching. The partition-based approach described below adopts the last option but avoids its problems.

3.2 Partition-based dissimilarities

The partition-based approach divides the descriptor vector into K non-overlapping parts: $\mathbf{u} = (\mathbf{u}[1], \dots, \mathbf{u}[K])$, and calculates the dissimilarity between two (full) descriptors from the dissimilarities between the corresponding parts. Here, we describe the partition and the dissimilarities. In Sec 3.3 we show how to use this partitioning to estimate the desired distribution. Then we explain why it avoids the problems of the standard approach.

To specify the partition-based dissimilarity, we need to make three choices: how to partition the descriptor, how to measure partial dissimilarities, and how to combine them into the final dissimilarity.

The partitioning - The descriptors may be partitioned in many ways, some of which are more practical and statistically consistent with the independence assumption (made in Sec. 3.3). We tried several finer and coarser options, and ended up with the natural partition of the SIFT vector to its 16 orientation histograms, which gave the best results.

Basis distance - The dissimilarity between parts is denoted $d(\mathbf{u}[j], \mathbf{v}[j])$, and, to distinguish it from the dissimilarity of the full descriptors, is called a *basis distance*. It may be the L_2 metric but may also be some other metric (e.g. L_1) or even a dissimilarity measure which is not a metric (e.g. EMD). In Sec 5, we experiment with the L_2 distance and two versions of the EMD [21, 19].

Dissimilarity measure - There are many ways to combine the part dissimilarities, and we consider three of them:

1. **Sum dissimilarity** - The dissimilarity between two full descriptors, already used and analyzed in [21], is the sum of the basis distances:

$$\delta^{\text{sum}}(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^K d(\mathbf{u}[j], \mathbf{v}[j]). \quad (8)$$

2. **Max-value dissimilarity** - The dissimilarity between the full descriptors, already used for a *contrario* analysis of shape recognition ([17, 15]), is the the maximal basis distance.

$$\delta^{\text{max}}(\mathbf{u}, \mathbf{v}) = \max_j d(\mathbf{u}[j], \mathbf{v}[j]). \quad (9)$$

3. **Multi-value dissimilarity** - Instead of summarizing all the basis distances associated with the parts with one scalar, we may keep them separate. Then the dissimilarity measure is a vector of K basis distances. This richer, vectorial, dissimilarity has the potential to exploit the non-uniformity in the descriptor's parts and to be more robust.

$$\bar{\delta}^{\text{multi}}(\mathbf{u}, \mathbf{v}) = (d(\mathbf{u}[1], \mathbf{v}[1]), d(\mathbf{u}[2], \mathbf{v}[2]), \dots, d(\mathbf{u}[K], \mathbf{v}[K]))^T. \quad (10)$$

This dissimilarity, being a vector, induces only a partial order. We follow the standard definition and say that one dissimilarity vector is smaller (not larger) than another if each of its components is smaller (not larger) than the corresponding component of the second, or

$$\bar{y}_1 \leq \bar{y}_2 \iff \forall j, (\bar{y}_1)_j \leq (\bar{y}_2)_j. \quad (11)$$

Note that this partial order suffices for defining a distribution because, for every vector \bar{y} , the set of all vectors not larger than \bar{y} is well defined.

For brevity, we use $\delta(\mathbf{u}, \mathbf{v})$ as a general notation that, depending on context, may refer to the different types of dissimilarity defined above, and to different basis distances $d(\cdot)$. Note that $\delta(\mathbf{u}, \mathbf{v})$ may take a vector value. Therefore, we refrain from referring to it as a distance function.

Not all dissimilarities may be represented using the partition-based schemes. Euclidean distance is one example. Note, however, that squared Euclidean distance is representable as a sum of basis distances, and as it is monotonic in the Euclidean distance, they are equivalent with respect to ranking.

3.3 Estimating the background models

Different background models are constructed for the three types of partition-based dissimilarities discussed above. Let $\mathbf{Y}^{\mathbf{a}_i}[1], \mathbf{Y}^{\mathbf{a}_i}[2], \dots, \mathbf{Y}^{\mathbf{a}_i}[K]$ be the K random variables corresponding to the K basis distances between the parts of a specific descriptor \mathbf{a}_i and a randomly drawn descriptor from \mathcal{B} .

The background models are estimated using the assumption that the set of random variables $\{\mathbf{Y}^{\mathbf{a}_i}[j]\}_{j=1}^K$, associated with unrelated sub-descriptors, is mutually statistically independent. For SIFT descriptors, this assumption is not completely justified due to the correlation between nearby image regions and because of the common normalization. Yet it seems to be a valid and common approximation. Later we comment on the validity of this assumption, test it empirically, and discuss a model which avoids it.

The background model for multi-value dissimilarity is characterized by the distribution $F_{\mathbf{Y}^{\mathbf{a}_i}}^{\text{multi}}(\bar{y})$. Here $\mathbf{Y}^{\mathbf{a}_i}$ is specified by the partition-based multi-value dissimilarity and is a vector variable. To specify its distribution, we use the partial order between these vectors specified in Eq. (11). Then,

$$F_{\mathbf{Y}^{\mathbf{a}_i}}^{\text{multi}}(\bar{y}) = \Pr\left(\bigcap_{j=1}^K \{\mathbf{Y}[j] \leq (\bar{y})_j\}\right). \quad (12)$$

Using the independence assumption, we get

$$F_{\mathbf{Y}^{\mathbf{a}_i}}^{\text{multi}}(\bar{y}) = \prod_{j=1}^K \Pr(\mathbf{Y}[j] \leq (\bar{y})_j). \quad (13)$$

Each term is independently estimated as empirical distribution, yielding

$$\hat{F}_{\mathbf{Y}^{\mathbf{a}_i}}^{\text{multi}}(\bar{y}) = \prod_{j=1}^K \left(\frac{1}{|\mathcal{F}|} \cdot |\{\mathbf{v} \in \mathcal{F} : d(\mathbf{a}_i[j], \mathbf{v}[j]) \leq (\bar{y})_j\}| \right). \quad (14)$$

The background model for sum dissimilarity is characterized by the distribution $F_{\mathbf{Y}^{\mathbf{a}_i}}^{\text{sum}}(y)$, which is estimated by convolving the estimated densities of the part dissimilarities; we refer the reader to [21] for details.

The background model for max-value dissimilarity is characterized by the distribution $F_{\mathbf{Y}^{\mathbf{a}_i}}^{\text{max}}(y)$. The max-value dissimilarity is a particular case of the multi-value dissimilarity and is similarly estimated; see also [17, 15].

As before, we use the notation $F_{\mathbf{Y}^{\mathbf{a}_i}}(y)$ for general reference to a dissimilarity distribution that describes the background model.

Partition-based background models have several advantages. First, the quantization of the distribution is fine even for the small descriptor set \mathcal{B} , as the multiplication of several distributions reduces the step size exponentially (in the number of parts). Moreover, the presence of a correct match in \mathcal{B} is less harmful because even if the corresponding descriptor is globally closest to \mathbf{a}_i , it doesn't necessarily mean that all its parts are closest to the respective parts $\mathbf{a}_i[j]$ as well.

The proposed matching algorithm, based on the probabilistic interpretation and the distribution estimation, is concisely specified in Algorithm 1.

Algorithm 1 Partition-based matching algorithm

Preliminary: Choose a basis distance d , a dissimilarity measure δ , and the type of false alarm probability (conditional or not).

Input: Two sets of descriptors, \mathcal{A}, \mathcal{B} , associated with two sets of image points.

Output: A set of matches, one for every descriptor $\mathbf{a}_i \in \mathcal{A}$, with corresponding estimates for probability of false match $\{((\mathbf{a}_i, \mathbf{b}_{j(i)}), PFA(\mathbf{a}_i, \mathbf{b}_{j(i)}))\}$.

Algorithm

1. For each descriptor $\mathbf{a}_i \in \mathcal{A}$,
 - (a) For each descriptor part calculate all basis distances $\{d(\mathbf{a}_i, \mathbf{b}_j) : j = 1, 2, \dots, |\mathcal{B}|\}$ and estimate a one-dimensional empirical distribution.
 - (b) Estimate the background model either by taking the convolution of the empirical densities (for sum dissimilarity [21]) or by just keeping them as a set of distributions (for max-value [17, 15] or multi-value dissimilarities (Eq. 14)).
 - (c) For each descriptor $\mathbf{b}_j \in \mathcal{B}$, calculate the dissimilarity $\delta(\mathbf{a}_i, \mathbf{b}_j)$ and use the background model to infer the probability of false alarm $PFA_{1-1}(\mathbf{a}_i, \mathbf{b}_j)$.
 - (d) Choose the descriptor associated with the lowest PFA_{1-1} : $\mathbf{b}_{j(i)} = \arg \min_{\mathbf{b}_j \in \mathcal{B}} PFA_{1-1}(\mathbf{a}_i, \mathbf{b}_j)$. Let $(\mathbf{a}_i, \mathbf{b}_{j(i)})$ be the resulting match.
 - (e) Augment the match $(\mathbf{a}_i, \mathbf{b}_{j(i)})$ with its unconditional PFA (Eq. 2) or the conditional PFA (Eq. 6).

4 Markov inequality based interpreting

Sec. 2 describes a method for using a distribution of false dissimilarities for matching. It justifies the ratio criterion [9] as an instance of this approach. Here we propose an alternative explanation of and justification for the ratio criterion.

Consider the set of N_B ordered dissimilarities (e.g. Euclidean distances) $\delta_1, \dots, \delta_{N_B}$ between a given descriptor $\mathbf{a}_i \in \mathcal{A}$ and all descriptors in \mathcal{B} . The hypothesized match $(\mathbf{a}_i, \mathbf{b}_{j(i)})$ corresponds to the smallest dissimilarity δ_1 . Our goal, again, is to estimate how likely is it that δ_1 was drawn from the same distribution as the other dissimilarities. To that end, we apply the Markov bound.

We assume that the dissimilarities associated with incorrect matches are independently drawn instances of the random variable Y . Let D, Z be two random variables, related to Y , where D assumes the minimal dissimilarity over a set of N samples, and $Z = 1/D$. Z is nonnegative, and by Markov's inequality, satisfies $\Pr(Z \geq a) \leq E[Z]/a$, or $\Pr(1/D \geq a) \leq E[1/D]/a$. Letting $d = 1/a$ and rearranging the inequality, we get

$$\Pr(D \leq d) \leq d \cdot E[1/D]. \quad (15)$$

Thus, knowing the expected value $E[1/D]$, we could bound the probability that the observed minimal value, δ_1 , belongs to the distribution of D . To estimate $E[1/D]$ from a single data set (δ_1 excluded), with a single minimal value, we use bootstrapping. Given a set of N samples, bootstrapping takes multiple samples of length N , drawn uniformly with replacement, and estimates the required

Table 1: mAP for *absolute* algorithms for the Mikolajczyk et al. [11] dataset. *PMV* is third from the right.

	Deterministic			Probabilistic								
	L2	CEMD	TMEMD	sum			max-value			multi-value		
				L2	CEMD	TMEMD	L2	CEMD	TMEMD	L2	CEMD	TMEMD
Blur (Bikes)	0.756	0.694	0.737	0.795	0.556	0.795	0.720	0.455	0.706	0.798	0.769	0.791
Blur (Trees)	0.513	0.460	0.499	0.524	0.483	0.522	0.419	0.318	0.408	0.516	0.481	0.512
Viewpoint (Graffiti)	0.334	0.317	0.337	0.354	0.359	0.373	0.281	0.247	0.277	0.383	0.368	0.388
Viewpoint (Wall)	0.686	0.666	0.689	0.697	0.456	0.708	0.587	0.515	0.594	0.714	0.688	0.713
Zoom + Rotation (Bark)	0.804	0.750	0.771	0.812	0.622	0.789	0.736	0.665	0.712	0.823	0.788	0.812
Zoom + Rotation (Boat)	0.789	0.761	0.783	0.813	0.797	0.815	0.725	0.676	0.714	0.829	0.808	0.826
Light (Leuven)	0.861	0.433	0.845	0.899	0.884	0.897	0.855	0.791	0.841	0.909	0.886	0.897
JPEG Compression (UBC)	0.874	0.864	0.874	0.878	0.698	0.882	0.852	0.447	0.854	0.876	0.874	0.878

statistics from the samples. It is easy to show that, for large $N = N_B$,

$$\hat{E}[1/D] = \sum_{k=2}^{N_B} p_k \frac{1}{\delta_k} \leq \frac{1}{\delta_2}, \quad (16)$$

where $(p_2, p_3, p_4, \dots) \approx (0.63, 0.23, 0.08, \dots)$. Combining Eqs. 15 and 16, we get

$$\Pr(D \leq \delta_1) \leq \delta_1 \cdot E[1/D] \approx \delta_1 \cdot \hat{E}[1/D] \leq \frac{\delta_1}{\delta_2}. \quad (17)$$

We can now justify Lowe’s relative criterion by observing that the ratio δ_1/δ_2 approximates a bound over the probability that a randomly drawn dissimilarity associated with incorrect matches is smaller or equal to the observed δ_1 . If the bound is small, then so is the probability that δ_1 is associated with an incorrect match. This makes the ratio δ_1/δ_2 a statistically meaningful criterion.

5 Experiments

5.1 Experimental setup

We experimented with 18 different variations of the proposed matching algorithm, corresponding to combinations of three different basis distances (L_2 , circular EMD (CEMD) [20], and thresholded modulo EMD (TMEMD) [19]), the three different dissimilarity measures (sum, max-value, and multi-value), and the unconditional and conditional matching criteria. The variant that uses CEMD distance, sum dissimilarity, and the unconditional criterion corresponds to the algorithm of [21]. We denote by *PMV* (probabilistic multi-value) and *PMV_c* (probabilistic multi-value conditional) the variations that use the L_2 basis distance and the multi-value dissimilarity. The probabilistic algorithms are compared to deterministic algorithms which use the same basis distances as dissimilarities between the descriptors (without partition). The deterministic algorithms are implemented in six versions corresponding to the three distance functions and to the two criteria based on distance and distance ratio (two of those variations correspond to [9] and [19]).

Table 2: mAP for *relative* algorithms for the Mikolajczyk et al. [11] dataset. PMV_c is third from the right.

	Deterministic			Probabilistic								
	L2	CEMD	TMEMD	sum			max-value			multi-value		
				L2	CEMD	TMEMD	L2	CEMD	TMEMD	L2	CEMD	TMEMD
Blur (Bikes)	0.812	0.792	0.810	0.814	0.383	0.812	0.740	0.649	0.727	0.812	0.776	0.806
Blur (Trees)	0.536	0.499	0.535	0.539	0.505	0.539	0.430	0.335	0.420	0.533	0.497	0.527
Viewpoint (Graffiti)	0.380	0.381	0.396	0.385	0.161	0.405	0.299	0.270	0.299	0.411	0.398	0.419
Viewpoint (Wall)	0.707	0.605	0.713	0.711	0.605	0.718	0.595	0.511	0.603	0.725	0.689	0.723
Zoom + Rotation (Bark)	0.816	0.762	0.787	0.816	0.769	0.789	0.739	0.656	0.713	0.822	0.787	0.811
Zoom + Rotation (Boat)	0.823	0.799	0.820	0.825	0.623	0.822	0.732	0.684	0.721	0.834	0.795	0.825
Light (Leuven)	0.907	0.676	0.908	0.909	0.678	0.909	0.874	0.800	0.862	0.917	0.898	0.908
JPEG Compression (UBC)	0.883	0.700	0.886	0.884	0.477	0.887	0.855	0.381	0.856	0.881	0.878	0.883

We experimented with the datasets of Mikolajczyk et al. [11] and of Fischer et al. [6]. Both datasets contain 5 image pairs, each composed of an original image and a transformed version of it. We used the evaluation protocol of [11] which relies on finding the homography between the images, for both sets. All the algorithms match SIFT descriptors, extracted using the VLFEAT package [26]. The partition-based algorithms divide the SIFT descriptor into 16 parts, corresponding to the 16 histograms of the standard SIFT. Different partitioning, into 2, 4, 8, and 32 parts, performed less well compared to the 16 part partition.

5.2 Matching Results

The results for the Mikolajczyk et al. [11] dataset are summarized using the mAP (mean average precision) per scene (over the first 4 transformation levels) in Table 1 and Table 2, and the mAP per transformation level in Fig. 1. In Table 1 we compare the proposed probabilistic algorithms that use unconditional probability (PFA_{1-1}) to Lowe’s first matching criterion, and to the other deterministic algorithms that use CEMD [20] and TMEMD [19]. The algorithms in this set are referred to as *absolute*. Table 2 reports on *relative* algorithms, including the probabilistic algorithms that use the conditional probability of false alarm, PFA_c , and the deterministic versions corresponding to Lowe’s distance ratio, as well as to similar ratios of the EMD variants. Fig. 1 compares the 4 versions corresponding to PMV , PMV_c , and Lowe’s first and second criteria.

The results for the absolute algorithms are clear: PMV obtains the best results on the average. The sum-value based algorithm follows closely. Both algorithms clearly outperform the deterministic algorithms.

The performance obtained by using the unconditional probabilistic criterion is comparable to that obtained by the non-probabilistic distance ratio criterion. This is not surprising because both algorithms rely on the context of other dissimilarities, besides that of the tested descriptor pair; see also [21].

While the three dissimilarities use the same context different results are achieved. The max dissimilarity, for example, performs even worse than the deterministic approaches. It seems that the multi-value algorithm is able to exploit the cell-dependent variability in the distances between the various SIFT

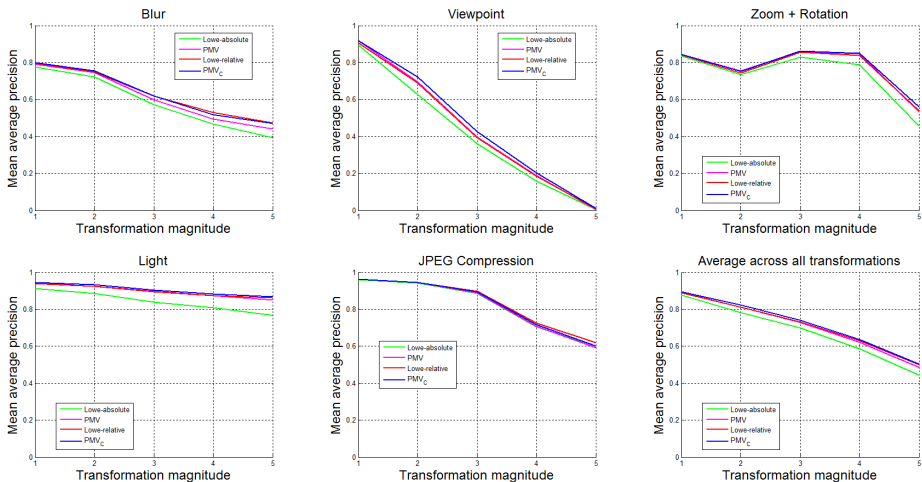


Fig. 1: mAP as a function of the transformation magnitude for PMV , PMV_c , and Lowe’s two criteria for the Mikolajczyk et al. [11] dataset.

components. This variability could result from the downweighting of the outer cells and the lower variability of their histograms, due to the larger distance of the outer cells from the most likely location of the interest point: the patch’s center. We believe that it is also due to the accidental nonuniformity of the distances between the parts, which is averaged by the sum distance, and leads to the poorer performance of the max-value dissimilarity measure, which is based on a uniform dissimilarity threshold.

The differences between the relative algorithms are small. PMV_c is slightly better, especially for viewpoint transformations which is the most common application for matching. We believe that the reason for PMV_c being less successful with the blur and JPEG compression is that these transformations add noise to the image, which has a greater impact on low-dimensional vectors. The results demonstrate the validity of the interpretations proposed in this paper.

Fig 2 summarizes the results obtained for the Fischer et al. dataset [6] (the 3 nonlinear transformations, which do not correspond to a homography, were ignored). The results are consistent, although with only a minor improvement compared to Lowe’s ratio criterion. Note that performance differences between Lowe’s first criterion and second criterion are not expressed here as well³.

The partition-based algorithms are more computationally expensive than the deterministic ones (slower by a factor of 8 to 15, in our current, unoptimized,

³ Different matching papers use different evaluation protocols, different feature points (even when using SIFT descriptors), and different criteria for counting recalls and false alarms. As such, the results cannot be directly compared with those reported in some other papers. Nevertheless, we use exactly the same evaluation protocol for all the algorithms tested here.

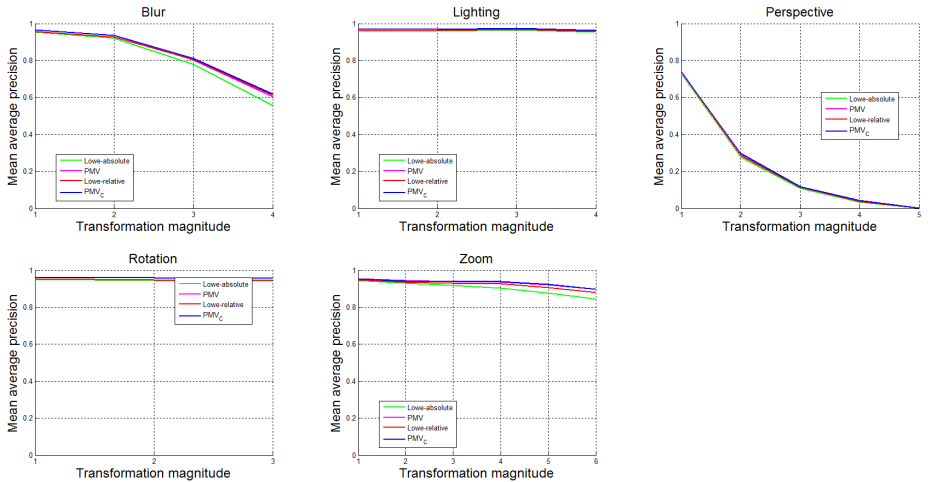


Fig. 2: mAP as a function of the transformation levels for PMV , PMV_C , and Lowe’s two criteria, for the Fischer et al. [6] dataset.

implementation). As our goal in the implementation was to demonstrate the validity of the probabilistic justification, and to show that the generality of the analysis possibly leads to diverse algorithms, we did not focus on runtime optimizations. Our algorithms can be optimized following common methods of efficient nearest neighbor search [16]. This is left for future work.

5.3 Complementary experiments

The importance of point-adapted distribution All the above experiments used *point-adapted* distributions. We experimented also with the *internal* and *external* distributions (defined in Sec. 3.1), using images from the Caltech 101 dataset [5] to compute the external distribution. We found that the algorithms performed best with the point-adapted distribution; see Fig. 3. The external distribution based versions performed worse than Lowe’s first distance criterion.

Relaxing the independence assumption We challenged the statistical independence assumption and found that some substantial correlations between the partial distances exist. Using a more complex distribution model, which relies on a dependence tree (following [18]), did not improve the ranking result.

6 Discussion and conclusions

The main contributions of this paper are two, independent, probabilistic interpretations of Lowe’s ratio criterion. This criterion, used to match SIFT descriptors, is very effective and widely used, yet so far it has only been empirically

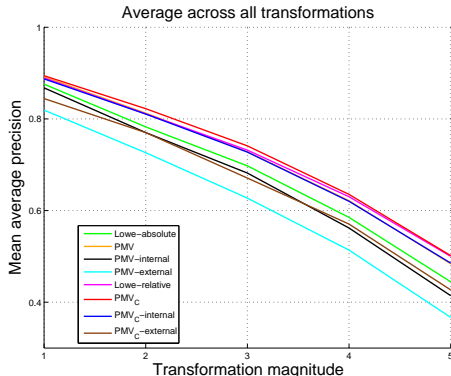


Fig. 3: Comparing external, internal, and point-adaptive distribution estimations (see Sec. 2.1 and Sec. 5.3) for the Mikolajczyk et al. [11] dataset.

based. Our interpretations show that, in a probabilistic setting, it corresponds either to a conditional probability that the match is incorrect, or to the Markov bound on this probability. To the best of our knowledge, this is the first rigorous justification of this important tool.

To test the (first) interpretation empirically, we used techniques of the *a contrario* approach to construct a distribution of the dissimilarities associated with false matches. Some dissimilarity functions were considered, including one that, unlike common measures, expresses the dissimilarity by a multi-value (vector) measure and can better capture the variability of the descriptor.

Using a matching criterion based on conditional probability and a multi-value dissimilarity measure led to state-of-the-art matching performance (for non-learning-based algorithms). The improvement over the previous, nonprobabilistic, algorithms is not very large. The experiments nonetheless support the probabilistic interpretation, as they demonstrate that explicit use of it leads to consistent and even better results.

We follow several aspects of the *a contrario* theory [3, 4]: we use background models that provide the PFA probability that an event occurs by chance, and the partition-based estimation technique. However, in order to use the probabilistic interpretation of the ratio criterion, and contrary to the *a contrario* approach, we refrain from using the expected number of false alarms (NFA) as the matching criterion and use the probability of false alarm instead.

One promising idea for future work is to combine our method of using only false positive error estimation in the decision-making process with models for false negative errors; see e.g. [10], which focuses on the differences between specific descriptors in different images.

References

1. Abdel-Hakim, A.E., Farag, A.A.: Csift: A sift descriptor with color invariant characteristics. *CVPR 2, 1978–1983* (2006)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). *Computer Vision Image Understanding* 110(3), 346–359 (2008)
3. Desolneux, A., Moisan, L., Morel, J.M.: Meaningful alignments. *International Journal of Computer Vision* 40(1), 7–23 (2000)
4. Desolneux, A., Moisan, L., Morel, J.M.: *From Gestalt Theory to Image Analysis: A Probabilistic Approach*, vol. 34. Springer Science & Business Media (2007)
5. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* 106(1), 59–70 (2007)
6. Fischer, P., Dosovitskiy, A., Brox, T.: Descriptor matching with convolutional neural networks: A comparison to sift. *arXiv preprint arXiv:1405.5769* (2014)
7. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: *CVPR04*. vol. 2, pp. II–506 (2004)
8. Lowe, D.G.: Object recognition from local scale-invariant features. *ICCV 2*, 1150–1157 (1999)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
10. Mikolajczyk, K., Matas, J.: Improving descriptors for fast tree matching by optimal linear projection. In: *ICCV07*. pp. 1–8 (2007)
11. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 27(10), 1615–1630 (2005)
12. Miksik, O., Mikolajczyk, K.: Evaluation of local detectors and descriptors for fast feature matching. *ICPR* pp. 2681–2684 (2012)
13. Moreels, P., Perona, P.: Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision* 73(3), 263–284 (2007)
14. Morel, J.M., Yu, G.: Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences* 2(2), 438–469 (2009)
15. Mottalli, M., Tepper, M., Mejail, M.: A contrario detection of false matches in iris recognition. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 442–449 (2010)
16. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1) 2*(331-340), 2 (2009)
17. Musé, P., Sur, F., Cao, F., Gousseau, Y., Morel, J.M.: An a contrario decision method for shape element recognition. *International Journal of Computer Vision* 69(3), 295–315 (2006)
18. Myaskovskiy, A., Gousseau, Y., Lindenbaum, M.: Beyond independence: An extension of the a contrario decision procedure. *International Journal of Computer Vision* 101(1), 22–44 (2013)
19. Pele, O., Werman, M.: A linear time histogram metric for improved sift matching. *ECCV* pp. 495–508 (2008)
20. Rabin, J., Delon, J., Gousseau, Y.: Circular earth movers distance for the comparison of local features. *ICPR* pp. 1–4 (2008)
21. Rabin, J., Delon, J., Gousseau, Y.: A contrario matching of sift-like descriptors. In: *ICPR*. pp. 1–4 (2008)
22. Sande, K.E.V.D., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32(9), 1582–1596 (2010)

23. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. *International Journal of Computer Vision* 37(2), 151–172 (2000)
24. Silverman, B.W.: *Density Estimation for Statistics and Data Analysis*, vol. 26. CRC press (1986)
25. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. *Foundations and Trends® in Computer Graphics and Vision* 3(3), 177–280 (2008)
26. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/> (2008)
27. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: *CVPR*. pp. 4353–4361 (2015)
28. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* 73(2), 213–238 (2007)