Geometric Consistency Checking for Local-Descriptor Based Document Retrieval

Eduardo Valle, David Picard, Matthieu Cord

Geometric Consistency Checking for Local-Descriptor Based Document Retrieval

Eduardo Valle
NPDI - ICEx / DCC / UFMG
31270-010 Belo Horizonte, Brazil
mail@eduardovalle.com

David Picard
LIP6 UPMC PARIS 6
Paris, France
david.picard@lip6.fr

Matthieu Cord
LIP6 UPMC PARIS 6
Paris, France
Matthieu.cord@lip6.fr

ABSTRACT

An efficient architecture, based on voting and local descriptors, has been proposed to retrieve multimedia documents. In this paper, we evaluate different geometric consistency schemes, which can be used in tandem with this architecture, and that, in many contexts, are essential to boost the retrieval performance. Our empirical results show however, that geometric consistency alone is unable to guarantee high-quality results in databases that contain too many non-discriminating descriptors.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.8 [**Image Processing and Computer Vision**]: Scene analysis—*motion, object recognition*

General Terms

Algorithms, Design, Experimentation.

Keywords

Image Retrieval, Local Descriptors, Geometric Consistency, Retrieval by Voting, CBIR.

1. INTRODUCTION

Many document retrieval applications are mainly based on "target identification", meaning that the user is not interested in retrieving a category of objects but in obtaining all information available about a specific target object. When dealing with visual documents (images, videos, 3D models, etc.) on this kind of context, there is a very effective architecture, based on a voting algorithm over local descriptors. This architecture was first described in [6] and has been extensively used throughout the literature (e.g. [5, 7]).

To explain the architecture in a nutshell: we start by taking each visual document in the database and describing it using local descriptors. When the user performs a query, we also compute the local descriptor for the query document. For each query descriptor, we look for the most similar descriptors in the database, and give a vote for the document to which it belongs. Then we count how many votes each document received and use this number as a criterion of similarity.

The method has several advantages. Compared to global descriptor techniques, it is robust because the descriptors are many: if some get too distorted or are completely lost, enough will remain to guarantee good results. Also, it is efficient, because pairwise image comparison is not necessary; we match the individual descriptors independently, and rely on the vote count to aggregate the results. This allows performing the search in sub-linear time.

On the other hand, the architecture is not without challenges. First and foremost, the multiplicity of descriptors penalizes the performance, since many descriptors must be matched in order to identify a single document. The whole scheme is feasible only if efficient indexing techniques are able to accelerate the similarity search used to match the descriptors. This problem has been extensively addressed in the literature, and today good indexing methods are available for multimedia descriptors [3, 1, 8].

Another problem arises when matches are made incorrectly, giving votes for the wrong documents. If the number of query descriptors is large and the fraction of incorrect matches is small, then any incorrectly identified document will receive significantly less votes than the correct one. Otherwise, the reliability of the voting mechanism is compromised. If a significant fraction of matches is known to be incorrect, the reliability can be greatly improved by enforcing geometric consistency constraints. The idea is to scrutinize the list of images obtained by the vote counting algorithm and check, for each image, if the matches are compatible with the expected geometric transformation. The matches which do not follow the general trend are removed, the votes are recounted and the list is re-ranked.

In this paper, we evaluate the effectiveness of geometric consistency strategies in scenarios with an overwhelming fraction of incorrect matches, a situation which arises when either the query or the database images are of poor quality, or the deformation suffered by the query image is too strong.

2. APPLICATION: ITOWNS PROJECT

This study was motivated by the iTowns Project, which is defining a new generation of multimedia web tools that mixes a broadband 3D geographic image-based browser with an image-based search engine [1]. The first goal of this new type of search engine is to retrieve, in the high-resolution database, the scene corresponding to a given query image.

In a possible scenario, the users are looking for information about a restaurant in front of them (e.g., feedback from other patrons). They can take a picture of the restaurant with a cell-phone and, using the iTowns web server, obtain the desired information.

In order to accomplish this goal, there are basically three steps to perform:

1. Match the query document with the corresponding document in the database;
2. Find information associated with the database document and related to the query;
3. Retrieve only relevant information regarding the user interests.

In this paper, we focus on the first step.

Since our problem involves target identification, we have decided to employ the efficient architecture based on voting over local-descriptors, which was described on the previous section. We have used SIFT [5] to describe the images, due to their robustness. In order to match the descriptors, we have used Multicurves, an indexing based on space-filling curves which is well adapted for very large databases [8].

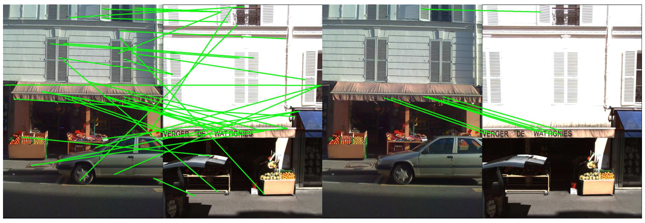

Figure 1: Answering a user query - we want to retrieve the database document (right) corresponding to the user query (left: image taken with a cell phone). The retrieval is performed by matching the local descriptors (green lines). Though many matches may be found (left half) few of them might be correct (right half). Geometric con sistency helps to filter out the incorrect matches.

3. CONSISTENCY STRATEGIES

There are several different ways to perform the geometric consistency, but all of them fall on two general strategies: (1) estimate the geometric transformation by applying an (usually robust) estimator, and eliminating the matches incompatible with that transformation; (2) computing a statistical distribution of some geometrical transformation parameter (rotation, scale) of each match, and eliminating the matches which deviate too much from the mode of that distribution.

The first strategy is more precise, but also more complex to implement. In order to apply it, one has to choose the kind of transformation model whose parameters will be estimated (e.g., scale change, similarity transformation, affine transformation, etc.) and the estimator used. Since many outliers are expected, the estimator has to be robust.

A usual combination is to use the RANSAC estimator [2] with a 2D affine transformation model. The RANSAC (RANdom Sample Consensus) is a Maximum Likelihood technique that is robust to the presence of a large fraction of outliers. It works by selecting (at random) a small set of samples and estimating the model parameters from them. The model so estimated is then used to count the inliers and the outliers. The process is iterated several times, selecting potentially different samples at each iteration. The model which generates the largest fraction of inliers is kept. The idea behind RANSAC is that, if it happens to select a sample exclusively composed of inliers, then there is a good chance that the estimated model will be compatible with all other inliers.

The 2D affine transformation model is a compromise between simplicity and comprehensiveness. It can model most, but not all, deformations suffered by the query. In particular, the viewpoint change, which is affine in the 3D spaces, generates a non-affine geometric transformation in the 2D plane. But the general 3D affine transformation is too complex, demanding a lot of samples to be reliably estimated, and thus it is ill adapted for RANSAC. Since viewpoint transformations can be locally modeled as 2D affine transformations, the model is able to match at least part of the image.

The second strategy, based on statistics, is less precise but much simpler. In our comparison we make the choices described in [4]. In order to apply it, we first choose which parameters of the transformation we will inspect — rotation and scale being the most common and then study the statistical behavior of that parameter. We inspect only the rotation (which is readily available as the difference between the principal direction of the query and target SIFT points), creating a histogram of the angles observed. We then consider as inliers only the matches corresponding to the highest peak (the mode) on that histogram.

4. EXPERIMENTS

We have compared the two consistency strategies (RANSAC and Rotation Histogram). To make the comparison more complete, we have also included the baseline method (using the brute vote count, without geometric consistency), and an alternative method using pairwise image comparison (as described in [5]). The methods compared were thus:

- Voting algorithm with the RANSAC over 2D affine transform consistency criterion (“RANSAC”).
- Voting algorithm with the Rotation Histogram consistency criterion (“Rotation Histogram”).
- Pairwise comparison using a criterion of contrast in the distance between descriptors (“Pairwise Matching”).
- Voting algorithm without geometric consistency (“Brute Vote”).

We have tested the methods on two subsets of the iTowns database. The query sets of both tests contained images taken by a mobile phone in front of some of the shops on

9[1] See http://itowns.ign.fr

the street. The first dataset consisted of 82 images of a single street (about 350 000 descriptors). Since the images (both in the query set and in the database) were direct views of the buildings, the transformation between query and its corresponding target images was simple, and we have considered this test easy. The second dataset contained 300 images of a large boulevard (about 3.5 millions of descriptors). As the vehicle taking the pictures was in the middle of the street, the target regions on the images (a shop, for instance) were very small. Thus, few descriptor on each image were actually meaningful. Since the transformations were severe (scaling, viewpoint changes), we have considered this test difficult.

For both sets, we have manually built the ground truth by annotating which images corresponded to each query. We have used two criteria for the evaluation. The first consisted in measuring the rank of the first relevant image retrieved (averaged over the query set). The second measure was the evolution of the number of relevant image in the retrieved set, as the size of this set increased.

An example for a single query can be appreciated on Figure 2. The left image shows the retrieval attempt using the Brute Vote technique. Since the query image presents several distortions (occlusions, scale change, small viewpoint change) the retrieval is not perfect, but a relevant image is found in the 6th rank. Using the Rotation Histogram technique (middle image) the relevant image rank is improved to the 3rd position. Using the RANSAC technique (right image) three relevant images are found on the first three ranks — an improvement on both the rank of the first relevant image and the overall precision.

4.1 Results for the First Dataset

We have computed the mean best rank of relevant images for all four methods. The results are shown on Table 1.

Method	mean best rank
Image matching	27.09
Brute vote	14
RANSAC	**1.09**
Rotation Histogram	7.91

Table 1: Mean best rank for the first dataset.

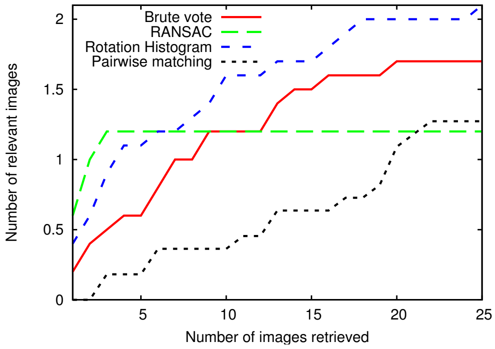

Figure 3: Evolution of the number of relevant images as the number of retrieved images grows.

As we can see, the Pairwise Matching shows the worst results. The method by Brute Force has a much more reasonable time, but the quality is still not good. The Rotation Histogram scheme was able to cut the mean rank almost by half, but it was RANSAC who had the most impressive results. In our application context, the performances it obtained would be satisfactory from the user point of view.

We also measured the evolution of the number of relevant images as the number of retrieved images increases on Figure 3. The RANSAC method outperforms the others for a small number of retrieved images, but then stops to progress at around 6 images. If the number of retrieved images is large, the Rotation Histogram technique is able to retrieve more relevant images than RANSAC. There are two reasons for the flat progression in RANSAC. First, it requires a minimum number of correct matches in order to evaluate the model. For the 2D affine transform, 7 matches are needed (though the strict minimum are 3 matches, this would lead to numerically unstable estimations). This means that if an image has less than 7 matches it is automatically discarded, since no model can be estimated. Second, if the number of outliers becomes too overwhelming, the probability of a reasonable model being estimated becomes too small and RANSAC fails.

4.2 Results for the Second Dataset

Like we have done for the first subset, we have compute the mean best rank of relevant images for all four methods, shown in Table 2.

Method	mean best rank
Image matching	80.67
Brute vote	98.80
RANSAC	**34.40**
Rotation Histogram	59.10

Table 2: Mean best rank for the second dataset.

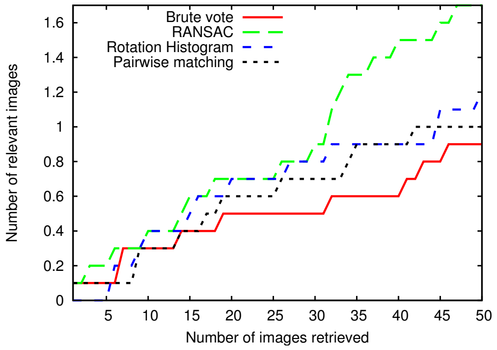

Figure 4: Evolution of the number of relevant images as the number of retrieved images grows.

The results are unsatisfactory, especially when contrasted with the ones obtained on the (easier) dataset (Table 1). RANSAC still shows the best results, but its impact is now much subdued. For all methods, the first relevant image is so low on the ranking that probably it won't even show in the first page of results. From the point of view of the user, the results are highly inadequate.

This low performance is confirmed on the graph of evolution of the number of relevant images (Figure 4). There is very

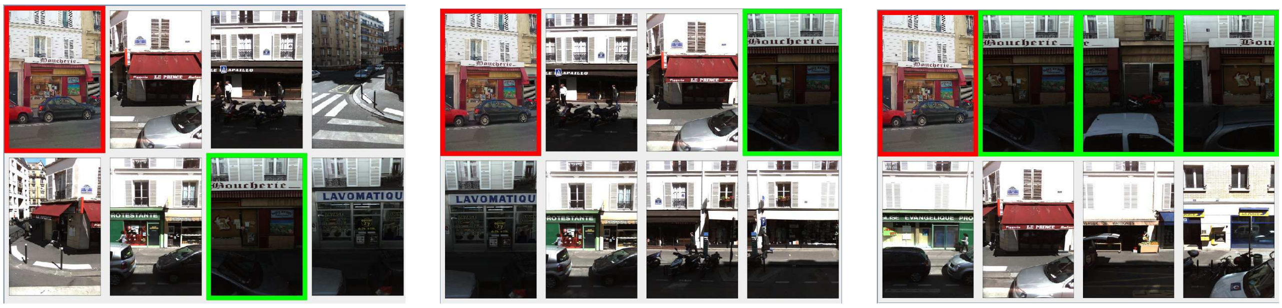

Figure 2: A sample query using the Brute Vote technique (left) and enhanced by geometric consistency by Rotation Histogram (middle) and RANSAC (right). The query image (red border) is followed by the ranked answers (right-left, top-down), some of which are relevant (green border).

little difference between the methods, and only as the number of retrieved images gets very large (over 30), RANSAC starts to show some advantage over the other methods.

5. DISCUSSION

Adding geometric consistency constraints seems to be essential to the success of document retrieval in our application context. Among the two strategies examined, RANSAC was the only able to provide satisfactory results. However, the technique seems to be highly dependent on the nature of the documents. In the case of a small database, with moderately distorted queries (like our first study) the results are good enough to be used in the intended application. In our second study (with more challenging conditions: larger database, problematic features, highly distorted queries) the quality of the results suffered.

It is probable that geometric consistency alone will not be enough to achieve high-quality results in contexts like ours, where a large number of non-discriminating descriptors are spawned by problematic image features like tree branches and complex shadows. Those descriptors, which match at random on the database, increase dramatically the number of false matches, inflating the rank of non relevant images (Figure 5). An important addition to the architecture would be the ability to recognize and eliminate those non-informative descriptors.

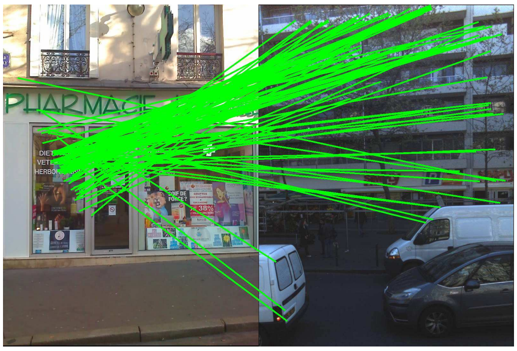

Figure 5: False matches between two images after geometric consistency check with RANSAC.

The fact that Pairwise Matching performed so badly was somewhat surprising, since it can benefit from a distance contrast criterion (explained in [5]) aimed at automatically rejecting bad matches. The most probable explanation is that it does not benefit from the “dilution” of incorrect matches on the entire database (since those matches are made at random, they tend to get evenly distributed among all images).

To conclude, we consider extending the voting-based architecture to our application very challenging. Advances in the state of the art are needed both in terms of descriptors quality and matching accuracy. We intend to share our databases with the community in order to allow the benchmarking of those tasks on this difficult context.

6. REFERENCES

[1] L. Amsaleg and P. Gros. Content-based retrieval using local descriptors: Problems and issues from a database perspective. *Pattern Anal. Appl.*, 4(2-3):108–124, 2001.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[3] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *The VLDB Journal*, pages 518–529, 1999.

[4] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In A. Z. David Forsyth, Philip Torr, editor, *European Conference on Computer Vision*, volume I of *LNCS*, pages 304–317. Springer, oct 2008.

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.

[6] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):530–535, 1997.

[7] E. Valle, M. Cord, and S. Philipp-Foliguet. Fast identification of visual documents using local descriptors. In *DocEng ’08: Proceeding of the eighth ACM symposium on Document engineering*, pages 173–176, New York, NY, USA, 2008. ACM.

[8] E. Valle, M. Cord, and S. Philipp-Foliguet. High-dimensional descriptor indexing for large multimedia databases. In *CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 739–748, New York, NY, USA, 2008. ACM.