Image point correspondences and repeated patterns

Frédéric Sur, Nicolas Noury, Marie-Odile Berger

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Image point correspondences and repeated patterns

Frédéric SUR — Nicolas NOURY — Marie-Odile BERGER

N° 7693

July 2011

Vision, Perception and Multimedia Understanding

apport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7693--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
NANCY - GRAND EST

Image point correspondences and repeated patterns

Frédéric SUR , Nicolas NOURY , Marie-Odile BERGER

Theme : Vision, Perception and Multimedia Understanding
Équipe-Projet Magrit

Rapport de recherche n° 7693 — July 2011 — 50 pages

Abstract: Matching or tracking interest points between several views is one of the keystones of many computer vision applications. The procedure generally consists in several independent steps, basically interest point extraction, then interest point matching by keeping only the "best correspondences" with respect to similarity between some local descriptors, and final correspondence pruning to keep those that are consistent with a realistic camera motion (here, consistent with epipolar constraints or homography transformation.) Each step in itself is a delicate task which may endanger the whole process. In particular, repeated patterns give lots of false correspondences in descriptor-based matching which are hardly, if ever, recovered by the final pruning step. We discuss here the specific difficulties raised by repeated patterns in the point correspondence problem. Then we show to what extent it is possible to address these difficulties. Starting from a statistical model by Moisan and Stival, we propose a one-stage approach for matching interest points based on simultaneous descriptor similarity and geometric constraint. The resulting algorithm has adaptive matching thresholds and is able to pick up point correspondences beyond the nearest neighbour. We also discuss Generalized RANSAC and we show how to improve Morel and Yu's ASIFT, an effective point matching algorithm to make it more robust to the presence of repeated patterns.

Key-words: Point correspondence problem, repeated patterns, perceptual aliasing, double nail illusion, a-contrario model, generalized RANSAC, SIFT, ASIFT.

Centre de recherche INRIA Nancy – Grand Est
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Mise en correspondance de points dans les images et motifs répétés

Résumé : L'appariement ou le suivi de points d'intérêt entre plusieurs images est la brique de base de nombreuses applications en vision par ordinateur. La procédure consiste généralement en plusieurs étapes indépendantes, à savoir : l'extraction des points d'intérêt, puis l'appariement des points d'intérêt en gardant les « meilleures correspondances » selon la ressemblance de descripteurs locaux, et enfin l'élagage de l'ensemble des correspondances pour garder celles cohérentes avec un mouvement de caméra (ici, cohérentes selon les contraintes épipolaires ou une homographie globale). Chaque étape est une tâche délicate qui peut compromettre le succès du processus entier. En particulier, les motifs répétés génèrent de nombreux faux appariements qui sont difficilement rattrapés par l'élagage final. Dans ce rapport nous discutons les difficultés spécifiques soulevées par les motifs répétés dans l'appariement de points. Ensuite nous montrons dans quelle mesure il est possible de dépasser ces difficultés. En reprenant un modèle statistique proposé par Moisan et Stival, nous proposons une nouvelle approche prenant en compte simultanément la ressemblance des descripteurs et la contrainte géométrique. L'algorithme a des seuils d'appariement adaptatifs et est capable de sélectionner des correspondances au delà du plus proche voisin. Nous discutons aussi RANSAC généralisé et nous montrons comment améliorer ASIFT de Morel et Yu pour le rendre robuste à la présence de motifs répétés.

Mots-clés : Mise en correspondance de points d'intérêt, motifs répétés, aliasing perceptuel, illusion des deux ongles, modèle a-contrario, RANSAC généralisé, SIFT, ASIFT.

Contents

1 Introduction

A large part of computer vision literature is based on the matching of interest points between several views. "Matching" means that one has to detect interest points across several images that correspond to the same actual 3D point. This is often achieved by taking into account local descriptors, i.e. an encoding of the grey values from the vicinity of an interest point. While only a rough matching is needed in e.g. the image retrieval context (it is accepted that some correspondences are not correct), photography stitching [5] and most multiple views structure and motion applications [22] (see e.g Snavely et al.'s Phototourism and Bundler [55]) call for an accurate matching step. In this report we focus on the problem of correspondence finding between two views. Let us consider two views from the same 3D scene taken by a moving camera. A popular way to tackle this problem consists in the following steps:

1. In both views, extract interest points along with a descriptor of the local photometry.
2. Match them by taking into account some (dis-)similarity measure over the descriptors.
3. Prune the correspondences by finding out the most consistent set with respect to the geometry imposed by a realistic camera motion.
4. Estimate the camera motion between the two views. Then make the set of correspondences "denser" by relaxing the matching step 2 and taking into account this estimation. (This step is often referred to as "guided matching".)
5. Estimate the refined camera motion based on this final set of correspondences.

Interest point extraction in **step 1** can be achieved by Harris-Stephens corner detector [20], extrema of the Laplacian or of the determinant of Hessian [27] in scale-space, etc. Following the seminal work by Mohr and Schmid [53] a large amount of methods have emerged to attach to each interest point a local photometric descriptor, (quasi-) invariant to contrast change and to a large class of deformations. One of the most successful algorithms is probably Lowe's SIFT [31], which is based on this idea. See the reviews [1, 34, 35, 38].

Step 2 is certainly one of the very shortcomings of the method. It is indeed difficult to endow the space of descriptors with a handy metric. Putting a threshold over the Euclidean distance between descriptors to define correspondences simply does not work. A popular way [31] to define a set of correspondences is instead to keep the nearest neighbour, and optionally impose that the ratio of the distances between the nearest and the second nearest neighbour is below some threshold (obviously smaller than 1.) The nearest neighbour is indeed all the more relevant as the ratio is low. It works quite well even using the Euclidean distance. However, since most descriptors (and especially SIFT) are made of gradient orientation histograms, some authors propose to change the Euclidean distance to some distance that is somewhat more adapted to histograms. We can mention (by increasing computational complexity) χ^2 distance, Ling and Okada's diffusion distance [28], Earth Mover's Distance (EMD, see the seminal work by Rubner et al. [49], Rabin et al. [45], or Ling and Okada [29].) For the sake of historical completeness, let us also mention correlation methods (e.g. [68] and more recently [57]), that do not need descriptors to build point correspondences. However,

these latter methods suffer from the lack of invariance and are preferentially kept for small baseline stereovision.

Once a set of tentative correspondences has been defined from both images, **step 3** aims at selecting a subset made of correspondences that are consistent with the underlying geometric model. In the pinhole camera model, the so-called *epipolar* geometry is encoded in the *fundamental* matrix (or the *essential* matrix if intrinsic parameters are known) [15, 22], except if the camera motion is restricted to a rotation around its optical center, or if the 3D points are coplanar, in which case a *homography* (a planar projective transformation) maps interest points from one image to the other one. Since correspondences are spoilt by outliers (that is, correspondences between parts of images that look alike, but do not correspond to the same actual 3D object), robust statistics are called for, such as e.g. LMedS or M-estimators [59, 67]. The most popular choice is certainly RANSAC [16] and methods derived from it (MSAC and MLESAC [60], MAPSAC [58], PROSAC [9] to only cite a few.) The RANSAC paradigm deserves some attention in this discussion. RANSAC is an iterative procedure, that is based on two steps: a) draw a minimal sample to estimate the model, and b) build a subset of correspondences that is consistent with this model. This latter set is called *consensus set*. In the end, the "most consistent" set is kept. Consistency is measured by basically counting the cardinality of the consensus set (original RANSAC) or by some more sophisticated fitness measure (MSAC, MLESAC.) When running RANSAC-like algorithms, the user needs to tune several parameters by hand, which may be quite tricky. Recently, Moisan and Stival [36] have proposed a new RANSAC-like procedure to estimate the two-view geometry. Their algorithm is based on a statistical measure which does not need parameter tuning and is shown to behave as well as state-of-the-art methods with large rates of outliers. We will come back to this in section 4.

Once a consensus set has been found, **step 4** consists in estimating the geometry between the two views, by computing the fundamental matrix or homography. Then an optional stage follows: new correspondences are found by searching them along the epipolar lines. This step gives a set of correspondences which is hopefully distributed across both images in a "denser" fashion. This should allow a more reliable re-estimation of the geometry, based on this final set of correspondences. This is the goal of **step 5** where many methods have been proposed [15, 22, 67]. We do not elaborate on these steps in the present paper. However, whatever the ingenuity of steps 4 and 5, the set of corresponding points from steps 1-3 has to be good enough so that camera motion can be reliably estimated.

The reader can easily see that putting all these steps together is practically a difficult task. It indeed involves setting a lot of parameters, and a wrong choice for one of them may endanger the whole process.

Besides, repeated patterns bring specific problems. Repeated patterns are common in man-made environments (just think of windows on a façade or manufactured goods as cars in outdoor environments.) If the matching step is just based on the nearest neighbour conditioned by the distance ratio between the nearest and second nearest neighbour (as in standard SIFT matching, see step 2), it is obvious that repeated patterns are likely to be discarded at this early stage (since the ratio would be always close to 1.) Moreover, there is no insurance that the correspondence that is still kept is correct, as illustrated on figure 1. Although some methods are better than other towards this point (for example [45]) it is still very difficult to match repeated patterns in a reliable fashion. This is a crucial issue in many applications, as e.g. in structure from motion where it yields "catastrophic failures" in the reconstruction of the scene [47].

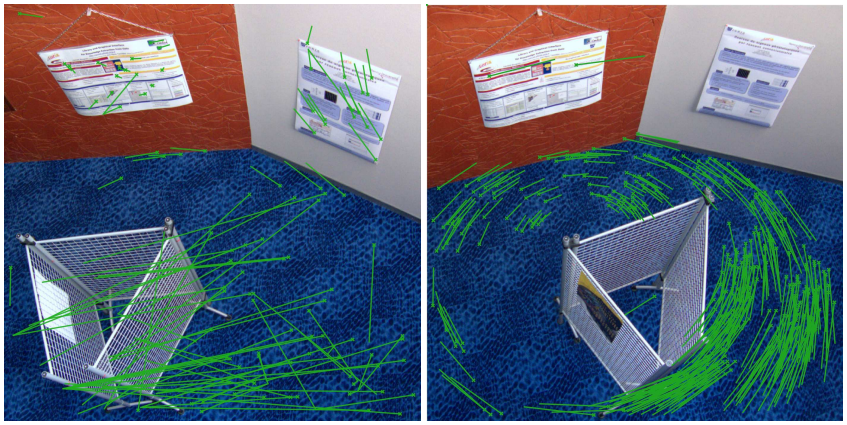

Figure 1: *Loria* image pair. SIFT interest points are marked with a cross, the green segment represents the apparent motion with the matching feature in the other image. On the left, the standard SIFT matching algorithm (NN-T in text) fails at identifying reliable matching interest points on the carpet. Thus, no subsequent pruning algorithm will succeed in drawing out the true correspondences. On the right, the proposed a-contrario approach using both photometric and geometric constraints (here a homography) finds correct correspondences in spite of the large number of repeated patterns.

The first contribution of this report is to discuss the specific problems raised by repeated patterns in the point correspondence problem (section 2.) Section 3 presents related work. The aim of the rest of the report is to illustrate how and to what extent the problems induced by repeated patterns can be overcome. Section 4 proposes a statistical a-contrario framework to replace steps 2 and 3 (and partly 4) to select correspondences beyond the nearest neighbours. The whole method is summarized in section 5 where the algorithmic choices are motivated. Section 6 is about experimental assessment and proof of concept. In particular, we also discuss a Generalized RANSAC which is also able to pick up non-nearest neighbour correspondences but necessitates tuning several parameters. We explain in section 7 how ASIFT [39], a very effective point matching algorithm robust to extreme viewpoint change, can be made more robust to repeated patterns. For the sake of completeness, the proofs of some propositions are given in appendix.

2 Repeated patterns: the curse of perceptual aliasing

As mentioned in the introduction, deciding correspondences between interest points based on a nearest neighbour criterion on descriptors is not sound when repeated patterns are present. The problem is that patches around interest points yield descriptors that are as much invariant as possible to viewpoint changes. Thus numerous descriptors from repetitive structures are very similar, making it impossible to infer correct correspondences only based on the conveyed information.

Let us remark that this phenomenon is strongly related to the so-called *perceptual aliasing* from the control of systems theory. This term was coined by Whitehead and Ballard [64] to describe the fact that a robot may possibly not distinguish between

different states of the world due to the limited accuracy of its sensors. Let us quote Whitehead and Ballard [64]: "*Perceptual aliasing can be a blessing or a curse. If the mapping between the external world and the internal representation is chosen correctly, a potentially huge state space (with all its irrelevant variation) collapses into a small simple internal state space. Ideally, this projection will group world situations that are the same with respect to the task at hand. But, if the mapping is not chosen carefully, inconsistencies will arise and prevent the system from learning an adequate control strategy.*" In the point correspondence framework, representing a world state (here: an actual 3D point) with a single internal state (here an invariant descriptor) is a desirable feature so that point matching under large viewpoint change is easily tractable. However, the presence of repeated structures makes different 3D points be internally represented by a set of very similar descriptors. It is then very hard, and sometimes impossible, to infer correspondences consistent with the true image registration from these confused representations.

For example, figure 2 shows point correspondences from which it is clearly impossible to get a consistent set. The left image in figure 1 is another example. Remark that we could increase the scale of the patches which yield the almost identical descriptors, until they are distinguishable from one another, as in e.g. [63]. However, this strategy would deteriorate the robustness to occlusion or clutter, as well as the invariance to viewpoint change, since this is based on the hypothesis that a similarity or affine transformation is a local approximation of a projective transformation.

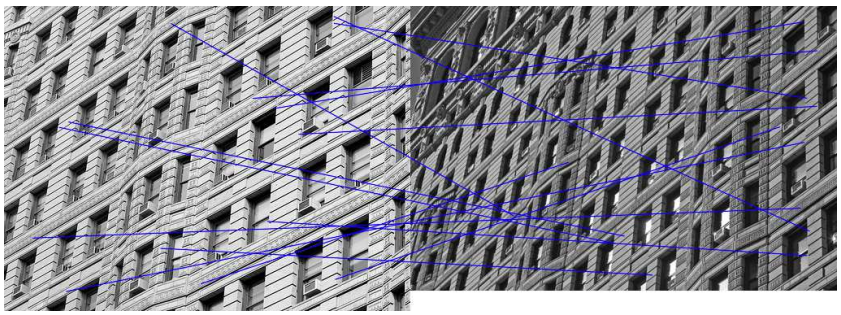

Figure 2: *Flatiron* image pair. SIFT correspondences (nearest neighbour + distance ratio criterion). It is simply impossible to extract a consistent set of correspondences.

Most of the time it is still possible to get a *consistent* set of correspondences from an image pair, in the sense that it can be the consequence of a plausible camera motion, just based on the information of the invariant descriptors, but not *correct* compared to the ground truth. Such an example can be seen on figure 3: the largest set of point correspondences consistent with a homography is actually not correct and consists of shifted patterns. In this case, the standard approach (considering photometric and geometric information independently) does not succeed in recovering from perceptual aliasing. The reason is that nearest neighbour matching tends to associate patches with the same absolute size (in pixels) because of the limited invariance of the descriptors to scale change.

Let us also remark that a certain amount of perceptual aliasing cannot be resolved just from images. For example, in figure 3 we implicitly assume that the correct motion is the minimal one, although a 90° rotation of the cube would be possible.

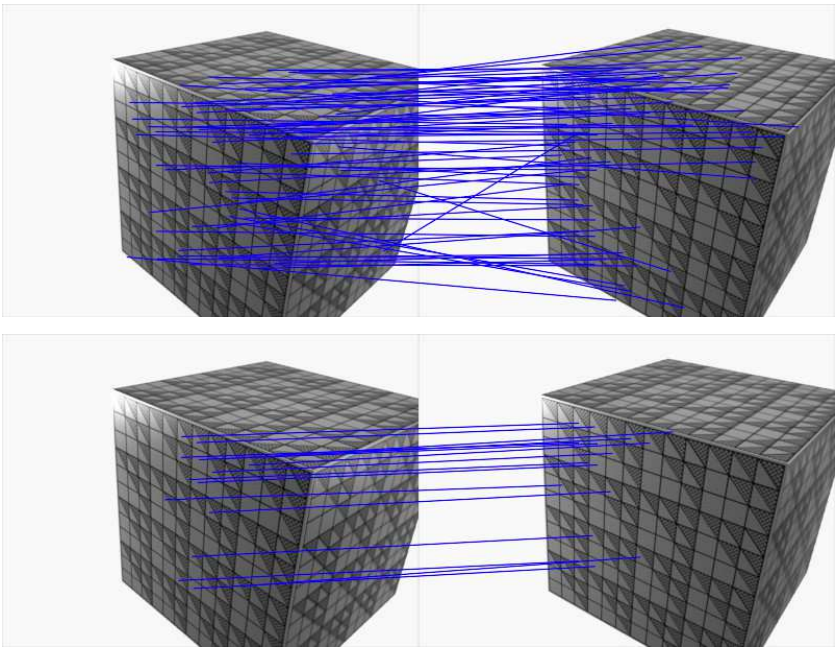

Figure 3: *Synthetic cube*. Top: SIFT correspondences (nearest neighbour + distance ratio criterion). Some correct correspondences can be seen. Bottom: the largest set consistent with a homography between the two images. As we can see, the 15 correspondences are not correct and map to shifted patterns.

Another situation is characteristic of scenes with repeated patterns. Even if the camera motion is properly estimated, lots of false correspondences can occur "by chance". For example in figure 4, we can see false correspondences on the carpet that are kept because they satisfy the nearest neighbour matching, and also the epipolar constraint, in the sense that each interest point lie on the associated epipolar line of its counterpart. This situation is particularly worrying in Structure from Motion applications: correspondences between repeated patterns may give a potentially large number of spurious reconstructed 3D points by triangulation.

This situation was known in the psychovision community from the eighties, and is named the *double nail illusion* after a nice article by Krol and van de Grind [23]. It is illustrated by figure 5 in the context of two-view geometry. We can see that any set of repeated patterns lying on an epipolar plane is likely to give arbitrary point correspondences. Some point correspondence problems are thus intrinsically ambiguous if the information is restricted to the invariant descriptors. It is illustrated by a purely theoretical reasoning here, but wrong correspondences due to repeated patterns are common in practical two-view problems, as in figure 4. Note that scale invariance discards any depth-related information (by making it impossible to decide based on relative size of the features), in a similar way as the *vergency trap* in the original double nail illusion [23]. Incorporating the scale information in the matching process as in [61] or the relative position of the interest points [52] should probably help disambiguating such situations. Another possibility would be to to group the repeating structure beforehand (in order to use the groups as non-repeated features) as in [26, 50, 51, 62]. To the best

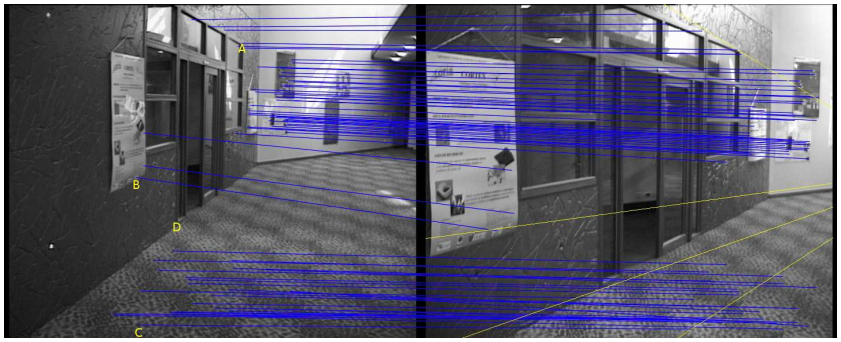

Figure 4: *Loria corridor*. Nearest neighbour matching, followed by epipolar RANSAC. Correct correspondences can be seen on the walls, and a large number of false ones on the carpet. The epipolar lines in the right image associated with points A and B (from the left image) naturally go through the corresponding point. The same holds for point C whose corresponding point in the right image is not correct. However, the camera motion is properly estimated: for example the epipolar line associated with point D actually go through the corner of the doorway in the other image, in spite that no correspondence is detected in its neighbourhood. Here, the correspondences on the carpet are kept because they satisfy simultaneously the nearest neighbour matching and the epipolar constraint.

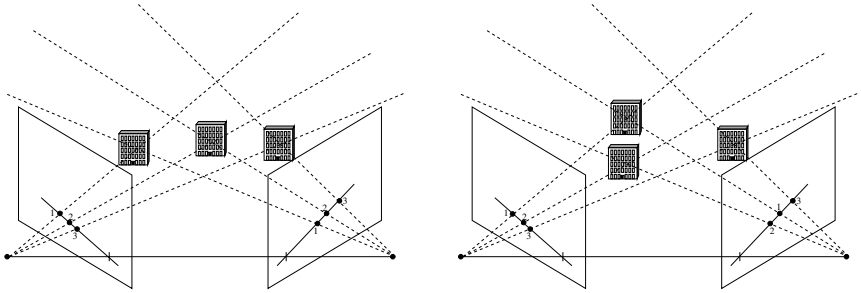

Figure 5: *The "multiple" nail illusion*. We consider here two views of a 3D scene with repeated patterns. The image plane is represented between the optical center and the scene. Even if the two cameras are calibrated (and thus the epipolar constraints is known), some situations cannot be distinguished just from the information contained in the invariant descriptors. Here, the three black dots in both images are interest points with almost identical descriptors (because they correspond to repeated patterns), and since they lie on a common epipolar plane they can be associated in any of the six possibilities, yielding any of the six possible respective positions in the 3D scene (two are shown here.) Of course, only a single possibility is physically feasible. Let us remind that the scale associated with the interest points is not taken into account in the descriptor based matching.

of our knowledge, this possibility has not been yet investigated further. The aim of the strategy used in section 7 for ASIFT is actually to disambiguate double nail illusion by enforcing relative positions in both images.

The double nail illusion yields false correspondences between repeated patterns. In some situations, it can even trap the matching algorithm and return a model with epipolar lines which do not correspond to the camera motion and which actually are vanishing lines (see figure 6.) The phenomenon is explained in figure 7. Note that this situation is related to the so-called *auto-epipolar matching* by Mundy and Zisserman [40]. In this latter paper, the authors argue that a single view of a duplicate object at two different positions is equivalent to two views of a single object, thus it is possible to define epipolar lines. Auto-epipolar matching is the case where corresponding points lie on the same epipolar line. Figures 6 and 7 actually correspond to auto-epipolar matching. Let us remind that the geometric model is usually fitted with RANSAC: the degenerate situation gives the largest consensus set and wins over smaller but correct sets of correspondences. The direction of the vanishing / epipolar lines corresponds to the direction of the longest 3D lines containing repeated patterns, yielding the largest auto-epipolar matching.

In this discussion we have pointed out that independently considering descriptor similarity and geometric constraint yields difficulties in the point correspondence problem. In the following section we present related works that aim at overcoming these difficulties. We come back to the double nail illusion in sections 6 and 7.

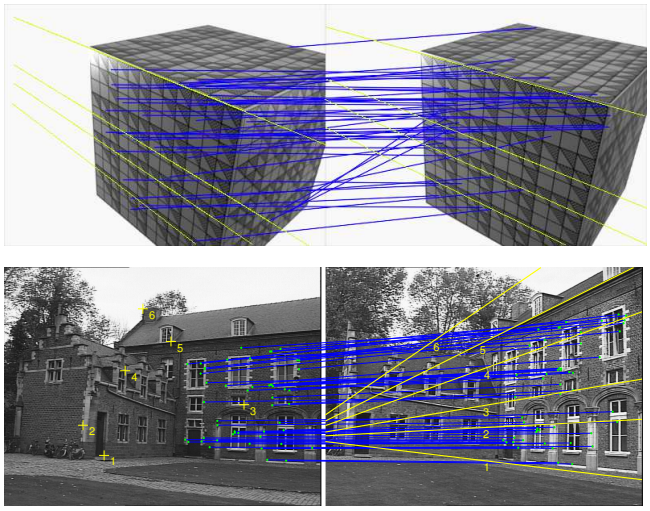

Figure 6: *Trapped by the double nail illusion: examples*. Correspondences here satisfy both descriptor similarity and epipolar constraints. However, many false correspondences can be seen (on the lower pair, correspondences can be seen with the fourth windows which is not visible in the first image), and the epipolar lines (in yellow) correspond to vanishing lines, yielding an incorrect camera pose estimation.

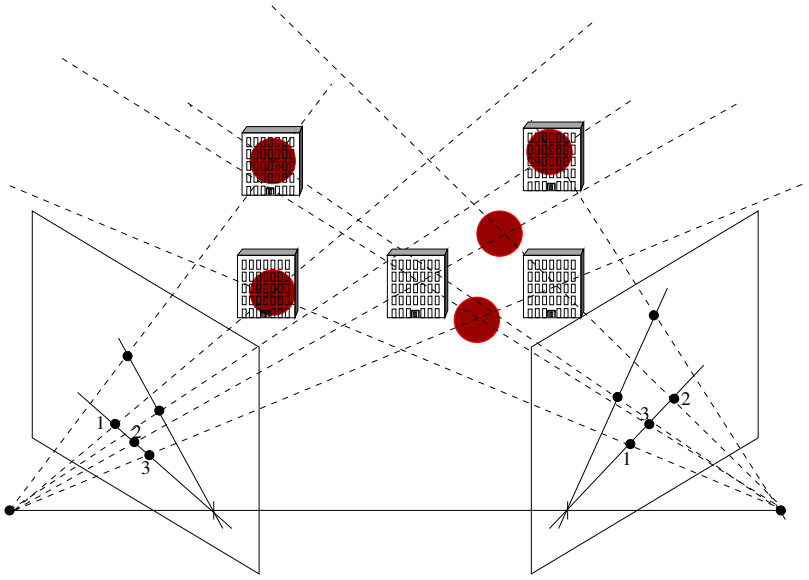

Figure 7: *Trapped by the double nail illusion: illustration*. Starting with the two views of the unknown 3D scene (repeated patterns on a vertical plane), the interest points (black dots) are not correctly associated. From these correspondences, the position of the 3D points is (incorrectly) inferred at the red disks. Consequently, the scene is not seen as planar, and there is no ambiguity on the retrieved epipolar geometry. Due to the auto-epipolar matching, the epipolar lines are here vanishing lines of the scene, and the epipoles are vanishing points.

3 Related work

Reliably matching interest points is a question which is often brought up by the literature. Several articles try to overcome the difficulties arising from the two-step correspondence finding (as described in the introduction) by circumventing it. To address the problem of "correspondence-free" structure from motion, a possibility is to use brute force techniques (i.e. considering all possible correspondences among extracted interest points), guided by some heuristics.

To the best of our knowledge, correspondence searching without prior photometric matching was for the first time extensively studied by Dellaert et al. [10]. However, their approach is purely combinatorial. They explicitly *"adopt the commonly used assumption that all features x_j are seen in all images, i.e. there are no spurious measurements and there is no occlusion"* [10]. We believe that the basic assumption is too restrictive to deal with occlusion and point misdetection, which often arise in practice. A recent paper [17] by Georgel et al. suggests how it could be possible to solve the two-view correspondence problem without any photometric information. Basically, it consists in smartly discretizing the set of essential matrices that enforce the geometric constraints. However, in their framework the combinatorial burden is seriously restricted by the set of possible camera motions between the two views, which restricts the degrees of freedom of the essential matrix from 5 to 2.

A way to reduce the computational complexity is to use photometric information along geometric constraints. Domke and Aloimonos [14] present a solution which consists in establishing an a priori probabilistic model for the correspondence distribution, computed for every image pixel. They do not need any preliminary matching step between interest points. Since the 5D space of all possible motions (in the calibrated case) must be explored, this approach has a heavy computational cost as in [10], although speeding up is possible when a preliminary motion estimation is known. The same basic idea was used before by Roy and Cox [48], who compute the photometric likelihood that points lie on the corresponding epipolar line, and aim at maximizing it over all possible motions. A similar idea is presented by Antone and Teller [3] in the context of omni-directional image networks. They indeed estimate the baseline between two views by considering all possible feature correspondences satisfying epipolar constraints. They propose to solve this high-complexity problem by constraining the search through feature similarity.

Another way of incorporating photometric and geometric constraint for structure and motion estimation has been investigated by Stein and Shashua in [56]. Optical flow provides photometric information but suffers from the well known *aperture problem* which is painful for scenes with long straight edges (and also suffers from the constant intensity assumption, which is on the contrary overcome by contrast invariant descriptors such as SIFT.) To avoid this, the authors of [56] propose to build the so-called *tensor brightness constraint* which is based on both the optical flow and the trifocal tensor which encodes the geometry between three views [22]. However, as every method based on the optical flow, this cannot be extended to large transformations between views, which would make the optical flow estimation unreliable. No explicit point correspondence is needed.

Some very recent works discuss the use of Radon transform for correspondence-free motion estimation. Lehmann et al. [24] focus on the determination of the affine fundamental matrix (that corresponds to the case of orthographic camera model.) The Fourier transform of different views are related through the parameters of the motion of the camera since image rotation and translation lead to spectrum rotation and phase

change. Motion parameters are retrieved by matching lines in the Fourier domain with a dedicated EM algorithm. Experimental results are promising; for the moment this approach is intrinsically restricted to the orthographic model which is less complex than the full epipolar model. However occlusions are not handled. The same idea is used by Makadia et al. [32]; the Fourier transform is used to generate a global likelihood function on the space of all observable camera motions, which appears to be (quite) easily tractable in the Fourier domain, although a careful discretization is needed. Results mainly concern catadioptric cameras.

All of the preceding "correspondence-free" models compute the camera motion by global view matching and are therefore not robust to occlusions and to small overlaps between images. Explicitly using interest points allows to deal with these latter shortcomings. From this point of view, we can cite recent works [9, 19, 57] which take into account photometric similarity to guide the search for correspondences that are consistent with camera motion. In these articles, which all propose improvements of the RANSAC algorithm, the goal is mainly to speed up the search for a consensus set. The common idea is to use the similarity between descriptors to guide the search: sampling is no more uniform as in classic RANSAC but is weighted by the similarity prior. However, a first step consisting in a photometric matching is still needed, and the problem of repeated patterns is not really tackled. Let us precise that in their paper about Guided-MLESAC [57], Tordoff and Murray mention the possibility of getting rid of the photometric matching step by incorporating the photometric information in the prior. Although no evidence is given, it should theoretically also improve their algorithm when facing repeated patterns.

Disambiguating correspondences when repeated patterns are present can be achieved by taking account of the local organization of the interest points. Deng et al. [11] associate SIFT descriptors with a region context descriptor which encodes the relative position of nearby interest points. Their so-called "reinforcement matching" directly takes into account geometric information from the "region context". As in our algorithm, matching is not restricted to nearest neighbours. Hence, to some extent, it should be able to disambiguate a certain amount of perceptual aliasing, although no evidence is given in [11]. In the same spirit, Aguilar et al. [2] propose a matching process which is based on the observation that the relative position of the interest points is preserved in both views, provided the viewpoint change is limited. Thus they define in each view a K-nearest neighbour graph. Point correspondences are obtained by matching the two graphs: outliers are detected if the graphs are locally non-isomorphic. This method seems promising and shows better results than the popular Softassign algorithm [18]: it is generic and robust to repeated patterns, although the geometric constraints are not explicitly enforced. This has been recently used to improve Morel and Yu's ASIFT [39] with respect to repeated patterns in [4].

While very few generic point matching algorithms explicitly tackle the repeated pattern problem, some authors first group features sharing the same aspect [26, 50, 51, 62]. For example, Schaffalitzky and Zisserman [50] determine vanishing lines in a single image, by pairing aligned repeated patterns. By detecting repeated structures in building façades, it is possible to achieve pose estimation in urban scenes [51].

The very recent work by Serradell et al. [54] shares several common points with the proposed a-contrario method. They make use of geometric and appearance priors to guide homography searching. While we also use appearance prior (which we call "photometric constraint" in section 4.3), our geometric prior is much simpler since it is derived by the random sampling of the tentative correspondences as in RANSAC. Although Serradell et al.'s method needs several algorithmic choices it gives promising

results with repeated patterns and even strong viewpoint changes. It is not limited to the nearest neighbour when matching interest points.

4 An a-contrario model for point correspondences under epipolar constraint and photometric consistency

In this section we propose a method based on a statistical a-contrario model. Since the seminal paper by Desolneux, Moisan and Morel [12], these models have been the subject of a large amount of literature. The books [8] and [13] and the references therein give a comprehensive account of their use in many different computer vision problems. In [7] and [41] several a-contrario models are designed for correspondence finding between two views. Nevertheless, these models deal with *geometrical shapes* under *affine transformations* instead of *interest points* under *epipolar and homographic constraints* as in the framework presented here.

The idea behind a-contrario models is that independent, structure-less random features can produce structured groups only with a very small probability. This claim is sometime called the *Helmholtz principle* in the a-contrario literature. The model proposed in this report is based on Moisan and Stival's a-contrario RANSAC [36] and on Rabin et al.'s a-contrario model for SIFT-like descriptor matching via an Earth Mover's Distance [45, 46]. The first paper [36] focuses on geometric constraints and assumes that correspondences between interest points are given by some prior step. It also gives an indication of how to find out correspondences based on geometry and photometry (the so-called "colored rigidity" criterion.) Our contribution consists in generalizing Moisan and Stival's algorithm to incorporate both epipolar constraint and photometric consistency. We also specify the implementation and build up heuristics to make the matching task tractable. The latter papers [45, 46] prove that Earth Mover's Distance is better than existing dissimilarity measures between SIFT features, and investigate several a-contrario approaches.

Let us give some notations. We assume that two views (images $\mathcal{I}_1$ and $\mathcal{I}_2$) from the same scene are given. For each image, some algorithm (for example SIFT) gives a set of interest points, along with a descriptor. Let us note $(x_i, D(x_i))_{1\leq i\leq N_1}$ (resp. $(y_j, D(y_j))_{1\leq j\leq N_2}$) the N_1 (resp. N_2) couples from $\mathcal{I}_1$ (resp. $\mathcal{I}_2$) such that x_i (resp. y_j) is the coordinate vector of an interest point, and $D(x_i)$ (resp. $D(y_j)$) is the corresponding local descriptor. Depending on the circumstances, we denote x_i the interest point itself, its pixel coordinates, or its homogeneous coordinates in the projective plane.

We assume to be within the scope of the pinhole camera model. In this framework, if x_i and y_j are the projections in $\mathcal{I}_1$ and $\mathcal{I}_2$ of the same 3D point, then y_j lies on the epipolar line associated with x_i. This line is represented by $F \cdot x_i$, where F is the fundamental matrix from $\mathcal{I}_1$ to $\mathcal{I}_2$. Conversely, x_i has to lie on the epipolar line $F^T \cdot y_j$ since the fundamental matrix from $\mathcal{I}_2$ to $\mathcal{I}_1$ is the transpose matrix F^T. However, if the camera has just been rotated around its optical center between the two views, or if interest points lie on a common plane, then the fundamental matrix is not defined. In this case, there is a 2D projective transformation (a homography) H such that $y_j = H(x_i)$ and $x_i = H^{-1}(y_j)$.

If the local descriptors were invariant to projective transformations, then $D(x_i)$ and $D(y_j)$ should be theoretically identical. However, such an invariance is practically unreachable without 3D information. With the additional assumption that the 3D scene

is locally planar, then invariance to homographies is just needed. Such an approach is used e.g. in [37]. Most of the time, a weaker invariance is still satisfying, namely invariance to affine transformations [35] or to zoom+rotation (similarity) transformations which is easier to handle in practice, as in Lowe's SIFT [31]. Consequently, the descriptors $D(x_i)$ and $D(y_j)$ are "similar" but not identical as soon as the viewpoint change does not exactly amount to an affine or similarity transformation.

The problem of interest is therefore to find a subset $\mathcal{S}$ of $\{1,\dots,N_1\}\times\{1,\dots,N_2\}$ and a fundamental matrix F or a homography H from $\mathcal{I}_1$ to $\mathcal{I}_2$ such that:

1. The distance between corresponding descriptors is below some threshold δ_D, ensuring that the local image patches are alike:

$$\forall (i,j) \in \mathcal{S},\ d_D(D(x_i), D(y_j)) \leq \delta_D. \tag{1}$$

2. The distance between a point and the epipolar line associated with the corresponding point is below some other threshold δ_G (and vice versa), ensuring that the epipolar constraint is satisfied:

$$\forall (i,j) \in \mathcal{S},\ d_G(x_i, y_j, F) := \max\{d_G(y_j, F\cdot x_i), d_G(x_i, F^T\cdot y_j)\} \leq \delta_G. \tag{2}$$

Alternatively for the homography constraint:

$$\forall (i,j) \in \mathcal{S},\ d_G(x_i, y_j, H) := \max\{d_G(y_j, H(x_i)), d_G(x_i, H^{-1}(y_j))\} \leq \delta_G. \tag{3}$$

Remark that symmetrization with respect to $\mathcal{I}_1$ and $\mathcal{I}_2$ in equations (2) and (3) could have been achieved in other manners. Here, the product-distance is used.

In the sequel we shall give a definition of both distances (or dissimilarity measures) d_D and d_G. The proposed statistical framework automatically balances geometry and photometry, and also automatically derives both thresholds δ_D and δ_G relatively to a set $\mathcal{S}$.

4.1 The a-contrario model

Before specifying distances d_D and d_G, we explain the statistical model that will help us in making decisions. In the a-contrario method, groups of features are said to be *meaningful* if their probability is very low under the hypothesis $\mathcal{H}_0$ that the features are independent. Independence assumption makes the probability computation easy, since joint laws are simply products of marginal laws which can be reliably estimated with a limited number of empirical observations. Without independence assumption, joint law estimation would indeed come up against the *curse of dimensionality*. In the statistical hypothesis testing framework, this probability is called a p-value: if it is low (typically below 5%), then it is likely that the group of interest does not satisfy independence assumption $\mathcal{H}_0$. There must be a better explanation than independence for this group, and this explanation should emphasize some common causality. Here, pairs of features form a meaningful group because interest points from a pair actually correspond to the same 3D point, and the motion of all interest points between the two views is consistent with the motion of the camera.

Let us assume that a set $\mathcal{S}$ of correspondences is given, as well as a fundamental matrix or a homography A and two thresholds δ_G and δ_D as in equations (1) and (2).

The probability that should be estimated is:

$$p(\mathcal{S}, A, \delta_G, \delta_D) := \Pr\big(\forall (i,j) \in \mathcal{S}, d_G(x_i, y_j, A) \leq \delta_G \text{ and } d_D(D(x_i), D(y_j)) \leq \delta_D \mid \mathcal{H}_0\big) \tag{4}$$

with d_G as in equations (2) or (3).

Let us also assume that the transformation A is estimated from a minimal subset of $\mathcal{S}$ as in the RANSAC paradigm. This means in the case where $A = F$ is a fundamental matrix that a subset s from $\mathcal{S}$ made of $m = 7$ correspondences is used to estimate F [67]. Remark that it would also be possible to use the 8-point linear method [30] with a slight adaptation. In the case where $A = H$ is a homography A, $m = 4$ points are needed in s. In the sequel, $S \backslash s$ is the set of correspondences in S that are not in s.

Definition 1 *Considering $(x_i, D(x_i))$ and $(y_j, D(y_j))$ as random variables, we define hypothesis $\mathcal{H}_0$ as:*

1. *$(d_D(D(x_i), D(y_j)))_{(i,j)\in\mathcal{S}}$, and $(d_G(x_i, y_j, A))_{(i,j)\in\mathcal{S}\backslash s}$ are mutually independent random variables.*
2. *$(d_G(x_i, y_j, A))_{(i,j)\in\mathcal{S}\backslash s}$ are identically distributed and their common cumulative distribution function is f_G*
3. *$(d_D(D(x_i), D(y_j)))_{(i,j)\in\mathcal{S}}$ are identically distributed and their common cumulative distribution function is f_D.*

Of course, $(d_G(x_i, y_j, A))_{(i,j)\in s}$ are also identically distributed but do not follow the same distribution function f_G as variables from $\mathcal{S}\backslash s$ since A is estimated from s, leading to the conditions $d_G(y_j, A(x_i)) \simeq 0$ and $d_G(x_i, A^{-1}(y_j)) \simeq 0$ for every $(i,j) \in s$.

As a consequence, it is possible to estimate the probability from equation (4):

Proposition 1

$$p(\mathcal{S}, A, \delta_G, \delta_D) = f_D(\delta_D)^k f_G(\delta_G)^{k-m} \tag{5}$$

where k is the cardinality of $\mathcal{S}$, and m the cardinality of s.

Proof: it is straightforward to derive:

$$p(\mathcal{S}, A, \delta_G, \delta_D) = \prod_{(i,j)\in\mathcal{S}\backslash s} \Pr\big(d_G(x_i, y_j, A) \leq \delta_G\big) \cdot \prod_{(i,j)\in\mathcal{S}} \Pr\big(d_D(D(x_i), D(y_j)) \leq \delta_D\big) \tag{6}$$

$$= f_D(\delta_D)^k f_G(\delta_G)^{k-m} \tag{7}$$

Equation (6) comes from point 1 in definition 1 and equation (7) from points 2 and 3. ■

In the hypothesis testing paradigm, the null hypothesis $\mathcal{H}_0$ is rejected as soon as $p(\mathcal{S}, A, \delta_G, \delta_D)$ is below the predetermined significance level (typically 5%.) However, it would mean here that, all things being equal, large groups $\mathcal{S}$ would be favoured since this yields small probabilities in equation (5). Following the a-contrario method, we do not directly deal with the probabilities but rather with the so-called *Number of False Alarms* (NFA), which permits to get rid of the arbitrary significance level. The NFA corresponds to the average number of groups consistent with A, δ_G, δ_D under hypothesis $\mathcal{H}_0$. The NFA is estimated by multiplying the probability of a false

alarm $p(S, A, \delta_G, \delta_D)$ by the number of possible events. Here, there are $\min\{N_1, N_2\} - m$ choices for $k \geq m$, $\binom{N_1}{k}$ choices for the interest points in image 1, $\binom{N_2}{k}$ choices for the interest points in image 2, $k!$ choices for the correspondences, $\binom{k}{m}$ choices for the minimal set to estimate A. Each minimal set s possibly leads to $Q = 3$ fundamental matrices (7-point algorithm [21]) or $Q = 1$ homography.

Definition 2 *We say that a set $\mathcal{S}$ of correspondences is ε-meaningful if there exists*

1. *two thresholds δ_G and δ_D such that:*

$$\forall (i,j) \in \mathcal{S},\ d_G(x_i, y_j, A) \leq \delta_G, \tag{8}$$

$$\forall (i,j) \in \mathcal{S},\ d_D(D(x_i), D(y_j)) \leq \delta_D, \tag{9}$$

2. *a transformation A evaluated from m points from $\mathcal{S}$;*

such that:

$$NFA(\mathcal{S}, A, \delta_G, \delta_D) := Q\,(\min\{N_1, N_2\} - m)\,k! \binom{N_1}{k}\binom{N_2}{k}\binom{k}{m} f_D(\delta_D)^k f_G(\delta_G)^{k-m} \leq \varepsilon \tag{10}$$

where k is the cardinality of $\mathcal{S}$, $m = 4$ and $Q = 1$ if A is a homography, $m = 7$ and $Q = 3$ if A is a fundamental matrix.

Since f_D and f_G are non-decreasing, the following proposition comes as a corollary of this definition.

Proposition 2 *A set $\mathcal{S}$ of correspondences is ε-meaningful if there exists a transformation A estimated from m correspondences among $\mathcal{S}$ such that:*

$$NFA(\mathcal{S}, A) := Q\,(\min\{N_1, N_2\} - m)\,k! \binom{N_1}{k}\binom{N_2}{k}\binom{k}{m} f_D(\delta_D)^k f_G(\delta_G)^{k-m} \leq \varepsilon \tag{11}$$

with $\delta_G = \max_{(i,j)\in\mathcal{S}} \max\{d_G(y_j, F \cdot x_i), d_G(x_i, F^T \cdot y_j)\}$,
$\delta_D = \max_{(i,j)\in\mathcal{S}}(d_D(D(x_i), D(y_j))$, and k the cardinality of $\mathcal{S}$.

The aim of the algorithm discussed in section 5 is to find the most (or a very) meaningful set of correspondences, that is to say the set of correspondences $\mathcal{S}$ with the lowest (or a very low) NFA($\mathcal{S}$). Equation (11) balances the trade-off between the probability $f_D(\delta_D)^k f_G(\delta_G)^{k-m}$ and the number of possible sets of size k among the N_1 interest points from image 1 and N_2 interest points from image 2. If δ_D and δ_G are fixed, when k grows, the first one vanishes while the latter one tends to increasing (see proposition 3 section 9)

Definition 2 was outlined in [36] (*colored rigidity*) but was neither investigated further nor implemented.

In the following sections we specify the choice for distances d_D and d_G, and associated cumulative distribution functions f_D and f_G. Note that the a-contrario framework, as it has been presented, is valid as long as f_D (resp. f_G) is a cumulative distribution function for distance d_D (resp. d_G).

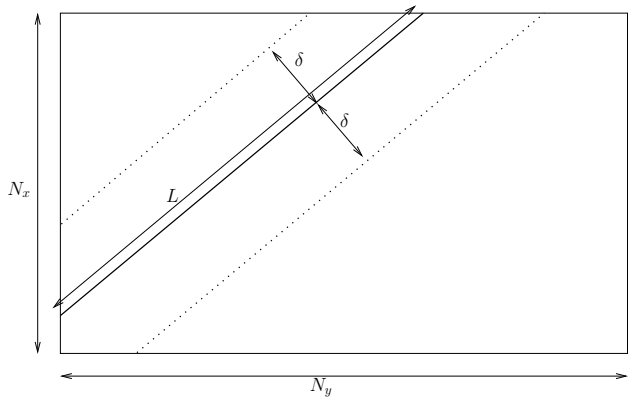

Figure 8: Moisan and Stival's model [36]: considering uniformly distributed points within an image $\mathcal{I}$ of size $N_x \times N_y$, the probability that a point falls at a distance $\leq \delta$ to a straight line with length L is approximately $2\delta L/(N_x N_y)$. If D denotes the length of the diagonal of $\mathcal{I}$ and S its surface area, then $S = N_x N_y$ and $L \leq D$. Consequently, the probability is bounded from above by $2D/S \cdot \delta$.

4.2 Modelling the geometric constraint

In the case of epipolar constraint ($A = F$), Moisan and Stival [36] propose to define $d_G(y, F \cdot x)$ as the Euclidean distance between y and the epipolar line $F \cdot x$. The function f_G is then defined as (with a slight abuse, see below):

$$f_G(d_{\text{euc}}(y, F \cdot x)) = \frac{2D}{S} d_{\text{euc}}(y, F \cdot x) \tag{12}$$

where D and S are respectively the diameter and surface area of both images (mildly assumed here to have the same size.) This choice comes from an a-contrario model which is more specific than the one from the previous section. In their paper, Moisan and Stival not only assume independence, but also that interest points are uniformly distributed in images. This leads them to estimate the probability that some random point (drawn from a uniform distribution) falls at a distance less than δ from an epipolar line. It is easy to see via a simple geometric argument that this probability is bounded above by $\frac{2D}{S}\delta$. See figure 8.

With equation (2), we derive here:

$$\begin{aligned} \Pr(d_G(x, y, F) \leq \delta_G) &= \Pr(\max\{d_G(y, F \cdot x), d_G(x, F^T \cdot y)\} \leq \delta_G) & (13) \\ &= \left(\frac{2D}{S}\delta_G\right)^2 & (14) \end{aligned}$$

by assuming that the Euclidean distances between y and $F \cdot x$ and between x and $F^T \cdot y$ are independent, following the a-contrario framework.

As we have seen earlier, $f_G(\delta_G)$ is balanced by the probability $f_D(\delta_D)$ related to the photometric constraint. We decide to parametrize the distribution f_G by actually using:

$$f_G(\delta_G) = \left(\frac{2D}{S}\delta_G\right)^{2\alpha} \tag{15}$$

We discuss the influence of this α parameter in section 6.1.1.

Let us note that $\frac{2D}{S}\delta_G$ may be larger than 1 since it is actually an upper bound of the cumulative distribution function. In order to speed-up the search, we decide to a priori eliminate groups such that this probability is larger than 5%. For typical 500×500 images, this corresponds to $\delta_G > 12.5$ pixels.

In the case of homography constraint ($A = H$), we have just to adapt the definition of $f_G(\delta_G)$ from a point-line correspondence (equation (15)) to a point-point correspondence, for example with:

$$f_G(\delta_G) = \left(\frac{\pi\delta_G^2}{S}\right)^{2\alpha}. \tag{16}$$

Indeed, $\pi\delta_G^2/S$ is the probability for a random point uniformly distributed across an image (surface area S) fall at a distance less than δ_G from a fixed point. The same kind of model was used in a different context in [6] and in [46].

4.3 Modelling the photometric constraint

We define here d_D and f_D, namely the distance between local photometric descriptors and the associated cumulative distribution function.

Since the space of descriptors is neither isotropic nor homogeneous, it is well known (see for example [31]) that it is a bad idea to measure the proximity between descriptors by a simple Euclidean distance. This observation leads to the nearest neighbour matching approach. Because of the above-mentioned heterogeneousness, any "good" metric over descriptors should not be evaluated as a norm as $||D(x) - D(y)||$. On the contrary, it should take into account the vicinity of $D(x)$ in order that the value of $d_D(D(x), D(y))$ has the same meaning in terms of "perceptual proximity" for every pair of descriptors $(D(x), D(y))$.

Rabin et al. [45] exploit this point of view by defining an a-contrario model dedicated to SIFT-like descriptor matching. Their approach has the advantage of automatically deriving distance thresholds that adapt to the descriptor of interest. Unlike the a-contrario model proposed in this report, they do not take into account the geometric constraints. Taking our inspiration from [45], and based on previous works [41, 42, 44], we define:

$$d_D(D(x), D(y)) = \phi_{D(x)}(\mathsf{dist}(D(x), D(y))) \tag{17}$$

where dist is some distance (or dissimilarity measure) over the descriptor space (Euclidean distance or a more sophisticated one as specified in the sequel), $\phi_{D(x)}$ is the cumulative distribution function of $\mathsf{dist}(D(x), D(\cdot))$ when $D(\cdot)$ spans the set of descriptors in image $\mathcal{I}_2$.

Note that, provided $\phi_{D(x)}$ is exactly known and $\mathsf{dist}(D(x), D(y))$ is actually a realization of the underlying random process, then $d_D(D(x), D(y))$ is uniformly distributed over the unit interval $[0, 1]$ (see proposition 5, section 9.) This distance therefore automatically adapts to the heterogeneousness of the descriptor space as a "contextual dissimilarity measure".

However, $\phi_{D(x)}$ is not known, and the SIFT descriptors have high dimensionality (typically 128.) In addition, SIFT descriptors are made of $N = 16$ histograms of dimension $m = 8$. Rabin et al. [45] exploit these remarks to reduce the dimensionality by using the following definition for the distance between descriptors, provided a suitable

distance $\widetilde{\mathsf{dist}}$ between histograms:

$$\mathsf{dist}(D(x), D(y)) = \sum_{i=1}^{N} \widetilde{\mathsf{dist}}(D^i(x), D^i(y)). \tag{18}$$

Let us note for every $i \in [1, N]$, $\varphi_{D^i(x)}$ the distribution function of $\widetilde{\mathsf{dist}}(D^i(x), D^i(\cdot))$. Following the discussion in [45] and under independence assumption, $\phi_{D(x)}$ is defined as:

$$\phi_{D(x)}(\delta) = \int_0^{\delta} \bigotimes_{i=1}^{N} \varphi_{D^i(x)}(t)dt \tag{19}$$

where $\otimes$ is the convolution product.

Indeed, $\mathsf{dist}(D(x), D(y))$ appears as the sum of N random variables whose probability distribution is indeed the convolution product of the N marginal distributions under independence assumption.

In practice, the distribution function $\varphi_{D^i(x)}$ is empirically estimated over the set of all $D^i(y)$ when y spans the set of the interest point extracted from image $\mathcal{I}_2$.

We have $d_D(D(x), D(y)) = \phi_{D(x)}(\mathsf{dist}(D(x), D(y)))$ from equation (17). In order to fulfill requirements of section 4.1, we still need to define the cumulative distribution function f_D. Since

$$f_D(t) = \Pr(\phi_{D(x)}(\mathsf{dist}(D(x), D(y))) \leq t) = t \tag{20}$$

if $f_{D(x)}$ is continuous and increasing (this is a classic property of cumulative distribution functions, see section 9), we simply set here $f_D(t) = t$.

We still have to define $\widetilde{\mathsf{dist}}$, that is to say the distance between sub-histograms of each descriptor (equation (18).)

Rabin et al. recently proposed in [45] to use Earth Mover's Distance (EMD), which is especially well adapted for histogram comparison. The intuitive meaning of EMD is that it corresponds to the minimum cost that an Earth Mover has to pay to reshape a histogram into another one, given the cost $c_{i,j}$ to move a unit of material from bin i to bin j. More specifically, most local photometric descriptors (and especially SIFT) are made of histograms of the gradient direction which is distributed along the *circular* interval $[0, 2\pi)$. Consequently, it is sound to use a metric that behave well with respect to these circular histograms. We use here the efficient circular EMD by Rabin et al. [45], denoted CEMD.

5 Discussion of the a-contrario model and algorithm

5.1 Discussing the NFA criterion

We can see from equation (11) that α in f_G (equations (15) or (16)) permits to balance between the geometric and photometric probabilities. These probabilities have not the same order of magnitude: the first one varies around 10^{-5} while the latter one may be around 10^{-20}. Thus α behaves as a normalization parameter, which is set once and for all to 5 after the discussion of section 6.1.1.

The sets with small NFA are the most relevant ones, as soon as the NFA is below 1. In this section we show that searching for an ε-meaningful set is realistic, given the

complexity of the problem and the probabilities at hand. This discussion completes the comments on the so-called *colored rigidity* in [36]. For the sake of simplicity, we consider here the epipolar constraint case and assume $N_1 = N_2 = N$.

Let us note

$$M(k, N) := 3(N-7)k! \binom{N}{k}^2 \binom{k}{7}. \tag{21}$$

Figure 9 shows the graph of $-\log_{10}(M(k,N))/k$ vs k for several typical values of N. From equation (11), this gives the maximal value for the logarithm of $f_D(\delta_D) f_G(\delta_G)^{1-7/k} \simeq f_D(\delta_D) f_G(\delta_G)$ so that the corresponding group $\mathcal{S}$ is 1-meaningful (in the case $N_1 = N_2 = N$.) One has indeed:

$$\text{NFA}(\mathcal{S}) \leq 1 \quad \text{iff} \quad \log_{10}\left(f_D(\delta_D) f_G(\delta_G)^{1-7/k}\right) \leq -\log_{10}(M(k,N))/k. \tag{22}$$

One can see that the NFA criterion meets two requirements that are naturally expected:

- When N is fixed, the smaller the k, the smaller the latter probability product should be. This situation can be met when dealing with a large rate of outliers and seeking meaningful groups with k small with respect to N. Since f_D and f_G are non-decreasing, this means that thresholds δ_D and δ_G must be stricter in this case.
- When k/N is fixed, the larger N, the smaller the probability product (and hence the stricter the thresholds.) This is handy when looking for correspondences in fixed size images: the denser tentative correspondences are, the more accurate they should be with respect to geometric and photometric criteria.

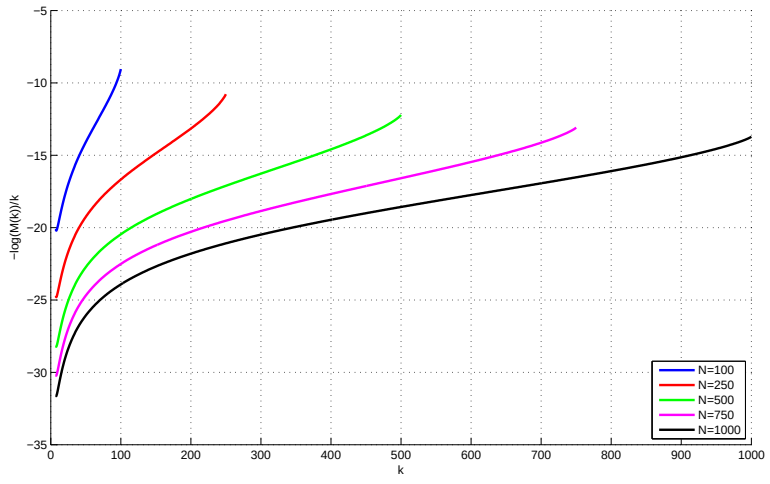

Figure 9: $-\log_{10}(M(k,N))/k$ vs k, for several values of N. This gives the order of magnitude of the logarithm of $f_D(\delta_D) \cdot f_G(\delta_G)$ so that it is still possible to find a 1-meaningful set of correspondences. (Best seen in color.)

5.2 Speeding up the search for meaningful sets

When looking for the most meaningful group of correspondences (either under fundamental matrix or under homography), a naive approach would consist in testing all possible sets of correspondences. However, if $N = 100$ features are extracted in each image, there are

$$\sum_{k=0}^{N} k! \binom{N}{k}^2 \simeq 10^{164} \tag{23}$$

such sets (as already remarked in [3].) Since testing all possible sets is out of question, a heuristic-driven search is called for. First (section 5.2.1), we use some (large) threshold on the photometric constraint to restrict the set of tentative correspondences for a given interest point from image $\mathcal{I}_1$. Since the set of possible correspondences is still huge, we use a random sampling method, i.e. a RANSAC-like heuristic (section 5.2.2.)

5.2.1 Combinatorial reduction

In order to reduce the computational burden, we do not consider all possible correspondences $y_1, \ldots, y_{N_2}$ in image $\mathcal{I}_2$ for an interest point x_i from image $\mathcal{I}_1$, but only the set of tentative correspondences $y_{j_1}, \ldots, y_{j_{N_i}}$ such that the distance between the associated descriptors is below some threshold. Of course, this matching threshold should just allow to prune the set of possibilities for algorithmic complexity purpose. Thus it should be large enough so that the true matching decision is not made at this step, while eliminating clearly non-relevant correspondences.

In order to avoid arbitrary thresholds, we use the handy a-contrario framework given by [45]. In this latter case y_j is a tentative correspondence to x_i if, with notations of equation (17):

$$N_1 N_2 d_D(D(x_i), D(y_j)) \leq \widetilde{\varepsilon}. \tag{24}$$

The value of $\widetilde{\varepsilon}$ does not depend on the experimental setup and is carefully discussed in [45]. Note that proposition 5 (section 9) argue about the auto-adaptability of this quantity to the considered descriptors. We set in this report $\widetilde{\varepsilon} = 10^{-2}$ which gives a reasonable amount of tentative correspondences. This choice is motivated in section 6.1.2. In practice, we get between 0 and 30 tentative correspondences for each x_i in a typical image.

5.2.2 Random sampling algorithm

At this stage each interest point x_i from image $\mathcal{I}_1$ is matched to a set of N_i tentative correspondences $y_{j_1}, \ldots, y_{j_{N_i}}$ in image $\mathcal{I}_2$. Now, the aim is to pick up one (or zero) $y_{j(i)}$ from this list. Since the algorithmic complexity is still too large, we use a random sampling algorithm. It is a two-step iterative algorithm, which we describe for the two cases of interest (fundamental matrix F or homography H):

A draw a sample made of seven correspondences for estimating F, or four for H

B look for the most meaningful group made from a subset of the preceding tentative correspondences, consistent with F or H.

A. Drawing a seven- or four-correspondence sample. Seven (or four) points x_i are uniformly drawn, and then are associated to a tentatively corresponding point $y_{j(i)}$. Since it gives good experimental results and reduces the computational burden, we use nearest neighbour matching (in the sense of the photometry.)

The fundamental matrix is then estimated via the non-linear "seven-point algorithm" [21, 22], and in the case of the homography, it is estimated by the Direct Linear Transform and the resulting linear system is solved by Singular Value Decomposition.

Remark that the SIFT algorithm may extract several keypoints at the same location but with different orientations or scales. In order to avoid degenerated cases, we check that the minimum sample does not contain such points. We have experimentally checked that these multiple points do not introduce noticeable bias in the computation of the NFA.

B. Seeking meaningful groups. Correspondences are added to the previous seven ones to form a group as meaningful as possible. We make use of the following heuristic, which consists in iterating the following stages.

1. For every x_i, select:

$$y_{j(i)} = \underset{y_{j_k}}{\operatorname{argmin}} \{f_D(d_D(D(x_i), D(y_{j_k}))) \cdot f_G(d_G(A, x_i, y_{j_k}))\} \tag{25}$$

   and sort correspondences $(x_i, y_{j(i)})$ in increasing order along this latter value, in order to obtain a series of nested groups made of $k = 7, 8, 9, \ldots, N_1$ correspondences.

   This step can produce correspondences between $N > 1$ x_i's to a single y_j, which should not happen. Therefore, we decide to keep among these correspondences a single one, namely $(x_i, y_{j(i)})$ such that the above-mentioned probability product is minimized.

2. Compute the NFA for each one of the above-mentioned nested groups and select the most meaningful one.

3. Sort correspondences $(x_i, y_{j(i)})$ in increasing order along $f_G(\delta_G(A, x_i, y_{j(i)}))$ to build up a new set of nested groups, compute the NFA and select the most meaningful one.

4. Return the most meaningful group found out by either step 2 or 3.

Steps 1 and 2 obviously do not ensure that the obtained group is the most meaningful one with a fixed F matrix (unlike the a-contrario RANSAC algorithm from [36] where the geometric criterion only is used.) This heuristic aims at driving the search. It is based on the fact that, provided k is fixed, the most meaningful group minimises the product $f_D(\delta_D)f_G(\delta_G)$. Note that Step 1 allows selecting correspondences among non-nearest neighbours. We have experimentally remarked that Step 3 often allows to discard false correspondences that are introduced with a low k in Step 1 because the photometric distance is very good and overwhelms the (poor) geometric distance. Using successive heuristics to test set of correspondences is sound since the lowest NFA is sought, whatever the way the group is built.

Let us remark that F (or H) is never reestimated over the whole set in this algorithm; the current F or H is estimated from a minimum sample. The aim is indeed here to get a set of correspondences, and not to estimate a registration from this set.

5.2.3 About the number of iterations

Note that in step A, we draw the minimum sample s only from the nearest neighbours. As in every RANSAC scheme, assuming that the outlier rate is r among the nearest neighbours, the number of iterations N should be

$$N \geq \frac{\log(1-p)}{\log(1-r^{|s|})} \tag{26}$$

where p is the probability to get at least one sample s made of inliers. Estimating r online has been the subject of a large literature; we will not elaborate on it in this report, and we decide to take $N = 20,000$ iterations in all experiments, so that the returned group is actually the most meaningful one with a high confidence.

In step A, we could also have avoided biasing the algorithm by the nearest-neighbour choice as in [66]. It would have been possible to pick up for each i the corresponding point $y_{j(i)}$ by drawing it randomly in the set $y_{j_1}, \ldots, y_{j_{N_i}}$ where y_{j_i} has weight $K/d_D(D(x_i), D(y_{j_i}))$ (K is a normalization parameter.) This scheme would preferably select nearest neighbours but also permits non-nearest neighbours among the minimum sample. However, the outlier rate is significantly larger for non-nearest neighbours (which can be verified in experiments, see table 3); this latest scheme thus needs much more iterations (although it is difficult to quantify) while it does not improve the results.

5.3 Algorithm

To sum up the discussion, the whole algorithm is given here. We consider that two views of the same 3D-scene are given.

1. Use SIFT algorithm to extract interest points and (zoom+rotation / contrast change) invariant descriptors from each view: $(x_i, D(x_i))_{i\in\{1,\ldots N_1\}}$ and $(y_j, D(y_j))_{j\in\{1,\ldots N_2\}}$.
2. For every $i \in \{1, \ldots N_1\}$,
   (a) build the empirical distance d_D (section 4.3, equation (17)),
   (b) define a set of tentative correspondences (section 5.2.1.)
3. Iterate ($N = 20,000$, see section 5.2.3):
   (a) choose seven (*resp. four*) points x_i and pick up the seven (*resp. four*) corresponding points $y_{j(i)}$ (heuristic A from 5.2.2),
   (b) compute the three possible fundamental matrices F from these seven correspondences and goes to (c) for each one of these matrices, (*resp. compute the homography H from these four correspondences and goes to (c)*)
   (c) select the most meaningful group (heuristic B from 5.2.2.)

In the end, return the most meaningful group ever encountered.

Let us remark that the proposed probabilistic model and algorithm are not specific to SIFT descriptors and can be easily adapted to other invariant histogram-based descriptors.

6 Experiments

We test now point matching under epipolar constraint (fundamental matrix) or homography when repeated patterns are present.

Computation time of the a-contrario algorithm (section 5.3) is about 30-40 seconds for typical 500×500 images ($\simeq 5$ seconds for 200×200 images.) Speeding-up would be possible via multi-core programming and improving estimation of the photometric distribution function (the $\varphi_{D^i(x)}$ in equation (19)), e.g. by subsampling the dataset in a Monte-Carlo estimation.

The experiments are organized in the following way. Section 6.1 investigates the role of the two parameters of the algorithm, namely α (section 4.2) and $\widetilde{\varepsilon}$ (section 5.2.1.) Section 6.2 discusses that the proposed a-contrario model permits to retrieve correspondences even when a large number of repeated patterns are present in the scene, and compare it to a Generalized RANSAC able to pick correspondences beyond the nearest neighbour.

All these experiments show that the matching thresholds (δ_D and δ_G) are automatically derived and actually vary, and that we are able to select correspondences that are not photometric nearest neighbours. Of course, these latter correspondences are never taken into account in the popular approach based on nearest neighbours described in section 1. Now, Generalized RANSAC with the correct parameters give similar results in most experiments.

In most illustrations of the a-contrario model, the correspondences are represented by a straight segment from the interest point, whose length is the apparent motion of the point in a view to the corresponding point in the other view.

6.1 Sensitivity of the a-contrario model to the parameters

We test here the influence of the parameters α (from section 4.2) and $\widetilde{\varepsilon}$ (from section 5.2.1.)

6.1.1 Influence of α

Equations (15) and (16) shows us that the smaller α, the smaller the contribution of the geometric constraint in the NFA. One could imagine that, starting from a group of correspondences with cardinality k, photometric threshold δ_D and geometric threshold δ_G, dividing α by two would yield an equally meaningful group of correspondences, that is to say with the same k, δ_D and a new δ'_G such that $2D/A \cdot \delta_G = (2D/A \cdot \delta'_G)^{1/2}$. However, this would imply $\delta'_G = 0.001$ pixel, for 500×500 images and for the standard value $\delta_G = 0.5$ pixel. Such a small value for δ'_G is simply not reachable. As a matter of fact, a smaller α yields a set of correspondences less constrained by the geometry, and more by the photometric resemblance of the descriptors. On the contrary, a larger α should yield groups of correspondences that meet the geometric constraint well, however the photometric constraint may be too loosened in this case. In this latter case the geometry may not correspond to the reality. Note also that the geometric and photometric constraints are balanced by the number of corresponding points as explained in section 5.1: a small group can win over a large one if the photometric (small α) or geometric (large α) constraint is tightly enforced. From this heuristic discussion, we see that a trade-off must be met. This situation is illustrated by figures 10 (epipolar constraint) and 11 (homography constraint) in a situation where the camera motion is a rotation around the center of the cube.

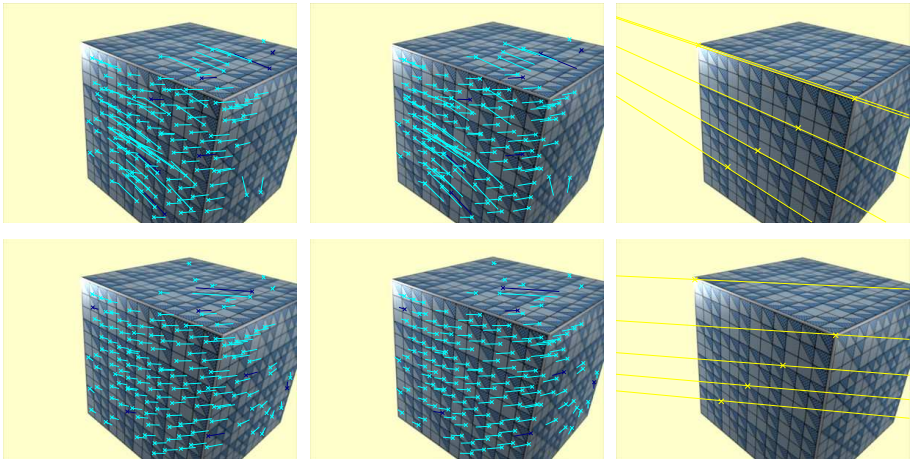

Figure 10: Influence of the α parameter, epipolar constraint. Top: $\alpha = 3$. There are 138 correspondences (between left and middle images; 31 of them are not nearest neighbours.) In this case, the geometric constraint is not enforced in a strong enough way. The most meaningful group corresponds to the situation where a lot of correspondences are found along the dominant direction of the lattice of repeated patterns (vanishing lines as shown in right image, see section 2.) Bottom: $\alpha = 5$. There are 148 correspondences (47 are not nearest neighbours), and the retrieved epipolar pencil is now consistent with the camera motion. Note a false correspondence on the top of the cube that is consistent with respect to photometry (matches between repeated patterns) and geometry (matches along an epipolar line) as in figure 6.

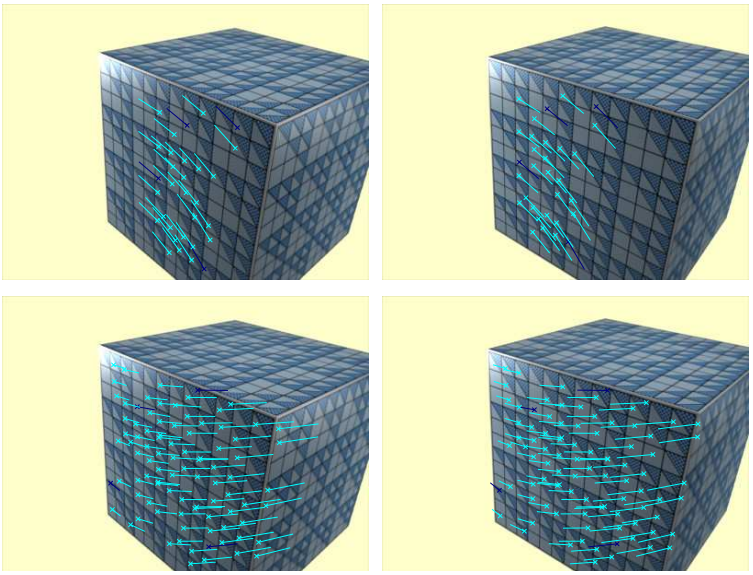

Figure 11: Influence of the α parameter, homographic constraint. Top: $\alpha = 1$. 31 correspondences are retrieved between left and right images (4 are not nearest neighbours.) We can see that they are correct (i.e. between perceptually similar features), but shifted. Let us recall that SIFT has not a perfect invariance to scale change. Here, the most similar descriptors between the two images are the ones with the same absolute scale. Such correspondences are possible here because of the large amount of repeated patterns. A standard pair of images under different viewpoints is less likely to show repeated patterns that have the same absolute scale. Here, the photometric constraint has a too strong influence on the NFA, and is not balanced by the geometric constraint. Bottom: $\alpha = 3$. 91 correspondences are retrieved (56 are not nearest neighbours): the influence of the geometry is now strengthen and the algorithm provides a larger group, which is now correct.

α	Nb. of points	geometry δ_G in pixel	photometry $\log(\delta_D)$
1	149.2 (1.8)	.71 (.12)	-40.4 (.5)
2	151.6 (3.8)	.70 (.14)	-39.5 (1.1)
3	162.1 (8.1)	.58 (.12)	-33.6 (3.2)
4	164.2 (5.2)	.49 (.13)	-30.5 (1.8)
5	172.3 (10)	.46 (.11)	-26 (4.2)
6	183.5 (6.2)	.49 (.09)	-21.2 (1.9)
7	182.9 (5.3)	.47 (.10)	-20.7 (.48)
8	183.7 (4.9)	.46 (.08)	-20.6 (.28)
9	184.5 (4.5)	.46 (.08)	-20.5 (.28)
10	185 (4.8)	.46 (.08)	-20.5 (.17)

Table 1: Influence of α on the retrieved sets of correspondences, with the same images as in figure 10. From left to right, α, *Nb. of points* is the cardinality of the retrieved set, δ_G, and $\log(\delta_D)$. Standard deviation are indicated between brackets (average over 100 runs.) Up to $\alpha = 6$, the larger α, the smaller δ_G. When α is between 6 and 10, the accuracy does not decrease anymore because there are only a few matches with a distance to the epipolar line less than 0.4. Then the balance in the NFA furnishes groups with still lower NFA, but with the same k, δ_G, δ_D. The same experiment with the homography constraint gives similar results (not shown here.)

Table 1 provides some statistics when α grows (and thus the geometric constraint has more importance.) One can see that when α grows from 1 to 5 the number of retrieved correspondences grows, the geometric accuracy is better (distance to the epipolar line), while the photometric constraint becomes looser. Note that α larger than 6 does not yield significant changes in the most meaningful set of correspondences. From this table and other experiments, we decide to set in all following experiments $\alpha = 5$ once and for all.

In the previous experiments, $\widetilde{\varepsilon}$ was set to 10^{-2}.

6.1.2 Influence of $\widetilde{\varepsilon}$

We test here the influence of the $\widetilde{\varepsilon}$ parameter (section 5.2.1). Since the SIFT descriptors are invariant to zoom+rotation only, a bias will appear in the probabilities as soon as the viewpoint change is too strong, as in every SIFT-based method. We therefore consider a small motion between two views, so that the non-invariance of SIFT to viewpoint interferes as least as possible. The test is led here with the fundamental matrix model. As an illustration, figure 12 shows some results with a varying $\widetilde{\varepsilon}$.

Table 2 gathers some statistics about this experiment. Concerning the influence of $\widetilde{\varepsilon}$ (namely the parameter of the "combinatorial reduction step"), one can see that reducing it subsequently reduces the number of tentative correspondences among which the most meaningful set is sought, while having almost no impact on the cardinality of this set. In other words, decreasing the value of $\widetilde{\varepsilon}$ speeds up the search while discarding mainly false correspondences. Note that at least 20-25% of the matches are not nearest neighbours. Once again, it would have been impossible to retrieve them with the classic SIFT nearest neighbour matching.

$\widetilde{\varepsilon}$	# tentative corr.	# most meaning. group	% of rank 1 corr.
1	2027	219.5 (6.4)	76.1
10^{-2}	1409	220.6 (4.9)	76.2
10^{-4}	999	218.3 (3.9)	76.6
10^{-6}	663	202.7 (3.0)	78.4
10^{-8}	407	172.8 (2.8)	85.5
10^{-10}	274	146.2 (2.2)	90.3

Table 2: *Synthetic* images. Influence of $\widetilde{\varepsilon}$. From left-most column to right-most one: the four distances between descriptors that are tested, the six values of $\widetilde{\varepsilon}$ in the range 1 - 10^{-10}, the number of tentative correspondences retrieved after the combinatorial reduction step (section 5.2.1), the cardinality of the most meaningful group (average over 100 runs, standard deviation in brackets), and the average proportion of nearest neighbours among this group of correspondences.

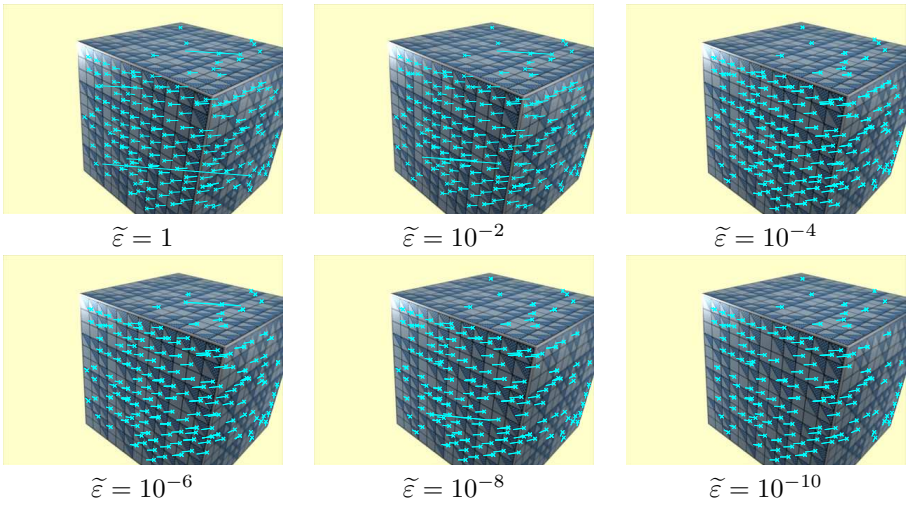

Figure 12: *Synthetic* images. In this experiment, we search for the most meaningful group consistent with a fundamental matrix between two for six values of $\widetilde{\varepsilon}$. We only show here the first view, the blue segment corresponds to the apparent motion of an interest point (localized by a cross) between the two views. One can see that some false correspondences are still retrieved. A careful examination shows that they actually lie along the associated epipolar line, and simply cannot be detected in a two-view matching. In all experiments (whatever the distance and $\widetilde{\varepsilon}$ as in table 2), the average distance to the epipolar line is about 0.2-0.3 pixel. 343 SIFT keypoints were extracted from image 1, and 321 from image 2.

From these results and other experiments on realistic images, we decide to set in the sequel $\widetilde{\varepsilon} = 10^{-2}$, which leads to a good trade-off between complexity reduction and size of the most meaningful group.

6.2 Point correspondences and perceptual aliasing

We consider here images with repeated patterns, in the light of section 2. In particular, we discuss the a-contrario algorithm and a Generalized RANSAC detailed in the following section.

6.2.1 A Generalized RANSAC algorithm

This section presents a Generalized RANSAC algorithm, inspired by Zhang and Kosecka [66].

Standard RANSAC takes as input the tentative one-to-one point correspondences given by a preliminary step based on descriptor similarity. If this step is relaxed so that each interest point x in the first image has K tentative correspondences $y_1 \dots y_K$ in the second, then the following algorithm is able to extract a set of correspondences consistent with a homography or the epipolar constraint. It consists in iterating the following operations: (here $y(x)$ is one of the $y_1 \dots y_K$ associated to y_i)

1. Draw a minimum sample $(x, y(x))$ to estimate H (or F).
2. Knowing H (or F) associate each x to a single $y(x)$ among the tentative correspondences.
3. Count the number of correspondences $(x, y(x))$ such that $d(x, y(x), H)$ (or $d(x, y(x), F)$, as in equations (2) or (3)) is less than a pre-determined threshold δ_{GR}.

In the end, the largest consensus set is returned.

We decide to associate in step 1 the x to their nearest neighbour $y(x)$ among the tentative correspondences y_i (in the sense of the descriptor proximity) and in step 2 to define $y(x)$ among the y_i as minimizing the distance d.

In [66] it is suggested to randomly sample the x and $y(x)$, and different strategies are given. We found that taking the nearest neighbour in step 2 significantly reduces the number of needed iterations. It is a sound hypothesis since in tractable cases, most correct correspondences are still among nearest neighbours, as we will see in the experiments.

Note that our aim here is not to compete with state-of-the-art RANSAC algorithms in terms of computational load, we just design a proof-of-concept algorithm to compare with the a-contrario models. Building a competitive Generalized RANSAC would necessitate strategies as like as in PROSAC [9] for example.

Let us discuss an example of the properties of Generalized RANSAC. As we have seen on figure 3, in some examples the nearest neighbour + threshold condition inevitably yields erroneous sets of correspondences. In figure 13, we show that when relaxing the tentative correspondences to the nearest neighbours (without any condition on the distance ratio), then it is possible to find a correct set of correspondences. Of course, a large number of iterations is needed since the outlier rate significantly increases. Here, about 31,000 iterations are needed to get a correct set (average over 10 runs). However, when searching among $K = 5$ first nearest neighbours, only about 13,000 iterations are needed. For example, the group shown in figure 13 (bottom) is made of 43 correspondences of rank 1, 25 of rank 2, 24 of rank 3, 14 of rank 4, 7 of rank 5. The subset of rank 1 correspondences is beaten by the erroneous set shown in figure 13 (top) which is made of 49 matches. Thus less iterations are needed when searching beyond the nearest neighbour. When searching among $K = 10$ first nearest

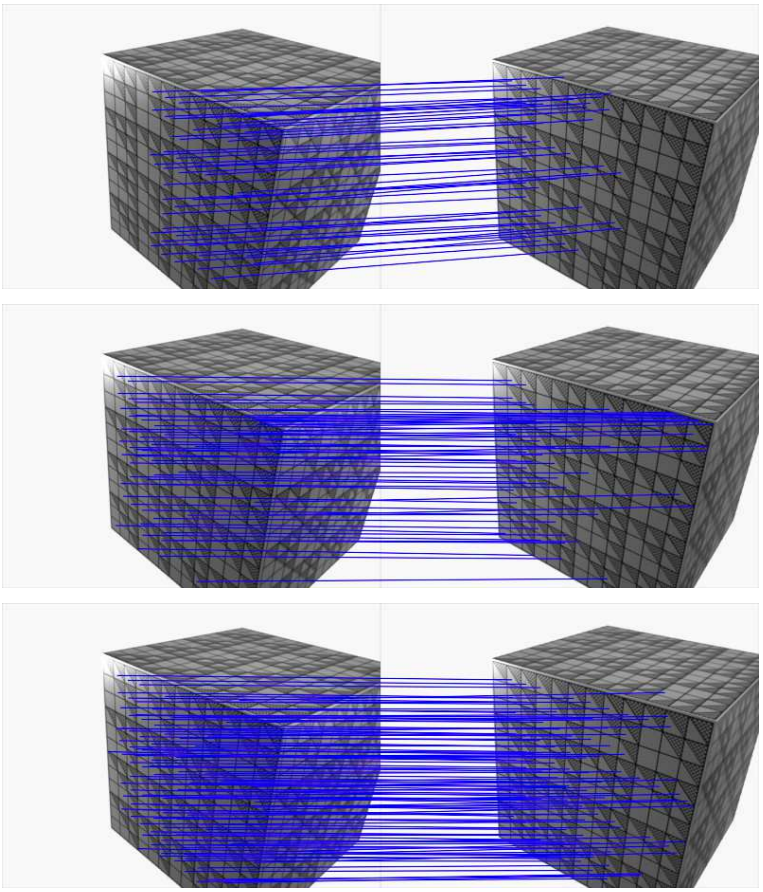

Figure 13: *Generalized* RANSAC. About 800 SIFT keypoints are extracted from both images. The distance threshold δ_{GR} is set to 1. Top and middle: $K = 1$ and homography constraint. False large consensus set can be retrieved, here 49 matches (top.) However, ensuring a large enough number of iterations in Generalized RANSAC makes it possible to find correct sets, here 60 matches (middle.) Bottom: $K = 5$. 113 correspondences are retrieved here, only 43 are ranked first; less iterations are needed.

neighbours (not shown), about $6,000$ iterations are needed to get a correct set among which less than 50% are nearest neighbours. In this case there are only 4 correspondences between 7th and 10th nearest neighbours. In spite that each pattern is repeated more than 10 times, this means that the invariant descriptors are actually not repeated more than 5-6 times. This is due to their limited invariance and also to the fact that their extraction scale is such that they convey information which makes them distinguishable.

Figure 14 shows the correspondences with the a-contrario model. The advantage is that the distance thresholds and number of tentative correspondences kept are automatically derived.

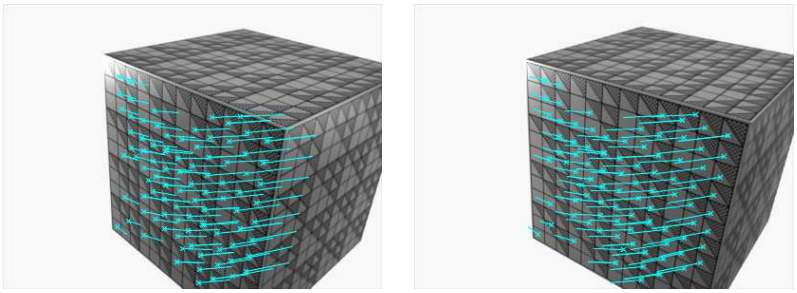

Figure 14: *A-contrario* RANSAC. Apparent motion of the interest points between the two views. All correspondences are correct. Here δ_G was determined as 2.6 pixels. Among the 111 matches, 67 are nearest neighbours, 20 second nearest neighbours, 10 third nearest neighbours, and the rest behind fourth and eleventh.

6.2.2 Point correspondences and the curse of perceptual aliasing

The aim of this section is to show that the proposed method allows us to obtain more correspondences than a standard robust matching criterion when confronted with repeated patterns. Indeed, a significant part of the matched interest points does not come from the nearest neighbour descriptor, but from correspondences with a higher rank. We compare the proposed algorithm with a usual method using steps 2 and 3 presented in section 1, that is: NN-T matching (Euclidean distance, threshold on the ratio set to 0.6 as in Lowe's code), followed by a robust selection with the a-contrario RANSAC from [36]. The use of this two-step scheme is called NN-T+O, our method AC for a-contrario. We also compare with Generalized RANSAC of section 6.2.1.

Repeated patterns and homography. We first use the homography as the geometric constraint. When confronted to repeated patterns, the number of matches selected with NN-T is small, as shown in the right image of figure 15: repeated features are generally discarded at this early stage, and of course cannot be retrieved by the subsequent RANSAC. Our method, as shown in the left image of figure 15, retrieves much more correspondences. The numerous extra correspondences coincide with matches which are not nearest neighbour for the descriptor distance. From table 3, we can see that while 128 features are matched, 42 are ranked first, and 86 have higher ranks.

Table 3 also shows that if the tentative correspondences were defined as nearest neighbours (without any limitation over the distance ratio), then it should be possible to get a consistent set of correspondences with classic RANSAC. There are indeed a good amount of correspondences with rank 1. However, the outlier rate is in this case very large, and would call for a large number of iterations. In addition, the RANSAC process could be trapped by shifted patterns as illustrated in section 2.

Let us also remind that the a-contrario algorithm automatically adapts the thresholds to the complexity of the scene: in the *Monkey* image pair ($1,500 \times 1,200$ images) δ_G was derived as 3.1 pixels, in *Loria* (800×800 images) as 2.1 pixels and in *Flat Iron* (500×400 images) as 7.8 pixels. Using these values as distance threshold δ_{GR} in Generalized RANSAC naturally yields similar results, but this parameter is hard to infer.

	Number of correspondences		
Rank	*Monkey*	*Loria*	*Flat Iron*
1	42	98	8
2	23	32	3
3	17	29	1
4	11	18	1
5	8	20	0
6	8	19	0
7	4	11	0
8	8	5	0
9	3	13	0
10	2	8	0
11	0	15	0
12	2	7	0
>12	0	37	0
Total	128	312	13

Table 3: Number of occurrences of the n-th nearest neighbours selected by the AC method (homography case.) *Monkey* corresponds to figure 15 (412 vs 445 extracted keypoints), *Loria* to figure 1 (2,562 vs 2,686), *Flat Iron* to figure 16 (756 vs 598.) Remark the strong perceptual aliasing in these pairs. As noted in figures 1 and 16 the NN-T+O method does not succeed at all in *Loria* and *Flatiron* experiments. Ranks larger than 2 are all the more frequent as the scene contains repeated patterns, and cannot be retrieved by the NN-T+O method, or any method limited to nearest neighbour matching.

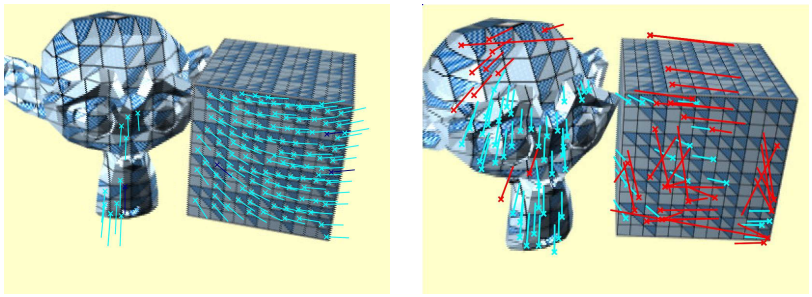

Figure 15: *Monkey*, homographic constraint. Two images with repeated patterns. On the left, the proposed AC model, most of the patterns lying on the dominant plane are detected (segments represent the apparent motion between the two views.) On the right, the second image but with correspondences from NN-T (both colors) and NN-T+O (inliers in blue, outliers in red.) Many more correspondences are retrieved with the AC algorithm. Generalized RANSAC gives similar results.

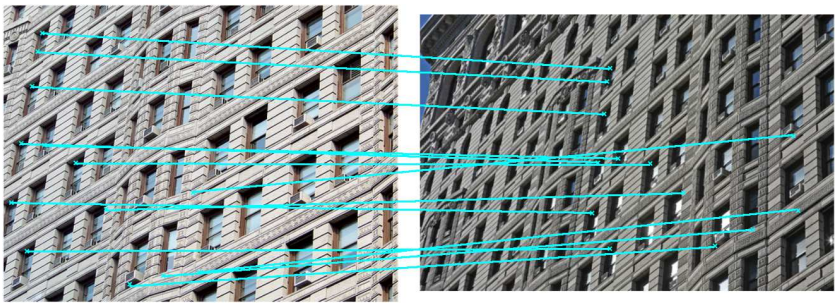

Figure 16: *Flat Iron*, homographic constraint. 13 matches can be found with AC method, all of them are correct. Here NN-T+O does not give any consistent group. Note the strong perceptual aliasing and the quite strong illumination and viewpoint changes.

Repeated patterns and epipolar constraint. In this section, we test the behaviour of the AC algorithm under epipolar constraint. Since the epipolar constraint acts more "gently" than the homographic one (it is a point/line constraint), some false correspondences are simply unavoidable, as formalized by the double nail illusion of section 2. Figure 17 shows a situation with a repetitive texture were almost no false correspondence. Small baseline matching gives good results thanks to the limited invariance of the descriptors. The nearest neighbour is more likely to be the correct one, unlike the larger baseline case shown in figure 4. However, as soon as the baseline grows, Generalized RANSAC as well as a-contrario matching gives false correspondences due to the confused distance between descriptors yielding interest points satisfying epipolar constraint "by chance". Large baseline matching is addressed in section 7.

Figure 18 shows the result of the AC method. It yields more correspondences than the NN-T+O method (not shown here, 92 vs 29), distributed in a denser fashion across the views. Nevertheless, a careful examination shows that many correspondences are not correct, in spite that the keypoints actually lie near their epipolar lines, because of the above-mentioned reason. This still permits an accurate estimation of the epipolar pencil. We select corresponding points (x, y) by hand in both views (especially in areas where almost no correspondence is retrieved with NN-T+O), and draw the associated epipolar lines. The line Fx (resp. $F^T y$) should meet the point y (resp. x.) Here F is re-estimated over the consensus set retrieved by NN-T+O or AC. The reestimation consists in minimizing the Sampson metric [22, 67].

In addition to the problem with false correspondences due to repeated patterns along epipolar lines, some point matching problems are very difficult to solve because the repeated patterns yield degenerate motion estimation, as explained in section 2. For example, in figure 19 one can see that the most meaningful group consists in wrong correspondences among points that match in a dominant plane along lines parallel to an edge of the cube. Let us also note that the stricter point-point constraint from the homography case (compare figure 19 to figure 15) enables to retrieve a consistent set, unlike in the epipolar constraint case.

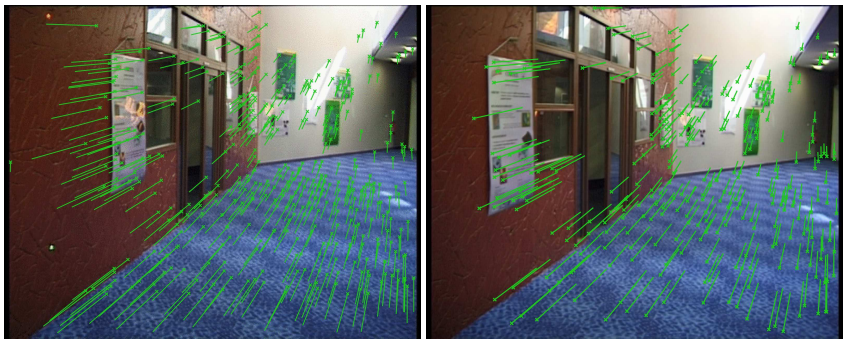

Figure 17: *Corridor*, epipolar constraint. AC method (left) retrieves 423 correspondences. 405 correspondences have rank 1, 13 rank 2, 2 rank 3, 1 rank 4, 1 rank 6, 1 rank 10. NN-T+O method (right) retrieves 295 out of 316 NN-T matches. The additional correspondences are on the carpet and on the wall. 1,269 keypoints were extracted from image 1, 1,360 from image 2. The Generalized RANSAC gives similar results.

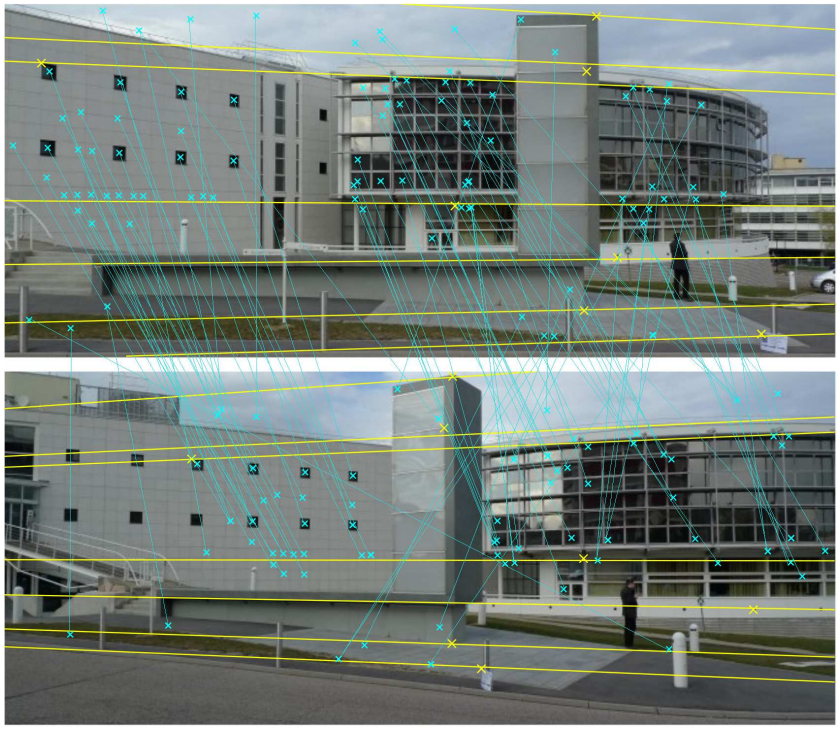

Figure 18: *Loria building*. AC method (epipolar constraint) between the top and bottom views. 92 correspondences can be seen. In particular, the repeated left-hand windows are all retrieved and not shifted. However, many "unavoidable" false correspondences are also retrieved, as the ones between the structures of the left-hand façade which are indeed shifted along the epipolar lines (compare to the position of the windows; the same phenomenon appears on the right-hand façade.) Nevertheless, one can see from the hand-picked correspondences (in yellow) that the associated (reestimated) epipolar lines are much closer. The distance is less than 5 pixels, except from one point on the parallelipedic structure on the foreground which is still at 15 pixels. It is logical to find a poorer accuracy on this structure since a very small amount of points is extracted from it, and since its apparent motion is quite different from the motion of the background.

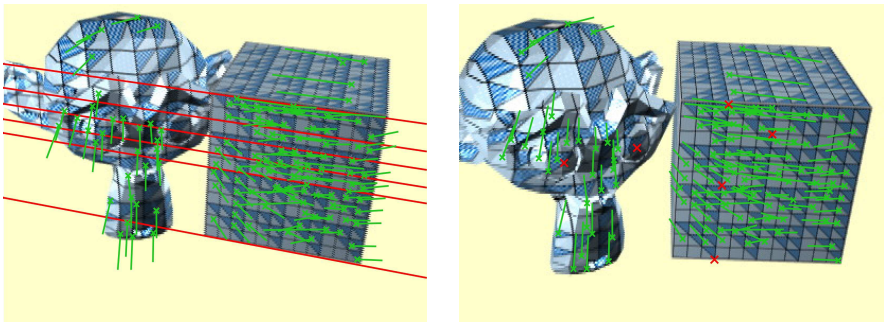

Figure 19: *Monkey*, epipolar constraint. Failure case study. The recovered geometry corresponds to the vanishing lines. In that case, the point / line constraint does not solve the ambiguity intrinsic to perceptual aliasing, and thus gives false correspondences. A few hand inserted points in red show the epipolar lines pencil, which corresponds to the pattern alignment along the vanishing lines, and not to the true motion. See the discussion of section 2.

7 Improving ASIFT with respect to repeated patterns

As mentioned earlier, small baseline matching is quite easy even in the presence of repeated patterns since the nearest neighbour matching already gives good results. Large viewpoint changes are much more challenging. Several assessment papers [1, 35, 38] have extensively compared the most standard interest points / region descriptors, from SIFT [31] to Harris / Hessian Affine [34] to MSER [33]. Now, the authors of [38] conclude: *"no detector/descriptor combination performs well with viewpoint changes of more than* 25-30°." This is confirmed by the authors of a very recent survey [1]. All of these methods are thus prone to fail at a certain point. A more successful approach has been recently proposed by several authors (*e.g.* [25, 39]), in which viewpoint simulation is used to attain affine invariance. These papers demonstrate that this dramatically improves the number of matches between two views compared to MSER or Harris/Hessian Affine, especially with a strong viewpoint change.

In Morel and Yu's ASIFT [39], affine invariance of image descriptors is attained by remarking from Singular Value Decomposition that any affine mapping A (with positive determinant) can be decomposed as

$$A = \lambda R_\psi \begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix} R_\phi \tag{27}$$

where $\lambda > 0$, R_ψ and R_ϕ are rotation matrices, $\phi \in [0, 180^{\text{o}})$, $t \geqslant 1$.

Since SIFT is scale and rotation invariant, a collection of affine invariant (ASIFT) descriptors of an image I is obtained by extracting SIFT features from the simulated images $I_{t,\phi}$ with

$$I_{t,\phi} = \begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix} R_\phi(I). \tag{28}$$

Indeed, the location of the SIFT keypoints is virtually covariant with any scale and rotation change λR_ψ applied to $I_{t,\phi}$, and the associated descriptor does not change. From [39], it is sufficient to discretize t and ϕ as: $t \in \{1, \sqrt{2}, 2, 2\sqrt{2}, 4\}$ and $\phi = \{0, b/t, \ldots, kb/t\}$ with $b = 72^{\text{o}}$ and $k = \lfloor t/b \cdot 180^{\text{o}} \rfloor$.

The next step is to match ASIFT features between two images I and I'. A two-scale approach is proposed in [39]. First, the $I_{t,\phi}$ and $I'_{t',\phi'}$ are generated from downsampled images (factor 3), then SIFT features extracted from each pair $(I_{t,\phi}, I'_{t',\phi'})$ are matched via the standard algorithm from [31], namely that nearest neighbours are selected provided the ratio of the distance between the nearest and the second nearest is below some threshold. The deformations corresponding to the M pairs (M typically set to 5) that yield the largest number of matches are used on the full-resolution I and I', giving new SIFT features that are matched by the same above-mentioned criterion. The interest points obtained from the correspondences are then placed in I and I', provided already-placed correspondences are at a distance larger than $\sqrt{3}$.

This strategy is used to limit the computational burden and also prevents redundancy between SIFT features from different deformations. A subsequent step consists in eliminating spurious correspondences with RANSAC by imposing epipolar constraints.

Since ASIFT is based on nearest neighbour matching, it is subject to the curse of double nail illusion, as well as any nearest neighbour based algorithm. See figure 20. However, the idea behind ASIFT is that the world is locally planar, and since affine transformations are first order approximations of homographies, affine simulation is expected to ease the matching of features lying on the same 3D plane. Instead of

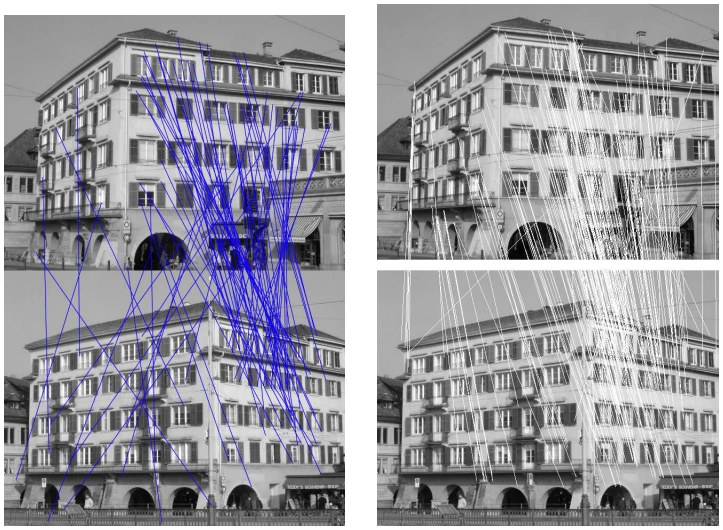

Figure 20: *False correspondences due to repeated patterns and epipolar constraint*. Left: standard nearest neighbour matching. Right: ASIFT (software from [65]). Nearly all correspondences on the left façade are not correct, as well as a lot on the right. The a-contrario method has a similar behaviour (not shown.) The image pair is all the more difficult as the aspect of the left façade changes a lot between the two views. Image pair from the Zürich Building Database.

using nearest neighbour matching without any geometric information, it seems natural to enforce homographic constraint. Actually, enforcing the homography constraint helps solving the double nail illusion, as explained in figure 21. Thus, we propose as a proof-of-concept to replace the nearest neighbour matching when comparing simulated images by a Generalized RANSAC (section 6.2.1) enforcing a homography constraint. Note that in [43] the use of the presented a-contrario matching was investigated, and the authors of [4] have independently suggested to use graph matching. Nevertheless, graph matching does not explicitly enforce a geometric constraint.

Two images I and I' being given, the proposed Improved-ASIFT algorithm would consist in the following steps:

1. Generate the $I_{t,\phi}$ and $I'_{t',\phi'}$ (viewpoint simulation.)
2. Extract SIFT features from all simulated images.
3. Match SIFT features: for each pair from step 1, extract a group of correspondences with Generalized RANSAC enforcing the homography constraint.
4. Keep only the matched SIFT keypoints from the $I_{t,\phi}$'s and $I'_{t',\phi'}$'s, among the N largest sets of correspondences.
5. Discard possible false correspondences: epipolar RANSAC.

The output is a set of corresponding points of interest.

The algorithm here works at a fixed resolution, and thus do not take advantage of the multi-resolution scheme implemented in original ASIFT. The main difference is to

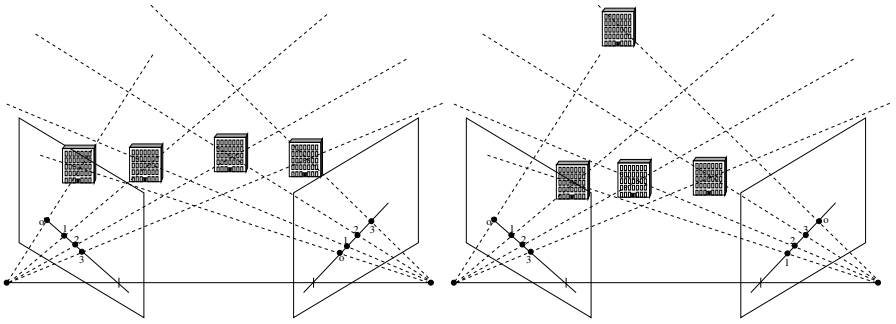

Figure 21: *Adding homography constraint helps recovering from the double nail illusion*. Compared to figure 5, adding a homography constraint between the interest points in the two views selects a single possibility. It comes down to imposing the relative positioning of the interest points from a view to the other one. Now, if one of the correspondences is seen as an outlier (marked as 'o' in the schemes), it is still possible to get (here) two shifted sets of correspondences which are equally plausible. In both cases, the three remaining correspondences are still consistent with the ordering imposed by the homography. However, this situation is possible if the subset of correspondences is only made of repeated patterns. In most realistic cases, non-repeated features incorporated in the group consistent with the homography help disambiguating the shifting.

replace in step 3 nearest neighbour matching by homography Generalized RANSAC. However, if there is a too dominant plane in the scene, then it is difficult to extract a consistent set of correspondences on a small piece of plane if the threshold δ_{GR} is set to a small value. Therefore, δ_{GR} is set here to, typically, 4-5 pixels. If some false correspondences are introduced here, they are likely to be discarded by step 5.

Figure 22 shows the result of improved ASIFT on a building image from the Zürich Building Database. Compared to image 20, we can see that the correspondences are this time correct.

In the very challenging image pair of figure 23 we are also able to get a lot of correspondences (all correct).

Figures 24 and 25 show the correspondences obtained with two views from Oxford's house sequence. The baseline is quite large, and nearest neighbour approaches do not give any correspondence. With the improved ASIFT we are able to recover from the double nail illusion, while standard ASIFT yield many false correspondences between the chimneys and the white bricks over the windows. For the improved ASIFT we have hidden the textured background and ground in order to get rid of the problem with the dominant planes mentioned above.

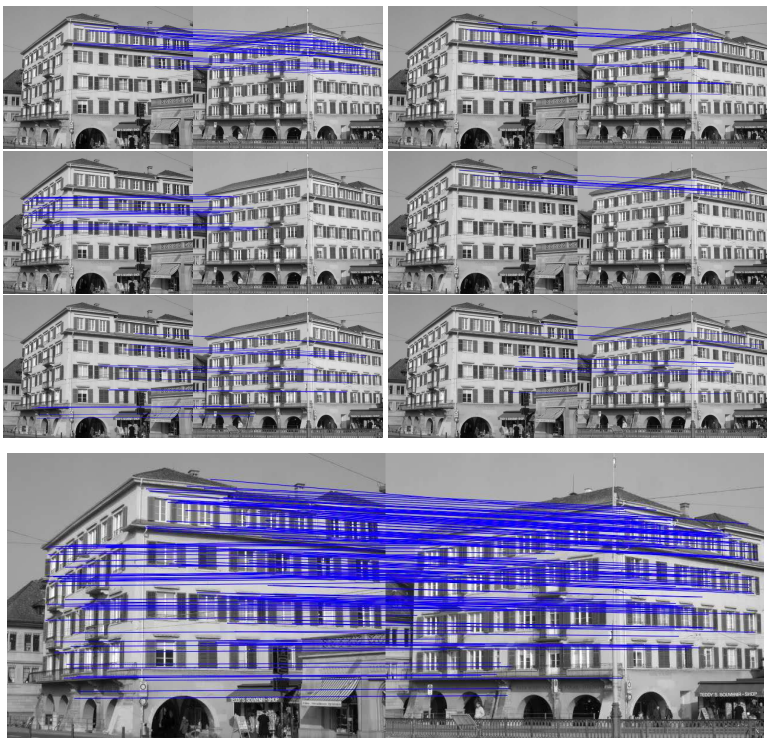

Figure 22: *Improved* ASIFT*: Zürich Building*. Improved ASIFT yields 151 correspondences, only 89 of them are nearest neighbours. The six largest set of correspondences retrieved by the homography Generalized RANSAC (coming from different pairs of simulated views) are shown on the top of the figure. They actually correspond to features lying on the same piece of plane.

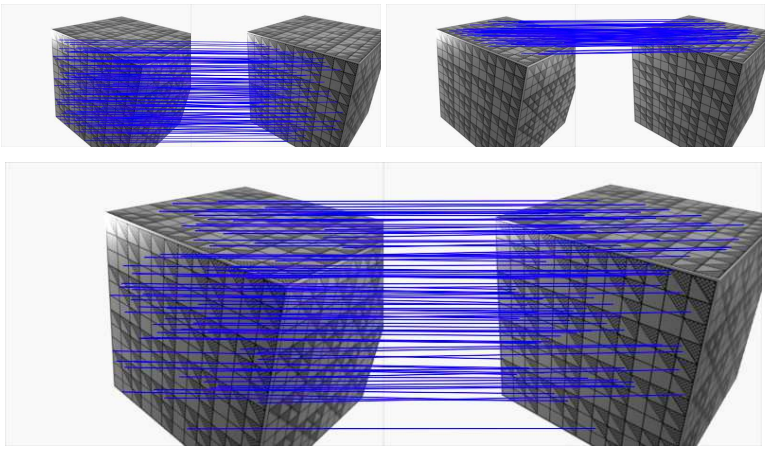

Figure 23: *Improved* ASIFT*: cube*. In this synthetic cube experiments, the two largest set of correspondences consistent with a homography are shown on the top. Each of them correspond to a side of the cube. On the bottom, 100 correspondences are shown (among a total of 324, only 142 are nearest neighbours.)

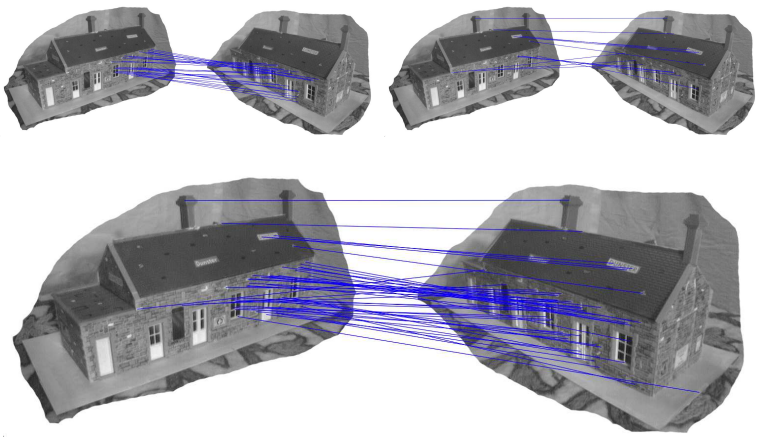

Figure 24: *Improved* ASIFT*: house*. Top: the two largest set of correspondences consistent with a homography. Bottom: the 49 correspondences (only 33 are nearest neighbours.) The textured background and ground are hidden in this proof-of-concept experiment.

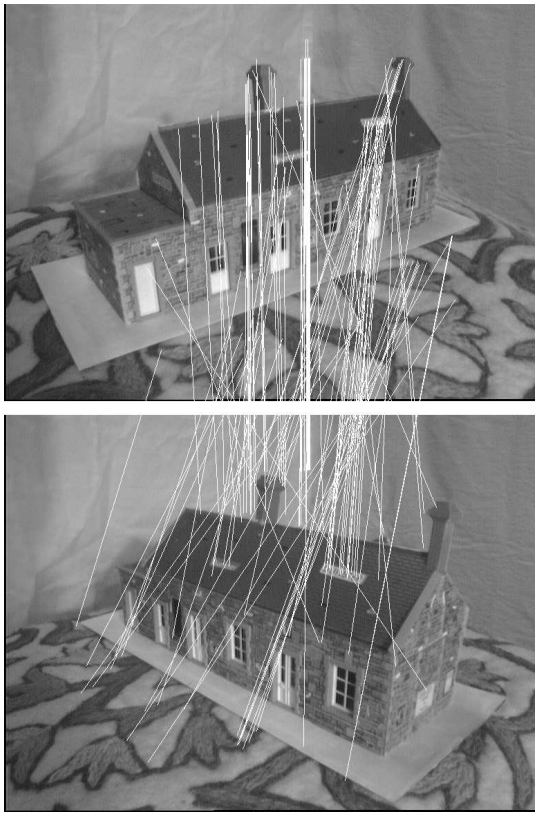

Figure 25: ASIFT*: house*. ASIFT [65] gives here 104 correspondences; many of them are not correct because of the double nail illusion, e.g. between interest points on the two chimneys, or on the light-coloured bricks over the windows.

8 Conclusion

In a discussion on the influence of repeated patterns on the two-view correspondence problem, we have emphasized the double nail illusion. We have designed an a-contrario model and a Generalized RANSAC for point matching which are both able to pick up correspondences beyond the nearest neighbours, permitting to get correct sets of correspondences in challenging situations. The a-contrario approach automatically balances in the NFA both photometric and geometric matching thresholds without any user intervention. It gives results similar to the Generalized RANSAC, but does not necessitate tuning various parameters.

The limited invariance of local image descriptors yield incorrect correspondences between repeated patterns when the viewpoint change is too strong. In this case, we have also explained how to improve ASIFT, yielding an algorithm robust to wide viewpoint changes, even with repetitive textures. We obtain more correspondences than in the standard SIFT matching which often simply fails. This can be helpful in e.g. object recognition (more correspondences means an increased confidence) or in structure and motion applications (for a denser 3D map.) However, the localization accuracy of these extra correspondences has still to be investigated, especially for ASIFT since they come from simulated views.

9 Appendix: some proofs

Proposition 3 *Let N be an integer and let us note for every integer $7 \leq k \leq N$:*

$$M(k,N) = 3(N-7)k!\binom{N}{k}^2\binom{k}{7}. \tag{29}$$

Then the series $(M(k,N))_k$ is increasing between $k=7$ to $k=k_0$, and decreasing for $k \geq k_0$, where

$$k_0 = \left(2N+1-\sqrt{4N-23}\right)/2 \quad \sim_{N\to+\infty} N-\sqrt{N}.$$

Proof: Computing the ratio between two consecutive terms,

$$\frac{M(k+1,N)}{M(k,N)} = \frac{(N-k)^2}{k-6}. \tag{30}$$

This ratio is larger than 1 if and only if $P(k) = k^2-(2N+1)k+N^2+6$ is positive, which is true provided $k < k_0$ where $k_0 = \left(2N+1-\sqrt{4N-23}\right)/2$ is the smallest root of P. The second root is indeed larger than N, and $k \leq N$.

This proposition justifies the remark just after proposition 2 (in the case $N_1 = N_2 = N$.) ■

In the text, we also make use of the following classic proposition.

Proposition 4 *If X is a real random variable and F is its cumulative distribution function, then for any non-negative real number x:*

$$\Pr(F(X) \leq x) \geq x \tag{31}$$

and the equality holds if F is continuous and increasing.

Proof: Let us denote $F^{-1}(x) = \arg\inf_{t\in\mathbb{R}}\{F(t) \geq x\}$, which exists because F is non-decreasing. Then one can see that $F(t) \leq x$ if and only if $t \leq F^{-1}(x)$.

Successively:

$$\Pr(F(X) \leq x) = \Pr(X \leq F^{-1}(x)) = F(F^{-1}(x)) \geq x \tag{32}$$

and the equality holds if F is continuous and increasing since in this case F^{-1} is the inverse of F. ■

Proof of the remark about d_D in section 4.3.

Proposition 5 *Suppose that the space of* SIFT *descriptors is endowed with a (arbitrary) metric* dist*. Let us consider a descriptor D and a random descriptor D' such that* dist(D,D') *is a random variable with cumulative distribution function f_D (supposed to be continuous and increasing.) Let us define the new metric $d(D,D') := f_D(\mathsf{dist}(D,D'))$. Then $d(D,D')$ is uniformly distributed on the unit interval $[0,1]$.*

Proof: One has indeed for every $t \in [0,1]$:

$$\Pr(d(D,D') \leqslant t) = \Pr(f_D(\mathsf{dist}(D,D') \leqslant t) = t \tag{33}$$

from proposition 4. ■

References

[1] H. Aanæs, A. Dahl, and K.S. Pedersen. Interesting interest points. *International Journal of Computer Vision*, 2011. To appear.

[2] W. Aguilar, Y. Frauel, F. Escolano, M. E. Martinez-Perez, A. Espinosa-Romero, and M. A. Lozano. A robust Graph Transformation Matching for non-rigid registration. *Image and Vision Computing*, 27(7):897–910, 2009.

[3] M. Antone and S. Teller. Scalable extrinsic calibration of omni-directional image networks. *International Journal Computer Vision*, 49(2-3):143–174, 2002.

[4] C. Le Brese, J.J. Zou, and B. Uy. An improved ASIFT algorithm for matching repeated patterns. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2949–2952, Hong Kong, 2010.

[5] M. Brown and D.G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.

[6] T. Buades, Y. Lou, J.-M. Morel, and Z. Tang. A note on multi-image denoising. In *Proceeding of the International Workshop on Local and Non-Local Approximation in Image Processing*, Tuusalu, Finland, 2009.

[7] F. Cao, J. Delon, A. Desolneux, P. Musé, and F. Sur. A unified framework for detecting groups and application to shape recognition. *Journal of Mathematical Imaging and Vision*, 27(2):91–119, 2007.

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

[9] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 220–226, San Diego, CA, USA, 2005.

[10] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Machine Learning*, 50(1-2):45–71, 2003.

[11] H. Deng, E. N. Mortensen, L. Shapiro, and T. G. Dietterich. Reinforcement matching using region context. In *Proceedings of the Beyond Patches workshop at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, USA, 2006.

[12] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.

[13] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt theory to image analysis: a probabilistic approach*. Interdisciplinary applied mathematics. Springer, 2008.

[14] J. Domke and Y. Aloimonos. A probabilistic notion of correspondence and the epipolar constraint. In *Proceedings of the International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, Chapel Hill, NC, USA, 2006.

[15] O. Faugeras, Q.-T. Luong, and T. Papadopoulou. *The Geometry of Multiple Images*. MIT Press, 2001.

[16] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[17] P. Georgel, A. Bartoli, and N. Navab. Simultaneous in-plane motion estimation and point matching using geometric cues only. In *Workshop on Motion and Video Computing (WMVC)*, Snowbird, UT, USA, 2009.

[18] S. Gold and A. Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):377–388, 1996.

[19] L. Goshen and I. Shimshoni. Balanced exploration and exploitation model search for efficient epipolar geometry estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1230–1242, 2008.

[20] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, pages 147–151, Manchester, UK, 1988.

[21] R. Hartley. Projective reconstruction and invariants from multiple images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(10):1036–1041, 1994.

[22] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[23] J.D. Krol and W.A. van de Grind. The double-nail illusion: experiments on binocular vision with nails, needles, and pins. *Perception*, 9(6):651–669, 1980.

[24] S. Lehmann, A. P. Bradley, I. V. L. Clarkson, J. Williams, and P. J. Kootsookos. Correspondence-free determination of the affine fundamental matrix. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):82–97, 2007.

[25] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1465–1479, 2006.

[26] T.K. Leung and J. Malik. Detecting, localizing and grouping repeated scene elements from an image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 546–555, Cambridge, UK, 1996.

[27] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[28] H. Ling and K. Okada. Diffusion distance for histogram comparison. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 246–253, New York, NY, USA, 2006.

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

[30] H.C. Longuet-Higgins. A computer program for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[31] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[32] A. Makadia, C. Geyer, and K. Daniilidis. Correspondence-free structure from motion. *International Journal of Computer Vision*, 75(3):311–327, 2007.

[33] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[34] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[35] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2006.

[36] 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(3):201–218, 2004.

[37] N. D. Molton, A. J. Davison, and I. D. Reid. Locally planar patch features for real-time structure from motion. In *Proceedings of the British Machine Vision Conference (BMVC)*, Kingston University, London, UK, 2004.

[38] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *International Journal of Computer Vision*, 73(3):263–284, 2007.

[39] J.-M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[40] J.L. Mundy and A. Zisserman. Repeated structures: Image correspondence constraints and 3D structure recovery. In J. Mundy, A. Zisserman, and D. Forsyth, editors, *Applications of Invariance in Computer Vision*, pages 89–106. Springer-Verlag, 1994.

[41] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel. An a contrario decision method for shape element recognition. *International Journal of Computer Vision*, 69(3):295–315, 2006.

[42] N. Noury, F. Sur, and M.-O. Berger. Fundamental matrix estimation without prior match. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 513–516, San Antonio, TX, USA, 2007.

[43] N. Noury, F. Sur, and M.-O. Berger. How to overcome perceptual aliasing in ASIFT? In *Proceedings of the International Symposium on Visual Computing (ISVC), part I*, volume LNCS 6453, pages 231–242, Las Vegas, NV, USA, 2010.

[44] N. Noury, F. Sur, and M.-O. Berger. Modèle a contrario pour la mise en correspondance robuste sous contraintes épipolaires et photométriques. In *Actes du congrÃ¨s Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, Caen, France, 2010.

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

[46] J. Rabin, J. Delon, Y. Gousseau, and L. Moisan. MAC-RANSAC : reconnaissance automatique d'objets multiples. In *Actes du congrÃ¨s Reconnaissance de Formes et Intelligence Artificielle (RFIA)*, Caen, France, 2010.

[47] R. Roberts, S.N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, 2011.

[48] S. Roy and I.J. Cox. Motion without structure. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 1, pages 728–734, Vienna, Austria, 1996.

[49] Y. Rubner, C. Tomasi, and L.J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[50] F. Schaffalitzky and A. Zisserman. Planar grouping for automatic detection of vanishing lines and points. *Image and Vision Computing*, 18(9):647–658, 2000.

[51] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert. Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, USA, 2008.

[52] C. Schmid. A structured probabilistic model for recognition. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2485–2490, Los Alamitos, CA, USA, 1999.

[53] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[54] E. Serradell, M. Özuysal, V. Lepetit, P. Fua, and F. Moreno-Noguer. Combining geometric and appearance priors for robust homography estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 3, pages 58–72, 2010.

[55] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.

[56] G.P. Stein and A. Shashua. Model-based brightness constraints: on direct estimation of structure and motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):992–1015, 2000.

[57] B.J. Tordoff and D.W. Murray. Guided-MLESAC: faster image transform estimation by using matching priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1523–1535, 2005.

[58] P. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):35–61, 2002.

[59] P. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3):271–300, 1997.

[60] P. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78:138–156, 2000.

[61] B. Triggs and P. Bendale. Epipolar constraints for multiscale matching. In *Proceedings of the British Machine Vision Conference (BMVC)*, Aberystwyth, UK, 2010.

[62] T. Tuytelaars, A. Turina, and L. Van Gool. Noncombinatorial detection of regular repetitions under perspective skew. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):418–432, 2003.

[63] A. Vedaldi and S. Soatto. Local features, all grown up. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1753–1760, New York, NY, USA, 2006.

[64] S.D. Whitehead and D.H. Ballard. Learning to perceive and act by trial and error. *Machine Learning*, 7(1):45–83, 1991.

[65] G. Yu and J.-M. Morel. ASIFT: An Algorithm for Fully Affine Invariant Comparison. *Image Processing On Line*, 2011.

[66] W. Zhang and J. Kosecka. Generalized RANSAC framework for relaxed correspondence problems. In *Proceedings of the International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, Chapel Hill, NC, USA, 2006.

[67] Z. Zhang. Determining the epipolar geometry and its uncertainty: a review. *International Journal of Computer Vision*, 27(2):161–195, 1998.

[68] Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, 78:87–119, 1995.

Centre de recherche INRIA Nancy – Grand Est
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399