Determining point correspondences between two views under geometric constraint and photometric consistency

Nicolas Noury, Frédéric Sur, Marie-Odile Berger

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Determining point correspondences between two views under geometric constraint and photometric consistency

Nicolas Noury — Frédéric Sur — Marie-Odile Berger

N° 7246

April 2010

Vision, Perception and Multimedia Understanding

ISSN 0249-6399 ISRN INRIA/RR--7246--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
NANCY - GRAND EST

Determining point correspondences between two views under geometric constraint and photometric consistency

Nicolas Noury , Frédéric Sur , Marie-Odile Berger

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

Rapport de recherche n° 7246 — April 2010 — 39 pages

Abstract: Matching or tracking points of interest between several views is one of the keystones of many computer vision applications, especially when considering structure and motion estimation. The procedure generally consists in several independent steps, basically 1) point of interest extraction, 2) point of interest matching by keeping only the "best correspondences" with respect to similarity between some local descriptors, 3) correspondence pruning to keep those consistent with an estimated camera motion (here, consistent with epipolar constraints or homography transformation). Each step in itself is a touchy task which may endanger the whole process. In particular, repeated patterns give lots of false matches in step 2) which are hardly, if never, recovered by step 3). Starting from a statistical model by Moisan and Stival [32], we propose a new one-stage approach to steps 2) and 3), which does not need tricky parameters. The advantage of the proposed method is its robustness to repeated patterns.

Key-words: Point correspondences, SIFT matching, a contrario model, RANSAC, perceptual aliasing.

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 entre deux images, sous contraintes géométriques et photométriques

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, en particulier lorsqu'il est question d'estimation de la structure et du mouvement. La procédure consiste généralement en plusieurs étapes indépendantes, à savoir : 1) extraction des points d'intérêt, 2) appariement des points d'intérêt en gardant les « meilleures correspondances » selon la ressemblance de descripteurs locaux, et 3) 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 dans l'étape 2) qui sont difficilement rattrapés par l'étape 3). En reprenant un modèle statistique proposé par Moisan et Stival [32], nous proposons une nouvelle approche intégrant les étapes 2) et 3), qui ne nécessite pas de paramètre critique. La méthode proposée présente l'avantage d'être robuste à la présence de motifs répétés.

Mots-clés : Mise en correspondance de points, appariement SIFT, modèle a contrario, RANSAC, aliasing perceptuel.

Contents

1 Introduction

A large part of computer vision literature is based on the matching of points of interest between several views. "Matching" means that one has to detect points of interest 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 a point of interest. While only a rough matching is needed in e.g. the image retrieval context (one accepts that some correspondences are not correct), structure and motion problems call for an accurate matching step. In these latter problems, the aim is indeed to estimate the change of pose of a camera along a video sequence by tracking points of interest, and to estimate the 3D location of the corresponding points.

The following section explains why this is an intrinsically difficult problem.

1.1 From images to geometry

In this report we focus on the problem of correspondence finding between two views. This task is the keystone of many computer vision problems and it has to be solved in most multiple views structure and motion applications [18], for instance in Snavely et al.'s Phototourism and Bundler softwares to cite a single representative work [50]. Let us consider two images from the same 3D scene taken by a moving camera. A popular way to tackle the problem consists in the following steps:

1. In both views, extract points of interest 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 of step 2 and taking into account this estimation. (This step is often referred to as "guided matching".)
5. In the end, estimate the refined camera motion based on this final set of correspondences.

Let us have a deeper look on this classic methodology.

Point of interest extraction in **step 1** can be achieved by Harris-Stephens corner detector [17] or extrema of Laplacian [24] in scale-space. Following the seminal work by Mohr and Schmid [49] a large amount of methods have emerged to attach to each point of interest a local photometric descriptor, (quasi-) invariant to contrast change and to a large class of deformations. These descriptors must at least be (quasi-) invariant to contrast change and to a large class of deformations. By considering the scale given by extrema of Laplacian across scale-space and main directions of the gradient within a circle whose radius is proportional to this scale, it is possible to define neighbourhoods for points of interest which are invariant to scale and rotation change. Descriptors in themselves are made of histograms that gather statistics over gradient direction in the previously defined neighbourhood. The gradient direction is indeed invariant to contrast change, unlike the gradient norm. Actually, if g is any smooth non-decreasing

contrast change and $u(x, y)$ is an image, then $\nabla(g \circ u) = g'(u) \cdot \nabla u$. One of the most successful algorithms is probably Lowe's SIFT [28], which is based on this idea. See for example Mikolajczyk et al.'s reviews [30, 31].

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 [28] to define a set of correspondences is instead to keep the nearest neighbour, provided the distance ratio 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 are made of gradient orientation histograms, some authors propose to change the Euclidean distance to some distance that is somewhat more adapted to histograms. One can mention (by increasing computational complexity) χ^2 distance, Ling and Okada's diffusion distance [25], Earth Mover's Distance (EMD, see the seminal work by Rubner et al. [45], Rabin et al. [41], or Ling and Okada [26]). For the sake of historical completeness, let us also mention correlation methods (e.g. [62]), 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 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. Let us assume for a while that the camera motion is not restricted to a rotation around its optical center, and that the 3D points do not lie on a plane. 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) [14, 18]. 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), and point of interest location is disrupted by noise, robust statistics are called for, such as e.g. LMedS or M-estimators [57, 61]. Most popular choice is certainly RANSAC [15] and methods derived from it (MSAC, MLESAC [58], MAPSAC [56] and other methods [6] 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 geometry, and b) build a subset of correspondences that is consistent with this geometry. 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 [32] 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 [32, 37]. We will come back to this in section 2.

Once a consensus set has been found, **step 4** consists in estimating the geometry between the two views, by computing the fundamental (or essential) matrix. 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 [14, 18, 61]. 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.

One 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 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 discarded at this early stage (since the ratio would be always close to 1). Moreover, there is no insurance that the correspondence that is kept is correct, as illustrated on figure 1. Although some methods are better than other towards this point (for example [41]) it is still very difficult to match repeated patterns in a reliable fashion.

This phenomenon is sometimes called *perceptual aliasing*. This term was coined by Whitehead and Ballard [59] 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 [59]: "*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.*"

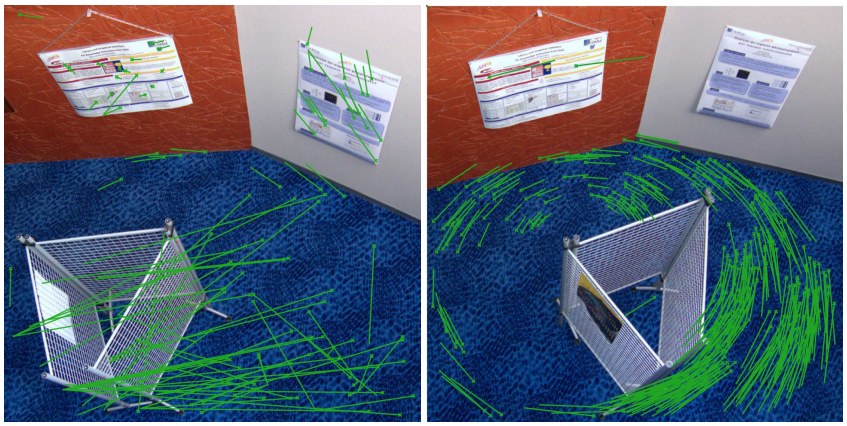

Figure 1: *Loria* image pair. SIFT points of interest are marked with a cross, the green segment represents the apparent motion with the matching feature in the other image. On the left, the standard SIFT matching algorithm (NN-T in text) fails at identifying reliable matching points of interest on the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

In this report we focus on matching points of interest between two views of the same 3D scene. The camera model which is considered here is the pinhole model,

so that the matching stage has to be consistent with epipolar geometry. We assume that the intrinsic parameters of the camera are unknown, thus the epipolar geometry is encoded either in a fundamental matrix, or in a homography if the position of the camera does not change or if a single plane can be seen in both views. We will show in experiments that recovering from perceptual aliasing in point correspondences permits a better accuracy for epipolar geometry estimation.

The contribution of this report is to replace steps 2 and 3 (and partly 4) by a all-in-one statistical framework. The proposed method allows repeated patterns matching and does not incorporate tricky parameters. We obtain much more correspondences than in the standard SIFT matching, which are distributed in a denser fashion across the scene. This can be helpful in e.g. object recognition (more correspondences means an increased confidence) or in structure and motion applications (for a better accuracy).

1.2 Related work

Reliably matching points of interest 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 previous section) by circumventing it. A possibility to address the problem of "correspondence-free" structure from motion is to use brute force techniques (i.e. considering all possible correspondences among extracted points of interest), 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. [8, 9]. 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"* [9]. The problem is solved by maximising the geometric likelihood of the very large number of possible correspondences. Their contribution is to design a specific EM algorithm with a simulated annealing step to avoid local minima. We believe that the basic assumption is too restrictive to deal with occlusion and point misdetection, which often arise in practice. Let us remark that the popular *Iterative Closest Point* (ICP) algorithm [2] has the same algorithmic purpose as the previous article. ICP indeed aims at matching point clouds related by a rigid mapping, without any prior correspondence. However, a reliable preliminary estimation of the mapping is needed and ICP is not robust to spurious measurements, although modern developments overcome these difficulties (see [21] for a single example among the huge literature about ICP).

A way to reduce the complexity is to use photometric information along geometric constraints. Domke and Aloimonos [13] 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 points of interest. Since the 5D space of all possible motions (in the calibrated case) must be explored, this approach has a heavy computational cost as in [9], although speeding up is possible when a motion estimation is known. The same basic idea was used before by Roy and Cox [44], 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 [1] 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 [51]. Optical flow [19] 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 [51] 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 [18]. However, as every method based on the optical flow, this cannot be extended to large transformations between views which makes 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. [22] 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. [29]; 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 points of interest allow to deal with these latter shortcomings. From this point of view, one can cite recent works [6, 16, 55] 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. Deng et al. [10] associate SIFT descriptors with a region context descriptor which encodes the relative position of nearby points of interest. 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 [10].

On the contrary, repeated patterns are explicitly considered by Schaffalitzky and Zisserman in [46]. The authors aim at determining vanishing lines in a single image, by pairing aligned repeated patterns. To carry on with Whitehead and Ballard's statement, perceptual aliasing is a "blessing" for them, as for Schindler et al. [47]. Indeed, repeated structures in building frontages are used in [47] to match them against a geolocalized building database, in order to achieve pose estimation in urban scenes.

A preliminary version of this work has appeared in conference proceedings [36, 38].

1.3 Organization of the report

Related work and motivation are presented in section 1. We explain the proposed statistical model in section 2. The whole method is summarized in section 3 where the algorithmic choices are motivated. Section 4 is about experimental assessment and proof of concept. We conclude with section 5. For the sake of completeness, the proofs of some propositions are given in section 6.

2 An *a contrario* model for point correspondences under epipolar constraint and photometric consistency

In this report we propose a method based on a so-called *a contrario* model. Since the seminal paper by Desolneux, Moisan and Morel [11], these models have been the subject of a large amount of literature. The books [5] and [12] and the references therein give a comprehensive account of their use in many different computer vision problems. In [4] and [35] several *a contrario* models are designed for correspondence finding between two views. Nevertheless, these models deal with *geometrical shapes* under *affine transformations* instead of *points of interest* 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. As pointed out as soon as in [11], the same idea governs Stewart's MINPRAN [52] that has been proposed as a RANSAC-like method (see Torr and Murray's survey article [57] for a discussion of MINPRAN as a robust estimation method).

The model proposed in this report is based on Moisan and Stival's *a contrario* RANSAC [32] and to some extent on Rabin et al.'s *a contrario* model for SIFT-like descriptor matching via an Earth Mover's Distance [41, 40, 39]. The first paper [32] focuses on geometric constraints and assumes that correspondences between points of interest 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 [41, 40, 39] prove that Earth Mover's Distance is better than existing dissimilarity measure between SIFT features, and investigate several *a contrario* approaches. However, geometric constraints are considered only as an additional step, while the emphasis is put on object matching.

Let us give some notations. One assumes 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 points of interest, 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 a point of interest, and $D(x_i)$ (resp. $D(y_j)$) is the corresponding local descriptor. Depending on the circumstances, we denote x_i the point of interest itself, its pixel coordinates, or its homogeneous coordinates in the projective plane.

We assume to be within the scope of the pinhole camera model. We also assume for a moment that the camera position has changed between the two views, and that

points of interest do not lie on a common plane. We will address these specific cases in section 3.3. 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 a normal vector whose expression is $F \cdot x_i$, 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. 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. If one makes the additional assumption that the 3D scene is locally planar, then one just needs invariance to homographies. Such an approach is used e.g. in [33]. Most of the time, one is satisfied with a weaker invariance, namely invariance to affine transformations [31] or to zoom+rotation (similitude) transformations which is easier to handle in practice, as in Lowe's SIFT [28]. Consequently, one wishes that $D(x_i)$ and $D(y_j)$ are "similar enough".

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 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:

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

Remark that symmetrization with respect to $\mathcal{I}_1$ and $\mathcal{I}_2$ in equation (2) 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, and thresholds δ_D and δ_G. The proposed statistical framework automatically balances geometry and photometry, and also automatically derives both thresholds relatively to a set $\mathcal{S}$.

2.1 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* methodology, 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 make 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, 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 points of interest from a pair actually correspond to

the same 3D point, and the motion of all points of interest 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 F and two thresholds δ_G and δ_D as in equations (1) and (2). The probability that should be estimated is:

$$\begin{array}{l} p(\mathcal{S}, F, \delta_G, \delta_D) := \\ \quad \Pr\big(\forall (i,j) \in \mathcal{S}, d_G(x_i, y_j, F) \leq \delta_G \text{ and } d_D(D(x_i), D(y_j)) \leq \delta_D \mid \mathcal{H}_0\big). \end{array} \tag{3}$$

Let us also assume that the fundamental matrix F is estimated from a minimal subset of $\mathcal{S}$ as in the RANSAC paradigm. This means that a subset s from $\mathcal{S}$ made of 7 correspondences is used to estimate F [61]. Remark that it would also be possible to use the 8-point linear method [27] with a slight adaptation.

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 $(d_G(x_i, y_j, F))_{(i,j)\in\mathcal{S}\setminus s}$ are mutually independent random variables.*
2. *$(d_G(x_i, y_j, F))_{(i,j)\in\mathcal{S}\setminus s}$ 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, $(d_G(x_i, y_j, F))_{(i,j)\in s}$ are also identically distributed but do not follow the same distribution function f_G as variables from $\mathcal{S}\setminus s$ since F is estimated from s, leading to the conditions $d_G(y_j, F{\cdot}x_i) \simeq 0$ and $d_G(x_i, F^T{\cdot}y_j) \simeq 0$ for every $(i,j) \in s$.

As a consequence:

Proposition 1

$$p(\mathcal{S}, F, \delta_G, \delta_D) = f_D(\delta_D)^k f_G(\delta_G)^{k-7} \tag{4}$$

where k is the cardinality of $\mathcal{S}$.

Proof: it is straightforward to derive:

$$\begin{aligned} p(\mathcal{S}, F, \delta_G, \delta_D) &= \prod_{(i,j)\in\mathcal{S}\setminus s} \Pr\big(d_G(x_i, y_j, F) \leq \delta_G\big) \cdot \prod_{(i,j)\in\mathcal{S}} \Pr\big(d_D(D(x_i), D(y_j)) \leq \delta_D\big) && (5) \\ &= f_D(\delta_D)^k f_G(\delta_G)^{k-7} && (6) \end{aligned}$$

Equation (5) comes from point 1 in definition 1 and equation (6) from points 2 and 3. ■

In the hypothesis testing paradigm, one would reject the null hypothesis $\mathcal{H}_0$ as soon as $p(\mathcal{S}, F, \delta_G, \delta_D)$ is below the predetermined significance level. However, it would mean here that, all things being equal, large groups $\mathcal{S}$ would be favoured. In the *a contrario* methodology, one does not directly deal with the probabilities but rather with the so-called *Number of False Alarms*. It corresponds to the average number of groups consistent with F, δ_G, δ_D under hypothesis $\mathcal{H}_0$.

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:*

$$\forall (i,j) \in \mathcal{S},\ d_G(x_i, y_j, F) \leq \delta_G, \tag{7}$$

$$\forall (i,j) \in \mathcal{S},\ d_D(D(x_i), D(y_j)) \leq \delta_D, \tag{8}$$

2. *a fundamental matrix F evaluated from 7 points from $\mathcal{S}$;*

such that:

$$\begin{aligned} &\mathit{NFA}(\mathcal{S}, F, \delta_G, \delta_D) := \\ &\quad 3\left(\min\{N_1, N_2\} - 7\right) k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{7} f_D(\delta_D)^k f_G(\delta_G)^{k-7} \leq \varepsilon \end{aligned} \tag{9}$$

where k is the cardinality of $\mathcal{S}$.

One can show [5, 12] that the average number of ε-meaningful sets is, under $\mathcal{H}_0$, bounded from above by ε. This justifies the expression "Number of False Alarms".

As noted in [41], this presentation is equivalent to the well known Bonferroni correction in the hypothesis testing framework: the confidence level is divided by the number of comparisons. Let us estimate this number. There are $\min\{N_1, N_2\} - 7$ choices for $k \geq 7$, $\binom{N_1}{k}$ choices for the points of interest in image 1, $\binom{N_2}{k}$ choices for the points of interest in image 2, $k!$ choices for the correspondences, $\binom{k}{7}$ choices for the minimal set to estimate F, and each minimal set possibly leads to three fundamental matrices [61].

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

Proposition 2 *A set $\mathcal{S}$ of correspondences is ε-meaningful if there exists a fundamental matrix F estimated from 7 correspondences among $\mathcal{S}$ such that:*

$$\mathit{NFA}(\mathcal{S}, F) := 3\left(\min\{N_1, N_2\} - 7\right) k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{7} f_D(\delta_D)^k f_G(\delta_G)^{k-7} \leq \varepsilon \tag{10}$$

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

The aim of the algorithm discussed in section 3 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 (10) balances the trade-off between the probability $f_D(\delta_D)^k f_G(\delta_G)^{k-7}$ and the number of possible sets of size k. 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 6)

Definition 2 was outlined in [32] (*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), that is to say a non-decreasing function over $[0, +\infty)$ such that $f_D(0) = 0$ and $f_D(+\infty) = 1$ (resp. $f_G(0) = 0$ and $f_G(+\infty) = 1$).

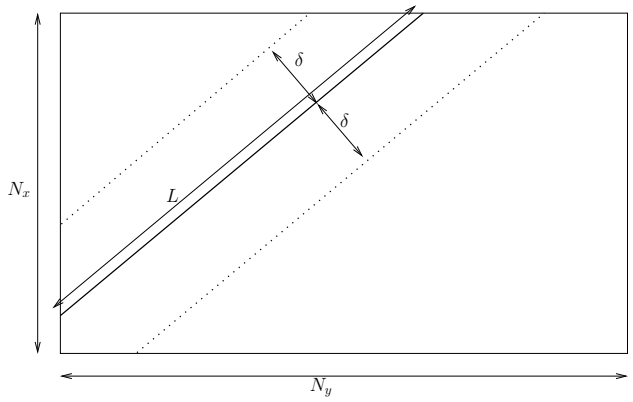

Figure 2: Moisan and Stival's model: considering uniformly distributed points within an image $\mathcal{I}$ of size $N_x \times N_y$, the probability that a point falls at a distance $\leq \delta$ to a straight line with length L is approximately $2\delta L/(N_x N_y)$. If D denotes the length of the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

2.2 Modelling the geometric constraint

2.2.1 Moisan and Stival's model [32]

Moisan and Stival [32] 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):

$$f_G(d_{\text{euc}}(y, F \cdot x)) = \frac{2D}{A} d_{\text{euc}}(y, F \cdot x) \tag{11}$$

where D and A are respectively the diameter and 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 article, Moisan and Stival not only assume independence, but also that points of interest are uniformly distributed in images. This leads them to estimate the probability that some random point (drawn from a uniform distribution) fall at a distance less than δ from an epipolar line. It is easy to see via a simple geometric argument that this probability is $\frac{2D}{A}\delta$ (it is actually an upper bound). See figure 2.

With equation (2), one derives here:

$$\begin{aligned} \Pr(d_G(x, y, F) \leq \delta_G) &= \Pr(\max\{d_G(y, F \cdot x), d_G(x, F^T \cdot y)\} \leq \delta_G) & (12) \\ &= \left(\frac{2D}{A}\delta_G\right)^2 & (13) \end{aligned}$$

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.

For a reason that will be made clear in section 3.1, we actually use:

$$f_G(\delta_G) = \left(\frac{2D}{A}\delta_G\right)^{2\alpha} \tag{14}$$

with $\alpha = 5$ once and for all.

Let us note that $\frac{2D}{A}\delta_G$ 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.

2.2.2 Taking account of point of interest location uncertainty [54]

Considering the normal vector of an epipolar line as a Gaussian random process with mean l and covariance matrix Σ_l, then, with probability p, realizations of this process lie within the hyperbola $C = ll^T - k^2\Sigma_l$ where $k = \Phi^{-1}(p)$ and Φ is the cumulative distribution function of the χ^2 law with two degrees of freedom (cf [18]).

In [54], we have proposed to use $d_G(y, F \cdot x) = k^2(y)$, where y lies on the conic given by the matrix:

$$C = ll^T - k^2(x)\Sigma_l \tag{15}$$

with $l = F \cdot x$ and Σ_l the covariance matrix of l. We use for f_G the cumulative distribution function of the χ^2 law with two degrees of freedom. The covariance matrix Σ_l is obtained in [54] after a careful derivation based on the assumption that point location comes with a Gaussian noise which propagates to fundamental matrix (via the 8-point method) and epipolar lines estimation. The case of homographies is investigated in [53]. To the best of our knowledge, covariance matrices for fundamental matrix and epipolar lines was first derived in [7] (see also [18] and references therein). Seeking points along epipolar line based on the uncertainty was also investigated in [3].

We do not elaborate on it further in this report, this possibility is mentioned for the sake of completeness.

2.3 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 [28]) 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 $d_D(D(x), D(y))$ has the same meaning in terms of "perceptual proximity" for every descriptor $D(x)$.

Rabin et al. [41] 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 [41], and based on previous works [35, 36, 38], we define:

$$d_D(D(x), D(y)) = \phi_{D(x)}(\mathsf{dist}(D(x), D(y))) \tag{16}$$

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 6). This distance therefore automatically adapts to the heterogeneousness of the descriptor space as a "contextual dissimilarity measure" (see [20] in a different context). The distance increases as the number of perceptually close descriptors grows.

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. [41] exploit these remarks to reduce the dimensionality by exploring two possible definitions for the distance between descriptors, provided a suitable distance $\widetilde{\mathsf{dist}}$ between histograms:

$$\mathsf{dist}(D(x), D(y)) = \sum_{i=1}^{N} \widetilde{\mathsf{dist}}(D^i(x), D^i(y)) \tag{17}$$

or:

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

where $(D^1(x), D^2(x), \ldots, D^N(x))$ is the set of N histograms from $D(x)$.

We denote dist-SUM (resp. dist-MAX) the dissimilarity measure from equation (17) (resp. (18)).

Let us note for every $i \in [1, N]$, $\phi_{D^i(x)}$ the empirical cumulative distribution function of $\widetilde{\mathsf{dist}}(D^i(x), D^i(y))$, when $D(y)$ spans the set of descriptors from image $\mathcal{I}_2$. Following the *a contrario* framework and under independence assumption, we define $\phi_{D(x)}$ as:

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

for the dist-SUM case (where $\otimes$ is the convolution product), and as:

$$\phi_{D(x)}(\delta) = \prod_{i=1}^{m} \int_0^{\delta} \phi_{D^i(x)}(t)dt \tag{20}$$

for the dist-MAX case.

Indeed, in the first case, $\mathsf{dist}(D(x), D(y))$ appears as the sum of N random variables (whose probability distribution is indeed the convolution product of the N marginal distributions under independence assumption), while in the latter case the distance appears as the maximum of the random variables (whose cumulative distribution function is the product of the N marginal cumulative distribution functions under independence assumption).

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

In both cases, $d_D(D(x), D(y)) = \phi_{D(x)}(\mathsf{dist}(D(x), D(y)))$ from equation (16). In order to fulfill requirements of section 2.1, one still needs to define the cumulative distribution function f_D. Since

$$f_D(t) = \Pr(\phi_{D(x)}(\mathsf{dist}(D(x), D(y))) \leq t) = t \tag{21}$$

if $f_{D(x)}$ is continuous and increasing (this is a classic property of cumulative distribution functions, see section 6), we simply set here $f_D(t) = t$.

The following paragraphs give definitions for the distance $\widetilde{\mathsf{dist}}$ between histograms.

2.3.1 Euclidean and Manhattan distances

A very basic idea with respect to descriptor matching is to simply use the Euclidean or the Manhattan distance. That is to say, if $D^i(x)_j$ is the value of the j-th bin of histogram $D^i(x)$:

$$\widetilde{\mathsf{dist}}(D^i(x), D^i(y)) = \left(\sum_{j=1}^{m} |D^i(x)_j - D^i(y)_j|^p \right)^{1/p} \tag{22}$$

with $p = 1$ (Manhattan) or $p = 2$ (Euclidean).

We refer to the corresponding choices for d_D (and f_D) as: EUC-MAX, EUC-SUM, MAN-MAX, or MAN-SUM, depending on the way the distance between histograms are aggregated to build the distance between descriptors (SUM or MAX).

2.3.2 χ^2 distance

Moreover, papers [25, 26] claims that χ^2 bin-to-bin comparison is better for descriptor matching than Euclidean distance, and does not need much more computation. We also test this distance:

$$\widetilde{\mathsf{dist}}(D^i(x), D^i(y)) = \chi^2(D^i(x), D^i(y)) = \sum_{j=1}^{m} \frac{\left(D^i(x)_j - D^i(y)_j\right)^2}{D^i(x)_j + D^i(y)_j} \tag{23}$$

We refer to this choice for d_D and f_D as: CHI2-MAX or CHI2-SUM.

2.3.3 Rabin et al.'s CEMD distance [41]

Another possibility, investigated in [26, 41, 45], is to use Earth Mover's Distance (EMD), which is especially well adapted for histogram comparison. 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 some metrics that behave well with respect to these circular histograms, such as the efficient distance recently proposed by Rabin et al. [41]. The remainder of this section is a digest of [41] to which we refer the reader for more details. We intentionally switch to the same notations as in [41].

Let us note for a while $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$ two (circular) histograms made of m bins. The circular EMD (denoted CEMD) between them is then defined as the solution of a linear program:

$$\text{CEMD}(x, y) = \min_{(\alpha_{i,j}) \in \mathcal{M}} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_{i,j} c(i, j), \tag{24}$$

where

$$\mathcal{M} = \left\{ (\alpha_{i,j}) \text{ s.t. } \alpha_{i,j} \geq 0, \sum_j \alpha_{i,j} = x_i, \sum_i \alpha_{i,j} = y_j \right\} \tag{25}$$

and $c_{i,j}$ is the so-called ground distance between bins x_i and y_j. This corresponds to the general definition of Earth Mover's Distance. For circular histograms, the authors of [41] choose the quite natural L^1 (circular) cost:

$$c_{i,j} = \frac{1}{N} \min(|i - j|, N - |i - j|). \tag{26}$$

The intuitive meaning of EMD is that it corresponds to the minimum cost that an Earth Mover has to pay to reshape histogram x into histogram y, given the cost $c_{i,j}$ to move a unit of material from bin i to bin j.

Rabin et al. prove in [42] that CEMD computation is easily tractable. They consider the cumulative histogram X^k of x starting at the k-th bin: for each $i \in \{1, \dots, N\}$, $X_i^k = \sum_{j=k}^{k+i} x_j$, where x is circularly indexed. Cumulative histogram Y^k is defined for y in the same way. Provided x and y are normalised (that is to say $\sum_i x_i = 1$), Rabin et al. show that

$$\text{CEMD}(x, y) = \min_{k=1\dots N} ||X^k - Y^k||_1. \tag{27}$$

Nevertheless, when considering photometric descriptors, orientation histograms x and y are **not** normalised (in SIFT, the *whole* descriptor is normalised). Anyway, following [41] we keep on using this formula to compute CEMD.

We refer to the corresponding choice for d_D and f_D as: CEMD-SUM and CEMD-MAX. Let us note that CEMD-SUM is intensively investigated in [41].

3 Summing-up and algorithm

3.1 Metric on correspondence sets

Given two sets of descriptors from two images, the most meaningful set of correspondences is sought. The aim is to find a set $\mathcal{S}$ and a matrix F estimated from a 7-point sample such that the Number of False Alarms (NFA) is the lowest among all possible sets. According to proposition 2, the NFA is defined as:

$$\text{NFA}(\mathcal{S}, F) := 3\,(\min\{N_1, N_2\} - 7)\,k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{7} f_D(\delta_D)^k f_G(\delta_G)^{k-7} \tag{28}$$

with

$$\delta_G = \max_{(i,j)\in\mathcal{S}} \max\{d_G(y_j, F \cdot x_i), d_G(x_i, F^T \cdot y_j)\} \tag{29}$$

and

$$\delta_D = \max_{(i,j)\in\mathcal{S}} (d_D(D(x_i), D(y_j)). \tag{30}$$

where f_G and d_G are fixed in section 2.2, and f_D and d_D are defined as one of the choices (EUC, MAN, CHI2, CEMD / MAX or SUM) described in section 2.3.

For example, the NFA associated with CEMD-MAX is:

$$\begin{aligned}\text{NFA}(\mathcal{S}, F) := 3\,(\min\{N_1, N_2\} - 7)\,k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{7} \cdot \\ \left(\max_{(i,j)\in\mathcal{S}} \prod_{k=1}^{N} \int_0^{d_{\text{CEMD-MAX}}(D(x_i), D(y_j))} \phi_{D^k(x_i)}(t) dt \right)^k \cdot \\ \left(\frac{2D}{A} \max_{(i,j)\in\mathcal{S}} \max\{d_{\text{euc}}(y_j, F \cdot x_i), d_{\text{euc}}(x_i, F^T \cdot y_j)\} \right)^{2\alpha(k-7)}\end{aligned} \tag{31}$$

One can see from equation (31) that α 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 behave as a normalization parameter, which is set once and for all to 5.

3.2 Discussing the NFA criterion

The sets with small NFA are the most relevant ones, especially when 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 [32]. For the sake of simplicity, we assume here $N_1 = N_2 = N$.

Let us note

$$M(k, N) := 3(N-7)k! \binom{N}{k}^2 \binom{k}{7}. \tag{32}$$

Figure 3 shows the graph of $-\log_{10}(M(k,N))/k$ vs k for several typical values of N. From equation (10), this gives the maximal value for the logarithm of the quantity $f_D(\delta_D) f_G(\delta_G)^{1-7/k} \simeq f_D(\delta_D) f_G(\delta_G)$ so that the corresponding group $\mathcal{S}$ is 1-meaningful (in the case $N_1 = N_2 = N$). One has indeed:

$$\text{NFA}(\mathcal{S}) \leq 1 \quad \text{iff} \quad \log_{10}\left(f_D(\delta_D) f_G(\delta_G)^{1-7/k}\right) \leq -\log_{10}(M(k,N))/k. \tag{33}$$

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

- When N is fixed, the smaller 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 putative correspondences are, the more accurate they should be with respect to geometric and photometric criteria.

3.3 Considering homographies

The previous *a contrario* model is about point correspondences under epipolar constraint. However, in degenerated cases (e.g. when the camera simply rotates around its optical center or if the 2D points lie on a common plane), the correspondences are linked through a homography.

Adapting proposition 2 (section 2.1) leads us to say that a set $\mathcal{S}$ of correspondences is ε-meaningful in this case if there exists a homography H estimated from 4 correspondences among $\mathcal{S}$ such that:

$$\text{NFA}(\mathcal{S}, H) = (\min\{N_1, N_2\} - 4)\, k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{7} f_D(\delta_D)^k f_g(\delta_G)^{k-4} \leq \varepsilon \tag{34}$$

with the same notations as in proposition 2.

One has just to adapt the definition of $f_G(\delta_G)$ from a point-line correspondence (equation (14)) to a point-point correspondence, i.e.:

$$f_G(\delta_G) = \left(\frac{\pi \delta_G^2}{A}\right)^{2\alpha}. \tag{35}$$

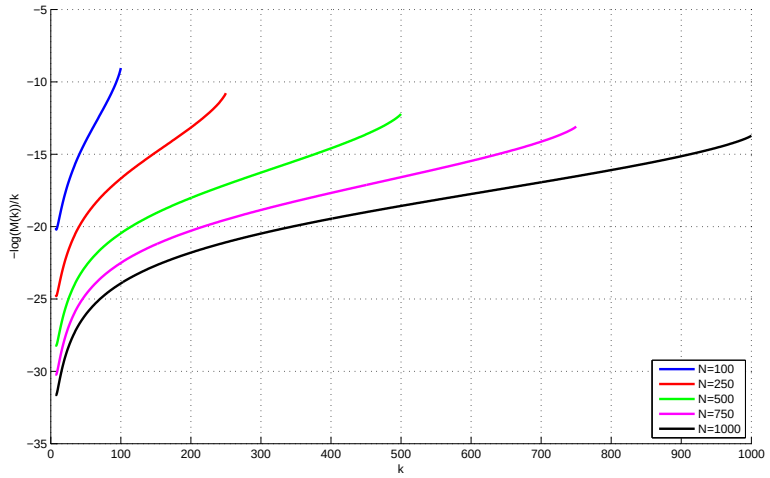

Figure 3: $-\log_{10}(M(k,N))/k$ vs k, for several values of N. This gives the order of magnitude of the logarithm of $f_D(\delta_D) \cdot f_G(\delta_G)$ so that it is still possible to find a 1-meaningful set of correspondences. (Best seen in color.)

Indeed, $\pi\delta_G^2/A$ is the probability for a random point uniformly distributed across an image (area A) fall at a distance less than δ_G from a fixed point. The same kind of model was used in a different context in [43].

As in the fundamental matrix case, $\alpha = 5$ once and for all.

Let us note that the framework can also be adapted to affine transformations (which need at least $m = 3$ points) or to zoom+rotation ($m = 2$).

The NFA would become:

$$\text{NFA}(S) := (\min\{N_1, N_2\} - m)\, k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{m} f_D(\delta_D)^k f_G(\delta_G)^{k-m} \leq \varepsilon \tag{36}$$

with $\delta_G = \max(d_G(x_i, f(y_j)))$, $\delta_D = \max(d_D(D(x_i), D(y_j))$, and k the cardinality of $\mathcal{S}$.

3.4 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

$$\sum_{k=0}^{N} k! \binom{N}{k}^2 \simeq 10^{164} \tag{37}$$

such sets (as already remarked in [1]). Since testing all possible sets is out of question, a heuristic-driven search is called for. First (section 3.4.1), we use some (large) threshold on the photometric constraint to restrict the set of putative correspondences for a given point of interest 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 3.4.2).

3.4.1 Combinatorial reduction

In order to reduce the computational burden, we do not consider all possible correspondences $y_1, \dots, y_{N_2}$ in image $\mathcal{I}_2$ for a point of interest x_i from image $\mathcal{I}_1$, but only the set of putative correspondences $y_{j_1}, \dots, y_{j_{N_i}}$ 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 [41]. In this latter case y_j is a putative correspondence to x_i if, with notations of equation (16):

$$N_1 N_2 d_D(D(x_i), D(y_j)) \leq \widetilde{\varepsilon}. \tag{38}$$

The value of $\widetilde{\varepsilon}$ does not depend on the experimental setup and is carefully discussed in [41] for the CEMD-SUM distance. Note that proposition 5 (section 6) 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 putative correspondences. In practice, we get between 0 and 30 putative correspondences for each x_i in a typical image.

3.4.2 Random sampling algorithm

At this stage each point of interest x_i from image $\mathcal{I}_1$ is matched to a set of N_i putative correspondences $y_{j_1}, \dots, y_{j_{N_i}}$ 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 in the case of the fundamental matrix:

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

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

A. Drawing a seven-correspondence sample. Seven points x_i are uniformly drawn, and then are associated to a putatively corresponding point $y_{j(i)}$. Since it gives good experimental results, we use nearest neighbour matching (in the sense of the photometry). We could also have not biased the algorithm by this choice, and instead pick up for each i the corresponding point $y_{j(i)}$ by drawing it randomly in the set $y_{j_1}, \dots, y_{j_{N_i}}$ 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. However, the outlier rate is significantly larger for non-nearest neighbours (which can be verified in experiments); this latest scheme thus needs much more iterations.

The fundamental matrix is then estimated via the non-linear "seven-point algorithm".

Remark that SIFT algorithm may extract several keypoints at the same location but with different orientations. In order to avoid degenerated cases, we check that the seven-point 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 of the undermentioned groups.

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:

$$y_{j(i)} = \underset{y_{i_k}}{\operatorname{argmin}} \{f_D(d_D(D(x_i), D(y_{j_k}))) \cdot f_G(d_G(F, x_i, y_{j_k}))\} \tag{39}$$

   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 x_is to a single y_j, which should not appear. 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 (following equation (28), with δ_G and δ_D given by equations (29) and (30)), and select the most meaningful one.

3. Sort correspondences $(x_i, y_{j(i)})$ in increasing order along $f_G(\delta_G(F, x_i, y_{j(i)}))$ 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 [32] 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.

Other random sampling strategies in a similar context are described in [60].

Let us note that in the case of homographies, building H needs 4 correspondences instead of 7 for F. Apart from this point, the algorithm is exactly the same.

3.5 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 points of interest 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 2.3, equation (16)),

(b) define a set of putative correspondences (section 3.4.1).

3. Iterate:

   (a) choose seven (*resp. four*) points x_i and pick up the seven (*resp. four*) corresponding points $y_{j(i)}$ (heuristic A from 3.4.2),

   (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 3.4.2).

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

The number of iterations in step 3 is set in section 4 to a large value (with respect to the estimated proportion of outliers) so that the returned group is actually the most meaningful one with a high confidence.

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.

4 Experimental assessment

The following experiments are all led with the algorithm described in section 3.5. We test correspondence finding under epipolar constraint (fundamental matrix) or homography.

4.1 An experiment on synthetic images

We test here the influence of the $\widetilde{\varepsilon}$ parameter (section 3.4.1), using the different distances between descriptors defined in section 2.3. 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 invariance of SIFT does not interfere. In this case, corresponding points should have descriptors that are actually alike. The test is led here with the fundamental matrix model. As an illustration, figure 4 shows some results with the CEMD-SUM distance. Note that the images have a large number of repeated patterns, the problem is thus very challenging.

Table 1 and table 2 gather 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 also reduces the number of putative 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. It would have been impossible to retrieve them with the classic SIFT matching algorithm (nearest neighbour matching, provided the ratio of the distances between the nearest and second nearest neighbour is below some threshold). In this particular experimental framework (small motion of the camera), these tables show that the most meaningful group is larger with CHI2-SUM, CEMD-SUM, CHI2-MAX, CEMD-MAX. With these distances, the ratio of nearest neighbour matches is significantly smaller with EUC-SUM, EUC-MAX and MAN-MAX, proving that these

latter distances do not sort corresponding features consistently. This confirms results from [26] and [41] in another framework.

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), and to use the CEMD-SUM metric. Actually we have noticed the CHI2 metrics give mixed results. For example, the proposed algorithm with CEMD-SUM gives a good result in the challenging experiment of figure 7, while no group at all is given by the other metrics.

distance	$\widetilde{\varepsilon}$	# putative corr.	# most meaning. group	% of rank 1 corr.
MAN-SUM	1	3094	211	79.1
	10^{-2}	2253	207	78.7
	10^{-4}	1733	206	79.6
	10^{-6}	1336	210	79.1
	10^{-8}	1001	205	79.5
	10^{-10}	808	207	80.1
EUC-SUM	1	10197	195	47.4
	10^{-2}	9365	181	54.6
	10^{-4}	7247	196	53.1
	10^{-6}	5374	185	51.4
	10^{-8}	3922	203	47.4
	10^{-10}	2865	210	53.3
CHI2-SUM	1	3091	225	79.1
	10^{-2}	2088	227	79.2
	10^{-4}	1638	233	78.5
	10^{-6}	1296	228	78.9
	10^{-8}	1020	227	78.9
	10^{-10}	794	229	80.4
CEMD-SUM	1	2027	232	75.8
	10^{-2}	1409	227	77.1
	10^{-4}	999	223	76.7
	10^{-6}	663	206	77.6
	10^{-8}	407	178	85.6
	10^{-10}	274	147	90.5

Table 1: *Synthetic* images. Comparison of descriptor distance and 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 putative correspondences retrieved after the combinatorial reduction step (section 3.4.1), the cardinality of the most meaningful group, and the proportion of nearest neighbours among this group of correspondences.

distance	$\widetilde{\varepsilon}$	# putative corr.	# most meaning. group	% of rank 1 corr.
MAN-MAX	1	10290	206	52.4
	10^{-2}	10290	219	50.6
	10^{-5}	10290	185	52.4
	10^{-10}	10270	197	58.8
	10^{-15}	5478	201	49.8
	10^{-20}	2529	206	50.0
EUC-MAX	1	10290	205	45.8
	10^{-2}	10290	213	43.6
	10^{-5}	10290	187	50.2
	10^{-10}	10019	173	54.3
	10^{-15}	5035	217	51.2
	10^{-20}	2248	196	59.1
CHI2-MAX	1	9984	227	81.5
	10^{-2}	7670	217	82.4
	10^{-5}	3783	222	84.7
	10^{-10}	1736	228	82.4
	10^{-15}	1018	228	81.5
	10^{-20}	606	226	81.4
CEMD-MAX	1	8096	225	76.9
	10^{-2}	4126	227	76.7
	10^{-5}	3215	226	76.5
	10^{-10}	1144	223	77.1
	10^{-15}	494	200	83.1
	10^{-20}	234	145	89.6

Table 2: *Synthetic* images. Comparison of descriptor distance and influence of $\widetilde{\varepsilon}$ (continued.) Some cases give 10290 correspondences, which corresponds to an implementation parameter: this simply means that $\widetilde{\varepsilon}$ prunes the set of correspondences to a constant-size set.

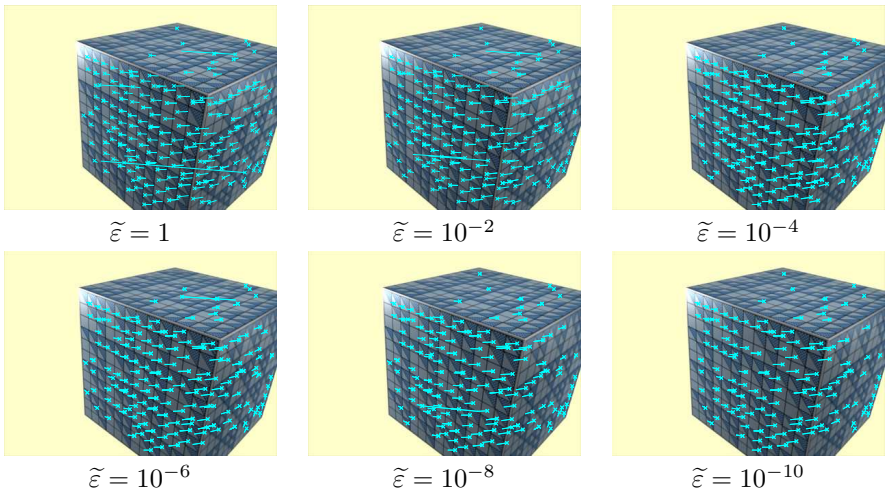

Figure 4: *Synthetic* images. In this experiment, we search for the most meaningful group consistent with a fundamental matrix between two views, with the CEMD-SUM distance, for six values of $\widetilde{\varepsilon}$. We only show here the first view, the blue segment corresponds to the apparent motion of a point of interest (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 tables 1 and 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.

4.2 Point correspondences and 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 points of interest 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 1.1:

- NN-T matching (Euclidean distance, threshold on the ratio set to 0.6 as in Lowe's code).
- Robust selection with the *a contrario* RANSAC from [32].

The use of this two-step scheme is called NN-T+O, our method AC for *a contrario*.

4.2.1 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 5: 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 5, retrieves much more correspondences. The numerous extra correspondences coincide with matches which are not nearest neighbour for the descriptor distance. Following table 3, we show that for 128 features matched, 42 have first rank, and 86 higher ranks.

	Number of correspondences			
Rank	*Monkey*	*Loria*	*Sears*	*Flat Iron*
1	42	98	532	8
2	23	32	24	3
3	17	29	12	1
4	11	18	3	1
5	8	20	5	0
6	8	19	3	0
7	4	11	3	0
8	8	5	2	0
9	3	13	1	0
10	2	8	0	0
11	0	15	0	0
12	2	7	1	0
>12	0	37	3	0
Total	128	312	589	13

Table 3: Number of occurrences of the n-th nearest neighbours selected by the AC method. *Monkey* corresponds to figure 5 (412 vs 445 extracted keypoints), *Loria* to figure 1 (2,562 vs 2,686), *Sears* to figure 6 (2,787 vs 2,217), *Flat Iron* to figure 7 (756 vs 598). Remark the strong perceptual aliasing in these pairs. As noted in figure 1 the NN-T+O method does not succeed at all in *Loria* experiment. 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.

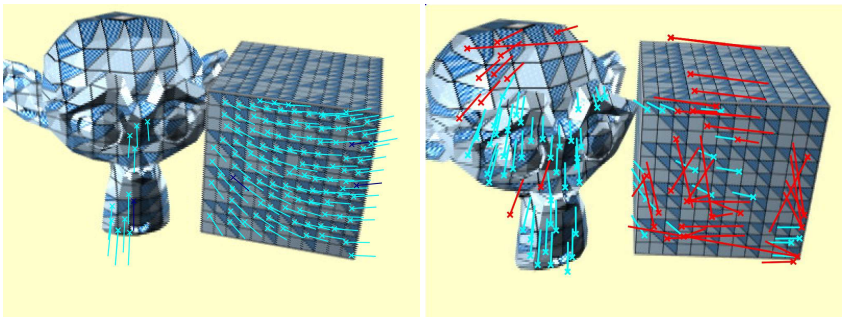

Figure 5: *Monkey*, homographic constraint. Two images with repeated patterns. On the left, the proposed AC model, most of the patterns lying on the dominant plane are detected (segments represent the apparent motion between the two views). On the right, the second image but with correspondences from NN-T (both colors) and NN-T+O (inliers in blue, outliers in red). Many more correspondences are retrieved with the AC algorithm.

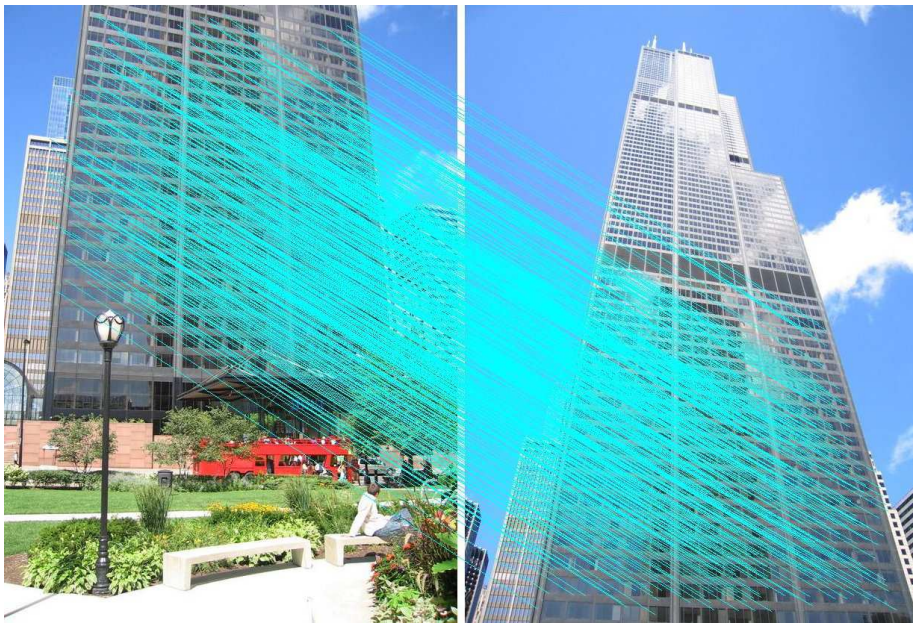

Figure 6: *Sears*, homographic constraint. 589 matches can be found with AC method.

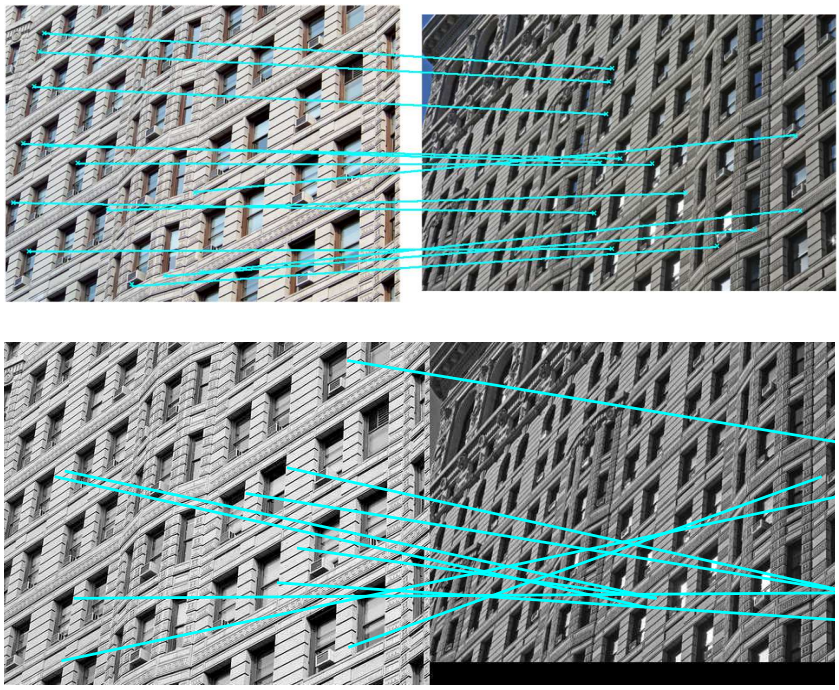

Figure 7: *Flat Iron*, homographic constraint. 13 matches can be found with AC method (top), all of them are correct. Here NN-T+O does not give any consistent group. Indeed, NN-T method (bottom, the threshold is relaxed to 0.7) provides only false correspondences. Every subsequent RANSAC will thus fail to find a consistent geometry. This is a very challenging problem: note the strong perceptual aliasing and the quite strong illumination change.

4.2.2 Repeated patterns and epipolar constraint

In this section, we test the behaviour of the AC algorithm under epipolar constraint. An example can be seen on figure 8. Since the epipolar constraint acts more "gently" than the homographic one (it is a point/line constraint), some false correspondences are simply unavoidable. Actually, two points can be associated because they fall "by chance" in the vicinity of the associated epipolar line, although they do not correspond to the same 3D point. Such a situation cannot be disambiguated from two views. This causes a serious problem when confronted to perceptual aliasing. If the repeated patterns are nearly aligned with the epipolar lines, then false correspondences between similar patterns are unavoidable, leading to an inconsistent meaningful set. One can realize that this phenomenon happens when repeated patterns are parallel to the baseline of the two cameras, which is a quite common situation. Such correspondences can be seen for example in figures 9 and 10.

These two latter figures show the comparison between the NN-T+O and the AC methods. Since the AC method is more robust to perceptual aliasing, we get more correspondences that are distributed in a denser fashion across the views. However, a careful examination shows that many correspondences are not correct, in spite that the keypoints actually lie near their epipolar lines. As one can see, this still allows us to get a more 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 $(Fx, F^T y)$. 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 by the Powell conjugate gradient algorithm [18, 61]. While a dramatic discrepancy can be seen with NN-T+O (figure 9), AC method permits a more accurate estimation.

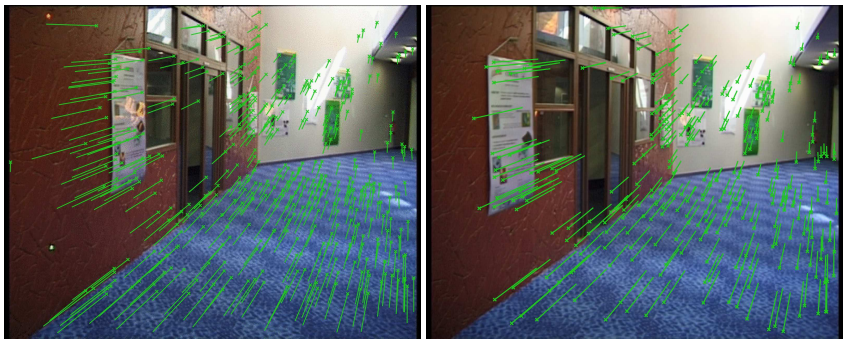

Figure 8: *Corridor*, epipolar constraint. AC method (on the left) retrieves 423 correspondences. 405 correspondences have rank 1, 13 rank 2, 2 rank 3, 1 rank 4, 1 rank 6, 1 rank 10. NN-T+O method (on the right) retrieves 295 out of 316 NN-T matches. The additional correspondences are on the carpet and on the wall. 1,269 keypoints were extracted from image 1, 1,360 from image 2.

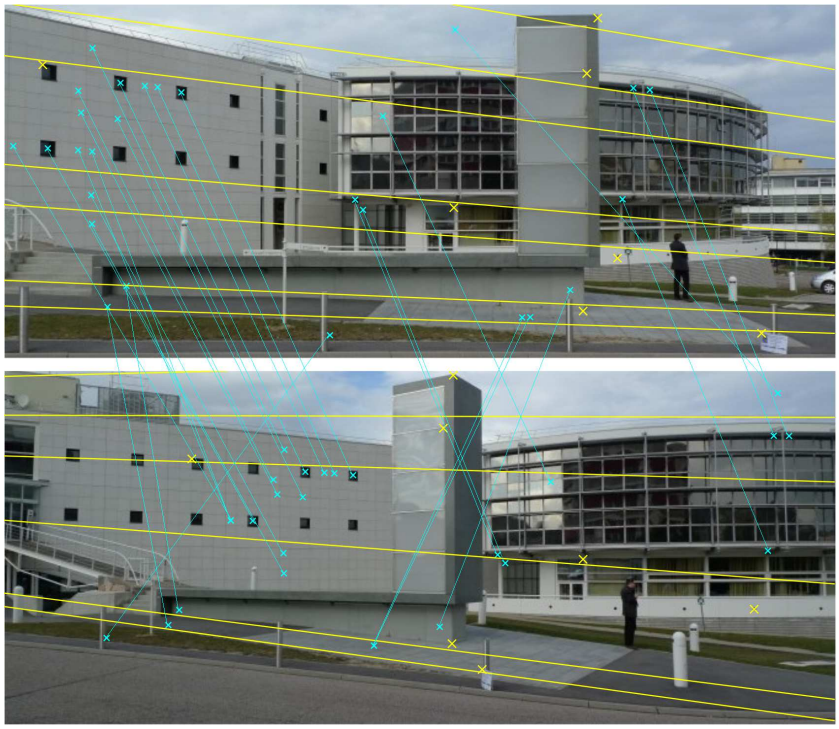

Figure 9: *Loria building*. NN-T+O method (epipolar constraint) between the top and bottom views. Blue segments represent the apparent motion of keypoints, marked by a cross. 630 keypoints were extracted from image 1, 603 from image 2. Some false correspondences can be seen. However, their distance to the corresponding epipolar line is actually small. Note that the apparent motion of the foreground (a parallelipedic structure) is quite different from the motion of the background. Some hand-picked pairs of points (in yellow) show that the epipolar pencil (reestimated from the whole consensus set) noticeably deviates from the actual one. The epipolar lines should indeed meet the yellow crosses. The distance is actually between 4 and 20 pixels.

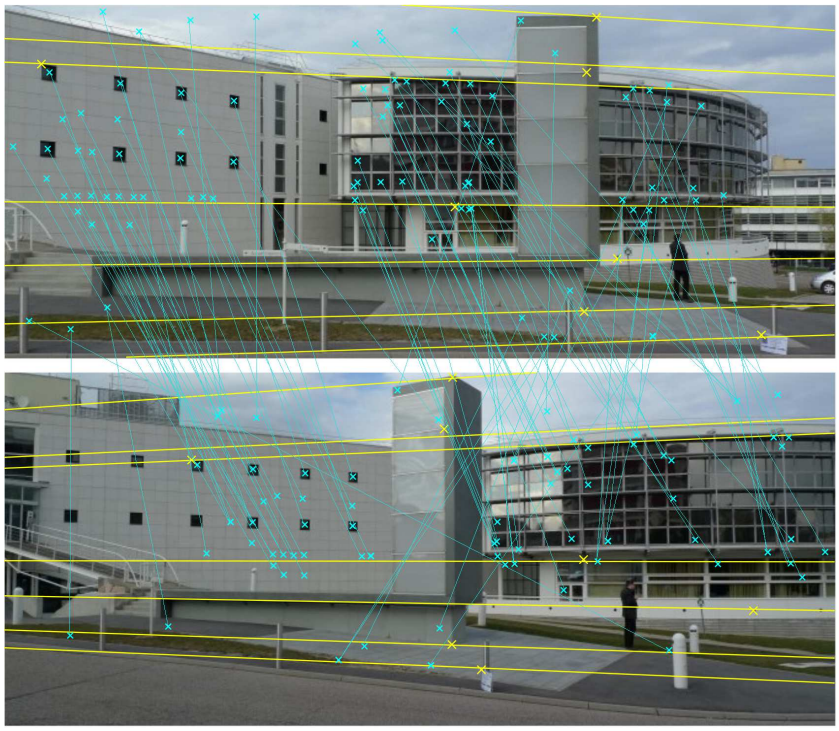

Figure 10: *Loria building*. AC method (epipolar constraint) between the top and bottom views. Compared to figure 9, more correspondences can be seen. In particular, the repeated left-hand windows are all retrieved. However, many "unavoidable" false correspondences are also retrieved, as the ones between the structures of the left-hand frontage which are indeed shifted along the epipolar lines (compare to the position of the windows; the same phenomenon appears on the right-hand frontage). 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 which is still at 15 pixels.

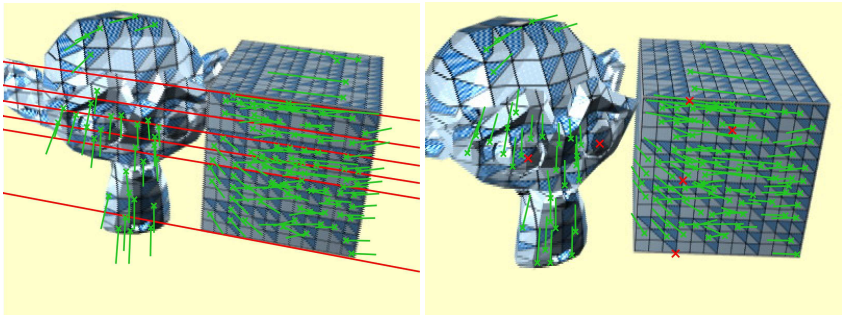

Figure 11: *Monkey*, epipolar constraint. Failure case study. The recovered geometry corresponds to the vanishing lines. In that case, the point / line constraint does not solve the ambiguity intrinsic to perceptual aliasing, and thus gives false correspondences. A few hand inserted points in red show the epipolar lines pencil, which corresponds to the pattern alignment along the vanishing lines, and not to the true motion.

4.3 When perceptual aliasing cannot be overcome

If the number of repeated patterns is very large, and if they are along a pencil of lines, then this leads to a dramatic failure of the AC algorithm in the epipolar case. Indeed, when looking for point correspondences under epipolar constraint, we look for the "most consistent" set (in the sense of the NFA) with respect to a pencil of straight line. These straight lines are hopefully the epipolar lines. However, in the considered situation, there are possibly too many false correspondences across the lines along which repeated patterns are distributed. The retrieved pencil then degenerates to a bunch of vanishing lines, and epipoles degenerate to vanishing points. For example, in figure 11 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. Note that if most features are not in a dominant plane, then the AC algorithm is not trapped by vanishing lines and is able to retrieve the correct geometry, even in the case where images are essentially made of repeated patterns (as in figure 4 where points are not concentrated over a single plane). Let us also note that the stricter point-point constraint from the homography case (compare figure 11 to figure 5) enables to retrieve a consistent set, unlike the epipolar constraint case. We end up pointing out that finding correspondences under epipolar constraint is simply impossible if the repeated patterns are distributed along lines in a dominant plane. This shortcoming is common to every RANSAC-like method.

4.4 Assessment of the *a contrario* model on unrelated images

Figure 12 shows the result of the AC algorithm (homography) on two views of different buildings and parking lots. One still retrieves 9 correspondences (error is smaller than 0.9 pixel) over 1,000 extracted keypoints. This corresponds to a true "false alarm" in the sense that the correspondences are not correct (the underlying 3D objects are not the same), but they are consistent with a physical truth. When looking for correspondences in pairs of unrelated images, in all cases the NFA of the most meaningful group is above 1 or this group is made of a very small number of correspondences.

Figure 12: Two unrelated images. The buildings and parking lots are *not* the same. 9 correspondences are retrieved by AC algorithm (homography case, segments represent the apparent motion). They are mainly made of SIFT keypoints aligned with the white line and the shadow under the car. Other keypoints are extracted with a very small scale over very contrasted edges, the associated descriptor is therefore not discriminative enough.

5 Conclusion

In this report we have presented a statistical *a contrario* model to determine correspondences between points of interest from two images. While most methods from the literature treat photometric consistency and geometric constraint separately, the proposed method integrates them in a single metric, namely a Number of False Alarms. The contribution of the method is a significant improvement in the number of correspondences that are obtained, especially when considering repetitive patterns (perceptual aliasing). Besides, it does not involve touchy parameters.

Some situations with repetitive patterns are still not tractable with this method. When pattern repetition is distributed along an epipolar line, it is simply impossible to correctly match them in the two-view geometry framework. Introducing a convenient third view would enable disambiguation in some cases. Taking into account some shape context as in [10] or the local distribution of the points of interest as in [48] could also help.

Let us also note that point correspondence finding is intrinsically limited by the lack of geometric invariance of SIFT keypoints, or of the (pseudo) affine invariant keypoints [30, 31] which could also be used. Methods like A-SIFT [34] or Lepetit et al [23] would be interesting in wide baseline matching since they provide a better invariance by generating affine transformations of the images.

6 Appendix: some proofs

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

$$M(k,N) = 3(N-7)k!\binom{N}{k}^2\binom{k}{7}. \tag{40}$$

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,

$$\frac{M(k+1,N)}{M(k,N)} = \frac{(N-k)^2}{k-6}. \tag{41}$$

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 propositon 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:*

$$\Pr(F(X) \leq x) \geq x \tag{42}$$

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 $t \leq F^{-1}(x)$.

One has successively:

$$\Pr(F(X) \leq x) = \Pr(X \leq F^{-1}(x)) = F(F^{-1}(x)) \geq x \tag{43}$$

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 2.3.

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 $t \in [0,1]$:

$$\Pr(d(D,D') \leqslant t) = \Pr(f_D(\mathsf{dist}(D,D') \leqslant t) = t \tag{44}$$

from proposition 4. ■

References

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

[2] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[3] S. Brandt. On the probabilistic epipolar geometry. *Image and Vision Computing*, 26(3):405–414, 2006.

[4] 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.

[5] 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.

[6] 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.

[7] G. Csurka, C. Zeller, Z. Zhang, and O. Faugeras. Characterizing the uncertainty of the fundamental matrix. *Computer Vision and Image Understanding*, 68(1):18–36, 1997.

[8] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. Structure from motion without correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, Hilton Head, SC, USA, 2000.

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

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

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

[19] B.K.P. Horn and B.G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[20] H. Jégou, C. Schmid, H. Harzallah, and J. Verbeek. Accurate image search using the contextual dissimilarity measure. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2009. to appear.

[21] S. Kaneko, T. Kondo, and A. Miyamoto. Robust matching of 3D contours using Iterative Closest Point algorithm improved by M-estimation. *Pattern Recognition*, 36(9):2041–2047, 2003.

[22] 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.

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

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

[25] 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 City, NY, USA, 2006.

[26] 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.

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

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

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

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

[31] 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.

[32] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3):201–218, 2004.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] N. Noury, F. Sur, and M.-O. Berger. Modèles statistiques pour l'estimation de la matrice fondamentale. In *Acte du Congrès francophone des jeunes chercheurs en vision par ordinateur ORASIS*, Obernai (France), 2007.

[38] 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.

[39] J. Rabin, J. Delon, and Y. Gousseau. Circular Earth Mover's Distance for the comparison of local features. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Tampa, FL, USA, 2008.

[40] J. Rabin, J. Delon, and Y. Gousseau. A contrario matching of sift-like descriptors. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Tampa, FL, USA, 2008.

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

[42] J. Rabin, J. Delon, and Y. Gousseau. Transportation distances on the circle. Preprint arXiv:0906.5499, 2009.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

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

[50] 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.

[51] 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.

[52] C.V. Stewart. MINPRAN: a new robust estimator for computer vision. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(10):925–938, 1995.

[53] F. Sur. Robust matching in an uncertain world. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010. To appear.

[54] F. Sur, N. Noury, and M.-O. Berger. Computing the uncertainty of the 8 point algorithm for fundamental matrix estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, pages 965–974, Leeds, UK, 2008.

[55] 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.

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

[57] 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.

[58] 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.

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

[60] 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.

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

[62] 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