A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry

Zhengyou Zhang, Rachid Deriche, Olivier Faugeras, Quang-Tuan Luong

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry

Zhengyou ZHANG

Rachid DERICHE - Olivier FAUGERAS - Quang-Tuan LUONG

N° 2273

May 1994

PROGRAMME 4

Robotique,

image

et vision

ISSN 0249-6399

A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry

Zhengyou ZHANG
Rachid DERICHE - Olivier FAUGERAS - Quang-Tuan LUONG

Programme 4 — Robotique, image et vision
Projet Robotvis

Rapport de recherche n˚ 2273 — May 1994 — 38 pages

Abstract: This paper proposes a robust approach to image matching by exploiting the only available geometric constraint, namely, the epipolar constraint. The images are uncalibrated, namely the motion between them and the camera parameters are not known. Thus, the images can be taken by different cameras or a single camera at different time instants. If we make an exhaustive search for the epipolar geometry, the complexity is prohibitively high. The idea underlying our approach is to use some classical techniques (correlation and relaxation methods in our particular implementation) to find an initial set of matches, and then use a robust technique—the Least Median of Squares (LMedS)—to discard false matches in this set. The epipolar geometry can then be accurately estimated using a well adapted criterion. More matches are eventually found, as in stereo matching, by using the recovered epipolar geometry. A large number of experiments have been carried out, and very good results have been obtained.

Regarding the relaxation technique, we define a new measure of matching support, which allows a higher tolerance to deformation with respect to rigid transformations in the image plane and a smaller contribution for distant matches than for nearby ones. A new strategy for updating matches is developed, which only selects those matches having both high matching support and low matching ambiguity. The update strategy is different from the classical "winner-take-all", which is easily stuck at a local minimum, and also from "looser-take-nothing", which is usually very slow. The proposed algorithm has been widely tested and works remarkably well in a scene with many repetitive patterns.

Key-words: Robust Matching, Epipolar Geometry, Fundamental Matrix, Least Median Squares (LMedS), Relaxation, Correlation.

(Résumé : tsvp)

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

Mise en correspondance robuste d'images non calibrées par recouvrement de la géométrie épipolaire

Résumé : Dans cet article, nous proposons une approche robuste au problème de la mise en correspondance de primitives dans le cas d'images non calibrées i.e on ne dispose ni du mouvement entre les caméras ni des paramètres intrinsèques associés à chacune des caméras. Les images peuvent ainsi être considérées comme prises par une même caméra à différents instants ou par un système stéréoscopique de 2 caméras. Dû à la complexité de la tâche, une recherche exhaustive de la géométrie épipolaire ne peut évidemment être entreprise. L'idée principale de l'approche que nous développons est d'utiliser, dans une première phase, des méthodes classiques de type corrélation et relaxation afin de trouver un premier ensemble d'appariements entre points à forte courbure des 2 images. Une seconde phase fait alors appel à des méthodes robustes de type - Moindres médianes des erreurs au carré (Least Median of Squares ou LMedS) - afin d'éliminer les éventuels faux appariements. La géométrie épipolaire est alors estimée à l'aide de cet ensemble d'appariements et de l'utilisation de critères de minimisation adéquats et bien adaptés au problème de sa détermination. Une ultime phase de recherche des appariements respectant la contrainte de la ligne épipolaire peut alors être effectuée afin d'améliorer encore les résultats. Un grand nombre d'expériences ont été menées et d'excellents résultats ont été obtenus.
Concernant la technique de relaxation utilisée, nous définissons une nouvelle mesure d'appariement qui associe une grande tolérance aux déformations rigides dans le plan image, et qui associe une faible (resp. grande) contribution aux appariements lointains (resp. proches). Une nouvelle stratégie pour la mise à jour des appariements est développée. Elle permet de ne sélectionner que les appariements qui ont un large support de correspondance et un faible score d'ambiguïté. Cette mise à jour est différente de l'approche classique dite "le vainqueur prend tout" (winner-take-all) sujette au problème des minima locaux, et aussi différente de l'approche "le perdant ne gagne rien" (looser take nothing) qui est très lente. L'algorithme proposé a été largement testé et fonctionne remarquablement bien sur des scènes à motifs périodiques.

Mots-clé : Mise en correspondance robuste, Géometrie épipolaire, Matrice fondamentale, Moindres médianes (LMedS), Relaxation, Corrélation.

Contents

List of Figures

1 Introduction

Matching different images of a single scene remains one of the bottlenecks in computer vision. A large amount of work has been carried out during the last 15 years, but the results are, however, not satisfactory. The only geometric constraint we know between two images of a single scene is the *epipolar constraint*. However, when the motion between the two images is not known, the epipolar geometry is also unknown. The methods reported in the literature all exploit some heuristics in one form or another, for example, intensity similarity, which are not applicable to most cases. The difficulty is partly bypassed by taking long sequences of images over short time interval [9, 58]. Indeed, as the time interval is small and object velocity is constrained by physical laws, the interframe displacements of objects are bounded, i.e., the correspondence of a token at the subsequent instant must be in the neighborhood of the first. However, in many cases, such as a pair of uncalibrated stereo images, this technique does not apply. Developing a robust image matching technique is thus very important.

Over the years numerous algorithms for image matching have been proposed. They can roughly be classified into two categories:

1. **Template matching.** In this category, the algorithms attempt to correlate the grey levels of image patches in the views being considered, assuming that they present some similarity [4, 15, 16, 7, 14]. The underlying assumption appears to be a valid one for relatively textured areas and for image pairs with small difference; however it may be wrong at occlusion boundaries and within featureless regions.
2. **Feature matching.** In this category, the algorithms first extract salient primitives from the images, such as edge segments or contours, and match them in two or more views. An image can then be described by a graph with primitives defining the nodes and geometric relations defining the links. The registration of two maps becomes the mapping of the two graphs: *subgraph isomorphism*. Common techniques are tree searching, relaxation, maximal clique detection, etc. Some heuristics such as assuming affine transformation between the two images are usually introduced to reduce the complexity. These methods are fast because only a small subset of the image pixels are used, but may fail if the chosen primitives cannot be reliably detected in the images. The following list of references is by no means exhaustive: [54, 50, 5, 6, 45, 35, 22, 55]

The approach we propose in this paper aims at exploiting the only geometric constraint, i.e., the epipolar constraint, to establish robustly correspondences between two perspective images of a single scene. We first extract high curvature points and then match them using a classical correlation technique followed by a new fuzzy relaxation procedure. More precisely, our algorithm consists of three steps:

- Establish initial correspondences using some classical techniques,
- Estimate robustly the epipolar geometry,
- Establish correspondences using estimated epipolar geometry as in stereo matching.

The basic idea is first to estimate robustly the epipolar geometry, and then reduce the general image matching problem to stereo matching. In the subsequent sections, we will first

review the epipolar geometry, and then describe in detail the three steps of the proposed approach. A preliminary version of this paper appeared in the proceedings of the third European Conference on Computer Vision [12].

A similar idea has been independently exploited by Xu et al. [57, 40], who also searched for image correspondences through the recovery of the epipolar geometry. There are however two main differences:

- The weak perspective camera model is used in their work, and a full perspective model is used in ours. The choice of the most appropriate criterion for the recovery of the epipolar geometry is not addressed in their work.
- The search for the epipolar geometry is carried out with an exhaustive strategy in their work. The complexity is prohibitively high even for a weak perspective model ($O(m^4n^4)$, where m and n are the number of points in the first and second image, respectively). The complexity is reduced by checking only a few closest points. In our work, some classical techniques are applied to find an initial set of correspondences.

We could apply the same strategy as that of Xu et al. [57, 40]. In fact, it has been applied to solve the correspondence problem between two sets of 3D line segments [59]. When applying it to the problem addressed in this paper, we need 8 point correspondences in order to estimate the epipolar geometry. The complexity is then $O(m^8n^8)$. Suppose both m and n are 100, the complexity is in the order of 10^{32}! Xu et al. [57, 40] deal with also the motion segmentation problem using epipolar constraint, which is not addressed in this paper.

Recently, computer vision researchers have paid much attention to the robustness of vision algorithms because the data are unavoidably error prone [17, 60]. Many the so-called *robust regression* methods have been proposed that are not so easily affected by outliers [25, 48]. The reader is referred to [48, Chap. 1] for a review of different robust methods. The two most popular robust methods are the *M-estimators* and the *least-median-of-squares* (LMedS) method (see Sect. 6.3). Kumar and Hanson [26] compared different robust methods for pose refinement from 3D-2D line correspondences, while Meer et al. [38], for image smoothing. Haralick et al. [18] applied M-estimators to solve the pose problem from point correspondences. Thompson et al. [51] applied the LMedS estimator to detect moving objects using point correspondences between orthographic views. Other recent works on the application of robust techniques to motion segmentation include [52, 42, 3].

Regarding the robust recovery of the epipolar geometry, our work is closely related to that of Olsen [43] and that of Shapiro and Brady [49]. Olsen uses a linear method to estimate the epipolar geometry, which has already been shown in one of our previous work [32] to be insufficiently accurate. He further assumes that knowledge of the epipolar geometry, as in many practical cases, is available. In particular, he assumes the epipolar lines are almost aligned horizontally. This knowledge is then used to find matches between the stereo image pair, and a robust method (the M-estimator, see Sect. 6.3) is used to detect false matches and to obtain a better estimate of the epipolar geometry. Shapiro and Brady also use a linear method. The camera model is however a simplified one, namely an affine camera. Correspondences are established by tracking corner features over time. False matches are

rejected through a *regression diagnostic*, which computes an initial estimate of the epipolar geometry over all matches, and sees how the estimate changes if a match is deleted. The match whose removal maximally reduces the residual is identified to be an *outlier* and is rejected. The procedure is then repeated with the reduced set of matches until all outliers have been removed. These two approaches (M-estimators and Regression diagnostics) work well when the percentage of outliers is small and more importantly when their derivations from the valid matches are not too large, as in the above two works. In the case described in this paper, two images can be quite different. There may be a large percentage of false matches (usually around 20%, sometimes 40%) using heuristic matching techniques such as correlation, and a false match may be completely different from the valid matches. The robust technique described in this paper deals with these issues and can theoretically detect outliers when they make up as much as 50% of whole data.

2 Notation

A camera is described by the widely used pinhole model. The coordinates of a 3-D point $M = [x, y, z]^T$ in a world coordinate system and its retinal image coordinates $\mathbf{m} = [u, v]^T$ are related by

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbb{P} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} ,$$

where s is an arbitrary scale, and $\mathbb{P}$ is a 3×4 matrix, called the perspective projection matrix. Denoting the homogeneous coordinates of a vector $\mathbf{x} = [x, y, \cdots]^T$ by $\tilde{\mathbf{x}}$, i.e., $\tilde{\mathbf{x}} = [x, y, \cdots, 1]^T$, we have $s\tilde{\mathbf{m}} = \mathbb{P}\tilde{M}$.

The matrix $\mathbb{P}$ can be decomposed as

$$\mathbb{P} = \mathbf{A} [\mathbf{R}\ \mathbf{t}] ,$$

where $\mathbf{A}$ is a 3×3 matrix, mapping the normalized image coordinates to the retinal image coordinates, and $(\mathbf{R}, \mathbf{t})$ is the 3D displacement (rotation and translation) from the world coordinate system to the camera coordinate system. The most general matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} -fk_u & fk_u \cot\theta & u_0 \\ 0 & -\frac{fk_v}{\sin\theta} & v_0 \\ 0 & 0 & 1 \end{bmatrix} , \tag{1}$$

where

- f is the focal length of the camera,
- k_u and k_v are the horizontal and vertical scale factors, whose inverses characterize the size of the pixel in the world coordinate unit,

- u_0 and v_0 are the coordinates of the principal point of the camera, i.e., the intersection between the optical axis and the image plane, and
- θ is the angle between the retinal axes. This parameter is introduced to account for the fact that the pixel grid may not be exactly orthogonal. In practice, however, it is very close to $\pi/2$.

As is clear, we cannot separate f from k_u and k_v. We thus have five intrinsic parameters for each camera: $\alpha_u = -fk_u$, $\alpha_v = -fk_v$, u_0, v_0 and θ.

The first and second images are respectively denoted by I_1 and I_2. A point $\mathbf{m}$ in the image plane I_i is noted as $\mathbf{m}_i$. The second subscript, if any, will indicate the index of the point in consideration.

3 Epipolar Geometry

Considering the case of two cameras as shown in Fig. 1. Let C_1 and C_2 be the optical centers of the first and second cameras, respectively. Given a point $\mathbf{m}_1$ in the first image, its corresponding point in the second image is constrained to lie on a line called the *epipolar line* of $\mathbf{m}_1$, denoted by $\mathbf{l}_{\mathbf{m}_1}$. The line $\mathbf{l}_{\mathbf{m}_1}$ is the intersection of the plane Π, defined by $\mathbf{m}_1$, C_1 and C_2 (known as the *epipolar plane*), with the second image plane I_2. This is because image point $\mathbf{m}_1$ may correspond to an arbitrary point on the semi-line C_1M (M may be at infinity) and that the projection of C_1M on I_2 is the line $\mathbf{l}_{\mathbf{m}_1}$. Furthermore, one observes that all epipolar lines of the points in the first image pass through a common point $\mathbf{e}_2$, which is called the *epipole*. $\mathbf{e}_2$ is the intersection of the line C_1C_2 with the image plane I_2. This can be easily understood as follows. For each point $\mathbf{m}_{1k}$ in the first image I_1, its epipolar line $\mathbf{l}_{\mathbf{m}_{1k}}$ in I_2 is the intersection of the plane Π^k, defined by $\mathbf{m}_{1k}$, C_1 and C_2, with image plane I_2. All epipolar planes Π^k thus form a pencil of planes containing the line C_1C_2. They must intersect I_2 at a common point, which is $\mathbf{e}_2$. Finally, one can easily see the symmetry of the epipolar geometry. The corresponding point in the first image of each point $\mathbf{m}_{2k}$ lying on $\mathbf{l}_{\mathbf{m}_{1k}}$ must lie on the epipolar line $\mathbf{l}_{\mathbf{m}_{2k}}$, which is the intersection of the same plane Π^k with the first image plane I_1. All epipolar lines form a pencil containing the epipole $\mathbf{e}_1$, which is the intersection of the line C_1C_2 with the image plane I_1. The symmetry leads to the following observation. If $\mathbf{m}_1$ (a point in I_1) and $\mathbf{m}_2$ (a point in I_2) correspond to a single physical point M in space, then $\mathbf{m}_1$, $\mathbf{m}_2$, C_1 and C_2 must lie in a single plane. This is the well-known *co-planarity constraint* or *epipolar equation* in solving motion and structure from motion problems when the intrinsic parameters of the cameras are known [29].

Let the displacement from the first camera to the second be $(\mathbf{R}, \mathbf{t})$. Let $\mathbf{m}_1$ and $\mathbf{m}_2$ be the images of a 3-D point M on the cameras. Without loss of generality, we assume that M is expressed in the coordinate frame of the first camera. Under the pinhole model, we have

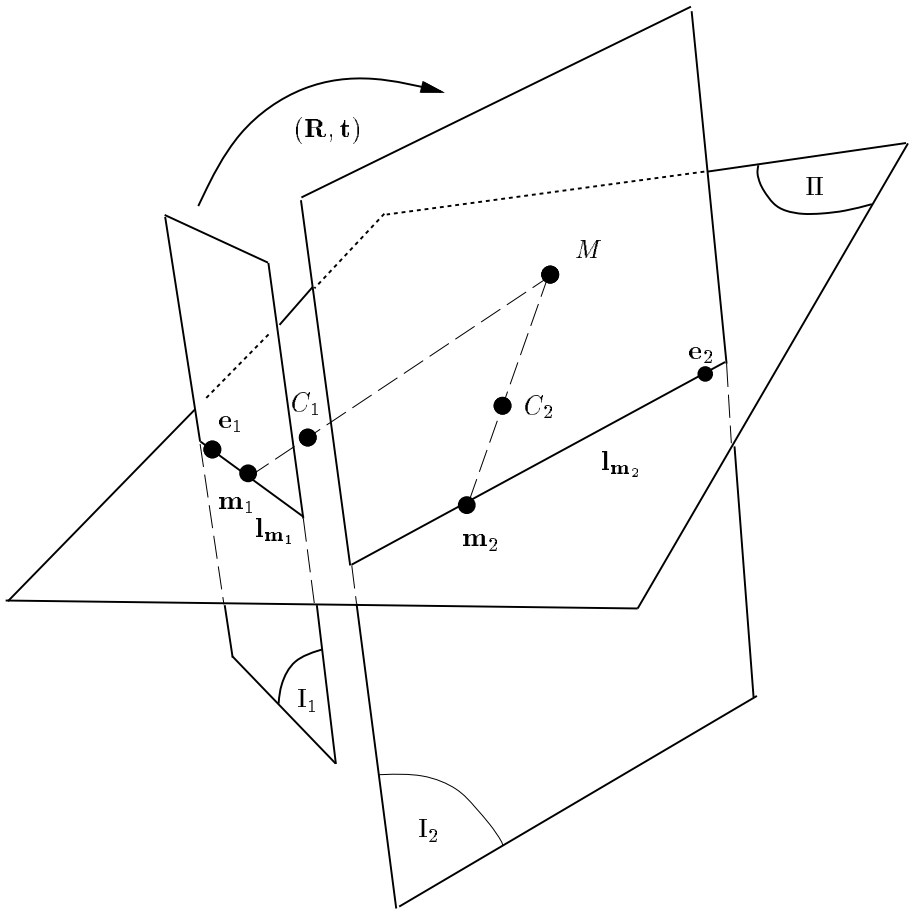

Fig. 1. The epipolar geometry

the following two equations:

$$s_1\tilde{\mathbf{m}}_1 = \mathbf{A}_1\,[\mathbf{I}\ \mathbf{0}]\begin{bmatrix}M\\1\end{bmatrix}$$
$$s_2\tilde{\mathbf{m}}_2 = \mathbf{A}_2\,[\mathbf{R}\ \mathbf{t}]\begin{bmatrix}M\\1\end{bmatrix}\ ,$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are the intrinsic matrices of the first and second cameras, respectively. Eliminating M, s_1 and s_2 from the above equations, we obtain the following fundamental equation

$$\tilde{\mathbf{m}}_2^T\mathbf{A}_2^{-T}\mathbf{T}\mathbf{R}\mathbf{A}_1^{-1}\tilde{\mathbf{m}}_1 = 0\ , \tag{2}$$

where $\mathbf{T}$ is an antisymmetric matrix defined by $\mathbf{t}$ such that $\mathbf{Tx} = \mathbf{t} \wedge \mathbf{x}$ for all 3-D vector $\mathbf{x}$ ($\wedge$ denotes the cross product).

Equation (2) is a fundamental constraint underlying any two images if they are perspective projections of one and the same scene. Let $\mathbf{F} = \mathbf{A}_2^{-T}\mathbf{T}\mathbf{R}\mathbf{A}_1^{-1}$, $\mathbf{F}$ is known as the fundamental matrix of the two images [31]. Without considering 3-D metric entities, we can think of the fundamental matrix as providing the two epipoles (i.e., the vertexes of the two pencils of epipolar lines) and the 3 parameters of the homography between these two pencils. This is the only geometric information available from two uncalibrated images [36, 31]. This implies that the fundamental matrix has only seven degrees of freedom. Indeed, it is only defined up to a scale factor and its determinant is zero. Geometrically, $\mathbf{F}\tilde{\mathbf{m}}_1$ defines the epipolar line of point $\mathbf{m}_1$ in the second image. Equation 2 says no more than that the correspondence in the right image of point $\mathbf{m}_1$ lies on the corresponding epipolar line. Transposing equation 2 yields the symmetric relation from the second image to the first image.

It can be shown that the fundamental matrix $\mathbf{F}$ is related to the essential matrix $\mathbf{E} = \mathbf{t} \times \mathbf{R}$ [29, 23] by

$$\mathbf{F} = \mathbf{A}_2^{-T}\mathbf{E}\mathbf{A}_1^{-1} \ .$$

It is thus clear that if the cameras are calibrated, the problem becomes the one of *motion and structure from motion* [29, 53, 39, 13, 1, 56, 24].

4 Finding Candidate Matches by Correlation

Before recovering the epipolar geometry, we must establish a few correspondences between images. To this end, a corner detector is first applied to each image to extract high curvature points. A classical correlation technique is then used to establish matching candidates between the two images.

4.1 Extracting Points of Interest

First, feature points corresponding to high curvature points are extracted from each image. A great deal of effort has been spent by the computer vision community on this problem, and several approaches have been reported in the literature in the last few years. They can be broadly divided into two groups: The first group consists in first extracting edges as a chain code, and then searching for points having maxima curvature [10, 2, 37] or performing a polygonal approximation on the chains and then searching for the line segment intersections [21]. The second group works directly on a grey-level image. The large number of techniques that have been proposed within this group are generally based on the measurement of the gradients and of the curvatures of the surface (see [11] for a review).

In our application, we use the corner detector [19], which is a slightly modified version of the Plessey corner detector [20, 41]. It is based on the following operator:

$$R(x, y) = \det[\hat{C}] - k \ \mathrm{trace}^2[\hat{C}] \ , \tag{3}$$

where $\hat{C}$ is the following matrix:

$$\hat{C} = \begin{bmatrix} \hat{I_x^2} & \widehat{I_x I_y} \\ \widehat{I_x I_y} & \hat{I_y^2} \end{bmatrix} , \tag{4}$$

where $\hat{I}$ denotes the smoothing operation on the grey level image $I(x, y)$. I_x and I_y indicate the x and y directional derivatives respectively.

We use a value of k equal to 0.04 to provide discrimination against high contrast pixel step edges. After that, the operator output is thresholded for the corner detection. It should be pointed out that this method allows us to recover a corner position up to pixel precision. In order to recover the corner position up to sub-pixel position, one uses the model based approach we have already developed and presented in [8], where corners are extracted directly from the image by searching the parameters of the parametric model that best approximate the observed grey level image intensities around the corner position detected. One can notice that such an approach is well adapted for scenes containing polyhedral objects, but not for most outdoor scenes.

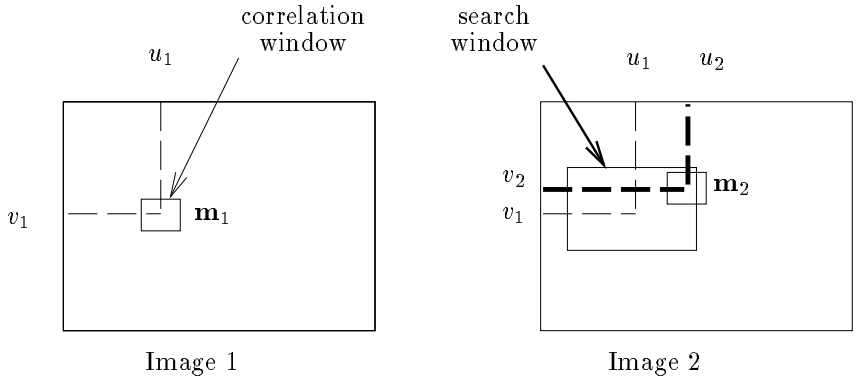

Fig. 2. Correlation

4.2 Matching Through Correlation

Given a high curvature point $\mathbf{m}_1$ in image 1, we use a correlation window of size $(2n+1) \times (2m+1)$ centered at this point. We then select a rectangular search area of size $(2d_u + 1) \times (2d_v + 1)$ around this point in the second image, and perform a correlation operation on a given window between point $\mathbf{m}_1$ in the first image and all high curvature points $\mathbf{m}_2$ lying within the search area in the second image. The search window reflects some *a priori* knowledge about the disparities between the matched points. This is equivalent to reducing the search area for a corresponding point from the whole image to a given window. The

correlation score is defined as

$$Score(\mathbf{m}_1, \mathbf{m}_2) = \frac{\sum_{i=-n}^{n} \sum_{j=-m}^{m} [I_1(u_1+i, v_1+j) - \overline{I_1(u_1, v_1)}] \times [I_2(u_2+i, v_2+j) - \overline{I_2(u_2, v_2)}]}{(2n+1)(2m+1)\sqrt{\sigma^2(I_1) \times \sigma^2(I_2)}}, \tag{5}$$

where $\overline{I_k(u,v)} = \sum_{i=-n}^{n} \sum_{j=-m}^{m} I_k(u+i, v+j)/[(2n+1)(2m+1)]$ is the average at point (u, v) of I_k $(k = 1, 2)$, and $\sigma(I_k)$ is the standard deviation of the image I_k in the neighborhood $(2n+1) \times (2m+1)$ of (u, v), which is given by:

$$\sigma(I_k) = \sqrt{\frac{\sum_{i=-n}^{n} \sum_{j=-m}^{m} I_k^2(u,v)}{(2n+1)(2m+1)} - \overline{I_k(u,v)}}\,. \tag{6}$$

The score ranges from -1, for two correlation windows which are not similar at all, to 1, for two correlation windows which are identical.

A constraint on the correlation score is then applied in order to select the most consistent matches: For a given couple of points to be considered as a candidate match, the correlation score must be higher than a given threshold. If the above constraint is fulfilled, we say that the pair of points considered is self consistent and forms a *candidate match*. For each point in the first image, we thus have a set of candidate matches from the second image (the set is possibly nil); and in the same time we have also a set of candidate matches from the first image for each point in the second image.

In our implementation, $n = m = 7$ for the correlation window, and a threshold of 0.8 on the correlation score is used. For the search window, d_u and d_v are set to a quater of the image width and height, respectively. It is thus very large (half of the whole image).

5 Disambiguating Matches Through Relaxation

Using the correlation technique described above, a point in the first image may be paired to several points in the second image (which we call *candidate matches*), and vice versa. Several techniques exist for resolving the matching ambiguities. The technique we use falls into the class of techniques known as *relaxation techniques*. The idea is to allow the candidate matches to reorganize themselves by propagating some constraints, such as continuity and uniqueness, through the neighborhood.

5.1 Measure of the Support for a Candidate Match

Consider a candidate match $(\mathbf{m}_{1i}, \mathbf{m}_{2j})$ where $\mathbf{m}_{1i}$ is a point in the first image and $\mathbf{m}_{2j}$ is a point in the second image. Let $\mathcal{N}(\mathbf{m}_{1i})$ and $\mathcal{N}(\mathbf{m}_{2j})$ be, respectively, the neighbors of $\mathbf{m}_{1i}$ and $\mathbf{m}_{2j}$ within a disc of radius R. If $(\mathbf{m}_{1i}, \mathbf{m}_{2j})$ is a good match, we will expect to see many matches $(\mathbf{n}_{1k}, \mathbf{n}_{2l})$, where $\mathbf{n}_{1k} \in \mathcal{N}(\mathbf{m}_{1i})$ and $\mathbf{n}_{2l} \in \mathcal{N}(\mathbf{m}_{2j})$, such that the position of $\mathbf{n}_{1k}$

relative to $\mathbf{m}_{1i}$ is similar to that of $\mathbf{n}_{2l}$ relative to $\mathbf{m}_{2j}$. On the other hand, if $(\mathbf{m}_{1i}, \mathbf{m}_{2j})$ is a bad match, we will expect to see only few matches, or even not any at all, in their neighborhood.

More formally, we define a measure of support for a match, which we call the *strength of the match* (SM for abbreviation), as

$$S_M(\mathbf{m}_{1i}, \mathbf{m}_{2j}) = c_{ij} \sum_{\mathbf{n}_{1k} \in \mathcal{N}(\mathbf{m}_{1i})} \left[\max_{\mathbf{n}_{2l} \in \mathcal{N}(\mathbf{m}_{2j})} \frac{c_{kl}\, \delta(\mathbf{m}_{1i}, \mathbf{m}_{2j};\ \mathbf{n}_{1k}, \mathbf{n}_{2l})}{1 + dist(\mathbf{m}_{1i}, \mathbf{m}_{2j};\ \mathbf{n}_{1k}, \mathbf{n}_{2l})} \right] ,$$

where c_{ij} and c_{kl} are the goodness of the candidate matches $(\mathbf{m}_{1i}, \mathbf{m}_{2j})$ and $(\mathbf{n}_{1k}, \mathbf{n}_{2l})$, which can be the correlation scores given in the last section, $dist(\mathbf{m}_{1i}, \mathbf{m}_{2j};\ \mathbf{n}_{1k}, \mathbf{n}_{2l})$ is the average distance of the two pairings, i.e.,

$$dist(\mathbf{m}_{1i}, \mathbf{m}_{2j};\ \mathbf{n}_{1k}, \mathbf{n}_{2l}) = [d(\mathbf{m}_{1i}, \mathbf{n}_{1k}) + d(\mathbf{m}_{2j}, \mathbf{n}_{2l})]/2$$

with $d(\mathbf{m}, \mathbf{n}) = ||\mathbf{m} - \mathbf{n}||$, the Euclidean distance between $\mathbf{m}$ and $\mathbf{n}$, and

$$\delta(\mathbf{m}_{1i}, \mathbf{m}_{2j};\ \mathbf{n}_{1k}, \mathbf{n}_{2l}) = \begin{cases} e^{-r/\varepsilon_r} & \text{if } (\mathbf{n}_{1k}, \mathbf{n}_{2l}) \text{ is a candidate match and } r < \varepsilon_r \\ 0 & \text{otherwise} \end{cases}$$

where r is the relative distance difference given by

$$r = \frac{|d(\mathbf{m}_{1i}, \mathbf{n}_{1k}) - d(\mathbf{m}_{2j}, \mathbf{n}_{2l})|}{dist(\mathbf{m}_{1i}, \mathbf{m}_{2j};\ \mathbf{n}_{1k}, \mathbf{n}_{2l})}$$

and ε_r is a threshold on the relative distance difference. The above definition of the strength of a match is similar in the form to that used in the PMF stereo algorithm [44].

Several remarks can be made regarding our measure of matching support.

- Firstly, the strength of a match actually counts the number of candidate matches found in the neighborhoods, but only those whose positions relative to the considered match are similar are counted.
- Secondly, the test of similarity in relative positions is based on the relative distance (the value of r). Indeed, the similarity in relative positions is justified by the hypothesis that an affine transformation can approximate the change between the neighborhoods of the candidate match being considered. This assumption is reasonable only for a small neighborhood. Thus we should allow larger tolerance in distance differences for distant points, and this is exactly what our criterion does.
- Thirdly, the contribution of a candidate match $(\mathbf{n}_{1k}, \mathbf{n}_{2l})$ to the strength of the match $(\mathbf{m}_{1i}, \mathbf{m}_{2j})$ is the exponential of the negative relative error r, which is strictly monotonically decreasing function of r. When r is very big, then $\exp(-r/\varepsilon_r) \rightarrow 0$, and the candidate match can be ignored. When $r \rightarrow 0$, i.e., the difference is very small, then $\exp(-r/\varepsilon_r) \rightarrow 1$, and the candidate will largely contribute to the match $(\mathbf{m}_{1i}, \mathbf{m}_{2j})$.