Joint Estimation of Correspondence and Motion Using Global Rigidity and Local Descriptors

Ricardo S. Cabral
Institute for Systems and Robotics / IST, Lisboa, Portugal
r.silveira.cabral@ist.utl.pt

Abstract

In this paper, we impose a rigidity constraint to the motion between points and correspondence candidates obtained by extracting local descriptors from an image pair. In doing so, we are able to couple the estimation of correspondences and motion within a single optimization problem, as a minimization of a cost function over a search grid in two parameters, therefore bypassing the combinatorial explosion inherent to the correspondence problem. The resulting algorithm has polynomial complexity and optimal properties when the camera parameters are known.

*We demonstrate through a series of synthetic and real data experiments the robustness of our method to outliers and occlusion. We show the versatility in our algorithm's input, by coupling it with discriminative feature selection algorithms (*e.g.*, SIFT).*

1. Introduction

Many problems in computer vision are solved by using redundancy. Whether obtained from multiple cameras or from a single camera over time and in different positions, most of the solutions available require the position of a set of points to be known in various images. The task of finding these *trajectories* along the images, known as the *correspondence problem*, is often relegated to a second role due to its intrinsic combinatorial nature: to each point in the first image corresponds a high cardinality set of respective candidates in subsequent images (Fig. 1(a)).

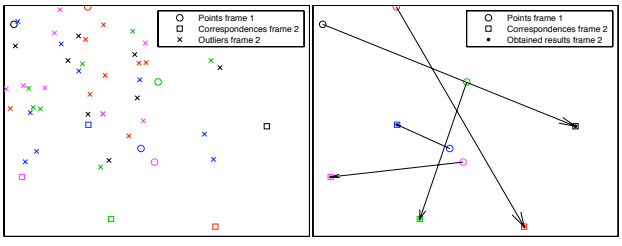
Recently, methods [4, 6] were discovered that extract local descriptors in images which not only possess interesting properties (*e.g.*, invariance to scale and rotation) but are highly discriminative, easing their correspondence in a pair of images by formulating it as a nearest neighbor problem. The success of these techniques within the frameworks in which they are currently used motivates its use in more general settings, as is the case of 3D motion.

In this paper, we impose a rigidity constraint to the motion between points and correspondence candidates obtained by extracting local descriptors from an image pair. In doing so, we are able to couple the estimation of correspondences and motion within a single optimization problem, as a minimization of a cost function over a search grid in two parameters, therefore bypassing the combinatorial explosion inherent to this problem. The resulting algorithm accurately estimates motion and correspondence with polynomial complexity, finding an optimal solution when the camera parameters are known.

The approach presented takes as input sets of point coordinates, specifically a set of points in the first image, each with a set of correspondence candidates in the second. With this data, we estimate motion between both cameras and select a number of most likely candidates (Fig. 1(b)) for each point in the first image (which, in particular, can be the case of 1-to-1 correspondence).

Besides rigidity, we assume an orthographic projection model is valid. Even though these assumptions are not restrictive, our method is able to cope objects exhibiting small perspective effects.

Although the deduction hereby presented is particularized for the case of candidate matches detected as having the same intensity (which assumes a Lambertian object), this method can easily be generalized to work with any feature extraction technique (we show examples of this in the experimental section). Specifically, it enables us to merge



(a) input data (b) our method's results

Figure 1. Points in a pair of images. The colors represent points with equal intensity value.

cues from brightness and color without being prone to the aperture problem.

The remainder of this paper is organized as follows: Sec. 2 formulates our goal within an optimization framework, after preliminary notions necessary to understand the problem. Sec. 3 describes the steps taken in order to derive the solution to the problem described in Sec. 2. Sec. 4 describes the experiments performed to assess the performance of the algorithms. Finally, Sec. 5 presents closing comments.

2. Problem formulation

We start by considering an image as an application

$$I: \qquad \mathbb{R}^2 \mapsto \mathbb{R}^3$$

that maps pixel coordinates into a certain intensity value. We distinguish the frame f from an image sequence of the same object by using the notation I_f, where f corresponds to the frame index in the sequence. We also refer to the i-th feature point present in the frame f by the coordinate vector $\mathbf{u}_i$ and to its intensity on frame f by $I_f(\mathbf{u}_i)$.

2.1. Motion

Provided points have the same intensity value in different frames, *i.e.*, the object is Lambertian, the motion of the i-th pixel between two frames I_1 and I_2 can be defined as a space shift of the image function along both of its axis as

$$I_1(x_i, y_i) = I_2(x_i + \Delta x_i, y_i + \Delta y_i),$$

in which coefficients Δx_i and Δy_i respectively account for the i-th pixel movement along the x and y axis.

2.2. Cameras

We model cameras as projection operators that transform points in 3D space onto an image. We consider in this paper that images are obtained according to an *orthographic* camera model, valid when the variation of the object depths is very small when compared to the distance between the camera and the object. In this model, motions are obtained by the composition of a matrix $\mathbf{R} \in \mathbb{R}^{2\times 3}$ and a translation vector $\mathbf{t}$ about the object coordinate system, comprising a total of 5 degrees of freedom, as

$$\mathbf{H} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}. \tag{1}$$

With $\mathbf{R}$ being an Euclidean coordinate change, it follows that its rows are orthogonal and have unit norm. Therefore, all possible matrices $\mathbf{R}$ constitute a Stiefel Manifold $\mathsf{O}(2,3)$ [2], the set of $\mathbb{R}^{2\times 3}$ matrices whose rows are orthonormal

$$\mathsf{O}(2,3) = \{\mathbf{R} \in \mathbb{R}^{2\times 3} : \mathbf{R}\mathbf{R}^\top = \mathbf{I}_2\}, \tag{2}$$

where $\mathbf{I}_2 \in \mathbb{R}^{2\times 2}$ is the identity matrix. Since the singular values of $\mathbf{R}$ are the square roots of the eigenvalues of $\mathbf{R}\mathbf{R}^\top$, each of its singular values is equal to 1, see for instance [3].

2.3. Structure from Motion (SfM)

Using the camera definition on (1), we write the position of N feature points along F image frames as the product of a matrix $\mathbf{M} \in \mathbb{R}^{2F\times 4}$ (obtained by stacking all camera matrices $\mathbf{H}$ responsible for each frame) by a matrix $\mathbf{S}^\top \in \mathbb{R}^{4\times N}$, containing the 3D coordinates of the N points tracked in the object coordinate system, as

$$\mathbf{W} = \mathbf{M}\mathbf{S}^\top. \tag{3}$$

We consider the case of two sets $\mathbf{W}_1$ and $\mathbf{W}_2$, containing the coordinates of N points on a pair of images ($F = 2$). Assuming the first camera coordinate system is aligned with the object coordinate system and perpendicular to the z axis [1], we can write $\mathbf{M}$ as

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{R} & \mathbf{t} \end{bmatrix},$$

where $\mathbf{R} \in \mathbb{R}^{2\times 3}$ models the orientation of camera 2 relative to the object coordinate system and $\mathbf{t} \in \mathbb{R}^2$ describes the translation between the camera's and object's points of origin.

By splitting $\mathbf{R}$ in the matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ and vector $\mathbf{b} \in \mathbb{R}^2$ as

$$\mathbf{R} = \begin{bmatrix} \mathbf{A} & \mathbf{b} \end{bmatrix}, \tag{4}$$

the transformation between images $\mathbf{W}_1$ and $\mathbf{W}_2$ can be interpreted as

$$\mathbf{W}_2 = \mathbf{A}\mathbf{W}_1 + \mathbf{b}\mathbf{z}^\top + \mathbf{t}\mathbf{1}^\top, \tag{5}$$

a three step transformation (Fig. 2) where $\mathbf{1}^\top \in \mathbb{R}^{1\times N}$ is a vector of ones.

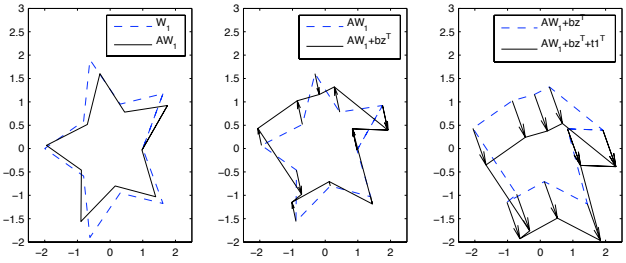


Figure 2. Image motion decomposed in three sub-transformations: the one applied by the operator $\mathbf{A}$ (left), translation along the vector $\mathbf{b}$ according to each point depth z_i (center) and translation $\mathbf{t}$ applied commonly to all points (right).

According to this formulation, in the SfM with correspondences case, one wishes to find, given a set of point trajectories along two frames, the unknowns $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{t}$ that define the camera motion, as well as the unknown object depths $\mathbf{z}$, such that (5) holds.

2.4. SfM without correspondences as an optimization problem

However, equations (3) and (5) are only valid for the case where correspondences are known. The problem we formulate, on the other hand, makes no such assumption, and is thus plagued with a large number of outliers in the form of candidate matches. This fact alone is what constitutes the main challenge of this problem, due to the **combinatoric explosion** arising from the existence of several possible matches.

We are, however, able to surpass this *curse of dimensionality* and **achieve an accurate solution with polynomial complexity**, by making use of the fact that the intensity is independent from the object's point of view (we assume a Lambertian object). This way, point correspondences between images must display the same intensity value. For the aforementioned case of two image frames $\mathbf{W}_1$ and $\mathbf{W}_2$, this yields

$$I_2(\mathbf{A}\mathbf{W}_1 + \mathbf{b}\mathbf{z}^\top + \mathbf{t}\mathbf{1}^\top) = I_1(\mathbf{W}_1),$$

a formulation that restricts the possible destinations each of the $i \in \{1, \dots, N\}$ points present in the image can take by unifying rigidity and intensity constraints. We are now in conditions to establish our problem statement.

Problem Statement Given $I_2, I_1, \mathbf{W}_1 = [\mathbf{u}_1 \ \dots \ \mathbf{u}_N]$, find $\mathbf{A}, \mathbf{b}, \mathbf{t}, \mathbf{z}$ such that

$$I_2(\mathbf{A}\mathbf{u}_i + \mathbf{b}z_i + \mathbf{t}) \in \mathcal{S}_i, \quad \forall i \in \{1, \dots, N\},$$

where $\mathcal{S}_i$ represents the level curve of I_2 with value given by the intensity in image I_1 at point $\mathbf{u}_i$. In practice, we have small variations due to illumination and shading, so we consider all points whose intensities are in the small polyhedron centered in $I_1(\mathbf{u}_i)$ and defined by the ℓ_1 norm, as

$$\mathcal{S}_i = \{\mathbf{v} \in \mathbb{R}^2 : ||I_1(\mathbf{u}_i) - I_2(\mathbf{v})||_1 \leq \xi\}. \tag{6}$$

This formulation suggests the optimization problem of finding correspondences such that the distance d of each point $\mathbf{A}\mathbf{u}_i + \mathbf{b}z_i + \mathbf{t}$ to its respective intensity level set $\mathcal{S}_i$ is minimized, or

$$\begin{aligned} &\text{minimize} && \Phi(\mathbf{A}, \mathbf{b}) \\ &\text{subject to} && \mathbf{A}\mathbf{A}^\top + \mathbf{b}\mathbf{b}^\top = \mathbf{I}_2, \end{aligned} \tag{7}$$

where

$$\Phi(\mathbf{A}, \mathbf{b}) = \inf_{\mathbf{t}\in\mathbb{R}^2, \mathbf{z}\in\mathbb{R}^N} \sum_{i=1}^{N} d(\mathbf{A}\mathbf{u}_i + \mathbf{b}z_i + \mathbf{t}, \mathcal{S}_i), \tag{8}$$

the set distance $d(\cdot, \cdot)$ is defined using the Euclidean norm as

$$d(\mathbf{x}_i, \mathcal{S}_i) = \min_{\mathbf{v}\in\mathcal{S}_i} ||\mathbf{x}_i - \mathbf{v}||_2$$

and the restriction $\mathbf{A}\mathbf{A}^\top + \mathbf{b}\mathbf{b}^\top = \mathbf{I}_2$ follows from (2) and (4).

In words, we wish to find the transformation of type (5) that best repositions points $\mathbf{u}_i$ in the first image with their respective level sets $\mathcal{S}_i$ in the second image. By doing this, we are not only able to assess the goodness of a camera model given image intensity data, but obtain from the level sets $\mathcal{S}_i$ the correspondence candidates that best fit that model.

3. Correspondence Estimation using Rigidity and local Descriptors (CERD)

The question naturally arising at this point is how the value of the cost function $\Phi(\mathbf{A}, \mathbf{b})$, itself an optimization problem, is calculated, given a known pair $(\mathbf{A}, \mathbf{b})$.

Let us start by decomposing $\mathbf{t}$ in two components

$$\mathbf{t} = \alpha\mathbf{b} + \gamma\mathbf{b}_\perp, \qquad ||\mathbf{b}_\perp||_2 = 1,$$

where $\mathbf{b}_\perp$ stands for the orthogonal complement of $\mathbf{b}$. In doing so, we implicitly assume that $\mathbf{b} \neq \mathbf{0}$, a condition equivalent to not having a pure planar motion, as the influence of object depths in motion is confined within subspace. Provided this condition is met, we are able to rewrite (8) as

$$\Phi(\mathbf{A}, \mathbf{b}) = \inf_{\substack{\alpha\in\mathbb{R}, \\ \gamma\in\mathbb{R}, \\ \mathbf{z}\in\mathbb{R}^N}} \sum_{i=1}^{N} d(\mathbf{A}\mathbf{u}_i + \mathbf{b}(z_i + \alpha) + \gamma\mathbf{b}_\perp, \mathcal{S}_i). \tag{9}$$

There is an intrinsic ambiguity to finding the value of the factors $z_i + \alpha$ from its sum alone. Thus, we collapse these two variables in z_i (*i.e.*, we assume $\alpha = 0$ for simplicity), rendering (9) as the optimization problem

$$\Phi(\mathbf{A}, \mathbf{b}) = \inf_{\gamma\in\mathbb{R}, \mathbf{z}\in\mathbb{R}^N} \sum_{i=1}^{N} d(\mathbf{A}\mathbf{u}_i + \mathbf{b}z_i + \gamma\mathbf{b}_\perp, \mathcal{S}_i),$$

which we further simplify by using the fact that the infimum over two variables can be decoupled as

$$\Phi(\mathbf{A}, \mathbf{b}) = \inf_{\gamma\in\mathbb{R}} \left[\inf_{\mathbf{z}\in\mathbb{R}^N} \sum_{i=1}^{N} d(\mathbf{A}\mathbf{u}_i + \mathbf{b}z_i + \gamma\mathbf{b}_\perp, \mathcal{S}_i) \right]. \tag{10}$$

As can be seen in Fig. 3, the effect of not knowing the object depths $\mathbf{z_i}$ is that every point in I_2 belonging to the line with direction given by the vector $\mathbf{b}$ that intersects the point $\mathbf{A}\mathbf{u}_i + \mathbf{t}$ is considered valid by the motion model. All points in this line intersecting the level curve $\mathcal{S}_i$ are, therefore, considered correct correspondences. Having said this, our problem becomes minimizing the distance of the level curve $\mathcal{S}_i$ to the line passing through the point $\mathbf{A}\mathbf{u}_i + \gamma\mathbf{b}_\perp$ with direction $\mathbf{b}$, or

$$\Phi(\mathbf{A}, \mathbf{b}) = \inf_{\gamma\in\mathbb{R}} \sum_{i=1}^{N} d\left((\mathbf{b}_\perp)^\top \mathbf{A}\mathbf{u}_i + \gamma, (\mathbf{b}_\perp)^\top \mathcal{S}_i\right). \tag{11}$$

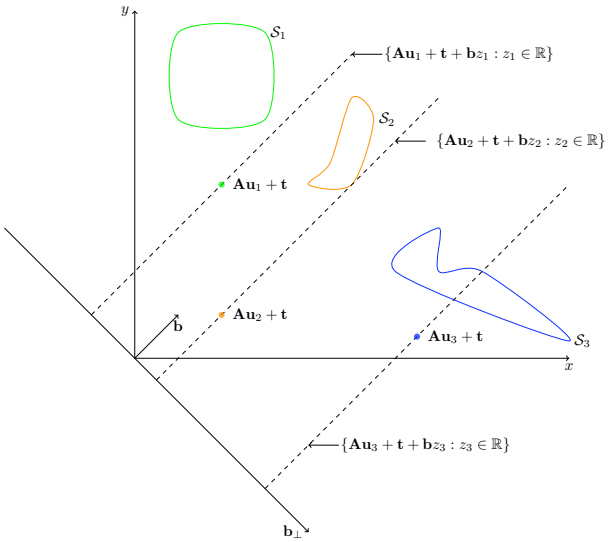


Figure 3. Ambiguity of correspondence for known $\mathbf{H}$ and unknown object depths $\mathbf{z}$. The dashed lines represent the variation of object depth $\mathbf{z}$, making each intersection of the lines passing along $\mathbf{A}\mathbf{u}_i$ with direction given by the vector $\mathbf{b}$ and level curves $\mathcal{S}_i$ a valid correspondence.

This new measure corresponds to projecting every point in $\mathcal{S}_i$ and $\mathbf{A}\mathbf{u}_i$ onto the subspace defined by $\mathbf{b}_\perp$. In this subspace, the entire line passing through the point $\mathbf{A}\mathbf{u}_i$ with direction $\mathbf{b}$ collapses in a single point, allowing a representation that is independent of the object depth z_i.

In practice, since the image has a finite resolution, and therefore a finite number of points, the level sets $\mathcal{S}_i$ degenerate into discrete sets

$$\mathcal{S}_i = \{\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,N_i}\},$$

where N_i represents the number of points in the second image in the level set $\mathcal{S}_i$. Hence, the set distance operator $d\left((\mathbf{b}_\perp)^\top \mathbf{A}\mathbf{u}_i + \gamma, \mathbf{b}_\perp{}^\top \mathcal{S}_i\right)$ is naturally defined as

$$\min_{k \in \{1,\cdots,N_i\}} \left|(\mathbf{b}_\perp)^\top \mathbf{A}\mathbf{u}_i + t - (\mathbf{b}_\perp)^\top \mathbf{v}_{i,k}\right|, \qquad (12)$$

the minimum distance between the point $(\mathbf{b}_\perp)^\top \mathbf{A}\mathbf{u}_i + \gamma$ and every element in the set $(\mathbf{b}_\perp)^\top \mathcal{S}_i$

Since $\mathbf{R} \in \mathsf{O}(2,3)$, we are able to fully parameterize its components $\mathbf{A}$ and $\mathbf{b}$ as

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & r \end{bmatrix} \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}, \qquad (13)$$

$$\mathbf{b} = \pm \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \sqrt{1-r^2}, \qquad \mathbf{b}_\perp = \pm \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}. \qquad (14)$$

For the purposes of this discussion, we restrict the pair (θ, ϕ) to the square $[0, 2\pi[\times[-\frac{\pi}{2}, \frac{\pi}{2}]$, the minimum interval necessary to span all possible matrices.

Using the definition of $\mathbf{A}$ and the expression for $\mathbf{b}_\perp$ in (13), the product $(\mathbf{b}_\perp)^\top \mathbf{A}$ in (12) is simplified, yielding the set distance $d\left((\mathbf{b}_\perp)^\top \mathbf{A}\mathbf{u}_i + \gamma, (\mathbf{b}_\perp)^\top \mathcal{S}_i\right)$ as

$$\min_{k \in \{1,\cdots,N_i\}} \left|\mathbf{u}_i^\top \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix} + \gamma - \mathbf{v}_{i,k}^\top \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}\right| \qquad (15)$$

where we have factored out the sign indetermination in $\mathbf{b}_\perp$.

Note that despite $\mathbf{R}$ being defined univocally by the parameters (r, θ, ϕ), (15) only depends on the last two and translation component γ. Besides that, the orthonormality constraint in (7) is embedded within this equation. This allows us to reformulate the optimization problem in (7) as an unconstrained minimization of (11) with the set distance $d(\cdot)$ defined as in (15).

3.1. Solving for translation

Since now we have defined how to calculate the value of $\Phi(\mathbf{A}, \mathbf{b})$ given $\mathbf{A}, \mathbf{b}$, let us assume that the camera parameters are available. Since the parameters θ and ϕ are known, the problem is reduced to

$$\inf_{\gamma \in \mathbb{R}} C(\gamma) \qquad (16)$$

where the cost function $C(\gamma)$ is

$$C(\gamma) = \sum_{i=1}^{N} d\left((\mathbf{b}_\perp)^\top \mathbf{A}\mathbf{u}_i + \gamma, (\mathbf{b}_\perp)^\top \mathcal{S}_i\right) \qquad (17)$$

and $d(.)$ is defined as in (15).

According to this formulation, we wish to obtain from the $i \in \{1, \ldots, N\}$ points in the first image and respective sets of candidates $\mathcal{S}_i$ (each with N_i prospective matches) in the second image, a candidate $\mathbf{v}_{i,k} \in \mathcal{S}_i$ for each point that corroborates best the motion given by the camera parameters $\mathbf{A}, \mathbf{b}$ and the translation component γ.

The solution we propose relies on the key observation that, in the absence of noise, γ is common to every correct correspondence, *i.e.*, it is given by the difference between the point in the first image and it's correct match in the second image, when projected onto the subspace defined by $\mathbf{b}_\perp$. As such, we restrict γ to forced correspondences between points in the first image and every possible candidates in the second, as

$$\gamma_{m,n} = \mathbf{v}_{m,n}^\top \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} - \mathbf{u}_m^\top \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix}, \qquad (18)$$

where $\mathbf{u}_m$ stands for the coordinates of the m-th point ($m \in \{1, \ldots, N\}$) in image $\mathbf{W}_1$ and $\mathbf{v}_{m,n}$ for the position of the n-th correspondence candidate for point n in the level set $\mathcal{S}_m$.

This results in the confinement of (16) to a search in the finite set of possible correspondences

$$\Phi(\mathbf{A}, \mathbf{b}) = \min_{(m,n) \in \{1,\cdots,N\} \times \{1,\cdots,N_m\}} C(\gamma_{m,n}) \qquad (19)$$

Note that this problem is equivalent to (16) in the absence of noise, since the set of forced matches includes the correct solution.

For the problem presented in (19), we present the following optimal solution: For each of the possible forced correspondences $\gamma_{i,j}$, we select from the level sets $\mathcal{S}_i$ the best match for each of remaining $N-1$ points in image 1 as the candidate with the least distance (*i.e.*, the nearest neighbor) according to the set distance $d(\cdot)$. We could have restricted the search to one point in the first image and its respective level set, but we chose to in order to account for motions that might occlude some of the points in the image.

In doing this, we have reduced a problem with a number of possible assignments that is combinatoric $\binom{N \times N_i}{N_i}$ to polynomial, with an algorithm that has a number of operations that approximates $\mathcal{O}\left(N^2 N_i^2\right)$.

In the case where the camera parameters are known and the candidates available in the second image differ from $\mathbf{W}_2$ only by a permutation matrix, such as in [5, 7], the optimal solution is given by attributing correspondences between each point and candidates according to the natural order of the sets. This can be achieved by sorting both constellations, a process with complexity of $\mathcal{O}\left(N \log N\right)$, where N is the number of points considered in each image.

It should be noted that if the transformations that obtain constellations 1 and 2 from $\mathbf{W}_1$ and $\mathbf{W}_2$ are not injective, two points will have the same projection on the $\mathbf{b}_\perp$ axis; in this case, the set order makes no distinction between the two, therefore making it impossible to distinguish between the possible correspondence solutions.

3.2. Solving for rotation

The cost function introduced in (17) assesses the extent to which correspondences can be explained by a rigid, orthographic motion, allowing us to select the best candidates for a motion given by a specific pair (θ, ϕ). Hence, the solution mentioned in (19) for γ can be subsumed under a broader algorithm that performs a discretized search on the parameter grid (θ, ϕ) and calculates for each tuple the candidates that best agree with the given motion. This method, which we call CERD, is summarized in Algorithm 1.

Since we perform a search throughout $\mathsf{O}(2,3)$, this method finds the global minimum of (19) up to a specified grid resolution, giving as a byproduct the motion model and point correspondences that generate it.

It should be noted that this algorithm can be categorized as having a polynomial complexity of $\mathcal{O}\left(N_p N^2 N_c^2\right)$, where N_p stands for the number of points the grid is discretized in.

4. Experiments

In this section, we perform experiments to assess the performance of CERD. We start by evaluating the algorithms with synthetic data, both for the case of known (Sec 4.1) and unknown (Sec. 4.2) camera parameters. Sec. 4.3 extends the results obtained in the previous sections using real data. Finally, in Sec. 4.4 we couple our method with existing state of the art feature extracting algorithms, such as SIFT.

For the synthetic experiments, we generate data as follows: we obtain a pair of random images, as follows: we generate parameters $\mathbf{t}$, r, θ, ϕ and the shape matrix $\mathbf{S}$ according to a uniform distribution, the latter bound to the interval $[0, w_s]$. We then specify points in the first image (the set $\mathbf{W}_1$) as the first and second rows of $\mathbf{S}^\top$ and obtain $\mathbf{W}_2$ according to the model in (5). We obtain each of the level sets $\mathcal{S}_i$ as the union of each point in $\mathbf{W}_2$ with an additional N_c candidates, also generated from a uniform distribution with window size w_s. Additionally, if occlusion is to be tested, we remove from a number $0 \leq N_o \leq N$ of randomly selected level sets the original match in $\mathbf{W}_2$.

4.1. Known θ, ϕ

In this section, we present experiences assessing the behavior of the minimization proposed in (19), for a given pair θ, ϕ, as described in Sec. 3.1.

Constellations on both image and on the projected axis. In this experiment, we generated a pair of images with $N = 5$ points in the first frame and N level sets in the second, each with $N_c = 10$ candidates.

We feed the algorithm with points $\mathbf{W}_1$ and level sets $\mathcal{S}_i$, as well as the camera parameters θ and ϕ, leaving the correct matching between points in both images to be obtained. Fig. 4 shows the variation of the cost function $C(\gamma_{m,n})$ for each of the forced correspondences, with minima occurring when each of the N points is matched with its correct pair.

The value of $C(\gamma)$ in the global minima shows, despite the considerable point distances between images 1 and 2

Algorithm 1 CERD — Correspondence Estimation using Rigidity and local Descriptors.

```
Initialize Global Cost C = ∞
for all −π/2 ≤ φ ≤ π/2 do
  for all 0 ≤ θ < 2π do
    Calculate Φ(A, b) as in (19)
    if Φ(A, b) better than C then
      Update Global Cost C = Φ(A, b)
    end if
  end for
end for
```

(Fig. 1), that inliers in both constellations fully overlap after being projected in the subspace defined by $\mathbf{b}_\perp$ and subtracting the translation component $\gamma_{m,n}$ yielding the minima. These results demonstrate the usability of this subspace not only for its dimensionality reduction but for its independence from unknown variables r and $\mathbf{z}$. **In the absence of noise, our method is able to find the optimal solution while bypassing the combinatoric nature of the problem and coping with a high number of outliers**.

Sensitivity to noise. In practice, images are subject to errors inherent not only to sensor noise but to a discretization in pixels. To illustrate the robustness to noise of our method, we start by introducing the concept of average point disparity, which we define as the average of the ℓ_1 norm of the distance between ground truth (with noise) and the algorithm's output, for each of the N correspondences, or

$$\frac{1}{N \times w_s} \sum_{i=1}^{N} ||\hat{\mathbf{v}}_i - \mathbf{v}_i||_1$$

where $\hat{\mathbf{v}}_i, \mathbf{v}_i$ stand respectively for each correct (ground truth) and calculated match for point $\mathbf{u}_i$ and w_s is the normalization constant that factors in the window size bounding generated data (for simplicity sake, however we select $w_s = 1$). Note that by making this a distance measure rather than a binary (match, mismatch) decision, the average point disparity takes into account the case where the points obtained are not the original candidate but a neighboring point.

We generate data with $N = 5$ points and $N_c = 10$ candidate matches. We apply to the each of the point inliers in the second constellation AWGN with standard deviation σ as a fraction of the window size w_s and perform the minimization in (19), measuring the variation of cost $C(\gamma_{m,n})$ (Fig. 5) and average point disparity (Fig. 6) with the standard deviation of noise σ.

The results present in Fig. 6 show that the method's ability to make correct matches decreases with the presence of noise, being able to estimate correspondences within a 5% disparity bound while subject to noise levels around $\sigma = 1 \times 10^{-3}$. Translating to a real case scenario, our method is able to withstand noise levels corresponding to having a measurement error greater than 6 pixels in approximately $32\ \%$ of all points registered in a 4000×4000 pixels (a 16 Megapixel image), a value by far greater than what is obtained with today's cameras.

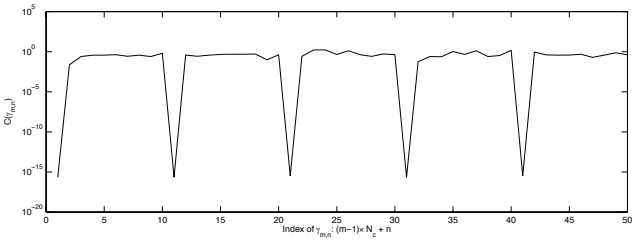


Figure 4. Behavior of $C(\gamma_{m,n})$ for $N = 5$ and $N_c = 10$.

On the other hand, the value of $C(\gamma_{m,n})$ increases with σ, allowing us to conclude that an increase in noise levels incurs in a fading of the separation between minima and the remaining values for the cost function, making the matching process more strenuous. In fact, Fig. 7, which shows the variation of the cost function $C(\gamma_{m,n})$ for $\sigma = 1 \times 10^{-3}$, illustrates this trend; while still showing 5 points with minimum cost, the value at these points is several orders of magnitude higher, when compared with the noiseless case presented in Fig. 4.

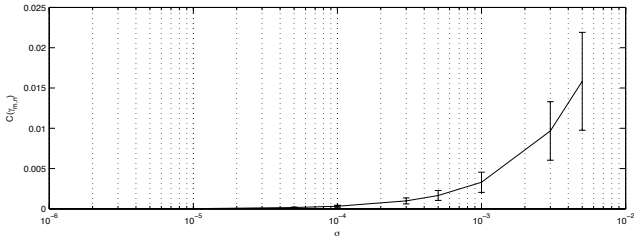


Figure 5. Variation of $C(\gamma_{m,n})$ with AWGN of standard deviation σ averaged over 6000 tests.

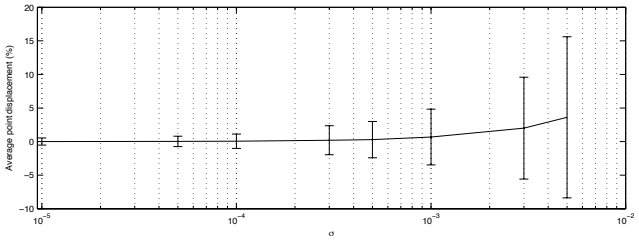


Figure 6. Variation of average point disparity with AWGN of standard deviation σ averaged over 6000 tests.

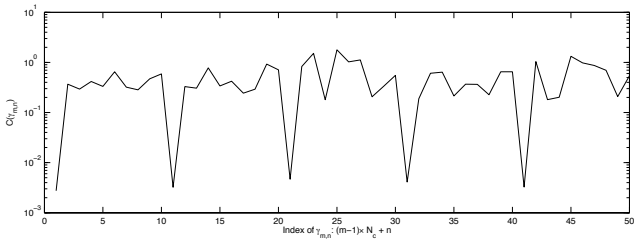
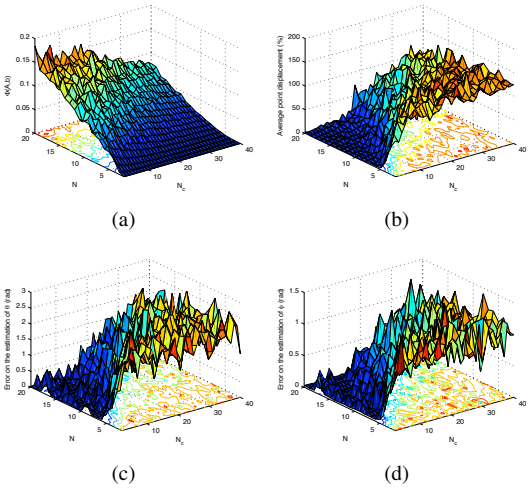


Figure 7. Variation of $C(\gamma_{m,n})$ for known camera parameters for $N = 5$ and $N_c = 10$ for noise with standard deviation of 1×10^{-3}.

4.2. Fully automatic SfM

Building up on the experiments of Sec. 4.1, we now test our algorithm as an automatic SfM recovery system. For



(a) (b) (c) (d)

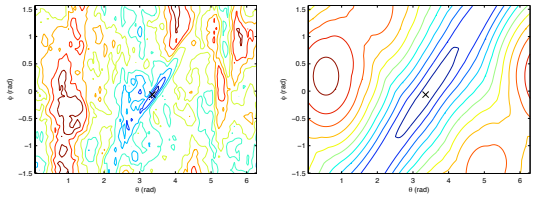
Figure 8. Variation of cost function $\Phi(\mathbf{A}, \mathbf{b})$, average point displacement and camera parameters θ and ϕ estimation errors with the number N of points and number N_c of candidate matches for a grid resolution of 2500 points: (a) variation of $\Phi(\mathbf{A}, \mathbf{b})$; (b) variation of average point disparity; (c) variation of estimation error for θ; and, (d) variation of estimation error for ϕ.

this purpose, we only feed the algorithm the inputs $\mathbf{W}_1$ and level sets $\mathcal{S}_i$ generated as before, leaving the camera parameters to be estimated.

Variation with number of points and candidates. In this experiment, we assess how the behavior of the cost function $\Phi(\mathbf{A}, \mathbf{b})$ and the average point disparity change when the algorithm is subject to a variation on the number of points N and the cardinality N_c of the level sets $\mathcal{S}_i$. Results are present in Fig. 8.

Several conclusions can be drawn from this experiment:

- The cost function $\Phi(\mathbf{A}, \mathbf{b})$ values are small, indicating a solution that respects both constraints is found for all cases;
- Although $\Phi(\mathbf{A}, \mathbf{b})$ values grow with the number of points N, this phenomena is due to the fact that the cost function is made of a larger set of residuals. This is a consequence of grid discretization, and should therefore lose its importance as resolution is increased;
- Having more points in the first frame adds information on the shape and possibly confines motion estimation, allowing for better estimation of correspondence and camera parameters θ and ϕ;
- On the other hand, an increase in the number of candidates N_c allows for a lower value of $\Phi(\mathbf{A}, \mathbf{b})$, since with more outliers, there are more possibilities of rearranging data into explanations other than the one given by ground truth, as can be inferred from the fact that the considerable average point disparity;
- The fact that the estimation of motion parameters θ and ϕ worsens as a the number of candidates N_c increases can be attributed to the same reasons.



(a) (b)

Figure 9. Variation of cost function $\Phi(\mathbf{A}, \mathbf{b})$ with camera parameters θ and ϕ for $N = 5$ points and $N_c = 10$ candidates for a grid resolution of 2500 points (the cross in the contours represents parameters θ, ϕ used to generate data): (a) contour obtained with multiple candidates; and, (b) contour obtained with ground truth.

These results allow us to define a region for this grid resolution of $N_c \leq N$ candidates to N points where the number of outliers is such that they are not easily coupled into an alternative valid motion model, thus encouraging the proposed grid search approach. Within this region, **our method is able to successfully estimate motion and correspondences from images alone in a non-combinatoric fashion**. To get a sense of what the cost function $\Phi(\mathbf{A}, \mathbf{b})$ profile is in this region, we show in Fig. 9 the cost function obtained for $N = 20$ points and $N_c = 10$ candidates and the cost function obtained by feeding only ground truth to Algorithm 1. This figure shows that both cost functions exhibit their global minimum in the same neighborhood, which coincides with the values of (θ, ϕ) used to generate data.

It should be noted, in this context, the feature extracting mechanism is the sole responsible for the value of N_c. The choice of local descriptors and their ability to discriminate between possible matches is, therefore, important to avoid multiple explanations for the data.

Grid resolution variation. For each of the test cases, we now vary the resolution of the grid and measure the results of the obtained camera parameters ϕ and θ against ground truth, as well as the value of the cost function $\Phi(\mathbf{A}, \mathbf{b})$ and the average point disparity, for different values of the number of candidates N_c.

A comparison of the cost function variation $\Phi(\mathbf{A}, \mathbf{b})$ (Fig. 10) with the average point disparity (Fig. 11) shows the existent correlation between the average point disparity and the cost value. As the number of candidates increases, however, this correlation becomes less evident for smaller scales. From this experiment, we conclude that the grid resolution should be chosen according to the cardinality of the

level sets $\mathcal{S}_i$, **therefore allowing the use of coarser resolutions without degrading the results of motion and correspondence estimation**, something to consider as a trade-off exists between using a finer resolution and computational effort. It should be noted, however, that while a finer scale allows a clearer distinction between some outlier arrangements and the correct match, multiple arrangements that respect rigidity constraints can still exist. In that case, all explanations within these conditions are accepted as correct.

4.3. Real images

In this section, we exemplify the use of our algorithm to find correspondences in a real case scenario.

Feature selection. For the real image experiments, we have used a corner detector [8, 9] to find points of interest in both images (in order to avoid the dimensionality explosion inherent to considering entire surfaces as possible matches). We randomly select, from the extracted corners, N points to be processed. Care is taken so as to avoid using the same point more than once, as is to discard points with a very small number of candidates in the second frame. For each of these points, we then collect the level sets $\mathcal{S}_i$ by selecting from the detected corners in image 2 according to intensity restrictions.

Lego blocks. In this experiment, we test our method against the 768×576 pixel image pair depicting Lego blocks in different poses present in Fig. 12. This stereo pair is a courtesy of the SYNTIM project database [10] and was

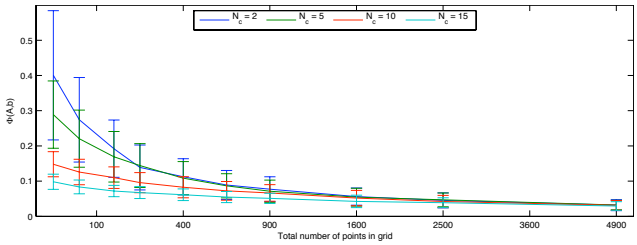


Figure 10. Variation of $\Phi(\mathbf{A}, \mathbf{b})$ with grid resolution and number of candidates N_c for $N = 10$ points, averaged over 200 tests.

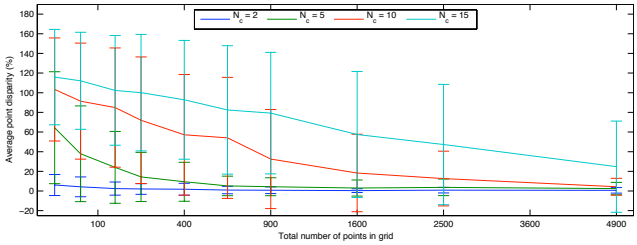
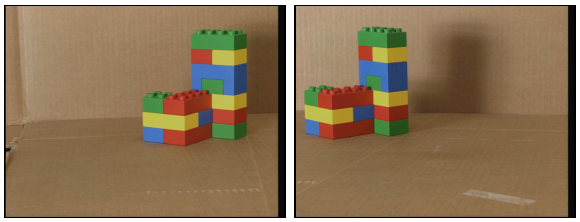
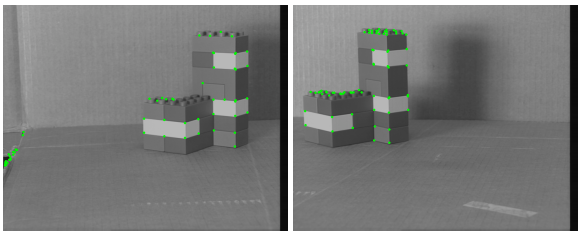


Figure 11. Variation of average point disparity with grid resolution and number of candidates N_c for a set of $N = 10$ points, averaged over 200 tests.



(a) Frame 1 (b) Frame 2

Figure 12. Lego blocks pair, INRIA © copyright.



(a) Frame 1 (b) Frame 2

Figure 13. Corners detected by the algorithm in Lego blocks pair for use in correspondence estimation.

captured with a calibrated setup. For this pair, the corner detector extracted 57 and 69 points from Frames 1 and 2, respectively. The distribution of these features through the images can be seen in Fig. 13.

We used a grid with a total of 2500 points. The estimated camera parameters and their error relative to ground truth (Tab. 1) shows the error values obtained are similar to the error due to grid discretization $\frac{\pi}{\sqrt{2500}} = 0.0628$ rad.

Table 1. Error of estimated parameters in Lego blocks pair.

Parameter	Value	Estimated	Error
r	0.9753	N/A	N/A
ϕ	1.5610	1.5080	0.0531
θ	1.6052	1.5708	0.0344

The matching results (Fig. 14) show correct matches for 23 out of 24 points selected.

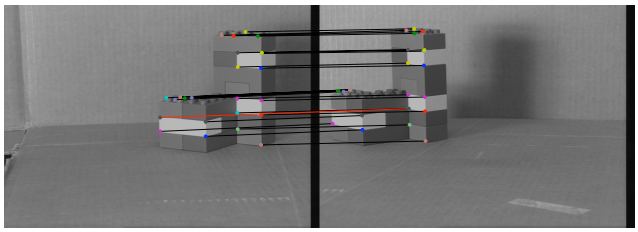


Figure 14. Matching results for Lego blocks pair. Red lines represent mismatches.

Lego blocks with occlusion. In this experiment, we take the points selected in the previous experiment and remove

the ground truth match from $\mathcal{S}_1$. We obtain the same pa-

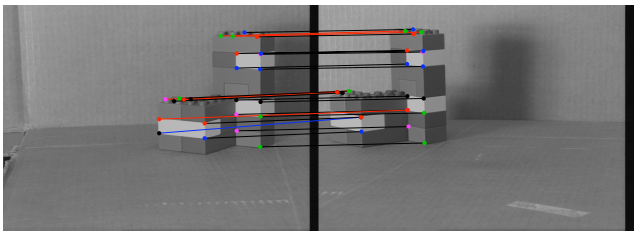
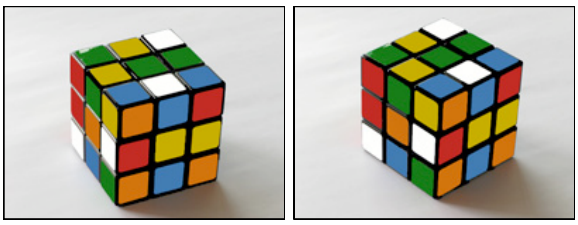


Figure 15. Matching results for Lego blocks pair with occluded points. Blue line represents correspondence obtained for occluded point. Red lines represent mismatches.

rameter estimation errors as before (Tab. 1) and the matching results (Fig. 15) still show correct matches for 18 out of 24 points. These results show that the existence of light occlusion does not seem to perturb the matching process for the majority of remaining points, allowing us to conclude that **our method is able to estimate correspondences and motion in a robust manner while being subject to a high number of outliers and light occlusion**.

Rubik's cube. In this experiment, we test our method against the 512×384 pixel image pair present in Fig. 16, depicting a Rubik's cube with slight perspective effects, captured by the author with an uncalibrated camera. We digitally edited the pictures to remove light reflections in the black grid of the cube and to harmonize face colors between frames. We selected 20 points in frame 1 and obtained level sets in frame 2 with 5 to 19 candidates. Since the corner detector didn't detect some correct points in frame 2 as possible matches, we added them to their respective level sets by hand. We used a grid with a total of 2500 points. The



(a) Frame 1 (b) Frame 2

Figure 16. Rubik's cube pair.

matching results (Fig. 17) show correct matches for 15 out of 20 points selected. Note that in this case, there are various explanations of the data by different physical models.

However, if we extend these results to include 5 point correspondences with the least cost for each point (Fig. 18), these now include the majority (4 out of 5) of the correct matches not found previously. Considering the best 8 matches, the remaining correct match is included.

This experiment shows that **in cases where correspondence estimation is hard, our method is still able to drastically reduce the cardinality of the set of possible correspondence matches between frames obtained using brightness constraints**.

4.4. Coupling with feature extraction methods

In this section, we present a joint utilization of our method with other features, to illustrate the versatility in having our method's input consist simply of bags of point coordinates, regardless of whence they were obtained from. For this purpose, we revisit the Rubik's cube image pair, this time in its original condition (Fig. 19). In this case corner detection is not appropriate, as it detects features in the cube grid and faces due to lighting variations (Fig. 20). This leads to a large amount of features (663 and 1590 for the first and second frame, resp.) and consequently, to a poor correspondence estimation (Fig. 21), with no correct matches in the set of 5 least cost matches.

We replace corners with SIFT features, proceeding as follows: we extract from each of the frames image descriptors using the implementation in [11] (Fig. 22); then, we

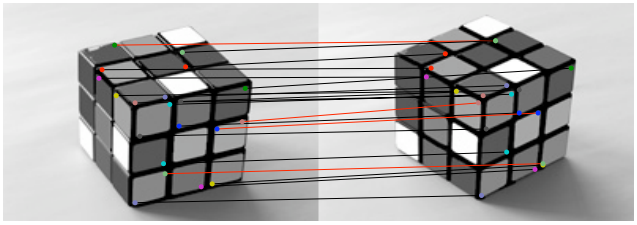


Figure 17. Matching results for Rubik's cube pair. Red lines represent mismatches.

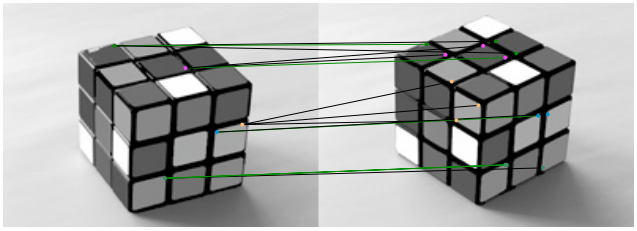
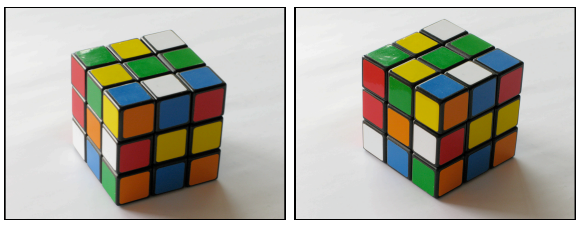
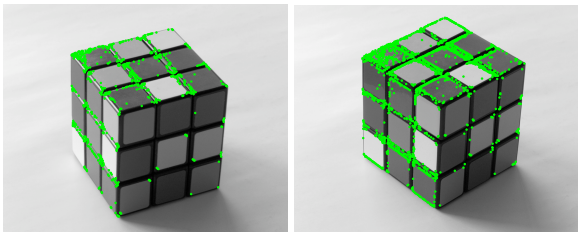


Figure 18. Results for best 5 matches in Rubik's cube pair. Green lines represent ground truth explanation.



(a) Frame 1 (b) Frame 2

Figure 19. Original Rubik's cube pair



(a) Frame 1 (b) Frame 2

Figure 20. Corners detected by the algorithm in original Rubik's cube pair for use in correspondence estimation.

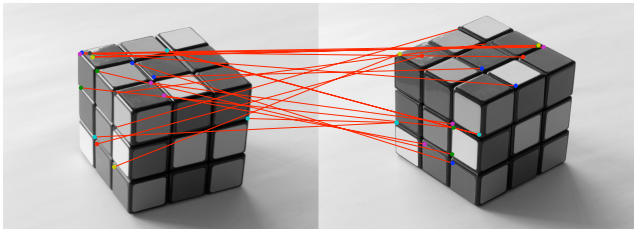
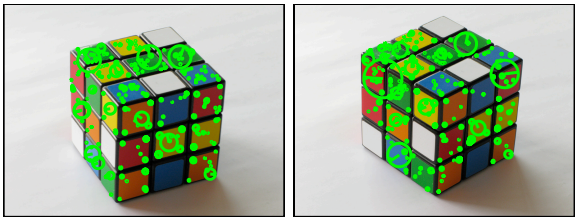


Figure 21. Correspondence results for original Rubik's cube pair using corners as features for a grid resolution of 2500 points.



(a) Frame 1 (b) Frame 2

Figure 22. Features detected by SIFT in the original Rubik's cube pair for use in correspondence estimation.

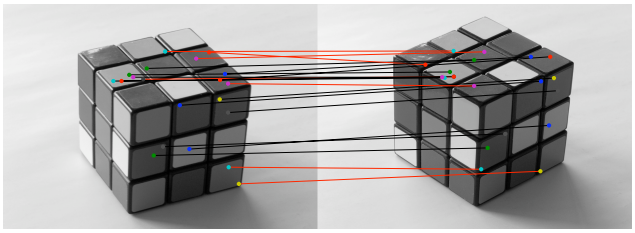


Figure 23. Correspondence results for original Rubik's cube pair using SIFT features for a grid resolution of 2500 points.

randomly select $N = 20$ points from the features in the first frame and form, for each of them, their respective level sets S_i from the features in the second frame. Note that in this experiment, we dismissed the scale and orientation information embedded in the features, since having considerable 3D motion does not allow a robust match based on these cues. Results (Fig. 23) show a significant improvement over using corners as features. Out of 20 points, 13 are included in the selection of 5 matches with least cost.

5. Conclusions

We have successfully designed and demonstrated through experimentation a method (CERD) to estimate motion and correspondence between a pair of images by imposing global rigidity constraints to points and correspondence candidates found using local descriptors. CERD has polynomial complexity, therefore bypassing the combinatorial explosion associated with the correspondence problem. Moreover, it is able to cope with large motions, mild occlusion and a large number of outliers.

The resulting algorithm has optimal properties when the camera parameters are known. When they are not, it performs a two-parameter grid search on the motion space to find the globally best possible explanations for the data.

The intrinsic ambiguity of having various possible explanations for the data prevents the distinction of a specific configuration as the "correct" one. This can be attenuated by using more discriminative local descriptors to find points and correspondence candidates.

References

[1] P. Aguiar and J. Moura. Rank 1 weighted factorization for 3d structure recovery: algorithms and performance analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1134–1149, Sept. 2003. 2

[2] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. 2

[3] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, February 1990. 2

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 1

[5] J. Mota and A. P.M.Q. Efficient methods for points matching with known camera orientation. Technical report, ISR/IST, 2008. 5

[6] A. Neri and G. Jacovitti. Maximum likelihood localization of 2-d patterns in the gauss-laguerre transform domain: theoretic framework and preliminary results. *IEEE Transactions on Image Processing*, 13(1):72–86, 2004. 1

[7] R. Oliveira, R. Ferreira, and J. P. Costeira. Optimal multiframe correspondence with assignment tensors. In *ECCV (3)*, pages 490–501, 2006. 5

[8] E. Rosten and T. Drummond. Fusing points and lines for high performance tracking. In *IEEE International Conference on Computer Vision*, volume 2, pages 1508–1511, October 2005. 8

[9] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision (to appear)*, May 2006. 8

[10] T. SYNTIM at INRIA. Stereograms (stereo images) of the Syntim team, INRIA. http://perso.lcpc.fr/tarel.jean-philippe/syntim/paires.html, 2004. 8

[11] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. 9