Using multiple-hypothesis disparity maps and image velocity for 3-D motion estimation

D. Demirdjian T. Darrell

Massachusetts Institute of Technology, AI Laboratory
Cambridge, MA 02139
demirdji@ai.mit.edu trevor@ai.mit.edu

Abstract

In this paper we explore a multiple hypothesis approach to estimating rigid motion from a moving stereo rig. More precisely, we introduce the use of Gaussian mixtures to model correspondence uncertainties for disparity and image velocity estimation. We show some properties of the disparity space and show how rigid transformations can be represented. An algorithm derived from standard random sampling-based robust estimators, that efficiently estimates rigid transformations from multi-hypothesis disparity maps and velocity fields is given.

1. Introduction

Estimation of rigid motion from a moving stereo rig is an important problem in computer vision. To solve it, one must both identify correspondences across viewpoint, i.e., disparity, and across time, i.e., displacement. Searching for the joint set of disparity and displacement estimates that obey the equations of rigid motion is the fundamental step in rigid-body motion stereo.

Unfortunately, it is well known that estimation of both disparity (stereo) and displacement (optical flow) correspondences is often extremely ambiguous. Many common image structures do not offer unique correspondence: *repetitive textures, low textures regions, ...*

One approach to this problem is to identify a core set of image features which are likely to have unique matches. An operator which selects "good features to track", e.g., [15] can locate points whose local contrast constrains tracking in both the horizontal and vertical direction. But many scenes will not have a sufficient number of such features, or they may be difficult to track over longer time periods. And to only use points at the image which pass such an interest operator ignores the constraints offered by the partially ambiguous points which could in some cases improve the estimation.

Indeed, the success of parametric optical flow estimation shows that reliable estimates of global motion can be obtained even when the local measurements are almost everywhere ambiguous [1, 13, 2]. Such gradient-based methods are appealing because they integrate locally linear constraints, rather than single point estimates of image displacement. Gradient-based motion-stereo has been demonstrated [9, 10], but requires coarse-to-fine processing with 3-D warping when motion yields greater than sub-pixel displacements.

We would like to capture displacement/disparity information at partially ambiguous points, but without requiring a coarse-to-fine method. Our approach is to develop a rigid motion stereo estimator that uses non-unimodal input representations. In most rigid motion stereo estimation algorithms, the input representation is an optical flow and disparity map. These representations capture only the mean estimate (and sometimes variance) of the displacement or disparity. In this paper we explore a multiple hypothesis approach to estimating rigid motion from a moving stereo rig. We use a disparity and displacement space representation in which a Gaussian mixture model represents the match surface over the range of possible offsets.

Organization

The paper is organized as follow. In Section 2, we introduce the use of Gaussian mixtures to model correspondence uncertainties for disparity and velocity (optical flow) estimation. An algorithm to estimate them is given in Section 3. In Section 4 we show some properties of the disparity space and show how rigid transformations can be represented. An algorithm, derived from standard random sampling-based robust estimators that efficiently estimates rigid transformations from multi-hypothesis disparity maps and velocity fields, is given in Section 5. Finally Section 6 shows some experiments with real data and the approach is discussed in Section 7.

2. Multiple-hypothesis disparity maps and image velocity fields

Estimating disparity maps from stereo images or image velocity fields [8] from two consecutive images of a sequence are difficult and challenging problems. Even if these problems are often tackled from different point of views (correlation-based methods, optical flow), they both consist in finding the most similar sub-images in two images.

In most of tracking and motion estimation algorithms, disparity maps and velocity fields are supposed to be uniquely identified (one image pixel is associated to one disparity and one velocity). However there are some cases (*e.g.* low-textured regions, repetitive textures, depth discontinuity) when the information extracted from the images is not sufficient enough to recover a unique correspondence. In order to lower some ambiguities, coarse-to-fine methods [14], and region growing techniques [12] are usually employed with some degree of success. Techniques involving the utilization of interest points [16, 7] are also interesting but provide only sparse range images.

All these approaches comes to use only part of the available information contained in the images or make hypothesis about the 3-D scene. In order to use all the available image information (dense approach), a probabilistic correspondence model allowing for multiple hypothesis is necessary.

Gaussian mixtures are well adapted to model multiple correspondence hypothesis. We propose to model the probability of a point correspondence (disparity maps, velocity fields) using mixtures of Gaussians, each component of the mixture corresponding to a possible correspondence.

In the case of disparity maps, the probability $p(d|x,y)$ of observing a disparity d of a pixel (x,y) is given by:

$$p(d|x,y) = \sum_{i=1}^{K_d} \pi_i \frac{e^{-\frac{(d-d_i)^2}{2\sigma_{d_i}^2}}}{\sqrt{2\pi}\sigma_{d_i}} \qquad (1)$$

The disparity d of (x,y) is then considered to be d_i with a probability π_i and a variance σ_i^2.

In the case of velocity fields, the probability $p(\boldsymbol{\tau}|x,y)$ of observing the image velocity $\boldsymbol{\tau} = (\delta x, \delta y)$ is given by:

$$p(\boldsymbol{\tau}|x,y) = \sum_{j=1}^{K_\tau} \rho_j \frac{e^{-(\boldsymbol{\tau}-\boldsymbol{\tau}_j)^\top \Sigma_{\tau_j}^{-1} (\boldsymbol{\tau}-\boldsymbol{\tau}_j)}}{\sqrt{2\pi|\Sigma_{\tau_j}|}} \qquad (2)$$

The velocity $\boldsymbol{\tau}$ of (x,y) is then considered to be $\boldsymbol{\tau}_j$ with a probability ρ_j and a covariance Σ_j.

These models describe well the correspondence ambiguities. It also enables to capture more visual information along the vision process and propagate it to the different levels of the system. In this paper, we show these correspondence models can be propagated and updated using a robust estimator in order to recover the motion of rigid objects observed by a stereo rig.

3. Gaussian mixtures estimation

We describe in this section an algorithm to estimate the Gaussian mixtures (1) associated with a disparity map from a rectified image pair and (2) associated with the velocity field between two consecutive images of a sequence.

3.1. Disparity maps

Let $\mathcal{C}^{(l,r)}(x,y,x',y')$ be a correlation function (*e.g.* SAD, SSD, ...), normalized such that its values are in $[0,1]$, between the points (x,y) in the left image and (x',y') in the right image. Let $\mathcal{C}^{(l,r)}_{min}(x,y)$ be the minimum correlation score $\mathcal{C}^{(l,r)}(x,y,x-d,y)$ over d values in a user-defined range. A standard approach consists in considering the disparity d associated with $\mathcal{C}^{(l,r)}_{min}(x,y)$ as the disparity of (x,y). Instead we consider that each disparity d_i such that:

$$\mathcal{C}^{(l,r)}(x,y,x-d_i,y) \leq \mathcal{C}^{(l,r)}_{min}(x,y) + \xi_{disp}$$

where ξ_{disp} is a tolerance threshold, is a potential disparity. Then such d_i is considered as the mean of a Gaussian component of the mixture and associated with a variance $\sigma_{d_i} = \sigma$ and weight π_i defined as:

$$\pi_i = \frac{1}{N_\pi} \frac{1 - \mathcal{C}^{(l,r)}(x,y,x-d_i,y)}{1 - \mathcal{C}^{(l,r)}_{min}(x,y)}$$

where N_π is a normalization factor such that: $\sum_i^{K_d} \pi_k = 1$.

3.2. Velocity fields

In order to estimate the velocity field between two consecutive images of a sequence, a similar approach as for disparity maps is used. For each point (x,y) in the first image, all velocities $\boldsymbol{\tau} = (\delta x, \delta y)$ in a user-defined range are considered and the correlation function $\mathcal{C}^{(l,l')}(x,y,x',y')$ is evaluated between the point (x,y) and $(x',y') = (x + \delta x, y + \delta y)$ between two consecutive images of a sequence. Let $\mathcal{C}^{(l,l')}_{min}(x,y)$ be the minimum of these values.

Each velocity $\boldsymbol{\tau}_j = (\delta x_j, \delta y_j)$ such that:

$$\mathcal{C}^{(l,l')}(x,y,x+\delta x_j, y+\delta y_j) \leq \mathcal{C}^{(l,l')}_{min}(x,y) + \xi_{veloc}$$

where ξ_{veloc} is a tolerance threshold, is considered as a potential velocity. Then such $\boldsymbol{\tau}_j$ defines the mean of a Gaussian of covariance $\Sigma_{\tau_j} = \sigma^2 \mathbf{I}$ and weight $\rho_j = \frac{1}{N_\rho} \frac{1-\mathcal{C}^{(l,l')}(x,y,\boldsymbol{\tau}_j)}{1-\mathcal{C}^{(l,l')}_{min}(x,y)}$ where N_ρ is a normalization factor.

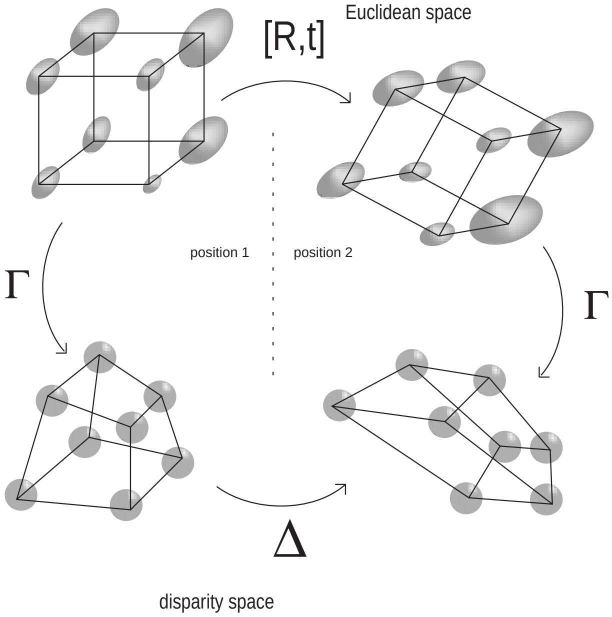

Figure 1: Euclidean reconstruction and motion of a cube *vs.* reconstruction and motion in the disparity space.

4. Rigid transformations in the disparity space

In this section, we introduce the transformation that maps two reconstructions of a rigid scene in the disparity space. We call this transformation *d-motion* and show how it is related to the rigid motion in the Euclidean space (see [4] for details). We also show how to estimate this transformation from a list of correspondences $\omega = (x, y, d) \mapsto \omega' = (x', y', d')$.

4.1. Properties of the disparity space

In this section we argue the use of disparity images for spatial representation of stereo data. We claim that this representation has nice geometric and topological properties that makes it ideal for spatial data representation and optimal motion estimation.

We show that (i) using homogeneous coordinates, the disparity image is a particular projective reconstruction of the 3-D observed scene; therefore, the disparity space is a projective space, and (ii) for parallel camera stereo rigs, the noise in the disparity space is isotropic.

4.2. Geometric feature

Let us consider a parallel stereo rig. Let f be the focal length, (u_0, v_0) the principal point coordinates associated with the stereo rig and B be the baseline of the stereo rig.

Let $\boldsymbol{M} = (X\ Y\ Z)$ be the 3-D coordinates of a point observed by a stereo rig. Let d be the disparity of the associated image point $\boldsymbol{m} = (x\ y)$ and $(\bar{x}\ \bar{y}) = (x - u_0\ y - v_0)$ the centered image point coordinates.

Then the relation between (X, Y, Z) and $(\bar{x}, \bar{y}, d)$ is:

$$\begin{cases} \bar{x} = x - u_0 = f\frac{X}{Z} \\ \bar{y} = y - v_0 = f\frac{Y}{Z} \\ d = \frac{fB}{Z} \end{cases} \qquad (3)$$

Let ω a vector such that $\omega = \begin{pmatrix} \bar{x} \\ \bar{y} \\ d \end{pmatrix}$. Using homogeneous coordinates and multiplying each term of the equations (3) by Z, we can show [5, 4] that:

$$\begin{pmatrix} \omega \\ 1 \end{pmatrix} = \frac{1}{Z}\mathbf{\Gamma} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \simeq \mathbf{\Gamma} \begin{pmatrix} \boldsymbol{M} \\ 1 \end{pmatrix} \qquad (4)$$

where ”$\simeq$” denotes the equality up to a scale factor and $\mathbf{\Gamma}$ is a 4×4 matrix such that:

$$\mathbf{\Gamma} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 0 & fB \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

The Eq.(4) demonstrates that there is a projective transformation $\mathbf{\Gamma}$ between the homogeneous coordinates $(X\ Y\ Z\ 1)$ in the 3-D Euclidean space and the homogeneous coordinates $(\bar{x}\ \bar{y}\ d\ 1)$. Therefore $(\bar{x}\ \bar{y}\ d\ 1)$ is a projective reconstruction of the scene. The disparity space is then a projective space.

4.3. Topological feature

An important feature of the disparity space is that the noise associated with $(\bar{x}\ \bar{y}\ d)$ is known:

- The noise associated with $\bar{x}$ and $\bar{y}$ is due to the *image discretization*. Without any *a priori* information, the variances $\sigma_{\bar{x}}$ and $\sigma_{\bar{y}}$ of this noise is the same for *all* image points. We can write $\sigma_{\bar{x}} = \sigma_{\bar{y}} = \sigma$ where σ is the pixel accuracy (typically $\sigma = 1$ *pix.*);
- The noise associated with d is related to the matching process and modeled as a Gaussian mixture as shown in Eq.(1). Each Gaussian of the mixture is characterized by a mean d_i (disparity hypotyhesis) and a variance σ (pixel accuracy).

It is clear that the noises associated with $\bar{x}$, $\bar{y}$ and d are independent. It is worth noticing that when the mixture describing d is composed of a single component the noise in d is a unimodal Gaussian. Therefore the noise in $(\bar{x}\ \bar{y}\ d)$ is isotropic and homogeneous.

4.4. Motion in the disparity space: d-motion

Let us consider a fixed stereo rig observing a moving point. Let $\boldsymbol{M} = (X\ Y\ Z)$ and $\boldsymbol{M'} = (X'\ Y'\ Z')$ be the respective 3-D Euclidean coordinates of this point before and after the rigid motion. Let $\mathbf{R}$ and $\boldsymbol{t}$ denote the rotation and translation of the rigid motion. Using homogeneous coordinates we have:

$$\begin{pmatrix} \mathbf{M}' \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{M} \\ 1 \end{pmatrix}$$

Replacing $\begin{pmatrix} \mathbf{M} \\ 1 \end{pmatrix}$ and $\begin{pmatrix} \mathbf{M}' \\ 1 \end{pmatrix}$ using Eq.(4) we have:

$$\Gamma^{-1} \begin{pmatrix} \omega' \\ 1 \end{pmatrix} \simeq \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \Gamma^{-1} \begin{pmatrix} \omega \\ 1 \end{pmatrix}$$

Let $\mathbf{H}_d = \Gamma \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \Gamma^{-1}$. Then we have:

$$\begin{pmatrix} \omega' \\ 1 \end{pmatrix} \simeq \mathbf{H}_d \begin{pmatrix} \omega \\ 1 \end{pmatrix} \tag{5}$$

Let $\bar{\mathbf{R}}$ be a 2×2 matrix, $\boldsymbol{r}$, $\boldsymbol{s}$ and $\bar{\boldsymbol{t}}$, 2-vectors and λ and μ scalars such that:

$$\mathbf{R} = \begin{pmatrix} \bar{\mathbf{R}} & \boldsymbol{r} \\ \boldsymbol{s}^\top & \lambda \end{pmatrix} \qquad \boldsymbol{t} = \begin{pmatrix} \bar{\boldsymbol{t}} \\ \mu \end{pmatrix}$$

Then $\mathbf{H}_d$ can be expressed as follow:

$$\mathbf{H}_d = \begin{pmatrix} \bar{\mathbf{R}} & \frac{1}{B}\bar{\boldsymbol{t}} & f\boldsymbol{r} \\ 0 & 1 & 0 \\ \frac{1}{f}\boldsymbol{s}^\top & \frac{\mu}{fB} & \lambda \end{pmatrix} \tag{6}$$

Using standard coordinates, Eq.(5) becomes:

$$\boxed{\omega' = \Delta(\omega) = \frac{1}{(\omega^\top\ 1)^\top \boldsymbol{\gamma}} (\mathbf{A}\omega + \boldsymbol{b})} \tag{7}$$

where $\mathbf{A}$ is a 3×3 matrix, $\boldsymbol{b}$ a 3-vector and $\boldsymbol{\gamma}$ a 4-vector such that:

$$\mathbf{A} = \begin{pmatrix} \bar{\mathbf{R}} & \frac{1}{B}\bar{\boldsymbol{t}} \\ 0 & 1 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} f\boldsymbol{r} \\ 0 \end{pmatrix} \quad \boldsymbol{\gamma} = \frac{1}{f} \begin{pmatrix} \boldsymbol{s} \\ \frac{\mu}{B} \\ f\lambda \end{pmatrix}$$

The transformation $\omega' = \Delta(\omega)$ is called *d-motion*. In homogeneous coordinates, it can be defined by the matrix $\mathbf{H}_d$. In standard coordinates, it can be defined by $\mathbf{A}$, $\boldsymbol{b}$ and $\boldsymbol{\gamma}$.

4.5. Motion estimation with d-motion

Let $\omega_i \rightarrow \omega_i'$ be a list of point correspondences. The problem of estimating the rigid motion between the points ω_i and ω_i' amounts to minimizing over Δ the following error:

$$E^2 = \sum_i \varepsilon_i^2 \tag{8}$$

where $\varepsilon_i^2 = ||\Delta(\omega_i) - \omega_i'||^2$

As demonstrated previously, if the focal length f and the baseline B of the stereo rig are known, Δ can be parameterized by $\mathbf{R}$ and $\boldsymbol{t}$. The error E^2 can therefore be minimized over $\mathbf{R}$ and $\boldsymbol{t}$.

In the case of small motion, the rotation $\mathbf{R}$ can be parameterized by:

$$\mathbf{R} = \mathbf{I} + \begin{pmatrix} 0 & -w_c & w_b \\ w_c & 0 & -w_a \\ -w_b & w_a & 0 \end{pmatrix} \tag{9}$$

The error ε_i^2 can be expressed in a quasi-linear way:

$$\begin{aligned} \varepsilon_i^2 &= ||\Delta(\omega_i) - \omega_i'||^2 \\ &= ||\frac{1}{(\omega_i^\top\ 1)^\top \boldsymbol{\gamma}} (\mathbf{A}\omega_i + \boldsymbol{b}) - \omega_i'||^2 \\ &= \nu_i^2 ||\mathbf{P}_i \boldsymbol{u} + \boldsymbol{v_i}||^2 \end{aligned}$$

where $\nu_i = \frac{1}{(\omega_i^\top\ 1)^\top \boldsymbol{\gamma}}$, $\boldsymbol{u} = (w_a\ w_b\ w_c\ \boldsymbol{t}^\top)^\top$. $\boldsymbol{v_i}$ is a 3-vector and $\mathbf{P}_i$ is a 3×6 matrix whose entries depend on ω_i and ω_i'.

The total error E^2 is then:

$$E^2 = \sum_i \nu_i^2 ||\mathbf{P}_i \boldsymbol{u} + \boldsymbol{v_i}||^2 \tag{10}$$

This form enables to perform the minimization of E^2 using an iterative weighted linear least square:

1. *Initialization:* Let $\nu_i(0) = 1$. Estimate $\boldsymbol{u}$ using Eq.(10);
2. *Evaluate* $\nu_i(k+1) = \frac{1}{(\omega_i^\top\ 1)^\top \boldsymbol{\gamma}}$ from the current solution for $\boldsymbol{u}$;
3. *Minimize:* $E^2(k+1)$ using a standard weighted linear least square method (ν_i fixed);
4. *Stop test:* if $|E^2(k+1) - E^2(k)| \leq \epsilon$ then stop, else return to step 2.

It is well known that this kind of approach is not guaranteed to converge to the correct minimum. However we observed that in practice, the minimization was correct and performed in few iterations.

In the case of larger motions, a global non-linear minimization must be performed over the 6 motion parameters.

4.6. Properties of the d-motion estimator

There are many theoretical advantages of estimating the motion from disparity space and d-motion:

- **Estimation accuracy**. Minimizing E^2 gives an accurate estimation of $\mathbf{R}$ and $\boldsymbol{t}$. The noise of points in the disparity space is homogeneous and isotropic (as stated in section 4.3). Therefore minimizing E^2 gives a (statistically) quasi-optimal estimation.

- **Generalization to the uncalibrated case**. The d-motion can be generalized in the uncalibrated case (f and B unknown). In that case, $\mathbf{A}$, $\boldsymbol{b}$ and $\boldsymbol{\gamma}$ can be represented by 12 general parameters such that:

  $$\mathbf{A} = \begin{pmatrix} \star & \star & \star \\ \star & \star & \star \\ 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} \star \\ \star \\ 0 \end{pmatrix} \quad \boldsymbol{\gamma} = \begin{pmatrix} \star \\ \star \\ \star \\ \star \end{pmatrix}$$

  In this case, the d-motion can be considered as a particular case of projective transformation [3] and has the same structure as an affine transformation [11]. $\mathbf{R}$ and $\boldsymbol{t}$ cannot be recovered, but the d-motion Δ can still be estimated and used for tasks requiring no Euclidean estimation, such as motion segmentation or motion detection.

5. Robust Structure and Motion estimation

In this part, we describe a method to estimate d-motions from *multiple-hypothesis* disparity maps and velocity fields.

A direct estimation of the d-motion using the method described in Section 4.5 is obviously not possible because of multiple correspondences but also because:

- there may be outliers (*e.g.* mismatched points due to occlusions);
- there may be many independently moving objects in the scene. Then all points don't satisfy a single motion.

A robust estimator is then necessary to estimate the d-motion. Though random sampling-based methods like RANSAC [6] or LMedS (Least Median Square) are appealing and have been used with success in the context of motion estimation, they cannot be directly applied (because of multiple hypothesis). However we introduce here an algorithm largely inspired from standard random sampling approaches but adapted for multiple hypothesis data.

Data samples are selected by a random sampling process. For each sample, a solution for the d-motion is estimated and a criterion is estimated over the entire data set. The solution yielding the best criterion is finally kept.

Random sampling process

A sample consists here of n image points associated each with a set of correspondences. A set of correspondences of a point (x, y) is composed of (i) a disparity $d_i(x, y)$ before motion, (ii) a velocity $\boldsymbol{\tau}_j(x, y) = (\delta x, \delta y)$ and (iii) a disparity $d'_k(x + \delta x, y + \delta y)$ after motion.

Image points (x, y) are first randomly chosen in a list of *support* points, composed of points associated with the least disparity and velocity ambiguities (*i.e.* $K_d(x, y)$ *and* $K_\tau(x, y)$ *small*. It could be appealing to chose as support points only points having no ambiguities at all $(i.e. K_d(x, y) = K_\tau(x, y) = K'_d(x + \delta_x, y + \delta y) = 1)$. However, in some difficult cases (see images 4 and reffig:experim3), there may be few such points (*e.g.* a black circle on a white background does not have any such points) and some ambiguous (multi-hypothesis disparity and velocity) points have to be chosen as well.

In the case where $K_d(x, y) = K_\tau(x, y) = K'_d(x + \delta_x, y + \delta y) = 1$, (x, y) has a unique set of correspondence (d, τ, d'). Otherwise the set of correspondences of (x, y) is chosen as follow:

- The disparity d_i (*resp.* velocity $\boldsymbol{\tau}_j$) is randomly chosen among the multiple components of the Gaussian mixture of the point (x, y) of the disparity map before motion (*resp.* velocity field) with a probability π_i (*resp.* ρ_j);
- The disparity d'_k is then randomly chosen as one of the components of the Gaussian mixture of the point $(x + \delta x, y + \delta y)$ of the disparity map after motion.

D-motion estimation

The d-motion is estimated from the sample of n points using Eq.(7). A minimum choice for n is 3. However in our experiments we chose $n \geq 5$ in order to have a better estimate.

Criterion

A point (x, y) is considered as *inlier* for a d-motion if a valid set of correspondences $(d_i, \boldsymbol{\tau}_j, d'_k)$ exists such that it is consistent with this d-motion, *i.e.* if the transfer error $\varepsilon^2 = ||\Delta(x\ y\ d_i)^\top - (x + \delta x\ y + \delta y\ d'_k)^\top||^2$ is smaller than a threshold.

Let ϕ be defined by:

$$\phi(x, y, d_i, \boldsymbol{\tau}_j, d'_k) = \begin{cases} \pi_i \rho_j \pi'_k & \text{if } \varepsilon^2 \leq \alpha\sigma^2 \\ 0 & \text{otherwise} \end{cases}$$

The function ϕ gives (i) a null score to sets of correspondences not consistent with a d-motion Δ and (ii) a score equal to the probability of observing $(d_i, \boldsymbol{\tau}_j, d'_k)$ for consistent ones.

The criterion φ associated with an image point is then given by:

$$\varphi(x, y) = \max_{i=1}^{K_d(x,y)} \max_{j=1}^{K_\tau(x,y)} \max_{k=1}^{K'_d(x+\delta x_j, y+\delta y_j)} \phi(x, y, d_i, \boldsymbol{\tau}_j, d'_k) \tag{11}$$

The criterion φ is equal to the score ϕ of the most probable consistent correspondence set.

Finally the total criterion is:

$$\Phi = \sum_{x,y} \varphi(x,y) \tag{12}$$

It is worth noticing that the number of iterations to estimate φ in (11) should be $K_d(x,y)K_\tau(x,y)\sum_{k=1}^{K_\tau(x,y)} K'_d(x+\delta x_k, y+\delta y_k)$.

Hopefully the complexity can be made much smaller. We assume that a table $(x,y,\delta x,\delta y) \mapsto \rho$ (*resp.* $(x',y',d') \mapsto \pi'$) that gives 0 if $\tau = (\delta x, \delta y)$ (*resp.* d') is not a mixture component of point (x,y) (*resp.* (x',y')) and its weight ρ (*resp.* π') otherwise. These tables can easily be filled during the Gaussian mixtures estimation.

Then, for each component d_i in the mixture, a theoretical transfer point $(\hat{x}', \hat{y}', \hat{d}')$ can be estimated. The search of consistent $(\delta x, \delta y, d')$ has just to be done by looking in the tables $(x,y,\delta x,\delta y) \mapsto \rho$ and $(x',y',d') \mapsto \pi'$ in cells such that $(\hat{x}' - x - \delta x)^2 + (\hat{y}' - y - \delta y)^2 + (\hat{d}' - d')^2 \leq \alpha\sigma^2$. The number of iterations to estimate ϕ is then proportional to $K_d(x,y)$.

Algorithm

The random sampling is then performed many times. For each sample, a d-motion is estimated and a criterion Φ is estimated over the entire image. The process is stopped when a fraction T of the image is considered as inliers (we chose T=50% in our experiments). Then the d-motion is estimated using all available $inliers$ weighed by $\varphi(x,y)$ and a final $inlier$ estimation is performed.

It worth noticing that, while inliers are identified, the set of correct correspondences (d,τ,d') associated with each inlier are identified as well. The weights π' of the Gaussian mixtures corresponding to consistent disparities d' are then reinforced (increment with a fixed value) and normalized.

This updated disparity map can then be used in order to estimate the next motion.

6. Experiments

6.1. D-motion accuracy

Experiments with simulated data are carried out in order to estimate the quality of the d-motion estimator. A synthetic 3-D scene of 100 points is generated. A random rigid motion is generated as well. The 3-D points of each position (before and after the rigid motion) are projected onto the cameras of a virtual stereo rig, and Gaussian noise with varying standard deviation (0.0 to 1.2 *pix*) is added to the image point locations. Two different methods are applied : (i) the method based on d-motion (direct minimization of Eq.(8)) and (ii) the quaternion-based algorithm [8]. In

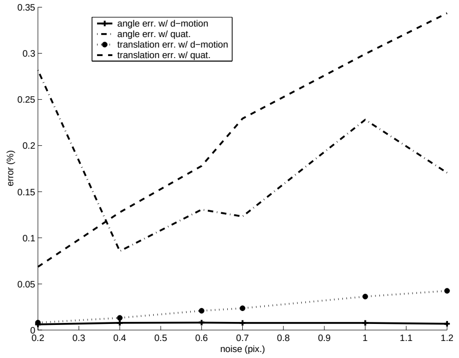

Figure 2: Estimation of the relative angle and translation errors for small motions

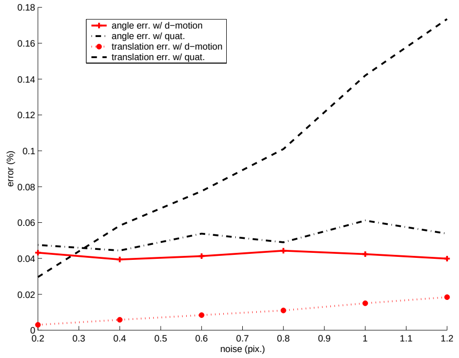

Figure 3: Estimation of the relative angle and translation errors for standard motions

order to compare the results, some errors are estimated between the expected motion and the estimated ones. The criterion errors are : the relation translation error, the relative rotation angle error and the angle between the expected and estimated rotation axis.

This process is iterated 500 times. The mean of each criterion error is shown Figures 2 and 3. Figure 2 shows the estimation of the relative angle and translation errors for small motions (rotation angle smaller than 0.1 rad.). Figure 3 shows the estimation of the relative angle and translation errors for ”standard” motions (rotation angle greater than 0.1rad.. Both figures show that the method gives accurate results even for high image noise (greater than 1.0 pix.).

6.2. D-motion using multi-hypothesis correspondences

Experiments with real data were conducted in order to justify the applicability and reliability of our approach.

We applied our approach to difficult scenes (see Figures 4 and 6). These scenes have a lot of ambiguous points (repetitive textures and horizontal textures for Figure 4 and nearly no textures and corners in Figure 6). We applied our technique from Section 5. In order to estimate the disparity maps and velocity fields, we used a SSD-based technique using 9×9 template windows.

Figure 4 (*resp.* Figure 5) shows[1] the most probable components of the velocity fields corresponding to a vertical ascending translation (*resp.* right-to-left rotation) of the stereo rig.

Figures 6 and 7 show the most probable components of the velocity fields obtained from two different motions of the stereo-rig.

These experiments show that our approach succeeds in finding the correct velocities even for ambiguous image points (such as the ones around the metallic curtains or on the border and center of black circles) where standard optical flow algorithms usually fail.

We also carried out an experiment using a sequence of image pairs in order to estimate the accuracy of the motion estimation. The sequence consists of 220 image pairs (see Figure 8) gathered by a moving stereo rig. The motion is first an ascending vertical translation (motion 1) of 15*cm* with constant speed and then a rotation around a vertical axis (motion 2) of 25*deg* with constant speed.

Figures 9 and 10 show the estimated location of the sensor during motion 1.

Figure 12 shows the error between the estimated rotation axis of the motion and the ground-truth rotation axis during motion 2. Figure 11 shows the estimated angle of motion 2.

[1] For a better visualization, only velocities of points on a grid are represented.

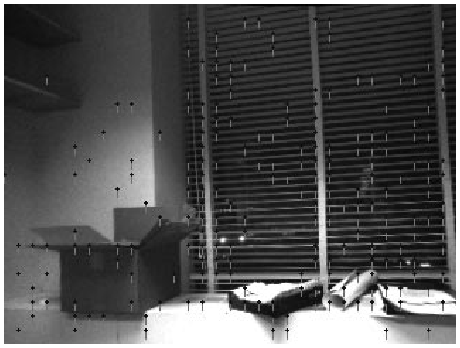

Figure 4: Velocity field estimated from ascending vertical translation

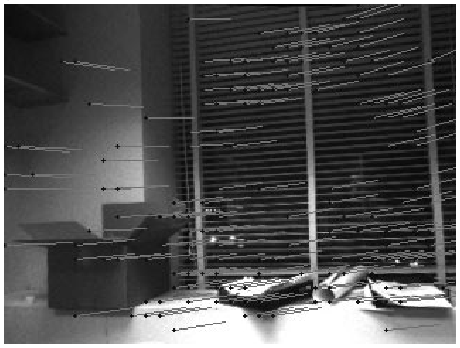

Figure 5: Velocity field estimated from right-to-left rotation

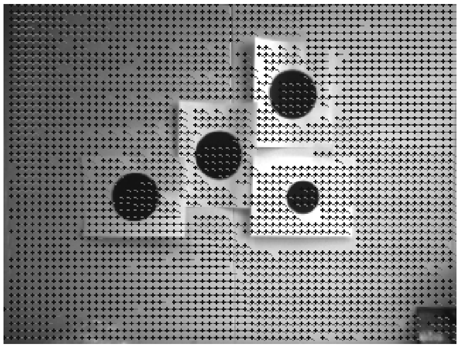

Figure 6: Velocity field estimated from translation

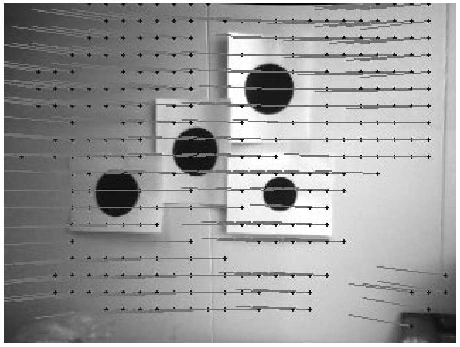

Figure 7: Velocity field estimated from rotation

All the results show that the motion estimation is quite consistent with the ground-truth data. In all the experiments, the motion estimation takes between 5 and 10 *sec.* for 1 image.

7. Discussion

In this paper, we have described a multiple hypothesis approach to estimating rigid motion from a moving stereo rig. We use a disparity and displacement space representation in which a Gaussian mixture model represents the match surface over the range of possible offsets. We also show that the disparity space was an adequate space to represent spatial data and introduced the rigid transformations associated with this space (d-motions). Finally we proposed a random sampling-based algorithm that estimates rigid transformations from multi-hypothesis disparity maps and velocity fields and gives at the same time the most consistent set of disparity and velocity hypothesis.

We show with experiments that our approach enabled to find correct velocity fields for difficult images (images with many ambiguous points) and accurately gave a motion estimation.

In this paper we used Gaussian mixtures where each component has the same variance σ. The main reason for

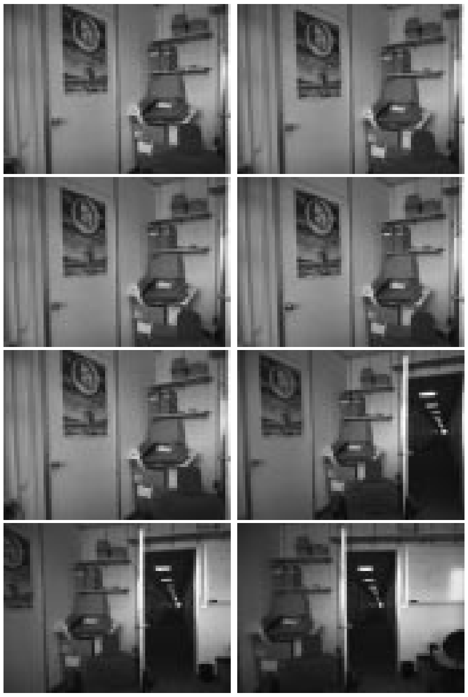

Figure 8: Images extracted from the sequence. The first 4 images correspond to the ascending vertical translation (motion 1). The last 4 images correspond to the rotation around a vertical axis (motion 2).

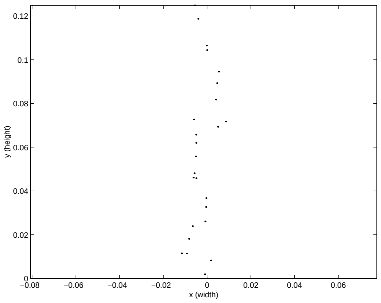

Figure 9: Estimated location of the rig in the xy-plane (fronto-parallel plane)

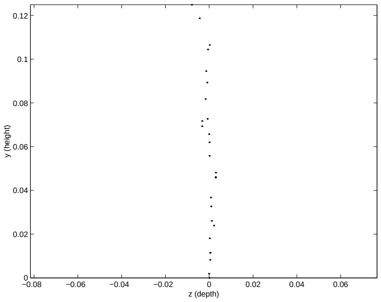

Figure 10: Estimated location of the rig in the yz-plane

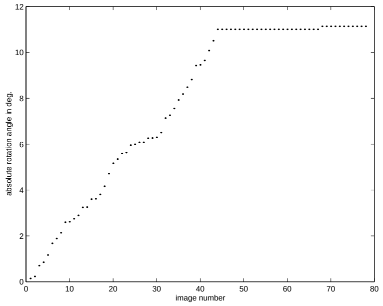

Figure 11: Estimated angle of the rotation

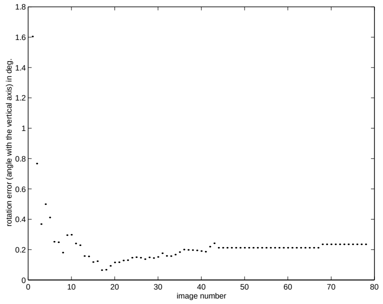

Figure 12: Error angle between the estimated and the ground-truth rotation axis

that is that their estimation is fast and easy using the algorithm from Section 3. However we are investigating the use of general Gaussian mixtures to model correspondences. Such Gaussian mixtures can be estimated using a standard EM-algorithm. Our future work consists in designing an algorithm to robustly estimate the d-motion from such models.

References

[1] J. Bergen , P. Anadan, K. Hanna, and R. Hingorami. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, pages 237–252, 1992.

[2] C. Bregler. Learning and Recognizing Human Dynamics in Video Sequences. In *Proc. Computer Vision and Pattern Recognition*, pp. 568–574, 1997.

[3] D. Demirdjian and R. Horaud. Motion-Egomotion Discrimination and Motion Segmentation from Image-pair Streams. *Computer Vision and Image Understanding*, volume 78, number 1, pages 53–68, April 2000.

[4] D. Demirdjian and T. Darrell. Motion Estimation from Disparity Images In *International Conference on Computer Vision*, pages 213–218, volume I, 2001.

[5] F. Devernay and O. Faugeras. From projective to Euclidean reconstruction. In *In Proc. Computer Vision and Pattern Recognition Conference*, pages 264–269, San Francisco, CA., June 1996.

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

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[8] B.K.P. Horn, H.M. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America*, 5(7): 1127–1135, July 1988.

[9] M. Harville, A. Rahimi, T. Darrell, G. Gordon and J. Woodfill. 3D Pose Tracking with Linear Depth and Brightness Constraints. In *Proc. International Conference on Computer Vision*, Bombay, pp. 206–213, 1999.

[10] G.P. Stein and A. Shashua. Direct Estimation of Motion and Extended Scene Structure from a Moving Stereo Rig. In *Proc. Computer Vision and Pattern Recognition*, 1998.

[11] J. Koenderink and A. van Doorn. Affine structure from motion. *Journal of the Optical Society of America A*, 8(2):377–385, 1991.

[12] M. Lhuillier and L. Quan. Robust Dense Matching Using Local and Global Geometric Constraints. *Proc. of the 16th International Conference on Pattern Recognition*, Barcelona, Spain, Vol. 1, pp. 968–972, 2000.

[13] M. Irani, B. Rousso, and S. Peleg. Computing Occluding and Transparent Motions. In *International Journal on Computer Vision*, pages 5–16, 1994.

[14] M. O'Neill and M. Denos. Automated system for coarse-to-fine pyramidal area correlation stereo matching. *Image and Vision Computing*, vol. 14, no. 3, pp. 225–236, 1996.

[15] J. Shi and C. Tomasi. Good features to track. In *Proc. Computer Vision and Pattern Recognition*, pages 593–600, IEEE Computer Society, Seattle, Washington, June 1994.

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