Tools for Geometric Computer Vision

Juho Kannala

May 3, 2002

Contents

1 Introduction

With our eyes we gather a lot of information from our surroundings. Human vision is of course more than just eyes collecting fotons since our ability to interpret visual information is remarkable. We can observe for instance distances, directions, shapes and colours. One goal of computer vision is to develop methods that would make it possible to obtain this kind of information automatically from digital images. In this work we will concentrate on computational methods that are needed to determine the three dimensional structure of a scene from two dimensional images.

Building the three dimensional reconstruction of a scene from multiple views is an important problem in computer vision. The multiple view geometry has been extensively studied during the last years, but there is still no universally applicable solution to the reconstruction problem. Though there exists a lot of 3D-modeling systems they usually require special scene set ups, special hardware, or manual feature extraction [6]. However, the techniques able to solve the structure from motion problem would have many applications. One application example could be an autonomous robot that moves in a changing environment and observes its surroundings by a camera.

In this report we consider the problem of determining the metric structure of a scene from two views. The starting point is that we have two images of a scene taken by two cameras at different positions. We assume the cameras are calibrated pinhole cameras and the scene structure is initially unknown. Our approach is to search a set of interest points from the images and compute the three dimensional coordinates for them.

The reconstruction problem can be divided in three different stages. The first stage is to automatically extract point matches between the images and is discussed in Section 4. The second stage is to estimate the two view geometry, incorporated in the fundamental matrix, using the point matches. Several methods for the fundamental matrix estimation are considered in Section 5. The final stage considers back-projecting the matched image points to reconstruct the 3D structure (Section 6).

By combining and implementing different techniques proposed in the computer vision literature we outline a procedure for making a rough reconstruction from two views. The methods studied were implemented in MATLAB.

2 Camera Model

Conventional digital cameras can usually be well modeled by the pinhole camera model that is just a central projection from the world points to the image plane. The pinhole camera geometry is shown in Figure 1. The origo of the camera coordinate system is the centre of the projection, also known as the camera centre. The z-axis is perpendicular to the image plane and points to the front of the camera. The point where the z-axis intersects the image plane is called the principal point and the distance between the camera centre ($\mathbf{C}$) and the principal point ($\mathbf{c}$) in the world coordinate frame is the focal length.

The central projection can be represented as a linear transformation in ho-

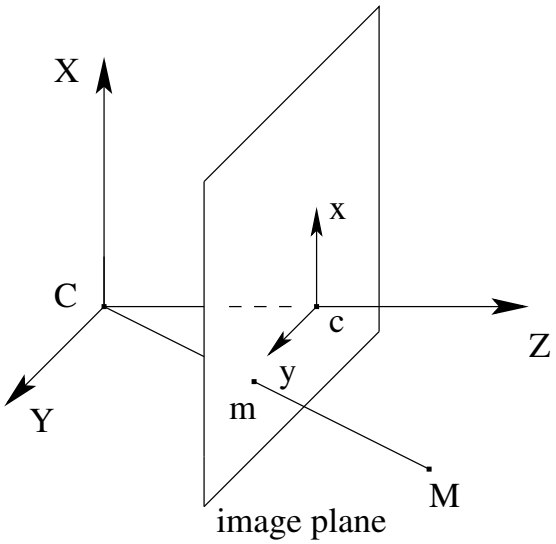

Figure 1: Pinhole camera model.

mogenous coordinates

$$\tilde{\mathbf{m}} = \mathbf{P}\tilde{\mathbf{M}} = \mathbf{K}(\mathbf{R}\ \mathbf{t}) \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \tag{1}$$

where the transformation matrix $\mathbf{P}$ is called the camera projection matrix and $\tilde{\mathbf{M}}$ is a 4-vector representing the homogenous coordinates of a point $\mathbf{M}$ in the world coordinate frame (we use $\tilde{\cdot}$ to denote the homogenous representation). The image of $\mathbf{M}$ is $\mathbf{m}$ and its homogenous pixel coordinates, $\tilde{\mathbf{m}} = (u\ v\ 1)^{\mathrm{T}}$, can be computed from (1). The 3×4 matrix $\mathbf{P}$ is decomposed into two matrices. Here the 3-vector $\mathbf{t}$ and rotation matrix $\mathbf{R}$ relate the camera position and orientation to the world coordinate system,

$$\mathbf{X}_{\mathrm{cam}} = \mathbf{R}\mathbf{X}_{\mathrm{world}} + \mathbf{t}. \tag{2}$$

The matrix $(\mathbf{R}\ \mathbf{t})$ in (1) transforms the world coordinates to the normalized image coordinates by assuming that the focal length is 1. The matrix

$$\mathbf{K} = \begin{pmatrix} fm_u & -fm_u \cot\theta & u_0 \\ 0 & fm_v/\sin\theta & v_0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} a_u & -a_u \cot\theta & u_0 \\ 0 & a_v/\sin\theta & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

contains the camera internal parameters and transforms the normalized image coordinates into the pixel image coordinates [2]. The parameter f means the focal length, θ is the angle between the pixel coordinate axis, m_u and m_v are the number of pixels per unit distance in image coordinates in the directions of u and v. The image of the principal point is $(u_0,\ v_0)$.

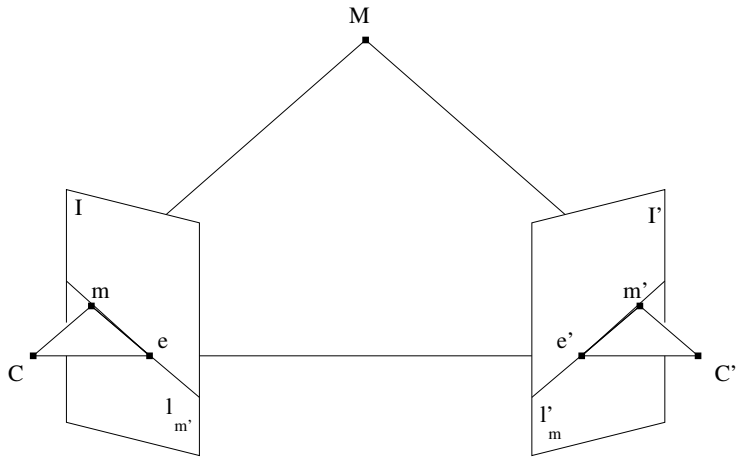

Figure 2: Epipolar geometry. $\mathbf{C}$ and $\mathbf{C}'$ are the camera centres and I and I' are the image planes of the cameras. World point $\mathbf{M}$ is projected to points $\mathbf{m}$ and $\mathbf{m}'$. The corresponding epipolar lines are $\mathbf{l}_{\mathbf{m}'}$ and $\mathbf{l}'_{\mathbf{m}}$.

3 Two View Geometry

The two view geometry, also called as epipolar geometry, is illustrated in Figure 2. The planes I and I' are the image planes of the cameras with camera centres $\mathbf{C}$ and $\mathbf{C}'$. The line joining the two camera centers intersects the image planes in epipoles $\mathbf{e}$ and $\mathbf{e}'$. Each world point $\mathbf{M}$ together with the camera centres defines a plane, where also the image points $\mathbf{m}$ and $\mathbf{m}'$ must lie. This coplanarity constraint characterizes the two view geometry and it can be represented with the essential matrix $\mathbf{E}$ by formula

$$\tilde{\mathbf{x}}'^{\mathrm{T}}\mathbf{E}\tilde{\mathbf{x}} = 0, \tag{4}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are the normalized image coordinates corresponding to $\mathbf{m}$ and $\mathbf{m}'$. The essential matrix is defined by

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}, \tag{5}$$

where translation vector $\mathbf{t}$ and the rotation matrix $\mathbf{R}$ are as in (2) with the first camera coordinate frame taken as the world coordinate frame. $[\mathbf{t}]_\times$ is a skew symmetric 3×3 matrix representing the cross-product with the vector $\mathbf{t}$ ($[\mathbf{t}]_\times \mathbf{x} \equiv \mathbf{t} \times \mathbf{x}$). Because the rank of $[\mathbf{t}]_\times$ is 2, the essential matrix is also a rank 2 matrix.

Equation (4) can be represented in the pixel image coordinates with the camera matrices $\mathbf{K}$ and $\mathbf{K}'$ since

$$0 = \tilde{\mathbf{x}}'^{\mathrm{T}}\mathbf{E}\tilde{\mathbf{x}} = \tilde{\mathbf{m}}'^{\mathrm{T}}\mathbf{K}'^{-\mathrm{T}}\mathbf{E}\mathbf{K}^{-\mathrm{T}}\tilde{\mathbf{m}} = \tilde{\mathbf{m}}'^{\mathrm{T}}\mathbf{F}\tilde{\mathbf{m}}. \tag{6}$$

The 3×3 matrix $\mathbf{F}$ defined above is called the fundamental matrix. The fundamental matrix is a rank 2 matrix and it is generally defined up to a scale factor. Therefore it has 7 independent parameters. For each point $\mathbf{m}$ in the first image there is a line $\mathbf{l}'_{\mathbf{m}}$ in the second image where the corresponding point $\mathbf{m}'$ lies.

The epipolar line $\mathbf{l}'_{\mathbf{m}}$ goes through the epipole $\mathbf{e}'$, because it is the projection of the line joining points $\mathbf{C}$ and $\mathbf{m}$. Likewise for each point in the second image there is a corresponding epipolar line in the first image. The lines $\mathbf{l}_{\mathbf{m}'}$ and $\mathbf{l}'_{\mathbf{m}}$ are related to the fundamental matrix by the equations [1]

$$\mathbf{l}'_{\mathbf{m}} = \mathbf{F}\tilde{\mathbf{m}}, \qquad \mathbf{l}_{\mathbf{m}'} = \mathbf{F}^{\mathrm{T}}\tilde{\mathbf{m}}'. \tag{7}$$

For the epipoles $\mathbf{e}$ and $\mathbf{e}'$ we have

$$\mathbf{F}\tilde{\mathbf{e}} = 0, \qquad \mathbf{F}^{\mathrm{T}}\tilde{\mathbf{e}}' = 0. \tag{8}$$

In order to determine the structure of a scene from two views we must solve the translation and rotation between the cameras. However, the structure can be determined only up to a similarity transformation. If we have enough point correspondences between the images, at least 7 in general position, we can estimate the fundamental matrix. If the cameras have been calibrated, the camera matrices $\mathbf{K}$ and $\mathbf{K}'$ are known, we can compute the essential matrix from (6). The translation and rotation can then be extracted from $\mathbf{E}$ up to an overall scale, which cannot be determined. Finally we have the camera projection matrices $\mathbf{P}$ and $\mathbf{P}'$ and we can back-project the image points to their corresponding 3D-points.

4 Image Matching

Suppose we have two different images of a single scene and we want to determine the structure of the scene. Before the unknown epipolar geometry can be estimated we have to first extract some point correspondences from the images. This is called image matching. Seven matches is the minimum for the F-matrix estimation, but there should be considerably more in order to minize the effect of noise and erroneous matches. We use the correlation and relaxation technique proposed by Zhang *et al.* to obtain the initial point matches [2, 3]. After we have estimated the fundamental matrix, we use the epipolar geometry to obtain a more reliable set of matches. That is produced by the multi-resolution matching technique proposed by Brandt and Heikkonen in [7]. The main ideas behind these techniques are shortly reviewed in the following two sections.

4.1 Initial Matches by Correlation and Relaxation

First, feature points corresponding to high curvature points are extracted from the views. This is done by the modified version of the Harris corner detector [2, 3]. After the corner points have been extracted from both images match candidates are obtained with the classical correlation technique. The correspondence of corner points is measured by computing the correlation between correlation windows around the corner points. The corner pairs having a correlation score high enough are considered as match candidates. Because in general there can be large image torsions, the correlation windows must be rotated and the correlation computed with several rotation angles. The angle which gives the highest correlations in average will be chosen and only the candidates with this particular angle will be considered in the relaxation process. For a corner point $\mathbf{m}_i$ in the first image several candidate points $\mathbf{m}'_j$ in the second image

may be found with the correlation technique. For choosing the best candidates Zhang proposes a relaxation procedure that utilizes the information of the corner point neighborhoods. It is based on the reasonable assumption that if $(\mathbf{m}, \mathbf{m}')$ is a good match the corner points near $\mathbf{m}$ in the first image should have a counterpart near $\mathbf{m}'$ in the second image, see [2] and [3] for details.

The correlation–relaxation technique has been implemented in the software ImageMatching available on the Internet [12]. This software can also be used to estimate the F-matrix robustly with the Least-Median-of-Squares (LMedS) method and, after that, to obtain a new set of matches containing less false matches. The result of the match extraction with the ImageMatching software is shown in Figures 3 and 4, where we have a pair of images of a house taken by the same camera from two different positions. In Figure 3 the matched points have been plotted to the first image and in Figure 4 to the second. In the latter figure the positions of the corresponding points in the first image are also shown. From the trajectories it is possible to identify the false matches. For example in Figure 4 there are two false matches in the lower right corner. There are 363 matches in total of which 10 are clearly incorrect.

4.2 Multi-Resolution Matching Utilizing Epipolar Geometry

After we have the final estimate for the fundamental matrix, we utilize it and its covariance matrix in image matching. The estimation of the F-matrix and its covariance will be concerned in Section 5. By using the epipolar geometry in image matching we try to obtain a large set of matches without false matches. In the reconstruction it is important that the essential parts of the images are densely covered with matches. This is because the reconstruction is computed by back-projecting the matched points. There are three stages in the method: (1) image rectification, (2) wavelet based matching and (3) handling of multiple match candidates. Next we give an overview, how the multi-resolution algorithm works in different stages. For a detailed study see [7]. In Figure 5 there is a plot of the final matches found by the multiresolution method. There are 801 matches and only one false was found (in the lower right corner).

4.2.1 Image Rectification

If the images to be matched are taken from widely differing viewpoints the scene may not look the same in both images. With the correlation method image torsions were taken into account by rotating correlation windows. Here we have another approach. Using the fundamental matrix we compute 2D projective tranformations $\mathbf{H}$ and $\mathbf{H}'$ which transform the epipolar lines to run horizontally in both images. Furthermore corresponding epipolar lines will have same y-coordinate after the transformation. Consequently, the disparities between the transformed and resampled images are in the x-direction only. Transformation $\mathbf{H}'$ for the second image is computed as explained in [1], Section 10.12. Then $\mathbf{H}$ is obtained from

$$\mathbf{H} = \mathbf{H}_{\mathrm{A}}\mathbf{H}'([\tilde{\mathbf{e}}']_{\times}\mathbf{F} + \mathbf{e}'\tilde{\mathbf{e}}^{\mathrm{T}}), \tag{9}$$

Figure 3: Matched points, obtained by ImageMatching, in the first image.

Figure 4: Matched points from the second (x) and the first (o) image plotted in the second image. Points are obtained by ImageMatching-software.

where $\mathbf{e}$ and $\mathbf{e}'$ are the epipoles and $\mathbf{H}_\mathrm{A}$ has three free parameters,

$$\mathbf{H}_\mathrm{A} = \begin{pmatrix} a & b & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{10}$$

Hartley and Zisserman propose that $\mathbf{H}_\mathrm{A}$ should be chosen by minimizing the disparity between transformed images [1]. This is done by minimizing the sum of squared distances

$$\sum_i d^2(\mathbf{H}\tilde{\mathbf{m}}_i, \mathbf{H}'\tilde{\mathbf{m}}'_i), \tag{11}$$

Figure 5: Final matches from the second (x) and the first (o) image plotted in the second image. Matches are obtained by the multi-resolution method.

where $(\mathbf{m}_i, \mathbf{m'}_i)$ are the initial matches. However we noticed that when the scene has considerable depth variation and the disparity between images is large this criterion can lead to severe distortion of the first rectified image. When the epipolar lines are almost horizontal better results are obtained when $\mathbf{H}_\mathrm{A}$ is chosen by minimizing the disparity between the rectified and original image. This is done by minimizing the following sum of squared distances

$$\sum_i d^2(\mathbf{H}\tilde{\mathbf{m}}_i, \tilde{\mathbf{m}}_i). \tag{12}$$

This is a linear least-squares minimization problem of three parameters, $\mathbf{H}$ is given by (9), and can be easily solved.

Image rectification, with the two different ways to compute the transformation $\mathbf{H}$, is illustrated in Figures 6 to 8. There is less distortion in the latter rectified image pair and the multi-resolution method found 436 matches. From the first rectified image pair it found only 34 matches.

4.2.2 Wavelet-Based Matching

Before the image rectification the corner points were extracted from the images with the Harris detector. After the corner extraction and rectification the images are decomposed to four resolution levels (original and three lower resolution levels) with wavelet decomposition. Then for each corner point in the first rectified image we compute the most probable point (15) on the epipolar line in the second rectified image . We take two correlation windows, a smaller template window around the corner and a bigger search window around the most probable point. Then we compute correlation between the template and the search window at each pixel shift. Correlation is computed first at the lowest resolution level. Then the local maxima, whose correlation coefficient is greater than 0.8, are extracted from the correlation image. In higher resolutions

Figure 6: Original pair of images

Figure 7: Rectified image pair. The rectifying transformation for the first image is obtained by minimizing the disparity between rectified images.

Figure 8: Rectified image pair. The rectifying transformation for the first image is obtained by minimizing the disparity between the original and rectified image.

the correlation is computed only at the neighbourhoods of these maxima. The points which have significant correlation at each resolution level are considered as match candidates. If there are several candidates the best is chosen using the disparity information contained in the fundamental matrix covariance.

4.2.3 Handling of Multiple Candidates

The uncertainty of epipolar lines can be characterized with the covariance matrix of the fundamental matrix. Brandt and Heikkonen have used the uncertainty of epipolar lines in characterizing the goodness of match candidates. The discussion here follows those represented in [7] and [2].

Let $\mathbf{m_0} = (u_0, v_0)^{\mathrm{T}}$ be a point in the first image and its covariance matrix $\mathbf{C_{m_0}}$. The corresponding epipolar line in the second image is $\mathbf{l}' = \mathbf{F}\tilde{\mathbf{m}}_\mathbf{0}$ and the first-order approximation for its covariance is obtained from [2]

$$\mathbf{C_{l'_0}} = \frac{\partial \mathbf{l'_0}}{\partial \mathbf{F}} \mathbf{C_F} \frac{\partial \mathbf{l'_0}}{\partial \mathbf{F}}^{\mathrm{T}} + \mathbf{F} \begin{pmatrix} \mathbf{C_{m_0}} & \mathbf{0}_2 \\ \mathbf{0}_2^{\mathrm{T}} & 0 \end{pmatrix} \mathbf{F} \tag{13}$$

where $\mathbf{F}$ in the first term is treated as a vector of 9 elements, $\mathbf{0}_2 = (0,0)^{\mathrm{T}}$. The first order approximation for the fundamental matrix covariance $\mathbf{C_F}$ is computed as in (26).

Because any point $\mathbf{m}' = (u', v')^{\mathrm{T}}$ on the epipolar line $\mathbf{l}'_0 = (l'_1, l'_2, l'_3)^{\mathrm{T}}$ satisfies $\tilde{\mathbf{m}}'^{\mathrm{T}}\mathbf{l}'_0 = l'_1u' + l'_2v' + l'_3 = 0$ we may use the parameterization [1] $\hat{\mathbf{l}}'_0 = (l'_1/l'_3, l'_2/l'_3)^{\mathrm{T}}$ for the epipolar line. The first order approximation for the covariance of $\hat{\mathbf{l}}'_0$ is computed in the same way as (13)

$$\mathbf{C_{\hat{l}'_0}} = \frac{\partial \hat{\mathbf{l}}'_0}{\partial \mathbf{l}'_0} \mathbf{C_{l'_0}} \frac{\partial \hat{\mathbf{l}}'_0}{\partial \mathbf{l}'_0}^{\mathrm{T}} \tag{14}$$

Now $\hat{\mathbf{l}}'_0$ may be considered as a random 2-vector having a Gaussian distribution with covariance $\mathbf{C_{\hat{l}'_0}}$. The probability distribution of possible epipolar lines given $\mathbf{m}_0$ can be understood as a two-dimensional Gaussian kernel in the dual space centered at the point $\mathbf{l}'_0$. The estimated epipolar line $\mathbf{l}'_0 = \mathbf{F}\tilde{\mathbf{m}}_0$ is therefore the mean of the distribution. (Points in the dual space correspond lines in the original space and vice versa [1].)

Let $\mathbf{m}'$ be a match candidate for $\mathbf{m}_0$. The goodness of this candidate can be characterized with the above Gaussian density in the dual space. $\mathbf{m}'$ corresponds to a line in the dual space and we use the integral of the Gaussian density over this line as a measure for the goodness of the candidate. The greater the value the better the candidate. There are also other ways to characterize the candidates as described in [7].

With $\mathbf{C_{\hat{l}'_0}}$ it is also possible to compute the most probable point $\mathbf{m}'_*$ in the second image. This is the point where the point correspondence for $\mathbf{m}'_0$ will most likely lie. $\mathbf{m}'_*$ corresponds to a line in the dual space, which has the direction of the largest variance of the Gaussian density. Therefore the point $\mathbf{m}'_*$ can be obtained from (see [7])

$$\tilde{\mathbf{m}}_* = \mathbf{l}'_0 \times \begin{pmatrix} \mathbf{v}^{\mathrm{T}} & 0 \end{pmatrix}^{\mathrm{T}}, \tag{15}$$

where $\mathbf{v}$ is the eigenvector of $\mathbf{C_{\hat{l}'_0}}$ corresponding to the largest eigenvalue.

[1] If $l'_3 = 0$ we may use $(l'_2/l'_1, l'_3/l'_1)^{\mathrm{T}}$ or $(l'_1/l'_2, l'_3/l'_2)^{\mathrm{T}}$

5 Estimation of the Fundamental Matrix

It is possible to compute the fundamental matrix from seven point correspondences. Then there may exist one or three real solutions [1]. In practice we have much more point correspondences. Because of noise in the measured image points the point correspondences $\mathbf{m}$ and $\mathbf{m}'$ will not satisfy (6) exactly. The goal of the estimation is to find the matrix that best approximates the true solution according to a given criterion. Several methods using different criterions have been proposed [1, 2, 4].

5.1 Linear Least-Squares Technique

Denoting by F_{ij} the coefficients of $\mathbf{F}$ and by $(u_i, v_i, 1)^{\mathrm{T}}$ and $(u'_i, v'_i, 1)^{\mathrm{T}}$ the homogenous image coordinates of $\mathbf{m}$ and $\mathbf{m}'$ the equation (6) may be rewritten as

$$\mathbf{a}^{\mathrm{T}}\mathbf{f}_9 = 0 \tag{16}$$

with

$$\begin{aligned} \mathbf{a} &= [u_i u'_i, v_i u'_i, u'_i, u_i v'_i, v_i v'_i, v'_i, u_i, v_i, 1] \\ \mathbf{f}_9 &= [F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{23}, F_{31}, F_{32}, F_{33}] \end{aligned} \tag{17}$$

The scalar equation (16) is linear and homogenous in $\mathbf{f}_9$. Thus eight correspondences allow us to determine $\mathbf{F}$ up to a nonzero scalar factor. With more than eight correspondences any linear least-squares technique can be used to solve $\mathbf{f}_9$ from

$$\mathbf{A}_n\mathbf{f}_9 = \mathbf{0} \tag{18}$$

where n is the number of point correspondences and

$$\mathbf{A}_n = \begin{pmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{a}_n^{\mathrm{T}} \end{pmatrix}.$$

In the singular value decomposition $\mathbf{A}_n = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$, the solution vector $\mathbf{f}_9$ that minimizes $\| \mathbf{A}_n\mathbf{f}_9 \|$ subject to $\| \mathbf{f}_9 \| = 1$ is the last column of $\mathbf{V}$ [1].

The advantage of the linear least-squares technique is that it leads to a noniterative computation method. However it is quite sensitive to noise, even with large data sets [2]. One reason to this is that the $\det(\mathbf{F}) = 0$ constraint is not necessarily satisfied. This constraint can be enforced by replacing $\mathbf{F}$ with the closest singular matrix $\hat{\mathbf{F}}$ under a Frobenius norm. $\hat{\mathbf{F}}$ is found using the singular value decomposition [2].

Another improvement to the linear least squares method is the normalization of the image coordinates as described in [1]. The normalization actually leads to a great improvement in the stability of the least squares problem and since the added complexity of the algorithm is insignificant, the normalization should be always done when linear least-squares technique is used.

5.2 Minimization of Geometric Distance

A problem with the above least-squares technique is that it does not minimize a geometric distance. The maximum likelihood estimate for the fundamental

matrix under the assumption of Gaussian image errors is obtained by minimizing the reprojection error

$$\sum_i d^2(\tilde{\mathbf{m}}_i, \tilde{\hat{\mathbf{m}}}_i') + d^2(\tilde{\mathbf{m}}_i', \tilde{\hat{\mathbf{m}}}_i') \tag{20}$$

where $\mathbf{m}_i$ and $\mathbf{m}_i'$ are the measured correspondences and $\hat{\mathbf{m}}_i$ and $\hat{\mathbf{m}}_i'$ are estimated noise free correspondences that satisfy $\tilde{\hat{\mathbf{m}}}_i'^{\mathrm{T}} \mathbf{F} \tilde{\hat{\mathbf{m}}}_i = 0$ exactly for some rank-2 matrix $\mathbf{F}$ [1]. The problem with this error function is that in order to estimate $\hat{\mathbf{m}}_i$ and $\hat{\mathbf{m}}_i'$ we have to estimate the noise free 3D-points $\hat{\mathbf{M}}_i$ at the same time with the fundamental matrix. To reduce the dimensionality of the minimization problem and to make the implementation simpler we use the first-order approximation for the reprojection error in (20). The cost function for the minimization of the first-order geometric error is

$$\sum_i \frac{(\tilde{\mathbf{m}}_i'^{\mathrm{T}} \mathbf{F} \tilde{\mathbf{m}}_i)^2}{(\mathbf{F}\tilde{\mathbf{m}}_i)_1^2 + (\mathbf{F}\tilde{\mathbf{m}}_i)_2^2 + (\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{m}}_i')_1^2 + (\mathbf{F}^{\mathrm{T}}\tilde{\mathbf{m}}_i')_2^2}, \tag{21}$$

where $(\mathbf{F}\tilde{\mathbf{m}}_i)_j^2$ represents the square of the j-th entry of the vector $\mathbf{F}\tilde{\mathbf{m}}_i$ [1]. With this cost function we do not need to estimate the 3D-coordinates $\hat{\mathbf{M}}_i$ because the noise free image coordinates $\hat{\mathbf{m}}_i$ and $\hat{\mathbf{m}}_i'$ do not appear in equation 21. Thus a minimization problem with $7 + 3n$ degrees of freedom is reduced to one with only 7 degrees of freedom.

Another possible geometric quantity to be minimized is the distance between points and their corresponding epipolar lines. This leads to a following cost function

$$\sum_i d^2(\tilde{\mathbf{m}}_i, \mathbf{F}^{\mathrm{T}}\tilde{\mathbf{m}}_i') + d^2(\tilde{\mathbf{m}}_i', \mathbf{F}\tilde{\mathbf{m}}_i) \tag{22}$$

Although this cost function is reported to give slightly inferior results to (21) it is used in the LMedS-implementation by Zhang, see Section 5.5.1.

5.3 Parameterization of the Fundamental Matrix

In order to minimize a nonlinear cost function like (21), we need a parameterization for the fundamental matrix that enforces the rank-2 constraint. We use the minimum parameterization of 7 parameters in order to compute the fundamental matrix covariance as described in next section. The problem is to find a parameterization that would be applicable with all possible camera configurations [5]. One possible parameterization is

$$\mathbf{F} = \begin{pmatrix} a & b & -ax - by \\ c & d & -cx - dy \\ -ax' - cy' & -bx' - dy' & (ax + by)x' + (cx + dy) \end{pmatrix} \tag{23}$$

where x and y are the coordinates of the first epipole and x' and y' the coordinates of the second. The scalars a,b,c and c are defined up to a scale factor, and we can obtain a minimal parameterization by dividing (23) with the largest of them. When the epipoles are at infinity above parameterization can not be used. One must thus switch between different choices of the two rows and two columns of the F-matrix to use as the basis. Along with four choices of which

of a,b,c and c to set to 1, there are a total of 36 maps to parameterize the fundamental matrix.

However, in [4] Zhang and Loop proposed a technique for estimating the fundamental matrix with a single parameterization like that in (23). The technique needs an initial estimate for the fundamental matrix and we use that given by the LMedS-method. The idea behind the technique of Zhang and Loop is to find a projective transformation in each image such that in the transformed image space the first element of the fundamental matrix has the largest value and the epipoles are not at infinity. The fundamental matrix is estimated in the transformed space and then transformed back to the original space. The projective transformations are found by using the initial F-matrix estimate as explained in [4].

5.4 The Covariance of the Fundamental Matrix

The fundamental matrix can be thought as a random vector $\mathbf{f}_7$ of $\mathbb{R}^7$ whose mean is the exact value we are looking for. Each estimation is a sample of $\mathbf{f}_7$ and its uncertainty can be characterized with the covariance matrix $\mathbf{C}_{\mathbf{f}_7}$. Csurka *et al.* proposed in [5], how an approximation for $\mathbf{C}_{\mathbf{f}_7}$ can be computed. In Section 4.2.3 we explained how the uncertainty of the fundamental matrix can be utilized in image matching.

As explained above in Section 5.2, we estimate the fundamental matrix by minimizing a cost function ϵ which is of the form [2]

$$\epsilon(\hat{\mathbf{m}}, \mathbf{f}_7) = \sum_{i=1}^{n} \epsilon_i^2(\mathbf{m}_i, \mathbf{m'}_i, \mathbf{f}_7), \tag{24}$$

where $\hat{\mathbf{m}} = (\mathbf{m}_1, \mathbf{m}'_1, \ldots, \mathbf{m}_n, \mathbf{m}'_n)$. Here n is the number of matches $(\mathbf{m}_i, \mathbf{m}'_i)$ in the images. In [5] it is shown that an approximation for $\mathbf{C}_{\mathbf{f}_7}$ is obtained from

$$\mathbf{C}_{\mathbf{f}_7} = \frac{2\epsilon_{\min}}{n-7}\mathbf{H}^{-\mathrm{T}}, \tag{25}$$

where $\epsilon_{\min}$ is the value of ϵ at the minimum and $\mathbf{H} = \frac{\partial^2 \epsilon}{\partial \mathbf{f}_7^2}$ the Hessian of ϵ with respect to $\mathbf{f}_7$ at the minimum. $\mathbf{C}_{\mathbf{f}_7}$ is thus computed using $\epsilon_{\min}$ and $\mathbf{H}$ which are given as a by-product of the fundamental matrix estimation. When the fundamental matrix is considered as a vector of 9 elements the first order approximation for the 9×9 covariance matrix $\mathbf{C}_{\mathbf{F}}$ is computed from

$$\mathbf{C}_{\mathbf{F}} = \frac{d\mathbf{F}}{d\mathbf{f}_7}\mathbf{C}_{\mathbf{f}_7}\frac{d\mathbf{F}}{d\mathbf{f}_7}^{\mathrm{T}}. \tag{26}$$

The 9×7 Jacobian $d\mathbf{F}/d\mathbf{f}_7$ is computed by differentiating the nine elements of (23) with respect to the seven parameters (parameter a is set to 1).

5.5 Robust Methods

In practice the point correspondences that are used to estimate the fundamental matrix contain not only noise but also false matches and badly localized matches.

[2] ϵ_i is called the residual of match i and ϵ is the sum of squared residuals.

Especially the presence of false matches will completely spoil the estimation process if we directly apply the methods described above. The methods, that are not so easily affected by false matches, are called robust. There are different robust methods for the fundamental matrix estimation, for example the M-estimators, the Least-Median-of-Squares (LMedS) method and the Bayesian method [2, 8]. We used the LMedS-method to obtain an initial estimate for the fundamental matrix and then the Bayesian method for the final estimate.

5.5.1 Least-Median-of-Squares-Method

The LMedS-method for robust fundamental matrix estimation is implemented in the ImageMatching software [12] and described in [2]. The method tries to find the fundamental matrix that minimizes the median of squared residuals, not the sum of squared residuals as described in Section 5.2.

Given n point correspondences, $(\mathbf{m}_i, \mathbf{m}'_i)$, a Monte Carlo type technique is used to draw m random subsamples of 7 different point correspondences. For each subsample, indexed by J, the fundamental matrix $\mathbf{F}_J$ is determined. Then for each $\mathbf{F}_J$ the median of the squared residuals, M_J, is computed from

$$M_J = \underset{i=1,\ldots,n}{\text{median}}[d^2(\tilde{\mathbf{m}}_i, \mathbf{F}_J^{\mathrm{T}}\tilde{\mathbf{m}}'_i) + d^2(\tilde{\mathbf{m}}'_i, \mathbf{F}_J\tilde{\mathbf{m}}_i)] \tag{27}$$

Note that here the residual is the distance between a point and the corresponding epipolar line. The LMedS-estimate is the $\mathbf{F}_J$ that minimizes M_J.

5.5.2 Bayesian Method

After we have the initial estimate for the fundamental matrix we use the Bayesian weighting principle proposed in [8] to obtain the final estimate. The idea behind the Bayesian weighting is to fit two normal distributions to the residual distribution with the maximum likelihood method. One distribution is for the relevant matches and the other for the false matches. The assumption that the residuals of relevant matches are normally distributed is reasonable because the residual is a physical quantity (in our case the first order approximation for the reprojection error) and the image errors are caused by many independent sources. The justification of the normal distribution assumption for the false matches is not so obvious, but it seems to give quite good results. The following mathematical formulation of the Bayesian weighting principle was first given in [8].

Given a fundamental matrix estimate $\hat{\mathbf{F}}$, let the random variables $\epsilon_{\mathrm{r},\hat{\mathbf{F}}}$ and $\epsilon_{\mathrm{f},\hat{\mathbf{F}}}$ correspond to the residuals of relevant and false matches. The variables are assumed to follow normal distribution, i.e. $\epsilon_{\mathrm{r},\hat{\mathbf{F}}} \sim N(\mu_{\mathrm{r},\hat{\mathbf{F}}}, \sigma_{\mathrm{f},\hat{\mathbf{F}}})$ and $\epsilon_{\mathrm{f},\hat{\mathbf{F}}} \sim N(\mu_{\mathrm{f},\hat{\mathbf{F}}}, \sigma_{\mathrm{f},\hat{\mathbf{F}}})$. The corresponding density functions of the residuals are $p(\epsilon|S_\mathrm{r})$ and $p(\epsilon|S_\mathrm{f})$, where S_r and S_f are the sets of relevant and false matches.[3] The density function of the residual of the whole set of matches is

$$p(\epsilon) = P_\mathrm{r}p(\epsilon|S_\mathrm{r}) + P_\mathrm{f}p(\epsilon|S_\mathrm{f}) \tag{28}$$

where P_r and $P_\mathrm{f} = 1 - P_\mathrm{r}$ are a priori probabilities of the relevant and false matches. The parameters of (28) can be obtained by maximizing the following

[3]The subscript $\hat{\mathbf{F}}$ is omitted for clarity

likelihood function

$$L = \prod_{i=1}^{n} p(\epsilon_i | P_{\mathrm{r}}, \mu_{\mathrm{r}}, \sigma_{\mathrm{r}}, \mu_{\mathrm{f}}, \sigma_{\mathrm{f}}), \tag{29}$$

where n is the number of matches. This optimization problem is relatively easily solved with standard optimization tools. When the maximum likelihood estimate for the distribution parameters is found, the Bayes rule can be used to compute a posteriori probability

$$P(S_{\mathrm{r}}|\epsilon) = \frac{P_{\mathrm{r}} p(\epsilon|S_r)}{P_{\mathrm{r}} p(\epsilon|S_{\mathrm{r}}) + P_{\mathrm{f}} p(\epsilon|S_{\mathrm{f}})}. \tag{30}$$

$P(S_{\mathrm{r}}|\epsilon_i)$ tells the probability for a match with residual ϵ_i to be relevant. It is also obvious that the relevant matches should have a greater weight in the fundamental matrix estimation. Therefore Brandt and Heikkonen suggested that when the fundamental matrix is estimated by minimizing a cost function like (21) the residuals should be weighted with the posteriori probabilities. Thus a new estimate for the fundamental matrix is computed by weighting the new residual by a posteriori probability of the old residual, i.e.

$$\min_{\hat{\mathbf{F}}_{\mathrm{new}}} \sum_i P(S_{\mathrm{r}}|\epsilon_{i,\hat{\mathbf{F}}}) \epsilon^2_{i,\hat{\mathbf{F}}_{\mathrm{new}}}. \tag{31}$$

This leads to an iterative process, where a new fundamental matrix estimate is computed until the Frobenius norm of the difference between two subsequent matrices is unchanged.

In our implementation we use the LMedS-estimate as an initial value for the Bayesian method. The fundamental matrix is estimated using the minimal parameterization described in Section 5.3. The Bayesian weighting seems to slightly improve the estimate given by the LMedS-method. Another advantage of the Bayesian method is that the computation of the fundamental matrix covariance is possible as described in Section 5.4. This is not directly possible with the LMedS-method.

We tested our implementation with four image pairs for which the true fundamental matrix was known through camera calibration. The test images were obtained from the INRIA-Syntim database [11]. In Table 1 there are the average differences between the true and estimated fundamental matrices (the difference is computed as explained in [2] pp. 116). For three image pairs the Bayesian method improves the result of LMedS, but for one pair the difference between the true and estimated matrices is large. One explanation could be that the number of matches was relatively low in this pair. The result of the fundamental matrix estimation for the first image pair (House) is illustrated in Figures 9-11. In Figure 9 is the first image and three chosen points. The epipolar lines corresponding to these points are plotted in Figures 10 and 11.

The result of estimating the fundamental matrix and its covariance is also illustrated in Figures 12 and 13. Three points are chosen from the first image and the corresponding epipolar lines are plotted to the second image. The most probable point (15) is also plotted to the second image. The hyperbolas on both sides of the epipolar lines are the 97 % confidence intervals computed from the covariance matrix [2]. The small error bounds indicate that the estimation has succeeded.

Figure 9: The first image and three chosen points. Copyright of the image pair belongs to INRIA Syntim.

Figure 10: The second image and epipolar lines. Solid line from LMedS-estimate, dashed from calibrated F-matrix.

Figure 11: The second image and epipolar lines. Solid line from Bayes-estimate, dashed from calibrated F-matrix.

image pair	LMedS	Bayes
House	8.17	5.89
Color	2.91	8.25
Sport	3.63	3.35
Tot	7.00	5.71

Table 1: The result of F-matrix estimation with the LMedS- and Bayesian-methods

Figure 12: Three points chosen from the first image and marked with 'x' .

Figure 13: The corresponding epipolar lines and their error bounds. The most probable position of the corresponding point is marked with 'x'.

6 Structure Computation

When the fundamental matrix is estimated it is possible to compute the camera projection matrices $\mathbf{P}$ and $\mathbf{P}'$. They can be determined up to a similarity transformation if the camera calibration matrices $\mathbf{K}$ and $\mathbf{K}'$ are known. Otherwise the camera projection matrices can be determined only up to a projective transformation and a projective reconstruction is possible. The projective reconstruction and the original scene are projectively equivalent. For example angles, parallelism and ratios of lengths and areas may not be preserved in the projective reconstruction but concurrency and collinearity are preserved.

If the camera calibration matrices are known, an estimate $\mathbf{E}_0$ for the essential matrix may be computed from (6) after the fundamental matrix is estimated. The actual essential matrix has only 5 degrees of freedom (three for translation and three for rotation minus one for overall scale), whereas the fundamental matrix has seven. The additional constraints for the essential matrix cause its two nonzero singular values to be equal [1]. Our estimate $\mathbf{E}_0$, obtained from (6), does not necessarily satisfy the additional constraints. Therefore we set the two singular values of $\mathbf{E}_0$ equal to their mean and obtain a new estimate $\mathbf{E}$ that is the best approximation to $\mathbf{E}_0$, in the sense of Frobenius norm, satisfying the constraints for an essential matrix [9].

The rotation $\mathbf{R}$ and translation $\mathbf{t}$ between the cameras can be determined from the essential matrix $\mathbf{E}$. This can be done easily with singular value decomposition and is described in [1]. We choose the first camera coordinate frame as the world coordinate frame and fix the overall scale by setting the distance between the camera centres to 1, i.e. $||\mathbf{t}|| = 1$. Then we have the camera projection matrices $\mathbf{P} = \mathbf{K}(\mathbf{I}\ \mathbf{0})$ and $\mathbf{P}' = \mathbf{K}'(\mathbf{R}\ \mathbf{t})$ up to a similarity transformation.

Next we compute the three dimensional correspondences for each match in the images. Since there are errors in the image measurements there will not exist a point $\mathbf{M}$ that exactly satisfies $\tilde{\mathbf{m}} = \mathbf{P}\tilde{\mathbf{M}}$ and $\tilde{\mathbf{m}}' = \mathbf{P}'\tilde{\mathbf{M}}$. Therefore we find a point $\hat{\mathbf{M}}$ that minimizes the sum of squared reprojection errors

$$\rho(\mathbf{m}, \mathbf{m}') = d^2(\mathbf{m}, \hat{\mathbf{m}}') + d^2(\mathbf{m}', \hat{\mathbf{m}}'), \tag{32}$$

where $\hat{\mathbf{m}}$ and $\hat{\mathbf{m}}'$ are the images of $\hat{\mathbf{M}}$ and satisfy $\tilde{\hat{\mathbf{m}}}'^{\mathrm{T}}\mathbf{F}\tilde{\hat{\mathbf{m}}} = 0$ exactly. The minimization of reprojection error was already discussed in Section 5.2, but here we assume the estimated fundamental matrix and camera projection matrices to be error free. We minimize (32) by first finding the points $\hat{\mathbf{m}}$ and $\hat{\mathbf{m}}'$ (for details see [1], Algorithm 11.1). We find two corresponding epipolar lines l and l' that minimize

$$d^2(\mathbf{m}, \mathrm{l}) + d^2(\mathbf{m}', \mathrm{l}'). \tag{33}$$

$\hat{\mathbf{m}}$ and $\hat{\mathbf{m}}'$ are the points on the epipolar lines l and l' that are closest to points $\mathbf{m}$ and $\mathbf{m}'$. Minimization of (33) can be reduced to finding the real roots of a polynomial of degree 6 since the epipolar lines l and l' can be parameterized with a single parameter when the fundamental matrix is known.

When solving $\hat{\mathbf{M}}$ we use cross product to eliminate the homogenous scale factors. By writing out equations $\tilde{\hat{\mathbf{m}}} \times (\mathbf{P}\tilde{\hat{\mathbf{M}}}) = 0$ and $\tilde{\hat{\mathbf{m}}}' \times (\mathbf{P}'\tilde{\hat{\mathbf{M}}}) = 0$ we get two linearly independent equations from each. These can be written in the form

$$\mathbf{A}\tilde{\hat{\mathbf{M}}} = \mathbf{0}, \tag{34}$$

where

$$\mathbf{A} = \begin{pmatrix} u\mathbf{p}_3^{\mathrm{T}} - \mathbf{p}_1^{\mathrm{T}} \\ v\mathbf{p}_3^{\mathrm{T}} - \mathbf{p}_2^{\mathrm{T}} \\ u'\mathbf{p}_3^{'\mathrm{T}} - \mathbf{p}_1^{'\mathrm{T}} \\ v'\mathbf{p}_3^{'\mathrm{T}} - \mathbf{p}_2^{'\mathrm{T}} \end{pmatrix}.$$

Here $\mathbf{p}_i^{\mathrm{T}}$ and $\mathbf{p}_i^{\mathrm{T}}$ are the rows of $\mathbf{P}$ and $\mathbf{P}'$; (u, v) and (u', v') are the coordinates of $\hat{\mathbf{m}}$ and $\hat{\mathbf{m}}'$. The system of equations (34) is redundant and $\tilde{\tilde{\mathbf{M}}}$ can be solved as the right null vector of $\mathbf{A}$. Thus we obtain the three dimensional coordinates corresponding to a match $(\mathbf{m}, \mathbf{m}')$.

The results of computing the scene structure as described above are shown in Figures 14 to 17. The images in Figures 12 and 13 were taken with a digital camera, whose internal parameters are known. The result of reconstruction is illustrated in Figures 14 and 15. The reconstruction was computed after the effect of lens distortion was removed from the images. This was possible because the distortion parameters for the camera were known. The median and average value of the reprojection error among all matches were 1.33 and 1.34.

The image pair *Sport*, obtained from INRIA-Syntim database with the calibration data and already introduced in Table 1, is shown in Figure 16. We computed two reconstructions of it, one with the calibrated F-matrix and the other with the estimated F-matrix, Figure 17. The median and average of the reprojection error were 0.31 and 0.50 with the calibrated F-matrix and 2.05 and 1.90 with the estimated F-matrix.

The real cameras are not exactly pinhole cameras. The deviation from the pinhole model is called lens distortion and can be observed in Figure 17. The back wall of the latter reconstruction is not straight although it evidently should be. Curiously the reconstruction with the estimated F-matrix seems better. However, because neither the orientation of walls and the optical axis of the first camera nor the accuracy of camera calibration are known it is difficult to draw any conclusions.

7 Conclusions

In this report we have reviewed different methods for the various tasks in automatic scene reconstruction. As shown in the previous section some quite interesting results were obtained. We discussed the fundamental matrix estimation a lot because it is a crucial part in the reconstruction process. The proposed methods are capable to robust estimation of the epipolar geometry. However, if the image correspondences do not contain enough depth variation or there are too few of them, the estimation may fail and the reconstruction is impossible. Thus the performance of the automatic reconstruction depends also on the scene. The result is better with richly textured scenes, because then the number of found correspondences between the views is greater in general.

The main drawback of our approach is the sparseness of the recovered depth information. When single points are matched between the views, the depth information can be computed only for the corresponding world points and the result is often a very sparse reconstruction. The advantage of our approach is that it is computationally fast. If one wants to compute a dense reconstruction a possible approach would be to first compute dense disparity maps for the

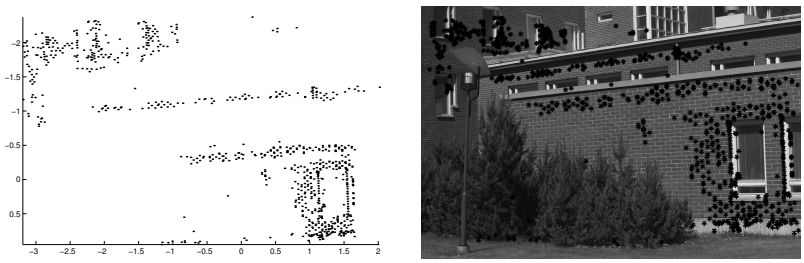

Figure 14:
Left: Front view of the reconstructed 3D-points.
Right: The measured image points and projected 3D-points plotted to the same image.

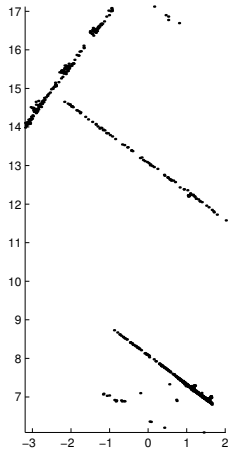

Figure 15: Top view of the reconstructed 3D-points

Figure 16: Image pair *Sport* and the matches obtained by the multi-resolution method. Copyright of the image pair belongs to INRIA-Syntim.

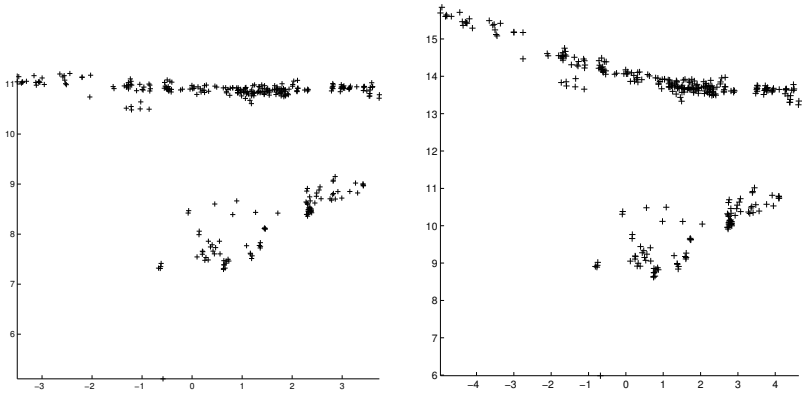

Figure 17: Top views of the two reconstructions from image pair *Sport*. On the left is the reconstruction computed by using the estimated F-matrix and on the right is the reconstruction computed by using the calibrated F-matrix.

images. These algorithms are computationally heavier, but for example Alvarez et al. have reported good results with this approach [10].

References

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

[2] G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition*, Kluwer, 1996.

[3] Z. Zhang, R. Deriche, O. Faugeras and Q.-T. Luong, *A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry*, Technical Report INRIA 2273, Sophia-Antipolis, France, 1994.

[4] Z. Zhang and C. Loop, "Estimating the fundamental matrix by transforming image points in projective space," *CVIU*, vol.82, no. 2, pp. 174–180, 2001.

[5] G. Csurka, C. Zeller, Z. Zhang and O. Faugeras, "Characterizing the uncertainty of the fundamental matrix," *CVIU*, vol. 68, no. 1, pp. 18–36, 1997.

[6] M. Johansson, F. Kahl and A. Heyden, "VISIRE: From Video to VRML," in *Proc. Scandinavian Conference on Image Analysis*, Bergen, Norway, 2001, pp. 446–453.

[7] S. Brandt and J. Heikkonen, "Multi-Resolution matching of uncalibrated images utilizing epipolar geometry and its uncertainty," in *Proc. ICIP*, Thessaloniki, Greece, 2001, pp. 213–216.

[8] S. Brandt and J. Heikkonen, "A Bayesian weighting principle for the fundamental matrix estimation,", *PRL*, vol. 21, no. 12, pp. 1081–1092, 2000.

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

[10] L. Alvarez, R. Deriche, J. Sánchez and J. Weickert, *Dense Disparity Map Estimation Respecting Image Discontinuities: A PDE and Scale-Space Based Approach*, Technical Report INRIA 3874, Sophia-Antipolis, France, 1994.

[11] INRIA-Syntim Image Database, <http://www-rocq.inria.fr/mirages/SYNTIM_OLD/analyse/paires-eng.html>.

[12] ImageMatching-software, <http://www-sop.inria.fr/robotvis/personnel/zzhang/softwares.html>.