Motion and Structure from Long Stereo Image Sequences

Ning Cui\textsuperscript{\dagger} Juyang Weng*\textsuperscript{\ddagger} Paul Cohen\textsuperscript{\dagger}

\textsuperscript{\dagger} Dept. of Electrical Engineering
Ecole Polytechnique de Montreal
Montreal PQ, H3C 3A7, Canada

*Centre de Recherche Informatique de Montreal
3744 Rue Jean-Brilliant
Montreal PQ, H3P 1P1, Canada

Abstract

The treatment of uncertainty in 3-D points determined by stereo triangulation affects significantly the reliability of the estimated 3-D motion and structure. In this paper, we first address this issue in the estimation of inter-frame motion. Two methods are discussed, a closed-form matrix-weighted least-squares solution and an iterative optimal solution. Using each of these two methods for two-view analysis, we proceed to deal with long image sequences. A recursive-batch approach is adopted to fuse multiple stereo views in order to achieve higher performance without suffering from excessive computational cost. Experiments with a real stereo image sequence have been presented to show the performance of the algorithms.

1 Introduction

The process of sensing and understanding the visual world may involve motion between image sensors and the scene. We consider in this paper a stereo system, with a fixed but general stereo configuration, moving in a static environment. The evolution of the viewing direction and position provides multiple images of the same part of a scene. Our objective is to estimate the unknown motion of the stereo system and the structure of the scene, by fusing partially overlapping views so that a more accurate and consistent description of the visual world can be derived.

Since real images are always contaminated by noise, the estimation of 3-D transformation and the fusion of multiple views should take into account the varying uncertainties in observations. Estimation based on proper error modeling can potentially improve the robustness of the estimates. Several researchers have dealt with this subject. Kiang et al\cite{2} modeled the uncertainties in observations using scalar weights. Simpler scalar weights, which are inversely proportional to depth, have also been used by Moravec in\cite{4}. But neither unweighted nor scalar weighted closed-form least-squares solution can properly account for the reliability differences of the components in 3-D points. In this paper, we introduce a closed-form matrix-weighted least-squares solution for motion parameters from 3-D point correspondences of two stereo image pairs. To further improve this result, an iterative optimization process is developed, using a low-dimensional search space.

In the case of long image sequences, we use a recursive-batch approach to estimate the motion and the 3-D structure. By this approach, observation data are subdivided into groups and estimation is done in a sequential fashion among groups of data. Within each data group, estimation is done in a batch fashion. In this way, advantages of sequential processing are kept and the performance of the algorithm is drastically improved compared to direct nonlinear iterated Kalman filtering\cite{6}. In addition, if there exist possible outliers in the data to be processed, the method of robust statistics\cite{10} can then be invoked to detect the outliers and suppress their harmful effects. We also investigate the representation of 3-D points in local and world coordinate systems. In self-guided navigation, the use of a local system is preferable. However, if the navigation has to refer to a map or to construct a global 3-D map of the sensed world, a representation in a world coordinate system is more appropriate. In order to assess the performance of the recursive-batch approach, experiments with a real stereo image sequence have been conducted.

2 Closed-Form Solution

Let the camera-centered coordinate system be fixed with the camera, the origin of the system coinciding with the optical center of the camera and the Z-axis being aligned with the optical axis pointing towards
the scene. We choose the coordinate system centered at the left camera as the local coordinate system. The orientation and position of the right camera, with respect to the left camera, is specified by a rotation matrix $M$, and a translation vector $B$. A vector $x_r = (x_r, y_r, z_r)^T$ represented in the right-camera-centered system is related to $x_l = (x_l, y_l, z_l)^T$ in the local system by
\[ x_l = M x_r + B, \]
where $M$ and $B$ are determined through camera calibration. We can define a normalized pin-hole camera model in which the focal length is equal to 1 and the image plane is at $z = 1$. For a point $x = (x, y, z)^T$, the corresponding image vector of the point $x$ is: $X = x/z$. The first two components of the right camera vector in the local coordinate system. The corresponding image vector of the point is related to $X$ is defined as the image coordinates of the point.

Thus, the depths $z_l$ and $z_r$ of a point, in the local and the right-camera-centered systems, respectively, can be determined from (1):
\[ X_{zl} = M X_{zr} + B. \] (2)

Equation (2) expresses the epipolar constraint: $X_l$, $MX_r$, and $B$ are coplanar. In the presence of noise, the epipolar constraint may be violated. Let the noise-contaminated projections in the left and right images be $\tilde{u}_l$ and $\tilde{u}_r$, respectively. We have $\tilde{u}_l = u_l + \delta_l$, $\tilde{u}_r = u_r + \delta_r$, where $\delta_l$ and $\delta_r$ are additive noise vectors in the image plane (from calibration, quantization, matching, etc.).

In this paper, the principle of the minimum variance estimation [3] [5] [1] and its nonlinear extension [10] will be used to formulate several objective functions.

2.1 Optimal Determination of a 3-D Point from a Pair of Noisy Stereo Projections

We assume that the correlation between the image error components is negligible. We also assume the same error variance in the different components. Let the estimated 3-D point $\hat{x}$ have its projections $u_l(\hat{x})$ and $u_r(\hat{x})$ in the left and right images, respectively. According to the minimum variance estimation principle, the optimal 3-D point $\hat{x}$ should minimize
\[ \| \tilde{u}_l - u_l(\hat{x}) \|^2 + \| \tilde{u}_r - u_r(\hat{x}) \|^2. \] (3)

This is a nonlinear minimization problem. The best linear estimate (in the least-squares sense) is obtained by solving for $z_l$ and $z_r$, which minimize:
\[ \| M \hat{x}_r + B - \hat{x}_l \|. \] (4)

The estimated $\hat{x}$ is then determined by
\[ \hat{x} = (M \hat{x}_r + B + \hat{x}_l) / 2. \] (5)

From this approximate solution, a few iterations can be performed to minimize (3). Based on (5), the error covariance matrix of $x$ can be easily derived [10]:
\[ \Gamma_x = \frac{\partial c(\hat{u}_l, \hat{u}_r)}{\partial u_l} \Gamma_u \frac{\partial c(\hat{u}_l, \hat{u}_r)^T}{\partial u_l} + \frac{\partial c(\hat{u}_l, \hat{u}_r)}{\partial u_r} \Gamma_u \frac{\partial c(\hat{u}_l, \hat{u}_r)^T}{\partial u_r}. \] (6)

2.2 Determination of Interframe Motion using a Matrix-Weighted Objective Function

The objective is to determine the rotation matrix $R$ and the translation vector $T$ from a sequence of point correspondences: $(x_l, x_r')$. Using the estimated 3-D positions $\hat{x}_l' = x_l' + \delta_l'$, and $\hat{x}_r = x_r + \delta_x$, we have
\[ \hat{x}_l' = R \hat{x}_l + T + \delta_l, \] (7)
where
\[ \delta_l = \hat{x}_l - R \hat{x}_l. \] (8)

Suppose that the errors in the observed points are uncorrelated between time $t_0$ and $t_1$. It follows from (8) that the residual vector $\delta_l$ has a covariance matrix
\[ \Gamma_l = \mathbf{E} \delta_l \delta_l^T = \Gamma_{x_l} + R \Gamma_{x_l} R^{-1}. \] (9)

The motion parameters should be determined by minimizing:
\[ \sum_{i=1}^n (R \hat{x}_l + T - \hat{x}_l') \Gamma_i^{-1} (R \hat{x}_l + T - \hat{x}_l'), \] (10)
where $n$ is the number of point correspondences. Letting a denotes a three dimensional vector consisting of the three independent parameters of the rotation matrix $R$, the expression (10) is a nonlinear function of a six-dimensional parameter vector $m = (a^T, T^T)^T$. With a small rotation, the rotation matrix is roughly equal to an identity matrix, $R \approx I$, and the weighting matrix in (9) does not depend very much on $R$ (if this is not the case, a closed-form scalar weighted least-squares solution [10] can be used to get an estimated $R$ first). So, the weighting matrix can be approximated by
\[ \Gamma_i = \Gamma_{x_i} + \Gamma_{x_i}. \] (11)

An approximated closed-form solution for the motion parameters $\hat{m}$, based on a simplified matrix-weighted objective function is thus obtained [10].
3 Iterative Optimal Solution from Two Stereo Pairs

Let the true (unknown) 3-D position of the i-th point at time \( t_k \), in the local coordinate system, be \( x_{i,k} \) and the collection of all such points at time \( t_k \) be \( X_{.,k} \). Given two stereo image pairs at instants \( t_0 \) and \( t_1 \), the parameter vector \( t \) to be estimated consists of the interframe motion parameter vector \( m \) and the structure of the points \( X_{.,0} \) at time \( t_0 \): \( t = (m^T, X_{.,0}^T) \) (equivalently, we can consider the structure at time \( t_1 \)). The set of image observation vectors consists of all noise corrupted version \( u_{i,j,k} \) of \( u_{i,j,k} \) (\( k \) is the time index, \( j=1 \) designates the left image and \( j=2 \) for the right image). The objective function to be iteratively minimized is

\[
 f(m, X_{.,0}) = \sum_{i=1}^{n} \sum_{j=1}^{2} \sigma^{-2} ||u_{i,j,k}(m, x_{i,k}) - \hat{u}_{i,j,k}||^2
\]

where \( u_{i,j,k}(m, x_{i,k}) \) is the noise-free projection computed from \( m \) and \( x_{i,k} \), and \( \sigma^2 \) is the image noise variance.

The above model is a natural model with two stereo pairs. Another alternative model, although less natural, is useful for our recursive estimation from long image sequences. If \( x_{i,0} \) is considered as an “observation” at time \( t_0 \), which is determined by the method in Subsection 2.1, the objective function to be minimized becomes

\[
 f(m, X_{.,0}) = \sum_{i=1}^{n} \{(x_{i,0} - \hat{x}_{i,0})^T \Gamma_{X_{.,0}^{-1}}(x_{i,0} - \hat{x}_{i,0}) + 2 \sigma^{-2} ||u_{i,j,1}(m, x_{i,0}) - \hat{u}_{i,j,1}||^2 \}
\]

Instead of performing a computationally expensive direct minimization of the above objective function, we reduce the dimension of the parameter space first. Since the objective functions are continuous, we have

\[
 f(m, \hat{x}_{.,0}) = \min_{m, X_{.,0}} f(m, X_{.,0}) = \min_{m} g(m, \hat{x}_{.,0})
\]

where

\[
 g(m, \hat{x}_{.,0}) = \min_{X_{.,0}} f(m, X_{.,0})
\]

is the smallest “cost”, computed by choosing the “best” structure \( X_{.,0} \), with a given motion parameter vector \( m \). This means that the space \( (m, X_{.,0}) \) is decomposed into two subspaces, corresponding to \( m \) and \( X_{.,0} \), respectively. In the subspace of \( m \), an iterative algorithm (e.g., Levenberg-Marquardt method or conjugate gradient method) is used. In the subspace of \( X_{.,0} \), an non-iterative method is used that gives the best \( x_{.,0} \) for any given \( m \). According to the decomposition shown in (13), the search space in

\[
 \min_{m} g(m, \hat{x}_{.,0})
\]

is just the 6-dimensional motion parameter space.

With a good very initial solution of \( m \) provided by the matrix-weighted solution discussed in Section 2, very few iterations are needed to reach the optimal solution.

Now we consider how to compute the best \( x_{.,0} \) in (14) without resorting to iterations. We have two sample data for the same parameter vector \( x_{i,0} \). One is \( p = \hat{x}_{i,0} \) with error covariance matrix \( \Gamma_p = \Gamma_{X_{.,0}} \) and the other is the point moved back from \( \hat{x}_{i,1} \): \( q = R^T \{ \hat{x}_{i,1} - T \} \) with the error covariance matrix \( \Gamma_q = R^T \Gamma_{X_{.,1}} R \). According to the minimum variance estimation principle, the optimal \( x_{i,0} \) should minimize

\[
 \{x_{i,0} - p\}^T \Gamma_p^{-1} \{x_{i,0} - p\} + \{x_{i,0} - q\}^T \Gamma_q^{-1} \{x_{i,0} - q\}
\]

Then for a given motion, the optimal \( x_{i,0}^* \) that minimizes (15) is directly computed (without iterations) by

\[
 x_{i,0}^* = p + \Gamma_p \{ \Gamma_p + \Gamma_q \}^{-1} \{q - p\}
\]

The expressions for the error covariance matrices of the estimated motion and 3-D structure can be easily derived based on the nonlinear extension of the minimum variance estimation [10].

4 A Recursive-Batch Approach for Long Image Sequences

Let the world coordinate system coincide with the local coordinate system at time \( t_0 \). In the world system, the attitude of the camera system at time \( t_k \) can be reached by moving camera from its original attitude at time \( t_0 \) by a rotation \( (R_k) \) about the origin followed by a translation \( (T_k) \).

The position of a scene point, \( x_{i,k} \), in the local coordinate system at \( t_k \), and its position \( x_{i,0} \) in the world coordinate system are related by:

\[
 x_{i,0} = R_k x_{i,k} + T_k
\]

If we define

\[
 R_{k+1,k} = R_k^{-1} R_{k+1}
\]

and

\[
 T_{k+1,k} = R_k^{-1} (T_k - T_{k+1})
\]

77
It can easily be shown that:

\[ x_{i,k+1} = R_{k+1,k} x_{i,k} + T_{k+1,k}. \]  

(20)

Since \( x_{i,k} \) and \( x_{i,k+1} \) are represented in the local coordinates system at \( t_k \) and \( t_{k+1} \), respectively, \( R_{k+1,k} \) and \( T_{k+1,k} \) express the inter-frame motion. They correspond to the motion parameters estimated in Section 2 and Section 3.

According to the discussion in the previous sections, the estimation of structure and interframe motions are closely related. The accuracy of the estimated structure influences the accuracy of the estimated motion parameters, and vice versa. Due to this type of interaction, a high estimation accuracy can be obtained if all the image frames are processed in a batch fashion. However, this is a computationally prohibitive task if the image sequence is long. In order to achieve good performance without suffering from excessive computational cost, we need batch processing only for those data that have considerable interactions. The above observations motivate our recursive-batch approach.

The model represented by the objective function (13) is very suitable for recursive-batch updating. With each new pair of stereo images, interframe motion is computed directly from the closed-form solution, and then the parameters are further optimized by the method presented in Section 3. Fusion of the estimated structure accumulated up to previous instant with the new stereo pair updates the structure for the next instant.

Once the interframe motion parameter vector \( m_{k+1,k} \) is estimated based on points \( \{ x_{i,k} \} \) before motion and \( \{ x_{i,k+1} \} \) after motion, we compute the covariance matrix of the motion parameter vector \( \Gamma m_{k+1,k} \) and covariance matrices of 3-D structure \( \Gamma x_{k+1} \) according to the procedure described in Section 3.

Using (18) and (19), we can update the global attitude of the stereo system from its previous position and current interframe motion, and compute its error covariance matrix.

5 Experiments

In our experiments with long image sequence, the stereo system consisted of two f=8.5mm CCD cameras, mounted on the tip of a high-precision six-joint robot arm. Each digital image grabbed from each camera has 480x512 pixels. After each stereo image pair was grabbed the manipulator was imposed a rotation by 2.25° (this is the only motion ground truth available) with a vertical revolute joint. The stereo cameras were calibrated in order to compensate for lens distortion, and compute internal and external parameters of each camera [9]. The relative configuration between the stereo cameras (\( M \) and \( B \)) was directly computed from the external parameters of the stereo cameras. The rotation angle was 11.86°, the rotation axis was \((0.99, 0.01, 0.04)^T\) and the translation vector \( B \) was \((0.00, 0.09, 0.01)^T\) (units are meters). Due to the relative configuration of the stereo cameras and the depth range of the scene, the common field of view of the stereo cameras was about 362-pixel wide. The average length of the vectors of the stereo displacement field was around 80 pixels, while the average length of the vectors of the temporal displacement field was around 50 pixels. Figure 1 shows the first and the last left images of the stereo sequence. The stereo image sequence used in the experiments consists of 10 consecutive stereo image pairs. The observed scene was approximately 1.2m far from the cameras.

The algorithm described in [7] was used to compute stereo and temporal matchings. The image matching algorithm establishes correspondence for each pixel. Since matching is more accurate where the image texture is abundant, the feature points used for motion computation consisted of a set of manually selected corner points in the first left image. These corners were then tracked automatically in consecutive left images by temporal matching, and the matching was refined by an intensity-based cross-correlation process. The depth map from the last stereo image pair is shown in Figure 2.

The feature points are classified into two categories: old and new. The old points are visible both in the current image and the previous one. The new points are visible in the current image but not in the previous one. A point is no longer considered as old if its neighborhood changed drastically due to motion. Only the old points (about 50 for each image pair) are included in the iterative optimization. The structure of the new points is estimated after motion.

In our experiment, two solutions were computed, the matrix-weighted close-form solution and the iterative optimal solution. These two solutions are very close but the iterative optimal solution is slightly better in terms of the root mean squared error in the estimated structure. This indicates that the matching is roughly correct on the corner points. Due to space limitation, only the first and last interframe motions, are listed in Table 1. The estimated angles of the interframe rotations from the optimization are listed in Table 2. Compared to the ground truth 2.25°, they are quite accurate. The error in rotation angles is within 0.17°.
In order to test the accuracy in the estimated structure, we manually measured the length $l_i$ of approximately 40 lines in the scene as ground truth.

From the estimated length $\hat{l}_i$, we can compute the root mean squared error of the line lengths as

$$\text{rmse} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (l_i - \hat{l}_i)^2},$$

and its mean error as

$$\text{me} = \frac{1}{n} \sum_{i=1}^{n} \| (l_i - \hat{l}_i) \|.$$

Table 2 illustrates for each view the rmse and me values for visible lines. We have noticed in the experiments that if point correspondences are well spread in stereo image pairs, covering as large common field of views as possible, then the corresponding motion and 3-D structure estimates are more accurate and stable.

It can be seen in Table 2 that the 3-D structure error increased with the frame number, one reason for this is that the corner points in the latter frames were not as well spread as in the first few frames, which may have resulted in a less accurate motion estimation (as the interframe translation ground truth was not available in the experiments, it is hard to assess the accuracy of the estimated motion for each frame). Moreover, the number and the distance of the visible lines varied from frame to frame, which also made the values of the structure error different.

Another point we may mention here is that we used wide angle lenses. This implies that a pixel corresponds to a larger area in the scene than with a normal lens or a tele-lens. Also, a short baseline was used (about the distance between human eyes). If a tele-lens or wider baseline were employed, the structure error would decrease significantly.

6 Summary

Our approach to motion and structure analysis through long image sequences is characterized by the following aspects:

1. A closed-form matrix-weighted solution is used to obtain a reliable solution for interframe motion.
2. To further improve the closed-form solution, an iterative optimization is formulated, using the space decomposition strategy.
3. A recursive-batch approach is used to process long image sequences. Careful experiments have been conducted to investigate the performance of our methods. The experimental results have been compared to the available ground truth.

Acknowledgements

This research was supported by the National Science and Engineering Research Council of Canada (under grant OGP00A3890) and the IRIS Network of Centers of Excellence.

References

Table 1: Motion estimations resulting from the matrix-weighted linear algorithm and the nonlinear optimization (unit m).

<table>
<thead>
<tr>
<th>motion parameters</th>
<th>linear</th>
<th>nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td>motion 1,0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>translation</td>
<td>(-0.003)</td>
<td>(-0.002)</td>
</tr>
<tr>
<td>(t_x)</td>
<td>(-0.044)</td>
<td>(-0.047)</td>
</tr>
<tr>
<td>(t_y)</td>
<td>(-0.007)</td>
<td>(-0.008)</td>
</tr>
<tr>
<td>length</td>
<td>0.045</td>
<td>0.048</td>
</tr>
<tr>
<td>rotation axis</td>
<td>(N_x)</td>
<td>0.872</td>
</tr>
<tr>
<td>(N_y)</td>
<td>0.112</td>
<td>0.858</td>
</tr>
<tr>
<td>(N_z)</td>
<td>-0.474</td>
<td>-0.502</td>
</tr>
<tr>
<td>rotation angle</td>
<td>(\theta^(*))</td>
<td>2.299</td>
</tr>
<tr>
<td>motion 9,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>translation</td>
<td>(-0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>(t_x)</td>
<td>(-0.045)</td>
<td>(-0.046)</td>
</tr>
<tr>
<td>(t_y)</td>
<td>(-0.009)</td>
<td>(-0.010)</td>
</tr>
<tr>
<td>length</td>
<td>0.046</td>
<td>0.047</td>
</tr>
<tr>
<td>rotation axis</td>
<td>(N_x)</td>
<td>0.853</td>
</tr>
<tr>
<td>(N_y)</td>
<td>0.072</td>
<td>0.855</td>
</tr>
<tr>
<td>(N_z)</td>
<td>-0.516</td>
<td>-0.516</td>
</tr>
<tr>
<td>rotation angle</td>
<td>(\theta^(*))</td>
<td>2.237</td>
</tr>
</tbody>
</table>

Figure 2: The depth map is shown as an intensity image from the last stereo image pair.

Table 2: Estimated motion and structure. \(k\): time index; \(\theta\): rotation angle (degree); \(rmse\): root mean square error (mm); \(me\): mean error (mm).

<table>
<thead>
<tr>
<th>(k)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\theta)</td>
<td>2.16</td>
<td>2.18</td>
<td>2.08</td>
<td>2.33</td>
<td>2.20</td>
<td>2.30</td>
<td>2.19</td>
<td>2.22</td>
<td>2.19</td>
</tr>
<tr>
<td>(rmse)</td>
<td>9.79</td>
<td>9.64</td>
<td>19.05</td>
<td>22.92</td>
<td>26.49</td>
<td>29.38</td>
<td>26.56</td>
<td>27.49</td>
<td>30.30</td>
</tr>
<tr>
<td>(me)</td>
<td>6.64</td>
<td>5.74</td>
<td>9.67</td>
<td>13.25</td>
<td>17.45</td>
<td>20.40</td>
<td>19.39</td>
<td>20.33</td>
<td>23.43</td>
</tr>
</tbody>
</table>