Variational methods for dense depth reconstruction from monocular and binocular video sequences

Geert De Cubber, E-mail: gdcubber@gmail.com

1 INTRODUCTION

1.1 Problem statement

The physical world can be regarded as a three-dimensional geometric space. This three-dimensional world is observed/perceived by humans by means of their five senses: hearing, touch, smell, taste and vision. Of all these sensing modalities, vision is the most powerful, which is shown by the fact that it occupies the most cortical space. The physics of vision is relatively well understood. However, once higher level representations are considered, the situation becomes less clear.

One important area of vision research concerns our ability to understand the three-dimensional nature of our environment. Indeed, the human eye can be regarded as a two-dimensional imaging device, meaning that the resulting image is (only) a 2D representation/projection of the 3D world. To recover the depth dimension which was lost during projection on the retina, the human visual system fuses multiple depth cues. These depth cues are grouped into two categories: monocular cues (cues available from the input of just one eye) and binocular cues (cues that require input from both eyes).

The computer vision community has been researching for a few decades now analogous methodologies mimicking the depth perception abilities of the human visual system. Most attention has hereby been focussed on the stereopsis-based (and thus binocular) depth cue. The result is that depth-from-stereo has now evolved to a mature research subject [1]. Motion-based depth reconstruction approaches like motion parallax and kinetic depth present a more promising research direction. Where stereo vision must be seen as a spatial integration of multiple viewpoints to recover depth, motion-based depth reconstruction can be seen as performing a temporal integration. The problem arising in this situation is known as the Structure from Motion problem and deals with extracting three-dimensional information about the environment from the motion of its projection onto a two-dimensional surface.

1.2 Research Objectives

In light of the previous work done in the field of 3D reconstruction, the main objective of this research work is to develop a dense structure from motion recovery algorithm. This algorithm should operate on monocular image sequences, using the camera movement as a main depth cue. The term dense indicates that a depth estimate for each pixel of the input images is required. We will show that an iterative variational technique is able to solve this 3D reconstruction problem. However, to converge to a solution, the iterative technique requires proper initialization. For this initialization process, a set of tools, based upon standard sparse structure from motion techniques, is employed. These classical methods are not capable of estimating a dense reconstruction, but they do suffice to estimate the camera motion parameters and the 3D positions of some feature points, which is used as an initial value for the iterative solver.

Dense structure from motion is a promising technology, but suffers from one major disadvantage: the processing time is in general very long. This is the case for most of the existing approaches and it is no different in our work. This drawback makes it less suited for a number of applications where the timely delivery of results is an issue (e.g. robotics). To deal with these issues, a second research objective was to integrate the developed dense structure from motion algorithm into a stereo reconstruction context. In this way, the processing time could be drastically reduced, as stereo adds a valuable constraint, which limits the search domain for solutions dramatically.
2 MONOCULAR RECONSTRUCTION

2.1 Problem statement

Dense structure from motion algorithms aim at estimating a 3D location for all image pixels. One could say that they seek to transform a normal camera from a 2D imaging device into a 3D imaging sensor. There are multiple approaches towards dense structure from motion. The most modern dense structure from motion algorithms minimize the optical flow constraint and enforce smoothness in the depth field in a variational framework. However, due to the noisiness of the optical flow and due to projection ambiguities (leading a.o. to occlusions), these algorithms are still not very robust when confronted with unconstrained 3D camera motion and changing illumination conditions. One could argue that these problems are due to the fact that dense structure from motion is a relatively new field of research that emerged recently thanks to the rise in computing power.

Sparse structure from motion, on the other hand, is a more mature research domain, which dates back to the early work of Longuet-Higgins [3] and Ullman [4]. Through the years, sparse structure from motion algorithms have been optimized and made more robust, notably by the work of Horn [5], Hartley and Zissermann [6] and Torr [7]. Hence, it would be beneficial to use the experience from sparse structure from motion in the dense case, which is the aim of this work.

2.2 Methodology

Variational formulation: To address the classical dense structure from motion shortcomings, we adopt a dual approach for dense structure estimation, trying to combine the robustness of sparse reconstruction techniques with the completeness of dense reconstruction algorithms. This is achieved by first solving the sparse reconstruction problem. These results then serve as initial guesses for the dense reconstruction process, which fuses the sparse data with dense information coming from a densely estimated optical flow field. The optical flow \( \mathbf{u} \) is a projection of the 3D motion field and is related to the structure and motion properties:

\[
\mathbf{u} = \mathbf{Q}_\omega \mathbf{\omega} + d\mathbf{Q}_t \mathbf{t},
\]

with the proximity \( d \) defined as \( d = \frac{1}{2} \) and the matrices \( \mathbf{Q}_\omega = \begin{bmatrix} \frac{xu}{I} & -f - \frac{x^2}{I} & y \\ \frac{yu}{I} & f + \frac{y^2}{I} & -x \end{bmatrix} \) and \( \mathbf{Q}_t = \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix} \). This relation between optical flow and structure and motion on one hand and the available sparsely reconstructed structure and motion parameters allow for integrated sparse-dense reconstruction, as sketched by Figure 1.

Here, a variational approach is presented to tackle this high-dimensional data fusion problem. This methodology formulates the problem of fusing dense image data - in the form of the image brightness constraint from the optical flow - with sparse data - in the form of the epipolar constraint of the sparse reconstruction - as an optimization problem.

The basic problem of the calculus of variations is to determine the function \( q(x, y) \) which minimizes or maximizes a functional

\[
J = \int F(q(x, y), q_x(x, y), q_y(x, y), x, y) \, dx \, dy,
\]

with \( q_x(x, y) = \frac{\partial q(x,y)}{\partial x} \) and \( q_y(x, y) = \frac{\partial q(x,y)}{\partial y} \). The integrand \( F(q(x, y), q_x(x, y), q_y(x, y), x, y) \) is in our case composed of two constraint equations. A first constraint, \( \phi_{\text{model}}(x,y) \), expresses the conformity of the current depth estimate at each pixel to the dense and sparse constraint models. Here, the expression of \( \phi_{\text{model}}(x,y) \) is based upon the image derivatives based optical flow constraint, expressed as \( I_xu + I_yv + I_t = 0 \). A second constraint, \( \phi_{\text{regularization}}(x,y) \), introduces anisotropic regularization to preserve the structure smoothness, while preserving depth discontinuities at boundary locations. \( J \) can thus be written as:

\[
J = \int_{\Omega} \phi_{\text{model}}(x,y) + \mu \phi_{\text{regularization}}(x,y) \, dx \, dy \quad (2)
\]

Automated diffusion parameter estimation: The diffusion parameter \( \mu \) is a positive constant, used to control the regularization during the diffusion process, i.e. to decide to which extent the disparity field is diffused at different locations of the disparity field \( \delta \). In the literature, the diffusion parameter is generally estimated empirically. There are several drawbacks related to this approach. First, the use of problem-dependent parameters with values which have to be empirically re-estimated for each set of input data, is a tedious process which hinders deterministic benchmarking of reconstruction algorithms. Second, considering that the solution of equation 9 must be sought using an iterative updating technique, there is no reason to assume that one and the same value for the regularization parameter \( \mu \) would be the optimal value throughout the iterative process. Therefore, it is beneficial to dynamically select the diffusion parameter.
Here, we use a method based upon the work of Yang who proposes in [8] a method to dynamically estimate \( \mu \). This estimation is based upon the assumption that pixels for which the gradient of the disparity field is above a certain percentage, for example 90% of the maximum of the histogram, are discontinuous locations. Here, the diffusion should stop or at least be largely reduced. This threshold is defined by Yang as \( T_g^k = \frac{(\mu)^2}{K^k + 2T_g^k} \), where \( K^k \) is the threshold of \( \| \nabla \delta^k \|_2 \) at time instance \( k \) associated with the location in the histogram of the disparity gradient where the percentage of the histogram reaches 90% of the maximum of the histogram. \( T_g^k \) is the corresponding value of the diffusion function at time instance \( k \). It is a small (e.g., 0.01), but known value, allowing us to determine the diffusion parameter \( \mu \) by:

\[
\mu^k = \sqrt{\frac{K^k T_g^k}{1 - 2T_g^k}},
\]

where \( \mu^k \) ensures that the diffusion function is less than \( T_g^k \) when the pixel locations belong to discontinuities.

Using this strategy, \( \mu^k \) can be automatically adjusted according to the disparity gradient distribution at each step of the diffusion process. The value of \( \mu^k \) is based upon the estimated disparity field during the diffusion process to better control the diffusion. This gives a more reasonable estimate of \( \mu \) compared to a fixed value for the whole diffusion process.

**Integration of sparse motion and dense flow data:**

The modeling term of equation 2, \( \phi_{model}(x, y) \), expresses the optical flow constraint. In this work, we based ourselves on the approach proposed by Alvarez in [9]. This formulation uses the fundamental matrix, which has as an advantage that, as the fundamental matrix is one of the most basic descriptors of the two-view geometry, it is more robust. Following this approach, \( \phi_{model}(x, y) \) can be written as:

\[
\phi_{model} = (I_{1,x} [a_1 \zeta + b_1] + I_{1,y} [a_2 \zeta + b_2] + I_{1,t}^2)
\]

The parameters \( a_1, a_2, b_1, b_2 \) in equation 4 depend on the two-view geometry expressed by the fundamental matrix \( F \) and \( \zeta \) represents the depth field to be estimated. Here, we assume the fundamental matrix to be known through a preliminary sparse reconstruction procedure, such that the depth field \( \zeta(x, y) \) can be estimated through minimization. In practice, the estimate for \( F \) is often not very accurate and it can be seen immediately from equation 4 that this would corrupt the depth estimation.

**Iterative motion parameter estimation:** To address this issue, we proposed a novel methodology for improving the estimate for the fundamental matrix \( F \). This can be done iteratively, as the depth estimation process is an iterative one. Indeed, the depth parameter \( \zeta_i(x, y) \) is in fact a function of the iteration number, expressed by the time parameter \( t \): \( \zeta_i(x, y) = \zeta^t_i(x, y) \), as with each iteration, a new value for \( \zeta^t_i(x, y) \) is estimated. It is thus possible to re-insert these results at each iteration to improve the estimation of the two-view geometry.

Considering that the depth parameter for each pixel \( \zeta_i(x, y) \) is known and the components of the fundamental matrix are the unknown parameters, the problem can be expressed as:

\[
\text{argmin}_{f_1, \ldots, f_N} \sum_{i=1}^{N} (\Gamma(f))^2,
\]

with \( f = f_{1..9} \) a vector consisting of the components of the fundamental matrix.

Equation 5 expresses a non-linear least squares problem, aiming at minimizing the sum of squared residuals \( r_1 = a_1 \zeta_1 \zeta^{t+1}_1 I_{1,x} + b_1 (f) I_{1,x} + a_2 (f) \zeta^t_1 I_{1,y} + b_2 (f) I_{1,y} + I_{1,t} \).

This least squares problem can be solved iteratively using the Gauss-Newton algorithm:

\[
f^{k+1} = f^k - (J^T J)^{-1} J^T r,
\]

with \( r \) the vector of residuals \( r_1 \) and \( J \) is the Jacobian matrix consisting of partial derivatives of \( r_1 \) to the different components of \( f \), such that \( J_{ij} = \frac{\partial r_i}{\partial f_j} \).

Following the fundamental matrix update equation 6, the two view geometry description (expressed by the fundamental matrix \( F \)) is iteratively updated by using the current estimate of the structure description (expressed by the depth parameter \( \zeta^t \)). Consecutive structure reconstruction results are used to improve the estimate of \( F \), which are on their turn used to improve the structure estimation.

**Anisotropic depth regularization:** Using only the constraint equations described up until now would lead to serious problems due to spatial (image) and temporal (movement) ambiguities. For example, matching of image intensities typically fails on monochrome surfaces, because due to the fact that there are multiple solutions, the numerical stability cannot be assured and the solver converges to random solutions or does not converge at all. What is needed to solve this problem is a regularization term which extrapolates and smooths the structural data over pixels which belong to the same physical object at the same distance. Multiple smoothing functions have been proposed and used in a structure from motion framework. The main problem these smoothing terms face is the preservation of discontinuities. Indeed, regularization should not over-smooth the solution such that depth discontinuities are no longer visible. Nagel and Enkelmann took into account this consideration and proposed in [10] an anisotropic smoothing term which preserves the depth discontinuities. The Nagel and Enkelmann regularization model has been proven successful in a range of independent experiments [9], [11] and formulates a regularization term of the form:

\[
\phi_{\text{regularization}} = (\nabla \zeta^t)^T D (\nabla I_1) (\nabla \zeta^t)
\]

Where \( D \) is a regularized projection matrix, assuring that smoothing takes place only along the direction of the boundaries and not across boundaries. Using this approach, discontinuities can be preserved while the energy functional is minimized.
Variational expression: The extremal functions of equation 2 can be obtained by expressing the Euler-Lagrange equations:

$$\frac{\partial F}{\partial \zeta} - \frac{d}{dx} \left( \frac{\partial F}{\partial x} \right) - \frac{d}{dy} \left( \frac{\partial F}{\partial y} \right) = 0, \quad (8)$$

with $F = \phi_{model}(x, y) + \mu \phi_{regularization}(x, y)$. $\phi_{data}(x, y)$ is defined by the constraint equation 4, while $\phi_{regularization}(x, y)$ defined by the regularization constraint of equation 7.

By defining $\Gamma = (I_{1,x}[a_1 \zeta + b_1] + I_{1,y}[a_2 \zeta + b_2] + I_{1,t})$, a formulation for the PDE-problem can be obtained:

$$\Gamma(a_1 I_{1,x} + a_2 I_{1,y}) - \mu \text{div}(D \nabla I_1) \nabla \zeta = 0 \quad (9)$$

Solving these partial differential equations for $\zeta(x, y)$ returns a solution for the depth field. However, solving this partial differential equation (PDE) in an analytical way is in general not possible. Hence, iterative approaches are required.

Numerical implementation: A semi-implicit scheme was chosen as a numerical scheme for solving the Euler-Lagrange problem stated by equation 9. The semi-implicit scheme can be seen as a compromise between the simplicity of the explicit scheme and the stability of the implicit scheme. Like the implicit scheme, the semi-implicit scheme is unconditionally stable, thus the time parameter $\Delta \sigma$ can be relatively large without producing any unstable nonphysical oscillations.

For this problem, the Euler-Lagrange equation can be solved, provided that an initial condition is given, by introducing a time parameter $\sigma$ and by calculating the asymptotic state for $\sigma \to \infty$:

$$\begin{align*}
\frac{\partial \zeta(x, \sigma)}{\partial \sigma} &= -\Gamma(a_1 I_{1,x} + a_2 I_{1,y}) + \mu \text{div}(D \nabla I_1) \nabla \zeta \\
\zeta(x, \sigma = 0) &= \zeta_0
\end{align*} \quad (10)$$

The system is solved by a least squares subspace trust region method based on the interior-reflective Newton method, as described in [12]. Each iteration involves the approximate solution of a large linear system using the method of preconditioned conjugate gradients. The iteration ends when the difference between 2 consecutive depth estimates falls below a chosen threshold.

The result of these iterative solvers depends strongly on the initialization method, which is why we also introduce a dense depth map initialization method which fuses sparse and dense data.

Initial guess estimation: To obtain an initial value for the depth field $\zeta_0$, it is sufficient to calculate an initial value for the flow field $u_0$, as both can be related. The calculation of $u_0$ is based upon the fusion of dense information from the optical flow and reconstructed sparse features. The reason for fusing both data streams is that both cues have their advantages and disadvantages:

- From a set of sparsely matched feature points $x^F(x^F, y^F)$ and $x^F(x^F, y^F)$, it is straightforward to calculate a feature flow magnitude map $\delta^F$. These sparse feature matches and the associated depth field can be estimated accurately, but they only contain sparse information.
- A densely estimated optical flow field $u$ has as a problem that, in general, this optical flow field is scaled, such that it is only possible to retrieve a relative estimate for the depth. To obtain an absolute measure, it is necessary to estimate a scale factor, $\sigma_s$, as $\sigma_s = \arg\min_{\sigma_s} \sum |u^F(x^F) - \sigma_s u(x^F)|^2$. The knowledge of the scale factor $\sigma_s$ between the flow fields as estimated by the dense optical flow and by the feature matching, allows to define a new flow field magnitude map, which is correctly scaled:

$$u^\text{flow}(x) = \sigma_s \sqrt{u(x)^2 + v(x)^2} \quad (11)$$

The remaining problem is that the estimated optical flow is in general less robust than the feature flow as calculated by point correspondences. It would thus be beneficial to combine the two types of data.

The problem when trying to fuse the estimated flow field magnitude map $u^\text{flow}(x)$ and the feature flow field magnitude map $u^F(x^F)$ is that the latter function is only defined at the feature points, whereas the former is a dense function defined at each image pixel. Therefore a region growing algorithm applied to the sparse disparity map $u^F(x^F)$.

For each pixel $x$, we estimate its probability to belong to the same region ($\Re(x^F)$) of a feature point $x^F$, taking into account the distance $\|x - x^F\|^2$ between the two points, and the difference between the flow magnitude values $u^\text{flow}(x)$ and $u^F(x^F)$. At each pixel location $x$, the region with the highest probability is selected, and the corresponding flow magnitude value $u^F(x^F)$ is associated to the dense feature-based flow field $u^\text{features}(x)$.

The initial flow magnitude map $u_0(x)$ can then be calculated by combining $u^\text{features}(x)$ and $u^\text{flow}(x)$:

$$u_0(x) = u^\text{features}(x) + \frac{u^\text{flow}(x) - \sum_i u^\text{flow}(x)}{\max(u^\text{flow}(x)) - \min(u^\text{flow}(x))} \quad (12)$$

2.3 Results

The presented algorithms were tested on a traditional benchmarking sequence, provided by Strecha et al. [13]. The fountain sequence consists of a series of shots from a water fountain, as shown on figure 2, recorded by a high-resolution camera. The reconstruction results from this sequence serve to compare the performance of the proposed method to existing state-of-the-art approaches. Figure 2 also shows the reconstructed depth maps for some frames of the Fountain sequence. It can be noted that the depth maps present a visually excellent reconstruction of the scene structure.

Quantitatively, the reconstruction results can be evaluated using the accuracy and completeness measures. Whereas the accuracy measures the relative error, the
completeness indicates the percentage of pixel within an error range of $3\sigma$ from the ground truth. The accuracy and completeness results can be compared to the results of other algorithms on the same sequence, as reported in [13]. Table 1 compares the numerical values of the 90% accuracy and 1.25mm completeness measures of the presented method with 3 other methods [13]. From this analysis, it is clear that the proposed algorithm is not the best one for the accuracy, as the Furukawa algorithm [14] scores better with a slightly lower relative error of 2.04, compared to 2.08 for our algorithm. Also for the completeness, the presented algorithm is not the best performer, as the algorithm by Strecha achieves a somewhat higher degree of completeness of 87.06%, compared to 85.6% for our algorithm. However, the presented algorithm presents overall the best compromise between accuracy and completeness, coming very close to the top performers in each category.

Compared to the other algorithms, the presented method shows relatively few very small and very large errors. This means that the presented method is not the most accurate in absolute terms, but it achieves at estimating a dense reconstruction where the errors are well constrained. The relative absence of very large errors means that the reconstruction results are robust and trustable.

Figure 3 shows some views of the reconstructed 3D model of the water fountain, textured with the original color information. It can be observed that the original 3D structure is visually well represented, as indicated above by the quantitative measures.

To further validate our approach, we used in a second phase a natural outdoor sequence as input data. This sequence was recorded using a simple commercial low resolution hand-held camera in outdoor conditions. The sequence, which is shown on Figure 4, was filmed by a person walking around a statue featuring a mixed background of trees, buildings and traffic. Due to the nature of this sequence, no ground truth data is presentable, so the performance of the reconstruction algorithm needs to be visually judged from the final 3D reconstruction.

Choosing evaluation sequences like this one means that the reconstruction algorithm needs to deal with a number of problems:

- Erratic, non-deterministic movement which is hard to estimate
Fig. 4. Some frames of the Natural Sequence and their corresponding estimated Depth Maps

- Low resolution and low-fidelity input data
- Difficult outdoor variable lighting conditions.
- Relatively large distance from camera to subject
- Complex background

Most existing benchmark sequences for multi-view image reconstruction nicely avoid these problems. Typically small, simple objects are filmed with a camera following a well-described motion path. In contrast, the Hands natural sequence presents a particularly difficult case for dense 3D reconstruction, as it consists of a large series of frames (400 frames in total), filmed with a low-quality commercial camera in challenging outside illumination conditions. Due to the nature of this sequence, no ground truth data is present, so only qualitative data is presented in this section to evaluate the presented reconstruction algorithm on this sequence.

Figure 4 shows the depth reconstruction result for some frames of the Hands natural sequence. It is apparent that the dense reconstruction technique succeeds in obtaining a correct depth estimate for this quite complex natural sequence. Relative depths have been well estimated, continuous areas have continuous depths and discontinuities are well preserved.

Fig. 5. Novel Views based on the Reconstructed 3D Model of the Statue of the Natural Sequence

Figure 5 shows some views of the reconstructed 3D model of the hands statue. It can be observed that the hands statue is clearly distinguishable from the background and that the reconstructed 3D structure of the statue corresponds to the real statue. It is in difficult sequences like this one that our approach shows its main advantage: that the overall reconstruction result has a high quality, without disturbing outliers.

3 BINOCULAR RECONSTRUCTION

3.1 Problem statement

The integration of the stereo and motion depth cues offers the potential of a superior depth perception, as the combination of temporal and spatial information makes it possible to reduce the uncertainty in the depth reconstruction result and to augment the precision. However, this requires the development of a data fusion methodology which is able to combine the advantages of each method, without propagating any errors induced by one of the depth reconstruction cues. Therefore, the mathematical formulation of the problem of combining stereo and motion data must be carefully considered.

The dense depth reconstruction problem can be casted as an energy minimization problem, as shown before by a number of researchers [17], [18]. The main problem in dense stereo - motion reconstruction is that the solving
methodology depends on the simultaneous evaluation of multiple constraints which have to be balanced carefully.

This is sketched on Figure 6, which shows the different constraints needed to be imposed for a sequence shot with a moving binocular camera. Consider a pair of rectified stereo images \((I_1^l, I_1^r)\) shot at time \(t = t_0\) and a stereo pair \((I_2^l, I_2^r)\) shot at time \(t = t_0 + t_k\), with \(t_k\) being determined by the framerate of the camera. A point \(x_1^l\) in the reference frame \(I_1^l\) can now be related to a point \(x_1^r\) via the stereo constraint, as well as to a point \(x_2^l\) via the motion constraint. Using the stereo and motion constraint in combination, the point \(x_1^l\) can even be related to a point \(x_2^r\), and this via a stereo + motion or a motion + stereo path. It is evident that, ideally, all these interrelations should be taken into consideration, and this for all the pixels in all the frames in the sequence. However, when confronted with long sequences of high-resolution video, this leads to prohibitively large systems of equations, which cannot be solved in an efficient way.

![Fig. 6. Motion and Stereo Constraints on a binocular sequence](image)

3.2 Methodology

The stereo-motion integration problem can be regarded as a high-dimensional data fusion problem. The presented approach solves this problem by posing it as an optimization problem. As such, the main problem is finding a suitable functional which minimizes the error on the dense reconstruction. In principle it is possible to relate every pixel from a reference image to a pixel in each of the other images of the sequence and vice versa. However, when confronted with a long sequence of high-resolution images, following this approach would lead to a massive amount of constraint equations to be solved simultaneously, which would make the problem impossible to solve in practice. Therefore, in order to limit the dimensionality of the problem, we adopt a different approach, where the sequence of stereo images is processed sequentially. Following this methodology, a pair of stereo images is related to a successive pair, as sketched in figure 7.

![Fig. 7. Processing strategy of a binocular sequence](image)

Figure 7 considers a binocular image stream consisting of left and right images of a stereo camera system. The left and right streams are processed individually, using the dense structure from motion algorithm proposed by De Cubber in [19], resulting in, respectively, a left and right proximity map \(d^l\) and \(d^r\). In parallel, the left and right images are combined using the stereo motion algorithm [20], [21], embedded in the Bumblebee stereo camera. As a result of this stereo computation, a new proximity map from stereo \(d^c\) can be defined. The reason for calling this proximity map \(d^c\) lies in the fact that it is defined in the reference frame of a virtual central camera of the stereo vision system.

Of course, there exist strong interrelations between the different proximity maps \(d^l\), \(d^r\) and \(d^c\), which need to be expressed to ensure consistency and to improve the reconstruction result. Therefore, we adopt an approach where, the left proximity map \(d^l\) is optimized, subject to two constraints, relating it to \(d^c\) and \(d^r\) respectively. In parallel, the right proximity map \(d^r\) is optimized, also subject to two constraints, relating it to \(d^c\) and \(d^l\).

The dense stereo - motion reconstruction problem can thus be stated as a constrained optimization problem:

\[
\text{Find } \min_{\mathbf{x} \in \Omega} E(\mathbf{x}) \text{ subject to: } \theta_i(\mathbf{x}) = 0 \text{ for } i = 1, ..., n \tag{13}
\]

where the function \(E(\mathbf{x})\) is the objective functional and \(\theta_i(\mathbf{x})\) express a number of constraint equations.

A traditional solving technique for constrained optimization problems as the one posed by equation 13 is the Lagrangian multiplier method, which converts a constrained minimization problem into an unconstrained minimization problem of a Lagrange function. In theory,
the Lagrangian methodology can be used to solve the stereo - motion reconstruction problem, however, to improve the convergence characteristics of the optimization scheme, it is better to use the augmented Lagrangian \( \mathcal{L}(x, \lambda) \). The augmented Lagrangian, which was presented by Powell and Hestenes in [22] and [23], adds a quadratic penalty term to the original Lagrangian:

\[
\mathcal{L}(x, \lambda) = E(x) + \sum_{i=1}^{n} (\lambda_i \theta_i(x)) + \frac{\rho}{2} \sum_{i=1}^{n} \theta_i(x)^2,
\]

(14)

with a penalty parameter \( \rho > 0 \).

In the context of dense stereo - motion reconstruction, we seek to simultaneously minimize 2 energy functions: \( E^l \) for the left image and \( E^r \) for the right image. These objective functions are a function of the proximity maps \( d_l^i \) and \( d_r^i \) (index “1” indicates the first image), which we seek to optimize. This optimization problem is subject to 4 constraint equations:

1. \( \theta_{lj}^l (d_l^i, d_l^j) = 0 \) relates \( d_l^i \) to the proximity map obtained from stereo \( d_l^c \).
2. \( \theta_{lj}^r (d_l^i, d_r^j) = 0 \) relates \( d_l^i \) to the proximity map of the right image \( d_r^j \).
3. \( \theta_{rj}^r (d_r^i, d_r^j) = 0 \) relates \( d_r^i \) to the proximity map obtained from stereo \( d_r^c \).
4. \( \theta_{rj}^l (d_r^i, d_l^j) = 0 \) relates \( d_r^i \) to the proximity map of the left image \( d_l^j \).

According to the Augmented Lagrangian theorem and the definition given by equation 14, we can write the augmented Lagrangian for the left image thus as follows:

\[
\mathcal{L}^l_i(d_l^i, \lambda_{lc}^i, \lambda_{rl}^i) = E^l_i(d_l^i) + \lambda_{lc}^i \theta_{lc}^l(d_l^i, d_l^c) + \frac{\rho}{2} \left[ \theta_{lc}^l(d_l^i, d_l^c) \right]^2 \\
+ \lambda_{rl}^i \theta_{rl}^l(d_l^i, d_l^j) + \frac{\rho}{2} \left[ \theta_{rl}^l(d_l^i, d_l^j) \right]^2
\]

(15)

For the right image we have in a similar fashion:

\[
\mathcal{L}^r_i(d_r^i, \lambda_{rc}^i, \lambda_{rl}^i) = E^r_i(d_r^i) + \lambda_{rc}^r \theta_{rc}^r(d_r^i, d_r^c) + \frac{\rho}{2} \left[ \theta_{rc}^r(d_r^i, d_r^c) \right]^2 \\
+ \lambda_{rl}^r \theta_{rl}^r(d_r^i, d_r^j) + \frac{\rho}{2} \left[ \theta_{rl}^r(d_r^i, d_r^j) \right]^2
\]

(16)

The energy function in equations 15 and 16 expresses the relationship between structure and motion between successive images and uses the formulations presented in section 2 for the monocular case.

The constraints \( \theta_{ij}^l(d_l^i, d_l^j) \) with \( (i,j)=\text{left, centre, right} \) express the similarity between an estimated proximity map \( d_l^i \) and another proximity map \( d_l^j \). In order to calculate this similarity, the second proximity map must be warped to the first one. This warping process is expressed by introducing a warping function \( \psi \), which is a function of the proximity \( d \) and the camera rotation and translation: \( \psi = \psi(x, d, \omega, t) \). The introduction of \( \psi \) allows to define the constraint equations, by comparing the proximity maps. This means that the constraint equations can be written in their most general form as:

\[
\theta_{ij}^l(d_l^i, d_l^j) = \left( d_l^i - d_l^j (x + \psi(x, d_l^i(x), \omega, t)) \right)^2
\]

(17)

As a numerical solving technique, we use the method presented by Brent in [24]. Brent’s method switches between inverse parabolic interpolation and golden section search [25]. This optimization method converges to a minimum within the search interval. Therefore, it is crucial that a good initial value is available for all status variables. To estimate this initial value for the proximity field, the dense disparity map from stereo is used. The reason for this is that the camera displacement between the left and right stereo frames is well known and is fixed over time. As such, it is possible to warp the stereo data in the virtual central camera reference frame towards the left and the right image with high accuracy.

The numerical scheme results in a methodology where there are, two functions which are optimized at the same time: one using \( \mathcal{L}^l_i \) which optimizes the left proximity map \( d_l^i \) and one using \( \mathcal{L}^r_i \) which optimizes the right proximity map \( d_r^i \). In the proposed algorithm, these functions are optimized alternatively, hereby always using the latest result for both proximity maps.

3.3 Results

The validation and evaluation of a dense stereo - motion reconstruction algorithm requires the use of an image sequence shot with a moving stereo camera. For this reason, we recorded a sequence, using a Bumblebee stereo vision camera. This sequence, shown in Figure 8, was shot in an indoor office environment. Due to the nature of this sequence, it is impossible to retrieve ground truth data about the depth field. The translation of the camera is mainly along its optical axis (Z-axis) and along the positive X-axis. The rotation of the camera is almost only along the positive Y-axis.

Fig. 8. Some frames of the binocular Desk sequence

It can be noted from Figure 8, that the used sequence consists of a cluttered environment, presenting serious challenges for any reconstruction algorithm:
• Cluttered environment with many objects at different scales of depth.
• Relatively large totally untextured areas making correspondence matching very difficult.
• Areas with specular reflection, violating the Lambertian assumption, traditionally made for stereo matching.
• Variable lighting and heavy reflections, causing saturation effects and incoherent pixel colors across different frames.

To compare our method to the state of the art, we implemented a more classical dense stereo-motion reconstruction approach. This approach defines classical stereo and motion constraints, based upon the constant image brightness assumption, alongside the Nagel-Enkelmann regularization constraint. These constraints are integrated into one objective function, which is solved using a traditional trust-region method. As such, this approach presents a relatively simple and straightforward solution. This methodology is used to serve as a base benchmarking method for the Augmented Lagrangian based stereo-motion reconstruction technique.

Applying this more classical technique to the Desk sequence shown in Figure 8 results in a depth reconstruction as shown in Figure 9. Overall, the reconstruction of the proximity field correlates with the physical reality, as imaged on Figure 8, but there are some serious errors in the reconstructed proximity fields, notably on the board in the middle of the image. This leads us to conclude that this method is not suitable for high-quality 3D modeling.

![Fig. 9. Proximity Maps for different frames of the Desk sequence using the Global Optimization Algorithm](image1)

Fig. 9. Proximity Maps for different frames of the Desk sequence using the Global Optimization Algorithm

The result of Figure 10 can be compared to Figure 9, which shows the same output, but using the global optimization approach. From this comparison, it is evident that the result of the Augmented Lagrangian-based reconstruction technique is far superior to the one using global optimization. The global optimization result features numerous problems: erroneous proximity values, under-regularized areas, over-regularized areas, erroneous estimation of discontinuities, ... None of those problems are present in the result of the Augmented Lagrangian, as shown on Figure 10.

To show the applicability of the presented technique towards 3D modeling, the individual reconstruction results were integrated to form one consistent 3D representation of the imaged environment. Figure 11 shows 4 novel views of the 3D model which was reconstructed as such. Figure 11 shows that a qualitative 3D model can be reconstructed using the Augmented Lagrangian-based stereo-motion reconstruction technique. Indeed, from the different novel viewpoints, the 3D structure of the office environment can be clearly deduced. There are no visible outliers and all items in the scene have been reconstructed, even those with very low texture.
The main contribution of this work lies in the development of a novel algorithm for dense 3D reconstruction from monocular image sequences. The general concept behind this algorithm is that it combines the robustness of traditional sparse structure from motion methods with the completeness of optical flow based dense reconstruction approaches. This was achieved by defining an integrated framework, formulating the problem of fusing dense image data as an optimization problem. A comparison to other state-of-the-art dense reconstruction techniques learned that the proposed dense reconstruction approach performs excellent. Without being the actual top performer for one specific quality measure, it succeeds at estimating a globally optimal reconstruction, which balances the accuracy and completeness measures. Our method delivers robust and reliable 3D reconstruction results, due to the relative absence of very large errors. As such, it proves to be a valuable candidate for the dense reconstruction of natural sequences.

Optimal depth perception requires the fusion of different depth cues. Therefore, we extended the monocular dense reconstruction algorithm to the binocular case, by integrating it in a dense stereo - motion reconstruction framework. The combination of temporal and spatial information makes it possible to reduce the uncertainty in the depth reconstruction result and to augment the precision, but this comes at the cost of an increased computational complexity, due to the large number of constraints to be considered. We presented a novel solution to this problem, by simultaneously optimizing the left and right proximity field, using the methodology of the Augmented Lagrangian. Experiments show that the quality of the results using the proposed methodology far exceeds those of more traditional global optimization - based methods. This allows us to conclude, that, also for the binocular case, the proposed 3D reconstruction algorithm presents an excellent reconstruction tool.

REFERENCES