INTEGRATING PRIOR KNOWLEDGE AND STRUCTURE FROM MOTION

N. Guilbert, H. Aanæs and R. Larsen

IMM
Technical University of Denmark (DTU)
2800 Lyngby, Denmark
E-mail: proj61@imm.dtu.dk

ABSTRACT

A new approach for formulating prior knowledge in structure from motion is presented, where the structure is viewed as a 3D stochastic variable, hereby priors are more naturally expressed. It is demonstrated that this formulation is efficient for regularizing structure reconstruction via prior knowledge. Specifically algorithms for imposing priors in the proposed formulation are presented.

1. INTRODUCTION

Structure from motion is a field of research which allows the reconstruction of the 3D structure and motion of a rigid body from a sequence of two or more images where 2D features of interest have been detected and correlated through the frames, see e.g. [3]. It is an essential tool in artificial intelligence and image understanding, applicable in robot vehicle navigation, content-based searches in video data or architectural visualization.

However, it is an inherent property of the approach, that the accuracy of 3D reconstruction is underpinned by the noise in the images, be it due to quantization or shortcomings of the feature extraction algorithm. Hence, enhancing the 3D structure via prior knowledge can improve the results considerably. For instance, knowing that a 3D object primarily consists of planes in right angles to each other, like eg. buildings, can greatly enhance its reconstruction. One might also envisage integrating a more sophisticated models along the line of [1, 2], whereby more complex priors can be expressed and applied.

Previously, Baker et al. [6] and Torr et al. [5] have addressed the problem of incorporating prior knowledge into structure from motion. These approaches identify *layers* in the images and constrain the reconstruction accordingly. These methods use backprojection onto the image(s) to determine whether regularization with a given prior is probable. Consequently, the noise is expressed in the 2D image domain, and it is implied that the camera matrices are determined perfectly.

In this paper, we address the problem of combining priors with structure from motion by formulating the estimated reconstruction as 3D a stochastic variable. This approach has the advantage of being intuitively accessible, since the 3D structure is computed explicitly. The explicite reconstruction also allows handling of the uncertainty of the camera motion, since our approach can incorporate that the noise on a given 2D feature potentially affects all of the 3D features. Finally, integration of more complex priors than planes (eg. splines or deformable templates) is natural.

We establish the mean of the 3D structure by standard structure from motion methods, see e.g. [3]. The dispersion of the resulting structure is then obtained by approximating the 2D to 3D transformation by a linear function, whereby the relations between noise on the 2D features and the 3D structure becomes apparent. In order to validate our approach, we then impose the prior of planar surfaces into structure from motion, applying a clustering technique.

The organisation of this paper is as follows: section 2 describes the stochastic structure modeling approach, section 3 describes the algorithm for fitting a plane to a set of points given heteroscedastic and anisotropic noise. Section 4 concerns the clustering of smaller planes into the final plane estimates.

2. STOCHASTIC STRUCTURE MODEL

A mentioned, we view the estimated 3D structure as a stochastic variable, derived from the the tracked 2D feature points in two images. We assume a calibrated pinhole camera:

$$\begin{bmatrix} x_{ij} \\ s_{ij} \end{bmatrix} = \mathbf{A}_i \begin{bmatrix} \mathbf{R}_i t_i \\ 0 \;\; 1 \end{bmatrix} \begin{bmatrix} X_j \\ 1 \end{bmatrix} = \mathbf{A}_i \mathbf{P}_i \begin{bmatrix} X_j \\ 1 \end{bmatrix} \qquad (1)$$

where the 2D feature x_{ij}, is the projection of the 3D feature X_j, in image i. Image i is described by its calibration matrix $\mathbf{A}_i$, its rotation $\mathbf{R}_i$ and translation t_i relative to the world coordinate system.

From the tracked 2D features, and this observation model (1), it is well known that the 3D structure, X_j and the camera motion, P_i, can be calculated via the epipolar geometry, see e.g. [3].

As such, assuming a well posed problem, the 3D structure:

$$\mathbf{X} = [X_1, \ldots, X_n]$$

can thus be seen as a function, $\mathcal{F}$, of the tracked 2D features:

$$\mathbf{x} = [x_{11}, \ldots, x_{1n}, x_{21}, \ldots, x_{2n}]$$

i.e.

$$\mathcal{F} : \mathbf{x} \to \mathbf{X}$$

So viewing $\mathbf{X}$ and $\mathbf{x}$ as stochastic variables with assumed Gaussian noise such that:

$$X_j \in N(\hat{X}_j, \Sigma_j) \quad x_{ij} \in N(\hat{x}_{ij}, \Sigma_{ij}^{2D})$$

we would like to estimate $\mathbf{X}$ given $\mathbf{x}$. Since we do not have $\mathcal{F}$ in closed form, we approach this problem by linearizing $\mathcal{F}$, by means of numerical gradients. Hence:

$$\mathbf{X} \in \mathcal{F}(\mathbf{x}) \approx \mathcal{F}(\hat{\mathbf{x}}) + \sum_k \frac{\delta \mathcal{F}}{\delta x_k} N(0, \Sigma_{ij}^{2D}) \Rightarrow (2)$$

$$Var(\mathbf{X}) \approx \sum_k \left(\frac{\delta \mathcal{F}}{\delta x_k} \right)^2 Var(x_k) \qquad (3)$$

$$E(\mathbf{X}) = \mathcal{F}(\mathbf{x}) \qquad (4)$$

denoting the composing features of $\mathbf{x}$ by x_k – i.e. k is valid configurations of ij – and assuming that the noise on $\mathbf{x}$ is independent. Hence from (3) the Gaussian noise structure of $\mathbf{X}$ can be derived.

As noted, this derived noise structure of $\mathbf{X}$, is based on the approximation of (independent) Gaussian noise, and of linearizing $\mathcal{F}$. These assumptions seam reasonable, in that the approach i capable of capturing, how variations in the image data, $\mathbf{x}$, affects the estimated structure, $\mathbf{X}$. These affects on the estimated structure, $\mathbf{X}$, are all what is reasonable to expect, since the noise structure on the 2D features is rarely known and as such all distributions – e.g. Gaussian – are approximations.

Our assumption of a calibrated camera – $\mathbf{A}_i$ known – is by no means necessary, as seen in [3], except if the imposed priors are non invariant to protective transformations, which they will seldom be.

3. ESTIMATING THE PLAN (PRIOR)

In order to test if a set of given 3D features lie on the same plane, and if so enforce that plane upon the the structure, we need to estimate the most likely plane. In this case it is non–trivial, since the 3D features have different or heteroscedstisk anisotropic noise. Implying that the Mahalanobis distance measure or norm for each 3D feature is different.

Hence given m 3D features $\{X_1, \ldots, X_m\}$, and corresponding Gaussian variance structures $\{\Sigma_1, \ldots, \Sigma_m\}$ we want to find the plane that minimizes the distance between 3D feature X_j and the plane, in the norm induced by Σ_j^{-1}. Let a plane be denoted by its normal vector, π, and its offset from origo, α_0, then any point, X, on the plane satisfies:

$$\pi^T \cdot X + \alpha_0 = 0$$

The most likely plane is estimated by iteratively approximating the noise structures by heteroscedastic *isotropic* noise, until convergence is achieved. In this isotropic case, the most like plane is given by:

$$\alpha_0 = -\frac{\sum_{j=1}^{m} \sigma_j X_j}{\sum_{j=1}^{m} \sigma_j}$$

$$\pi = \min_{\pi'} \sum_{j=1}^{m} ||\sigma_j (X_j + \alpha_0)^T \cdot \pi'||_2^2 \qquad (5)$$

where σ_j denotes the isotropic variance of Σ_j. Given a plane, the σ_j are given as the ratio between the length of the minimum distance between 3D feature and plane, $\tilde{r}_j$, in the induced norm relative to the minimum distance in the 2-norm, r_j, i.e.

$$\sigma_j = \frac{\tilde{r}_j^T \Sigma_j^{-1} \tilde{r}_j}{r_j^T r_j} \qquad (6)$$

In short, the algorithm is, where q denotes iterations:

1. **Initialize** $q = 0$, $\forall j \ \sigma_j^0 = det \Sigma_j$.
2. **Estimate Plane** $\pi^q, \alpha_0^q$with isotropic noise, for (5).
3. **Update Isotropic Noise**σ_j^q via (6)
4. **If not Stop**, $q = q + 1$, goto 2. The stop criteria is

$$\max_j (\sigma_j^q - \sigma_j^{q-1}) < tolerance.$$

It is seen, that the σ_j are updated such that the given plane has the right likelihood, with the original heteroscedstic *an*isotropic noise. So if $\forall j \ \sigma_j^q = \sigma_j^{q-1}$ then the optimal plane with the isotropic noise is also optimal in the *an*isotropic case.

4. IMPOSONG THE PRIOR

As mentioned, we validate our approach by imposing the prior of planar surfaces onto our structure. This is done by first triangulating the estimated 3D structure, whereby a set of three–point planes are constructed. A recomended methods for triangulation is [4] by Morris and Kande.

As such, this triangulation does not reguaralize the 3D structure, but it serves as a initialization for an algorithm where neighboring planes (e.g triangles) which are likely to be the same plane are merged. A statistical test for coplanarity is described below.

4.1. Test for Coplanarity

Given a set of 3D features, as described above, located on a plane, and this plane is estimated optimally as described in Section 3, then the residuals are realizations of the respective noise, $N(0, \Sigma_j)$, as estimated in Section 2. Hence a test of two sets of 3D features $\mathbf{X}_1, \mathbf{X}_2$ being located in the same

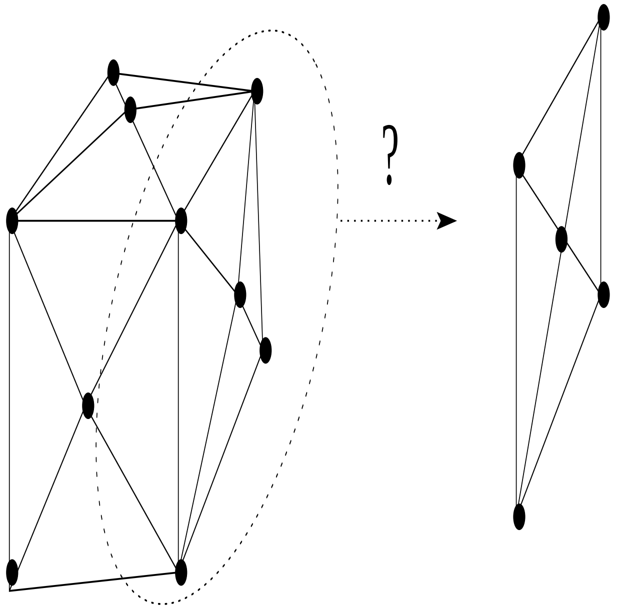

Fig. 1.

plane is to compare the residuals for the plane fitted to one of the sets and the plane fitted to the two sets combined.

More formally, the entities that should be compared are the residuals, $\tilde{r}_j$ between a 3D feature and the plane, normed with the respective variance, Σ_j. Since, if the assumption of coplanarity holds, these residuals should be elements in a $N(0,1)$ distribution. Hence a F-test can be used, or formally if the assumption that $\mathbf{X}_1$ and $\mathbf{X}_2$ are part of the same plane holds, then:

$$\frac{m}{n}\frac{\sum_{i=1}^{n} \tilde{r1}_j^T \Sigma_j^{-1} \tilde{r1}_j}{\sum_{i=1}^{m} r\tilde{12}_j^T \Sigma_j^{-1} r\tilde{12}_j} \in \mathbf{F}(n,m) \tag{7}$$

where $\tilde{r1}$ are the n residuals from $\mathbf{X}_1$ and $r\tilde{12}$ are the m residuals from $\mathbf{X}_1 \cup \mathbf{X}_2$.

4.2. Clustering Algorithm

The algorithm for combining neighboring planes is a greedy clustering algorithm in that at any given time, the two neighboring planes with the must likelihood of being the same plane are combined, if this likelihood is above a given threshold. Hence the triangles achieved by triangularization, are clustered into planes.

When these clusters have been calculated, the derived planes are enforced upon the 3D features, and the 3D features moved onto the plane, whereby the prior of planar structures is imposed. See **??**

5. RESULTS

6. CONCLUSION

7. ACKNOWLEDGMENTS

Street respect to Allan Åsbjerg Nielsen for enlightened discussions on plane fitting with heteroscedastisc anisotropic noise.

8. REFERENCES

[1] A. Blake and M. Isard. *Active Contours*. Springer–Verlag, London, UK., 1998. 352 pp.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision, Graphics and Image Processing*, 61(1):38–59, January 1995.

[3] R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK, 2000.

[4] D. Morris and T. Kanade. Image-consistent surface triangulation. In *IEEE Conf. Computer Vision and Pattern Recognition'2000*, pages 332–338, 2000.

[5] A. R. Dick P. H. S. Torr and R. Cipolla. Layer extraction with a bayesian model of shape. In *European Conf, Computer Vision'00*, pages 273–289, June 2000.

[6] R. Szeliski S. Baker and P. Anandan. A layered approach to stereo reconstruction. In *IEEE Conf. Computer Vision and Pattern Recognition'98*, pages 434–441, 1998.