MVA2000 IAPR Workshop on Machine Vision Applications, Nov. 28-30, 2000, The University of Tokyo, Japan

13—24

A Method for Estimating Multiple Motion Parameters and Planar Surface Parameters without Feature Points Correspondence

Akihiro Minagawa * Yutaka Horie * Norio Tagawa * Toshiyuki Tanaka *
Graduate School of Engineering
Tokyo Metropolitan University

Abstract

A method estimating motion and planar surface parameters based on pixel intensity is proposed. Conventional approaches for estimation of these parameters are based on detected edge or feature points correspondence. Thus the problem is one of searching for any point which corresponds to feature points in other images. These approaches provide a solution from sparse points, however the shape cannot be estimated using all the pixels in an image. In this paper, by assuming that the multiple planar surfaces consist of objects, we propose a method that can estimate the depth of all pixels in input images without feature points correspondence. This approach places no restriction on the number of input images, which is different from conventional stereo vision with the exception of the factorization method. In the proposed method, parameters can be estimated as the ML (maximum likelihood) estimator, and the depth as the MAP (maximum a posteriori) estimator.

1 Introduction

Motion parameters estimation and the three-dimensional depth recovery problem are two of the most important problems in the field of computer vision. Recently, these two problems have been combined with the camera calibration problem, and several methods have been proposed. However, these methods mainly depend on feature points such as edges, zero-cross points and so on. Tomasi and Kanade have proposed a fast and simple method known as the factorization method, which is the most basic method[8] of all the methods proposed to date. Although several extensions to this method have been proposed[2],[9], methods based on the factorization method work well only for images with many feature points with correspondence. Therefore, when discussing these methods, many feature points with precise correspondence are required. Note that there is a tradeoff between the number of feature points and the accuracy of correspondence. Another method proposed by Szeliski[7] generates a Mosaic, in which objects consist of a planar surface and motion are estimated from the all the pixels in images. This method enables high density estimation. In this paper, we extend the method proposed by Szeliski to multiple planar surfaces and multiple input images in a Bayes' framework. We begin, in Section 2, by defining a probabilistic model. Since this model includes the hidden label which represents the owner plane, this problem is in NP-complete. Therefore, in Section 3 we propose an algorithm that uses the EM (Expectation Maximization) algorithm including MFA (mean field approximation) for the reduction of the computational cost. In Sections 4, 5 and 6, the effectiveness of the proposed method including these techniques is confirmed by experiment and discussed.

2 Definition of probabilistic models

Figure 1 shows the motion parameters of frame k, in which the camera moves by rotation R_k and shift $\boldsymbol{s}_k$ to the position of frame 0. R_k is a 3×3 matrix and $\boldsymbol{s}_k \equiv s(k) = [A_k, B_k, C_k]^\top$ is a 3×1 vector, where the subscript $\top$ denotes the transpose of the vector or matrix. When the planar surface j is represented as $Z_0 = p_j\, X_0 + q_j\, Y_0 + r$, the translational matrix T_j with a pinhole camera is described as

$$wT_{jk} = \left[r_j I + \boldsymbol{m}_j \boldsymbol{s}_k^\top \right] R_k \tag{1}$$

where w is the scale ambiguity. These parameters can be described as $\Theta_{t_k} = [A_k, B_k, C_k, \theta_k, \phi_k, \psi_k]^\top$, $\Theta_{p_j} = [p_j, q_j, r_j]$, where k represents the frame number, and j represents the plane number. Then, assuming that the intensity g_{ik} of pixel i in input image k has noise $\mathcal{N}(0, \sigma_k)$, the probability density function (p.d.f.) conditioned by planar surface j is defined as,

$$p(g_{ik} | \boldsymbol{L}_{ik}; \Theta_{t_k}, \Theta_{p_j}, \sigma_k^2)$$

*Address: 1-1 Minami-osawa, Hachioji, Tokyo 192-0397 Japan. E-mail: {akihiro@, yutaka@elena., tagawa@, tanaka@}eei.metro-u.ac.jp.ac.jp

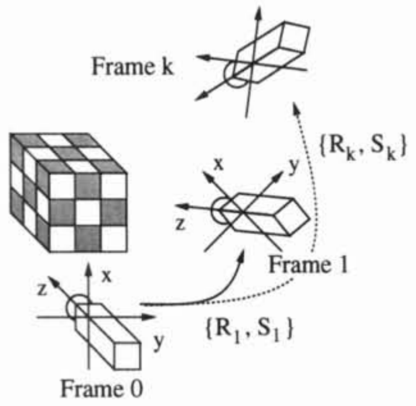


Figure 1: Multiple-Planar-Surface and Motion Parameters Estimation

$$= \frac{1}{2\pi\sigma_k^2} \exp - \frac{L_{ijk}(g_{ik} - g'_{(ijk)})^2}{2\sigma_k^2}, \qquad (2)$$

where, $g'_{(ijk)}$ represents the pixel intensity of the point in image 0 for which the pixel i of image k is transformed inversely according to planar surface j, and $L_{(ijk)} \in \{0, 1\}$ is the jth component of label vector $\boldsymbol{L}_{ik}$ which represents the label that describes the owner-plane, the sum of the components of which is equal to 1.

In this case, for the prior of labels $\{\boldsymbol{L}_{ik}\}$, in addition to considering spatial connection using MRF, temporal restriction is also considered. Thus, a hierarchical labeled model is proposed. Figure 2 shows the networks of this model. First, label $\boldsymbol{L}_{ik}$ and $\boldsymbol{L}'_i$ is provided, where, $\boldsymbol{L}_{ik}$ is considered as the label of input image k excluding 0, and $\boldsymbol{L}_{ik}$ as the label of input image 0. Then, the pixels of image 0 and image k connected by the transformation decided by parameter $\{\Theta_{t_k}, \Theta_{p_j}\}$ should be labeled as the same plane. This means that both label $\boldsymbol{L}_{ik}$ and $\boldsymbol{L}'_{m(ijk)}$ of pixels connected by the current parameter take similar vectors. The relationship of this can be written as,

$$p(\boldsymbol{L}_{ik}|\boldsymbol{L}'; \Theta_{t_k}, \Theta_p, \gamma_k) = \frac{\exp \sum_j \gamma_k L_{ijk} L'_{(ijk)j}}{\sum_{\{\boldsymbol{L}_{ik}\}} \exp \sum_{j'} \gamma_k L_{ij'k} L'_{(ij'k)j'}}. \qquad (3)$$

Moreover, as spatial connection MRF is defined as

$$p(\boldsymbol{L}'; \beta) = \frac{1}{Z_L} \exp \sum_{m \in \mathcal{I}_0} \sum_{m' \in \mathcal{R}_m} \beta \boldsymbol{L}'^{\top}_m \boldsymbol{L}'_{m'}. \qquad (4)$$

where, Z_L represents a partition function, $\mathcal{I}_0$ represents the image area in input image k, and $\mathcal{R}_m$ represents the neighborhood of pixel m.

This model is equal to HME (hierarchical mixture of experts) proposed by Jordan[4] under the MFA, and thus has a theoretical background.

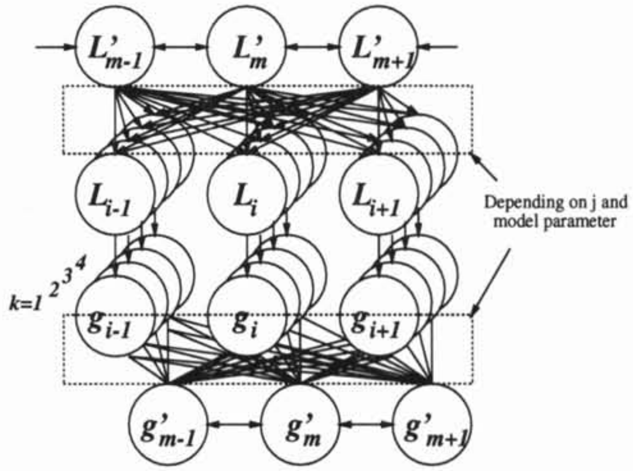


Figure 2: Hierarchical label networks

The ML estimator can be found by maximizing the following likelihood function with respect to the parameter

$$\{\hat{\Theta}_{t_k}, \hat{\Theta}_{p_j}, \{\hat{\sigma}_k^2\}, \hat{\beta}, \{\hat{\gamma}_k\}\} = \arg \left[\max L(\Theta_{t_k}, \Theta_p, \sigma_k^2, \beta, \{\gamma_k\}; \boldsymbol{g}_k) \right] \quad (5)$$

where, the likelihood function can be obtained from Eqns. 2,3, and 4 by taking the marginal distribution with respect to $\boldsymbol{L}_{ik}$ and $\boldsymbol{L}'_m$ as follows

$$p(\boldsymbol{g}; \Theta_t, \Theta_p, \{\sigma_k^2\}, \beta, \gamma_k) = \sum_{\{\boldsymbol{L}'\}} \left[\prod_k \prod_i \left\{ \sum_{\{\boldsymbol{L}_{ik}\}} p(\boldsymbol{g}_{ik}|\boldsymbol{L}_{ik}; \Theta_{t_k}, \Theta_p, \sigma_k^2) \times p(\boldsymbol{L}_{ik}|\boldsymbol{L}'; \gamma_k) \right\} p(\boldsymbol{L}'; \beta) \right]. \qquad (6)$$

When the ML estimator $\{\hat{\Theta}_t, \hat{\Theta}_p, \{\hat{\sigma}_k^2\}, \hat{\beta}, \{\hat{\gamma}_k\}\}$ can be found, the MAP estimator can be obtained by maximizing the posterior probability

$$p(\boldsymbol{L}, \boldsymbol{L}'|\boldsymbol{g}; \hat{\Theta}_t, \hat{\Theta}_p, \{\hat{\sigma}_k^2\}, \hat{\beta}, \{\hat{\gamma}_k\}) = \frac{p(\boldsymbol{g}|\boldsymbol{L}; \hat{\Theta}_t, \hat{\Theta}_p, \{\hat{\sigma}_k^2\}) p(\boldsymbol{L}|\boldsymbol{L}'; \{\hat{\gamma}_k\}) p(\boldsymbol{L}'; \hat{\beta})}{p(\boldsymbol{g}; \hat{\Theta}_t, \hat{\Theta}_p, \{\hat{\sigma}_k^2\}, \hat{\beta}, \{\hat{\gamma}_k\})}. \qquad (7)$$

Only the numerator needs to be maximized, in order to maximize the above probability . Thus the MAP estimator $\{\boldsymbol{L}_{ik}, \boldsymbol{L}'_m\}$ can be obtained by maximizing the numerator only.

3 Calculation methods

The maximization of Eqn. 7 involves an exponential number of terms, thus ways to reduces the computational cost must be considered. To solve this problem, We apply MFA[1],[10], and Factorizable forms of the posterior probability is obtained. Then the divergence between the real form and factorizable form is measured. Thus, the factorizable form which minimizes the divergence can be regarded as the optimal approximation. Minimizing the divergence generates an equation called the mean field equation. In the proposed model, two nested equations are generated. To solve these equations iteratively, the mean of each label can be calculated from the minimization of the divergence as

$$\bar{L}_{ijk} = \frac{\exp\left(-w_{ijk} + \gamma_k \bar{L'}_{mj}\right)}{\sum_{j'} \exp\left(-w_{ijk} + \gamma_k \bar{L'}_{mj'}\right)}, \tag{8}$$

$$\bar{L'}_{mj} = \frac{\exp\left(\sum_k \gamma_k \bar{L}_{ijk} + 2\sum_{m'} \beta \bar{L'}_{m'j}\right)}{\sum_{j'} \exp\left(\sum_k \gamma_k \bar{L}_{ij'k} + 2\sum_{m'} \beta \bar{L'}_{m'j'}\right)}. \tag{9}$$

In addition, maximizing Eqn. 6 is also difficult. Thus, we apply EM algorithm[3],[5]. Here, the detail of the algorithm is omitted due to the restrictions of this paper. In the E-step (Expectation step), using the value calculated by MFA from Eqns. (8) and (9), the conditional expected function of the complete data $\{\boldsymbol{g}, \boldsymbol{L}, \boldsymbol{L'}\}$ can be generated. In the M-step (Maximization step) the conditional expected function generated in the E-step is maximized with respect to each parameter. For a detailed explanation of the proposed algorithm, see [6].

4 Simulation Results

Figure 3 shows the input image frame 0 used for the simulation. This image is 280 × 280(pix.), and focus is 400(pix.). The number of planes is three and the number of frames is five. The camera moves from frame 0 by rotation R_k and shift $\boldsymbol{s}_k$, and images are taken at each camera point. The noise is $\sigma^2 = 9$ for correct intensity of these images, and these images are used as the input images. Note that each frame k (k=1...5) does not cover all the area of frame0. Using these six input images, model parameters are estimated and each pixel is clustered. Figure 4(a) shows the clustering results of the first iteration and Fig. 4(b) shows the convergence results (after 150 iterations of the EM algorithm) of clustering of each pixel for frame 0. Since the initial probability of each pixel to each plane is the same, the clustering is disconnected at the results of the first iteration. However, the convergence results show that almost all pixels are clustered to the correct plane, and thus prove that this algorithm is effective for clustering.

Figure 5 shows the input images for frame k (k=1...5), Fig. 6(a) shows the clustering results of the first iteration of the clustering and Fig. 6(b) shows the clustering results of each input image of frame k. The clustering results of frame k are very similar to the results of clustering for frame 0, indicating the validity of the proposed model.

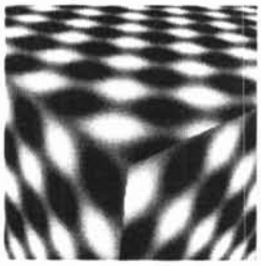
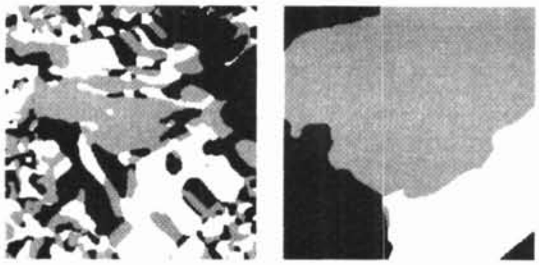


Figure 3: Input Image Frame 0



(a) First Iteration Results (b) Convergence Results

Figure 4: Clustering Results for Frame 0

5 Experimental Results

Next, we apply the proposed this algorithm to real images. The number of planes is three and the number of frames is three. All the images used in this experiment were obtained using a CCD camera. Figure 7(a) shows the input images of frame 0 (270 × 270 pix.) and clustering results. Figure 7(b) shows the clustering results of first iteration for frame 0. Figure 7(c) shows the convergence results of clustering for frame 0. As shown in Fig. 7(c), almost all pixels are clustered correctly. Since all the parameters are estimated based on this clustering results, it can be said that the proposed algorithm provides a good estimation of each parameter. These results prove that the proposed algorithm can be applied to real images.

6 Conclusion

In this paper, an estimation method for multiple-planar-surface parameters and multiple motion parameters has been proposed. The estimation method is derived by solving a hierarchical probabilistic model constructed to handle multiple input images. To solve this model, the EM algorithm with

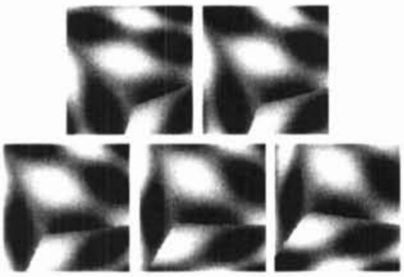
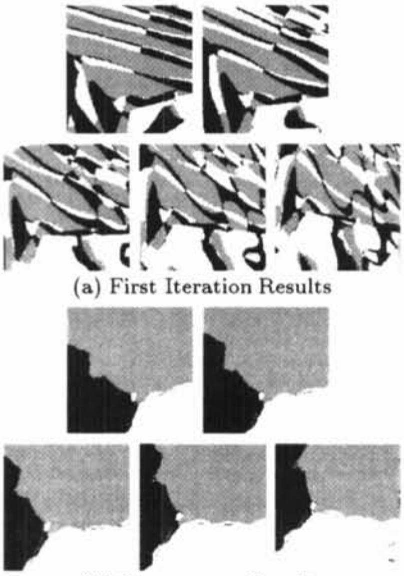


Figure 5: Input Images of Frame k



(a) First Iteration Results

(b) Convergence Results

Figure 6: Clustering Results for Frame k

MFA is applied. We have shown that the frameworks derived in this approximation are the same as hierarchical bayesian expert networks, and the effectiveness and stable convergency of the proposed method is confirmed using real images.

Acknowledgements

The authors would like to express thier appreciation to Sun Microsystems Inc. in Japan and Prof. Teramachi of Polytechnic University for their support in respect of computational environment.

References

[1] G. L. Bilbro, W. Snyder, S. Garnier, and J. Gault. Mean field annealing: A formalism for constructing GNC-like algorithms. *IEEE Trans. on Neural Networks*, vol. 3, no. 1, pp. 131–138, 1992.

[2] K. Deguchi. Factorization method for structure from perspective multi-view images. *IEICE Trans. of Japan, Inf. & syst.*, vol. E81-D, no. 11, pp. 1281–1289, 1998.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[4] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, no. 6, pp. 181–214, 1994.

[5] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering.* Deckker, New York, 1988.

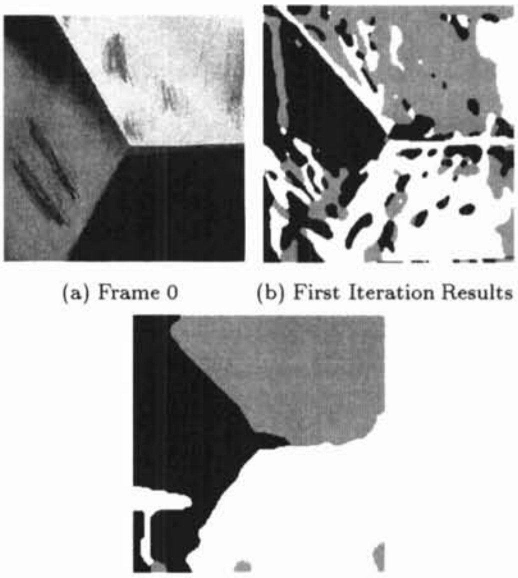
[6] A. Minagawa, Y. Horie, N. Tagawa, and T. Tanaka. Parameters estimation of multiple motions and planar surfaces based on pixel intensity. *Tech. Report of IEICE*, no. PRMU2000-13, pp. 7–14, 2000.

[7] R. Szeliski. Video mosaics for virtual environments. *IEEE Trans. on Comp. Graph. and Appl.*, vol. 16, no. 2, pp. 22–31, 1996.

[8] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[9] T. Ueshiba and F. Tomita. Factorization method for projective and euclidean reconstruction from multiple perspective views via iterative. *Proc. 5th ECCV*, vol. 1, pp. 296–310, 1998.

[10] J. Zhang. The mean field theory in EM procedures for Markov random fields. *IEEE Trans on Signal Processing*, vol. 40, no. 10, pp. 2570–2583, 1992.



(a) Frame 0

(b) First Iteration Results

(c) Convergence Results

Figure 7: Real Image and Results