

Direct 3-D Shape Recovery from Image Sequence Based on Multi-scale Bayesian Network

Norio Tagawa Junya Kawaguchi Shoichi Naganuma Kan Okubo
Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
tagawa@sd.tmu.ac.jp

Abstract

We propose a new method for recovering a 3-D object shape from an image sequence. In order to recover high-resolution relative depth without using the complex Markov random field (MRF) that includes a line process, we construct a recovery algorithm based on a belief propagation scheme using a multi-scale Bayesian network. With this algorithm, relative 3-D motion between a camera and an object can be determined together with relative depth, and the maximum a posteriori expectation-maximization (MAP-EM) algorithm is effectively used to determine a suitable approximation.

1. Introduction

We propose a method for obtaining 3-D depth information using a gradient based scheme with two successive images. In this field of study, spatially dense and stable detection is strongly required [1]–[3], and the aperture problem and the alias problem need to be completely solved [4]. Usually, either local optimization or global optimization is used to avoid the aperture problem. To avoid the alias problem, components of low spatial frequency are extracted by low-pass filtering and used to compute optical flow. However, these techniques lower the resolution of the obtained optical flow and hence, relative depth.

In this study, we attempt to directly recover 3-D depth information without explicitly detecting optical flow, and apply Bayesian network spreading to a resolution direction by decomposing the original image into multi-scale images. Unknown parameters are represented as a node as well as depth to be estimated and observed image information. We call this graphical model a multi-scale Bayesian network. If the parameters, including relative 3-D motion parameters, are determined in advance, the inference of depth in this

network is realized by Kalman filtering. Especially, for optical flow detection, Simoncelli [4] introduced the multi-scale Bayesian network, which considers optical flow as a node with parameters assumed to be known, and proposed the Kalman filter-based algorithm. In our study, we attempt to estimate the depth and parameters simultaneously from observations.

The parameters to be estimated are common to all multi-scale images, and hence, we have to adopt a suitable approximation to simplify the inference. In most tractable approximations, the parameters are considered to be independent between the multi-scale images. However, the information for the parameters obtained in a low-resolution image is not directly propagated; that is, it is implicitly propagated through the propagation of depth information. We propose a stable procedure using the maximum a posteriori expectation-maximization (MAP-EM) algorithm, which can directly propagate the parameters' information.

2. Gradient method for recovering depth

2.1. Projection model and optical flow

We use perspective projection as our camera-imaging model. The camera is fixed with an (X, Y, Z) coordinate system, where the viewpoint (lens center) is at origin O and the optical axis is along the Z -axis. The projection plane (image plane) $Z = 1$ can be used without any loss of generality, which means that the focal length equals 1. A space point (X, Y, Z) on the object is projected to image point (x, y) . At each (x, y) , the optical flow $[v_x, v_y]^T$ is formulated with an inverse depth $d(x, y) \equiv 1/Z(x, y)$ and the camera's translational and rotational vectors $\mathbf{u} = [u_x, u_y, u_z]^T$, and $\mathbf{r} = [r_x, r_y, r_z]^T$, respectively, as follows:

$$v_x = xy r_x - (1 + x^2) r_y + y r_z - (u_x - x u_z) d, \quad (1)$$

$$v_y = (1 + y^2) r_x - xy r_y - x r_z - (u_y - y u_z) d. \quad (2)$$

In the above equations, d is an unknown variable at each pixel position, and \mathbf{u} and \mathbf{r} are unknown common parameters. In the following, Eqs. 1 and 2 are rewritten as follows:

$$v_x = v_x^r(\mathbf{r}) + v_x^u(\mathbf{u})d, \quad (3)$$

$$v_y = v_y^r(\mathbf{r}) + v_y^u(\mathbf{u})d. \quad (4)$$

2.2. Gradient equation for rigid motion

The optical flow constraint equation, which is called the ‘‘gradient equation,’’ is the 1st approximation of the assumption that intensity is invariable with respect to the relative 3-D motion. At each pixel (x, y) in the images, the gradient equation is formulated with the partial derivatives f_x , f_y and f_t (where t denotes time) of the image intensity $f(x, y, t)$ and the optical flow, as follows:

$$f_t = -f_x v_x - f_y v_y. \quad (5)$$

By substituting Eqs. 3 and 4 into Eq. 5, the gradient equation representing rigid motion can be derived explicitly.

$$\begin{aligned} f_t &= -(f_x v_x^r + f_y v_y^r) - (f_x v_x^u + f_y v_y^u)d \\ &\equiv -f^r - f^u d. \end{aligned} \quad (6)$$

In this study, we use an integration kernel defined as a derivative of the Gaussian function to calculate f_x and f_y accurately, and f_t is detected as the finite difference using two successive frames. Hence, we suppose that only f_t contains the observation error, and we use Eq. 6 as the observation equation.

3. Depth recovery on the Bayesian network

3.1. Probabilistic model

In the proposed method, each image is decomposed into several images having different resolutions, and the depth information estimated using the low-resolution image is propagated to the high-resolution depth estimation in the belief propagation (BP) framework.

In the estimation process at each resolution image, we adopt a local optimization strategy with low computational cost to avoid the aperture problem. Therefore, we assume that the depth in the spatial local region is constant. We let $l = 1, 2, \dots, L$ be the index indicating the resolution, and $l = 1$ be the lowest resolution. Due to the differences of resolution, the size of the above-mentioned local region N_l needs to satisfy the relation $N_{l_1} < N_{l_2}$ for the case that $l_1 > l_2$.

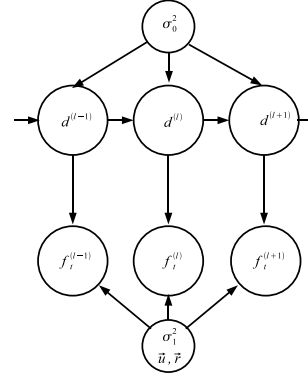


Figure 1. Bayesian network representation of the probabilistic model.

The resolution of $d^{(l)}$ becomes finer as l increases, and so the transition from $d^{(l)}$ to $d^{(l+1)}$ is modeled using a linear interpolation operator $E^{(l)}$, as follows:

$$d^{(l+1)} = E^{(l)} d^{(l)} + n_0^{(l)}, \quad (7)$$

where $n_0^{(l)}$ is independent Gaussian noise with zero mean and variance σ_0^2 , which is common to resolution l . Based on this model, the probability density of $d^{(l+1)}$ is also a Gaussian distribution with mean $\bar{d}^{(l+1)} = E^{(l)} \bar{d}^{(l)}$ and variance $\sigma_d^{2(l+1)} = E^{(l)2} \sigma_d^{2(l)} + \sigma_0^2$, which are recurrently represented. The covariances of $d^{(l)}$ and its spatial neighbors are omitted. $E^{(l)2}$ is the linear interpolation operator, and each weight coefficient of $E^{(l)2}$ is defined as the square of the corresponding weight coefficient in $E^{(l)}$.

The observation equation with respect to $f_t^{(l+1)}$, based on Eq. 6, is assumed as follows:

$$f_t^{(l)} = -f^r^{(l)} - f^u^{(l)} d^{(l)} + n_1^{(l)}. \quad (8)$$

By assuming $n_1^{(l)}$ as the Gaussian random variable with zero mean and variance σ_1^2 which is common to l , the conditional probability density of $f_t^{(l)}$ is a Gaussian distribution with mean $-f^r^{(l)} - f^u^{(l)} d^{(l)}$ and variance σ_1^2 .

The probabilistic variables $d^{(l)}$ and $f_t^{(l)}$ can be represented as the Bayesian network shown in Fig. 1. In this network, the parameters $\Theta \equiv \{\mathbf{u}, \mathbf{r}, \sigma_0^2, \sigma_1^2\}$, which are also shown, are regarded as probabilistic variables estimated through BP, described in the following section.

3.2. BP for depth and 3-D motion

In the Bayesian network shown in Fig. 1, the marginal posterior $p(d^{(L)} | \{f_t^{(L)}\}, \dots, \{f_t^{(1)}\})$

must be computed successively through $p(d^{(l)}|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\})$ ($l = 1, \dots, L - 1$). The symbol $\{a\}$ indicates the set of a for all pixels. If the values of the parameters, including the 3-D motion parameters, are known in advance, this propagation process can be accomplished by computing only the mean $\tilde{d}^{(l+1)}$ and the variance $\tilde{\sigma}_d^{2(l+1)}$ of $p(d^{(l+1)}|\{f_t^{(l+1)}\}, \dots, \{f_t^{(1)}\}, \Theta)$ successively, since $p(d^{(l+1)}|\{f_t^{(l+1)}\}, \dots, \{f_t^{(1)}\}, \Theta)$ is Gaussian. Although discussions of $\tilde{d}^{(l+1)}$ and $\tilde{\sigma}_d^{2(l+1)}$ are omitted in this paper due to space limitations, these can be analytically formulated and $\tilde{d}^{(l)}$ can be regarded as a minimum variance unbiased estimator $\hat{d}^{(l)}$. To avoid the alias problem in the formulation of $\tilde{d}^{(l+1)}$, the following definition of $f_t^{(l+1)}$ is adopted using $\hat{v}^{(l)}$ estimated based on $\hat{d}^{(l)}$, as follows [4]:

$$f_t^{(l+1)} = -f_x^{(l+1)}E^{(l)}\hat{v}_x^{(l)} - f_y^{(l+1)}E^{(l)}\hat{v}_y^{(l)} + \frac{\partial}{\partial t}\mathcal{W}(f^{(l+1)}, E^{(l)}\hat{v}^{(l)}). \quad (9)$$

In Eq. 9, finite difference is used instead of the partial derivative $\partial/\partial t$, and image warping is defined as

$$\mathcal{W}(f, \mathbf{v})(\mathbf{x}, t + \delta t) \equiv f(\mathbf{x} - \mathbf{v}\delta t, t + \delta t). \quad (10)$$

The abovementioned recurrent processing is consistent with the Kalman filter, but the true values of Θ have to be known in advance. In order to determine Θ properly, the BP for Θ is also needed. The formulation at $l + 1$ resolution can be written recurrently, as follows:

$$\begin{aligned} & p(\Theta|\{f_t^{(l+1)}\}, \{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}) \\ &= \frac{1}{Z} \int p(\{f_t^{(l+1)}\}|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}, \Theta, \{d^{(l+1)}\}) \\ & \quad \cdot p(\{d^{(l+1)}\}|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}, \Theta) \\ & \quad \cdot p(\Theta|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}) d\{d^{(l+1)}\} \\ &= \frac{1}{Z} p(\{f_t^{(l+1)}\}|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}, \Theta) \\ & \quad \times p(\Theta|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}), \end{aligned} \quad (11)$$

where $Z = p(\{f_t^{(l+1)}\}|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\})$. In this recurrent computation, the variance of the initial probabilistic density $p(\Theta)$ can be sufficiently large, if there is no prior for Θ .

For the exact BP with respect to $d^{(l)}$, $p(d^{(l+1)}|\{f_t^{(l+1)}\}, \dots, \{f_t^{(1)}\})$ must be computed using Eq. 11 and $p(d^{(l+1)}|\{f_t^{(l+1)}\}, \dots, \{f_t^{(1)}\}, \Theta)$, for which the mean and the variance are obtained as

$$p(d^{(l+1)}|\{f_t^{(l+1)}\}_{\mathcal{N}}, \{f_t^{(l)}\}, \dots, \{f_t^{(1)}\})$$

$$\begin{aligned} &= \int p(d^{(l+1)}|\{f_t^{(l+1)}\}_{\mathcal{N}}, \{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}, \Theta) \\ & \quad \times p(\Theta|\{f_t^{(l+1)}\}, \{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}) d\Theta. \end{aligned} \quad (12)$$

3.3. Approximate BP using the MAP-EM algorithm

We can use the EM algorithm [5] in order to approximately execute the BP mentioned above. Especially, the MAP-EM algorithm can be applied so that the parameters are also treated as random variables.

For each resolution l , the MAP-EM algorithm is executed. At the E-step, the following function with respect to Θ , which is generally called the Q function, is constructed.

$$\begin{aligned} Q(\Theta; \hat{\Theta}) &= \text{E} \left[\ln p(\{f_t^{(l)}\}, \{d^{(l)}\}|\{f_t^{(l-1)}\}, \dots, \{f_t^{(1)}\}, \Theta) \right. \\ & \quad \left. + \ln p(\Theta|\{f_t^{(l-1)}\}, \dots, \{f_t^{(1)}\})|\{f_t^{(l)}\}, \hat{\Theta} \right], \end{aligned} \quad (13)$$

where $\text{E}[\cdot]$ denotes the expectation operator, and $\hat{\cdot}$ indicates the estimate or the variable depending on the estimate, as determined in the corresponding iteration. Because $p(\Theta|\{f_t^{(l-1)}\}, \dots, \{f_t^{(1)}\})$ has a complex form, the Laplace approximation is adopted (i.e., it is approximated by a Gaussian distribution). To determine the variance of the approximated Gaussian distribution, we can evaluate $(-\partial^2 \ln p(\Theta|\{f_t^{(l-1)}\}, \dots, \{f_t^{(1)}\})/\partial\Theta^2)^{-1}$ at the Θ^* for which $p(\Theta^*|\{f_t^{(l-1)}\}, \dots, \{f_t^{(1)}\})$ takes the maximum value.

At the M-step, the above Q function is maximized with respect to Θ , and $\hat{\Theta}$ is updated by these values for the next iteration. This maximization cannot be done analytically; therefore, the generalized MAP-EM scheme is used.

After convergence of the above two steps at each l , $p(\{d^{(l)}\}|\{f_t^{(l)}\}, \dots, \{f_t^{(1)}\}, \hat{\Theta})$ is determined. The actual required density corresponds to Eq. 12, and this integration is numerically computed. In this study, we justify the solution of the above MAP-EM algorithm using the saddle point approximation, as follows:

$$\begin{aligned} & p(d^{(l)}|\{f_t^{(l)}\}_{\mathcal{N}}, \dots, \{f_t^{(1)}\}) \\ & \approx \int p(d^{(l)}|\{f_t^{(l)}\}_{\mathcal{N}}, \dots, \{f_t^{(1)}\}, \Theta)\delta(\Theta - \hat{\Theta})d\Theta \\ & = p(d^{(l)}|\{f_t^{(l)}\}_{\mathcal{N}}, \dots, \{f_t^{(1)}\}, \hat{\Theta}). \end{aligned} \quad (14)$$

4. Numerical experiments

To confirm the effectiveness of the proposed method, we conducted numerical experiments using artificial

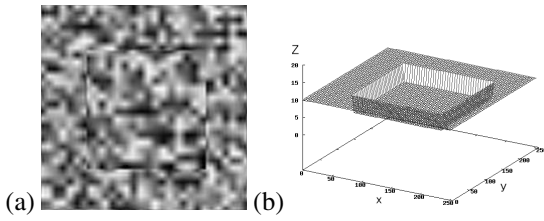


Figure 2. Data used in the experiments: (a) artificial image; (b) true depth map.

images. Figure 2(a) shows the first image, with 256×256 pixels generated by a computer graphics technique using the depth map shown in Fig. 2(b) and a random texture. The second image was generated at a different viewpoint, which was assumed to move with $\mathbf{u} = [0.1, 0.0, 0.0]^T$ and $\mathbf{r} = [0.0, 0.0, 0.0]^T$. In this situation, the theoretically calculated norm of the optical flow between the two images was approximately 2 pixels on average for the whole image. These images were decomposed into four layers with different resolutions by using band-pass filters.

The estimated depth maps are shown in Fig. 3. It is obvious from Eq. 6 that $\|\mathbf{u}\|$ and $|d|$ cannot be uniquely determined, and the scale of the depth in Fig. 3 is scaled to the true value of $\|\mathbf{u}\|$. The mesh size in Fig. 3 indicates N_l , and for example, $N_1 = 32 \times 32$ pixels. In the estimations, the variances of d and Θ prior to $l = 1$ were sufficiently large. The result obtained using the all the observed information corresponding to the four layers at the same time without the BP is shown in Fig. 4. For this result, the local region size was 8×8 pixels. From these results, we can confirm that stable recovery of the depth is achieved by the proposed method.

The above results were derived for noise-free images. Therefore, $n_1^{(l)}$ in Eq. 8 corresponds to the 1st approximation error of the gradient equation. We confirmed through experiments that, for the Gaussian image noise with a standard deviation of 5% with respect to the dynamic range of the image intensity, the proposed method has almost the same performance as that shown in Fig. 3. Additionally, we omitted the BP of Θ , and found that the root mean square error (RMSE) of the depth is one and a half times as large as that using the BP for Θ . This result is due to the estimation bias.

5. Conclusions

In this paper, we propose a method for stably recovering object shape as a depth map. The method is based on the multi-scale Bayesian network and approximate BP using the MAP-EM algorithm. The effective-

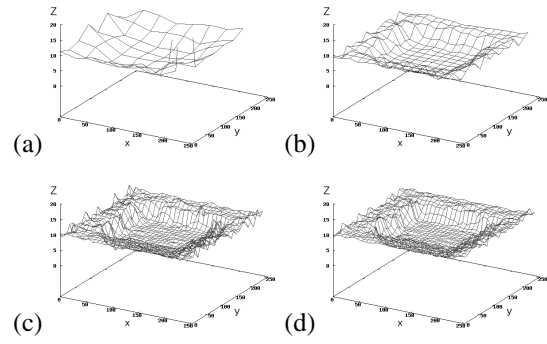


Figure 3. Estimated depth map with BP; (a) $l = 1$; (b) $l = 2$; (c) $l = 3$; (d) $l = 4$.

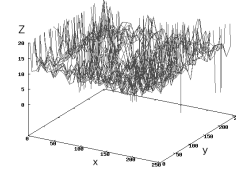


Figure 4. Estimated depth map without BP.

ness and applicability of the proposed method were confirmed through numerical experiments. In the future, the performance for real images needs to be examined, and a quantitative evaluation of the accuracy is required.

Acknowledgements: The authors are most grateful for the reviewers' constructive comments.

References

- [1] G. Farneback, "Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field," ICCV, pp. 171–177, 2001.
- [2] T. Brox, A. Bruhn, N. Papenber, J. Weickert, "High accuracy optical flow estimation based on a theory for warping," ECCV, vol. 4, pp. 25–36, 2004.
- [3] A. Bruhn and J. Weickert, "Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods," Int. Journal Comput. Vision, vol. 61, no. 3, pp. 211–231, 2005.
- [4] E.P. Simoncelli, "Bayesian multi-scale differential optical flow," Handbook of Computer Vision and Applications, Academic Press, pp. 397–422, 1999.
- [5] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum likelihood from incomplete data," J. Roy. Statist. Soc. B, vol. 39, pp. 1–38, 1977.