MVA2011 IAPR Conference on Machine Vision Applications, June 13-15, 2011, Nara, JAPAN

Reconstructing 3D Land Surface From a Sequence of Aerial Images

Shinya Mizoe, Yuichi Yaguchi, Kazuaki Takahashi, Kazuhiro Ota and Ryuichi Oka
University of Aizu
Aizuwakamatsu, Fukushima, Japan
{m5131121, d8101109, s1140131, m5131125, oka}@u-aizu.ac.jp

Abstract

This paper proposes a method for reconstructing a 3D surface landscape from an aerial image sequence captured by a single noncalibrated camera. Reconstructing a 3D surface landscape is more difficult than constructing a landscape of buildings or objects in a room because of the lack of available information about camera parameters, the need for mosaicking of 3D surface elements, and the introduction of nonrigid objects. Therefore, conventional methods are not directly applicable. In order to solve these problems, we apply so-called 2-Dimensional Continuous Dynamic Programming (2DCDP) to obtain full pixel trajectories between successive image frames in a sequence of aerial images. Then we apply Tomasi–Kanade Factorization to the full pixel trajectories to reconstruct the 3D surface. We also develop a mosaicking technique for connecting all of the partially reconstructed surfaces. The experimental results show that our proposed method is very promising for reconstructing 3D surfaces, including a forest, a mountain, a lake and several houses. We conduct experiments to compare our method against a SIFT-based method using two sets of data, namely, artificial and real image sequence data.

1 Introduction

In computer vision, many image-based 3D modeling methods have been developed, such as the stereo method [3], shape from shading [5], photometric stereo [13], the baseline matching method using epipolar geometry [9], Tomasi–Kanade Factorization methods [12], and shape from silhouettes [2]. Most of these methods achieve their goals under specific conditions and require extra input information, such as internal and external camera parameters or light source position (see Table 1). All of them, except the Factorization method, need precise camera parameters, which are contained in a fundamental matrix. Normally, the fundamental matrix can be extracted from several calibrated images [15]. An effective approach, quasi-dense baseline matching, developed by J. Kannala and S.S. Brandt [7], uses the fundamental matrix and seeds provided by the SIFT or KLT tracker to obtain more matching points for 3D object reconstruction. However, materials such as movies or photos taken by ordinary people are difficult to calibrate. Another approach, the Factorization method, does not require the fundamental matrix or calibrated images to reconstruct 3D objects, so it is still a practical method for this situation.

SIFT [8] and KLT [11] are used as pixel matching techniques in the Factorization method. They require a small angular variation in sequences of input images, and can obtain only a small number of matching pixels. For this reason, Factorization needs to use numerous input images to increase the quantity of matching pixels.

Meanwhile, landscape reconstruction from aerial image sequences is one of the most important applications of 3D reconstruction in computer vision. The researches of 3d reconstruction are using epipolar geometry [1], LIDAR data [10], and laser scan data from ground [4].

However, most of this research is for urban area reconstruction. Reconstruction for forestland is more difficult than that of urban areas because of the plain textures and elusive matching points. Moreover, and as mentioned previously, internal and external camera parameters are needed for accurate reconstruction.

In this paper, we propose a 3D reconstruction system, which is applicable for nonartificial landscapes, by using the 2DCDP [14] algorithm. 2DCDP was introduced by Yaguchi, Iseki, and Oka in 2008 [14]. This method maintains 2D pixel correlation and assures continuity and monotonicity in the input image, therefore giving numerous matching points. Additionally, 2DCDP uses the whole reference image for image registration. Thus, we can obtain proper matching points even if there are areas in which the texture is plain. This fact is an advantage over Factorization, which requires many reliable matching points to calculate an object's shape.

Section 2 presents an overview of this system, and details of the 2DCDP algorithm, Factorization method, and merging for 3D points. Section 3 shows the experimental results for artificial data and real forestland image sequences. Finally, our conclusion is included in Section 4.

2 System Overview

In this section, we propose a method that obtains a 3D landscape surface from an aerial image sequence. Matching points are obtained by using 2DCDP, and then a 3D coordinate of matching points is calculated by using the Factorization method. This method requires three or more images taken from distinct positions. To satisfy this condition, first, we select a frame, F_i, from the image sequence, and set a region of interest (ROI). From this, we call F_i a reference frame of an object. Next, we select two or more frames that completely contain the ROI of F_i. These frames can be considered as the images that are taken from distinct positions of the ROI. We can obtain corresponding points of the ROI within selected images by using 2DCDP [14], with the ROI as reference and F_i and the other frames as input. After that, a 3D surface of the ROI is reconstructed by using the Factorization method [12, 6]. Finally, 3D surfaces, which are ob-

Table 1. 3D reconstruction methods summary: 'x' indicates (a) Camera parameter, (b) Fixed camera position, (c) Fixed light position, (d) Camera distance, (e) Nonpeculiarity scale matrix, (f) Corresponding points, (g) Minimum number of required images

Method	Characteristic	(a)	(b)	(c)	(d)	(e)	(f)	(g)	Note
Stereo method	Parallax + Triangular surveying	x			x		x	2	Principle of human eye
Shape from shading	Reflection coefficient map	x	x	x			x	1	Smooth object
Photometric stereo	Reflection coefficient difference	x	x				x	3	Lambertian surface model
Baseline matching	E/F Matrix + Camera motion	x					x	2	Weak matching noise
Factorization method	Pixel correspondence + Motion separation					x	x	3	Affine camera model
Shape from silhouettes	Back projection + Voting	x	x		x			4+	Convex object only

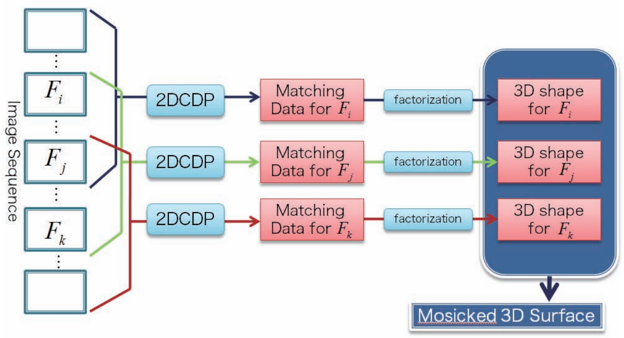

Figure 1. The processing flow of the system

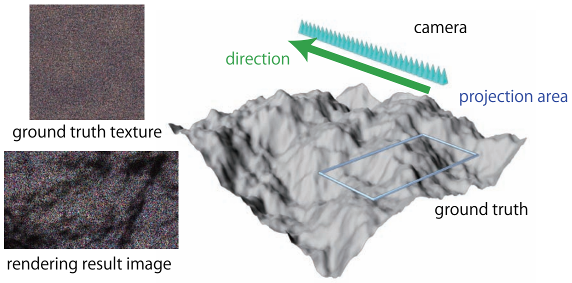

Figure 2. The landscape, camera position, and texture of artificial data.

tained by applying the previous steps on each F_i, are mosaicked using the RANSAC method (Section 2.1). Figure 1 illustrates the processing flow of the system.

2.1 Mosaicking for 3D Surfaces

In this section, we show a method of stitching together two reconstructed 3D surfaces. First, we obtain corresponding points between reconstructed 3D surfaces. Let F_A, F_B be reference images of objects A and B, respectively. Corresponding points between A and B are obtained by 2DCDP after putting F_A into input frames of B, and F_B into input frames of A. After that, we calculate the Affine matrix that minimizes the sum of the error between corresponding points with the RANSAC method. Let $C_A = \{{}_A\mathbf{x}_1, \cdots, {}_A\mathbf{x}_N\}, C_B = \{{}_B\mathbf{x}_1, \cdots, {}_B\mathbf{x}_N\}$, where N is the number of corresponding points between A and B, and $\mathbf{x}_i = (x_i \; y_i \; z_i)^T$. We choose three integers k, l, m randomly ($1 \leq k, l, m \leq N$, $k \neq l \neq m$), and calculate the following equations:

$$({}_A\mathbf{x}_k \; {}_A\mathbf{x}_l \; {}_A\mathbf{x}_m) = M({}_B\mathbf{x}_k \; {}_B\mathbf{x}_l \; {}_B\mathbf{x}_m), \tag{1}$$

where $M = (m_{ij})$ is a 3 × 3 matrix.
And, we define the affine matrix M' as follows:

$$\begin{aligned} M' &= \begin{pmatrix} m_{11} & m_{12} & 0 & m_{13} \\ m_{21} & m_{22} & 0 & m_{23} \\ 0 & 0 & \pm 1 & Z_A \mp Z_B \\ m_{31} & m_{32} & 0 & m_{33} \end{pmatrix}, & (2) \\ Z_A &= ({}_A z_k + {}_A z_l + {}_A z_m)/3, & (3) \\ Z_B &= ({}_B z_k + {}_B z_l + {}_B z_m)/3, & (4) \end{aligned}$$

where M' represents the rotation and skew of the x and y coordinates, and translation of x, y and z. Practically, it is always true that $m_{31} = m_{32} = 0$, and $m_{33} = 1$.
Then we transform C_B by M':

$$\begin{pmatrix} {}_B\mathbf{x}'_1 & \cdots & {}_B\mathbf{x}'_N \\ 1 & \cdots & 1 \end{pmatrix} = M' \begin{pmatrix} {}_B\mathbf{x}_1 & \cdots & {}_B\mathbf{x}_N \\ 1 & \cdots & 1 \end{pmatrix}. \tag{5}$$

Finally, we calculate the error between C_A and the transformed C_B:

$$D = \sum_{k=1}^{N} d({}_A\mathbf{x}_k, {}_B\mathbf{x}'_k), \tag{6}$$

where $d(x, y) = \sqrt{x^2 + y^2}$. The process of choosing k, l, m in Equation (1) is repeated a predefined number of times, and the M' that minimizes D is selected as the Affine matrix to join B into A.
In this process, we allow the rotation in the x and y directions only because the error grows rapidly with each join if we permit the rotation in the z direction. Moreover, M' is based on the assumpotion that camera direction is perpendicular to 3D surface. Note that $\pm 1, Z_A \mp Z_B$ in M' are enantiomorphs of the Factorization method.

3 Experiments

3.1 Conditions

This section describes the experimental data preparation and results for two sets of aerial image sequences that are generated by CG, and taken by general video

Table 2. Experimental conditions of artificial and real image sequence data.

Original video size		1280 × 720
2DCDP	Video size	320 × 240
	ROI size	240 × 120
SIFT	Video size	1280 × 720
	ROI size	960 × 480
The number of input images		7
The position of reference frame		4
Iteration times		1000

Table 3. The result of calculations of evaluation functions.

	2DCDP	SIFT	SIFT
ROI size	240 × 120	240 × 120	960 × 480
The number of points	28800	44	17122
D	0.002275	0.263603	0.025563
V	0.9861	-	1.90575

camera. A comparison between the proposed system and a system that adopts SIFT [8] as its matching method is also performed in this section. Table 2 describes experimental conditions of artificial and real image sequence data. Note that, SIFT has an advantage over 2DCDP at video and ROI size. Figure 2 shows the landscape, camera position, and texture of CG data. In the real landscape, the distance between two images is 10 frames.

3.2 Results

Figure 3 (a) shows reconstructed objects by using SIFT and 2DCDP, used images, and 2DCDP matching results. The result of 2DCDP + Factorization has more points, and the surfaces of regional points are smoother than SIFT. It would appear that SIFT is a feature-point-based method, so each corresponding point is independent. On the other hand, 2DCDP is a pixel-based method, therefore each point has the relationship. Figure 3 (b),(c) describe the combination of the objects of the artificial and real landscapes. Its error rate is similar to the single object error rate, despite restricting the Affine matrix to Equation (2).

3.3 Evaluations

In this section, we define evaluation functions and compare results of 2DCDP + Factorization with those of SIFT + Factorization in artificial data, as shown in Figure 3 (a). Two evaluation functions are described as follows:

$$D = \frac{\sum_{i=1}^{N}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{N}\hat{z}_i^2}, V = \frac{1}{k}\sum_{k}\frac{\mathrm{Var}(Z_k)}{\mathrm{Var}(\hat{Z}_k)}, \qquad (7)$$

where z_i, $\hat{z}_i$ represent the z coordinate of the i-th corresponding point of the reconstructed and ground truth landscapes, and Z_k is a subset of z_i. In this experiment, we divided a land surface into 100 local areas (equal segregation in both the x and y directions) as Z_k. The undulation of each local area is indicated by $\mathrm{Var}(Z_k)$. The 3D surface is close to ground truth when V is nearly one. The 3D surface made by SIFT + Factorization is like a pinholder, therefore V becomes far removed from one (see Table 3). We cannot calculate the result of 2DCDP+factorization under ROI size 960 ×480 because the computational complexity of 2DCDP is $O(n^4)$. If we have sufficient memory size for calculation, we will obtain $960 \times 480 = 460800$ points.

4 Conclusion

This paper proposed a method for reconstructing the 3D surfaces of landscapes from aerial imagery using the 2DCDP and Factorization methods, which needs only aerial imagery shot by a single noncalibrated camera, and requires no intrinsic or extrinsic camera parameters. The affine matrix for joining objects is restricted. However, we show that the increase in error is low if input video data have sufficient texture information, and we can obtain corresponding points that have less matching error. In our method, we hypothesized an orthographic projection that is a zero-order approximation of a perspective projection in the Factorization method. We need to compare our results with a weak perspective projection and a paraperspective projection.

References

[1] C. Baillard, C. Schmid, and A. Fitzgibbon. Automatic line matching and 3D reconstruction of buildings from multiple views. *IAPRS*, 32(3):69–80, 1999.

[2] H. Baker. Three-dimensional modelling. In *Proceedings of IJCAI 1977*, pages 649–655. Citeseer, 1977.

[3] S. Barnard and M. Fischler. Computational stereo. *ACM Computing Surveys (CSUR)*, 14(4), 1982.

[4] C. Fruh and A. Zakhor. Constructing 3D city models by merging aerial and ground views. *IEEE CGA*, 23(6):52–61, 2003.

[5] B. Horn and M. Brooks. *Shape from shading.* MIT press Cambridge Massachusetts, 1989.

[6] K. Kanatani and Y. Sugaya. Complete recipe for factorization. *IEICE technical report. Neurocomputing*, 103(391):19–24, 2003.

[7] J. Kannala and S. Brandt. Quasi-dense wide baseline matching using match propagation. In *2007 IEEE Conference on CVPR*, pages 1–8. IEEE, 2007.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[9] R. Mohr, L. Quan, and F. Veillon. Relative 3D reconstruction using multiple uncalibrated images. *IJRR*, 14(6):619, 1995.

[10] T. Schenk and B. Csathó. Fusion of Lidar Data and Aerial Imagery for a More Complete Surface Description. *ISPRS*, 34(3/A):310–317, 2002.

[11] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, Citeseer, 1991.

[12] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992.

[13] R. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.

[14] Y. Yaguchi, K. Iseki, and R. Oka. Optimal Pixel Matching between Images. *Advances in Image and Video Technology*, pages 597–610, 2009.

[15] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on PAMI*, 22(11):1330–1334, 2000.

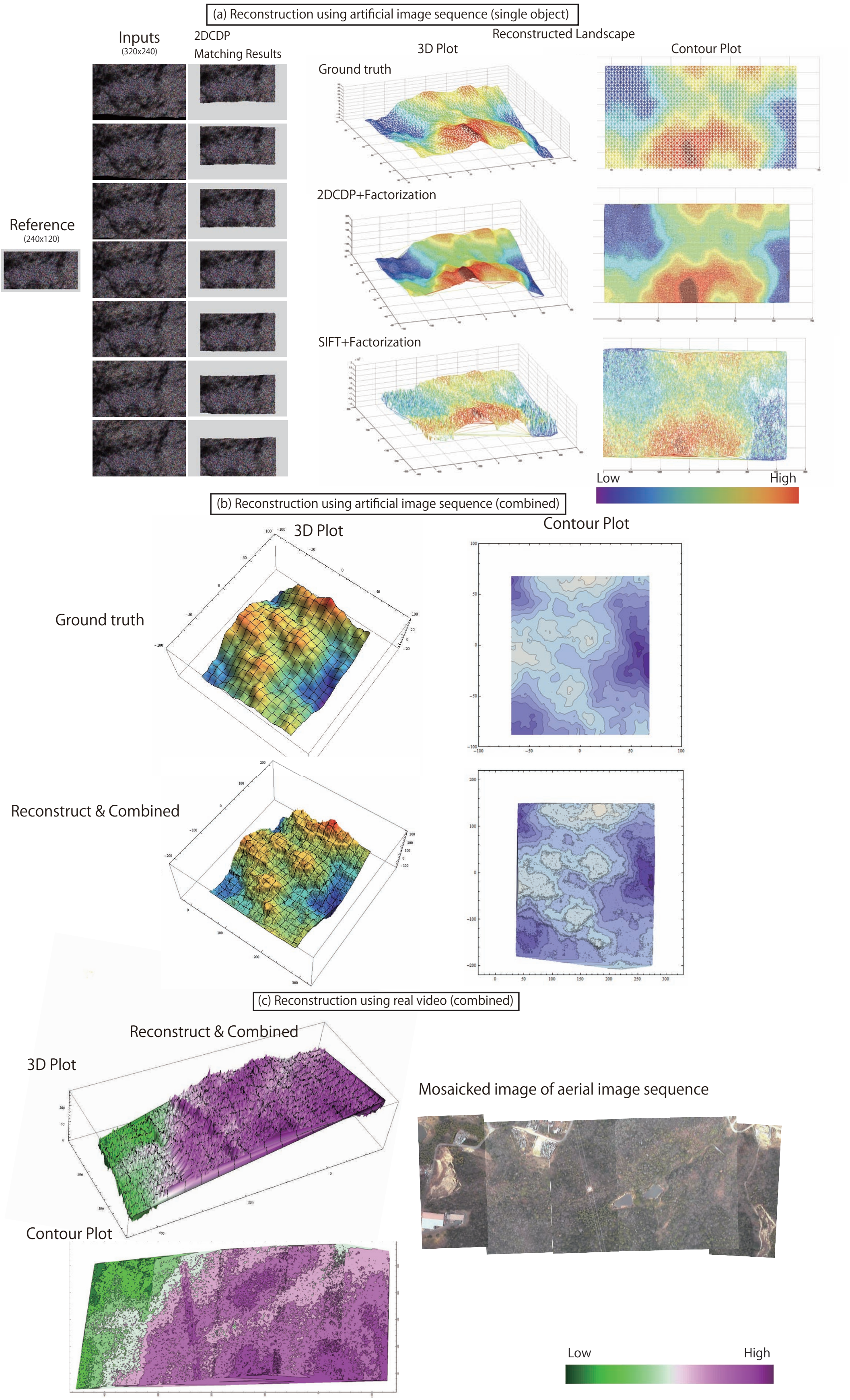

Figure 3. The results of reconstruction of 3D land surface. Figure (a) is the result of a single object using the artificial landscape of 2DCDP + Factorization and SIFT + Factorization. The left images of (a) are inputs to 2DCDP and its matching results. (b) and (c) shows the results of combining objects using artificial and real landscape image sequences. The right image of (c) is combined real images, and made by manual translation and rotation. Note that this manual handling has incompleteness because the Affine matrix (2) includes a skew transform for the x and y coordinates