Initializing 3-D Reconstruction from Three Views Using Three Fundamental Matrices

Yasushi Kanazawa[1], Yasuyuki Sugaya[1], and Kenichi Kanatani[2]

[1] Department of Computer Science and Engineering,
Toyohashi University of Technology, Aichi 441-8105 Japan
kanazawa@cs.tut.ac.jp, sugaya@iim.cs.tut.ac.jp

[2] Okayama University, Okayama 700-8530 Japan
kanatani2013@yahoo.co.jp

Abstract. This paper focuses on initializing 3-D reconstruction from scratch without any prior scene information. Traditionally, this has been done from two-view matching, which is prone to the degeneracy called "imaginary focal lengths". We overcome this difficulty by using three images, but we do not require three-view matching; all we need is three fundamental matrices separately computed from image pairs. We exploit the redundancy of the three fundamental matrices to optimize the camera parameters and the 3-D structure. We do numerical simulation to show that imaginary focal lengths are less likely to occur, resulting in higher accuracy than two-view reconstruction. We also test the degeneracy tolerance capability of our method by using endoscopic intestine tract images, for which the camera configuration is almost always nearly degenerate. We demonstrate that our method allows us to obtain more detailed intestine structures than two-view reconstruction and hence leads to new medical applications to endoscopic image analysis.

Keywords: Initialization of 3-D reconstruction, imaginary focal length degeneracy, three views, three fundamental matrices.

1 Introduction

Today, 3-D reconstruction from images is a common technique of computer vision thanks to various reconstruction tools available on the Web. The basic principle is what is known as *bundle adjustment*, computing from point correspondences over multiple images all 3-D point positions and all camera parameters by searching the high-dimensional parameter space. The search is done so as to minimize the discrepancy, or the *reprojection error*, between the observed images and the projections of the estimated 3-D points computed by the estimated camera parameters. The best known bundle adjustment software is *SBA* of Lourakis and Argyros [13]. Snavely et al. [15, 16] combined it with feature point detection and matching as a package called *bundler*. Bundle adjustment is an iterative process, requiring an initial solution, which is usually computed by choosing from among the input images pairs of well matched views. This is because the 3-D shape and the camera parameters are easily computed from two views, and various practically high-accuracy techniques have been presented [11].

However, it is well known that two-view reconstruction fails if the two cameras are in a "fixating" configuration, i.e., their optical axes intersect in the scene [3, 8]. This configuration is very natural when one takes images of the same object from two different positions. Another problem is that irrespective of the camera configuration, the information obtained from two views is minimal, resulting in the same number of equations as the number of unknowns. This may be an advantage in that the solution can be obtained analytically, but often the solution that satisfies all equations does not exist for noisy data. Typically, the square of some expressions containing the focal lengths become negative; this problem is known as the "imaginary focal length degeneracy".

The purpose of this paper is not so much to achieve yet higher reconstruction accuracy. Rather, we focus on preventing degeneracy. Namely, we want to initialize 3-D reconstruction stably from scratch, i.e., without requiring any prior information about the scene structure or the camera positions. There have already been some such attempts. Observing that fixating configurations occur when the principal point of one image matches to that of the other image, Hartley and Silpa-Anan [4] used the regularization approach to minimally moved the assumed principal points so that the imaginary focal lengths do not arise, but the solution depends on the regularization parameter. Kanatani et al. [9] proposed random resampling of matching points to avoid imaginary focal lengths, but a sufficient number of correspondences are necessary. Goldberger [2] adopted the projective reconstruction framework, computing the camera matrices up to projectivity from fundamental matrices and epipoles computed from image pairs. For Euclidean reconstruction, however, more information is required [14].

In this paper, we impose a strict constraint on the cameras so that the Euclidean structure results from minimum information, yet extra degrees of freedom remain to be adjusted to suppress imaginary focal lengths. This is made possible by using three images, but we do not require three-view matching; all we need is three fundamental matrices separately computed from image pairs. We do numerical simulation and observe that imaginary focal lengths are less likely to occur, resulting in higher accuracy than two-view reconstruction. Then, we show a novel medical application: we reconstruct the 3-D structure from endoscopic intestine tract images. This provides a good testbed for the degeneracy tolerance capability of our method, because the camera configuration is very pathological: the camera moves almost in one direction in intestine tracts and hence always in a near fixation configuration, which is very likely to cause imaginary focal lengths.

2 The Task

For two-view reconstruction, the cameras must be such that 1) the principal point is known, 2) the aspect ratio is 1, and 3) no image skew exists [5, 11]. This constraint stems from the fact that the available information from two views is limited. We could relax this for three views [2, 4, 14], but since our intention is to exploit the redundancy of three-view information to do optimization, we adopt the same constraint. This is no big restriction in practice, because today's cameras mostly satisfy the requirements or can easily be so calibrated beforehand. We define an xy image coordinate system such that the origin o is at the principle point (at the frame center by default) with the x-axis upward and the y-axis rightward. This is necessary for the x- and y-axes together with the optical axis regarded as the z-axis to constitute a right-handed system for 3-D rotation computation (for this purpose, we could instead take the x-axis rightward and the y-axis downward).

We capture three images of the same scene by three cameras (or equivalently by moving one camera). We call these images the 0th, 1st, and 2nd views, and the corresponding cameras the 0th, 1st, and 2nd cameras, respectively. Suppose a point (x, y) in the 0th view corresponds to (x', y') in the 1st view. We write the epipolar equation [5] between them in the form

$$(\mathbf{x}, \mathbf{F}_{01}\mathbf{x}') = 0, \qquad \mathbf{x} = \begin{pmatrix} x/f_0 \\ y/f_0 \\ 1 \end{pmatrix}, \qquad \mathbf{x}' = \begin{pmatrix} x'/f_0 \\ y'/f_0 \\ 1 \end{pmatrix}, \tag{1}$$

where $\mathbf{F}_{01}$ is the fundamental matrix between the 0th and 1st views. We write $(\mathbf{a}, \mathbf{b})$ for the inner product of vectors $\mathbf{a}$ and $\mathbf{b}$. The scaling constant f_0 is for stabilizing numerical computation; we take it to be an approximate focal length of the cameras and call it the *default focal length* (we set it to 600 pixels in our experiment). The fundamental matrix $\mathbf{F}_{02}$ between the 0th and 2nd views and the fundamental matrix $\mathbf{F}_{12}$ between the 1st and 2nd views are similarly defined. Fundamental matrices are uniquely computed from eight or more point correspondence pairs (theoretically seven points are sufficient, but the solution may not be unique). In our experiment, we use the EFNS (Extended Fundamental Numerical Scheme) of Kanatani and Sugaya [10], which can compute an exact reprojection error minimization solution.

We regard the XYZ coordinate system of the 0th camera, the origin O being at the lens center with the Z axis along the optical axis, as the world coordinate system. Let $\mathbf{t}_1$ and $\mathbf{t}_2$ be the lens centers of the 1st and the 2nd cameras, respectively, and $\mathbf{R}_1$ and $\mathbf{R}_2$ their rotations relative to the 0th camera. Let f, f', and f'' be the focal lengths of the 0th, 1st, and the 2nd cameras, respectively. The fundamental matrices $\mathbf{F}_{01}$, $\mathbf{F}_{02}$, and $\mathbf{F}_{12}$ ideally (i.e., if they are exact) satisfy the identities

$$\begin{aligned}
\mathbf{F}_{01} &\simeq \mathrm{diag}(1, 1, \frac{f}{f_0})\Big(\mathbf{t}_1 \times \mathbf{R}_1\Big)\mathrm{diag}(1, 1, \frac{f'}{f_0}), \\
\mathbf{F}_{02} &\simeq \mathrm{diag}(1, 1, \frac{f}{f_0})\Big(\mathbf{t}_2 \times \mathbf{R}_2\Big)\mathrm{diag}(1, 1, \frac{f''}{f_0}), \\
\mathbf{F}_{12} &\simeq \mathrm{diag}(1, 1, \frac{f'}{f_0})\Big((\mathbf{R}_1^\top(\mathbf{t}_2 - \mathbf{t}_1)) \times (\mathbf{R}_1^\top\mathbf{R}_2)\Big)\mathrm{diag}(1, 1, \frac{f''}{f_0}),
\end{aligned} \tag{2}$$

where the symbol $\simeq$ denotes equality up to a nonzero constant and $\mathrm{diag}(a, b, c)$ denotes the diagonal matrix with a, b, and c as the diagonal elements in that order. For a vector $\mathbf{v}$ and a matrix $\mathbf{A}$, we define $\mathbf{v} \times \mathbf{A}$ to be the matrix whose columns are the vector products of $\mathbf{v}$ and the corresponding columns of $\mathbf{A}$. The task of this paper is to compute f, f', f'', $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{R}_1$, and $\mathbf{R}_2$ from given fundamental matrices $\mathbf{F}_{01}$, $\mathbf{F}_{02}$, and $\mathbf{F}_{12}$, considering the fact that the computed $\mathbf{F}_{01}$, $\mathbf{F}_{02}$, and $\mathbf{F}_{12}$ may not be exact.

3 Focal Length Computation

Instead of computing f, f', and f'', we compute the following x, y, and z:

$$x = \left(\frac{f_0}{f}\right)^2 - 1, \qquad y = \left(\frac{f_0}{f'}\right)^2 - 1, \qquad z = \left(\frac{f_0}{f''}\right)^2 - 1. \tag{3}$$

It is known [9] that x and y ideally minimize, in the neighborhood of the solution, the quadratic polynomial in x and y

$$\begin{aligned}
&K_{01}(x,y) = \\
&(\mathbf{k}, \mathbf{F}_{01}\mathbf{k})^4 x^2 y^2 + 2(\mathbf{k}, \mathbf{F}_{01}\mathbf{k})^2 \|\mathbf{F}_{01}^\top \mathbf{k}\|^2 x^2 y + 2(\mathbf{k}, \mathbf{F}_{01}\mathbf{k})^2 \|\mathbf{F}_{01}\mathbf{k}\|^2 x y^2 \\
&+\|\mathbf{F}_{01}^\top \mathbf{k}\|^4 x^2 + \|\mathbf{F}_{01}\mathbf{k}\|^4 y^2 + 4(\mathbf{k}, \mathbf{F}_{01}\mathbf{k})(\mathbf{k}, \mathbf{F}_{01}\mathbf{F}_{01}^\top \mathbf{F}_{01}\mathbf{k}) xy \\
&+2\|\mathbf{F}_{01}\mathbf{F}_{01}^\top \mathbf{k}\|^2 x + 2\|\mathbf{F}_{01}^\top \mathbf{F}_{01}\mathbf{k}\|^2 y + \|\mathbf{F}_{01}\mathbf{F}_{01}^\top\|^2 \\
&-\frac{1}{2}\Big((\mathbf{k}, \mathbf{F}_{01}\mathbf{k})^2 xy + \|\mathbf{F}_{01}^\top \mathbf{k}\|^2 x + \|\mathbf{F}_{01}\mathbf{k}\|^2 y + \|\mathbf{F}_{01}\|^2\Big)^2,
\end{aligned} \tag{4}$$

where $\mathbf{k} = (0,0,1)^\top$, and that the minimum is 0. If quadric polynomials $K_{02}(x,z)$ and $K_{12}(y,z)$ are similarly defined, x and z minimizes $K_{02}(x,z)$, and y and z minimize $K_{12}(y,z)$; their minimums are 0. Hence, we can determine x and y from $K_{01}(x,y)$, y and z from $K_{12}(y,z)$, and z and x from $K_{02}(x,z)$. Moreover, the solution is analytically computed by the *Bougnoux formula* [5, 9]. In the presence of noise, however, the analytically obtained solutions are in general inconsistent to each other. Here, we adopt the solution x, y, and z that minimize

$$F(x,y,z) = K_{01}(x,y) + K_{02}(x,z) + K_{12}(y,z). \tag{5}$$

In our experiment, we used Newton iterations starting from $x = y = z = 0$, which is equivalent to $f = f' = f'' = f_0$. Then, f, f', and f'' are given from Eq. (3) in the form

$$f = \frac{f_0}{\sqrt{1+x}}, \qquad f' = \frac{f_0}{\sqrt{1+y}}, \qquad f'' = \frac{f_0}{\sqrt{1+z}}. \tag{6}$$

Note that if any of x, y, and z are equal to or less than -1, the computation fails. This is the so called "imaginary focal length problem", which frequently occurs in two-view reconstruction. One of the causes of this phenomenon is that the analytical solution relies on the fact that the solution not only minimizes $K_{01}(x,y)$, $K_{02}(x,z)$, and $K_{12}(x,z)$ but also their minimums are exactly 0, which does not hold for real data. Here, we are not assuming that their minimums are

0, so we expect that the imaginary focal length problem will be alleviated, if not completely avoided. In fact, we never encountered imaginary focal lengths in our three-view reconstruction experiments.

It is known [9] that if two cameras, say the 0th and the 1st, are in a fixating configuration, the minimum of $K_{01}(x, y)$ in Eq. (4) degenerates to a curve in the xy plane so it does not have a unique minimum. If we assume that $f = f'$, the solution is uniquely determined as the intersection of that curve with the line $x = y$. However, if the two cameras are in an "isosceles" configuration (fixating with equal distance), the minimum curve of $K_{01}(x, y)$ is "tangent" to the line $x = y$ and hence no clear intersection is defined. The same holds for the other pairs of cameras. However, our three-view formulation can uniquely determine the solution even when fixating camera configurations are included, unless the three cameras are in a simultaneous fixating configuration, in which case the Hessian of $F(x, y, z)$ in Eq. (5) becomes singular at the minimum, making numerical minimization unstable (we omit the details).

4 Translation Computation

The relative camera translation can be computed from the fundamental matrix between two views [11]. Hence, the three fundamental matrices $\mathbf{F}_{01}$, $\mathbf{F}_{02}$, and $\mathbf{F}_{12}$ can determine the translations between all the camera pairs. However, their signs and scales are indeterminate. Although we cannot fix the absolute scale as long as images are used, we can fix their relative scales from the "triangle condition", requiring that the three translations form a closed triangle. However, as we show shortly, the triangle condition involves camera rotations, so, unlike two-view reconstruction, translations cannot be determined separately. Here, we introduce a procedure for computing the translations and rotations at the same time.

Using the computed focal lengths f, f', and f'', we define the *essential matrices* $\mathbf{E}_{01}$, $\mathbf{E}_{02}$, and $\mathbf{E}_{12}$ by

$$\mathbf{E}_{01} \equiv \mathrm{diag}(1, 1, \frac{f_0}{f})\mathbf{F}_{01}\mathrm{diag}(1, 1, \frac{f_0}{f'}), \qquad \mathbf{E}_{02} \equiv \mathrm{diag}(1, 1, \frac{f_0}{f})\mathbf{F}_{02}\mathrm{diag}(1, 1, \frac{f_0}{f''})$$
$$\mathbf{E}_{12} \equiv \mathrm{diag}(1, 1, \frac{f_0}{f'})\mathbf{F}_{12}\mathrm{diag}(1, 1, \frac{f_0}{f''}), \tag{7}$$

From Eqs. (2), they ideally satisfy

$$\mathbf{E}_{01} \simeq \mathbf{t}_1 \times \mathbf{R}_1, \qquad \mathbf{E}_{02} \simeq \mathbf{t}_2 \times \mathbf{R}_2, \qquad \mathbf{E}_{12} \simeq \mathbf{t}_{12} \times \mathbf{R}_1^\top \mathbf{R}_2, \tag{8}$$

where

$$\mathbf{t}_{12} = \mathbf{R}_1^\top(\mathbf{t}_2 - \mathbf{t}_1), \tag{9}$$

is the lens center of the 2nd camera viewed from the 1st camera. The triangle condition means enforcing this equation. However, it involves $\mathbf{R}_1$, which is unknown yet. We resolve this as follow. Since Eqs. (8) imply that $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_{12}$ are, respectively, null vectors of $\mathbf{E}_{01}^\top$, $\mathbf{E}_{02}^\top$, and $\mathbf{E}_{12}^\top$ in the absence of noise, we compute those translations $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_{12}$ that minimize $\|\mathbf{E}_{01}^\top\mathbf{t}_1\|^2$, $\|\mathbf{E}_{02}^\top\mathbf{t}_2\|^2$, and $\|\mathbf{E}_{12}^\top\mathbf{t}_{12}\|^2$, respectively. The solution is given by the eigenvectors of $\mathbf{E}_{01}\mathbf{E}_{01}^\top$, $\mathbf{E}_{02}\mathbf{E}_{02}^\top$, and $\mathbf{E}_{12}\mathbf{E}_{12}^\top$ for their smallest eigenvalues. At this sage, the scales and the signs of $\mathbf{t}_1$, $\mathbf{t}_2$, are $\mathbf{t}_{12}$ are indeterminate. As in the case of two-view reconstruction [11], we choose their signs so that

$$\sum_\alpha |\mathbf{t}_1, \mathbf{x}_\alpha, \mathbf{E}_{01}\mathbf{x}'_\alpha| > 0, \qquad \sum_\alpha |\mathbf{t}_2, \mathbf{x}_\alpha, \mathbf{E}_{02}\mathbf{x}''_\alpha| > 0, \qquad \sum_\alpha |\mathbf{t}_{12}, \mathbf{x}'_\alpha, \mathbf{E}_{12}\mathbf{x}''_\alpha| > 0, \tag{10}$$

where $|\mathbf{a}, \mathbf{b}, \mathbf{c}|$ is the scalar triplet product of $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. The vectors $\mathbf{x}_\alpha$, $\mathbf{x}'_\alpha$, and $\mathbf{x}''_\alpha$ are the coordinates of the αth point represented by vectors as in Eqs. (1) with the default focal length f_0 replaced by the computed f, f', and f''. The summations run over the image pairs from which that point is visible. Equations (10) state that almost all points are "in front" of the three camera pairs *provided* the signs of $\mathbf{E}_{01}$, $\mathbf{E}_{02}$, and $\mathbf{E}_{12}$ are correct (this issue is discussed shortly). Note that the epipolar equation of Eq. (1) holds even if the point is "behind" the cameras and that the signs of the essential matrices in Eqs. (7) are indeterminate, inheriting the sign indeterminacy of the fundamental matrixes in Eqs.(2).

Once the signs of $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_{12}$ are determined, we can determine the rotations $\mathbf{R}_1$ and $\mathbf{R}_2$ (next section). Then, substituting the computed $\mathbf{R}_1$ into the triangle condition of Eq. (9), we minimize not $\|\mathbf{E}_{01}^\top\mathbf{t}_1\|^2$, $\|\mathbf{E}_{02}^\top\mathbf{t}_2\|^2$, and $\|\mathbf{E}_{12}^\top\mathbf{t}_{12}\|^2$ separately but their sum

$$\|\mathbf{E}_{01}^\top\mathbf{t}_1\|^2 + \|\mathbf{E}_{02}^\top\mathbf{t}_2\|^2 + \|\mathbf{E}_{12}^\top\mathbf{t}_{12}\|^2 = (\begin{pmatrix}\mathbf{t}_1\\ \mathbf{t}_2\end{pmatrix}, \mathbf{G}\begin{pmatrix}\mathbf{t}_1\\ \mathbf{t}_2\end{pmatrix}), \tag{11}$$

where we define the 6×6 matrix $\mathbf{G}$ by

$$\mathbf{G} = \begin{pmatrix} \mathbf{E}_{01}\mathbf{E}_{01}^\top + \mathbf{R}_1\mathbf{E}_{12}\mathbf{E}_{12}^\top\mathbf{R}_1^\top & -\mathbf{R}_1\mathbf{E}_{12}\mathbf{E}_{12}^\top\mathbf{R}_1^\top \\ -\mathbf{R}_1\mathbf{E}_{12}\mathbf{E}_{12}^\top\mathbf{R}_1^\top & \mathbf{E}_{02}\mathbf{E}_{02}^\top + \mathbf{R}_1\mathbf{E}_{12}\mathbf{E}_{12}^\top\mathbf{R}_1^\top \end{pmatrix}. \tag{12}$$

Equation (11) is minimized by the unit eigenvector $\begin{pmatrix}\mathbf{t}_1\\ \mathbf{t}_2\end{pmatrix}$ of $\mathbf{G}$ for the smallest eigenvalue, which is normalized to $\|\mathbf{t}_1\|^2 + \|\mathbf{t}_2\|^2 = 1$. The sign is adjusted so that the recomputed $\mathbf{t}_1$ and $\mathbf{t}_2$ align to their original orientations. After $\mathbf{t}_1$ and $\mathbf{t}_2$ are thus updated, we compute $\mathbf{t}_{12}$ in Eq. (9). From these $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_{12}$, we update $\mathbf{R}_1$ and $\mathbf{R}_2$ (next section). Using the resulting $\mathbf{R}_1$, we compute the unit eigenvector of $\mathbf{G}$ in Eq. (12) to update $\mathbf{t}_1$ and $\mathbf{t}_2$. We repeat this until they converge; usually, a few iterations are sufficient.

5 Rotation Computation

Given $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_{12}$, we compute $\mathbf{R}_1$ and $\mathbf{R}_2$ that satisfy Eqs. (8) by minimizing

$$\|\mathbf{E}_{01} - \mathbf{t}_1 \times \mathbf{R}_1\|^2 + \|\mathbf{E}_{02} - \mathbf{t}_2 \times \mathbf{R}_2\|^2 + \|\mathbf{E}_{12} - \mathbf{t}_{12} \times \mathbf{R}_1^\top\mathbf{R}_2\|^2. \tag{13}$$

It can be shown [7] that this minimization is equivalent to maximizing

$$J = \mathrm{tr}[\mathbf{K}_{01}^\top\mathbf{R}_1] + \mathrm{tr}[\mathbf{K}_{02}^\top\mathbf{R}_2] + \mathrm{tr}[\mathbf{K}_{12}^\top\mathbf{R}_1^\top\mathbf{R}_2], \tag{14}$$

where tr$[\,\cdot\,]$ denotes the trace of a matrix and we define

$$\mathbf{K}_{01} = -\mathbf{t}_1 \times \mathbf{E}_{01}, \qquad \mathbf{K}_{02} = -\mathbf{t}_2 \times \mathbf{E}_{02}, \qquad \mathbf{K}_{12} = -\mathbf{t}_{12} \times \mathbf{E}_{12}. \tag{15}$$

For maximizing Eq. (14), we make use of the fact [7] that if $\mathbf{K} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^\top$ is the singular value decomposition of matrix $\mathbf{K}$, the rotation $\mathbf{R}$ that maximizes $\mathrm{tr}[\mathbf{K}^\top\mathbf{R}]$ is given by $\mathbf{R} = \mathbf{V}\mathrm{diag}(1,1,\det(\mathbf{V}\mathbf{U}^\top))\mathbf{U}^\top$. First, we compute the rotation $\mathbf{R}_1$ that maximizes $\mathrm{tr}[\mathbf{K}_{01}^\top\mathbf{R}_1]$. Equation (14) can be rewritten as

$$J = \mathrm{tr}[\mathbf{K}_{01}^\top\mathbf{R}_1] + \mathrm{tr}[(\mathbf{K}_{02} + \mathbf{R}_1\mathbf{K}_{12})^\top\mathbf{R}_2]. \tag{16}$$

Using the computed $\mathbf{R}_1$, we determine the rotation $\mathbf{R}_2$ that maximizes $\mathrm{tr}[(\mathbf{K}_{02} + \mathbf{R}_1\mathbf{K}_{12})^\top\mathbf{R}_2]$. Equation (14) can also be rewritten as

$$J = \mathrm{tr}[\mathbf{K}_{02}^\top\mathbf{R}_2] + \mathrm{tr}[(\mathbf{K}_{01} + \mathbf{R}_2\mathbf{K}_{12}^\top)^\top\mathbf{R}_1]. \tag{17}$$

Using the computed $\mathbf{R}_2$, we determine the rotation $\mathbf{R}_1$ that maximizes $\mathrm{tr}[(\mathbf{K}_{01} + \mathbf{R}_2\mathbf{K}_{12}^\top)^\top\mathbf{R}_1]$. We iterate this, each time J increasing, until J ceases to increase.

For this computation, however, we need to resolve a critical issue: the signs of $\mathbf{E}_{01}$, $\mathbf{E}_{02}$, and $\mathbf{E}_{12}$ in Eq. (7) are indeterminate. The condition of Eqs. (10) merely ensures that the signs of $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_{12}$ are compatible with the signs of $\mathbf{E}_{01}$, $\mathbf{E}_{02}$, and $\mathbf{E}_{12}$. Here, we assume that the sign of $\mathbf{E}_{01}$ is correct (this will be checked later). For selecting the signs of $\mathbf{E}_{02}$ and $\mathbf{E}_{12}$, we note that we should ideally have $\mathbf{E}_{12}^\top\mathbf{R}_1^\top(\mathbf{t}_2 - \mathbf{t}_1) = 0$ and $\mathbf{E}_{12} \simeq \mathbf{t}_{12} \times \mathbf{R}_1^\top\mathbf{R}_2$ and introduce the following two rules, which resolve the problem (we omit the details):

– If $\|\mathbf{E}_{12}^\top\mathbf{R}_1^\top(\mathbf{t}_2 - \mathbf{t}_1)\| > \|\mathbf{E}_{12}^\top\mathbf{R}_1^\top(\mathbf{t}_2 + \mathbf{t}_1)\|$, we change the signs of $\mathbf{t}_2$ and $\mathbf{E}_{02}$.
– If $\|\mathbf{E}_{12} - \mathbf{t}_{12} \times \mathbf{R}_1^\top\mathbf{R}_2\| > \|\mathbf{E}_{12} + \mathbf{t}_{12} \times \mathbf{R}_1^\top\mathbf{R}_2\|$, we change the sign of $\mathbf{K}_{12}$.

6 3-D Position Computation

Using the computed translations $\mathbf{t}_1$ and $\mathbf{t}_2$ and rotations $\mathbf{R}_1$ and $\mathbf{R}_2$, we recompute the essential matrices $\mathbf{E}_{01}$, $\mathbf{E}_{02}$, and $\mathbf{E}_{12}$ as follows:

$$\mathbf{E}_{01} = \mathbf{t}_1 \times \mathbf{R}_1, \qquad \mathbf{E}_{02} = \mathbf{t}_2 \times \mathbf{R}_2, \qquad \mathbf{E}_{12} = \Big(\mathbf{R}_1^\top(\mathbf{t}_2 - \mathbf{t}_1)\Big) \times \mathbf{R}_1^\top\mathbf{R}_2. \tag{18}$$

We optimally correct $\mathbf{x}$, $\mathbf{x}'$, and $\mathbf{x}''$ (the image coordinates represented by vectors as in Eqs. (1) with the default focal length f_0 replaced by the computed f, f', and f'') to $\hat{\mathbf{x}}$, $\hat{\mathbf{x}}'$, and $\hat{\mathbf{x}}''$, respectively in such a way that $\|\hat{\mathbf{x}} - \mathbf{x}\|^2 + \|\hat{\mathbf{x}}' - \mathbf{x}'\|^2 + \|\hat{\mathbf{x}}'' - \mathbf{x}''\|^2$ is minimized subject to $(\hat{\mathbf{x}}, \mathbf{E}_{01}\hat{\mathbf{x}}') = (\hat{\mathbf{x}}, \mathbf{E}_{02}\hat{\mathbf{x}}'') = (\hat{\mathbf{x}}', \mathbf{E}_{12}\hat{\mathbf{x}}'') = 0$. For two views, this is nothing but the optimal triangulation procedure of Kanatani et al. [10, 12], which can be straightforwardly extended to three views (we omit the details).

The projection matrices $\mathbf{P}$, $\mathbf{P}'$, and $\mathbf{P}''$ of the three cameras have the form

$$\mathbf{P} = \mathrm{diag}(1,1,\frac{f_0}{f})\begin{pmatrix}\mathbf{I} & \mathbf{0}\end{pmatrix}, \qquad \mathbf{P}' = \mathrm{diag}(1,1,\frac{f_0}{f'})\begin{pmatrix}\mathbf{R}_1^\top & -\mathbf{R}_1^\top\mathbf{t}_1\end{pmatrix},$$
$$\mathbf{P}'' = \mathrm{diag}(1,1,\frac{f_0}{f''})\begin{pmatrix}\mathbf{R}_2^\top & -\mathbf{R}_2^\top\mathbf{t}_2\end{pmatrix}. \tag{19}$$

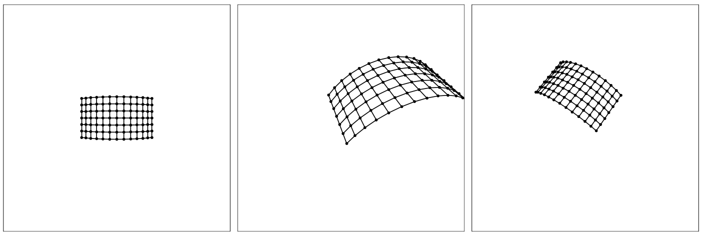


Fig. 1. The 0th, 1st, and 2nd views a simulated curved grid surface. The 0th and the 2nd cameras are nearly in a fixating configuration.

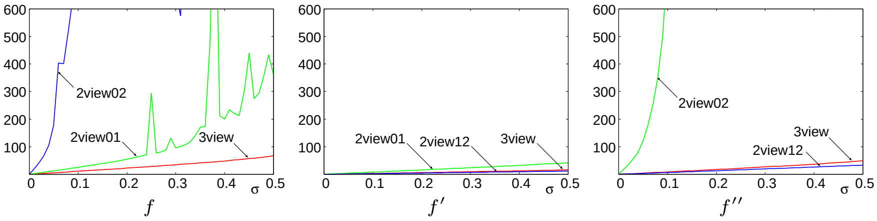


Fig. 2. The RMS error of focal length computation for σ, where "2view01", etc. denote the values computed from the 0th-1st image pair, etc., and "3view" means the value computed from the three views.

Let $\mathbf{X}_\alpha = (X_\alpha, Y_\alpha, Z_\alpha)^\top$ be the 3-D position of the αth point, and $\hat{\mathbf{x}}_\alpha$, $\hat{\mathbf{x}}'_\alpha$, $\hat{\mathbf{x}}''_\alpha$ its 2-D positions in the 0th, 1st, and 2nd views, respectively, after the optimal correction. The following projection relationships hold:

$$\hat{\mathbf{x}}_\alpha \simeq \mathbf{P}\begin{pmatrix}\mathbf{X}_\alpha \\ 1\end{pmatrix}, \qquad \hat{\mathbf{x}}'_\alpha \simeq \mathbf{P}'\begin{pmatrix}\mathbf{X}_\alpha \\ 1\end{pmatrix}, \qquad \hat{\mathbf{x}}''_\alpha \simeq \mathbf{P}''\begin{pmatrix}\mathbf{X}_\alpha \\ 1\end{pmatrix}. \tag{20}$$

These define in total six linear equations in $\mathbf{X}_\alpha$. Since Eqs. (20) *exactly* hold due to the optimal correction procedure, we can choose any three equations to solve for $\mathbf{X}_\alpha$ (or all equations by least squares). If the point is visible only in two views, we choose three equations from their corresponding projection relationships.

So far, we have assumed that the sign of $\mathbf{E}_{01}$ is correct (Section 5). If its sign is wrong (hence the signs of $\mathbf{E}_{02}$ and $\mathbf{E}_{12}$ are also wrong), the reconstructed shape is a mirror image of the true shape locating *behind* the cameras [5, 7]. Hence, if $\sum_\alpha^N \mathrm{sgn}(Z_\alpha) < 0$, for the visible points from the 0th camera, where $\mathrm{sgn}(x)$ returns 1, -1, and 0 according to $x > 0$, $x < 0$, and $x = 0$, respectively, we reverse the signs of all $(X_\alpha, Y_\alpha, Z_\alpha)^\top$.

7 Simulation Experiments

Figure 1 shows three simulated views (0th, 1st, and 2nd from left) of a grid surface. The frame size is assumed to be 800 × 800 pixels and the focal lengths $f = f' = f'' = 600$ pixels. We added independent Gaussian random noise of mean 0 and standard deviation σ pixels to the x and y coordinates of each grid

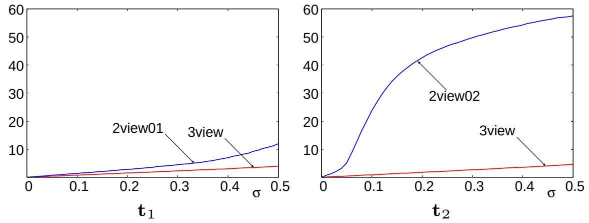


Fig. 3. The RMS error (in degree) of translation computation for σ. "2view01", etc. denote the values computed from the 0th-1st image pair, etc., and "3view" means the value computed from the three views.

point and conducted calibration and 3-D reconstruction. For a computed focal length f, we evaluated the difference $\Delta f = f - \bar{f}$ from its true value $\bar{f}$. If the computation failed ("imaginary focal lengths"), we let $f = 0$. Since the absolute scale of translation is indeterminate, we evaluated for a computed translation $\mathbf{t}$ the angle $\Delta\theta = \cos^{-1}(\mathbf{t}, \bar{\mathbf{t}})/\|\mathbf{t}\| \cdot \|\bar{\mathbf{t}}\|$ (in degree) it makes from its true value $\bar{\mathbf{t}}$. If the computation failed due to imaginary focal lengths, we let $\Delta\theta = 90°$. For a computed rotation $\mathbf{R}$, we evaluated the angle $\Delta\Omega$ (in degree) of the relative rotation $\mathbf{R}\bar{\mathbf{R}}^\top$ from the true value $\bar{\mathbf{R}}$. If the computation failed due to imaginary focal lengths, we let $\Delta\Omega = 90°$. Then, we evaluated the RMSs

$$E_f = \sqrt{\frac{1}{K}\sum_{a=1}^{K}\Delta f_a^2}, \qquad E_{\mathbf{t}} = \sqrt{\frac{1}{K}\sum_{a=1}^{K}\Delta\theta_a^2}, \qquad E_{\mathbf{R}} = \sqrt{\frac{1}{K}\sum_{a=1}^{K}\Delta\Omega_a^2}, \tag{21}$$

over $K = 10000$ independent trials, each time using different noise, where the subscript a indicates the value of the ath trial.

Figure 2 compares the accuracy of focal lengths computed from two views and from three views. We see that f and f'' computed from the 0th-2nd image pair have large errors. This is because the 0th and 2nd cameras are nearly in a fixating configuration. The large fluctuations of the plots indicate the occurrence of imaginary focal lengths. However, we can obtain accurate values for all the focal lengths if we use three images. In this noise range, no imaginary focal lengths occurred for three-view computation. Figures 3 and 4 compare the accuracy of translation and rotation. The error is large for the values computed from the 0th-2nd image pair due to the low accuracy of the focal length computation from them. As we see, however, we can obtain accurate values by using three views despite the fixating camera configuration of the 0th and 2nd cameras.

8 Endoscopic Image Experiments

Figure 5 shows two sets of three consecutive frames of intestine tract images taken by an endoscope receding along the tract. It is well known that if a camera is moved forward or backward, two-view reconstruction frequently fails because any two camera positions are nearly in a fixating configuration, frequently resulting in imaginary focal lengths. Hence, this is a good testbed for examining the

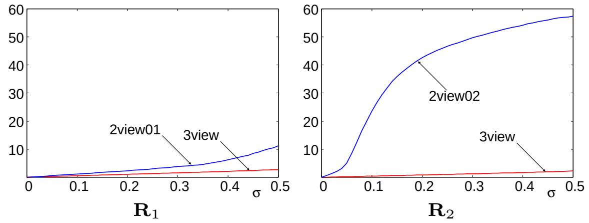


Fig. 4. The RMS error (in degree) of rotation computation for σ. "2view01", etc. denote the values computed from the 0th-1st image pair, etc., and "3view" means the value computed from the three views.

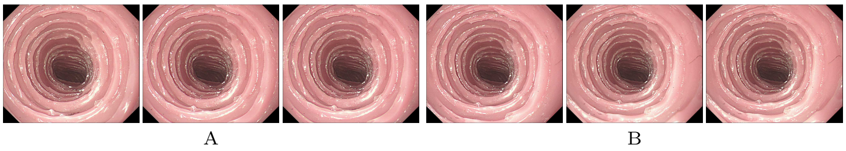


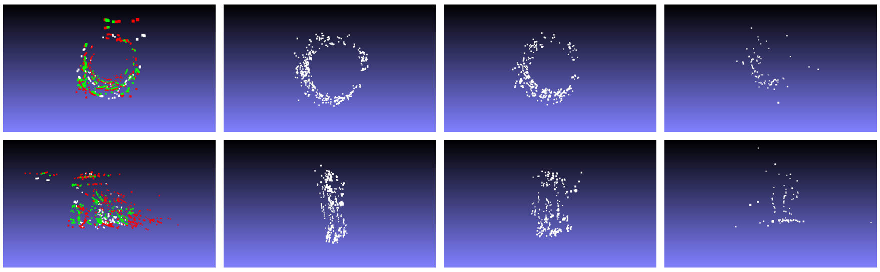
Fig. 5. Two sets of three consecutive frames of endoscopic intestine tract images.

degeneracy tolerance capability of our method. At the same time, our method, if successful, would bring about a new medical application of reconstructing 3-D structures from endoscopic images.

We extracted feature points and matched them between each pair of frames, using the method of Hirai, et al. [6]. Figure 6(a) shows the reconstruction from the three frames of the data set A in Fig. 5. For comparison, Fig. 6(b), (c), (d) shows the two-view reconstructions from the 0th-1st frame pair, the 0th-2nd frame pair, and the 1st-2nd frame pair, respectively; only those points viewed in the corresponding image pairs are reconstructed.

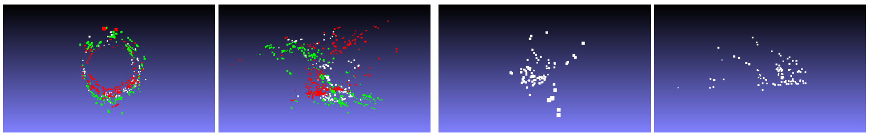
Since the ground truth is not known, we cannot tell which of (a), (b), (c), and (d) is the most accurate. As we can see, however, the three-view reconstruction (a) provides a detailed shape in a longer range along the tract with a larger number of points than the two-view reconstructions (b), (c), and (d). Ideally, the superimposition of (b), (c), and (d) should coincide with (a) if we correctly adjust the scale of the two-view reconstructions in (b), (c), and (d) (recall that the scale is indeterminate in each reconstruction). For real data, however, the two-view reconstructions do not necessarily agree with the three-view reconstruction. In this sense, our three-view reconstruction can be viewed as automatically adjusting the scales of two-view reconstructions and optimally merging them into a single shape.

Figure 7 shows the reconstruction from the data set B in Fig. 5. Figure 7(a) shows the resulting three-view reconstruction. In this case, two-view reconstruction was possible only from the 1st-2nd frame pair (Fig. 7(b)); the computation failed both for the 0th-1st frame pair and for the 0th-2nd frame pair due to imaginary focal lengths. Yet, using three images, we can accurately compute the 3-D positions of all pairwise matched points and obtain a detailed structure in a longer range along the tract.



(a) (b) (c) (d)

Fig. 6. Front views (above) and side views (below) of the 3-D reconstruction from the data set A in Fig. 5. (a) Using the three frames. Different colors indicate different image pairs they originate from. (b) Using the 0th-1st frame pair. (c) Using the 0th-2nd frame pair. (d) Using the 1st-2nd frame pair.



(a) (b)

Fig. 7. Front views and side views of the 3-D reconstruction from the data set B in Fig. 5. (a) Using the three frames. (b) Using the 1st-2nd frame pair. Reconstruction from the 0th-1st frame pair and reconstruction from the 0th-2nd frame pair both fail.

9 Concluding Remarks

We have presented a new method for initializing 3-D reconstruction from three views, generating a candidate solution to be refined later. Our main focus is to prevent the imaginary focal length degeneracy, which two-view reconstruction frequently suffers. Our method does not require correspondences among the three images; all we need is three fundamental matrices of image pairs. We exploited the redundant information provided by the three fundamental matrices to optimize the camera parameters and the 3-D structure. We conducted numerical simulation and observed that imaginary focal lengths never occurred in the experimented noise range while two-view computation frequently failed, resulting in higher average accuracy of our method than two-view reconstruction. We also tested the degeneracy tolerance capability of our method by using endoscopic intestine tract images, noting that the camera configuration is almost always near degeneracy. We observed that unlike two-view reconstruction our three-view computation never failed in our experimented instances (not all shown here) and that even when two-view reconstruction did not fail, our method produced a more detailed structure in a wider range than pairwise two-view reconstructions combined. Thus, our method is expected to bring about new medical applications to endoscopic image analysis.

Acknowledgments: This work was supported in part by JSPS Grant-in-Aid for Young Scientists (B 23700202), for Scientific Research (C 24500202), and for Challenging Exploratory Research (24650086) and the 2013 Material and Device Joint Project (2013355).

References

1. S. Bougnoux, From projective to Euclidean space under any practical situation, a criticism of self-calibration, *Proc. 6th Int. Conf. Comput. Vis.*, pp.790–796, Jan. 1998.
2. J. Goldberger, Reconstructing camera projection matrices from multiple pairwise overlapping views, *Comput. Vis. Image Understanding*, **97** (2005), 283–296.
3. R. Hartley, Estimation of relative camera positions for uncalibrated cameras, *Proc. 2nd European Conf. Comput. Vis.*, May 1992, Santa Margehrita Ligure, Italy, pp.579–587.
4. R. Hartley and C. Silpa-Anan, Reconstruction from two views using approximate calibration, *Proc. 5nd Asian Conf. Comput. Vis.*, Jan. 2002, Melbourne, Australia, pp.338–343.
5. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge University Press, Cambridge, U.K., 2004.
6. K. Hirai, Y. Kanazawa, R. Sagawa and Y. Yagi, Endoscopic image matching for reconstructing the 3-D structure of the intestines, *Med. Imag. Tech.*, **29**-1 (2011-1), 36–46.
7. K. Kanatani, *Geometric Computation for Machine Vision*, Oxford University Press, Oxford, U.K., 1993.
8. K. Kanatani and C. Matsunaga, Closed-form expression for focal lengths from the fundamental matrix, *Proc. 4th Asian Conf. Comput. Vis.*, January 2000, Taipei, Taiwan, Vol.1, pp.128–133.
9. K. Kanatani, A. Nakatsuji and Y. Sugaya, Stabilizing the focal length computation for 3-D reconstruction from two uncalibrated views, *Int. J. Comput. Vis.*, **66**-2 (2006-2), 109-122.
10. K. Kanatani and Y. Sugaya, Compact fundamental matrix computation, *IPSJ Trans. Comput. Vis. Appl.*, **2** (2010-3), 59–70.
11. K. Kanatani, Y. Sugaya and Y. Kanazawa, Latest algorithms for 3-D reconstruction from two views, in C. H. Chen (Ed.), *Handbook of Pattern Recognition and Computer Vision*, 4th ed., World Scientific Publishing, 2009, pp. 201-234.
12. K. Kanatani, Y. Sugaya, and H. Niitsuma, Triangulation from two views revisited: Hartley-Sturm vs. optimal correction, *Proc. 19th British Machine Vis. Conf.*, September 2008, Leeds, U.K., pp. 173–182.
13. M. I. A. Lourakis and A. A. Argyros, SBA: A software package for generic sparse bundle adjustment, *ACM Trans. Math. Software*, **36**-1 (2009-3), 2:1–30.
14. M. Pollefeys, R. Koch, and L. V. Cool, Self-calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters, *Int. J. Comput. Vis.*, **32**-1 (1999-1), 7–25.
15. N. Snavely, S. Seitz and R. Szeliski, Photo tourism: Exploring photo collections in 3d, *ACM Trans. Graphics*, **25**-8 (1995), 835–846.
16. N. Snavely, S. Seitz and R. Szeliski, Modeling the world from Internet photo collections, *Int. J. Comput. Vis.*, **80**-2 (2008-11), 189–210.