Projective 3D-reconstruction of Uncalibrated Endoscopic Images

P. Faltin, A. Behrens

Abstract

The most common medical diagnostic method for urinary bladder cancer is cystoscopy. This inspection of the bladder is performed by a rigid endoscope, which is usually guided close to the bladder wall. This causes a very limited field of view; difficulty of navigation is aggravated by the usage of angled endoscopes. These factors cause difficulties in orientation and visual control. To overcome this problem, the paper presents a method for extracting 3D information from uncalibrated endoscopic image sequences and for reconstructing the scene content. The method uses the SURF-algorithm to extract features from the images and relates the images by advanced matching. To stabilize the matching, the epipolar geometry is extracted for each image pair using a modified RANSAC-algorithm. Afterwards these matched point pairs are used to generate point triplets over three images and to describe the trifocal geometry. The 3D scene points are determined by applying triangulation to the matched image points. Thus, these points are used to generate a projective 3D reconstruction of the scene, and provide the first step for further metric reconstructions.

Keywords: 3D reconstruction, uncalibrated camera, epipolar geometry, trifocal geometry, bladder, cystoscopy, endoscopy.

1 Introduction

With about 68 810 new cases in 2008 in the United States [1], bladder cancer is a common disease of the urinary system. Tumors are usually inspected and treated by endoscopic interventions. Urological intervention of the bladder and urethra is also called cystoscopy. The cystoscope is inserted into the bladder through the urethra, which allows an inspection of the bladder wall. The inspection is usually performed close to the bladder wall, which is why the field of view is very limited. A possible way to improve the difficult orientation is e.g. by using an image mosaicking algorithm [2] to provide a panoramic overview image, or by generating a 3D model of the bladder. This paper presents a method for extracting 3D information from an uncalibrated endoscopic image sequence, which is then used for a projective 3D bladder reconstruction. In further steps, this information can be used for auto calibration of the camera, which leads to the desired metric reconstruction.

The paper is organized as follows: In section 2 the image preprocessing, the mathematical reconstruction and the reconstruction algorithms are described. In section 3 the evaluated image sequences and the results are presented. Finally section 4 summarizes the results and gives prospects for future work.

2 Reconstruction

2.1 Imaging

The image sequences are acquired by a rigid video endoscope system, in this case an Olympus EVIS EXERA II platform. At the ocular of the cystoscope, a CCD camera is attached, which delivers the data to a workstation. To illuminate the organ, a light source is coupled into the cystoscope. To increase the field of view, endoscopes usually have a fish-eye optic. A typical setup is shown in fig. 1. The RealTimeFrame software framework [3] is used for real-time data processing of endoscopic data. This software allows a very rapid prototyping of algorithms.

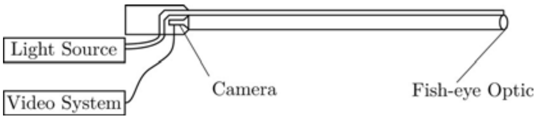


Fig. 1: A schematic view of a rigid cystoscope

2.2 Distortion correction

Cystoscope optics produces a strongly radial distorted image, which has to be corrected to extract valid 3D information. To compensate this distortion, the method of Hartley and Kang [4] is applied to each image in a preprocessing step. The radial distortion is modeled by the function

$$\vec{x}_d = \vec{z} + \lambda(r) \cdot (\vec{x}_u - \vec{z}) \tag{1}$$

with distorted point $\vec{x}_d$, center of distortion $\vec{z}$, a function depending on the radius $\lambda(r)$ and corrected point $\vec{x}_u$. $\lambda(r)$ is not based on a fixed model function but is dynamically determined, resulting in very precise distortion correction. An example using the implementation from [8] is shown in fig. 2.

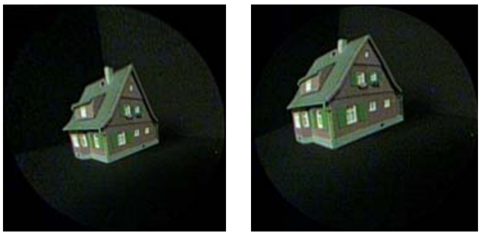


Fig. 2: Image distorted (left) by an endoscope, and image after correction (right)

2.3 Feature detection

Feature detection is accomplished by the SURF-algorithm [5], which extracts and describes distinctive points in each image independent of its scale, position and rotation. To detect points of interest, a Hessian matrix containing the approximated second order partial derivatives of a Gaussian function from fig. 3 is used. The extracted features are described by an analysis of the surrounding area via Haar wavelets. The results are stored in the descriptor vectors of the features. A simple comparison of different features can be made by summing the absolute summed differences of these vectors.

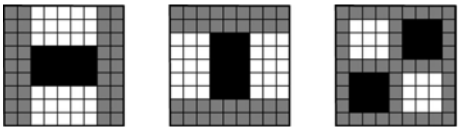


Fig. 3: Box filters in Hessian matrix

2.4 2-View correspondences

A simple method for generating correspondences over two images is brute matching. During this process, each feature f_1 from the first image is compared with every feature f_2 from the second image, and the f_2 which minimizes the difference between the feature vectors is chosen. A correspondence is identified, if the difference is less than a given threshold. This process usually results in a high number of wrong correspondences. Thus, advanced matching is applied.

In addition to forward matching, where the best f_2 for each f_1 is chosen, backward matching is applied by choosing the best f_1 for each f_2. A correspondence is identified only if these two matchings are equal. To improve the robustness, a slight restriction of the scale and orientation of the features by factor two, respectively 45°, is also applied. This assumption is valid since the position of the endoscope does not change much between two sequential images in a real bladder inspection. The last check is whether the detected best correspondence is reliable by comparing its distance d_{best} with the distance d_{2ndbest} from the second-best one via looking at their ratio $d_{\mathrm{best}}/d_{\mathrm{2ndbest}} > \tau$.

2.5 Epipolar geometry

Epipolar geometry describes the setup of two cameras looking at the same scene from different points of view. While in this section only the basic fundamentals are described, more details can be found in [14]. An exemplary camera setup showing the camera centers $\vec{C}$ and $\vec{C}\,'$ is given in fig. 4. The 3D-point $\vec{X}$ is projected onto the image planes, resulting in points $\vec{x}$ and $\vec{x}\,'$. The points $\vec{e}$ and $\vec{e}\,'$ are called epipoles, and they represent the projected camera centers on the image planes. The position, orientation and properties of the two cameras are described by the fundamental matrix $\boldsymbol{F}$. It has seven degrees of freedom and rank 2. Only the intrinsic geometry is described, which is why the fundamental matrix is independent of the scene content. A central equation for understanding epipolar geometry is the epipolar relation

$$\vec{x}\,'^{\,T} \cdot \boldsymbol{F} \cdot \vec{x} = 0 \tag{2}$$

which connects an image point from one image plane with its corresponding point on the other image plane. Epipolar lines for each image point can be calculated using the fundamental matrix. This line passes through the position of the corresponding point on the other image plane. All these lines intersect in the epipoles. They can be calculated by

$$\vec{l}\,' = \boldsymbol{F} \cdot \vec{x} \quad \text{or} \quad \vec{l} = \boldsymbol{F}^T \cdot \vec{x}\,' \,. \tag{3}$$

A geometric interpretation of eq. 3 is visualized in fig. 5.

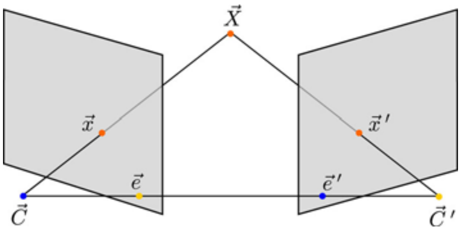


Fig. 4: Epipolar geometry

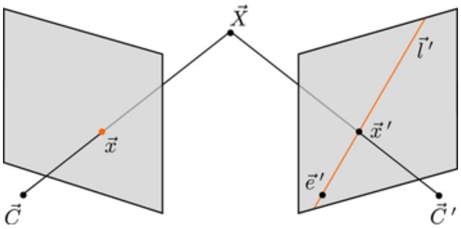


Fig. 5: Epipolar line

The different frames of the sequence are interpreted as different views. This is correct if there is a camera movement between the frames and if the scene is stationary.

The RANSAC-algorithm [6] is applied to estimate the fundamental matrix. This algorithm generates a random set of correspondences and calculates the fundamental matrix. Using backprojection, each corresponding point is classified as an inlier or an outlier. To be classified as an inlier, the reprojection error of a point pair has to be smaller than a threshold. For an acceptable computation time, the Sampson-Approximation [9] is used to determine the error. This process is repeated iteratively for other random samples. Finally, the inliers of the fundamental matrix, which yields to the largest number of inliers, are chosen to calculate the final fundamental matrix, and all outliers are eliminated. The RANSAC-algorithm uses the 7-point algorithm to calculate the fundamental matrices. The final matrix is estimated using the 8-point algorithm [10], which can handle eight or more points and provides a least squares solution. Both algorithms use a system of equations constructed with eq. 2.

2.6 2-View camera matrices

To perform a reconstruction at least two camera matrices are required. If the first camera is chosen with $\boldsymbol{P} = [\boldsymbol{I}|\vec{0}]$ the second camera matrix is defined by

$$\boldsymbol{P'} = [[\vec{e}\,']_\times \cdot \boldsymbol{F} + \vec{e}\,' \cdot \vec{v}^{\,T} | \lambda \cdot \vec{e}\,'] \,. \tag{4}$$

The fundamental matrix $\boldsymbol{F}$ and the epipole $\vec{e}\,'$ are known, but the scalar λ and the vector $\vec{v}$ are unknown. Correspondingly, there are four degrees of freedom to choose the second camera. Therefore a scene reconstruction based on eq. 4 is subjected to a projective transformation compared to the original scene, as shown in [11]. Fig. 6 shows an example. Without any camera calibration or additional scene information, no metric reconstruction from two views is feasible.

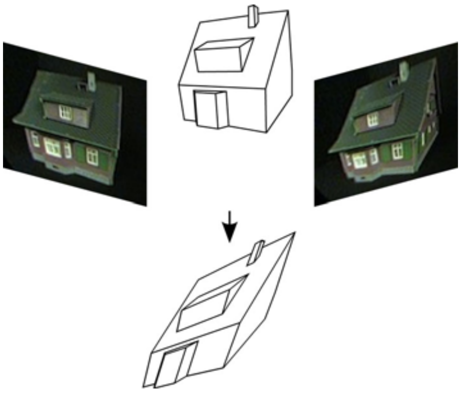


Fig. 6: Reconstruction with projective transformation

2.7 3-View correspondences

It seems to be a straightforward process to connect two matched image pairs with one common image to an image triplet using the two view correspondences. But in practise the SURF-algorithm cannot detect identical features in the three images, because of view changes and noisy image data. Additionally, not all correspondences could be identified. This results in situations where not every feature could be retrieved in all images. This is visualized in fig. 7. As can easily be seen, only a small number of correspondences share the same corresponding point in the image $i + 1$ indicated by surrounding circles. To increase the number of correspondences over three images, an additional matching process from the first to the third view is performed. This step may induce new incorrect matches, which have to be considered. Thus, an advanced RANSAC-algorithm is used to join the set of tracked correspondences and the set of directly matched correspondences, and to detect outliers. The set of tracked correspondences contains a high amount of valid ones. To benefit from this fact, the RANSAC-algorithms fills the samples with a higher probability from the tracked set than from the directly matched set of correspondences. But even if the tracked correspondences are verified by epipolar matching, they should not be chosen definitely, because they could still be wrong, as fig. 8 shows. $\vec{x}\,'_1$ and $\vec{x}\,'_2$ are located on the epipolar line, which corresponds to the point $\vec{x}_1$. This implies the epipolar relation is fulfilled, and a correspondence between $\vec{x}_1$ and $\vec{x}\,'_1$ or between $\vec{x}_1$ and $\vec{x}\,'_2$ appears to be correct in the second view. Only in the third view it is possible to identify the wrong correspondence $\vec{x}_1$ and $\vec{x}\,''_2$.

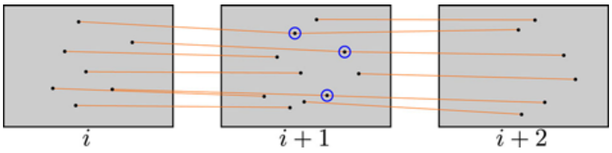


Fig. 7: Three image correspondences

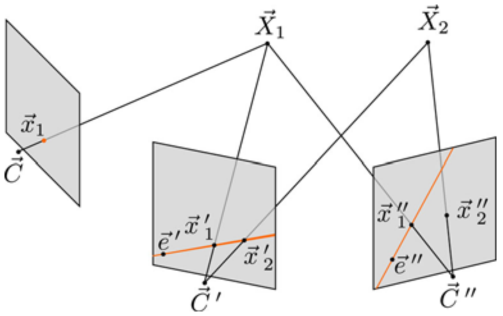


Fig. 8: Wrong correspondence detected in three views

2.8 Trifocal geometry

The mathematical description of trifocal geometry uses tensor notation. A good introduction to this topic can be found in [7] and [14]. In this paper, tensors are written in calligraphic letters and the Einstein notation is used.

Trifocal geometry describes the setup of three different views for the same scene. Like epipolar geometry, trifocal geometry is also intrinsic and thus independent from the scene content. A sample configuration is shown in fig. 8 with the camera centers $\vec{C}$, $\vec{C}\,'$ and $\vec{C}\,''$. The 3D point $\vec{X}_1$ is projected to the image points $\vec{x}_1$, $\vec{x}\,'_1$ and $\vec{x}\,''_1$. The epipoles $\vec{e}\,'$ and $\vec{e}\,''$ represent the projected camera center of the first camera on the image planes of the second and third view. The first camera is defined by $\boldsymbol{P} = [\boldsymbol{I}|\vec{0}]$, whereby the second camera is $\boldsymbol{P}' = [\boldsymbol{A}|\vec{e}\,']$ and the third camera is $\boldsymbol{P}'' = [\boldsymbol{B}|\vec{e}\,'']$. The properties and the relation of these cameras are described by the trifocal tensor $\mathcal{T}$. This is a $3 \times 3 \times 3$ third-order tensor with 18 degrees of freedom. The reduction from 27 parameters to 18 degrees of freedom is caused by the internal constraint

$$\mathcal{T}_i^{jk} = \mathcal{P}'^i_j \mathcal{P}''^k_4 - \mathcal{P}'^j_4 \mathcal{P}''^k_i \tag{5}$$

with the camera matrices $\boldsymbol{P}'$ and $\boldsymbol{P}''$ in tensor notation $\mathcal{P}'$ and $\mathcal{P}''$.

By analogy with the epipolar relation $\vec{x}\,'^T \cdot \boldsymbol{F} \cdot \vec{x} = 0$ of two view geometry, trifocal geometry yields to

$$\mathcal{X}^i \mathcal{X}'^j \mathcal{X}''^k \mathcal{E}_{jpr} \mathcal{E}_{kqs} \mathcal{T}_i^{pq} = 0_{rs}\,. \tag{6}$$

The tensor $\mathcal{E}$ in eq. 6 – called the Levi-Cevita symbol – represents the constant third-order $3 \times 3 \times 3$ tensor

$$\mathcal{E}_{rst} = \begin{cases} 0 & \text{if } r, s, t \text{ not distinctive} \\ +1 & \text{if } (r, s, t) \text{ is an even permutation} \\ -1 & \text{if } (r, s, t) \text{ is an odd permutation} \end{cases} \tag{7}$$

with $r, s, t \in \{1, 2, 3\}$.

The trifocal tensor can also be written in matrix notation using the three 3×3 matrices

$$\boldsymbol{T}_{i\ jk} = \mathcal{T}_i^{jk} \text{ mit } i, j, k \in \{1, 2, 3\}\,. \tag{8}$$

This notation can be used to extract the fundamental matrices between two different views from the trifocal tensor using the equations

$$\boldsymbol{F}_{21} = [\vec{e}\,']_\times \cdot [\boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_3] \cdot \vec{e}\,'' \tag{9}$$

and

$$\boldsymbol{F}_{31} = [\vec{e}\,'']_\times \cdot [\boldsymbol{T}_1^T, \boldsymbol{T}_2^T, \boldsymbol{T}_3^T] \cdot \vec{e}\,'\,. \tag{10}$$

Here the notation $\vec{a}^{\,T} \cdot [\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3] \cdot \vec{b}$ represents the vector $(\vec{a}^{\,T} \cdot \boldsymbol{M}_1 \cdot \vec{b}, \vec{a}^{\,T} \cdot \boldsymbol{M}_2 \cdot \vec{b}, \vec{a}^{\,T} \cdot \boldsymbol{M}_3 \cdot \vec{b})$, and $[\vec{x}]_\times$ denotes the skew-symmetric matrix, which for a vector $\vec{a}$ is given by

$$[\vec{a}]_\times = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix}. \tag{11}$$

The trifocal tensor is calculated by the normalized linear algorithm [14] including algebraic minimization. The basic idea is to solve a system of equations generated from eq. 6 and to enforce the inner constrains given by eq. 5.

2.9 Triangulation

After calculating the camera matrices, the 3D points can be reconstructed using triangulation. The concept is that the projection lines through the camera centers $\vec{C}$ and $\vec{C}\,'$ and the image points $\vec{x}$ and $\vec{x}\,'$ intersect in the 3D point $\vec{X}$, as shown in fig. 5. To calculate $\vec{X}$ the equation

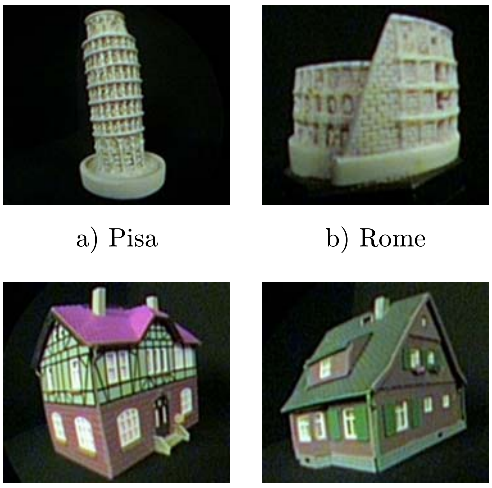
$$\begin{pmatrix} \bar{x} \cdot \vec{p}^{\,3T} - \vec{p}^{\,1T} \\ \bar{y} \cdot \vec{p}^{\,3T} - \vec{p}^{\,2T} \\ \bar{x}' \cdot \vec{p}\,'^{3T} - \vec{p}\,'^{1T} \\ \bar{y}' \cdot \vec{p}\,'^{3T} - \vec{p}\,'^{2T} \end{pmatrix} \cdot \vec{X} = \vec{0} \tag{12}$$

is solved, where $\vec{p}^{\,iT}$ and $\vec{p}\,'^{iT}$ are the i-th row vector of $\boldsymbol{P}$ and $\boldsymbol{P}'$.

Since subpixel positions can only be determined by interpolation and additional distortion is induced by the camera system, the detected image points are noisy. This results in the effect that two projection lines do not meet in space and instead form two skew lines. To overcome this problem, the image points $\vec{x}$ and $\vec{x}\,'$ are adjusted to meet the epipolar relation and are called $\bar{\vec{x}}$ and $\bar{\vec{x}}\,'$. Simultaneously, the sum of the Euclidean distance sum $d(\vec{x}, \bar{\vec{x}})^2 + d(\vec{x}\,', \bar{\vec{x}}\,')^2$ is minimized.

3 Results

The four different endoscopic video sequences from fig. 9 are used to analyze the different steps of the algorithm.



a) Pisa b) Rome

c) Train Station d) Wooden House

Fig. 9: Test sequences

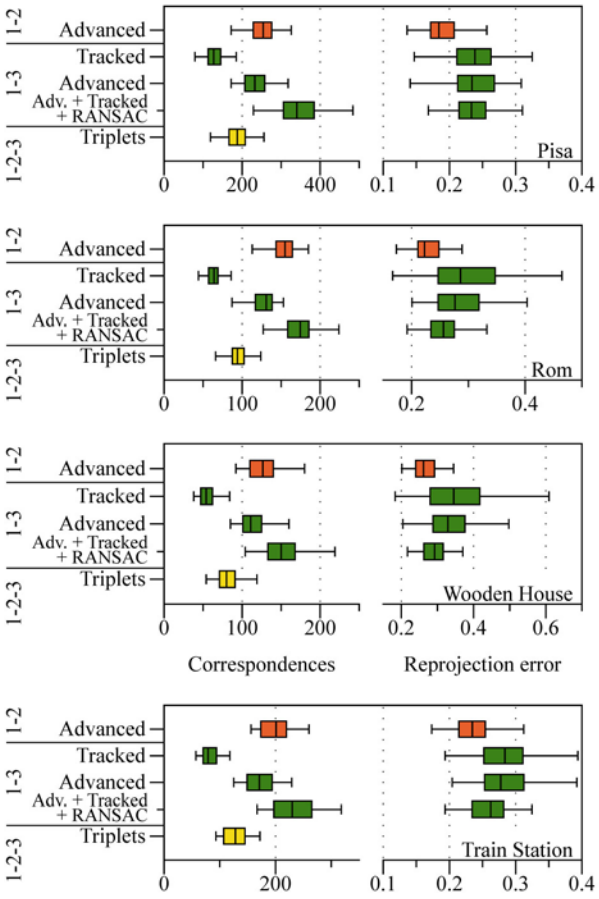


Fig. 10: Correspondences for two views

Boxplots are used to compare the data statistically over all frames of the sequences. 25 % or 75 % of all measured values are below the values indicated by the box borders. The line inside the box is the median and the whiskers indicate the 5 % and 95 % percentile.

Fig. 10 analyses the matching process for three views, as described in section 2.7. The number of correspondences is shown on the left side of the boxplots. On the right side, the related reprojection error is shown. The first row shows the results from the two-view process using the advanced matching from section 2.5 compared to the three-view process. Analysing the "Wooden House" sequence it is observable that the number of tracked correspondences (1–3) of about 60 is significantly lower than the number representing the directly identified correspondences (1–2), which is of about 125. The advanced matching between the first and third view (1–3) is only slightly inferior than the matching from the first view to the second view (1–2). This can be explained by the higher temporal distance between the images, which leads to higher variation of the image data. In the fourth row, the tracked and newly matched correspondences are joined using the advanced RANSAC-algorithm (1–3), which leads to the higher number of about 150 correspondences, compared to the matching among the first and second view. Finally only correspondences between all three views (1–2–3), called triplets, are selected, which reduces the total number to about 90. The mean error of about 0.3 pixels is constantly low. Only the tracked correspondences (1–3) show slightly higher error and variation, before applying the RANSAC-algorithm. This is caused by sporadic wrong correspondences, as described in section 2.7.

Finally, the reprojection error from the estimated trifocal tensor is analyzed. For this step, two fundamental matrices are calculated from the trifocal tensor by eq. 9 (1–2) and 10 (1–3). Fig. 11 shows for all sequences nearly the same subpixel error of about 0.3 pixels (1–2) or ∼ 0.6 pixels (1–3), respectively. Compared to epipolar geometry, the temporal distance has a stronger impact. Since no RANSAC-algorithm has yet been applied for estimating the trifocal tensor, outliers have a direct impact on the error value.

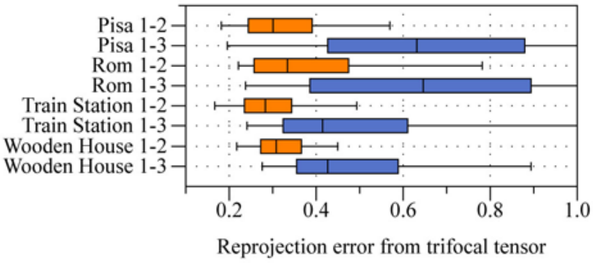


Fig. 11: Trifocal error

An exemplary reconstruction of the "Pisa" sequence is shown in fig. 12. In the left image all detected features are shown on the image plane, and in the right image a 3D reconstruction from these points is shown. Corresponding points have the same color. The reconstruction is compressed in the x-direction, caused by the projective transformation described in section 2.6.

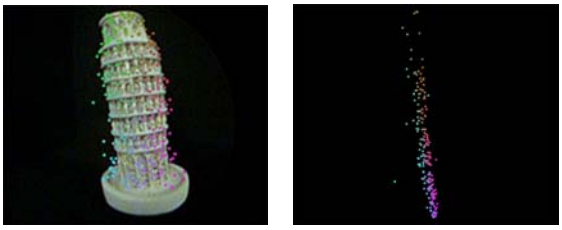


Fig. 12: Image from sequence "Pisa" and related reconstruction

4 Summary and prospects

This paper presents a method for reconstructing 3D scene content from uncalibrated endoscopic sequences based on SURF-features. The different steps yield robust results by using the RANSAC-algorithm in adapted forms. It has been shown that an epipolar geometry and a trifocal geometry can be extracted with high precision, whereby subpixel-precise reconstruction is possible. An important application for trifocal geometry is for extracting consistent camera matrices for the whole sequence by a linear method, like in [12] or [13]. Subsequently these cameras can be used for auto calibration [14], which allows metric reconstruction.

Acknowledgement

We would like to thank Prof. Dr.-Ing. Til Aach, Institute of Imaging and Computer Vision, RWTH Aachen University, Germany for supervising this project.

Additionally we would like to thank Olympus & Winter Ibe GmbH for providing a video endoscope system.

References

[1] American Cancer Society: *Cancer Facts & Figures 2008.* American Cancer Society, 2008.

[2] Behrens, A., Bommes, M., Stehle, T., Gross, S., Leonhardt, S., Aach, T.: A Multi-Threaded Mosaicking Algorithm for Fast Image Composition of Fluorescence Bladder Images. *Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling*, Vol. **7 625**, San Diego, CA, USA, 2010.

[3] Gross, S., Behrens, A., Stehle, T.: Rapid Development of Video Processing Algorithms with RealTimeFrame. *Conference Book Biomedica*, Liege, Belgium, 2009, pp. 217–220.

[4] Hartley, R., Kang, S. B.: Parameter-Free Radial Distortion Correction with Center of Distortion Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, Vol. **29**, No. 8, pp. 1 309–1 321.

[5] Bay, H., Ess, A., Tuytelaars, T., Gool, L. V.: Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 2008, Vol. **110**, No. 3, pp. 346–359.

[6] Fischler, M. A., Bolles, R. C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981, Vol. **24**, No. 6, pp. 381–395.

[7] Triggs, W.: The geometry of projective reconstruction I: Matching constraints and the joint image. *Proc. International Conference on Computer Vision*, Boston, MA, USA, 1995, pp. 338–343.

[8] Stehle, T., Truhn, D., Aach, T., Trautwein, C., Tischendorf, J.: Camera Calibration for Fish-Eye Lenses in Endoscopy with an Application to 3D Reconstruction. *Proceedings IEEE International Symposium on Biomedical Imaging*, Washington, D.C., USA, 2007.

[9] Sampson, P. D.: Fitting Conic Sections to 'Very Scattered' Data: An Iterative Refinement of the Bookstein Algorithm. *Computer Graphics and Image Processing. Fitting conic sections to scattered data*, 1982, Vol. **18**, No. 1, pp. 97–108.

[10] Hartley, R.: In Defense of the Eight-Point Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, Vol. **19**, No. 6, pp. 580–593.

[11] Pollefeys, M.: *Self-calibration and metric 3D reconstruction from uncalibrated image sequences.* PhD Thesis, Leuven, Belgium, 1999.

[12] Triggs, B.: Linear Projective Reconstruction from Matching Tensors. *Image and Vision Computing*, 1997, Vol. **15**, Issue 8, pp. 617–625.

[13] Avidan, S., Shashua, A.: Threading Fundamental Matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, Vol. **23**, No. 1, pp. 73–77.

[14] Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision.* 2nd ed. Cambridge University Press, 2003.

About the authors

Peter FALTIN was born in 1983 in Cologne, Germany. He studied Computer Engineering at RWTH Aachen University and received his Dipl.-Ing. in 2010. Since then he has held a PhD position at the Institute of Imaging & Computer Vision at RWTH Aachen University. His research focuses on medical image processing, video processing, signal processing and computer vision.

Alexander BEHRENS was born in Bückeburg, Germany in 1980. He received a Dipl.-Ing. degree in electrical engineering from the Leibniz University of Hannover, Hannover, Germany, in 2006. After receiving the degree, he worked as a Research Scientist at the university's Institut für Informationsverarbeitung. Since 2007, he has been a Ph.D. candidate at the Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany. His research interests are in medical image processing, signal processing, pattern recognition, and computer vision.

Peter Faltin
Alexander Behrens
E-mail: Peter.Faltin@lfb.rwth-aachen.de
Alexander.Behrens@lfb.rwth-aachen.de
Institute of Imaging & Computer Vision
RWTH Aachen University
D-52056 Aachen, Germany