MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
8-3

Measuring the Shape of Sewer Pipes from Video

Juho Kannala
Machine Vision Group
Dept. of Electrical and Information Engineering
P.O. Box 4500, 90014 University of Oulu, Finland
Juho.Kannala@ee.oulu.fi

Sami S. Brandt
Laboratory of Computational Engineering
Helsinki University of Technology
P.O. Box 9203, 02015 TKK, Finland
Sami.Brandt@tkk.fi

Abstract

In this paper we address the problem of automatic measurement of the shape of sewer pipes. We describe a method for recovering the interior shape of a sewer pipe from a video sequence that is acquired by a fish-eye lens camera moving inside the pipe. The method is based on solving the general structure-from-motion problem by tracking interest points across successive video frames. Here the interest points are points where the image intensity changes rapidly due to irregularities in the surface texture of the pipe. The experiments with real videos of concrete pipes show that the shape of a sewer pipe can be recovered solely from the video. The proposed method can be additionally used in other applications to recover the scene structure from video sequences taken by a calibrated fish-eye lens camera.

1. Introduction

The condition assessment of sewer pipes is usually carried out by visual inspection of sewer video sequences. However, the manual inspection has a number of drawbacks such as subjectivity, varying standards and high costs. Therefore several approaches for automation of sewer surveys have been suggested. For example, automatic detection of pipe joints and surface cracks from digital sewer images has been investigated [2]. In [13], a method for automatic detection of pipe joints and their shape analysis was proposed. An idea of recovering the three-dimensional shape of a surveyed pipe from survey videos was presented in [3], where a method for determining the pose of the camera relative to the central axis of the pipe was additionally proposed. Unfortunately that method is restricted to brick sewers with visible mortar lines.

Different kinds of sewer robots have been developed and some of them contain additional sensors, such as range cameras, besides the video camera [7, 11]. While the additional sensors of multisensoric robots provide additional information, they also lead to a more complex and expensive construction.

In this paper, we propose a method for recovering the shape of a sewer pipe solely from a video sequence that is acquired by a fish-eye lens camera. Our approach is to solve the structure-from-motion problem in the case of fish-eye image sequences by tracking interest points across successive images. In Section 2, we give an overview of our method and, in Section 3, we describe its differences to conventional structure from motion approaches in more detail. Results of the experiments with a real sewer video are reported in Section 4.

2. Overview of the Method

A typical sewer inspection system consists of a video camera and a remote controlled tractor. The sewer robot we used had a fish-eye lens camera whose wide field of view makes it possible to obtain a high resolution scan of the whole pipe by a single pass. Our approach to structure recovery follows mainly the framework presented in [5]. However, since the usual pinhole camera model is not a valid approximation to a fish-eye lens several important modifications are proposed. In the following, we briefly describe the different steps in our method.

2.1 Camera calibration

Although the modern approach to structure recovery is often uncalibrated [5], we adopt the traditional photogrammetric principle of camera calibration prior to measurements. One reason for this is the peculiarity of the fish-eye lens and the other is the requirement of high accuracy. The calibration is done by viewing a planar calibration object [10].The calibration gives the transformation $\mathcal{T}$ that warps the original image to a perspective one, i.e., transforms the fish-eye image coordinates $\mathbf{m}$ to $\tilde{\mathbf{x}} = \mathcal{T}(\mathbf{m})$, which are the normalised image coordinates of a pinhole camera [10].

2.2 Feature extraction and matching

The tracked features are interest points detected by the Harris corner detector, which is widely used in this kind of ap-

plications. The experiments with sewer video sequences showed that there are plenty of such features in eroded concrete pipes. The detected interest points are initially matched between each successive image pair through intensity cross-correlation of the neighbourhood.

2.3 Feature tracking

The putative point correspondences contain almost unavoidably some false matches. In the tracking step, the aim is to use the geometric constraints between successive view pairs and view triplets to guide the matching and to discard the false matches. This requires some modifications to the usual way of estimating the multiple view geometry [5]. The modifications are needed to ensure a justified distribution of estimation error when the image correspondences are measured from the original fish-eye images and the two and three view relations, defined by the essential matrix and the trifocal tensor, hold between the corrected images. We will give a detailed describtion of the proposed modifications in Section 3.

2.4 Reconstruction

The final step is to recover the structure by computing the three-dimensional coordinates of the tracked points. If enough interest points can be tracked and reconstructed, the arrangement of the corresponding three-dimensional points should be tubular allowing to estimate the shape of the pipe.

Here we use a hierarchical method that is similar to [6] and optimally distributes the reconstruction error over the whole sequence. The idea is to start by computing the camera motion and the 3D points for each image triplet and then hierarchically registrate the triplets into longer subsequences which are bundle adjusted at each level. The difference between our implementation and [6] is that we use the calibrated approach. This makes the algorithm simpler since then the registration of two overlapping subsequences requires finding a similarity transformation instead of a general 3-space homography where the transform may be computed by a non-iterative algorithm [12].

3. Geometry of Fish-Eye Views and Tracking

Let $\mathbf{m}_j^i$ be the measured coordinates of point correspondence i in view j. Given these measured correspondences and assuming that the image measurement errors obey a zero-mean isotropic Gaussian distribution, the optimal way of estimating the camera motion is minimising

$$\sum_i \sum_j d(\mathbf{m}_j^i, \hat{\mathbf{m}}_j^i)^2, \tag{1}$$

where $d(\cdot,\cdot)$ is the distance between two image points and $\hat{\mathbf{m}}_j^i$ are the estimated correspondences

$$\hat{\mathbf{m}}_j^i = \mathcal{P}_j(\hat{\mathbf{X}}^i). \tag{2}$$

Here $\mathcal{P}_j$ is the imaging function of the fish-eye camera in view j and $\hat{\mathbf{X}}^i$ represent the estimated 3D coordinates of point i. Since the camera is calibrated, the values of the internal camera parameters [10] are known and the cost (1) should be minimised over the external camera parameters in $\mathcal{P}_j$ and the 3D coordinates $\hat{\mathbf{X}}^i$. However, the direct minimisation of (1) requires a good initialisation and does not tolerate false matches. Hence, we implemented the RANSAC algorithm for the robust estimation of camera motion between view pairs and triplets. The implementation follows the general recommendations in [8] but the adaptation to the fish-eye case is our own and is described in the following.

3.1 Two views

Consider the case of two views, $j = \{1,2\}$, and assume that there is a set of putative point correspondences, $\mathbf{m}_1^i \leftrightarrow \mathbf{m}_2^i$. The transformed coordinates are $\tilde{\mathbf{x}}_j^i = \mathcal{T}(\mathbf{m}_j^i)$ and the two view constraint between the transformed images is expressed by the essential matrix [8].

In RANSAC, we randomly select samples of seven point correspondences and each sample gives one or three candidates for the essential matrix [8]. Then, given an essential matrix candidate $\mathbf{E}$ and the transformed correspondences, $\tilde{\mathbf{x}}_1^i \leftrightarrow \tilde{\mathbf{x}}_2^i$, there is a non-iterative algorithm [8] for computing such points $\hat{\mathbf{x}}_1^i$ and $\hat{\mathbf{x}}_2^i$ that minimise the geometric distance

$$\sum_i d(\tilde{\mathbf{x}}_1^i, \hat{\mathbf{x}}_1^i)^2 + d(\tilde{\mathbf{x}}_2^i, \hat{\mathbf{x}}_2^i)^2 \tag{3}$$

in the transformed image plane subject to the constraint

$$\hat{\mathbf{x}}_2^{i\top} \mathbf{E} \hat{\mathbf{x}}_1^i = 0.$$

By transforming the points $\hat{\mathbf{x}}_j^i$ to the original image, one obtains the points

$$\hat{\mathbf{m}}_j^i = \mathcal{T}^{-1}(\hat{\mathbf{x}}_j^i), \tag{4}$$

which may be used as approximations to the optimal exact correspondences in (1). We use (4) to compute the distances

$$d(\mathbf{m}_1^i, \hat{\mathbf{m}}_1^i)^2 + d(\mathbf{m}_2^i, \hat{\mathbf{m}}_2^i)^2 \tag{5}$$

and classify the correspondences into inliers and outliers. It is important that this distance is measured in the original image since the transformation $\mathcal{T}$ is highly non-linear. As usual, the $\mathbf{E}$ which has most inliers is chosen and gives our first estimate of the camera motion.

Since the essential matrix may be parameterised with the rotation and translation parameters [8], the equation (4) implicitly defines the points $\hat{\mathbf{m}}_j^i$ as a function of the external camera parameters and the measured correspondences. Hence, by substituting the points (4) into (1) one may write

the minimisation problem in the form

$$\min_{\mathbf{z}} \sum_i C_i(\mathbf{y}_i, \mathbf{z})^2, \tag{6}$$

where the vectors $\mathbf{y}_i$ contain the measured correspondences in both views and $\mathbf{z}$ is the 5-vector containing the parameters of the essential matrix. We refine our motion estimate by minimising (6) using only the inlier correspondences.

The cost function in (6) has such a form that as a by-product of the minimisation one can compute an estimate for the covariance of the parameters $\mathbf{z}$. This is described in detail in [4]. The estimated covariance $\mathbf{\Lambda_z}$ may be used to compute the epipolar envelopes which constrain the search region for new correspondences.

Given a point $\mathbf{m}$ in the first image, the corresponding epipolar line in the transformed image plane of the second image is

$$\mathbf{l} = \frac{\mathbf{E}(\mathbf{z})\mathcal{T}(\mathbf{m})}{||\mathbf{E}(\mathbf{z})\mathcal{T}(\mathbf{m})||}, \tag{7}$$

and its covariance is approximated by

$$\mathbf{\Lambda_l} = \left(\frac{\partial \mathbf{l}}{\partial \mathbf{E}}\frac{\partial \mathbf{E}}{\partial \mathbf{z}}\right)\mathbf{\Lambda_z}\left(\frac{\partial \mathbf{l}}{\partial \mathbf{E}}\frac{\partial \mathbf{E}}{\partial \mathbf{z}}\right)^\top + \frac{\partial \mathbf{l}}{\partial \mathbf{m}}\mathbf{\Lambda_m}\frac{\partial \mathbf{l}}{\partial \mathbf{m}}^\top, \tag{8}$$

where $\mathbf{\Lambda_m}$ is the covariance of $\mathbf{m}$ and the Jacobians are computed from (7). Assuming that $\mathbf{l}$ is a random line obeying a Gaussian distribution with the mean at the estimated value and covariance (8) the epipolar envelope is the conic

$$\mathbf{C} = \mathbf{l}\mathbf{l}^\top - k^2\mathbf{\Lambda_l}, \tag{9}$$

which represents an equal-likelihood contour bounding some fraction of all instances of $\mathbf{l}$ [8]. If $F_2(k^2)$ represents the cumulative χ^2_2 distribution and k^2 is chosen such that $F_2^{-1}(k^2) = \alpha$, then a fraction α of all lines lie within the region bounded by $\mathbf{C}$.

We illustrate the estimated two view geometry in Fig. 1 where two successive images of a sewer video sequence are shown. We have chosen two points from the first image and plotted the corresponding epipolar curves and envelopes to the second image by transforming the epipolar lines (7) and the hyperbolas (9) to the original fish-eye image. The yellow crosses in the second image are the narrowest points of the envelopes [1]. The narrow envelope of the vertical curve is the 95 % confidence interval used in our experiments to constrain the search region.

3.2 Three views

The two view constraint significantly reduces the occurrence of false matches but the three view constraint is even more effective. In the three-view case, we first robustly estimate the camera motion for view pairs (1,2) and (1,3).

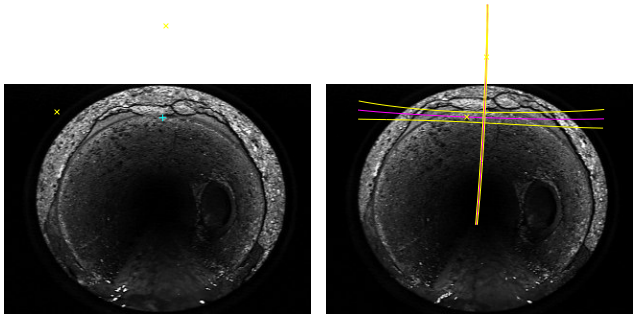

Figure 1: Estimated epipolar geometry for two fish-eye images. Two points in the first image are chosen (yellow crosses) and their epipolar curves (magenta curves) are plotted to the second image. The yellow curves are the epipolar envelopes. The envelope of the horizontal curve is broad because a very large value of $k^2 = 1000$ was chosen in (9) in order to better illustrate the error bounds. The narrow confidence interval of the vertical curve is the 95 % envelope that corresponds to a value $k^2 = 5.99$.

Then the only quantity that is left undetermined is the relative scale of the two translations, $\mathbf{t}_{1,2}$ and $\mathbf{t}_{1,3}$. We use the RANSAC procedure to determine this ratio from the three-view correspondences. At minimum only one additional sample correspondence needs to be drawn [9]. The distance measure used for the classification of inliers is

$$d(\mathbf{m}_1^i, \hat{\mathbf{m}}_1^i)^2 + d(\mathbf{m}_2^i, \hat{\mathbf{m}}_2^i)^2 + d(\mathbf{m}_3^i, \hat{\mathbf{m}}_3^i)^2, \tag{10}$$

where $\hat{\mathbf{m}}_1^i$ and $\hat{\mathbf{m}}_2^i$ are computed exactly as in (5) and $\hat{\mathbf{m}}_3^i$ is the point that is obtained by transferring the correspondence $\hat{\mathbf{m}}_1^i \leftrightarrow \hat{\mathbf{m}}_2^i$ to the third view with the trifocal point transfer.

The final estimate of the camera motion over each triple of views is refined by minimising (1) over both the motion parameters and the 3D coordinates of the inliers. We additionally iterate between (i) least-squares fit to inliers and (ii) re-classification of inliers; until convergence.

The robustly estimated three view geometry is used to guide the matching when establishing the final correspondences. Since the geometric constraint discriminates the false matches, a weaker similarity threshold can be employed for the correlation windows. By accepting only such correspondences that are found from at least three successive images, the occurrence of false matches becomes very improbable. The estimated camera motion for each image triplet also provides a basis for the final reconstruction as described in Sec. 2.4.

4. Experiments

We experimented a sewer video sequence scanned in an eroded concrete pipe. The uncompressed digital video was captured from an analog NTSC video signal at a resolution of 320×240. The experimented image sequence contained 159 fish-eye views and covered about two meters of the pipe. The total number of tracked interest points was 6864. With our current implementation we did not bundle

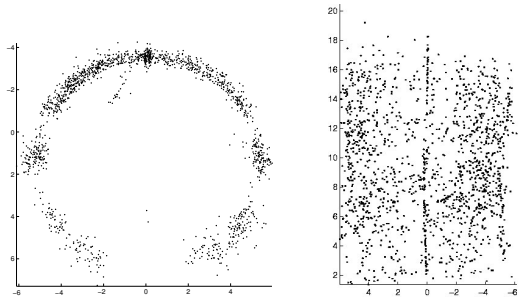

Figure 2: Front and top views of the reconstructed 3D points computed from a sequence of 35 images.

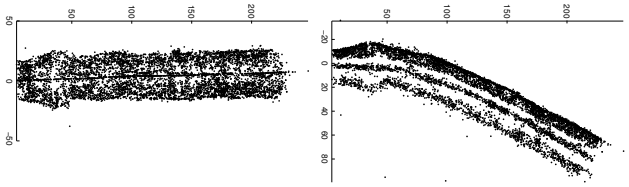

Figure 3: Top and side views of the reconstructed points for the sequence of 159 images. The thick part near the beginning of the pipe is a pipe socket in a displaced pipe joint.

adjust the entire sequence since the number of parameters would have been too large for a medium-scale Levenberg-Marquardt implementation. Sparse optimisation methods would obviously give a significant advantage, but we here computed the reconstruction by simply concatenating partial reconstructions that were bundle adjusted separately.

In Fig. 2, there is a three-dimensional reconstruction of points computed from correspondences over a subsequence of 35 images. There are 1512 points while the RMS projection error after the final bundle adjustment was 0.26 pixels. There are few reconstructed points in the bottom part of the pipe since it is difficult to find correspondences from the water region. The dense point cluster on the top is due to a sharp-edged crack in the middle of the roof. The points inside the pipe near the roof correspond to a root that is hanging from the roof.

To obtain a reconstruction of the whole pipe section, covered by the sequence of 159 views, we concatenated six partial reconstructions like shown in Fig. 2. As there is a three view overlap between each part, the partial reconstructions have common points and could be transformed into a common coordinate frame [12]. The result is shown in Fig. 3. The side view shows that the pipe is bent downwards. The bending is probably exaggerated here due to the accumulation of error in the concatenation because the concatenated reconstructions have an overlap of only three views.

5. Conclusions

We have proposed a novel method for recovering the scene structure from fish-eye image sequences and applied it to shape measurements of sewer pipes. The experiments show that the shape of a sewer pipe may be recovered solely from a video sequence that is scanned by a single pass through the pipe. In addition, the proposed framework is directly applicable to structure recovery from any video sequence taken by a calibrated camera suffering from severe lens distortion as the camera model [10] can flexibly model different kinds of distortions. The proposed method can be hence used in a wide range of potential applications.

Aknowledgements

We are grateful to Technical Research Centre of Finland and DigiSewer Productions Ltd. for cooperation and providing the sewer video sequences.

References

[1] S. S. Brandt. On the probabilistic epipolar geometry. In *Proc. BMVC*, pages 107–116, 2004.

[2] M. J. Chae and D. M. Abraham. Neuro-fuzzy approaches for sanitary sewer pipeline condition assessment. *J. Comput. Civil Eng.*, 15(1):4–14, January 2001.

[3] D. Cooper, T. P. Pridmore, and N. Taylor. Towards the recovery of extrinsic camera parameters from video records of sewer surveys. *Mach. Vis. and Appl.*, 11:53–63, 1998.

[4] G. Csurka, C. Zeller, Z. Zhang, and O. Faugeras. Characterizing the uncertainty of the fundamental matrix. *Comput. Vis. Image Underst.*, 68(1):18–36, 1997.

[5] A. W. Fitzgibbon and A. Zisserman. Automatic 3D model acquisition and generation of new images from video sequences. In *Proc. European Signal Processing Conference*, pages 1261–1269, 1998.

[6] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. ECCV*, pages 311–326, 1998.

[7] R. M. Gooch, T. A. Clarke, and T. J. Ellis. A semi-autonomous sewer surveillance and inspection vehicle. In *Proc. IEEE Intelligent Vehicles*, pages 64–69, 1996.

[8] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2000.

[9] J. Kannala. Measuring the shape of sewer pipes from video. Master's thesis, Helsinki University of Technology, 2004.

[10] J. Kannala and S. Brandt. A generic camera calibration method for fish-eye lenses. In *Proc. ICPR*, 2004.

[11] H.-B. Kuntze and H. Haffner. Experiences with the development of a robot for smart multisensoric pipe inspection. In *Proc. IEEE Robotics and Automation*, pages 1773–1778, 1998.

[12] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4):376–380, 1991.

[13] K. Xu, A. R. Luxmoore, and T. Davies. Sewer pipe deformation assessment by image analysis of video surveys. *Pattern Recognit.*, 31(2):169–180, 1998.