ISPRS Journal of Photogrammetry and Remote Sensing 64 (2009) 529–540

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Monocular 3D scene reconstruction at absolute scale

Christian Wöhler [a,*], Pablo d'Angelo [a], Lars Krüger [a], Annika Kuhl [b], Horst-Michael Groß [c]

[a] Daimler Group Research, Environment Perception, P. O. Box 2360, D-89013 Ulm, Germany

[b] Curtin University of Technology, Department of Computing, GPO Box U1987, Perth, Western Australia 6845, Australia

[c] Technische Universität Ilmenau, Faculty of Computer Science and Automation, P. O. Box 100565, D-98684 Ilmenau, Germany

ARTICLE INFO

Article history:
Received 27 July 2006
Received in revised form
23 March 2009
Accepted 23 March 2009
Available online 28 April 2009

Keywords:
Close-range photogrammetry
Reconstruction
Metrology
Bundle adjustment
Industry

ABSTRACT

In this article we propose a method for combining geometric and real-aperture methods for monocular three-dimensional (3D) reconstruction of static scenes at absolute scale. Our algorithm relies on a sequence of images of the object acquired by a monocular camera of fixed focal setting from different viewpoints. Object features are tracked over a range of distances from the camera with a small depth of field, leading to a varying degree of defocus for each feature. Information on absolute depth is obtained based on a Depth-from-Defocus approach. The parameters of the point spread functions estimated by Depth-from-Defocus are used as a regularisation term for Structure-from-Motion. The reprojection error obtained from bundle adjustment and the absolute depth error obtained from Depth-from-Defocus are simultaneously minimised for all tracked object features. The proposed method yields absolutely scaled 3D coordinates of the scene points without any prior knowledge about scene structure and camera motion. We describe the implementation of the proposed method both as an offline and as an online algorithm. Evaluating the algorithm on real-world data, we demonstrate that it yields typical relative scale errors of a few per cent. We examine the influence of random effects, i.e. the noise of the pixel grey values, and systematic effects, caused by thermal expansion of the optical system or by inclusion of strongly blurred images, on the accuracy of the 3D reconstruction result. Possible applications of our approach are in the field of industrial quality inspection; in particular, it is preferable to stereo cameras in industrial vision systems with space limitations or where strong vibrations occur.

1. Introduction

The knowledge of three-dimensional (3D) structure plays an important role in many fields such as navigation, mapping, obstacle avoidance, and object detection. Depth-from-Stereo (Scharstein et al., 2002) was one of the first methods for recovering depth information as it is inspired by human vision. The known geometry of the cameras is used to triangulate the spatial position of corresponding points from two images that are acquired from different viewpoints. The disadvantage of classical stereo vision systems is their need for a pair of precisely calibrated cameras, making it complex and costly for many applications. Therefore spatial scene reconstruction using monocular camera systems is often preferable. Structure-from-Motion (SfM) is such an alternative: From corresponding points in at least two images acquired sequentially at different camera positions the spatial positions of the points are recovered. The problem is that the scene can be reconstructed only up to a scaling factor as long as the camera positions are unknown.

Methods to establish point correspondences from different images require the detection and assignment of salient object features. Harris and Stephens (1988) propose image features that serve well for tracking algorithms. Widely used methods are SIFT features (Lowe, 2004), involving the extraction of scale-invariant features using a staged filtering approach, or the Kanade–Lucas–Tomasi (KLT) feature detector (Shi and Tomasi, 1994), which is based on the Harris corner detector and takes into account affine deformation.

A different approach to scene reconstruction utilises position-variant appearance, e.g. Depth-from-Defocus (Chaudhuri and Rajagopalan, 1999) and Depth-from-Focus (Subbarao, 1989; Ens and Lawrence, 1993; Subbarao and Choi, 1995). Depth-from-Focus uses images taken by a single camera at different focus settings to compute depth. The focus settings for the image depicting a point with minimal blurring are used to compute the absolute depth (Grossmann, 1987), relying on an appropriate calibration procedure. Depth-from-Defocus (DfD) methods rely on the fact that a real lens blurs the observed scene before the imaging device records it. The amount of blurring depends on the actual lens, but also on the distance of the observed object to the lens. Pentland (1982) uses this property to estimate depth simultaneously for

* Corresponding author. Tel.: +49 731 505 2148; fax: +49 731 505 4105.
E-mail address: christian.woehler@daimler.com (C. Wöhler).

0924-2716/$ – see front matter © 2009 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2009.03.004

all scene points from only one or two images. Depth information is extracted from a single image showing sharp discontinuities of intensity by Pentland (1987). A survey of existing methods is given by Chaudhuri and Rajagopalan (1999). Watanabe et al. (1995) propose a method that computes the DfD in real time using structured lighting.

So far, no attempt has been made to combine the precise relative scene reconstruction of SfM with the absolute depth data of DfD. A work related to this paper was published by Myles and da Vitoria Lobo (1998), where a method to recover affine motion and defocus simultaneously is proposed. However, the spatial extent of the scene is not reconstructed by their method, since it requires planar objects.

The main contribution of this article consists of a novel combination of SfM (a geometric method) with DfD (a real-aperture method). We will show that the combination of these methods yields a 3D scene reconstruction of high absolute metric accuracy based on an image sequence acquired with a monocular camera.

2. SfM and DfD

SfM recovers the spatial scene structure using a monocular camera. An initial step of SfM is the geometric calibration of the camera in terms of estimating the internal parameters, i.e. principal point, principal distance, pixel size, pixel skew, and lens distortion parameters (McGlone et al., 2004). An accurate value of the principal distance is required in Section 3 for calibration of the DfD method. Specifically, we use the semi-automatic approach for calibration rig detection proposed by Krüger et al. (2004). Subsequently, salient feature points are extracted and tracked across the sequence. The motion of these features relative to the camera is then used in a bundle adjustment (Triggs et al., 2000; McGlone et al., 2004) minimising the error term

$$E_{\mathrm{SfM}}\left(\{T_j\},\{X_i\}\right)=\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\mathcal{P}\left(T_jX_i\right)-x_{ij}\right]^2 \quad (1)$$

with respect to the M camera transforms T_j and the N scene points X_i. Here, x_{ij} denotes the 2D pixel coordinates of feature i in image j. The function $\mathcal{P}$ denotes the projection of 3D scene points to image coordinates and T_j the transform of the camera coordinate system of image j with respect to an arbitrary world coordinate system.

DfD directly recovers the spatial scene structure using a monocular camera. The depth D of the tracked feature points is calculated by measuring the amount of defocus, expressed for example by the standard deviation σ of the Gaussian-shaped point spread function (PSF) that blurs the image. An exact description of the PSF due to diffraction of light at a circular aperture is given by the radially symmetric Airy pattern $A(r)\propto[J_1(r)/r]^2$, where $J_1(r)$ is a Bessel function of the first kind (Pedrotti, 1993). For practical purposes, however, particularly when a variety of additional lens-specific influencing quantities (e.g. chromatic aberration) is involved, the Gaussian function is a reasonable approximation to the PSF (Chaudhuri and Rajagopalan, 1999). In the following, σ will be referred to as the "radius" of the PSF.

Measuring σ is the most important part of the depth estimation by DfD. The classical DfD approach uses two images of the same object taken at two different focal settings (Chaudhuri and Rajagopalan, 1999). Pentland (1987) shows that a priori information about the image intensity distribution, e.g. the presence of sharp discontinuities (edges), allows the computation of the PSF radius σ based on a single image. This is achieved by estimating the value of σ that generates the observed intensity distribution from the known ideal intensity distribution.

Since in our scenario such a priori information is not available, we suggest the empirical determination of a so-called Depth-Defocus Function. We assume that local features in the scene are tracked across a sequence of images and that for each feature the image is determined in which the feature appears best focused. Based on a calibration procedure, the radius σ of the Gaussian PSF which transforms the best-focused version of the feature into the currently observed pattern is determined as a function of depth D.

3. Spatial scene reconstruction by combining SfM and DfD

3.1. The Depth-Defocus Function

The Depth-Defocus Function $\mathcal{S}(D)=\sigma$ expresses the radius σ of the Gaussian PSF as a function of depth D, i.e. the distance between the object and the lens plane. It is based upon the lens law (Pedrotti, 1993):

$$\frac{1}{v}+\frac{1}{D}=\frac{1}{f}. \quad (2)$$

An object at distance D is focused if the principal distance is v, with f denoting the focal length of the lens. Varying the principal distance by a small amount Δv causes the object to be defocused as the light rays intersect before or behind the image plane. In the geometric optics approximation, a point in the scene is transformed into a so-called circle of confusion of radius $c=|\Delta v|/(2\kappa)$ in the image plane, where κ is the f-stop number expressing the focal length in terms of the aperture diameter. Empirically, we found that for small $|\Delta v|$ the resulting amount F of defocus can be modelled by a zero-mean Gaussian, which is symmetric in Δv:

$$F(\Delta v)=\frac{1}{\phi_1}\mathrm{e}^{-\frac{1}{\phi_2}\Delta v^2}+\phi_3. \quad (3)$$

The parameters ϕ_1 and ϕ_2 are normalising constants and ϕ_3 denotes the defocus level for very strongly blurred images. Setting $F=\sigma$ leads to the so-called Depth-Defocus Function as described below. To determine the best-focused version of a tracked feature, F is represented by other measures such as the grey value variance or the high-frequency integral of the amplitude spectrum of the image or part of it (cf. Sections 3.2 and 3.3). The radius c of the circle of confusion and the PSF radius σ are related to each other in that σ is a monotonically increasing nonlinear function of c. Hence, the symmetric behaviour of $c(\Delta v)$ apparent from Fig. 1(a) implies a symmetric behaviour of $\sigma(\Delta v)$. Depending on the constructional properties of lenses different from those we used in our experiments, analytic forms different from Eq. (3) but also symmetric in Δv may better match the observed behaviour of the PSF.

However, the Depth-Defocus Function expresses the relation between the depth of an object and its defocus. That is, the image plane is assumed to be fixed while the distance D of the object varies by the amount ΔD, such that $\Delta D=0$ refers to an object that is well focused. The dependence of c on ΔD is asymmetric, as shown in Fig. 1(b). Since neither D nor ΔD is known, the functional relation needs to be modelled with respect to Δv:

$$\frac{1}{v+\Delta v}+\frac{1}{D}=\frac{1}{f}. \quad (4)$$

A value of $\Delta v\neq 0$ refers to a defocused object point. Solving Eq. (4) for Δv and inserting Δv in Eq. (3) yields the Depth-Defocus Function

$$\mathcal{S}(D)=\frac{1}{\phi_1}\mathrm{e}^{-\frac{1}{\phi_2}\left(\frac{fD}{D-f}-v\right)^2}+\phi_3. \quad (5)$$