Computer Vision and Image Understanding 134 (2015) 105–115

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Cross-calibration of time-of-flight and colour cameras

Miles Hansard [a,b], Georgios Evangelidis [a], Quentin Pelorson [a], Radu Horaud [a,*]

[a] INRIA Grenoble Rhône-Alpes, 38330 Montbonnot Saint-Martin, France
[b] School of Electronic Engineering and Computer Science, Queen Mary, University of London, Mile End Road, United Kingdom

ARTICLE INFO

Article history:
Received 20 January 2014
Accepted 1 September 2014
Available online 5 November 2014

Keywords:
Camera networks
Time-of-flight cameras
Depth cameras
Camera calibration
3D reconstruction
RGB-D data

ABSTRACT

Time-of-flight cameras provide depth information, which is complementary to the photometric appearance of the scene in ordinary images. It is desirable to merge the depth and colour information, in order to obtain a coherent scene representation. However, the individual cameras will have different viewpoints, resolutions and fields of view, which means that they must be mutually calibrated. This paper presents a geometric framework for the resulting multi-view and multi-modal calibration problem. It is shown that three-dimensional projective transformations can be used to align depth and parallax-based representations of the scene, with or without Euclidean reconstruction. A new evaluation procedure is also developed; this allows the reprojection error to be decomposed into calibration and sensor-dependent components. The complete approach is demonstrated on a network of three time-of-flight and six colour cameras. The applications of such a system, to a range of automatic scene-interpretation problems, are discussed.

1. Introduction

The segmentation of multi-view video data, with respect to physically distinct objects of interest, is an essential task in automatic scene-interpretation. Visual segmentation can be based on colour, texture, parallax and motion information (e.g. [1,2]). The task remains very difficult, however, owing to the combined effects of non-rigid surfaces, variable lighting, and occlusion. It has become clear that depth cameras can make an important contribution to scene understanding, by enabling direct *depth segmentation*, based on the measured scene-structure (see CVIU special issue [3]). This approach is also highly effective for dynamic tasks, such as body tracking and action recognition [4]. Furthermore, if depth and colour information can be merged into a single representation, then a complete 3D representation is possible, in principle. This is clearly desirable, because colour and texture data are essential to many other aspects of scene-understanding, such as identification and tracking [5].

There are two major obstacles to the construction of a complete scene representation, from a multi-modal camera network. Firstly, typical depth sensors are unable to capture RGB data [6]. This means that the depth and colour cameras will have different viewpoints, and so the raw data are *inconsistent*. Secondly, typical TOF and RGB cameras have limited fields of view, and so the depth and colour data are *incomplete*. This paper addresses both of these problems, by showing how to estimate the geometric relationships in a multi-view, multi-modal camera network. This task will be called *cross-calibration*.

In order to constrain the problem, two practical constraints are imposed from the outset. Firstly, the system will be based on *time-of-flight* (TOF) cameras, in conjunction with ordinary RGB cameras. The TOF cameras are compact, can be properly synchronized, and are industrially specified, e.g., [6]. Secondly, a *modular* network of TOF + RGB units is required. This is so that individual units can be added or removed, in order to optimize the scene-coverage.

1.1. Overview

Time-of-flight cameras can, in principle, be geometrically calibrated by standard methods [7]. This means that each pixel records an estimate of the scene-distance (range) along the corresponding ray. The 3D structure of a scene can also be reconstructed from two or more ordinary images, via the *parallax* (e.g. binocular disparity) between corresponding image points. There are many advantages to be gained by combining the range and parallax data. Most obviously, each point in a parallax-based reconstruction can be mapped back into the original images, from which colour and texture can be obtained. Parallax-based reconstructions are, however, difficult to obtain, owing to the difficulty of putting the image points into correspondence. Indeed, it may be impossible to find any correspondences in untextured regions. Furthermore, if a

* Corresponding author.
E-mail address: Radu.Horaud@inria.fr (R. Horaud).

http://dx.doi.org/10.1016/j.cviu.2014.09.001
1077-3142/

Euclidean reconstruction is required, then the cameras must be calibrated. The accuracy of the resulting reconstruction will also tend to decrease with the distance of the scene from the cameras [8].

The range data, on the other hand, are often corrupted by noise and surface-scattering. The spatial resolution of current TOF sensors is relatively low, the depth-range is limited, and the luminance signal may be unusable for rendering and for classical image processing. It should also be recalled that TOF cameras, of the type used here, cannot be used in outdoor lighting conditions. These considerations lead to the idea of a *mixed* colour and time-of-flight system, as described in [9]. Such a system could, in principle, be used to make high-resolution Euclidean reconstructions, including photometric information [10,11].

In order to make full use of a mixed range/parallax system, it is necessary to find the exact geometric relationship between the different devices. In particular, the reprojection of the TOF data, into the colour images, must be obtained. This paper is concerned with the estimation of these geometric relationships. Specifically, the aim is to align the range and parallax reconstructions, by a suitable 3D transformation.

1.2. Previous work

Multi-view depth and colour camera-networks, of the kind used here, produce data-streams that are subject to a variety of geometric relationships [12]. These relationships depend on the calibration state, relative orientation, and fields of view of the different cameras. It follows that a variety of calibration strategies can be adopted. These are discussed below, with reference to the literature, and contrasted with the approach presented here.

Perhaps the simplest way to combine RGB and TOF data is to perform an essentially 2D registration between the images and depth maps, as reviewed in [13]; see also [14–16]. This 2D approach, however, can only provide an instantaneous solution, because changes in the scene-structure produce corresponding changes in the image-to-image mapping. Moreover, owing to the different viewpoints, a complete registration will usually be impossible. If the depth camera also produces a reliable intensity image, then photo-consistency can be used as a 3D calibration criterion. For example, Beder et al. [17] (see also [18,19]) reproject the *intensities* of the depth data into the colour images, and optimize the camera parameters with respect to the photo-consistency.

Zhu et al. [20] (see also [21]) present a sensor-fusion framework for the integration of TOF depth and binocular disparity information. This method assumes that a dense disparity map is being computed on-line, which is not required by the method presented in this paper. Furthermore, the geometric calibration method [20] requires manual identification of corresponding points, and is based on a weak perspective camera model. In contrast, our method is automatic, and is based on the more appropriate perspective camera model. However, [20] is complementary to our method, in the sense that their sensor-fusion framework (along with a dense stereo-matcher) could be combined with the projective calibration method described below. Wang and Jia [22] describe a related sensor-fusion framework for Kinect (rather than TOF) depth data and colour images.

Another approach to the multi-modal calibration problem is to apply standard methods, as far as possible, to the depth cameras. Wu et al. [23] describe an example of this approach. Lindner et al. [9] analyze the applicability of standard methods to TOF cameras, as well as characterizing the accuracy of the depth data. Mure-Dubois and Hügli [24] describe the Euclidean alignment of multiple TOF point clouds, having calibrated the cameras by standard methods.

Silva et al. [25] describe a cross-calibration methodology that is based on the identification of 3D lines in the TOF data, which are then projected to corresponding 2D lines in the RGB images. This method does not require a chequerboard or other calibration pattern; it does, however, require the existence and detection of straight depth-edges throughout the scene. The approach of Silva et al. also involves a non-trivial correspondence problem, which in turn influences the calibration accuracy. Our method uses a standard chequerboard pattern, with a known number of vertices, for which the correspondence problem is relatively straightforward.

Zhang and Zhang [26] present cross-calibration methodology that is based on plane constraints, as given in [27]. This has the advantage of not requiring 2D features to be detected in the (low resolution) TOF images. However, this method cannot address the crucial issue of lens distortion, which is considerable in typical TOF cameras [28]. A related Kinect-based calibration system is described by Herrera et al. [29], again using the plane-based method of [27]. Herrera et al. give a careful analysis of the Kinect intrinsic parameters, including lens and depth distortion. The latter is analyzed in more detail by Teichman et al. [30]. Our method does require features to be detected in the TOF images, but this also makes it straightforward to estimate the lens parameters, using standard techniques.

Mikhelson et al. [31] describe an automatic method for registering a Euclidean point-cloud (obtained from a Kinect device) to its 2D image-projections. This method, like ours, is based on a chequerboard target. However, Mikhelson et al. perform Euclidean 2D/3D registration, in contrast to the more general projective 3D/3D registration that is described below. Finally, there are methods that perform 3D/3D registration of dense data, subject to point-wise adjustments [32]. This strategy can achieve very close registrations, but introduces a more complex optimization problem, which is not fully compatible with the standard calibration pipeline.

1.3. Paper organization and contributions

This paper is organized as follows. Section 2.1 briefly reviews some standard material on projective reconstruction, while Section 2.2 describes the representation of range data in the present work. The chief contributions of the subsequent sections are as follows: Section 2.3 describes a point-based method that maps a classical multi-view reconstruction (projective or Euclidean) of the scene onto the corresponding TOF representation. The data are obtained from a TOF + RGB system, as shown in Fig. 1. This does not require the colour cameras to be calibrated (although it is necessary to correct for lens distortion). It is established that this model includes a projective-linear approximation of the systematic TOF depth-error.

Section 3 addresses the problem of multi-system alignment, which is necessary for complete scene-coverage. It is shown that this can be achieved in a way that is compatible with the individual TOF + 2RGB calibrations. The complete cross-calibration pipeline, given a collection of chequerboard images, is fully automatic.

Section 4 contains a detailed evaluation of these methods, using several large data-sets, captured by three TOF + 2RGB systems (i.e. a nine-camera network). In particular, Section 4.1 extends the usual concepts of reprojection error [12] to the multi-modal case. Section 4.2 then introduces a new metric for mixed TOF/RGB systems, which measures instantaneous sensor noise, as well as calibration error. The appropriateness of the 3D homography transformation, as opposed to a similarity transformation, is tested in Section 4.3. Section 4.4 discusses possible applications of these systems, including some real 3D reconstruction examples. Conclusions and future directions are discussed in Section 5.

The system presented here is based on the approach introduced by Hansard et al. [33,34]. The earlier work has been improved, and extended to the case of multiple TOF and colour cameras. In