Arab J Sci Eng (2017) 42:3451–3463
DOI 10.1007/s13369-017-2541-z

RESEARCH ARTICLE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

Mutual Information-Based Tracking for Multiple Cameras and Multiple Planes

Zhuoman Wen[1,2,3] · **Arjan Kuijper**[4,5] · **Matthieu Fraissinet-Tachet**[4] · **Yanjie Wang**[2] · **Jun Luo**[6]

Received: 19 December 2016 / Accepted: 10 April 2017 / Published online: 25 April 2017

Abstract Based on mutual information (MI), this paper proposes a systematic analysis of tracking a multi-plane object with multiple cameras. Firstly, a geometric model consisting of a piecewise planar object and multiple cameras is setup. Given an initial pose guess, the method seeks a pose update that maximizes the global MI of all the pairs of reference image and camera image. An object pose-dependent warp is proposed to ensure computation precision. Six variations of the proposed method are designed and tested. Mode 1, i.e., computing the 2nd-order Hessian of MI at each step as the object pose changes, leads to the highest convergence rates; Mode 2, i.e., computing the 1st-order Hessian of MI once at the beginning, occupies the least time (0.5–1.0 s). For objects with simple-textured planes, applying Gaussian blur first and then use Mode 1 shall generate the highest convergence rate.

Keywords Computer vision · Multi-camera tracking · Image registration · Mutual information · Nonlinear optimization

✉ Arjan Kuijper
arjan.kuijper@igd.fraunhofer.de

1 Southwest China Research Institute of Electronic Equipment, China Electronics Technology Group Corporation, Chengdu 610036, China

2 Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

3 University of the Chinese Academy of Sciences, Beijing 100049, China

4 Fraunhofer IGD, 64283 Darmstadt, Germany

5 Technische Universität Darmstadt, 64283 Darmstadt, Germany

6 Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China

1 Introduction

Single camera tracking methods are widely used in many applications. However, sometimes the object cannot be tracked by only one camera due to occlusion, bad imaging condition, and camera's limited field of view. Binocular cameras systems [1] are obtaining increasingly more attentions, and multi-camera tracking [2,3] is becoming a trend. With the development of the digital processors, organizing multiple cameras [4,5] in one system is a wise choice for object tracking that requires high precision.

The tracking of a 3D object that can be described as a set of 2D planes is a problem of warping the planes through a set of homographes depending on the pose of the object. The properties of homographes ensure that the warped image is computable without the need to render the model anew [6]. To track a piecewise planar object with multiple cameras requires computing all the warps between each plane of the object and each camera. Tracking methods which ensure high precision imply large amounts of calculation, making it impractical to apply such methods to real-time applications. Computation complexity is a huge challenge.

A visual tracking [7] methodology generally matches a large set of distinctive elements from a model surface to the current image. These elements can be local features, such as feature points, lines and small geometrical shapes, or global features, like area of the object surface.

SIFT [8] (scale invariant features transform) and SURF [9] (speeded up robust features) are two of the popular local feature detectors. However, the possible presence of outliers (false identifications) can lead to a less stable and precise pose estimation in these methods. Robust statistics methods like RANSAC [10] (Random sample consensus) and M-estimators [11] can remove these outliers, whereas a stable result is guaranteed only when the reliable features subset

is sufficiently large. These methods are also challenged when the object surface shows a significant curvature or the surface has an overall particular texture pattern, but a few distinctive local key points.

Template-based registration finds the pose parameters that optimizes of a function that describes the similarity or difference between a template and the current image. The most simple registration function is SSD [12] (sum of squared differences), whereas it is very sensible to scene variations. More complex algorithms include SCV [13] (sum of conditional variance) and NCC [14] (normalized cross-correlation). They perform better than SSD for global illumination variations, but are not robust against occlusion. MI [15–17] (mutual information) as the registration function is quite insensitive to changes in lighting condition and to partial occlusions, but are complex to implement [18]. MI measures the amount of information that one image contains of the other. The MI value of the corresponding images is maximal if the images are correctly geometrically aligned.

Reported work on MI focused on image registration [19], image quality index [20], motion segmentation [21], and monocular camera tracking [22,23]. However, Fraissinet-Tachet [24] proposed a multi-camera and multi-plane object tracking method using MI. In this method, the MI derivatives with respect to the pose parameters do not depend on the relative pose between the planes and the cameras; therefore, its convergence rate is low and drops severely as the angle between planes increases.

This paper proposes schemes to increase convergence rate by implementing an object pose-dependent warp. We first present a geometrical model which is consisted of multiple cameras and a multi-plane object. A pose-dependent warping update method is then proposed. Using MI, we illustrate an incremental method for finding the object pose with respect to each of the cameras. The complexity of the algorithm is reduced by simultaneously computing all the warps between the planes and cameras. Six variations of the proposed algorithm are tested with both intricate-textured images and simple-textured images.

Note that our previous paper published in JMIV has theoretically proven the appropriate form for the second derivative of MI [25]. This paper thoroughly explains the calculation process of second derivative of MI; proposes an approach to largely increase the convergence radius for objects with simple-textured planes, that is to apply Gaussian blur on images first, and then use Mode 1, i.e., computing 2nd-order Hessian of MI at each step as the object pose updates.

The structure of this paper is as follows. Section 2 describes the geometric model of the entire system. Section 3 explains the proposed algorithm in detail. Section 4 illustrates the architecture of the proposed algorithm and the six different variations of the method. Section 5 analyzes the performance of the proposed algorithm from different perspectives, and Sect. 6 concludes the paper.

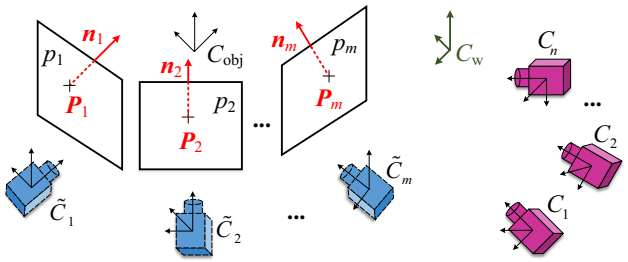

Fig. 1 Geometric model of the system

2 Geometric Model

The geometric model of our system is illustrated in Fig. 1. It consists of one piecewise planar object C_{obj}, m virtual local cameras $\tilde{C}_1$ to $\tilde{C}_m$, and n practical cameras C_1 to C_n ($m, n \in \mathbf{N}$ and $m, n \geq 1$).

Object C_{obj} includes m planes p_i, each defined by a normal vector $\boldsymbol{n}_i$ and a point on the plane $\boldsymbol{P}_i (i \in \{1, 2, \ldots, m\})$. Each plane p_i has a virtual local camera $\tilde{C}_i$ looking at it from the negative space defined by the plane. The relative position and orientation (pose) between the plane and its associated local camera is fixed. Therefore, the captured image of each local camera always remains the same, independent from the displacement of the object. All of the local cameras share the same intrinsic parameters (including focal length, principal point, and pixel size), and their relative pose to the associated planes is identical. With appropriate parameters, the images taken by the local cameras will be exactly the planes of the object. Note that these local cameras do not exist physically; they are modeled to facilitate the understanding and C^{++} implementation of the system. The set of cameras, C_1 to C_n, do exist physically. They are fixed with respect to the world coordinate C_W, and their intrinsic and extrinsic parameters are known.

The system includes m local cameras and n cameras, and the relative pose between each pair of local camera and camera has to be calculated. In total, there are $m \cdot n$ pairs to be considered.

3 Proposed Algorithm

3.1 Image Warping

Take the pair of local camera $\tilde{C}_m$ and camera C_n for instance, we denote $I^*_{\tilde{C}_m}$ as the image captured by local camera $\tilde{C}_m$, i.e., the reference image, and I_{C_n} the image seen by camera C_n. Given a pose T, there exists a warp ω_T [see Eq. (1)] that