Information Fusion 33 (2017) 15–28

Contents lists available at ScienceDirect

Information Fusion

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

A novel system for object pose estimation using fused vision and inertial data

Juan Li*, Juan A. Besada, Ana M. Bernardos, Paula Tarrío, José R. Casar

ETSI Telecomunicación, Universidad Politécnica de Madrid, Avenida Complutense 30, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 14 May 2015
Revised 22 April 2016
Accepted 24 April 2016
Available online 27 April 2016

Keywords:
Pose estimation
Error propagation
Augmented reality
Sensor fusion
Visual sensors
Inertial systems

ABSTRACT

Six-degree-of-freedom (6-DoF) pose estimation is of fundamental importance to many applications, such as robotics, indoor tracking and Augmented Reality. Although a number of pose estimation solutions have been proposed, it remains a critical challenge to provide a low-cost, real-time, accurate and easy-to-deploy solution. Addressing this issue, this paper describes a multisensor system for accurate pose estimation that relies on low-cost technologies, in particular on a combination of webcams, inertial sensors and a printable colored fiducial. With the aid of inertial sensors, the system can estimate full pose both with monocular and stereo vision. The system error propagation is analyzed and validated by simulations and experimental tests. Our error analysis and experimental data demonstrate that the proposed system has great potential in practical applications, as it achieves high accuracy (in the order of centimeters for the position estimation and few degrees for the orientation estimation) using the mentioned low-cost sensors, while satisfying tight real-time requirements.

1. Introduction

The term ‘pose’ is usually employed to refer to the combined information on position and orientation of a moving target (i.e., an object or a human). Position is represented by the three-dimensional location of the object, while orientation may be expressed as a set of consecutive rotations. Determining the pose of a target in 3D space is an important task in many traditional application fields, such as robotics [1–3] (e.g., for robot guidance, object manipulation, etc.), indoor tracking and activity estimation, or interaction [4].

In particular, in recent years, an attractive application area requiring accurate pose estimation is indoor Augmented Reality (AR). AR has been widely explored in training, entertainment, education and tourism to facilitate a novel way for the users to interact with their surroundings [4–9]. Ideally, an AR system should be able to overlay the virtual information upon the real world with no error and no latency, thus it needs a perfectly estimated pose of the target relative to the real world. Despite the progress that has been made to date, current technologies for indoor deployments are not able to achieve these performance goals. Better said, they still offer limited performance in terms of accuracy, computational cost, usability, robustness, on-board power consumption and easiness of deployment. In this context, this paper describes a multisensor solution for accurate pose estimation using low-cost technologies. The designed system provides pose estimation in real time and may be easily adapted to different environments.

The enabling apparatus is simple: (a) one or more infrastructure vision sensors (commercial off-the-shelf cameras), which are fixed and calibrated beforehand, (b) a three-axis accelerometer in the object to be tracked (e.g., embedded accelerometers in mobile devices), (c) a printable colored marker to be stuck on the object and (d) a server. The pose calculation process is implemented on the server side, leaving computing power of the client side for applications. The proposed fiducial has a linear thin stripe-like geometry; it is thus different when compared against the conventional square fiducials that are used in ARToolKit [10] or ARTag [11]. Linear fiducials may better adapt to final services, because they are less invasive to the environment than the square fiducials, due to their smaller dimensions. The thinness also allows them to be attached to a small surface, for example, borders of mobile devices, hats or eyeglasses frames. Therefore, the proposed solution has the potential for indoor person tracking, robot tracking, mobile AR and interaction in smart spaces. With respect to pure vision-based approaches, the fusion of vision data with accelerometer measurements reduces the number of unknown pose parameters; therefore, robustness and computational efficiency are enhanced. Moreover, gravitational acceleration measurements are used to aid in the pose estimation, but no acceleration integration process is

* Corresponding author. Fax: +34 91 3365876.
E-mail address: li.juan@grpss.ssr.upm.es (J. Li).

http://dx.doi.org/10.1016/j.inffus.2016.04.006
1566-2535/

performed. Therefore, our proposal generates zero-drift solutions and eliminates the requirement of having an initial state.

Within a bounded space, the system can work with a single active camera (monocular approach), or with two cameras (stereo vision approach), with the latter resulting in increased accuracy. To equip a room-like space with our pose estimation technology, more than two cameras may be needed to cover the whole space. The issues related to multi-camera management, such as object tracking and camera selection, will not be studied in this work. In our previous work [12], we proposed a six degree of freedom pose estimation system that fuses acceleration data and stereo vision. The system was evaluated by comparing to real measurements and a state-of-the-art marker-based system. Experimental results showed that the proposed stereo vision system provides high accuracy. This article introduces a new strategy to estimate 6-DoF pose by fusing data from the target object's accelerometer with input from one camera. Each component of the system is analyzed thoroughly. Besides, a complete pose estimation analytical error model for both the monocular and the stereo vision system is derived and validated by real tests. This paper provides more extensive experimental results and a thorough comparison to the state of the art to evaluate the system qualitatively and quantitatively. From our simulations and real tests of the system, it will be shown that the proposed pose estimation system has great potential in practical applications, as it achieves high accuracy (in the order of centimeters for the position estimation and few degrees for the orientation estimation) in real-time, using the mentioned low-cost sensors. Furthermore, the possible applications and the guidelines for the practical implementation of the system are addressed.

The rest of the article is organized as follows. Section 2 includes a review of previous work on object pose estimation systems. Section 3 states the mathematical formulation and explains the contributions of each sensing technology. Section 4 is dedicated to describe the pose estimation strategy, which fuses data from inertial and vision sensors. Section 5 models the errors of the system. Accuracy and computational load are assessed and the sensor error model is validated by experimental results in Section 6. The performance of the proposed approach applied to pointing applications is evaluated in Section 7. Finally, Section 8 concludes the paper and describes further lines of work.

2. Related work

Object pose estimation has been studied over the past several decades and a wide range of technologies have been explored [1–11,13–30]. Depending on the sensing technology, the available approaches may be classified into three main categories: sensor-based, vision-based and hybrid approaches. The existing literature on these categories is described below.

2.1. Sensor-based methods

Inertial sensors including accelerometers and gyroscopes have been widely used for robots [1], aircrafts and vehicles navigation [13]. The principle for determining position and orientation using these sensors is based on Newton's laws. Accelerometers measure the linear acceleration in the inertial reference frame, which is integrated to get the velocity and then integrated again to get the position. Gyroscopes measure the angular velocity and by integrating once, rotation angles can be calculated. Inertial Measurement Units (IMU) are maturely developed units for motion tracking which typically contain three orthogonal accelerometers and three orthogonal gyroscopes. They run at a high rate, therefore they are able to track fast and abrupt movements. Furthermore, they are not influenced by illumination and visual occlusion. On the downside, they suffer from a severe drift problem caused by accumulation of measurement errors, thus a periodic re-calibration is required. Several methods have been proposed to minimize the drift problem. For example, in [14], relative measurements were used instead of absolute measurements to reduce the drift error. It is worth mentioning that in inertial-based methods, the initial state is needed to calculate the absolute pose.

Magnetometers are used to get the heading angle by sensing the earth magnetic field [15]. In order to get a full pose estimation, they need to be combined with other technologies. The algorithm in [15] integrates inertial sensors with magnetometers and keeps the tracking results within about 2m of the true track throughout the entire in-building run. However, the measurements provided by magnetometers can be corrupted due to the presence of metallic objects in the surroundings, which is quite usual in indoor scenarios [16].

A different approach to positioning is the use of Radio-frequency (RF) technologies. They aim at locating moving objects (smartphones, robots, etc.) through diverse techniques(refer to e.g. [30] for a survey): WiFi, Bluetooth or ZigBee-based solutions usually rely on fingerprinting techniques (e.g. [31]) or channel modeling (e.g. [32]) to achieve a limited accuracy (3–4m in average). In addition, RF positioning systems do not generally support orientation estimation, therefore not providing a full pose estimation.

2.2. Vision-based methods

Visual sensing technologies try to interpret the environment through observations from cameras. Most of the available proposals estimate the spatial relationship between the camera and the object by finding the correspondence between 2D image points and 3D scene points. According to the tracked features, most of the methods can be grouped into marker-based, ready to decode a known external visual reference, and markerless methods, not needing any previously known symbol.

2.2.1. Marker-based methods

Marker-based methods recover the transformation between the fiducial (artificial) marker and the vision sensor by extracting the feature points previously defined in the marker. Several available libraries use planar fiducial for tracking, such as ARToolKit [10], ARTag [11], Studierstube [28], AprilTag [17] and OpenCV [29]. ARToolKit was developed in 1999 by Hirokazu Kato and has been widely used. Based on ARToolKit, ARTag was later developed to provide improved performance. The extended version of ARToolKit is ARToolKitPlus, which added more features over the ARToolKit. However, it is no longer developed and has a successor: Studierstube Tracker. It supports mobile phones as well as PCs and has low memory requirements. However, it is not open source. AprilTag has been recently developed for PCs and further improves accuracy and robustness. OpenCV is an open source cross-platform toolkit for image processing that supports PCs as well as mobile platforms. This library is still in development and has a large community of users. In [18], a chessboard pattern is tracked by OpenCV to implement mobile AR. Markers used in these libraries are black-white and have high contrast, so they are easily recognizable. On the downside, contrast-based detection is sensitive to lighting. Generally speaking, marker-based methods can provide high accuracy. However, the marker size and the distance as well as the viewing angle to the marker will affect the accuracy. These aforementioned markers need a big, flat surface to be placed. Therefore, they are unsuitable to be attached to a small object to be tracked, such as a mobile device. Instead, by using an on-body camera to track markers placed in known locations it is feasible to estimate the object's