Low-Cost Vision-Based Positioning System

Jingjing YAN: Jingjing.Yan@nottingham.edu.cn
Gengen HE: Gengen.He@nottingham.edu.cn
Craig HANCOCK: Craig.Hancock@nottingham.edu.cn

International Doctoral Innovation Centre, University of Nottingham Ningbo China, Ningbo 315100, China

Abstract. This study will introduce a prototype design of indoor positioning system with the combination of vision-based pedestrian tracking and digital map information in WGS 1984 and in future will combine with INS. A trial has been finished for visual tracking with integration of digitized floor plan, taking place in PMB Building in University of Nottingham. The result is promising for future development of positioning system.

Keywords. Indoor positioning, Pedestrian detection, Deep learning, Human tracking, Sensor fusion

1. Introduction

Indoor positioning plays an important role in indoor navigation process. Current popular solutions are Bluetooth, Wi-Fi, RFID and UWB (Gunduz et al. 2016), requiring installation of facilities or regular updating of databases (Basiri et al. 2017). With the consideration of cost and flexibility, INS-based Pedestrian Dead Reckoning (PDR) is a better answer as its sensors have already been built in smartphones (Harle 2013, Basiri, Lohan et al. 2017). However, this solution is still problematic as it accumulates errors during measurements and it only provides relative positioning, requiring additional positioning system for absolute locations (Harle 2013).

Visual tracking with the integration of georeferenced digital map can help to solve this problem and provide a relatively accurate absolute position to calibrate location information supplied by PDR. The methods for vision-based positioning can be divided as fixed and mobile camera systems (Torres-Solis et al. 2010, Mautz & Tilch 2011). Previous research figured out using image feature matching method for sensor fusion by attaching a camera to an Inertial Measurement Unit (IMU) on smartphone (Hide et al.
2010). However, this solution is still not perfect as the video taken by smartphone is energy consuming and cannot last for long.

The system in this project is mainly designed for public areas which have already installed surveillance cameras and it will utilize the existing indoor surveillance system to get relative positions after processing. Then it will combine with map information for absolute position. A pilot study is finished for vision-based tracking, showing promising result for future establishment of positioning system. In future, INS system will also be integrated into this system after being processed for step segmentation and then this information will be calibrated by previously acquired absolute position information. The following sections will first introduce overall design of positioning system, then concentrate on vision-based pedestrian tracking system by deep learning and map integration, and finally show achieved result in the pilot study.

2. Positioning System

2.1. Overall Design

The overall positioning system can be divided into four modules as pedestrian detection, depth information extraction, map information integration and combination of INS (Figure 1). This study mainly focused on the vision-based positioning with the support of map to provide absolute position information (in red dash-line box), which in future can be used as constraints to INS for drift calibration and external positioning system by application of particle filter based on same time step.

![Figure 1](http://doi.org/10.3929/ethz-b-000225587)

*Figure 1.* The overall design of positioning system (Visual tracking in red dash-line box).

2.2. Vision-Based Pedestrian Tracking

The conventional methods for optical pedestrian tracking have utilized background subtraction for foreground detection to identify people in im-
age sequences (e.g. (Tsai et al. 2016, Zhou et al. 2016)). However, these methods are limited to certain models of image matching (e.g. Gaussian Mixed Model) with lower accuracy. The state of the art methods for object detection are based on Region Proposal Network (RPN) and Region-Based Convolutional Neural Networks (R-CNNs), which is more flexible and robust to handle different datasets with higher accuracy (Ren et al. 2015). In this study, a Faster R-CNN algorithm will be applied for pedestrian detection in video, the RPN is used for Bounding Box (BB) prediction and objectness evaluation (classification) while a Fast R-CNN with a VGG-16 model is the detector which uses previous proposed regions (BBs) for object detection. The whole system is a single and unified network by feature sharing learned by Four-Step Alternating Training (Ren, He et al. 2015). The overall process of object detection is listed in Figure 2.

![Figure 2](image)

**Figure 2.** The architecture of Faster R-CNN network for object detection.

With this network, the accuracy of human detection is between 82.3% and 84.1% by using training data MS COCO and PASCAL VOC 2007 plus 2012 (Ren, He et al. 2015). This study will utilize pre-trained human detector from Faster R-CNN to extract BBs from acquired video. Then the central points of BBs' lower boundaries are extracted as feet positions of human for user path production (**Figure 3**). Although this study only has one single user in system trial, the proposed positioning system has the potential for multi-user tracking by a single camera.

![Figure 3](image)

**Figure 3.** Extracted user path from video data.
2.3. Depth Information
The depth information here is referred to distance to camera and this is achieved by a pinhole camera model (Dollar et al. 2012) as in (1):

\[
\frac{h_i}{f} = \frac{H_p}{d_i} (i = 1, 2, \ldots, n)
\]

Where \( h_i \) represents the pixel height of human in \( i^{th} \) frame extracted from video, \( H_p \) is the real height of person, \( d_i \) denotes the real distance of human to camera at \( i^{th} \) position and \( f \) is the focal pixel length. During practice, \( f \) is determined by pixel height of first frame as the initial distance to camera can be pre-determined according to map information by setting the starting point of human movement. With known height of participant and its pixel height extracted 1st frame, the focal pixel length of camera is then determined. The following \( d_i \) is proportional to \( h_i \) and a series of relative distances of human to camera are then estimated from the above information.

2.4. Map Integration
In this study, the georeferenced digitized floor plan is regarded as a prepared digital basemap to provide absolute positions for positioning and simple visual semantic representations (Figure 5). The coordinate selected for this study is WGS 1984 (with pre-settled 4th floor height at 9.5 meters), which is the common global coordinate for GPS localization and is beneficial to later combination with outdoor system. Then, the extracted user path in image coordinates from video data will be translated to the real geographical coordinates based on the marked points in the map as ground truth.

3. Experiments
In this pilot study, the video positioning system is tested in one corridor on the 4th floor of PMB building in university with a single camera installed at ceiling outside Room 416 and the result is in Figure 4. The average accuracy is below 1 meter. However, the two ending parts (from Room 401 to Room 403 and from Room 411 to 416) are problematic as the target is too far or too close to camera and the accuracy of these two parts are about 2 meters. In future, when coming into whole system with multiple cameras, if people cannot viewed by the camera, it will cause difficulties for human tracking and need support from inertial sensor data calibrated by digital map data. Meanwhile, the re-identification of people during movement in coverage area from one camera to another may also be supported by inertial sensor data and known camera positions to link the tracks together, which can identify the starting point of each people when...
entering the next camera by sharing same ending positioning recorded from the previous camera. It is also beneficial to track multiple people moving in the whole building area.

![User Path Extracted from Video Data](image)

**Figure 4.** The extracted user path in real coordinates visualized by positioning system (where 4F_door represents all doors at the 4th floor, the 4F_room represents all offices at 4th floor, 4F_apdt represents the non-office rooms and spare spaces at 4th floor and 4F_con represents the staircases and elevators which are regarded as connection parts in building).

4. Conclusion

This study has designed a low-cost positioning system for pedestrian positioning and tracking in indoor area with a satisfied performance during the trial. The future work will focus on sensor data fusion for multi-user detection in the entire building with the assistance of inertial sensor for invisible parts and track linking, and Faster R-CNN for multi-people detection, supported by digital basemaps with semantic information. This proposed system has the advantages of low cost by utilizing the existing indoor infrastructures and users’ smartphones, and can provide stable and reliable re-
results. In addition, it is also beneficial for the future development of a seamless outdoor-indoor positioning system by sharing same geographical coordinate system (WGS 1984).

Acknowledgement

The author acknowledges the financial support from the International Doctoral Innovation Centre, Ningbo Education Bureau, Ningbo Science and Technology Bureau, and the University of Nottingham. This work was also supported by the UK Engineering and Physical Sciences Research Council [grant number EP/L015463/1].

References


