People Detection and Tracking via Multiple 3D Range Sensor

Xiaowei SHAO[1], Huijing ZHAO[2], Ryosuke SHIBASAKI[1]
[1] Earth Observation Data Integration and Fusion Research Initiative, the University of Tokyo, Japan
435 General Research Building, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan
Tel: (81)- 4-7136-4307 Fax: (81)- 4-7136-4292
shaoxw@iis.u-tokyo.ac.jp, shiba@csis.u-tokyo.ac.jp
[2] Key Laboratory of Machine Perception, Peking University, China
zhaohj@cis.pku.edu.cn

ABSTRACT

People detection and tracking plays an important role in machine intelligence and has been intensively studied for many years. Nowadays traditional video camera based surveillance systems have a good performance when handling limited number of people, but the performance may decrease sharply in the case of crowded scenes where people are very close to each other and occlusions happen frequently, such as the congested crowd in front of a stair inside a railway station.

In this research, we propose a novel framework for detection and tracking of people in the surveillance area by using multiple Kinect sensors, which are set at the height of 3-5m and designed to scan the objects below it. Kinect is a new type of sensing input device released by Microsoft in 2010 and becomes very popular in recent years. It is small and light-weighted, and has very powerful 3D range sensing ability (640×480 pixels, 30 FPS). In this way, the 3D point data of people are collected from the top and the problem of occlusion can be solved to some extent. To enlarge the size of surveillance area, multiple Kinect sensors are employed.

First range image sequences are converted into real-world 3D coordinates and the calibration procedure is performed. Background points as well as moving foreground points are extracted and clustering technique is exploited to find possible candidates of people. However, due to the limited field view of Kinect, in many cases the observation of people is partial and incomplete. Conventional kernel based clustering method suffers from the variety of the 3D appearance of people and leads to many false positive targets. To address this issue, we improved the mean-shift algorithm by enhancing the weight of the local peak. The clustering is performed on x-y plane and the heighted information is utilized to generate corresponding weight. When clustering, the calculation is based on 2D grids and therefore it is very efficient. When the candidate targets are available, tracking is conducted by analyzing the spatial-temporal relation of targets.

When integrating the information from multiple sensors, localization error can be quite large due to sensing error, lens distortion, different viewpoints, partial measurement, synchronization error, and calibration error. Therefore, we didn't integrate the information at the raw data level as many researchers did, but instead we combine the tracking results from sensors by matching their temporal-spatial consistency.

An experiment was conducted inside a station where more than 30 people anticipated. Crowd were divided into several groups and moved according to some designed patterns. Total 6 Kinect sensors were utilized to cover the most congested area in the hall. Experimental results show the efficiency and effectiveness of our proposed method and the average tracking accuracy is 93.7%.

Keywords: People Detection; Tracking; 3D Range Scanner; Weighted Mean-shift Clustering; Sensor Integration