Pattern Recognition Letters 50 (2014) 149–158

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

3D people surveillance on range data sequences of a rotating Lidar ☆

Csaba Benedek *

Distributed Events Analysis Research Laboratory, Institute for Computer Science and Control, Hungarian Academy of Sciences, H-1111, Kende u. 13-17, Budapest, Hungary

ARTICLE INFO

Article history:
Available online 24 April 2014

Keywords:
Rotating multi-beam Lidar
MRF
Motion segmentation
Re-identification

ABSTRACT

In this paper, we propose an approach on real-time 3D people surveillance, with probabilistic foreground modeling, multiple person tracking and on-line re-identification. Our principal aim is to demonstrate the capabilities of a special range sensor, called rotating multi-beam (RMB) Lidar, as a future possible surveillance camera. We present methodological contributions in two key issues. First, we introduce a hybrid 2D–3D method for robust foreground–background classification of the recorded RMB-Lidar point clouds, with eliminating spurious effects resulted by quantification error of the discretized view angle, non-linear position corrections of sensor calibration, and background flickering, in particularly due to motion of vegetation. Second, we propose a real-time method for moving pedestrian detection and tracking in RMB-Lidar sequences of dense surveillance scenarios, with short- and long-term object assignment. We introduce a novel person re-identification algorithm based on solely the Lidar measurements, utilizing in parallel the range and the intensity channels of the sensor, which provide biometric features. Quantitative evaluation is performed on seven outdoor Lidar sequences containing various multi-target scenarios displaying challenging outdoor conditions with low point density and multiple occlusions.

1. Introduction

Moving people detection, localization and tracking are important issues in intelligent surveillance applications, such as person counting, activity recognition or abnormal event detection. However, these tasks are still challenging in crowded outdoor scenes due to uncontrolled illumination conditions, irrelevant background motion, and occlusions caused by various moving and static scene objects.

Vision algorithms in surveillance systems often follow a sequential approach [12], starting from low level classification of the observed environment, until object level and event level analysis of the scene. Foreground segmentation is a crucial initial step [3], since apart from highlighting the regions of interest, accurate object-silhouette masks can directly provide useful information for the scene interpretation modules, like biometric descriptors or various indicators of human behavior. Errors in the extracted foreground mask may also effect the consecutive person localization [21] and tracking [2] steps, especially in scenes with strong vegetation motion and occlusion. Model-based person tracking algorithms are widely used in the literature. An approach on 3D estimation of human pose from a monocular video was proposed by Brubaker et al. [6], which adopts a physics-based model. In [14], a model-based technique has been introduced to extract the silhouettes of moving people from stereo video sequences, and synthesizing realistic 3D person models. In both cases, however, a single person can be observed in each video frame, which condition is often not valid for outdoor surveillance scenes. Shu et al. [17] introduced a part-based human detector, which builds on person-specific SVM classifiers capturing the articulations of the human bodies in dynamically changing appearance and background. For such black-box models, an extensive training set selection is a crucial step.

Person re-identification is a fundamental task both for connecting the erroneously broken trajectories of the short term tacker module, and for identifying people who temporarily leave the Field of View (FoV) and re-appear later. Numerous methods in the literature address person re-identification in optical videos [1,7,15], however, their objectives are often notably different from the needs in our focused application. In the referred works, people identification is fulfilled within a large database (>100 people) using a ranking system, and the applied evaluation metric favors already, if the correct match is included within the first few candidates. This condition is acceptable if a manual verification follows the automated identification step (e.g., search in a police database), but in a fully automated surveillance system each person should be labeled with a single unambiguous identifier in real-time. On the other hand, we only deal with a few (6–8) pedestrians within a

☆ This paper has been recommended for acceptance by Dmitry Goldgof.

* Tel.: +36 12796106.
E-mail address: benedek.csaba@sztaki.mta.hu

http://dx.doi.org/10.1016/j.patrec.2014.04.010

scenario, which enables us to use weak biometric features for identification. Previously, Baltieri et al. [2] introduced a complete 3D video surveillance system implementing model based person tracking with re-identification based on multiple camera inputs, however it uses a computationally expensive Marked Point Process based approach for the localization, which currently does not enable real-time performance. Another practical problem is that multiple camera systems should usually be carefully fixed and calibrated beforehand, which makes quick temporary installation difficult for applications monitoring customized events.

Range image sequences offer significant advantages versus conventional video flows for scene analysis, since geometrical information is directly available [16], which can provide more reliable features than intensity, color or texture values [22,4]. Using Time-of-Light (ToF) cameras [16] or scanning Lidar sensors [8] enable recording range images independently of the illumination conditions and we can also avoid artifacts of stereo vision techniques. From the point of view of data analysis, ToF cameras record depth image sequences over a regular 2D pixel lattice, where established image processing approaches, such as Markov Random Fields (MRFs) can be adopted for smooth and observation consistent segmentation and recognition [4]. However, such cameras have a limited Field of View (FoV), which can be a drawback for surveillance and monitoring applications.

Rotating multi-beam Lidar systems (RMB-Lidar) provide a 360° FoV of the scene, with a vertical resolution equal to the number of the sensors, while the horizontal angle resolution depends on the speed of rotation (see Fig. 1). Each laser point of the output point cloud is associated with 3D spatial coordinates and a calibrated intensity value of the laser reflection which is related to the material and surface properties of the target point. For efficient data processing, the 3D RMB-Lidar points are often projected onto a cylinder shaped range image [8,9]. However, this mapping is usually ambiguous: On one hand, several laser beams with slight orientation differences are assigned to the same pixel, although they may return from different surfaces. As a consequence, a given pixel of the range image may represent different background objects at the consecutive time steps. This ambiguity can be moderately handled by applying multi-modal distributions in each pixel for the observed background-range values [8], but the errors quickly aggregate in case of dense background motion, which can be caused e.g., by moving vegetation. On the other hand, due to physical considerations, the raw data of distance, pitch and angle provided by the RMB-Lidar sensor must undergo a strongly non-linear calibration step to obtain the Euclidean point coordinates [13], therefore, the density of the points mapped to the regular lattice of the cylinder surface may be inhomogeneous. To avoid the above artifacts of background modeling, [9] has directly extracted the foreground objects from the range image by mean-shift segmentation and blob detection. However, we have experienced that if the scene has simultaneously several moving and static objects in a wide distance range, the moving pedestrians are often merged into the same blob with neighboring scene elements.

Instead of projecting the points to a range image, another way is to interpret the scene in the spatial 3D domain. MRF-like techniques based on 3D spatial point neighborhoods are frequently applied in remote sensing for point cloud classification [11], however the accuracy is low in case of small neighborhoods, otherwise the computational complexity rapidly increases. In [18,19] methods have been introduced for 3D pedestrian detection and tracking in point cloud streams of a mobile RMB-Lidar sensor, where the main challenge was to distinguish the pedestrians from other street objects within a large FoV with compensating the sensor motion. In this paper, we address significantly different scenarios: we use the RMB-Lidar sensor in a fixed position, and monitor a dense scene with several moving people in a compact outdoor environment, such as a courtyard or a small square. We expect high occlusion rate between the observed people due to crossing trajectories, and the considered pedestrians may leave the FoV and re-appear at any time during the inspection.

The main contributions of our method are twofold. *Firstly*, we introduce a hybrid 2D–3D approach (partially presented in [3]) for dense foreground–background segmentation of RMB-Lidar point cloud sequences obtained from a fixed sensor position. Our technique solves the computationally critical spatial filtering steps in the 2D range image domain by an MRF model, however, ambiguities of discretization are handled by joint consideration of true 3D positions and back projection of 2D labels. By developing a spatial foreground model, we significantly decrease the spurious effects of irrelevant background motion, which principally caused by moving tree crowns and bushes. For quantitative point level evaluation, we have developed a 3D point cloud Ground Truth (GT) annotation tool, and compared the detection results of the proposed model to three reference methods.

Secondly, we propose a real-time method for moving pedestrian detection and tracking in RMB-Lidar sequences for dense surveillance scenarios, with short- and long-term object assignment. Our tracker is non-model-based, using the assumption that people movements are expected in the monitored scene. During the Short-Term Assignment (STA) the different people are separated in the foreground regions of the point cloud frames, and the corresponding centroid positions are assigned to each other over the consecutive time frames. The Long-Term Assignment (LTA) is responsible for connecting the broken trajectories caused by STA errors and identifying the re-appearing people. This step is accomplished by extracting simple discriminative features from the tracked object sequences, and these descriptors are archived if the object disappears from the FoV. For newly appearing objects the descriptors are extracted over an initialization period, then re-activation is based on matching a given new object with its possible archived or temporarily invisible predecessors. As a consequence, in our system the STA of the tracking process can be obtained in real-time, while the identification information is

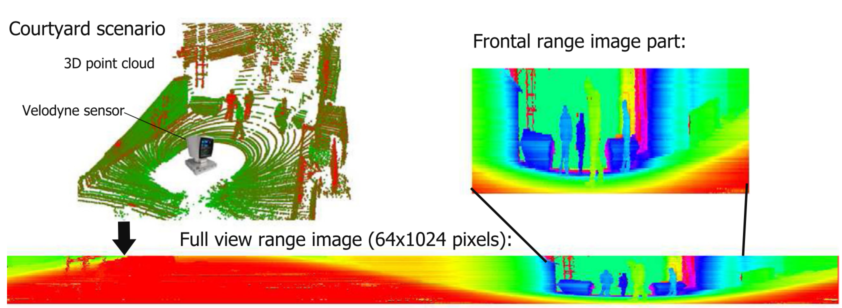

Fig. 1. Point cloud recording and range image formation with a Velodyne HDL-64E RMB-Lidar sensor.