Available online at www.sciencedirect.com

ScienceDirect

CAAI Transactions on Intelligence Technology 1 (2016) 210–224
http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Original Article

Scene-adaptive hierarchical data association and depth-invariant part-based appearance model for indoor multiple objects tracking☆

Hong Liu*, Can Wang, Yuan Gao

Key Laboratory of Machine Perception (Ministry of Education) and Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School, Peking University, China

Available online 1 November 2016

Abstract

Indoor multi-tracking is more challenging compared with outdoor tasks due to frequent occlusion, view-truncation, severe scale change and pose variation, which may bring considerable unreliability and ambiguity to target representation and data association. So discriminative and reliable target representation is vital for accurate data association in multi-tracking. Pervious works always combine bunch of features to increase the discriminative power, but this is prone to error accumulation and unnecessary computational cost, which may increase ambiguity on the contrary. Moreover, reliability of a same feature in different scenes may vary a lot, especially for currently widespread network cameras, which are settled in various and complex indoor scenes, previous fixed feature selection schemes cannot meet general requirements. To properly handle these problems, first, we propose a scene-adaptive hierarchical data association scheme, which adaptively selects features with higher reliability on target representation in the applied scene, and gradually combines features to the minimum requirement of discriminating ambiguous targets; second, a novel depth-invariant part-based appearance model using RGB-D data is proposed which makes the appearance model robust to scale change, partial occlusion and view-truncation. The introduce of RGB-D data increases the diversity of features, which provides more types of features for feature selection in data association and enhances the final multi-tracking performance. We validate our method from several aspects including scene-adaptive feature selection scheme, hierarchical data association scheme and RGB-D based appearance modeling scheme in various indoor scenes, which demonstrates its effectiveness and efficiency on improving multi-tracking performances in various indoor scenes.

Keywords: Multiple objects tracking; Scene-adaptive; Data association; Appearance model; RGB-D data

1. Introduction

In recent years, due to the growing demand for smart home applications [1–3], issues like multi-tracking [4,5], and re-identification [6–8] attract more and more attention from researchers in computer vision field, and also have many problems to be solved. Multiple objects tracking has been an active research topic in computer vision within a long period of time. It aims to locate moving objects, maintain their identities and retrieve their trajectories [4]. However, this is highly challenging in crowd environments with frequent occlusion, targets having similar appearances and complicated interaction.

Most previous methods that focus on multiple objects tracking can be organized into two main categories: One category takes information from future frames [9–13] to get better association via global analysis, like global trajectory optimization [9], network flows [11], hierarchical tracklets association [14], etc. However, they are based on the prerequisite that detection responses in all frames are given, both

☆ This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046), National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (JCYJ20130331144631730, JCYJ20130331144716089), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011).
* Corresponding author.
E-mail addresses: hongliu@pku.edu.cn (H. Liu), canwang@pku.edu.cn (C. Wang), ygao@sz.pku.edu.cn (Y. Gao).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.10.008

from past and future, so it is not suitable for time-critical applications and is relatively computation-consuming when performing a global optimization. The other category only considers past and current frames to make association decisions [15–18]. They usually relies on Kalman [19] or particle filter [20] to handle data association. Because of their recursive nature, this category is suitable for time-critical applications, but it may easily lead to irrecoverable wrong data association in crowded scene with similar appearance and complicated interactions. This requires the system not only has enough ability to discriminate all targets on current frame, but also has stable representation for each target in consecutive frames.

In order to increase the discriminating power, many pervious works [4,14,22] usually combine a bunch of features to represent detection responses and calculate the affinity matrix between them and existing tracklets. But they are with unsatisfactory performances on handling relatively challenging scenes for two reasons: First, feature representation of a same target may exhibit large variation due to illumination variation and wide range of poses. This indicates that stable representation of the target is hard to obtain. Second, observation errors of targets representation are common in cluttered scenes. For example, positions of detection responses may not be exactly on the center of targets due to frequent view-truncation and partial occlusion (as shown in Fig. 1), especially in indoor scenes field-of-view of sensors is relatively limited compared with outdoor scenes. In addition, accuracy of a detection response also relies on the detector's performance in the applied scene, which may vary a lot in different scenes. This leads to that same feature representation may have different reliability in different scenes. Therefore, combining a bunch of features may not contribute to a better association between detection responses and existing tracklets. On the contrary, features have lower reliability or discriminating power may bring adverse effect on reliable and discriminating features, and also unnecessary computational cost. Moreover, currently applications on network cameras harshly require more general and scene-adaptive schemes to handle the variety and diversity of the widespread scenes.

Therefore, motivated by properly handling the above problems and in order to achieve a time-critical indoor multi-tracking system, our work focus on accurate data association, scene-adaptive feature selection, and better appearance model. Our main contribution lies in three aspects: First, a novel hierarchical data association scheme based on the hierarchical feature space is proposed. Features are gradually combined during data association procedure according to the need of discriminating ambiguous detection responses, this avoids unnecessary computation cost and reduce error accumulation compared to simultaneously fusing bunch of features; Second, a scene-adaptive feature selection scheme is proposed, which measures features' reliability and disciminability of features used for targets representation in the applied scene, and selects relatively reliable and discriminative features for data association. This makes the algorithm much more general for various scenes in the camera network; Third, a novel depth-invariant appearance model is proposed as a high level feature for target representation, which properly handles server scale problem on 2D image plane, frequent view-truncation and partial occlusion which occur commonly in indoor environments. Experiments conducted in a variety of scenes demonstrate the effectiveness of reliable feature selection and hierarchical data association. The depth-invariant appearance model based on RGB-D data also shows its effectiveness when dealing with occlusion and scale change in complex indoor scenes.

2. Related work

2.1. Data association

Data association based multi-tracking methods become increasingly popular driven by the recent progress in object

Fig. 1. Tough problems for multi-tracking tasks. The first row shows various indoor and outdoor scenes with various illumination conditions, crowdedness and angle of views. The bottom two rows are detection responses obtained from our indoor datasets with large scale variation, frequent truncation by the field-of-view, partial occlusion, and wider range of poses than outdoor pedestrians [21].