Computer Vision and Image Understanding 000 (2016) 1–11

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Multi-object tracking via discriminative appearance modeling

Shucheng Huang[*], Shuai Jiang, Xia Zhu

Jiangsu University of Science and Technology, No.2, Mengxi Road, Zhenjiang City, Jiangsu, China

ARTICLE INFO

Article history:
Received 31 August 2015
Revised 4 June 2016
Accepted 9 June 2016
Available online xxx

Keywords:
Multi-object traking
Multiple hypotheses tracking
Metric learning
Multi-cue fusion
Appearance modeling

ABSTRACT

Tracking multiple objects is important for automatic video content analysis and virtual reality. Recently, how to formulate data association optimization more effectively to overcome ambiguous detected responses and how to build more effective association affinity model have attracted more concerns. To address these issues, we propose a metric learning and multi-cue fusion based hierarchical multiple hypotheses tracking method (MHMHT), which conducts data association more robustly and incorporates more temporal context information. The association appearance similarity is calculated using the distances between feature vectors in each associated tracklet and the salient templates of each track hypothesis, which is then fused with the dynamic similarity calculated according to Kalman filter online to get association affinity. To make appearance similarity more discriminative, the spatial-temporal relationships of reliable tracklets in sliding temporal window are used as constraints to learn the discriminative appearance metric which measures the distance between feature vectors and salient templates. The salient templates of generated track hypotheses are updated using an incremental clustering method, considering the high order temporal context information. We evaluate our MHMHT tracker on challenging benchmark datasets. Qualitative and quantitative evaluations demonstrate that the proposed tracking algorithm performs favorably against several state-of-the-art methods.

1. Introduction

Tracking is a fundamental task for video understanding in computer vision and pattern recognition. Video data increase rapidly every day due to the development of mobile terminal, digital camera networks and communication technology. The vast amount of data contains a lot of redundant information, which needs to be effectively presented and abstracted. The ability to simultaneously track multiple objects is important for automatic video analysis and virtual reality, which can provide global trajectory and pose information for higher level analysis and decision, and have numerous applications including target identification, intelligent surveillance, video coding, video analysis and human action recognition.

Multi-object tracking aims at inferring trajectories for each object from video sequence, which can be considered as a dynamic incremental spatiotemporal clustering problem. All image regions are clustered as certain specific object or background. In real senses, there are many challenges such as variance of number of objects, objects with similar appearances, variance of object appearance, complex motion, long time occlusions and clutter background, which generate the uncertainty and make multi-object tracking difficult. Recently with the significant progresses in research on object detection and classification, methods adopting tracking by detection framework become more and more popular. These approaches are much more flexible and robust in complex scenes where the camera may move or zoom in/out, and they are completely automatic without manual initialization which is important in practical applications.

There are two important parts in multi-object tracking based on tracking by detection framework: (1) association optimization model and (2) association affinity model. To execute association optimization, formulation of the first-order Markov model which is widely used in single object tracking can be extended in multi-object tracking, such as Joint Probabilistic Data Association Filter (JPDA) (Bar-Shalom et al., 2009) and methods based on particle filter (Breitenstein et al., 2009; Khan et al., 2005; Qu et al., 2007). However, it is usually much more effective to overcome the ambiguities in tracking process by considering information of past and future feedback simultaneously, which is the main conception of the current methods, such as multiple hypotheses tracking (MHT) (Cox and Hingorani, 1996; Miller et al., 1997; Ryoo and Aggarwal, 2008), markov chain monte carlo (MCMC) data association (Benfold and Reid, 2011; Oh et al., 2009; Yu and Medioni, 2009), and network flow graph (Ben Shitrit et al., 2011; Pirsiavash et al., 2011; Zhang et al., 2008). Recently, some work utilizes reliably associated tracklets as elements instead of detection results

[*] Corresponding author.
E-mail address: schuang2015@gmail.com (S. Huang).

http://dx.doi.org/10.1016/j.cviu.2016.06.003
1077-3142/© 2016 Elsevier Inc. All rights reserved.

to gradually reduce association ambiguities level by level (Brendel et al., 2011; Huang et al., 2008; Xing et al., 2009), which is robust in complex scene. However, in many association frameworks, especially the hierarchial methods, temporal context is not effectively utilized, and assignments between assumed trajectories and current measurements are only based on detected responses or tracklets in previous adjacent time.

Moreover, association affinity calculation plays an important role in multi-object tracking. Data association meets the challenges caused by detection error, occlusion, and similar appearance among multiple objects. The experiential affinity model used in most previous methods cannot evaluate the data association well. Some recent work focuses on building more effective affinity calculation model to achieve better results. Li et al. (2009) propose a learning association approach by training offline hybrid-boosted classifier to formulate affinity calculation as a joint problem of ranking and classification. This method needs ground truth in the same scene to generate samples. Kuo et al. (2010) propose an online learned discriminative appearance model which uses the spatial-temporal relationships of reliable associated tracklets as sampling constraints. They also propose an improved vision (Kuo and Nevatia, 2011) which incorporates target-specific appearance model for tracklets confirmed belonging to an object. These approaches only model data association as bipartite graph of tracklets. To decide whether a tracklet is associated to an object, it needs additional algorithms and heuristic information such as occupancy map. Although many proposed affinity models have been proposed, there are still two major issues: how to adapt appearance similarity model in video, and how to effectively embed appearance similarity in affinity calculation by considering other cues, such as dynamic similarity.

To address the above issues, we propose a hierarchical MHT framework based on reliable generated tracklets, which combines the merits of the hierarchial methods and MHT, conducting data association level by level to hierarchically reduce the ambiguities to form object tracks, while incorporating more temporal context information. The association affinity is calculated by fusing the dynamic similarity which is calculated using the Kalman filter with the appearance similarity online via logistic regression. The appearance similarity is defined on the distances between the salient templates of each track hypothesis and feature vectors which are extracted in the detected responses of associated tracklet. An incremental clustering method is adopted to attract and update the salient templates for generated track hypotheses, considering the high order temporal context. To enhance the discrimination of appearance similarity calculation, the distance metric is learned according to the constraints formed by the spatial-temporal relationships of tracklets. We summarize the contributions of this work in three folds:

- A hierarchial MHT association framework, which has the merits of hierarchial manner and MHT, is proposed to gradually link the short tracklets to obtain object track by considering more temporal context information.
- A robust appearance model is proposed to calculate the similarity of the salient template set and consider the high order temporal context of the generated tracks. The appearance similarity is then fused with the dynamic similarity by logistic regression with reliable tracklets.
- The discriminative appearance metric is learned using the spatial-temporal relationships of tracklets as constraints to measure the similarity between feature vectors and salient templates for appearance modeling.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the MHT algorithm in details. Section 4 gives the overview of the proposed tracking framework. Section 5 shows the proposed multiple information fusion based MHT Algorithm. Section 6 reports the experimental results on the widely used representative datasets. We conclude this paper in Section 7.

2. Related work

Visual tracking includes single object tracking (Zhang et al., 2013a,2015a,b) and multi-object tracking (Breitenstein et al., 2009; Khan et al., 2005; Milan et al., 2014; Qu et al., 2007). Research in multi-object tracking filed concentrates on the construction of affinity model and the strategy of association optimization. The first-order Markov process, broadly used in modeling single object tracking problem, can be extended to formulate data association in multi-object tracking. Although many interaction models are incorporated (Breitenstein et al., 2009; Khan et al., 2005; Qu et al., 2007), it is usually effective to overcome the ambiguities of occlusions, spurious measurements and missing measurements by considering information of past and future feedback simultaneously to estimate the current state. A kind of classic method is multiple hypotheses tracking (MHT) (Cox and Hingorani, 1996; Zhang and Xu, 2014) which is first used in radar signal processing. MHT maintains possible associations with higher probability (Miller et al., 1997) in a period of time to find out approximate optimal solution when receiving new information. This method generates many redundant hypotheses and relies on prune strategy to reduce the computational complexity. Recently, an observe-and-explain paradigm (Ryoo and Aggarwal, 2008) of MHT is proposed. This method enumerates tracking possibilities only when the system has enough information to evaluate them, which avoids exponential number of possibilities due to insufficient data. However, some additional heuristic decision procedures should be conducted to maintain and generate hypotheses.

Inspired by the MHT algorithm, many approaches have been proposed. MCMC (Benfold and Reid, 2011; Oh et al., 2009; Yu and Medioni, 2009) data association uses Metropolis-Hastings algorithm to construct an irreducible and aperiodic Markov chain, which is employed to efficiently sample from the posterior distribution of state space and obtain approximate optimal solution. These methods need a large number of iterations, and lead to higher time complexity. Some approaches construct network flow graph models (Ben Shitrit et al., 2011; Pirsiavash et al., 2011; Zhang et al., 2008) to map the maximum-a-posteriori (MAP) data association problem into cost-flow network with non-overlap constraint on trajectories. Zhang et al. (2008) propose an explicit occlusion model which gradually enlarges the original node set with occluded hypotheses and imposes lower bound constraints after each iteration of network optimization. The number of tracks needs to be pre-set, and the solution is obtained if all tracks are formed. The paper (Pirsiavash et al., 2011) shows that the global solution including the number of tracks can be obtained using the shortest path computation on a flow network, and gives a near-optimal algorithm based on 2-pass dynamic programming. However, the cost-flow network can only address trajectory affinity functions, which can be decomposed as the multiplication of affinity functions between two adjacent instants, restricting the presentation power.

There are also approaches addressing global association with hierarchical manner (Brendel et al., 2011; Huang et al., 2008; Xing et al., 2009; Zhang et al., 2012). In the paper (Huang et al., 2008), a three-level hierarchial association approach is proposed, which generates reliable tracklets at the low level, formulates the association as bipartite graph at the middle level, and estimates entries, exits and scene occluders using the already computed tracklets at the high level. The paper (Xing et al., 2009) adopts particle filter in local stage to generate reliable tracklets, which is buffered when