Computer Vision and Image Understanding 143 (2016) 25–37

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Global optimization for coupled detection and data association in multiple object tracking☆

Zheng Wu*, Margrit Betke

Computer Science Department, Boston University, Boston, MA02215, USA

ARTICLE INFO

Article history:
Received 11 February 2015
Accepted 6 October 2015
Available online 22 October 2015

Keywords:
Sparsity-driven detection
Network-flow data association
1-bit de-quantization
Dual decomposition
Multiple object tracking

ABSTRACT

We present a novel framework for tracking multiple objects imaged from one or more static cameras, where the problems of object detection and data association are expressed by a single objective function. Particularly, we combine a sparsity-driven detector with the network-flow data association technique. The framework follows the Lagrange dual decomposition strategy, taking advantage of the often complementary nature of the two subproblems. Our coupling formulation avoids the problem of error propagation from which traditional "detection-tracking approaches" to multiple object tracking suffer. We also eschew common heuristics such as "non-maximum suppression" of hypotheses by modeling the joint image likelihood as opposed to applying independent likelihood assumptions. Our coupling algorithm is guaranteed to converge and can resolve the ambiguities in track maintenance due to frequent occlusion and indistinguishable appearance between objects. Furthermore, our method does not have severe scalability issues but can process hundreds of frames at the same time. Our experiments involve challenging, notably distinct datasets and demonstrate that our method can achieve results comparable to or better than those of state-of-art approaches.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The interpretation of the motion of large groups of individuals is a challenging problem in computer vision. A complete multi-object tracking system typically consists of three phases: object detection, temporal data association (i.e., the assignment of current observations to object tracks) and an optional step of state estimation for each object. These modules are usually arranged sequentially in the system so each subproblem itself is optimized independently. However, due to the lack of robust methods for object detection in uncontrolled environments and poor scalability to large numbers of objects, current tracking systems do not generalize well on challenging datasets. Most previous efforts have therefore followed two distinct directions of research: building stronger object detectors and designing better data association methods. As a result, almost all existing tracking systems use a "detection-tracking design" with two separate modules to address the detection and data association tasks.

The detection-tracking design has the inherent weakness that it requires the output of the detection module to be reliable in order for the data association module to work properly. Detection errors such as "false alarms" and "missed detections" otherwise propagate to the data association module and false matches need to be corrected later.

In contrast, we show that error propagation from detection to data association can be avoided if both tasks, detection and data association, are combined into a single module and solved simultaneously by optimizing a single objective function. In addition, we show that temporal information could significantly improve the performance of an object detector in our coupling framework. The coupling idea appears attractive but introduces new challenges as well:

1. What type of objective function should be used? Many existing detection methods have not even been formalized as an optimization problem.
2. How can the new objective function be solved? Many current data association methods are complicated, approximate solutions to intractable problems. A new objective function that couples detection and data association might be even more difficult to optimize.
3. How can scalability of the proposed method be ensured? Computer vision systems face demands for being able to track large numbers of objects in dense formations. Given such large input sizes, an efficient algorithm to optimize the new objective function must be found.

In this paper, we address all the questions above with a formulation of a coupling function and a method to optimize it. In particular, we propose a detection method with the classic sparse-signal recovery technique [1] for the dense-object tracking scenario when a foreground estimation technique is available. This method can be used to

☆ This paper has been recommended for acceptance by Vittorio Murino.

* Corresponding author.
E-mail address: wuzheng1127@gmail.com, wuzheng@bu.edu (Z. Wu).

http://dx.doi.org/10.1016/j.cviu.2015.10.006
1077-3142/© 2015 Elsevier Inc. All rights reserved.

detect objects moving on the ground plane as well as objects moving in free 3D space. The sparsity constraint is important here because it can significantly reduce the number of false alarms and serves as a replacement of the heuristic technique of non-maximum suppression of hypotheses. We have to take care, however, that the approach does not lead to overly sparse results, that is, missed detections. To further boost the detection accuracy, we also impose a smoothness constraint from the data association aspect where we assume the state of each object follows a first-order Markov process and adopt the classical network flow formulation [2].

Unlike previous coupling formulations that rely on a coordinate descent technique [3], our overall objective function has a simple form and can be solved through Lagrange dual decomposition. The method distributes the coupling formulation to subproblems and coordinates their local solutions to achieve a globally optimal solution. For each subproblem, a very efficient off-the-shelf algorithm is available. The proposed paradigm also permits distributed computing. Finally, our method was tested both for monocular and multi-view videos, and achieved consistent robustness across several challenging, notably distinct datasets. In summary, our contributions are:

1. A novel and flexible framework for coupling the subproblems of detection and data association of multiple-object tracking in a single objective function, which permits using a straightforward global optimization technique.
2. A new sparsity-driven object detection method that takes binary foreground/background segmentation as input, infers mutual occlusion relationships, and achieves a high detection rate even when objects are severely occluded.
3. A general mechanism to suppress false alarms without "non-maximum suppression" either in the detection or in the data association stage.

2. Related work

An important motivation for designing the coupling framework is to perform occlusion reasoning, a major bottleneck of current tracking systems. In this paper, we focus on tracking multiple objects that are imaged with limited resolution from one or more static cameras and may partially occlude each other. It is also possible to encode a scene-specific occlusion model [4] into our formulation.

2.1. Occlusion reasoning in detection

There are two main challenges for occlusion reasoning in object detection. First, when an object becomes occluded and there is no or only a partial observation of the object on the image plane, model-based detectors must deal with difficult-to-predict uncertainty. Second, the heuristic "non-maximum suppression" technique, adopted by most detection methods, which aims to cluster close hypotheses, also explains away true detections.

The first challenge may be addressed by a part-based detector, which may be able to detect certain partially occluded objects, as long as the visible part appears with sufficient resolution [5–7], and the detector can be adjusted online [8–11]. Part-based detectors fail when objects are completely occluded or the resolution of an object is too small. Even for pedestrians, an object category well-studied in the computer vision community, the performance of the current state-of-art method [12] drops significantly under partial occlusion and degrades catastrophically for small resolution. Moreover, direct modeling of occlusion is difficult in general, as the degree of partial occlusion needs to be explicitly expressed in the object model. However, a detailed object model/detector is not necessary for many surveillance applications. Sometimes, it is not even useful due to limited resolution or challenging imaging conditions. As an alternative, when a reasonable background subtraction method is available, a common idea is to fit binary shape templates to the observations with the help of scene knowledge, such as camera calibration or multiview geometry [13–16]. These methods all rely on a background subtraction preprocessing step (which can be a difficult problem to develop). Therefore, they are sensitive to the quality of background subtraction and the degree of partial occlusion.

The side effect of the non-maximum suppression technique to filter out true detections is particularly undesirable when objects have a large overlap on the image plane. Instead of letting this heuristic step make ad-hoc decisions or trying to tune parameters for it, a number of recent works have shown that it is beneficial to formulate the object detection problem as a global optimization problem with a minimum description length (MDL) constraint or a context prior [16–18], and let the optimization process determines which hypotheses to select. Our detection methods used in the coupling framework fall into this category.

2.2. Occlusion reasoning in tracking

Despite efforts to detect partially occluded objects, missed detection/false positives are still inevitable, and any ambiguity could be resolved in the data association stage. Research efforts for multiple object tracking typically treat occluded objects as missed detection events or track occluded objects all together with a single tracker and iteratively grow or stitch tracklets (track fragments) before and after occlusions [19–28]. The main effort in this research direction is to design a discriminant similarity measure for the tracklet matching problem so that occlusion is implicitly resolved by filling gaps after stitching tracklets. It is important to note that all these approaches follow the "detection-tracking strategy" and therefore rely on good detectors for initialization. The limited ability through tracking to correct the error propagated from the detection stage typically implies that the missed detection/false positive events are assumed to occur rarely. However, if this assumption does not apply, hoping that the data association module itself will fix all detection errors is not promising.

Explicit occlusion modeling (OM) also appears in an optimization work by Andriyenko et al. [29], who integrated an occlusion model in their global objective function. As the objective function becomes more and more complicated, it becomes highly non-convex, and the optimization relies on good initialization as well as ad-hoc sampling heuristics to avoid local minima. Our formulation avoids these complications, is mathematically rigorous and simpler to optimize.

Another category of occlusion reasoning is to combine both the detector's output and motion flow estimation [30,31]. The intuition is that the dense motion flow is usually less affected by partial occlusion than an object detector so that clustering the low-level trajectories can help improve the overall tracking performance when objects are only partially visible. However, motion flow itself is not able to distinguish individuals in a crowd with coherent motion. It also significantly increases the amount of computation.

2.3. Relation to the proposed method

As occlusion cannot be resolved solely in the detection or data association phases, a natural extension is to consider combining these two subproblems into a single framework and take advantage of the often complementary nature of the two subproblems. A generative part-based model was proposed [32] that combines tracking and detection, and models both the approximate articulation of each person as well as the temporal coherency within a walking cycle. An extended part-based model has also been proposed to build a "joint people detector" that combines a state-of-the-art single person detector with a detector for pairs of people [11]. While such detailed part-based models offer a principled way to handle inter-person

Download English Version:

https://daneshyari.com/en/article/527285

Download Persian Version:

https://daneshyari.com/article/527285

Daneshyari.com