Computer Vision and Image Understanding 136 (2015) 103–116

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Identifying multiple objects from their appearance in inaccurate detections [☆]

Julian F.P. Kooij, Gwenn Englebienne, Dariu M. Gavrila [*]

Intelligent Autonomous Systems Group, Informatics Institute, University of Amsterdam, Sciencepark 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 19 April 2014
Accepted 20 March 2015

Keywords:
Object recognition
Segmentation
Generative model
Unsupervised learning
Latent Dirichlet Allocation
Video surveillance

ABSTRACT

We propose a novel method for keeping track of multiple objects in provided regions of interest, i.e. object detections, specifically in cases where a single object results in multiple co-occurring detections (e.g. when objects exhibit unusual size or pose) or a single detection spans multiple objects (e.g. during occlusion). Our method identifies a minimal set of objects to explain the observed features, which are extracted from the regions of interest in a set of frames. Focusing on appearance rather than temporal cues, we treat video as an unordered collection of frames, and “unmix” object appearances from inaccurate detections within a Latent Dirichlet Allocation (LDA) framework, for which we propose an efficient Variational Bayes inference method. After the objects have been localized and their appearances have been learned, we can use the posterior distributions to “back-project” the assigned object features to the image and obtain segmentation at pixel level. In experiments on challenging datasets, we show that our batch method outperforms state-of-the-art batch and on-line multi-view trackers in terms of number of identity switches and proportion of correctly identified objects. We make our software and new dataset publicly available for non-commercial, benchmarking purposes.

1. Introduction

Traditional tracking-by-detection approaches contain a data association step in which detections are matched to inferred object properties such as appearance, size, and location. However, this poses problems when objects are temporarily (partially) occluded or undetected. Wrong data association can deteriorate the learned object appearances, which further affects future associations. This exposes a chicken-and-egg problem: To localize objects in a scene, image observations need to be correctly associated to candidate objects, which requires knowledge of the object-specific identifying properties. Learning such properties from the observations, however, requires prior knowledge of the presence and location of the objects in the scene. Additionally, camera calibration may be unreliable (i.e. camera orientation may have changed) or unavailable, thus predefined detection windows may not necessarily fit the targets, and objects may exhibit unusual pose or size, resulting in low confidence detections that complicate association even more.

This paper focuses on these related problems, and presents a novel graphical model to determine the number of objects, their appearance, and their location per frame, from possibly *inaccurate* detections. To exploit detections that contain multiple partially occluding objects and background, we seek to loosen the traditional one-to-one relation between detections and objects, and instead infer which of the low-level appearance features present in a detection belong to what target. Key to our method is the adaptation of Latent Dirichlet Allocation (LDA) to “unmix” a number of consistent object appearances in the comparatively large number of detected regions, which are represented as a bag-of-features. This allows us to infer the presence and appearance of an object, even when a single object is responsible for multiple detections and when a single detection spans multiple objects, as often happens in the case of partial occlusion of one object by another. Hence, occluding objects are separated at the feature level and we eliminate the need for special treatment of assignments and appearance updates under occlusion. Additionally, we exploit that in a single frame an object is local to a part of that frame, so that non-overlapping detections are unlikely to both contain the same object. This spatial constraint is enforced by modeling each frame as a mixture of objects whose feature locations have a Gaussian distribution centered at an object’s image location. In post-processing, we can optionally “back-project” the feature labels in all images, and segment individual targets. These steps are illustrated in Fig. 1 on a frame from a challenging fight scene.

[☆] This paper has been recommended for acceptance by Nikos Paragios.

[*] Corresponding author.

E-mail addresses: julian.kooij@gmail.com (J.F.P. Kooij), g.englebienne@uva.nl (G. Englebienne), d.m.gavrila@uva.nl (D.M. Gavrila).

http://dx.doi.org/10.1016/j.cviu.2015.03.012
1077-3142/© 2015 Elsevier Inc. All rights reserved.

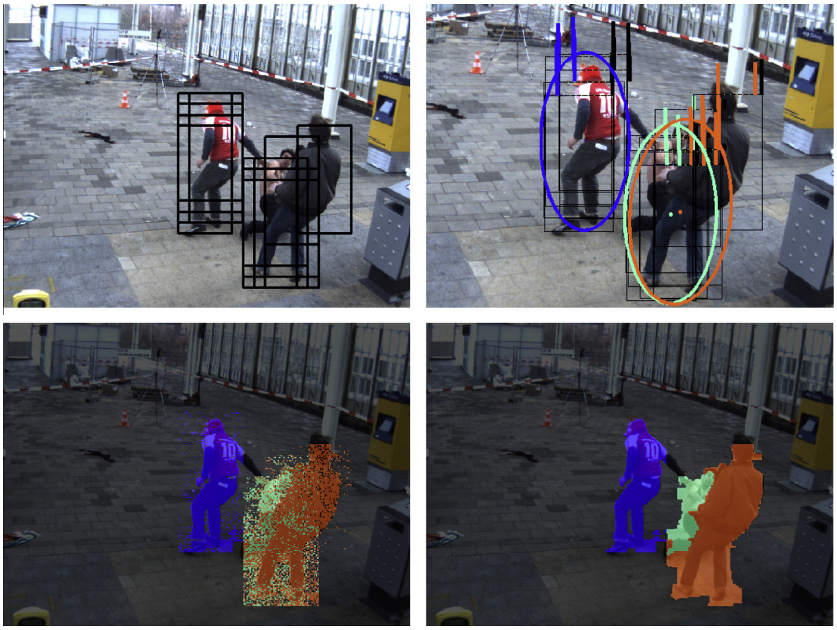

Fig. 1. Best viewed in color. Top left: example illustrating the challenges we address: detections (shown as black boxes) are inaccurate due to challenging poses, perspective and occlusions. Top right: inferred object presence and location. For each detection window the proportion of pixels associated with an object ID is indicated by a color-coded bar graph on top, black bars indicate background. Bottom left: for each pixel we sample an ID to illustrate the mixing proportions (optional post-processing). Bottom right: per pixel object IDs after global image segmentation (optional post-processing). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To demonstrate the appearance "unmixing" paradigm, we address multi-target association (up to the pixel level) in a batch of frames for scenes with static background and a reasonable upper limit on the expected object count, but where detections from an external object detector may be inaccurate. Our method is compatible with traditional tracking frameworks where motion models reduce the positional uncertainty, since our model incorporates a prior distribution over each object's location. Many different approaches have been proposed to enforce temporal consistency (e.g. merging tracklets [32], searching the space–time volume for globally consistent paths [13,22] or particle filters [11,31]), and state-of-the-art trackers have used strong motion models and the explicit specification of entry/exit regions to push performance. Therefore, our work focuses on appearance without dictating how to exploit such additional cues, and can be seen as complementing recent work on tracking under occlusion [2,22] which solely relies on temporal information. As a result, this paper treats video as an unordered collection of frames without temporal information, using the same positional prior for each individual frame, though for evaluation the output of our model will be compared to that of complete multi-target trackers.

2. Previous work

Our proposed method is related to tracking, image segmentation and object-recognition methods. The body of literature on these is extensive, and here we can only discuss a small set of papers which are most directly relevant.

Tracking-by-detection employs some method to detect target objects in video frames, and combine the detections into consistent tracks. For example, person detectors can be trained on HOG features [12], or by reasoning about spatial occupancy to explain background/foreground masks [5,13,21]. In terms of tracking, we can distinguish between on-line trackers that incorporate new observations on a frame-by-frame basis [11,15,21,31], and *batch* methods that perform global optimization for multiple frames at once [1–3,22,5,24,33].

Multi-view trackers observe targets simultaneously from various overlapping views, and are therefore more robust against occlusion in a single view. Using Probabilistic Occupancy Maps (POM) [13] for detection in calibrated views, [5] applies global appearance constraints to formulate a network flow optimization problem over all frames. This requires defining *a priori* appearance templates for distinct object classes, rather than learning the appearance of individuals from data. The on-line multi-view tracker of [21] relies on background subtraction and voxel carving to find candidate object locations in 3D. Using the Hungarian method [18], candidates are assigned to tracks or labeled as 'ghosts', i.e. faulty correspondences of foreground from different views, based on similarity scores for appearance, size, and Kalman filtered position. Back-projecting voxels to the images yields per-view object masks that take inter-object occlusion into account, which are used to learn the object appearances.

In the single view tracker of [24], appearances are first learned by clustering body part patches from a generic part-based person detector. The trained model is then used to track an individual and makes it possible to reason about self-occlusion. In [33] a part-model is used to deal with other types of occlusions, and fuses tracklets (i.e. trajectory fragments) into consistent tracks while learning a discriminative appearance model for each person. Another use of part-based models is to exploit the dynamics of parts to disambiguate tracks and recover after occlusion, e.g. [1] distinguishes multiple people seen in side-view from their articulated leg pose within a walking cycle.

Part-based models are not the only way to deal with occlusion. [2] adds occlusion reasoning to a continuous energy minimization framework [3] that globally optimizes detection-to-track associations under temporal constraints. In [22] this framework is extended to mixed discrete–continuous optimization for improved data association, with additional global constraints to consider track dynamics and exclude collisions, and keep overlapping tracks separated.

Others treat occlusions as temporary occurrences where no association can reliably be made. For instance, [15] weighs a large