ISPRS Journal of Photogrammetry and Remote Sensing xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Probabilistic multi-person localisation and tracking in image sequences

T. Klinger *, F. Rottensteiner, C. Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, D-30167 Hanover, Germany

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form 6 November 2016
Accepted 16 November 2016
Available online xxxx

Keywords:
Dynamic Bayesian Network
Gaussian process
Linear programming
Pedestrians
Tracking
Video

ABSTRACT

The localisation and tracking of persons in image sequences in commonly guided by recursive filters. Especially in a multi-object tracking environment, where mutual occlusions are inherent, the predictive model is prone to drift away from the actual target position when not taking context into account. Further, if the image-based observations are imprecise, the trajectory is prone to be updated towards a wrong position. In this work we address both these problems by using a new predictive model on the basis of Gaussian Process Regression, and by using generic object detection, as well as instance-specific classification, for refined localisation. The predictive model takes into account the motion of every tracked pedestrian in the scene and the prediction is executed with respect to the velocities of neighbouring persons. In contrast to existing methods our approach uses a Dynamic Bayesian Network in which the state vector of a recursive Bayes filter, as well as the location of the tracked object in the image, are modelled as unknowns. This allows the detection to be corrected before it is incorporated into the recursive filter. Our method is evaluated on a publicly available benchmark dataset and outperforms related methods in terms of geometric precision and tracking accuracy.

1. Introduction

Visual localisation and tracking of persons is one of the most active research topics in the fields of photogrammetry and computer vision. The generated trajectories carry important information for the semantic analysis of scenes and thus form a crucial input for many applications in fields such as autonomous driving, field robotics and visual surveillance. Many available systems apply object detection in single frames, an association step for linking detections to trajectories, and recursive filtering to find a synthesis between image-based measurements and a motion model.

In a multi-person tracking scenario, persons often need to react to the motion of other persons, so that the accuracy of a prediction can be improved if such interactions are taken into account. Considering information about interactions, which is often referred to as *motion context* (c.f. Yoon et al., 2015), in the motion model allows to generate more plausible predictions in the absence of measurements.

While most methods for tracking are concerned with a correct assignment of objects, only few papers address the geometric accuracy of a detection. However, in some applications, e.g. in driver assistance systems, where one has to decide whether a pedestrian does actually enter the vehicle path or not, the geometric accuracy of a trajectory is crucial.

This article investigates a Dynamic Bayesian Networks (DBN) (Dean and Kanazawa, 1989) for the purpose of recursive filtering. A DBN takes account of the structure of a probabilistic model in a factorised form and inference can be conducted efficiently using belief propagation (Pearl, 1988). The proposed DBN models both, the state vector of a recursive Bayes filter and the location of the tracked object in the image as unknowns. By modelling the parameters related to the pedestrian position in the image by hidden variables, our method carries out the update step of the recursive filter with an improved detection result, leading to an improved geometric accuracy of the posterior position. The model jointly considers camera orientation and scene geometry, a dynamic model, a generic state-of-the-art pedestrian detector and a classifier trained on individual persons.

For multi-person tracking we use a dynamic model which is based on Gaussian Process Regression. We formulate a new covariance function taking the spatial distance and the angular displacement of two trajectories into account. The output of the covariance function is used as a measure for the interaction between pedestrians. The covariance matrix contains the covariances of all pairs of pedestrians and is updated at every time step.

Tracking is carried out in monocular image sequences taken by potentially moving cameras. We assume that the calibration and

* Corresponding author.
E-mail address: klinger@ipi.uni-hannover.de (T. Klinger).

http://dx.doi.org/10.1016/j.isprsjprs.2016.11.006
0924-2716/

orientation parameters of the cameras are given at every time step. The 2D image coordinates are related to the world coordinates by the collinearity equations. Tracking is applied in a common 3D object coordinate system, leading to scale and viewpoint independence that enables the joint modelling of the motion parameters. We restrict the positions of pedestrians to a ground plane.

This article builds upon our previous work. A Dynamic Bayesian Network was proposed for the joint inference of the hidden state and location of the pedestrians in the image in Klinger et al. (2015); A new approach for the incorporation of motion context based on Gaussian Process Regression was proposed in Klinger et al. (2016).

The work presented in this paper makes four scientific contributions:

- A new model for the assessment of similarities between detections and tracked targets in a joint probabilistic data association framework is developed;
- Two new models for the generation of the feature vector using the Random Forest classifier are proposed and evaluated;
- A comprehensive study of the impact of the individual system components is carried out using image sequences from both, static and dynamic cameras;
- An empirical evaluation of the detection and non-maximum suppression strategy in comparison to state-of-the-art pedestrian detectors is given.

The new method is evaluated on a publicly available benchmark dataset. The results show that the geometric accuracy is superior to related work which directly incorporates single frame detections into the recursive filter as measurements. By using the proposed predictive model, the numbers of identity switches, false positive and false negative detections are reduced significantly compared to the related work.

The remainder of this paper is structured as follows. In Section 2 an overview over related work on the topics of this paper is given. In Section 3 we describe the methodology of our approach. Section 4 presents the experiments and in Section 5 we conclude our work and give an outlook on future work.

2. Related work

Existing approaches for tracking can be characterised according to the way in which the images are processed: frame by frame (online), in sections (e.g. tracklet-based approaches, leading to a delayed trajectory generation), or all at once (offline). Our work is set in the context of online recursive Bayesian estimation. In the following, we focus on related work on the topics of detection and localisation, data association, and motion context, as these topics are the central building blocks of our approach.

2.1. Detection and localisation

In recursive detection-based approaches for tracking, the measurements typically stem from classifiers trained offline on large sets of pedestrian images to generalise well across a wide range of different pedestrian characteristics (Dalal and Triggs, 2005; Felzenszwalb et al., 2010; Dollár et al., 2014). Although these approaches provide a solution to the recognition and localisation problem, the results are not particularly reliable and precise. In a comprehensive study, Dollár et al. (2011) show that the recall rate of 16 different pedestrian detectors decreases rapidly if the intersection-over-union score threshold is increased. When a mis-aligned detection is used in a tracking application, the generated trajectories are prone to be updated towards wrong positions. When applied in systems capable for online processing, these effects accumulate and the trajectories easily drift away from the target. This motivates a refinement of the detected object in image space prior to the actual updating of the filter.

A better alignment of the detection result to the real object location is for instance achieved by finding regions that best coincide with appearance features representative for a target. To this end, classifiers are used to learn target-specific models of appearance based on variants of Random Forests (Breiman, 2001; Saffari et al., 2009), Hough Forests (Gall and Lempitsky, 2013) and boosting (Breitenstein et al., 2011). The possiblity of adaptation to appearance changes makes these approaches more applicable to complex scenes with a wide range of depth, temporary occlusions, and changing lighting conditions. However, these approaches are quickly distracted from the actual target if the training data are derived from mis-aligned samples. Breitenstein et al. (2011) point out that the performance of online multi-person tracking with particle filtering can be improved when using specialised classifiers to weight the particles. With the classifiers trained by boosting, a single binary classifier must be learnt for every tracked person, so that each classifier only discriminates against the background and not directly against other persons. Random Forests are inherently applicable to multi-class problems, and are comparably fast in training and classification. For this reason, we follow Klinger et al. (2015) and use a Random Forest classifier for the tracking and refinement of multiple persons.

The reliability of automatic object detection is often affected by the fact that acceptable detection rates are only achieved if many false positive detections are accepted as well. To lower the risk of false positive detections additional clues like foreground information (Stauffer and Grimson, 2000) or shape (Leibe et al., 2005) are evaluated prior to further processing. When processing the available image-based observations sequentially, errors committed in one step cannot be corrected later. To this end, Hoiem et al. (2008), Schindler et al. (2010) and Choi et al. (2013) integrate different sources of information in the framework of probabilistic graphical models (Bishop, 2006; Förstner, 2013). Hoiem et al. (2008) refine the non-maximum suppression (NMS) of available object detectors by evaluating 3D information about the expected objects and the scene. The authors keep track of the distribution of the non-maximum detections, giving rise to a more sophisticated measure of uncertainties of the final detections. We follow Hoiem et al. (2008) and apply NMS in 3D object space and keep track of the distribution of the detections obtained at locations and scales in the vicinity of the optimal position. Different from that work, we further integrate prior knowledge about the scene into the process of NMS for scenes in which training data are available. For that purpose we use the generated detections and the results of a Random Forest classifier as one of the building blocks of the update procedure of recursive state estimation.

Schindler et al. (2010) make use of directed graphical models, i.e. Bayesian networks, for the joint inference of unknown parameters that are related to the correctness of a detection, object position, ground plane parameters and other variables. The positions of pedestrians are modelled as hidden variables in a Bayesian network, which are evaluated together with observations stemming from a pair of stereo cameras and from automatic object detection. The authors perform tracking in an additional step separately for every pedestrian. We follow Schindler et al. (2010) and optimise the pedestrian position in 3D space, with the difference that we combine the filtering of the pedestrian state and its localisation in the image in a single Dynamic Bayesian Network. This allows to represent the joint probability density function of the hidden and observed variables in factorised form and to incorporate the dynamic model.

Please cite this article in press as: Klinger, T., et al. Probabilistic multi-person localisation and tracking in image sequences. ISPRS J. Photogram. Remote Sensing (2017), http://dx.doi.org/10.1016/j.isprsjprs.2016.11.006