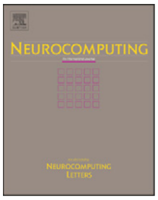
Neurocomputing 83 (2012) 165–175

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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



Multi-target tracking with occlusions via skeleton points assignment

Huan Ding*, Wensheng Zhang

Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Beijing 100190, China

ARTICLE INFO

Article history:
Received 19 May 2011
Received in revised form
6 December 2011
Accepted 10 December 2011
Communicated by Tao Mei
Available online 29 December 2011

Keywords:
Skeleton points assignment
Occlusion segmentation
Occlusion compensation
Multi-target tracking

ABSTRACT

Multiple-target tracking in complex scenes is one of the most complicated problems in computer vision. Handling the occlusion between objects is the key issue in multiple-target tracking. This paper introduces the method of motion segmentation into the object tracking system, and presents a SPA (Skeleton Points Assign, SPA) based occlusion segmentation approach to track multiple people through complex situations captured by static monocular cameras. In the proposed method, we first select the skeleton points and evaluate their occlusion states by low-level information like optical flow; then we assign these points to different objects using advanced semantic information, such as appearance, motion and color; finally, a dense classification of foreground pixels are taken advantages of to accomplish occlusion segmentation and a blob-based compensation strategy is utilized to estimate the missing information of occluded objects. Object tracking is handled by a particle filter-based tracking framework, in which a probabilistic appearance model is used to find the best particle. Experiments are conducted on the public challenging dataset PETS 2009. Results show that this approach can improve the performance of the existing tracking approach and handle dynamic occlusions better.

1. Introduction

Despite a lot of attention being dedicated to tracking multiple people in video sequences over the last 20 years, the problem remains very concerning in many computer vision and video processing tasks, such as video surveillance and event inference. In particular, the major challenge of multi-targets tracking is the frequent presence of visual occlusions. Occlusions make the current observation totally or partially unavailable for some time intervals. The problem of dealing with occlusions correctly is still an open subject.

Solutions to managing occlusions can be decomposed into two groups, depending on whether a scene is captured by a single stationary camera or by multiple cameras. Multi-view methods fuse information from multi-views to localize people on multiple scene planes [1]. These methods work well, except for special geometries of the camera views and people locations. These special geometries provide insufficient information to generate a unique signal for tracking, due to visual occlusions. In these cases more specialized tracking methods need to maintain track identity. In addition, for some applications, multiple views are not always available. Thus, this paper focuses on designing a monocular tracking methodology, with the goal of handling relatively complex occlusion scenarios.

A considerable amount of research has reported on the treatment of occlusions from a stationary monocular camera during the last decade. Most of them consider occlusion segmentation as part of the object model, and embed it into tracking process. These works build object models using color [2], appearance [2–5] and motion [6–8] information. Unfortunately, these models are learned for describing the postures or actions of the tracking targets, and do not fit well with occlusion segmentation.

There are also many attempts to deal with the problem of occlusions explicitly. Hai et al. [9] decomposed video frames into coherent 2D motion layers and introduced a complete dynamic motion layer representation in which spatial and temporal constraints on shape, motion and appearance are estimated using the EM algorithm. His method has been applied in an airborne vehicle tracking system and examples of tracking vehicles in complex interactions are demonstrated. Qian et al. [10] proposed a framework for treating the general multiple target tracking problem, which was formulated in terms of finding the best spatial and temporal association of observations that could maximize the consistency of both motion and appearance of object trajectories. Papadourakis et al. [11] presented a robust object tracking algorithm which could automatically build appropriate object representations by color and handles spatially extended and temporally long object occlusions. The majority of the above methods are under a simple assumption that object color satisfies either a single Gaussian or a Gaussian mixture distribution model.

* Corresponding author. Tel.: +86 010 82614489; fax: +86 010 62545229.
E-mail addresses: huan.ding@ia.ac.cn (H. Ding), wensheng.zhang@ia.ac.cn (W. Zhang).

0925-2312/$ - see front matter Crown Copyright © 2011 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.12.003

Meanwhile, these approaches do not take the historical information in the scene into account. Thus all the methods mentioned above lead to the poor performance of occlusion segmentation and multiple targets tracking for objects with similar color.

Recently, some important issues in utilizing motion segmentation in tracking system are discussed. In the context of motion segmentation, the literature can be divided in two kinds: direct methods and feature-based methods [12]. Direct methods recover the unknown parameters directly from measurable image quantities at each pixel in the image. This is in contrast with the feature-based methods, which first extract a sparse set of distinct features from each image separately, and then recover and analyze their correspondences in order to determine motion. Feature-based methods minimize an error measure that is based on distances between a few corresponding features, while direct methods minimize a global error measure that is based on direct image information collected from all pixels in the image. It is important to observe that with direct methods the pixel correspondence/classification is performed directly with the measurable image quantities at each pixel, while in feature-based methods this is done indirectly, based on independent feature measurements in a set of sparse pixels. An important property of the direct methods is that they can successfully estimate global motion even in the presence of multiple motions and/or outliers [13]. However, computational time is wasted by including in the minimization a large number of pixels where no flow can be reliably estimated. On the other hand, feature-based methods initially ignore areas of low information, resulting in a problem of fewer parameters to be estimated, with good convergence even for long sequences. Hence, considering the tracking efficiency, we only focus on the feature-based methods.

Feature-based methods for motion segmentation usually consist of two independent stages: (1) feature selection and/or correspondence and (2) motion parameter estimation [13]. The second stage is often performed through factorization methods [14], although some simpler clustering strategy can be used [15]. Several methods have been proposed for sparse feature selection and/or correspondence, and among them, the most popular are the Harris Corner Detector [16,17], and SIFT [18]. However, these sparse feature-based methods compute feature correspondences independently. Thus, they are very sensitive to outliers, making them susceptible to errors in motion parameter estimation/segmentation. Moreover, homogeneous regions of a frame may present none or few features, which results in the motion estimation/segmentation difficult (or even impossible) in large areas of the video frames. In object tracking field, Papadakis et al. [19] utilized a graph cuts approach and separated each object into visible and occluded parts using an energy function, which contains terms based on position and motion information. Silva et al. [12] obtained a pixel-wise segmentation by clustering a set of adaptively sampled points in space and time domains. These methods still tend to emphasize the motion segmentation, which focus on the low-level information of the pixels. They rarely use the high-level semantic features associated with tracking target. Thus it leads to a low performance of tracking and a high cost of computation.

In this paper, we present a novel approach for multi-target tracking where the scene is captured by a stationary monocular camera. Based on skeleton points assignment (SPA), our approach combines the advantages of feature-based motion segmentation methods and the probabilistic appearance-based particle filter tracking framework, as described next. Initially, a set of sparse points are computed, which we call the skeleton points. Instead of computing point correspondences independently (as done in many feature-based methods), neighboring particles are treated as they were linked, reducing the occurrence of outliers and avoiding the aperture problem. Moreover, the density of skeleton points is adaptive, and denser distributions are used in regions where precision is more important, so we can save computation without neglecting homogeneous regions. To compute point correspondences in a video sequence, we use the approach proposed by Sand and Teller [20]. After the skeleton points are selected, we classify them to different types according to their presence in three successive frames. The classified points are then treated separately with respect to their types. Each point is assigned to an appropriate tracking target, where high-level information, such as appearance, motion and color distribution, is utilized. After that, a pixel-wise dense clustering strategy is utilized, and every foreground pixel in the scene is clustered to the skeleton points. Then, the missing parts of occluded objects are estimated by a blob-based compensation approach. Finally, we get the tracking result in a probabilistic appearance-based particle filter tracking framework.

Compared with the existing methods, our contribution is stated as follows: firstly, we define the matching strategy and the state transition matrix of the skeleton points, to decide their states during tracking process. Secondly, we establish the skeleton points assignment model, based on the high-level features such as appearance, color distribution and motion. Finally, we densely cluster the foreground with skeleton points as kernel, which get the fully segmentation of occlusions with less computational cost. The proposed approach selects the skeleton points by the low-level motion information; meanwhile it treats the assignment problem of skeleton points with the high-level information. The proposed method potentially has the ability to efficiently generate a robust object appearance through complex occlusions, which is adequate for tracking.

This paper is organized as follows. Section 1 discusses the state of the art in the methods of multi-target tracking with occlusions, and contextualizes our work. An overview of the proposed approach is presented in Section 2. In Section 3, we describe the strategy of skeleton points extraction and matching. The estimation of the occlusion states of skeleton points along the video sequence is described in Section 4, as well as the classification method which is used to assign these points to corresponding targets. In Section 5, we present the dense representation for the moving targets. Section 6 introduces a rough compensation approach to estimate the missing parts of the occluded objects. Section 7 presents some experimental results obtained with the proposed approach. Finally, Section 8 summarizes the main contributions of this paper, and discusses some limitations of the proposed approach for future work.

2. Method overview

The structure of the proposed multi-target tracking approach can be divided into three main parts, which is shown in Fig. 1.

1. Background modeling and foreground detection.
2. Occlusion segmentation via skeleton points assignment.
3. Multi-target tracking in a probabilistic appearance-based particle filter tracking framework.

All the stages of the proposed approach are processed sequentially. Every stage is performed for the entire video before going to the next stage.

The first part concerns the foreground detection in the scene. We utilize the Gaussian mixture model [21] to estimate the moving points. This stage takes as input the original video frames, and output a set of foreground points.