Journal of Visual Communication and Image Representation 52 (2018) 45–57

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Multiple-target tracking on mixed images with reflections and occlusions☆

Ting-hao Zhang[a], Chih-Wei Tang[b,*]

[a] HTC Corporation, No. 23, Xinghua Rd., Taoyuan Dist., Taoyuan City 330, Taiwan
[b] Department of Communication Engineering, National Central University, Jhongli 32001, Taiwan

ARTICLE INFO

Keywords:
Multi-target tracking
Reflections
Occlusions
Multi-cue integration
Data association

ABSTRACT

Measurements arose from strong reflections combined with occlusions significantly degrade accuracy of multi-target tracking. Few methods have addressed this problem, and thus this paper proposes a robust multi-target tracker for mixed images with occlusions. For multi-cue integration using co-inference tracking, moving object detection significantly enhances motion cue based correction in the presence of reflections. Target templates are represented by sets of color and spatiality histograms. Joint likelihoods referring to both the target motion trajectory and appearance model of the co-inference fused state are computed. Thus each optimized particle weight with the criteria of maximum joint likelihood is more reliable in the face of reflections and inter-object occlusions. State estimation is achieved with the sample-based data association probability and occlusion confidence indicator. Experimental results show that the proposed tracker outperforms the-state-of-the-art multi-target trackers on images with strong reflections and inter-object occlusions.

1. Introduction

Visual tracking plays an important role in computer vision. For example, video based human pose recovery methods are to generate pleasing and semantically correct human skeletons (e.g., [1,2]). For human pose recovery, tracking helps keep the temporal consistency among poses and improve accuracy of pose estimation [3]. Recently, several issues of visual tracking are raised. For example, some trackers focus on robust long-term tracking (e.g., [4]), where the challenge is to overcome the significant change of appearance models of targets due to deformation, abrupt motion, and heavy occlusions. On the other hand, the rapid rise of deep learning boosts the advance of computer vision. The advantage of deep networks mainly relies on the fact that it can get high-dimensional abstract features from low-level features of images. For visual tracking using deep learning, the pioneer is proposed by Wang et al. [5], and Feng et al. provide a comprehensive review of deep learning based trackers in [6]. Since a mixed image contains both reflections from the transparent surface of the glass and the transmitted scene behind the surface, the appearance model of a target in the mixed images usually changes significantly due to strong reflections. Accordingly, accuracy of state estimation for visual tracking is decreased. Currently, several visual tracking schemes have been designed to solve the reflection interference problem [7–10]. However, the problem will be much more complex in multi-target tracking if the appearance model of the occluded target is partially measurable and both the appearance models of the occluder and the occluded target have been changed by strong reflections combined with inter-object occlusions.

Several occlusion detection methods have been designed for vison based multi-target tracking. In Kwak et al. [11], targets are divided into regular grid cells where each cell is associated with several patches. Based on the feature vector composed of likelihoods of patches, occlusion detection on each cell is realized by classification. Classification will be more reliable if the test image and the training images are captured in the same environment. In [12], the Histogram of Oriented Gradient (HOG)-based motion descriptor of an object is extracted from the frame difference. An object is taken as being occluded if the Hellinger distance of the motion descriptors in the current and previous frame is large enough. Occlusion handling can be completed with motion models, observation models, or data association for single camera based multiple object tracking. In Yang et al. [13], the object trajectory in surveillance videos is assumed to be similar to that of the entire group in the presence of occlusions. Thus, the group is tracked instead of an individual object during occlusions. For occlusion reasoning and handling at the correction stage, the particle filter based tracker in Liu et al. preliminarily locate a target by switching between two modes corresponding to the slight occlusion and severe occlusion, respectively [14]. Then the tracker precisely locates the target based on the similarity of corners between the tentative region and the template. For partial occlusions of dissimilar targets, Xiao et al. compute likelihood of an occluded target using visible parts of targets, and these

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: taylor355109@gmail.com (T.-h. Zhang), cwtang@ce.ncu.edu.tw (C.-W. Tang).

https://doi.org/10.1016/j.jvcir.2018.02.001
Received 26 December 2016; Received in revised form 15 July 2017; Accepted 3 February 2018
Available online 05 February 2018
1047-3203/

parts will be extrapolated to all particles in the subsequent frames [15]. For similar targets with partial occlusions, particle weights are adjusted using the motion direction of the target and the size of the bounding box of each particle [15].

Data association finds the correspondences between multiple measurements and multiple targets for multi-target tracking. Mahler proposes the finite-set statistics (FISST) based data association-free multi-sensory-multitarget Bayes filter in [16] and it has attracted considerable related research in the last decade. In contrast, data association based trackers can be either single-scan based (e.g., joint probabilistic data association filter (JPDAF)) or multi-scan based (e.g., multiple hypothesis tracking (MHT)) [17]. The single-scan schemes make the data association decision sequentially at each time step while the multi-scan schemes may revisit and revise previous decisions when a new set of measurements is available. Thus the former need less intensive computations than the latter. JPDAF uses the Kalman filter for state estimation [18,19]. The work in [20] proposes the JPDAF with merged measurements (JPDAM), allowing one measurement associates to two targets as targets are close. Since the particle filter is applicable to non-linear transitions and non-Gaussian states, the sample-based joint probability data association filter (SJPDAF) extending from the JPDAF provides a sample-based approach for data association and state estimation of multiple targets [21]. For data association coping with occlusions, Shi et al. utilize the occlusion information, derived from the geometry based occlusion detection, to reduce the number of feasible events of JPDAF [22]. In [23], measurements obtained from background subtraction are grouped into clusters by variational Bayesian clustering. In addition to adopting the SJPDAF, both the location and features of clusters are referred to assign clusters to targets.

Currently, several tracking methods have been proposed for mixed images. The single object tracker in [7] extracts the foreground layer (i.e., dynamic layer) by layer separation [24]. In order to alleviate the impact of background and reflection scene on the tracking performance, the single object tracking method in [8] obtains the motion cue of a moving object using the dynamic layer extracted by layer separation [24]. Based on the framework of particle filter with compensated motion model [25], the correction stage optimizes each particle weight using maximum likelihood where the more reliable cue is selected at each time instant. To reduce contamination of edges in the background and reflections on the dynamic layer, camera motion compensation is applied before layer separation in [9]. To maximize the observation likelihood in the presence of reflections, the proposed scheme combines co-inference [26] and maximum likelihood for visual cue integration. Co-inference tracking of multiple modalities uses the structured variational inference and infers the variational parameters of one modality by the other modalities to maximize the observation likelihood [26]. On the other hand, few multi-target tracking methods are dedicated to mixed images with reflections although many works have been proposed for multi-target tracking. In [10], a multi-target tracking scheme for mobile mixed images is proposed based on the framework of multi-cue integration in [9]. The co-inference predicted states are taken as measurements, validated by the sample-based joint probabilistic data association filter (SJPDAF) [21]. Then these measurements are used for computation of particle weights, and state estimation is achieved with the aid of SJPDAF based association probability. The major problem with [10] is that it fails in the presence of reflections combined with occlusions. In addition, the improper object representation of motion cue extracted by inaccurate moving object detections degrades tracking accuracy.

Thus, extended from our preliminary work in [10], this paper proposes a multi-target tracker for mixed images combined with occlusions. The major contributions are stated as follows. (1) With improved moving object detection at the pre-processing stage, accuracy of multi-cue integration based correction is enhanced in the presence of reflections. (2) The proposed spatiality histogram of motion cues greatly enhanced tracking accuracy in the presence of reflections. (3) The proposed occlusion confidence indicator combined with the SJPDAF based association probability improves accuracy of state estimation in the presence of inter-object occlusions. (4) The joint likelihoods that refer to both the target trajectory and the appearance model of the co-inference fused state are maximized to select the more reliable visual cue, improving tracking performance under reflections and occlusions. (5) Target templates are updated to keep track of the slight variation of the target appearance and avoid adapting a target template to a new one contaminated seriously by reflections or occlusions. The remainder of this paper is organized as follows. Section 2 provides an overview of the sample based joint probabilistic data association filter (SJPDAF) [21]. Section 3 proposes a multi-target tracking scheme that combines multi-cue integration, sample based data association, and occlusion reasoning for mixed images with reflections combined with inter-object occlusions. Section 4 analyzes experimental results, and Section 5 concludes this paper.

2. Overview of the sample-based joint probability data association filter

The sample based joint probability data association filter (SJPDAF) [21] combines the advantages of the particle filter [27] and JPDAF [18–20]. The particle filter uses a set of particles with associated weights to approximate the probability density function (pdf) of the state. It can be applied to non-Gaussian, multi-model, and non-linear state estimation. The particle filter recursively implements the Bayesian filter, consisting of two stages: prediction and update [27]. For state estimation of a single target, the prediction stage uses the motion model to predict the prior pdf of the state, $\mathbf{x}_k$, at time k by

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1},\mathbf{u}_{1:k}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{u}_k)p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1},\mathbf{u}_{1:k-1})d\mathbf{x}_{k-1}, \tag{1}$$

where $\mathbf{z}_{1:k-1} = \{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_{k-1}\}$ is a set of measurements up to time k-1, and $\mathbf{u}_{1:k-1} = \{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_{k-1}\}$ is a set of control terms up to time k-1. The update (i.e., correction) stage uses the measurements to compute the posterior pdf by

$$p(\mathbf{x}_k|\mathbf{z}_{1:k},\mathbf{u}_{1:k}) = p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1},\mathbf{u}_{1:k})/p(\mathbf{z}_k|\mathbf{z}_{1:k-1}), \tag{2}$$

where $p(\mathbf{z}_k|\mathbf{z}_{1:k-1})$ is the normalization constant. The weight of the ith sample $\mathbf{x}_k^{(i)}$ is updated by

$$\pi_k^{(i)} = \pi_{k-1}^{(i)}p(\mathbf{z}_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)},\mathbf{u}_{1:k})/q(\mathbf{x}_k^{(i)}|\mathbf{x}_{1:k-1}^{(i)},\mathbf{z}_{1:k}), \tag{3}$$

where the importance density $q(\mathbf{x}_k^{(i)}|\mathbf{x}_{1:k-1}^{(i)},\mathbf{z}_{1:k})$ can be chose to be the prior density [27].

In the presence of data association uncertainty of multi-target tracking, the JPDAF selects measurements for the update stage of the Kalman filter, evaluates the conditional probability of joint association events, and incorporates the association probability into state estimation of the Kalman filter based tracking [18–20]. Assume N_T targets are tracked with the set of states$X(k) = \{x_1(k),\ldots,x_{N_T}(k)\}$ at time k. Let $Z(k) = \{z_1(k),\ldots,z_{m(k)}(k)\}$ denote the set of measurements at time k, where $m(k)$ is the number of measurements. Let Z^k denote the sequence of all measurements up to time k · A validation matrix $\Omega = [\omega_j^t]$ is constructed using the validated measurements [18–20], where $j = 1,\ldots, m(k)$ and $t = 0,1,\ldots N_T\cdot$ ω_j^t indicates whether the jth measurement is inside the validation region of the tth target or not and $\omega_j^t = \{0,1\}$. By summing the probability of all feasible events $\theta(k)$ at time k where feasible events $[\hat{\omega}_j^t(\theta(k))]$ are extracted from the validation matrix, the posterior probability that the jth measurement comes from the tth target is computed by [18–20]

$$\beta_j^t(k) = \sum_{\theta} p(\theta(k)|Z^k)\hat{\omega}_j^t(\theta(k)), \tag{4}$$

Extends from the JPDAF, Schulz et al. proposes the SJPDAF to provide a sampled based approach for data association and uses the particle filter for state estimation of multi-targets [21]. By summing