Pattern Recognition 48 (2015) 812–826

Contents lists available at ScienceDirect

Pattern Recognition

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

Novelty detection in human tracking based on spatiotemporal oriented energies

Ali Emami [a,c,*], Mehrtash T. Harandi [d,b], Farhad Dadgostar [a], Brian C. Lovell [a]

[a] The University of Queensland, School of ITEE, QLD 4072, Australia
[b] NICTA (CRL), Locked Bag 8001, Canberra, ACT 2601, Australia
[c] NICTA (QRL), GPO Box 2434, Brisbane, QLD 4001, Australia
[d] Australian National University, Canberra, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form
14 May 2014
Accepted 4 July 2014
Available online 27 July 2014

Keywords:
Occlusion modeling
Novelty detection in tracking
Spatiotemporal oriented energy
Image motion analysis
Video surveillance

ABSTRACT

Integrated analysis of spatial and temporal domains is considered to overcome some of the challenging computer vision problems such as 'Dynamic Scene Understanding' and 'Action Recognition'. In visual tracking, 'Spatiotemporal Oriented Energy' (SOE) features are successfully applied to locate the object in cluttered scenes under varying illumination. In contrast to previous studies, this paper introduces SOE features for occlusion modeling and novelty detection in tracking. To this end, we propose a Bayesian state machine that exploits SOE information to analyze occlusion and identify the target status in the course of tracking. The proposed approach can be seamlessly merged with a generic tracking system to prevent template corruption (for example when the target is occluded). Comparative evaluations show that the proposed approach could significantly improve the performance of a generic tracking system in challenging occlusion situations.

1. Introduction

This paper introduces a framework to exploit motion dynamics of non-rigid human targets for the purpose of novelty detection and occlusion modeling. Although an extensive amount of research has been dedicated to overcome the well-known challenges of visual tracking such as illumination changes [1], comparatively little attention has been paid to robust schemes for template updating. Template corruption is Achilles' heel in almost any tracking system and our motivation in this paper is to address this under-researched area, especially for video surveillance technology.

Since their initial introduction by Fahle and Poggio [2], Adelson and Bergen [3] and Heeger [4], integrated analysis of videos in spatial and temporal domains is the de facto standard in various computer vision applications. The SOE feature set is an integrated and modern framework, proposed for analysis of dynamic patterns based on their constituent space–time orientation structure in the video. This framework has been successfully applied in various computer vision applications such as 'Dynamic Texture Recognition and Scene Understanding', 'Action Recognition' and 'Visual Tracking', to name a few [5–11].

As for visual tracking (which is our primary concern in this paper), modeling moving objects by SOE features was first explored by Cannons et al. [10–12].[1] Though SOE trackers have been shown to yield good performance in certain realistic situations [10–12], their intrinsic limitations may prove a hindrance in some others.

One example is the tailing effect in SOE features, which is an intrinsic consequence of temporal filtering. Basically the presence of the moving target in various locations of n successive frames appears as an energy tail in the SOE features. This makes the blobs of spatiotemporal energies look larger than the actual target size. Therefore, the target bounding box oscillates around the moving target over time, even for normal and non-occluded situations, which could lead to incorrect or unstable target localization (*cf.* experimental results of the surveillance video in [10,13]).

* Corresponding author at: The University of Queensland, School of ITEE, QLD 4072, Australia.
E-mail addresses: a.emami@uq.edu.au (A. Emami), mehrtash.harandi@nicta.com.au (M.T. Harandi), f.dadgostar@uq.edu.au (F. Dadgostar), lovell@itee.uq.edu.au (B.C. Lovell).

[1] In this paper, we refer to this category of visual tracking systems as SOE trackers for simplicity.

http://dx.doi.org/10.1016/j.patcog.2014.07.004
0031-3203/

Obviously, larger intra-frame target movements produce longer energy tails and higher inaccuracies in localization.

Another problem appears in scenarios where multiple targets move together in the same direction. In such situations the targets' energy blobs may be confused with each other and cause tracking errors. This is mainly according to the similar motion dynamics of the moving targets and the tailing effect of the SOE features. Moreover, if the motion direction of the targets rapidly changes in the video, their SOE based template will be invalid afterwards, due to the slow nature of the updating scheme in [10] and the fast evolution of motion dynamics.

Unlike the previous studies that utilize SOE features as fundamental tracking cues, in this paper we propose a framework based on SOE information for occlusion modeling and detection of tracking novelties.[2] The proposed method makes the template updating mechanism more reliable and prevents template corruption in novelty situations, which leads to significant performance improvement in tracking under challenging occlusion situations. In addition, we propose a Bayesian state machine that enables us to achieve two goals simultaneously. Firstly, it distinguishes non-occluded parts of the target and hence tracking can be done using the non-occluded target pixels. Secondly, it discriminates between the 'Partial Occlusion' and 'Full Occlusion' states of the target. We will see that in 'Full Occlusion' a different strategy for tracking should be used. It is worth mentioning that our proposed framework can be seamlessly fused with various tracking schemes, as a means to protect the target template or the bag of templates.

1.1. Contribution

The proposed occlusion detection system is designed for non-rigid human targets and works based on motion dynamics of the targets distinguished through the 'Spatio-temporal Oriented Energies'. Our contributions in this paper are five-folds:

(i) We propose to utilize SOE features in a 'Bayesian model' to determine state of the target in the course of tracking and discriminate between various types of occlusion, namely 'Partial' and 'Full' occlusions. This is extremely helpful in tracking crowds, since short term occlusions frequently happen in such video surveillance scenarios.
(ii) The tracking strategy in 'Full Occlusion' situation is temporarily changed to a 'tracking by prediction' method, due to the invisibility of the target. Meanwhile the system explores for the SOE based energy blob of the disappeared target in the probable areas.
(iii) A "Novelty Detection" method is suggested based on the established 'Bayesian model' which prevents template corruption in occlusions and drift situations through an adaptive template update mechanism.
(iv) An occlusion model is also proposed to generate a mask and determine the non-occluded target pixels in occlusion situation. This approach enhances the tracking performance in novelty situations, while avoiding the inaccuracies caused by the confusion of similar SOE blobs in SOE trackers.
(v) The proposed framework is extensively evaluated on several publicly available surveillance datasets in short-term/long-term occlusions. We also compared our proposal against other state of the art trackers, to demonstrate the attained gain for video surveillance applications.

To the best of our knowledge this is the first work that utilizes SOE features for the purpose of occlusion analysis. An early version of this study appeared previously [14]. In the current paper, occlusion reasoning is performed within a solid mathematical framework with fewer tuning parameters. Furthermore the system is evaluated against more experimental scenarios from various publicly available datasets to demonstrate the efficacy of the framework.

2. Related work

Handling occlusion in modern visual tracking is not overlooked by any means but the theme adopted in many studies is not a dedicated approach. Generally speaking, occlusion can be handled implicitly or explicitly in a tracker. In implicit solutions, the tracker mostly relies on a discriminative template and is usually able to handle short term partial occlusions [15–27]. On the contrary, in explicit solutions a dedicated mechanism is foreseen to handle and guide tracker during severe occlusion [28–31]. In the following text, we briefly overview some examples from each category.

Ross et al. suggested an incremental visual tracker (IVT) that incrementally updates a low-dimensional subspace representation of the target [15]. This is achieved by utilizing an incremental principal component analysis (PCA) algorithm to adapt the holistic appearance model to lighting and pose variations in the course of tracking. Partial occlusions can be considered to some extent in IVT, by involving a forgetting factor which reduces the effect of new data in constructing the low-dimensional subspace. The larger the forgetting factor, the lower is the sensitivity to short-term occlusions. Hence the system preserves the non-occluded representation for a short period of time. Babenko and Yang [17] proposed to generate a bag of templates from blocks around the current estimation of target and multiple instance learning framework for object tracking (MIL-Track). MIL-Track shows robustness against various appearance changes and partial occlusions to some extent. This is due to capturing multiple positive instances around the target which slightly incorporates the target background in the appearance model. Consequently the surrounding blocks can guide the tracker during the short-term occlusions. Such approaches could result in decent performance for short-term occlusions, but obviously fail in more challenging long-term and heavy occlusions.

A few recent studies adopted sparse representation for visual tracking, where the target appearance is modeled sparsely through a set of templates [20–27]. In these so-called ℓ_1 trackers such as proposed in [20], noise and corruption, occlusion and changes in background are directly addressed by means of the positive/negative trivial templates and the ℓ_1 minimization formulation. After all, occlusion is modeled at pixel level in the ℓ_1 tracker and though it might be handled to some extent (as a result of sparse formulation) a dedicated mechanism for detecting occlusions was not foreseen. Hence, the tracker could still suffer from long-term occlusions, which could introduce occluded templates into the tracker and cause drifting. A 'Structural Local Sparse Appearance Model' is introduced in [22] which exploits the strength of both the sparse representation (ℓ_1 tracker) and the incremental subspace learning (IVT tracker). The so-called IVT-ℓ_1 tracker adapts its positive/negative templates to the changes in target's appearance with an incremental updating strategy. To this end, a linear combination of the PCA basis vectors and the ℓ_1 templates are utilized for modeling the estimated target. To avoid contaminating the target model with occluded frames, the template set is updated by reconstructed target images using only the PCA basis vectors rather than the raw estimated target.

[2] Novelty in this work is considered as partial occlusion events when the occluding objects have different motion dynamics (*cf.* Section 4.3.2).