Computer Vision and Image Understanding 117 (2013) 573–586

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

Segmentation-based tracking by support fusion ☆

Markus Heber *, Martin Godec, Matthias Rüther, Peter M. Roth, Horst Bischof

Institute for Computer Graphics and Vision, Graz University of Technology, Inffeldgasse 16/II, 8010 Graz, Austria

ARTICLE INFO

Article history:
Received 22 August 2012
Accepted 1 February 2013
Available online 10 February 2013

Keywords:
Tracking output fusion
Object segmentation
Iterative refinement

ABSTRACT

In this paper we present a novel fusion framework to combine the diverse outputs of arbitrary trackers, which are typically not directly combinable, allowing for significantly increasing the tracking quality. Our main idea is first to transform individual tracking outputs such as motion inliers, bounding boxes, or specific target image features to a shared pixel-based representation and then to run a fusion step on this representation. The fusion process additionally provides a segmentation, which, in turn, further allows for a dynamic weighting of the specific trackers' contributions. In particular, we demonstrate our fusion concept by combining three diverse heterogeneous tracking approaches that significantly differ in methodology as well as in their reported outputs. In the experiments we show that the proposed fusion strategy can successfully handle highly complex non-rigid object scenarios where the individual trackers and state-of-the-art (non-rigid object and fusion based) trackers fail. We demonstrate high performance on a large number of challenging sequences, where we clearly outperform the individual trackers as well as state-of-the-art tracking approaches.

1. Introduction

In the last decade, visual object tracking has been a vital field of research in computer vision including many applications, such as surveillance, augmented reality, or assistance systems. Tracking generally describes the task of detecting and following an object over a sequence of images, where different strategies such as simple template tracking, salient image feature based tracking, or highly adaptive on-line learning based tracking are used. Each method exhibits specific advantages or invariants, allowing for tracking objects in different real-world scenarios. Facing challenging problems such as non-rigid object transformations, severe appearance changes, abrupt illumination variations, or extremely fast motion, a recent trend is to combine several trackers, where each tracker solves a special facet of the overall problem. Usually, this fusion is either handcrafted and based on heuristics [1,2], or is based on a simple combination of a large number of similar trackers [3–5]. In general, such methods are not satisfactory, because the trackers are either coupled too tightly (e.g., by a cascade [2]) or all trackers are run independently and only a late fusion is applied. Obviously, it is reasonable to combine different tracking cues as the goal is to combine advantages of diverse approaches while compensating for individual weaknesses. However, this is not a trivial task due to diverse typically not combinable outputs. Additionally, a tracking approach should also deliver a segmentation. This allows on the one hand for more precise updates of internal tracking states or models. Thus, adaption of noise and background, e.g., during on-line learning, or its incorporation into templates gets significantly reduced, resulting in more robust and stable tracking. On the other hand, object tracking in video requires for accurate object segmentations as the object, e.g., needs to be accurately extracted from the image in some applications.

In this work, we present a novel method to fuse heterogeneous tracking approaches within a common, pixel-based representation as illustrated in Fig. 1. A mapping of the individual trackers' outputs to a common representation is defined. Based on the performance of the individual contributing approaches, we apply a weighted combination and regularize the fusion results via iterative energy minimization. Finally, the overall result given by a segmentation gets back-propagated to the individual approaches and is used for updating. Thus, the trackers benefit on the one hand from the fusion as the combined advantages allow the individual trackers, e.g., to recover in error cases. On the other hand, the fine-grained object segmentation within each frame allows for more precise state and model updates than by using, e.g., bounding boxes. Fig. 2 illustrates our proposed tracking support fusion framework that combines an arbitrary number of heterogeneous trackers. Additionally, the obtained segmentation fits the tracking result to the current image data, where the individual trackers benefit from the fine-grained description of the object during updating.

☆ This paper has been recommended for acceptance by Y. Aloimonos.

* Corresponding author.
E-mail addresses: mheber@icg.tugraz.at (M. Heber), godec@icg.tugraz.at (M. Godec), ruether@icg.tugraz.at (M. Rüther), pmroth@icg.tugraz.at (P.M. Roth), bischof@icg.tugraz.at (H. Bischof).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.02.001

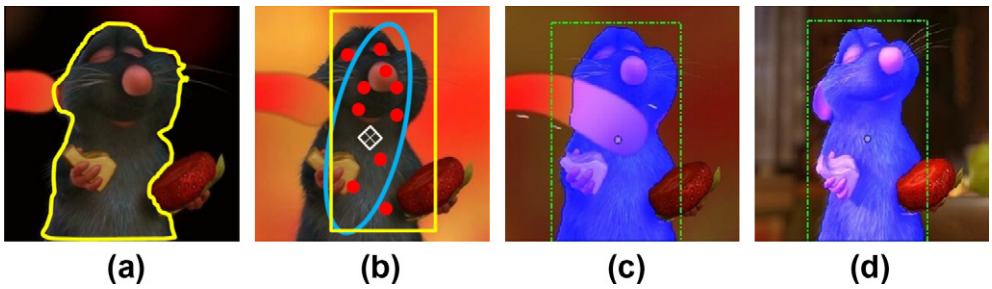

Fig. 1. Tracking support fusion (TSF): Based on an initially delimited region (a) an object is tracked by fusing the diverse outputs of individual trackers (e.g., bounding boxes, covariance ellipses, foreground pixels, object center votes) (b). The object's segmentation (blue overlay) which is then computed from the fused output in each frame (c and d) finally defines the track. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

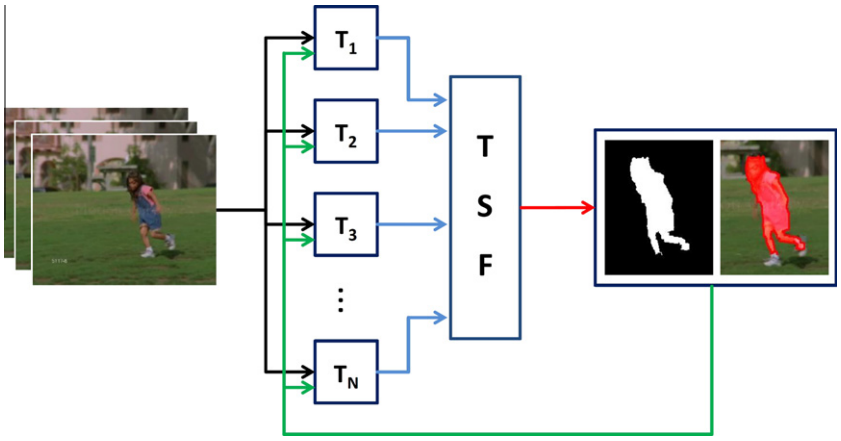

Fig. 2. Tracking fusion framework: The diverse outputs of different trackers $\mathbf{T}_i$ are fused into a common representation, followed by an iterative segmentation. The looped back binary segmentation allows for more precise object model or state updates for each tracker, respectively.

The major difficulty of any fusion method is that these individual tracking results (e.g., center point, bounding rectangle, kernel, or segmentation) strongly differ in their output as well as in the reported confidence values (e.g., probabilistic output, confidence range, error distances, or normalization). On the other hand, all visual tracking approaches have to be coupled to the image domain which we can exploit to define a common representation we refer to as *tracking support*. Tracking support is a quite general concept and defines a set of pixels in the image domain that support the trackers' result and corresponding likelihoods:

- Individual pixels having a high foreground probability.
- Keypoints sharing motion with the object.
- Image patches that are classified as foreground.
- Image regions that share similarities with the object (e.g., texture, color).
- ...

To fuse these results, first each tracker individually finds the object position. The tracking output is then transformed into tracking support sets and subsequently combined. Instead of a simple union of the tracking support sets, the fusion takes the recent performance of the individual trackers into account. The weighted support sets are then utilized within an iterative segmentation procedure. The segmentation determining both, the fusion result and the current congruences of the approaches allows for aligning data and determines the tracking result. Finally, the segmentation result is provided to the individual tracking approaches to ensure high quality updates of the individual states. This design makes the proposed fusion scalable in the number of contributing trackers. Moreover, it is completely parallelizable within the tracking stages while improving the granularity of the final result.

As a proof of concept we demonstrate our fusion framework using three complementary tracking methods, namely a template tracking method that is based on image blending updates, a recently presented discriminative tracking approach that is based on the generalized Hough transform [6], and a well known kernel tracking method based on feature histograms [7]. For these trackers we define the tracking support as projective homography inliers, as back-projected Hough votes that support the actual center object position, and as covariance ellipses around the object center point, respectively. Subsequently, we define the estimation of the object's segmentation as an iterated energy minimization problem that is solved using an extended version of the GrabCut [8] algorithm. In the experiments we show that the proposed fusion approach significantly increases the overall tracking performance on sequences where the individual trackers fail. Furthermore, we obtain a reasonable segmentation of the tracked object within each frame. Consequently, each tracker additionally benefits from the feedback of the fused result, since the given object segmentations allow more precise state and model updates within each tracking iteration. Although, we do not focus on accurate object segmentations or high segmentation quality, ground truth evaluations demonstrate competitive performance to recent methods that especially focus on object segmentation [9,10].

The paper is organized as follows. In Section 2 we recapitulate single target tracking and fusion concepts in computer vision. Our proposed fusion framework as well as the iterative segmentation algorithm are introduced in Section 3. Section 4 summarizes three heterogeneous trackers and their practical combination as