Clustering of Static-Adaptive Correspondences for Deformable Object Tracking

Georg Nebehay[1,2], Roman Pflugfelder[2]
[1]Institute for Computer Graphics and Vision, Graz University of Technology. [2]Digital Safety and Security Department, Austrian Institute of Technology.

This paper deals with the problem of tracking unknown objects (model-free tracking) by formulating a novel keypoint-based approach. As keypoint matching is ambiguous, robust methods such as RANSAC are often used to evaluate the fitness of the matches. Typically, strong assumptions about the motion model are made, of which the rigidity assumption probably is the most common one. However, one of the biggest challenges for tracking are articulated objects, often invalidating this assumption. Aiming at allowing for deformations to be handled, a strand of research has recently emerged in object recognition, studying how to incorporate spatial constraints in addition to photometric constraints [1] into the matching of keypoints. In our work, we adopt similar ideas in the context of model-free object tracking and formulate the following approach.

In each frame t, our aim is to identify the matches $\mathcal{L}_t = \{m_1, \ldots, m_n\}$ that represent the object of interest as accurately as possible. The individual steps of our approach are shown in Figure 1. We employ a static appearance model (top left) that is based solely on the initial appearance of the object, composed of the descriptors around all keypoints x_i^0, which are obtained by an interest point detector. We employ a global search in order to establish matches between keypoints x_i^0 from the initial frame and candidate interest points x_j^t in the current frame t (top center). The static model is robust and handles for instance the re-detection of keypoints after occlusions. However, it does not adapt to new object appearances. In contrast, our adaptive model is updated in every frame, comprising the image patches around all $x_i^{t-1} \in \mathcal{L}_{t-1}$ (top right). By estimating sparse optic flow from frame $t-1$ to frame t, we establish correspondences efficiently by means of a local optimization [4]. We then employ a pairwise dissimilarity measure D between correspondences m_i and m_j based on their geometric compatibility, directly reflecting the deformation of the object of interest. D is then used to partition $\mathcal{L}_t^*$ into subsets by employing a standard agglomerative clustering algorithm using single linkage, where a cutoff threshold δ is used in order to form flat clusters. The parameter δ steers the degree of tolerated deformation, where 0 means complete rigidity. We define D to be

$$D(m_i, m_j) = \left\| (x_i^t - Hx_i^0) - (x_j^t - Hx_j^0) \right\|, \quad (1)$$

where $\|.\|$ denotes the Euclidean distance and H is a similarity transform that is estimated from $\mathcal{L}_t^*$. Note that D is invariant to translations of x_i^t and x_j^t by a common displacement vector. It is therefore sufficient to estimate H up to scale s and rotation α, which is accomplished by two heuristics [2, 5]. We assume that the largest cluster $\mathcal{L}_t^+$ contains the correspondences relevant for the object, while correspondences of all other clusters belong to clutter (bottom left). Similar descriptors appearing on multiple parts of the object or in the background pose a major problem in descriptor matching. Based on $\mathcal{L}_t^+$, we disambiguate correspondences by excluding candidate keypoints that are geometrically dissimilar to $\mathcal{L}_t^+$ in a second matching round (bottom center). Finally, a rotated bounding box is computed using H (bottom right).

For our experiments, we detect and describe interest points by using BRISK [3], due to its invariance to scaling and rotation. For a quantitative assessment of tracking performance we employ the tracking dataset of Vojir et al. that is composed of 77 sequences. The sequences are a compilation of datasets that have been widely used in the evaluation of various tracking approaches. Most of the objects of interest in this dataset are non-rigid, thus rendering it suitable for evaluating our approach. We compare tracker output b_T to ground truth data b_{GT} using the overlap measure

$$\phi(b_T, b_{GT}) = \frac{b_T \cap b_{GT}}{b_T \cup b_{GT}}. \quad (2)$$

We perform a comparison of our method to five state-of-the-art trackers. The result is shown in Figure 2. On the left, the evaluation according to [6] is shown, visualizing the distribution of per-frame overlap measurements. On the right, the evaluation according to [5] is shown, visualizing the per-sequence recall, obtained by employing a threshold of $\phi > 0.5$ on each frame. Our method dominates both evaluations, demonstrating that our algorithm is applicable to a wide variety of object classes and scenes.

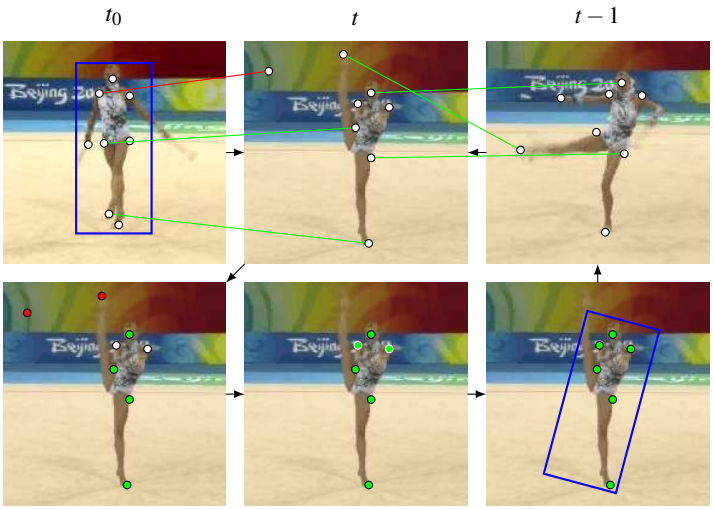

Figure 1: Outline of our approach.

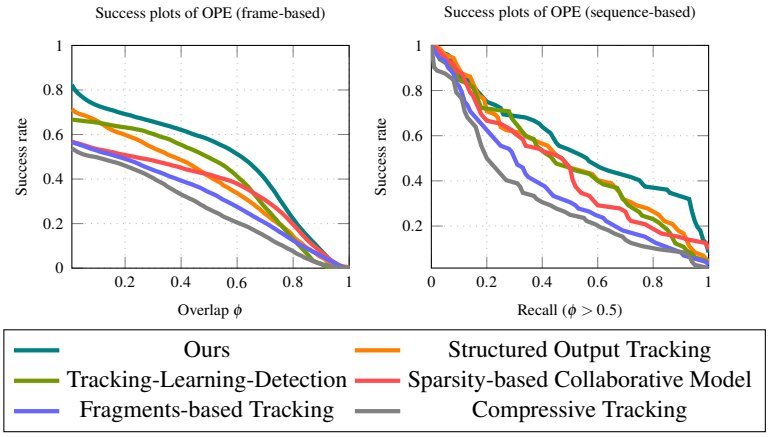

Figure 2: Comparison to five state-of-the-art trackers.

[1] M. Cho, J. Lee, and J. Lee. Feature correspondence and deformable object matching via agglomerative correspondence clustering. In *ICCV*, 2009.

[2] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-Backward Error: Automatic Detection of Tracking Failures. In *ICPR*, 2010.

[3] S. Leutenegger, M. Chli, and R. Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *ICCV*, 2011.

[4] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.

[5] G. Nebehay and R. Pflugfelder. Consensus-based matching and tracking of keypoints for object tracking. In *WACV*, 2014.

[6] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

This is an extended abstract. The full paper is available at the CVF webpage.
The C++ code is available at http://www.gnebehay.com/cmt.
The dataset is available at http://cmp.felk.cvut.cz/~vojirtom/dataset/.