Universal Hough dictionaries for object tracking

Fausto Milletari[1]
fausto.milletari@tum.de
Wadim Kehl[1]
kehl@in.tum.de
Federico Tombari[12]
tombari@in.tum.de
Slobodan Ilic[14]
slobodan.ilic@in.tum.de
Seyed-Ahmad Ahmadi[3]
ahmadi@cs.tum.edu
Nassir Navab[1]
navab@cs.tum.edu

[1] Computer Aided Medical Procedures
Technische Universität München
München, Germany

[2] Computer Science Department (DISI)
University of Bologna
Bologna, Italy

[3] Department of Neurology
Ludwig-Maximilians-Universität
München, Germany

[4] Siemens AG
München, Germany

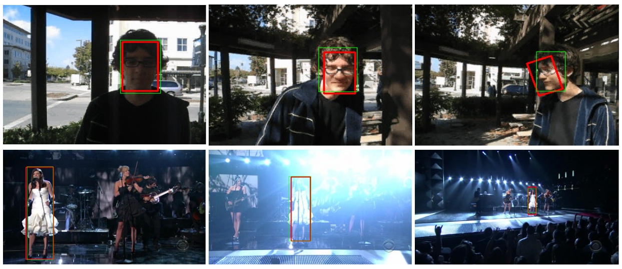

Figure 1: Few qualitative results evaluated on the benchmark dataset [2].

We propose a novel approach to online visual tracking that combines the robustness of sparse coding with the flexibility of voting based methods. Our algorithms is trained offline from a large set of patches extracted from images unrelated to the test sequences. In this way we obtain basis functions, also known as atoms, that can be sparsely combined to reconstruct local image content. During online tracking we adapt the generic knowledge learned by the dictionary to the specific object being tracked, by associating a set of votes and local object appearances to each atom. In each frame of the sequence the object's bounding box position is retrieved through a voting strategy.

1 Method

In the following we describe the three main steps of our approach.
— *Offline Dictionary learning* — we learn a dictionary of visual words from a large set of randomly sampled image patches, with the goal of obtaining a set of basis functions (i.e., *atoms*) capable of reconstructing a large variety of local image appearances. We collect a large set $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_n\}$ of image patches from generic images downloaded from the Internet and we obtain the dictionary $\mathbf{D} = \{\mathbf{d}_1, ..., \mathbf{d}_k\}$ containing k atoms by optimising the following problem with respect to $\mathbf{D}$:

$$\arg\min_{\mathbf{D}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} ||\mathbf{t}_i - \mathbf{D}\alpha_i||_2^2 + \lambda ||\alpha_i||_1. \tag{1}$$

Such *universal* dictionary encodes knowledge acquired from an amount of data that is well beyond what is available to other tracking approaches, which usually rely on single frame initialisation. As a result, our method is capable of reconstructing portions of the target object using a sparse combination of visual words while taking into account the large range of appearances that can be found in real-world situations, as supported by the findings of [1].
— *Tracker Initialisation* — we adapt the generic knowledge captured in the dictionary to the target object. We achieve this by storing votes to the bounding box centroid $\mathbf{c}$ and associated local object appearances, in correspondence to each dictionary atom. The content of the initial, manually placed, bounding box is reconstructed, patch-wise, through the dictionary $\mathbf{D}$. Image patches $\mathbf{p}_i$ are collected from the first frame of the sequence at hand in correspondence of the coordinates $\mathbf{x}_i$ and are reconstructed through $\mathbf{D}$ by solving the l_1-sparse optimisation problem

$$\arg\min_{\alpha_i} \frac{1}{2} ||\mathbf{p}_i - \mathbf{D}\alpha_i||_2^2 + \lambda ||\alpha_i||_1 \tag{2}$$

yielding the sparse coefficients α_i, and the reconstructions $\hat{\mathbf{p}}_i = \mathbf{D}\alpha_i$. Each dictionary atom is associated with a list of votes and appearances that is updated during online tracking. Using the non-zero coefficients of the vector α_i, we identify the atoms that contribute to each reconstruction and we add the votes $\mathbf{v} = \mathbf{c} - \mathbf{x}_i$ and the appearances $\hat{\mathbf{p}}$ to the relative lists.
— *Online Tracking* — We track the object across the sequence by retrieving the position of the bounding box centroid in each frame through the voting strategy. We extract image patches $\mathbf{p_i}$ from the area surrounding the last known position of the bounding box and we reconstruct them using the dictionary $\mathbf{D}$ and solving equation 2. The obtained sparse codes α_i are employed to identify the atoms involved in each reconstruction and retrieve the associated votes $\mathbf{V}_k = [\mathbf{v}_1, ..., \mathbf{v}_N]$ and appearances $\mathbf{A}_i = [\mathbf{a}_1, ..., \mathbf{a}_N]$. Each vote $\mathbf{v}_i$ contributes to the vote map after being weighted by the reciprocal of the sum of squared difference (SSD) between the reconstructions $\hat{\mathbf{p}}_i = \mathbf{D}\alpha_i$ and the appearances $\mathbf{a}_i$. The new bounding box centroid is found by identifying the location of the highest peak in the vote map. The votes and appearances associated with the dictionary atoms are therefore updated.

1.1 Results

Our method exhibits robustness towards occlusions, sudden local and global illumination changes as well as shape changes (Figure 1). We test our method on 50 standard sequences obtaining results comparable or superior to the state of the art, as shown in Figure 2.

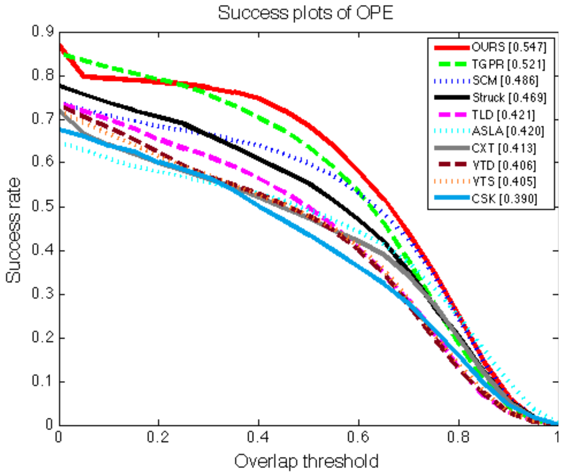

Figure 2: Success plot of our method in comparison with top performing algorithms on the 50 sequences from the CVPR13 Visual Tracking Benchmark [2]. Area under curve (AUC) is reported in brackets.

[1] Ivana Tosic and Pascal Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27–38, 2011.

[2] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.