Single target tracking using adaptive clustered decision trees and dynamic multi-level appearance models

Jingjing Xiao[1], Rustam Stolkin[2], Aleš Leonardis[3]
[1]School of EESE, [2]School of Mechanical Engineering, [3]School of Computer Science, University of Birmingham.

The paper presents a method for single target tracking of arbitrary objects in challenging video sequences. Targets are modelled at three different levels of granularity (pixel level, parts-based level and bounding box level), which are cross-constrained to enable robust model relearning. An overall schematic for the tracker is shown in Fig. 1. The tracker is first propagated by foreground matching at the top level, which generates a candidate image region for the middle level. Next, this candidate region is segmented, [2], into equally sized superpixels. We next propose a continuously-adaptive clustered tree method, which efficiently associates middle level target parts (in the form of patches) from the previous frame, onto candidate superpixels in the new frame, using the minimum number of features needed to achieve a confident match for each such patch. Lastly, to cope with target deformation and appearance changes, old patches are adaptively updated while another decision tree is used to choose the best new superpixels for creating new middle level parts to replace those that have been removed (e.g. due to drifting).

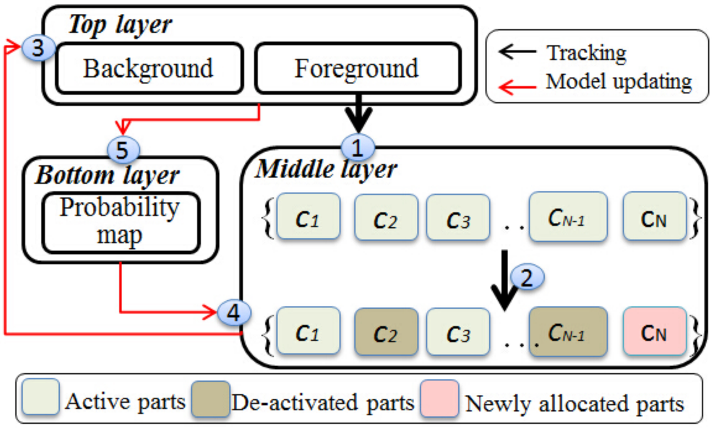

Figure 1: Block diagram of tracking process. 1- tracker propagation by foreground information matching at the top level; 2- middle level matching of target parts to superpixels in the new image, using clustered decision tree procedure; 3- update the top level foreground model by fusing data from all mid-level parts models; 4- update the bottom (pixel) level information using top level models; 5- use updated bottom (pixel-level) model to identify drifting mid-level parts and choose new patches (selected from available superpixels) to replace them.

The main contribution of this work is an adaptive clustered decision tree method, illustrated in Fig. 2. The purpose of this decision tree is to robustly match target parts from the previous frame onto candidate superpixels in the new frame. Each successive row on the tree represents a new kind of feature. If the first feature is sufficiently discriminative to distinguish the candidate superpixel from others, then a match is made, and the tree stops growing. Alternatively, similar scoring superpixels in that feature modality are clustered, and then a different feature modality must be used to grow this cluster into a new row of the tree. This procedure continues until each target part has been confidently matched to a unique superpixel in the new image.

If any candidate superpixel is both i) distinct enough from others that it forms its own leaf and does not lie inside a cluster (e.g. S_6 or S_N in Fig. 2) and ii) strongly matches the target patch, then the decision tree ceases growing and the middle level target part is labelled as matching that candidate superpixel. If, after exhausting all features (tree levels), no match can be found, then that target part becomes regarded as occluded and is temporarily switched off.

Table 1: VOT challenge results: comparing against best 5 trackers

Metrics	VOT 2013 (16 sequences)					
	Ours	PLT13	LGTp	EDFT	FoT	LTFLO
Failures	0	0	1.53	14	22	27.33
Accuracy	0.59	0.58	0.57	0.58	0.63	0.60
	VOT 2014 (25 sequences)					
	Ours	PLT14	DGT	DSST	SAMF	KCF
Failures	1	4	25	29	32	33
Accuracy	0.52	0.56	0.58	0.62	0.61	0.62

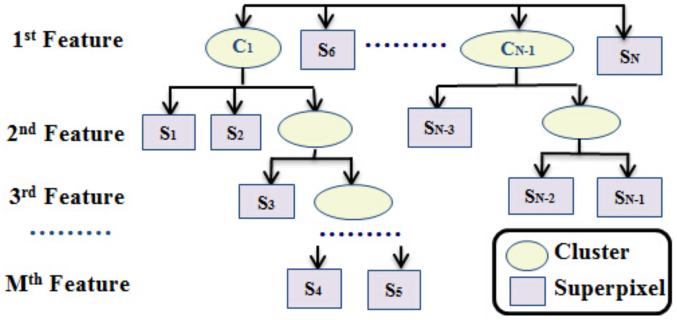

Figure 2: Clustered decision tree. Superpixels form leaves, which are clustered on successive tree-levels, each level corresponding to a different feature.

We first evaluate our tracker using the ICCV2013, and ECCV2014 “VOT challenge", [1], testbeds. Tab. 1 compares the performance of our tracker against the best 5 VOT trackers.

We visualize the results of some challenging sequences in Fig. 3, which shows the superior tracking performance of the tracker while handling the challenges: deformation, background clutter, and occlusion, etc.

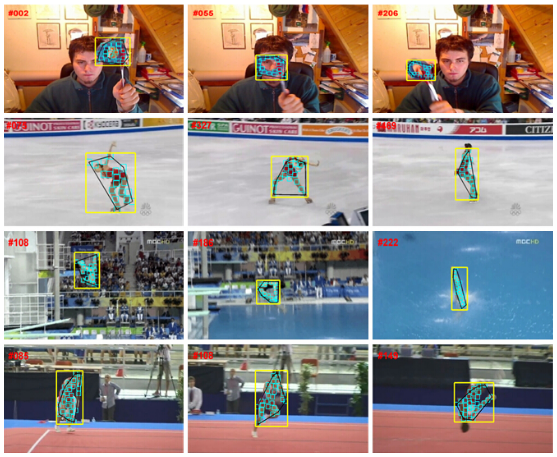

Figure 3: Results of the sequences: *Torus*, *Iceskating*, *Diving*, *Gymnastics*.

[1] The VOT challenge. http://www.votchallenge.net/.

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.