Part-based Visual Tracking with Online Latent Structural Learning: Supplementary Material

Rui Yao[1], Qinfeng Shi[2], Chunhua Shen[2], Yanning Zhang[1], Anton van den Hengel[2]
[1] School of Computer Science, Northwestern Polytechnical University, China
[2] School of Computer Science, The University of Adelaide, Australia
yaorui@mail.nwpu.edu.cn, {javen.shi,chunhua.shen,anton.vandenhengel}@adelaide.edu.au

In this supplementary material, we provide more related work, more details of structured output tracking and online latent structural learning for visual tracking, and more experimental results (both qualitative and quantitative). The experimental results includes demonstration videos, more quantitative results of our tracking algorithm with and without parts in Tab. 2 and Fig. 1, more quantitative comparisons of different part initialisations in Tab. 3 and Fig. 2, visual tracking result on representative frames in Fig. 3 to Fig. 15, and more CLE and VOR plots in Fig. 16 and Fig. 17.

1. More related work

Here we provide more related work which may not fit in the main text. Shu *et al.* [?] propose a part-based multiple-person tracker to handle partial occlusions. They only consider three possible subsets of parts: *head only, upper body parts*, and *all body parts*, which restricts its application to our general object tracking case where most objects are not human. Also data association plays an important role in their case, whereas it is not applicable to our single object tracking case. Similarly, Yang and Nevatia [?] propose a part-based multiple-person tracker, where they use rigid parts to model human: 1 part for *head*, 6 parts for *upper body* and 8 parts for *lower body*. Instead of learning the locations of the parts, they fix the locations, and focus on learning data association (as for multiple-person tracking). Overall, both [?] and [?] use parts specifically designed for human bodies, and they much emphasise on data association for multiple-person tracking, thus they are not applicable to our single general object tracking.

2. Demonstration videos

We show all tracking results of our tracker and competing trackers over all frames on thirteen sequences. In our experiments, we ran nine competing trackers, for clarity only the results of five trackers are marked in the videos. We also marked the part boxes (with index number) of our tracker on the demonstration videos.

3. More details of online latent structural learning for visual tracking

In this section, we show more details of visual tracking via online latent structural learning. In Section 3.1, we will show the overall procedure of "tracking-by-detection" approach with structured output learning, and in Section 3.2, we will describe the latent structural SVM for visual tracking and how to get the objective function for online learning shown in the main text of this paper.

3.1. Tracking-by-detection with structured output learning

The proposed algorithm is a novel tracking-by-detection approach to track generic objects. To give readers a better understanding of this kind of approach, we describe the overall procedure of visual tracking with structured output learning in this section. There are several papers related to structured output tracking. Hare *et al.* [?] applied structured learning to online visual tracking, which builds upon its previous successful application to object detection [?]. Yao *et al.* [?] use weighted online structural learning to deal with the inevitable changes in target appearance over time.

A structural tracking-by-detection approach maintain a structured output classifier trained online to distinguish the target object from the background. During tracking, the prediction of object location is obtained by the combination of the object

Experimental sequence	Demonstration video of tracking result
faceocc1	faceocc1.mp4
faceocc2	faceocc2.mp4
threemen	threemen.mp4
fskater	fskater.mp4
dollar	dollar.mp4
david	david.mp4
trellis	trellis.mp4
board	board.mp4
sylvester	sylvester.mp4
girl	girl.mp4
coke	coke.mp4
tiger1	tiger1.mp4
tiger2	tiger2.mp4

Table 1. The illustration of demonstration videos.

location estimated in previous frame and the offset of object location in current frame, where the offset is estimated by searching for the maximum classification score of candidate offsets using the structured output classifier.

Suppose at the t-th frame, the object location is represented by bounding box $B_t = (c_t, r_t, w_t, h_t)$, where c_t, r_t are the column and row coordinates of the upper-left corner, and w_t, h_t are the width and height. The offset is $\mathbf{y}_t = (\Delta c_t, \Delta r_t, \Delta w_t, \Delta h_t) \in \mathcal{Y}$. The bounding box on the target object is given in the 1st frame of the video, and the tracker is then required to track the object from the 2nd frame to the end of the video. Given the t-th frame $\mathbf{x}_t \in \mathcal{X}$, and the bounding box B_{t-1} in the $(t-1)$-th frame, the bounding box B_t in the t-th frame can be obtained via

$$B_t = B_{t-1} + \mathbf{y}_t^*, \;\; \mathbf{y}_t^* = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}_t, \mathbf{y}). \tag{1}$$

Here $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the discriminant function, it can be learned via structured SVM [**?**] with $\mathbf{y} \neq \mathbf{y}_t$ generated randomly in the vicinity of the true $\mathbf{y}_t$ by optimising

$$\begin{aligned}
&\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{t=1}^{T} \xi_t && (2)\\
&\text{s.t. } \forall t : \xi_t \geq 0 \; ;\\
&\quad \forall t, \mathbf{y} \neq \mathbf{y}_t : \langle \mathbf{w}, \Psi(\mathbf{x}_t, \mathbf{y}_t) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_t, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_t, \mathbf{y}) - \xi_t,
\end{aligned}$$

where the label cost is

$$\Delta(\mathbf{y}_t, \mathbf{y}) = 1 - \frac{(B_{t-1} + \mathbf{y}_t) \bigcap (B_{t-1} + \mathbf{y})}{(B_{t-1} + \mathbf{y}_t) \bigcup (B_{t-1} + \mathbf{y})}, \tag{3}$$

which was introduced in [**?**] to measure the VOC bounding box overlap ratio. In practice, $\mathbf{y}$'s are uniformly sampled in the vicinity of the $\mathbf{y}_t$ in [**?**].

As mentioned above, visual tracking is an online process which deals with a dynamic stream data. It is necessary for tracking algorithm to maintain and update the parameter for adapting new data. In [**?**, **?**], they updated the dual variables online following the work of [**?**, **?**]. Following the formulation in [**?**], Eq. (2) can be converted as:

$$\begin{aligned}
&\max_{\boldsymbol{\beta}} - \sum_{t,\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}_t) \beta_t^{\mathbf{y}} - \frac{1}{2} \sum_{t,\mathbf{y},s,\bar{\mathbf{y}}} \beta_t^{\mathbf{y}} \beta_s^{\bar{\mathbf{y}}} k((\mathbf{x}_t, \mathbf{y}), (\mathbf{x}_s, \bar{\mathbf{y}})) && (4)\\
&s.t. \; \forall t, \forall \mathbf{y} : \beta_t^{\mathbf{y}} \leq \delta(\mathbf{y}, \mathbf{y}_t) C;\\
&\quad \forall t : \sum_{\mathbf{y}} \beta_t^{\mathbf{y}} = 0,
\end{aligned}$$

Algorithm 1: Tracking-by-detection with structured output learning

```
Input: C, searching radius $s$ for tracking, searching radius $r$ for training, the bounding box of target object in the first
    frame $B_1$, the frame set $\{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$
Output: $B_1, \cdots, B_T$
Initialisation
$S_0 = \emptyset$, $B_0 = B_1$;
for $t = 1, 2, \cdots, T$ do
    Estimate the bounding box $B_t$
    $B_t = B_{t-1} + \mathbf{y}_t^*$;
    $\mathbf{y}_t^* = \text{argmax}_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_t, \mathbf{y})$, where $\mathcal{Y} = \{(\Delta c_t, \Delta r_t, 0, 0)|(\Delta c_t)^2 + (\Delta r_t)^2 < s^2\}$, the offsets $\mathbf{y}$s are sampled via
    searching $\mathcal{Y}$ exhaustively;
    Update the dual variables and the set of support vectors $S_t$
    Sample a set of example pairs $\{\mathbf{x}_t, \mathbf{y}_{t,i}\}_{i=1}^N$, where $\mathbf{y}_{t,i}$ is sampled via searching
    $\{(\Delta c_t, \Delta r_t, 0, 0)|(\Delta c_t)^2 + (\Delta r_t)^2 < r^2\}$ on a polar grid.
    Perform OLaRank algorithm on $\{\mathbf{x}_t, \mathbf{y}_{t,i}\}_{i=1}^N$ using $S_{t-1}$ and the dual variables estimated in the $(t-1)$-th frame.
end
```

where $\delta(\mathbf{y}, \mathbf{y}_t)$ is 1 when $\mathbf{y} = \mathbf{y}_t$ and 0 otherwise, and $k(\cdot, \cdot)$ is the kernel function. The discriminant function f becomes

$$f(\mathbf{x}_t, \mathbf{y}) = \sum_{s,\bar{\mathbf{y}}} \beta_s^{\bar{\mathbf{y}}} k((\mathbf{x}_s, \bar{\mathbf{y}}), (\mathbf{x}_t, \mathbf{y})). \tag{5}$$

OLaRank algorithm introduced in [**?**] is used to online update the dual variables and the set of *support vectors* S_t at the t-th frame, where *support vectors* refers to those pairs $(\mathbf{x}_t, y)$ for which $\beta_t^{\mathbf{y}} \neq 0$. OLaRank is an SMO-style [**?**] optimisation method, it use three basic operations. At each iteration, the OLaRank algorithm perform one PROCESSNEW step. Then perform ten REPROCESS step, which is the combination of one PROCESSOLD step followed by ten OPTIMIZE steps. We refer the readers to [**?**, **?**] for more details of OLaRank algorithm.

By using the above-mentioned online structured output learning, a visual tracking algorithm can be obtained. We assume $B_0 = B_1$ *i.e.* $\mathbf{y}_1 = (0, 0, 0, 0)$ and sample some $\mathbf{y} \neq \mathbf{y}_1$. The sampled $\mathbf{y}$s, and $\mathbf{y}_1$, are provided to our online learner to update the dual variables and the set of *support vectors* S_1. We then use this dual variables to predict $\mathbf{y}_2$ to obtain tracking result for the 2nd frame *i.e.* the bounding box B_2. We then take the predicted $\mathbf{y}_2$ as the true label, sample $\mathbf{y} \neq \mathbf{y}_2$, and feed them into the online learner to update dual variables and S_2, and so on. Algorithm 1 shows the complete tracking procedure.

3.2. Latent structural SVM in primal form

In this section, we show how to get the unconstrained form of latent pegasos defined by Eq. (8) in the main text of this paper. We use $\mathbf{z}_t = (\mathbf{z}_t^1, \cdots, \mathbf{z}_t^M) \in \mathcal{Z}$ to represent the offsets of M part boxes at the t-th frame. $\mathbf{z}_t$ can be treat as a latent variable, which is not provided by user. Here we use the notation in the main text of this paper. Latent variables have been used with SVMs previously [**?**, **?**, **?**] in a batch learning scenario.

We incorporate latent variables into structural SVM, and learn the discriminant function $f : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ defined by Eq. (3) in the main text via latent structural SVM. Since visual tracking is an online process, we propose latent pegasos to allow the use of latent variables in an *online* manner. Suppose we have T many frames, and we sample N many offsets at each frame *i.e.* $\{\mathbf{y}_{t,i} \neq \mathbf{y}_t\}_{i=1}^N$, the discriminant function f can be learnt via:

$$\begin{aligned} &\min_{\mathbf{w}} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\xi_{t,i} \\ &s.t.\ \forall t, i : \xi_{t,i} \geq 0, \\ &\quad \forall t, i, \max_{\mathbf{z}} f(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}) - \max_{\mathbf{z}'} f(\mathbf{x}_t, \mathbf{y}_{t,i}, \mathbf{z}') \geq \Delta(\mathbf{y}_t, \mathbf{y}_{t,i}) - \xi_{t,i}. \end{aligned} \tag{6}$$

However, the above is a batch learning process; whereas in tracking we need an online learning algorithm. Same as typical online learning settings in [**?**], at each iteration/frame, the algorithm only has access to the samples from the current frame. Denote $\mathbf{w}_t$ the parameter that predicts $\mathbf{y}_t$ for the t-th frame *i.e.* $\mathbf{y}_t = \text{argmax}_{\mathbf{y}} \max_{\mathbf{z}} f(\mathbf{x}_t, \mathbf{y}, \mathbf{z}; \mathbf{w}_t)$, we sample offsets

$\{\mathbf{y}_{t,i} \neq \mathbf{y}_t\}_{i=1}^{N}$ near the predicted $\mathbf{y}_t$. The online learning algorithm is supposed to update the parameter $\mathbf{w}_{t+1}$. By converting Eq. (6) at t-th frame into unconstrained form, we have the following (mini-batch) online learning objective similar to [**?**],

$$\min_{\mathbf{w}} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{N}\sum_{i=1}^{N}\left[\Delta(\mathbf{y}_t,\mathbf{y}_{t,i}) + \max_{\mathbf{z}'}\langle \mathbf{w}, \Phi(\mathbf{x}_t,\mathbf{y}_{t,i},\mathbf{z}')\rangle - \max_{\mathbf{z}}\langle\mathbf{w},\Phi(\mathbf{x}_t,\mathbf{y}_t,\mathbf{z})\rangle\right]_+. \tag{7}$$

Eq. (7) is the same as Eq. (8) in the main text of this paper for estimating $\mathbf{w}_{t+1}$ at t-th frame. The proposed online learning algorithm uses gradient descent to update $\mathbf{w}_{t+1}$ with an appropriate step size η_t as described in the main paper. From *Lemma 1* and *Lemma 2* in [**?**], we know that for appropriate step sizes, the online learning method using Eq. (7) as objective has a well known online regret bound of $O(\log T)$, since the objective function in Eq. (7) is Lipschitz (*i.e.* bounded (sub)gradient) and λ-strongly convex.

4. Evaluation of our tracking algorithm with and without part

In this section, we show more results of our tracking algorithm with and without part. Tab. 2 shows the VOC overlap ratio(VOR), centre location error (CLE) and success rate (SR) of three methods on sequence *board, david, sylvester, fskater*. Fig. 1 shows the VOR plots of tracking result on those four sequences.

	board			david			sylvster			fskater		
	VOR	CLE	SR	VOR	CLE	SR	VOR	CLE	SR	VOR	CLE	SR
Struck Linear	0.59	72.1	0.65	0.36	60.1	0.38	0.66	10.7	0.72	0.73	11.7	0.99
Ours without part	0.63	45.0	0.70	0.40	53.7	0.34	0.70	8.6	0.75	0.68	12.9	0.87
Ours with part	**0.83**	**14.1**	**0.98**	**0.81**	**7.6**	**1.00**	**0.76**	**5.7**	**0.95**	**0.81**	**7.9**	**1.00**

Table 2. Performance of the proposed tracking with/without part and Struck with linear kernel on four sequences.

5. Evaluation of different part initialisations

In this section, we show more results of our tracking algorithm with and different part initialisations. Tab. 2 shows the VOC overlap ratio(VOR), centre location error (CLE) and success rate (SR) of three methods on sequence *sylvester, fskater*. Fig. 1 shows the VOR plots of tracking result on those two sequences.

	sylvster			fskater		
	VOR	CLE	SR	VOR	CLE	SR
Initialise Part Manually	**0.76**	5.7	0.95	**0.81**	**7.9**	**1.00**
Initialise Part Automatically	**0.76**	**5.8**	**0.97**	0.80	8.0	**1.00**

Table 3. Performance of the proposed tracking with different part initialisations on two sequences.

6. Qualitative visual tracking results

In this section, we show the qualitative visual tracking results of the competing trackers over several representative frames of thirteen sequences in Fig. 3 to Fig. 15. The frame index numbers of these representative frames are drawn in the top-left corner of image.

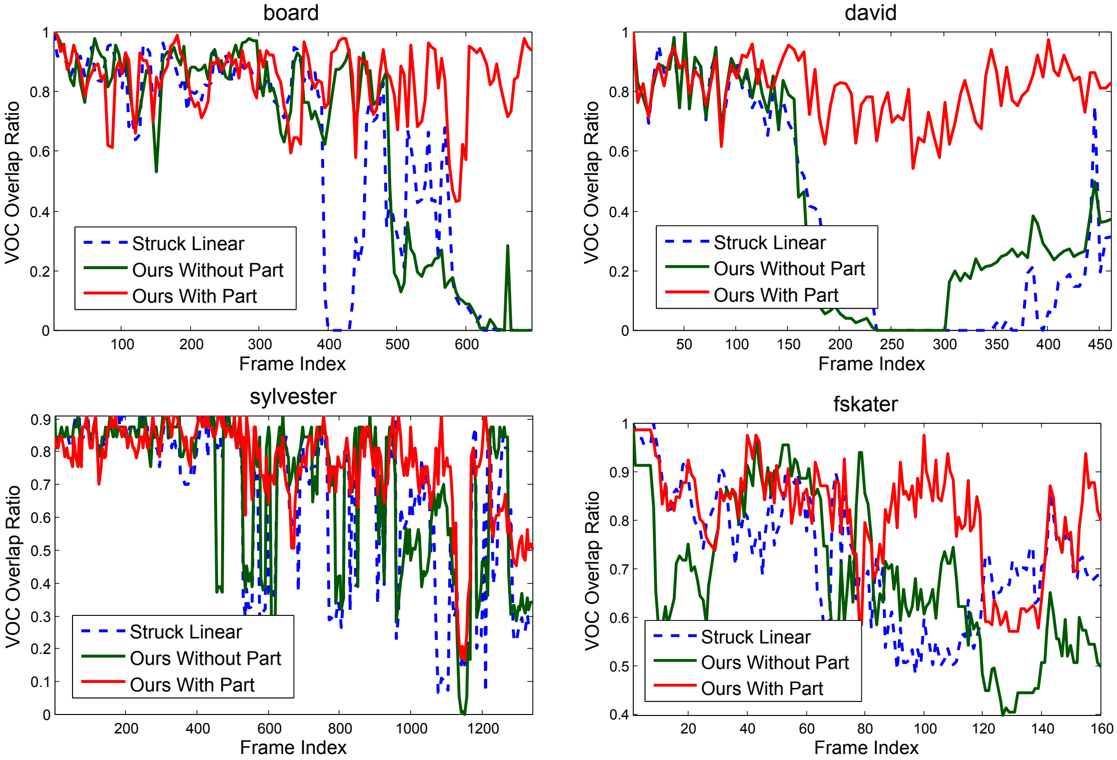

Figure 1. Evaluation of the proposed tracking with/without part and Struck with linear kernel in VOC overlap ratio plots on four sequences.

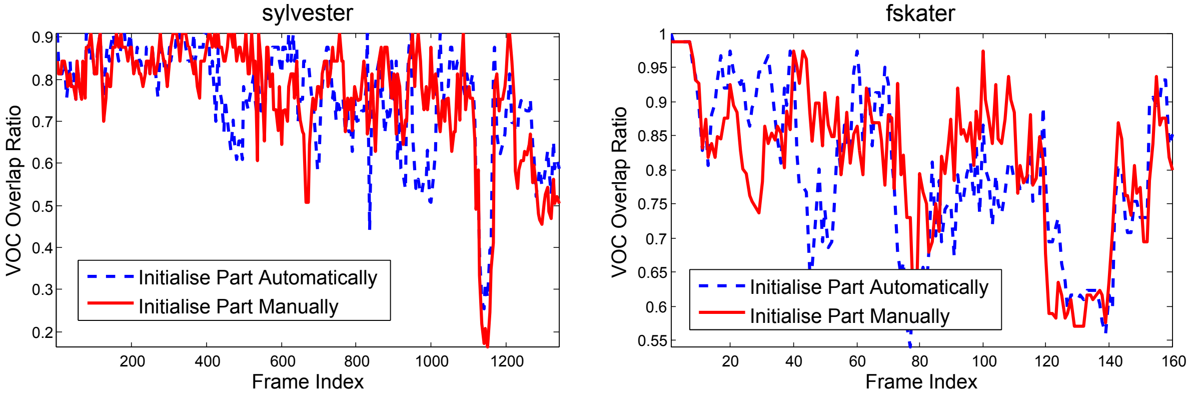

Figure 2. Evaluation of different part initialisations in VOC overlap ratio plots on two sequences.

7. Quantitative Comparison of competing tracking algorithms

In this section, we show the frame-by-frame CLE and VOR plots of eleven tracking algorithms on thirteen sequences in Fig. 16 and Fig. 17.

References

[1] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *Proc. ECCV*, volume 5302, pages 2–15, 2008.

[2] A. Bordes, L. Bottou, P. Gallinari, and J. Weston. Solving multiclass support vector machines with larank. In *Proceedings of International Conference on Machine Learning*, pages 89–96, 2007.

[3] A. Bordes, N. Usunier, and L. Bottou. Sequence labelling svms trained in one pass. In *Proc. ECML/PKDD*, pages 146–161, 2008.

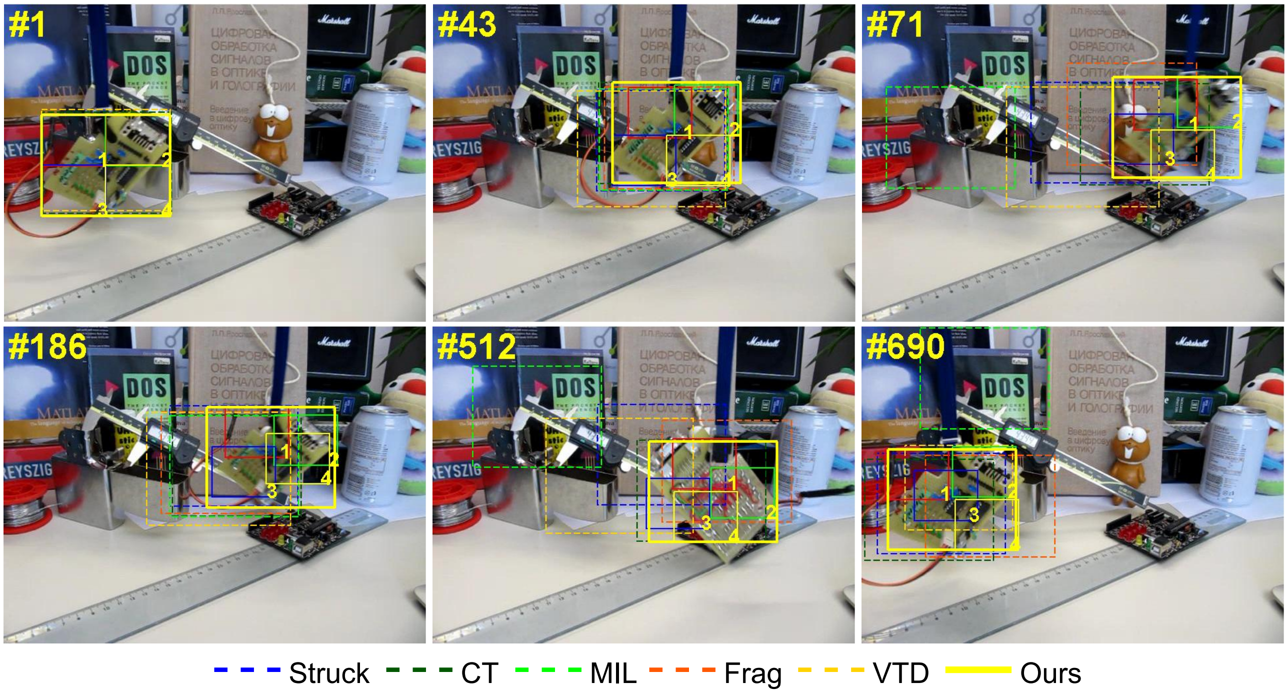

Figure 3. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *board* sequence. Note that the number marked on small rectangles represents the index of object parts.

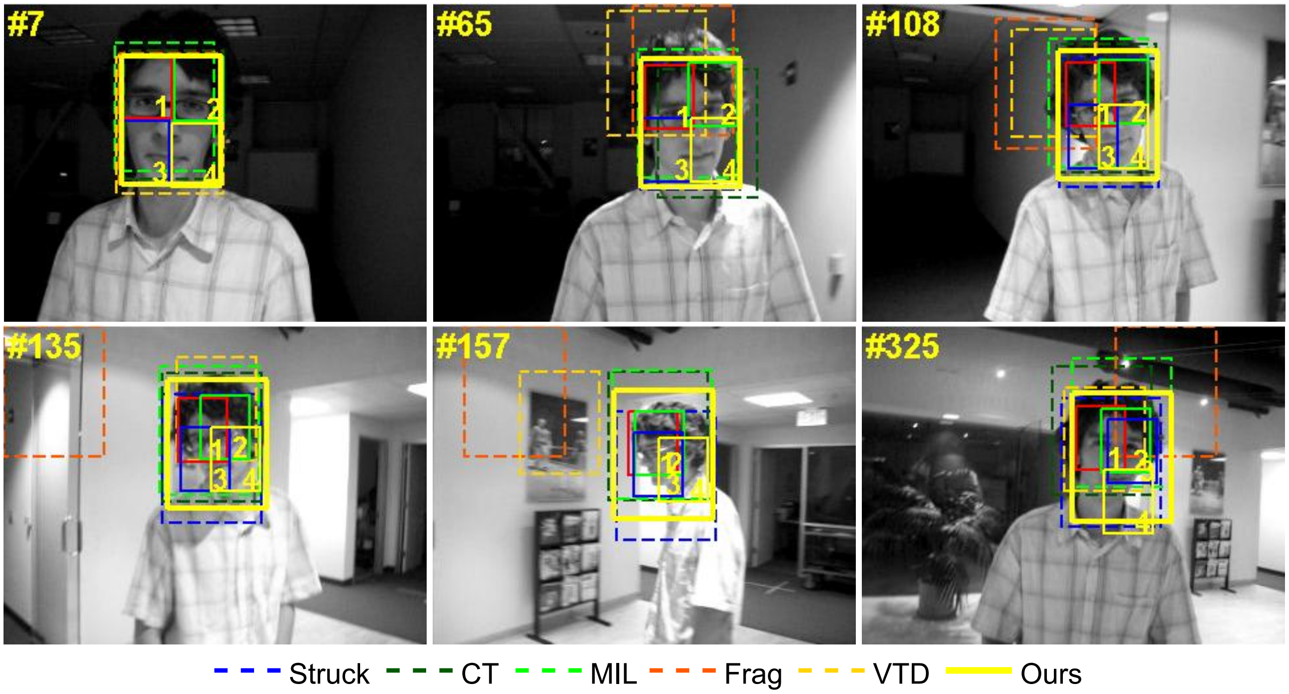

Figure 4. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *david* sequence. Note that the number marked on small rectangles represents the index of object parts.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

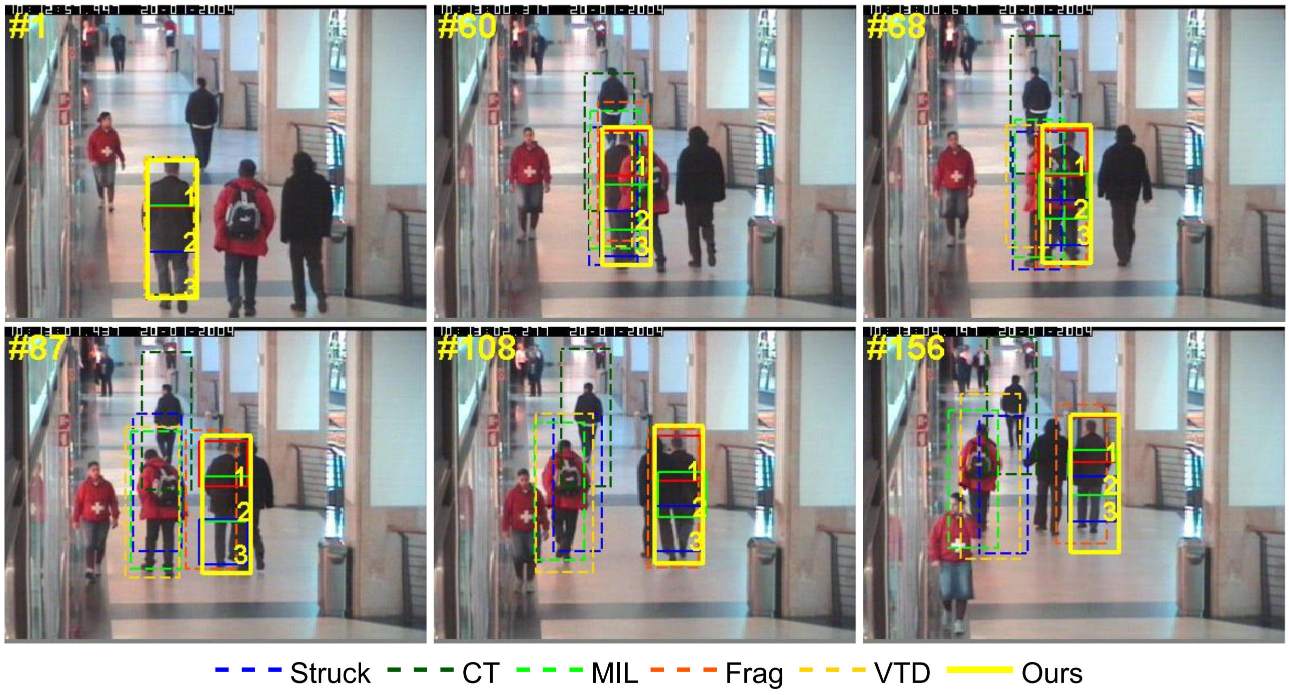

Figure 5. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *threemen* sequence. Note that the number marked on small rectangles represents the index of object parts.

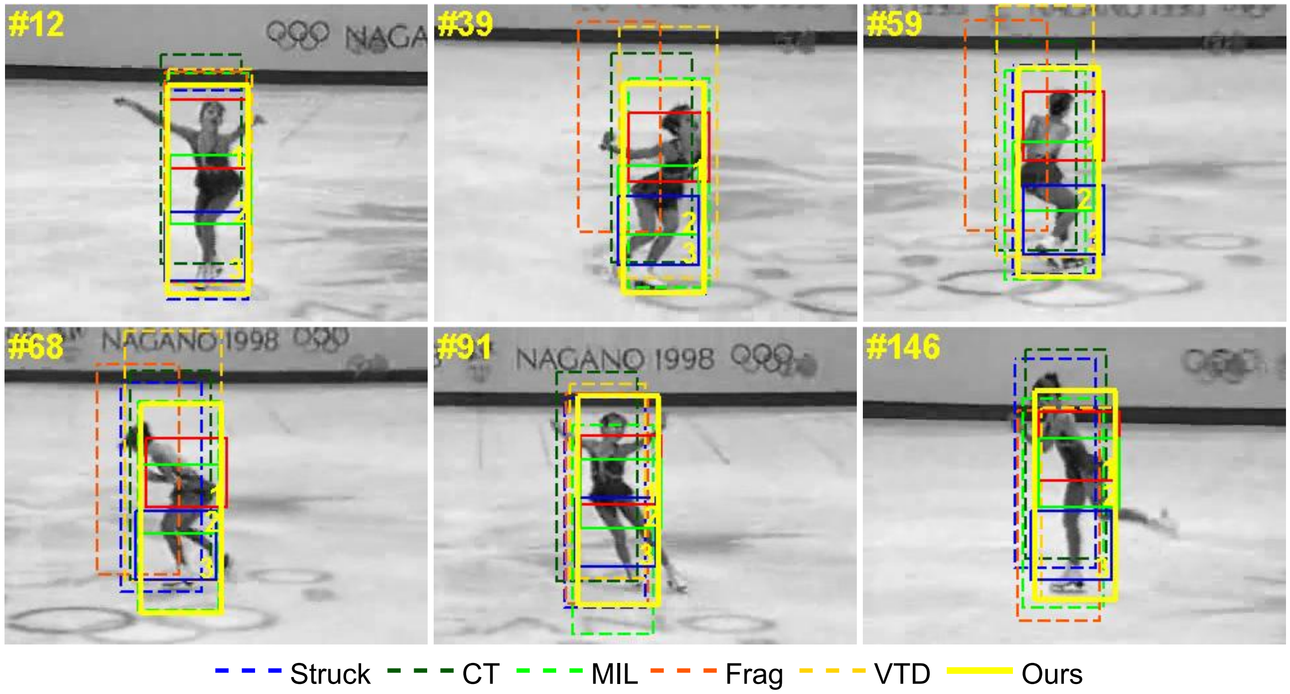

Figure 6. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *fskater* sequence. Note that the number marked on small rectangles represents the index of object parts.

[5] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained

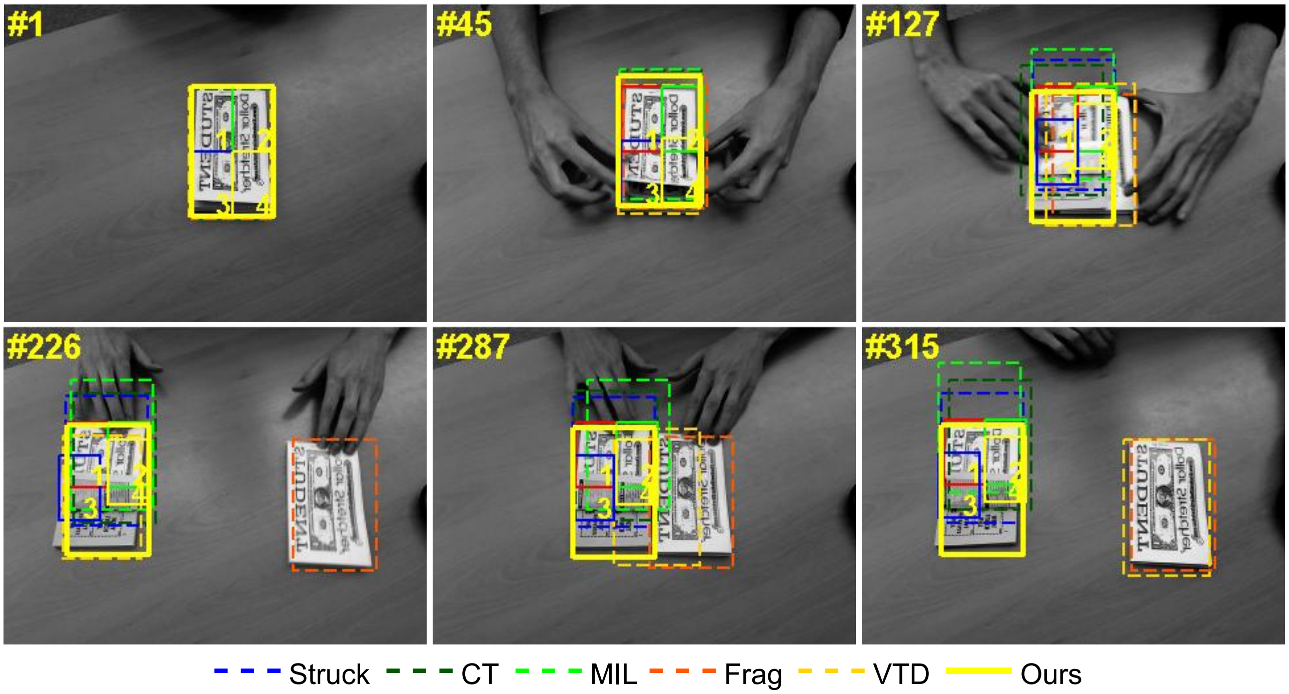

Figure 7. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *dollar* sequence. Note that the number marked on small rectangles represents the index of object parts.

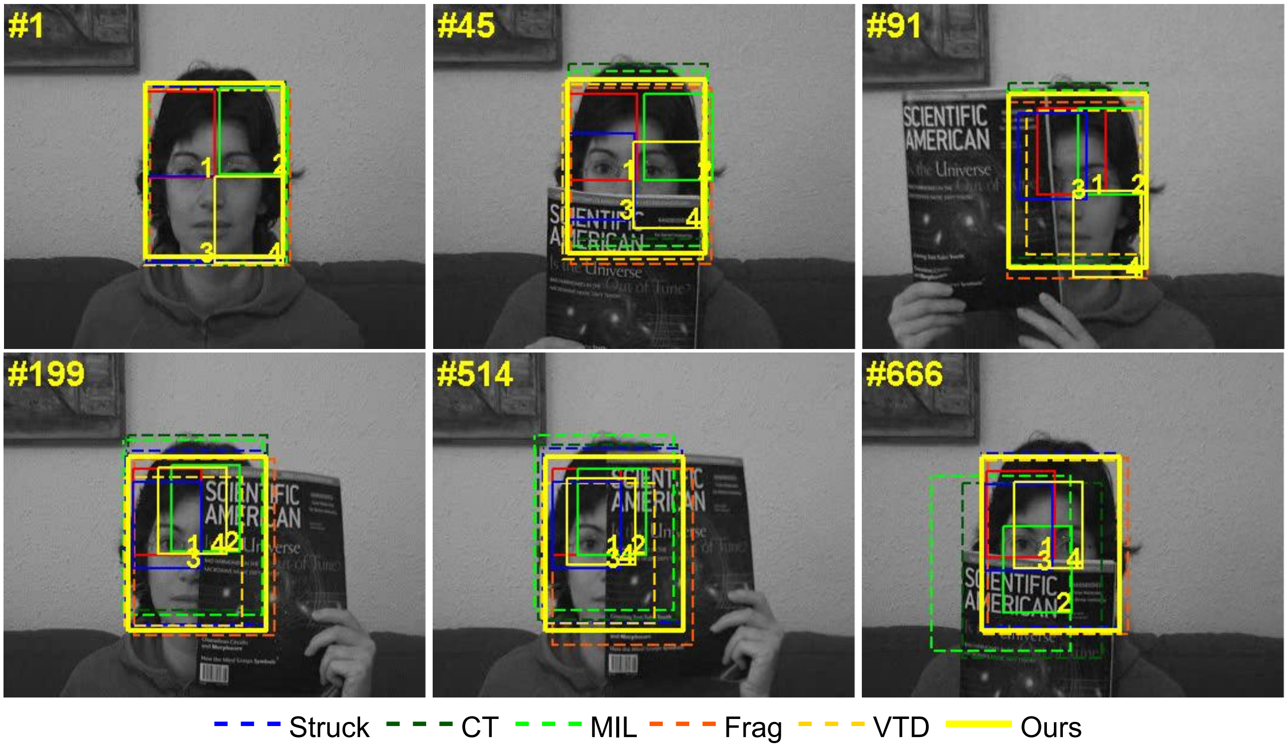

Figure 8. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *faceocc1* sequence. Note that the number marked on small rectangles represents the index of object parts.

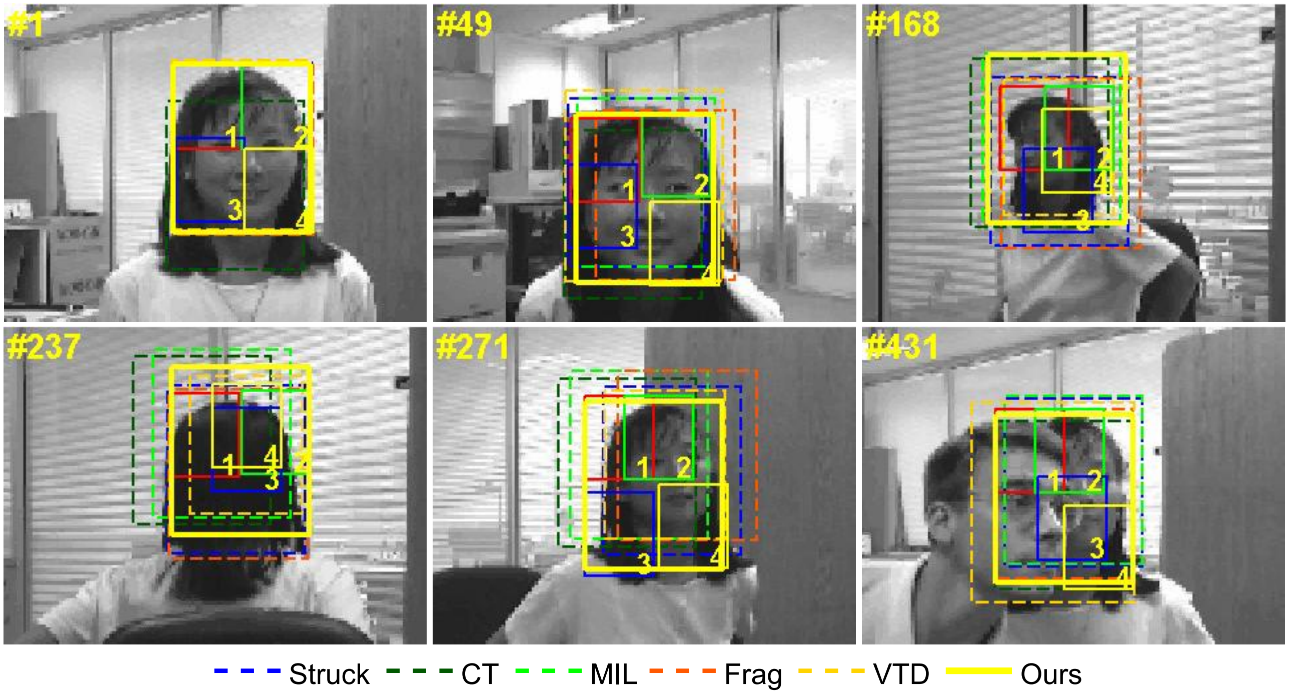

Figure 9. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *gril* sequence. Note that the number marked on small rectangles represents the index of object parts.

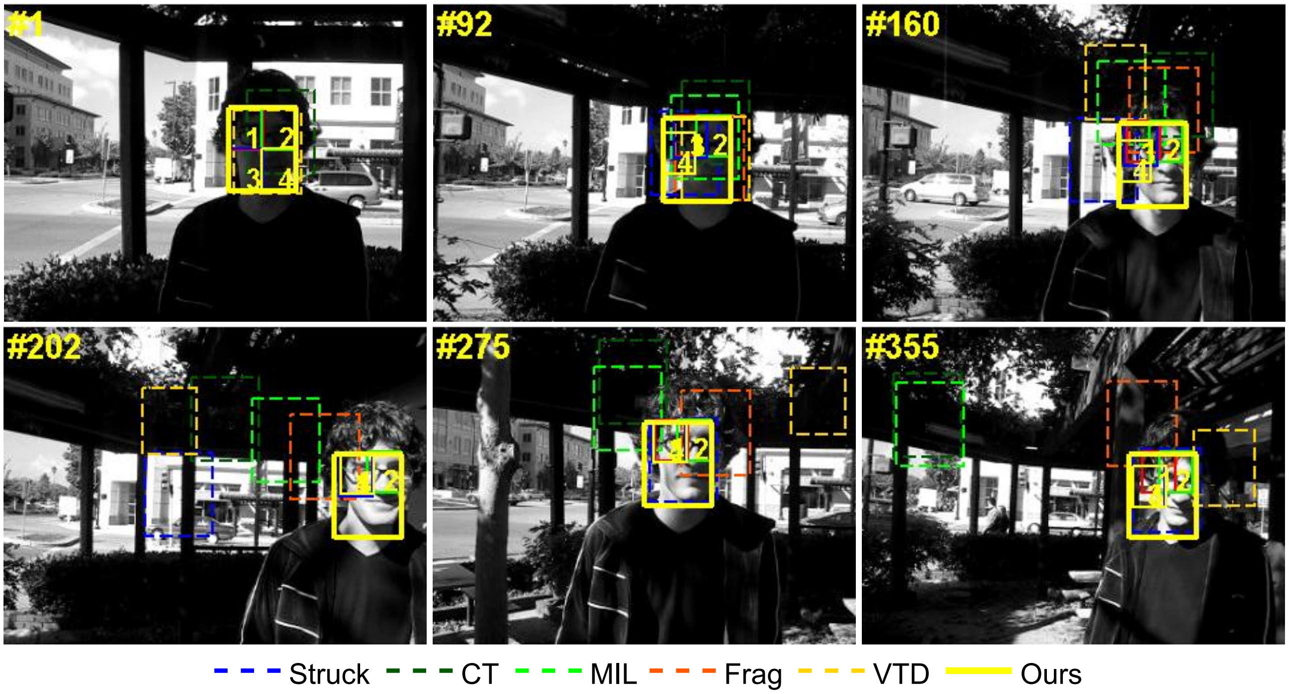

Figure 10. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *trellis* sequence. Note that the number marked on small rectangles represents the index of object parts.

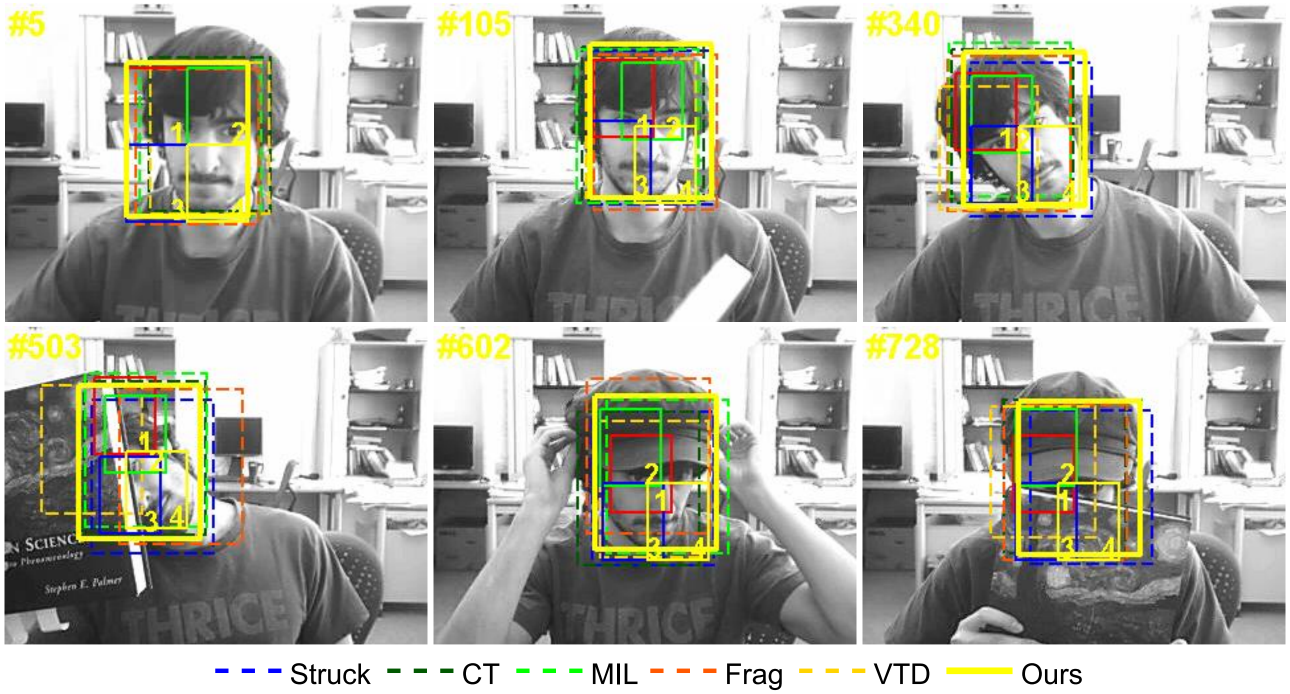

Figure 11. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *faceocc2* sequence. Note that the number marked on small rectangles represents the index of object parts.

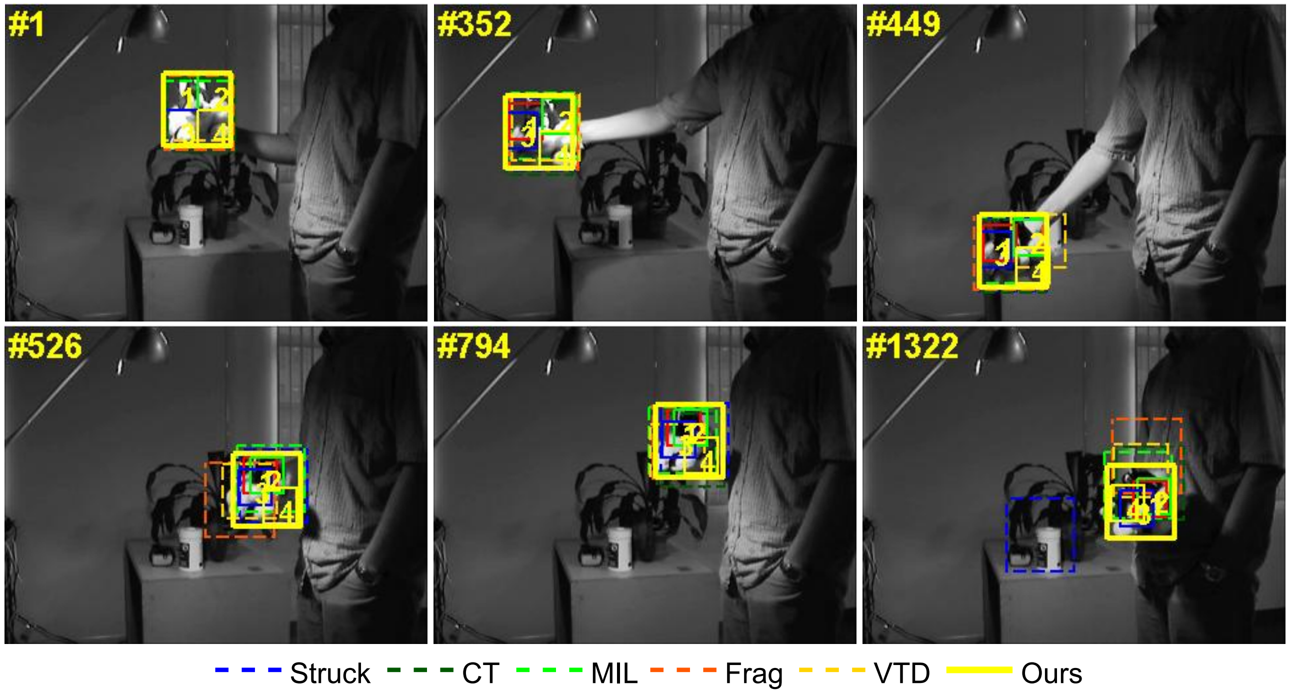

Figure 12. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *sylvester* sequence. Note that the number marked on small rectangles represents the index of object parts.

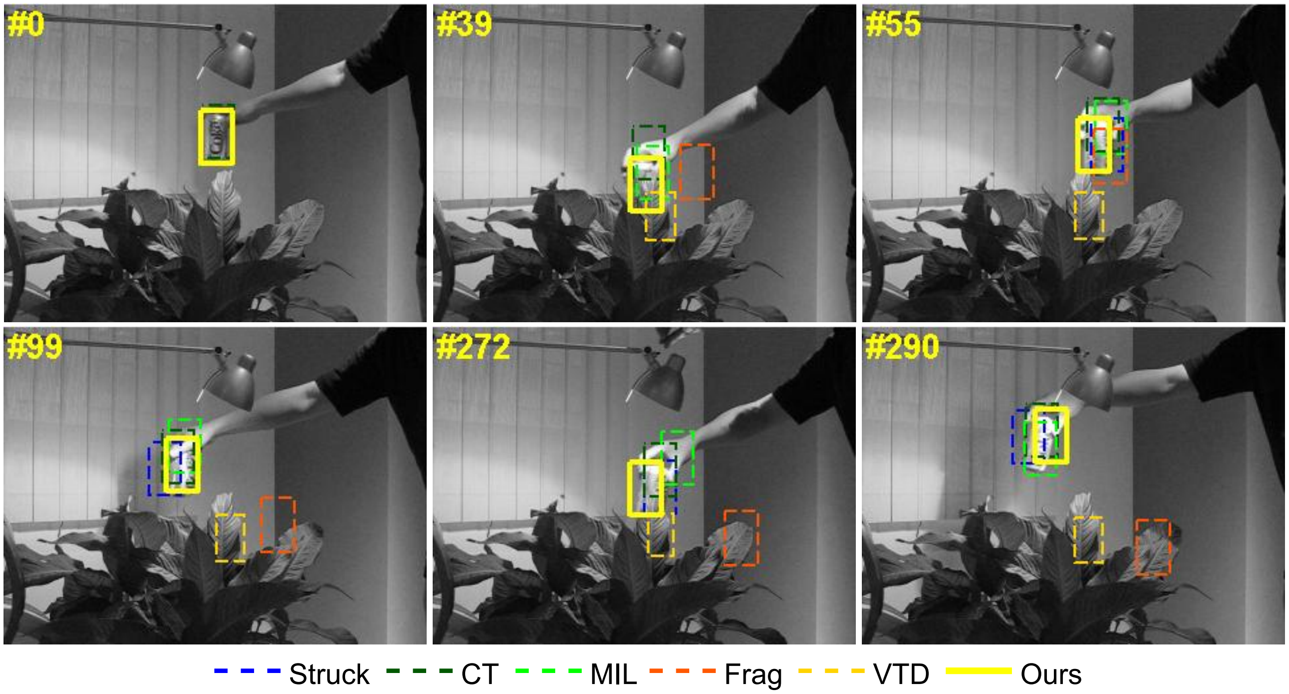

Figure 13. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *coke* sequence.

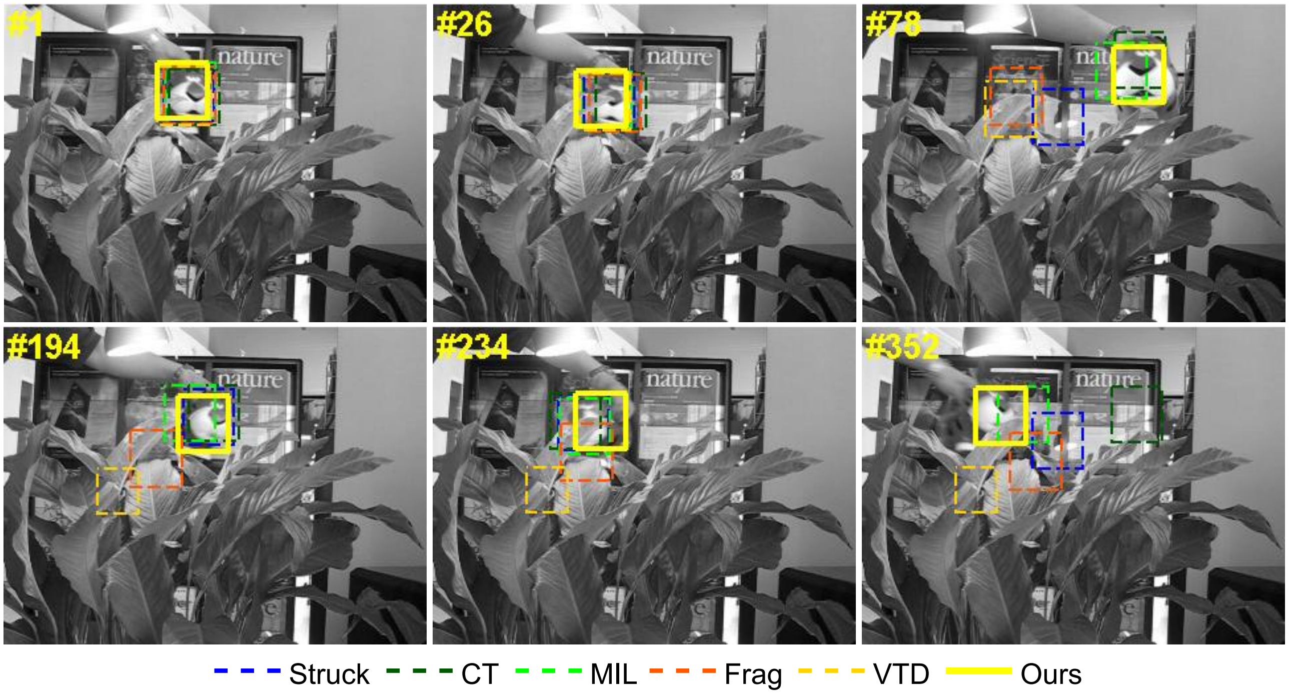

Figure 14. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *tiger1* sequence.

part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[6] S. Hare, A. Saffari, and P. Torr. Struck: Structured output tracking with kernels. In *Proc. ICCV*, pages 263–270, 2011.

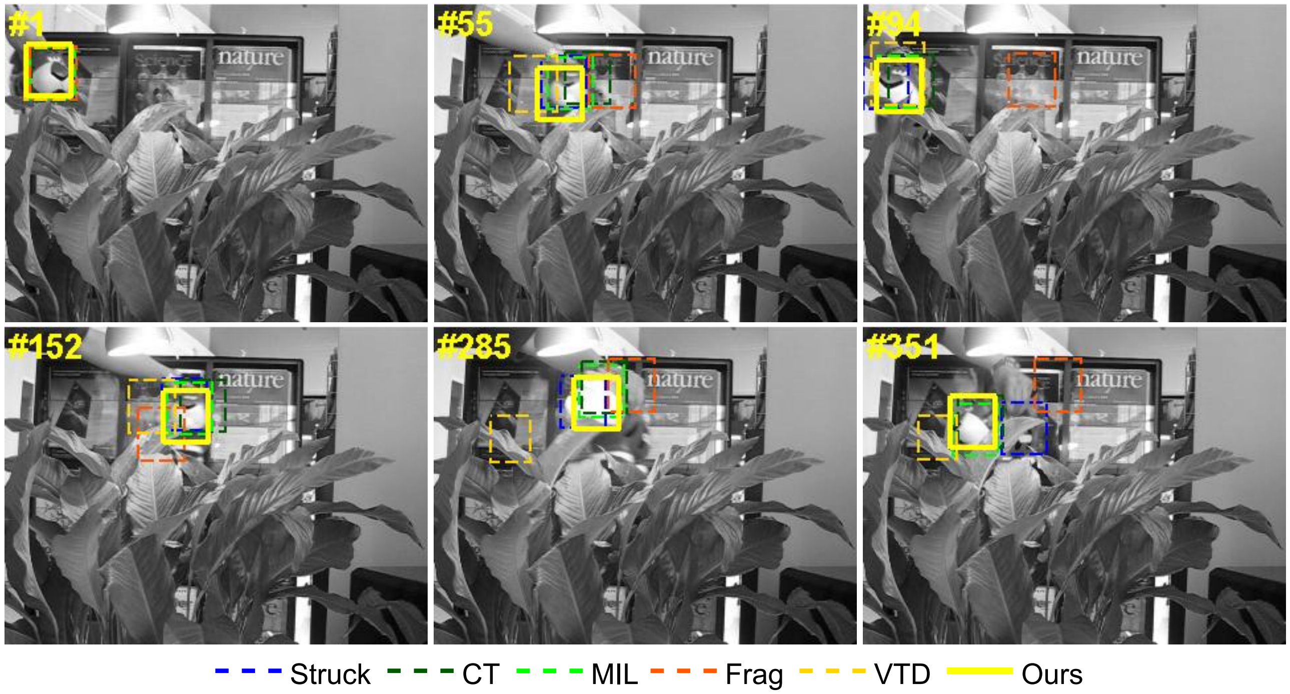

Figure 15. Qualitative visual tracking results of our tracker and competing trackers over representative frames of *tiger2* sequence.

[7] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[8] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Math. Program.*, 127(1):3–30, 2011.

[9] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *Proc. CVPR*, pages 1815–1821, 2012.

[10] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

[11] B. Yang and R. Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. In *Proc. ECCV*, pages 484 – 498, 2012.

[12] R. Yao, Q. Shi, C. Shen, Y. Zhang, and A. van den Hengel. Robust tracking with weighted online structured learning. In *Proc. ECCV*, 2012.

[13] C. N. J. Yu and T. Joachims. Learning structural SVMs with latent variables. In *Proceedings of International Conference on Machine Learning*, pages 1169–1176, 2009.

[14] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman. Latent hierarchical structural learning for object detection. In *Proc. CVPR*, pages 1062–1069, 2010.

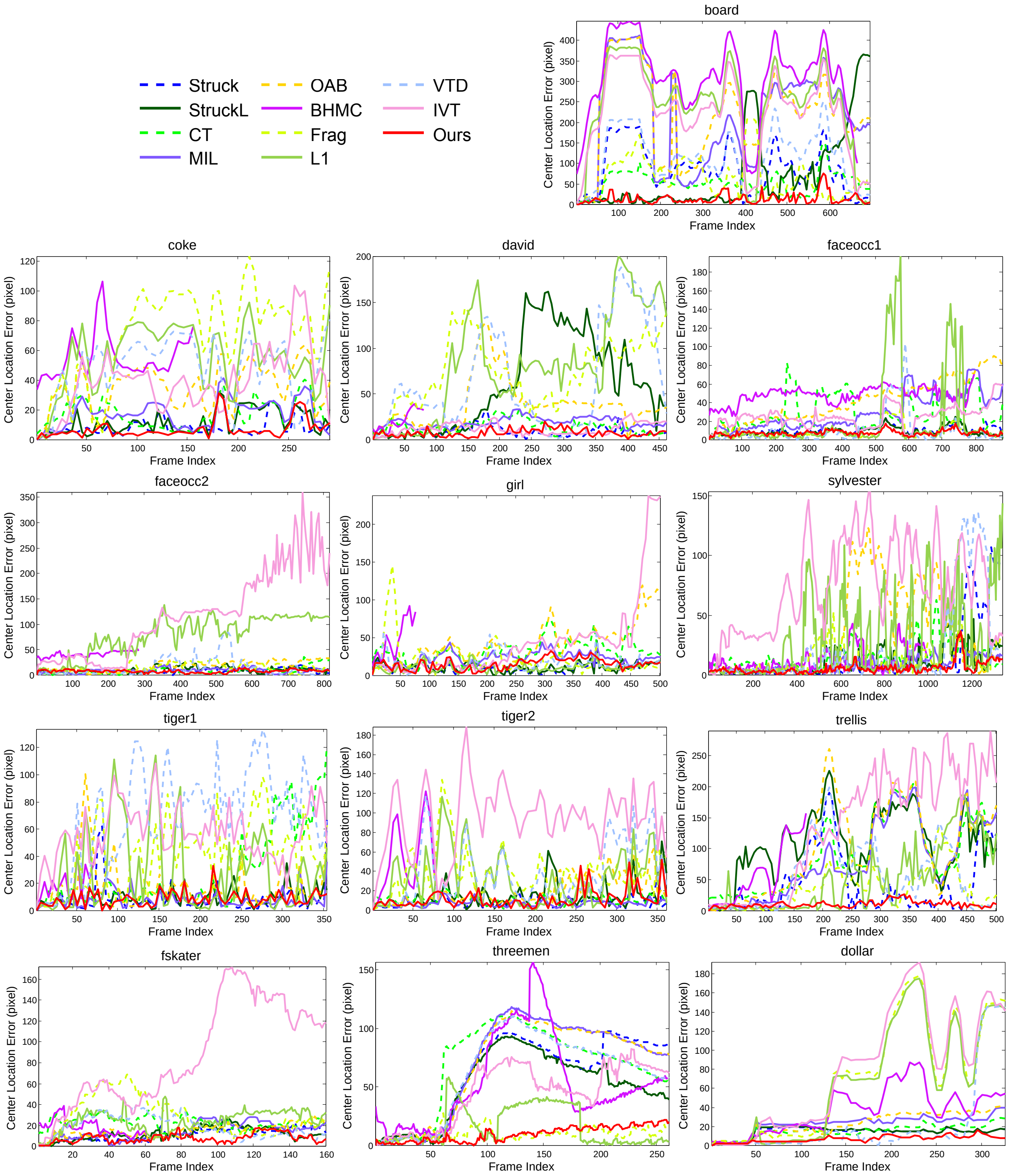

Figure 16. Quantitative evaluation of different trackers in centre location error plots on thirteen sequences.

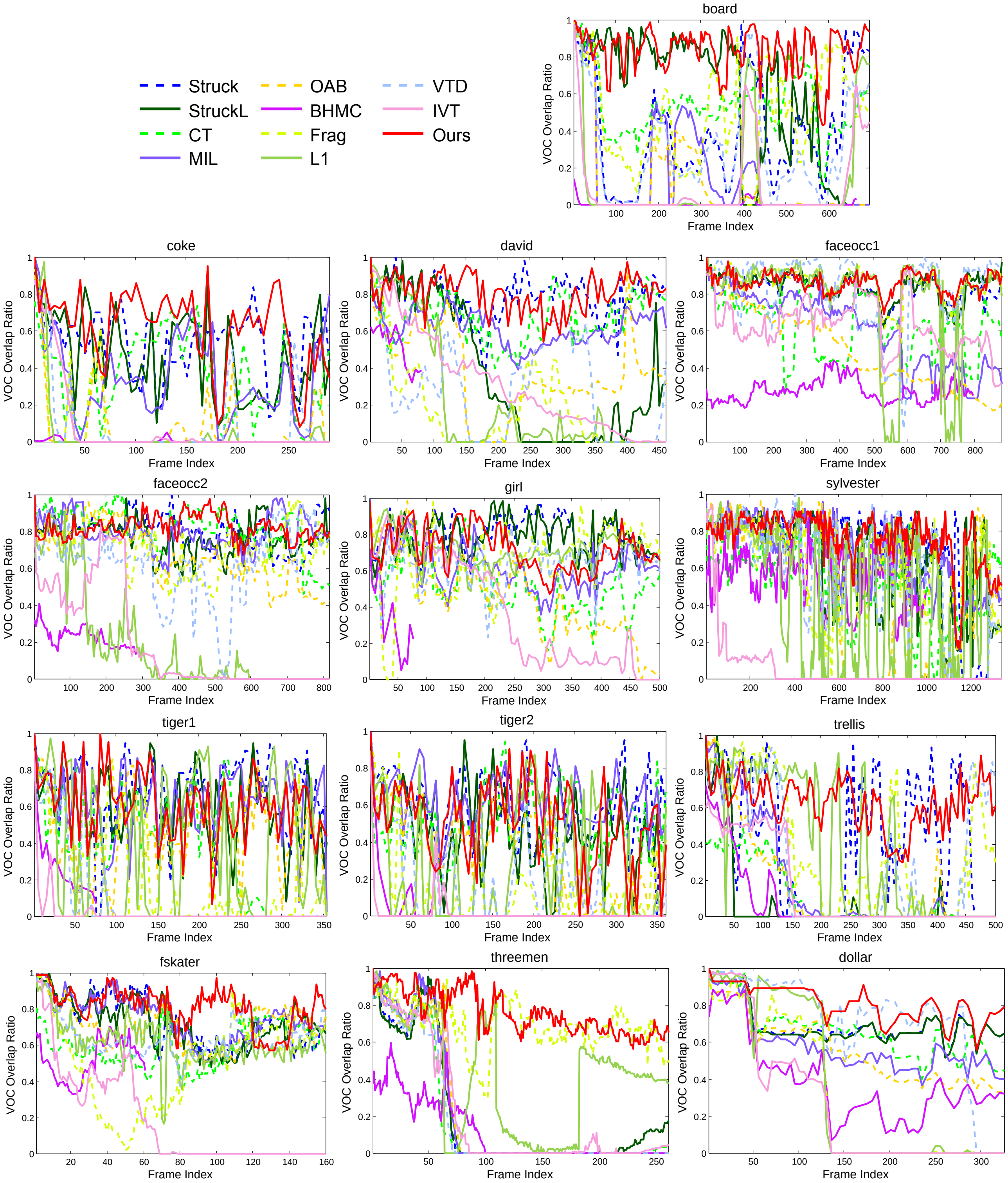

Figure 17. Quantitative evaluation of different trackers in VOC overlap ratio plots on thirteen sequences.