Superpixel Tracking

Shu Wang[1], Huchuan Lu[1], Fan Yang[1] abnd Ming-Hsuan Yang[2]
[1]School of Information and Communication Engineering, University of Technology, China
[2]Electrical Engineering and Computer Science, University of California at Merced, United States

This document includes the illustrations of the training and tracking procedures. Some experimental results are also presented in this document in high resolution. In addition the detail of the Gaussian motion model is presented here.

Our algorithm is implemented in MATLAB and is run on a machine of 2.2GHz duo core CPU with 2GB memories. It takes less than 5 seconds per frame during tracking, and 20 seconds or more during every updating procedure. The most time-consuming part is due to the use of the mean shift clustering algorithm. All the code and data sets will be made available at `http://faculty.ucmerced.edu/mhyang/pubs/iccv11a.html`.

The detail of our motion model:

The motion (or dynamical) model of our tracker is assumed to be Gaussian distributed:

$$p(X_t|X_{t-1}) = \mathcal{N}(X_t; X_{t-1}, \Psi) \tag{1}$$

where Ψ is a diagonal covariance matrix whose elements are the standard deviations for location and scale, i.e., σ_c and σ_s. The values of σ_c and σ_s dictate how the proposed algorithm accounts for motion and scale change. Furthermore, σ_c and σ_s are related with the target size and resolution of a certain video. According to these factors and Gaussian probability-density function, we calculate the motion estimate $p(X_t^{(l)}|X_{t-1})$ of N target candidates $\{X_t^{(l)}\}_{l=1}^N$ as follows:

$$\begin{aligned} & p(X_t^{(l)}|X_{t-1}) \\ & = \prod_{i=c,s} (2\pi {\sigma_i}^2)^{-\frac{1}{2}} \times \exp[-\tfrac{1}{2} \times (\tfrac{|X_t^{i,(l)} - X_{t-1}^i|}{\sigma_i \times \lambda_r})^2] \\ & \qquad\qquad\qquad\qquad , \ \forall\, l = 1, ..., N \end{aligned} \tag{2}$$

Parameter λ_r represents the resolution of a certain video, and it is obtained by:

$$\lambda_r = \frac{v_h + v_w}{2} \tag{3}$$

where parameter v_h and v_w denotes the height and the width of the frame image, respectively.

Figure 1 represents a part of the experimental results comparison of our tracker and other state-of-the-art trackers (IVT tracker, Visual Tracking Decomposition (VTD) tracker, MIL tracker, Frag tracker, PDAT and PROST tracker).

Figure 1. Four common challenges encountered in tracking. The results by our tracker, IVT, VTD, PROST, FragTrack and PDAT methods are represented by yellow, red, white, green, cyan, and magenta rectangles. Existing state-of-the-art trackers are not able to effectively handle heavy occlusion, large variation of pose and scale, and non-rigid deformation, while our tracker gives more robust results.

Figure 2 illustrates the whole tracking process, and Figure 3 explains for Equation 11 in our paper.

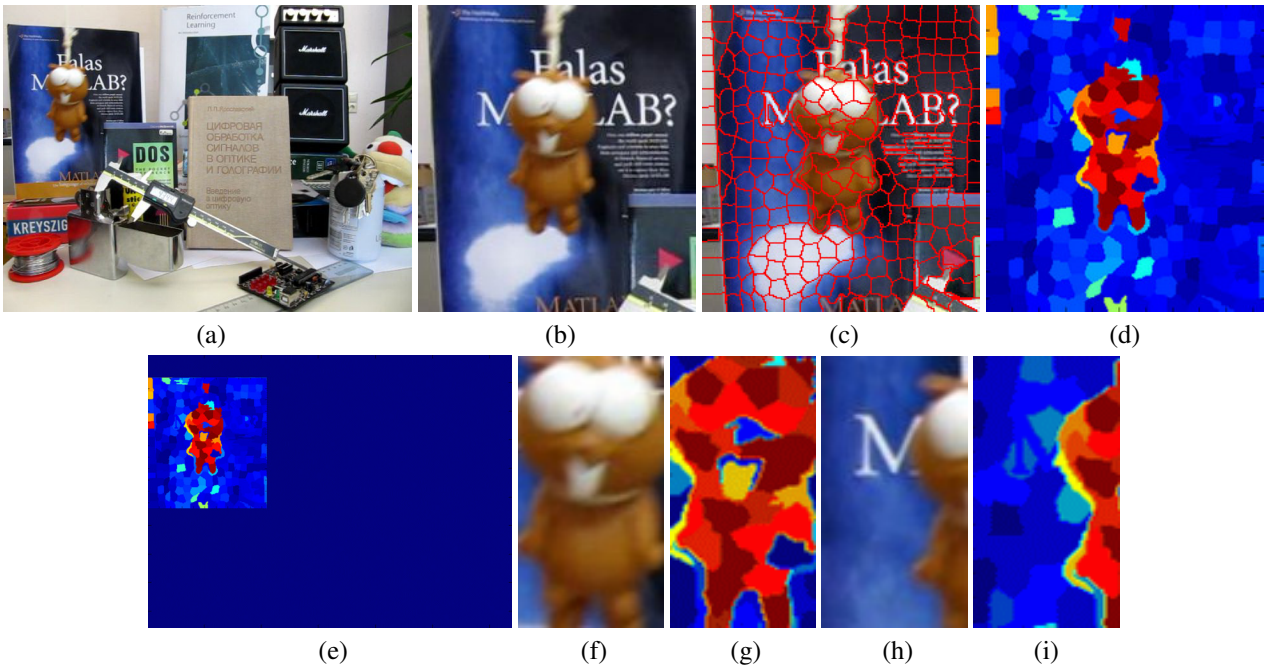

Figure 2. Illustration of the use of confidence map for state prediction in tracking process. (a) a new frame at time t. (b) neighborhood of the target location in the last frame, i.e., at state X_{t-1}. (c) segmentation result of (b). (d) the computed confidence map of superpixel. The superpixels colored with red indicates strong likelihood of belonging to the target and those colored with dark blue indicate strong likelihood of belonging to background. (e), (f) and (g), (h) show two target candidates with high and low confidence, respectively.

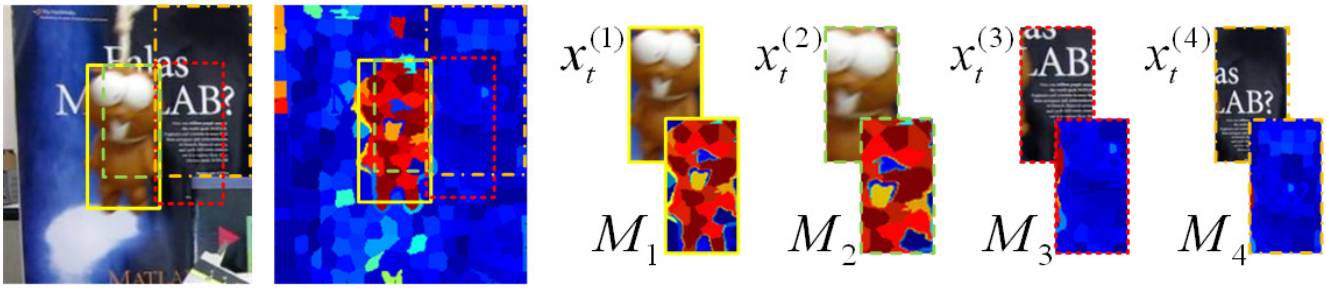

Figure 3. Confidence map. Four target candidate regions corresponding to states $X_t^{(i)}$, $i = 1, \ldots, 4$ are shown both in warped image and the confidence map. These candidates' confidence regions M_i, $i = 1, \ldots, 4$ have the same canonical size (upper right) after normalization. Based on corresponding equation (Equation 10 in our paper), Candidates $X_t^{(1)}$ and $X_t^{(2)}$ have similar positive confidence C_1 and C_2. Candidates $X_t^{(3)}$ and $X_t^{(4)}$ have similar negative confidence C_3 and C_4. However, candidate $X_t^{(2)}$ covers less target area than $X_t^{(1)}$, and $X_t^{(4)}$ covers more background area than $X_t^{(3)}$. Intuitively, target-background confidence of $X_t^{(1)}$ should be higher than $X_t^{(2)}$, while confidence of $X_t^{(4)}$ should be lower than $X_t^{(3)}$. These two factors are considered in computing confidence map in the tracking stage.

Figure 4 illustrates the entire training process, and Figure 5 shows the tracking results with comparisons to color-based trackers of mean shift tracker and adaptive color-based particle filter.

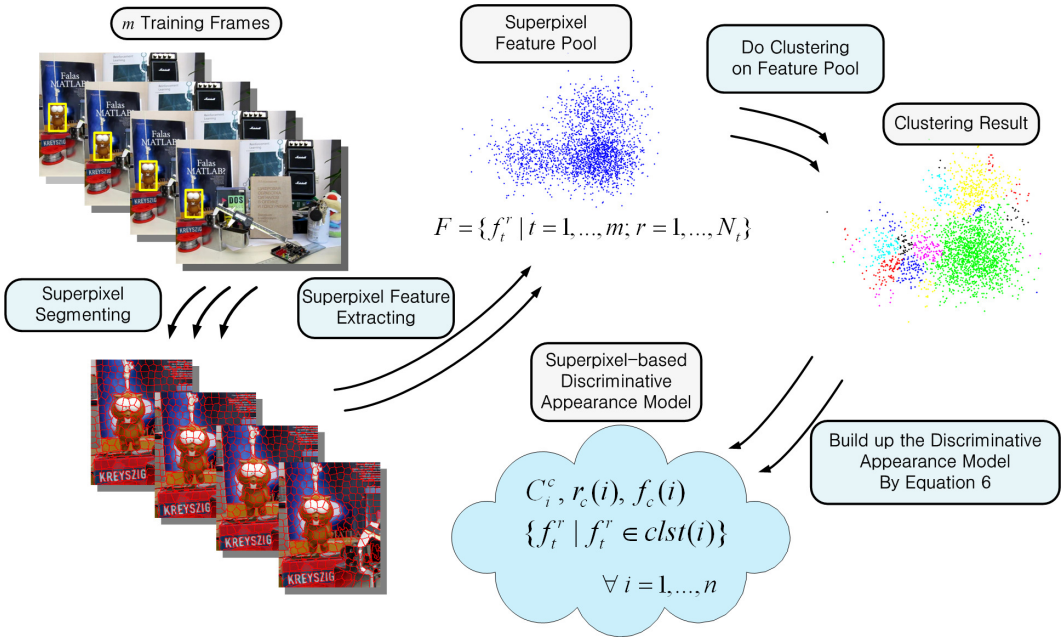

Figure 4. Illustration of the training process.

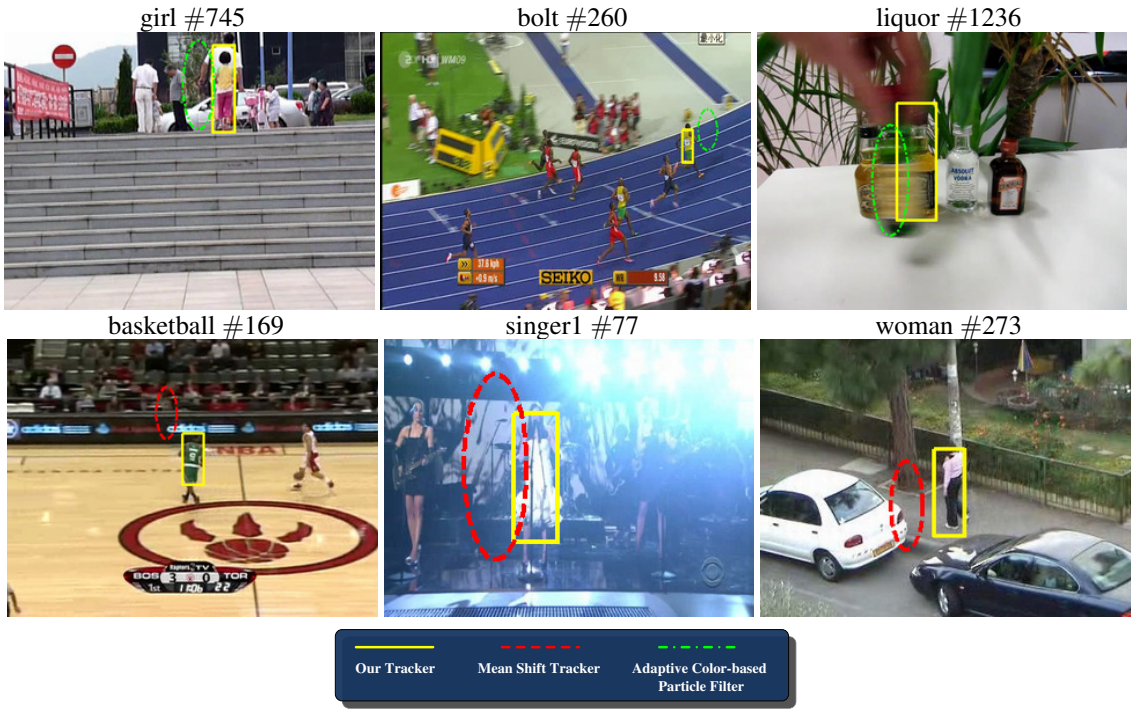

Figure 5. Tracking results with comparisons to color-based trackers. The results by the MS tracker, PF method and our algorithm are represented by red ellipse, green ellipse and yellow rectangles. It is evident that our tracker is able to handle cluttered background (*girl* and *basketball* sequences), drastic movement (*bolt* sequence), heavy occlusion (*liquor* and *woman* sequences) and lighting condition change (*singer1* sequence).

Figure 6 and Figure 7 represents the experimental results comparison of our tracker and other state-of-the-art trackers (IVT tracker, Visual Tracking Decomposition (VTD) tracker, MIL tracker, Frag tracker and PROST tracker).

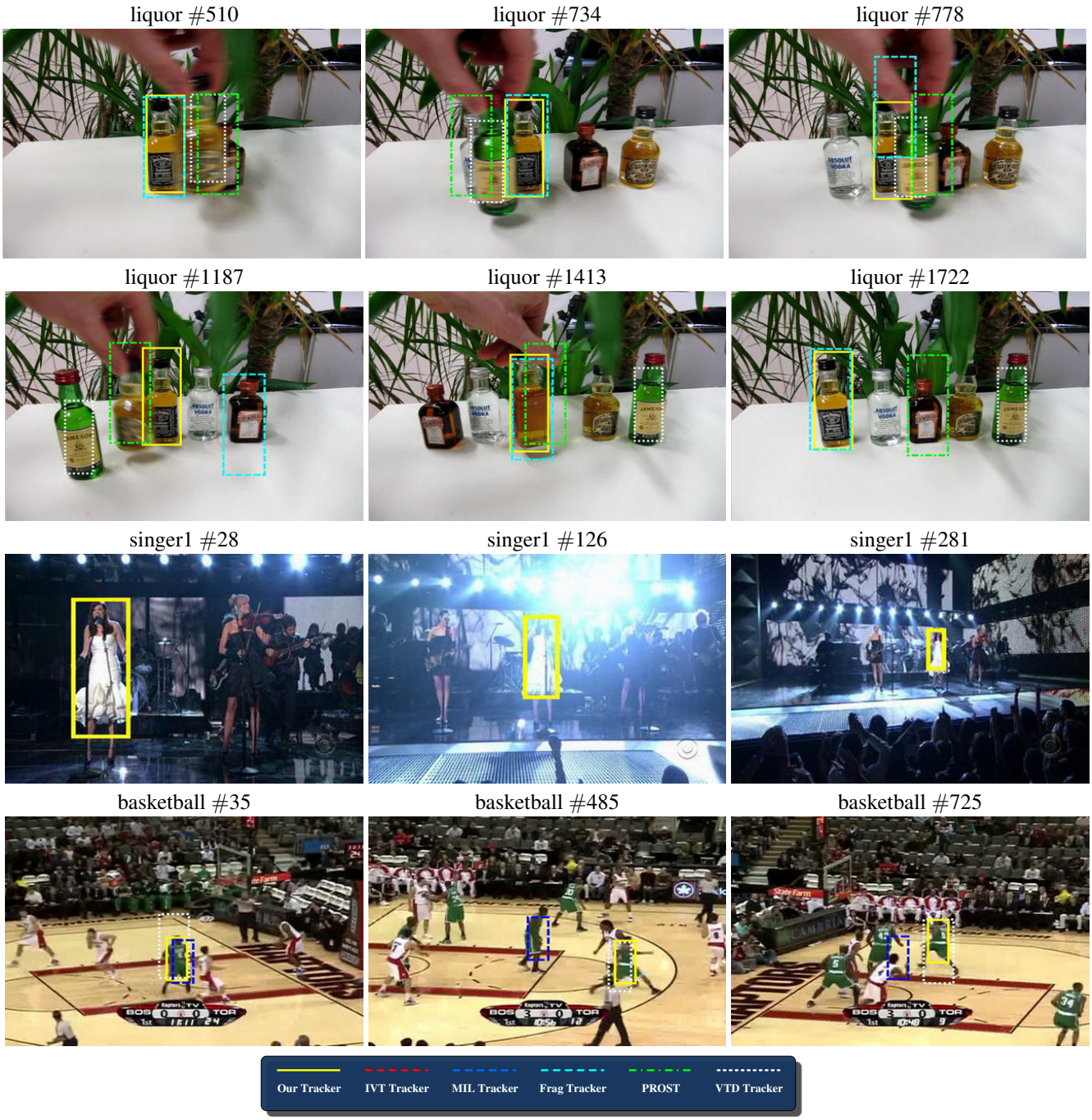

Figure 6. Tracking results with comparisons to other state-of-the-art trackers. The results by our tracker, IVT, VTD, PROST, MILTrack and FragTrack methods are represented by yellow, red, white, green, blue and cyan rectangles.

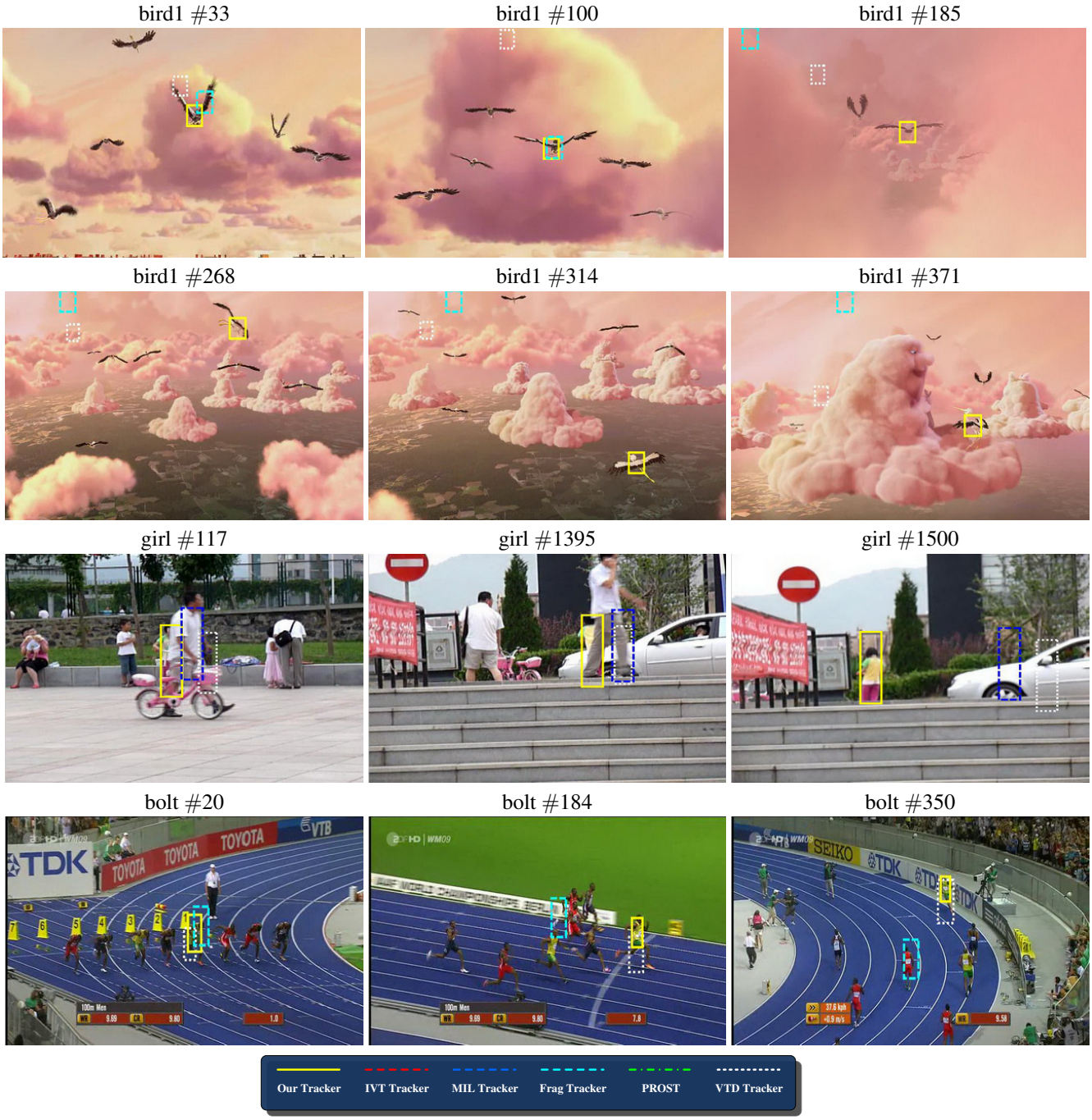

Figure 7. Tracking results with comparisons to other state-of-the-art trackers. The results by our tracker, IVT, VTD, PROST, MILTrack and FragTrack methods are represented by yellow, red, white, green, blue and cyan rectangles.

Figure 8 shows the tracking results of figure/ground segmentation across frames on sequence *liquor*. The 1st row are original images from six different frames, and the 2nd row are the local regions of target based on previous tracking results. The 3rd row are the confidence maps of correspondent local regions, which is obtained by using the appearance model. The 4th row are the figure/ground segmentation results. We obtain the segmentation results (the 4th row in Figure 8) by adopting a simple adaptive threshold-based segmentation method. The 5th row are the final tracking results of each frame.

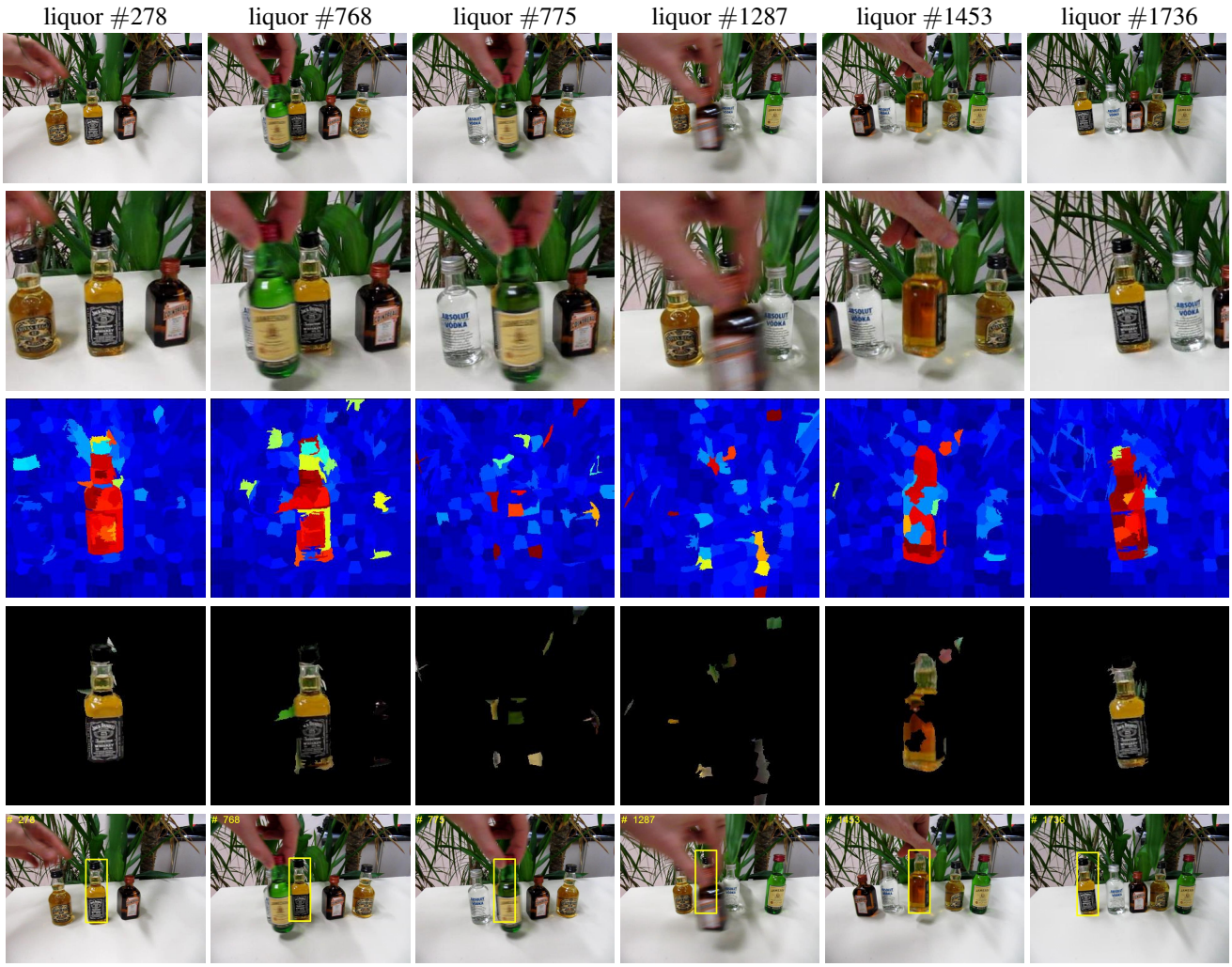

Figure 8. The tracking results of figure/ground segmentation across frames on sequence *liquor*. The 1st row are original images from six different frames, and the 2nd row are the local regions of target based on previous tracking results. The 3rd row are the confidence maps of correspondent local regions, which is obtained by using the appearance model. The 4th row are the figure/ground segmentation results. The 5th row are the final tracking results of each frame.

Frame 278 shows that the target appearance is well learned by our appearance model. A simple adaptive threshold-based segmenting on the confidence map (the 3rd row) of target local region (the 2nd row) gives finest figure/ground segmentation result (the 4th row).

Frames 768 and 778 show a full procedure when the target experiences a full occlusion by a similar target. It can be clearly seen from the confidence map (the 3rd row) that our appearance model gives high measurement to the superpixels that belong to target. Further, when a heavy occlusion occurs, the target disappears from our confidence map.

Frame 1287 shows another example of successful tracking under full occlusion. Frame 1453 shows that our tracker is robust to target pose changes that it still gives fine tracking result under severe pose variance (the 4th and the 5th row). Frame 1736 shows that our tracker survives all challenging factors in sequence *liguor* and our appearance model gives accurate segmentation result by the end of this sequence.

The plots from the paper in greater resolution are also reproduced in Figure 9.

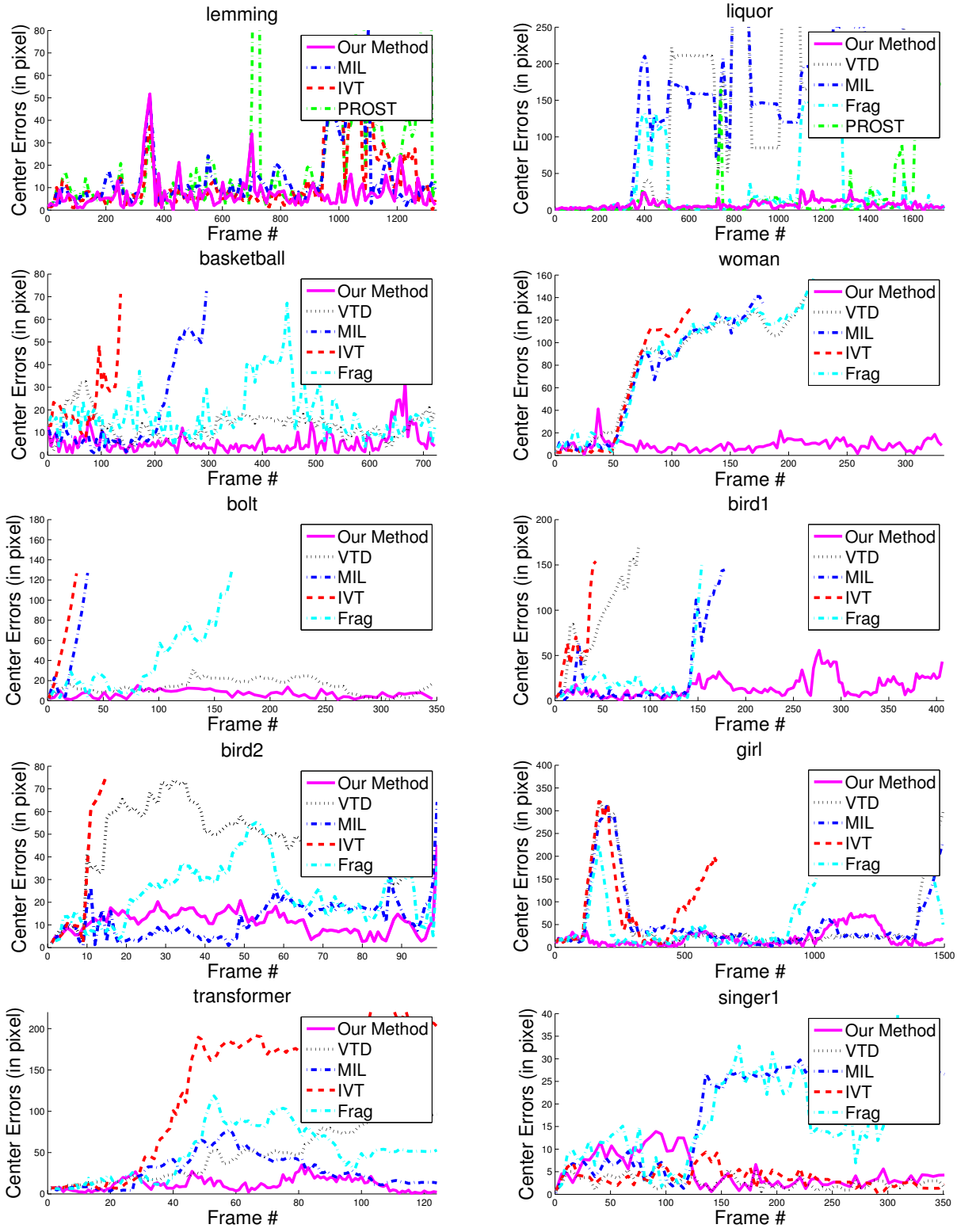

Figure 9. Quantitative evaluations of our tracker and other state-of-the-art trackers. Trackers of ours, IVT, Visual Tracking Decomposition (VTD), MILTrack, FragTrack, PROST are presented by color magenta, red, black, blue, cyan and green.