Supplementary Material to Robust Visual Tracking using Multi-Frame Multi-Feature Joint Modeling

Peng Zhang*, Shujian Yu*, *Student Member, IEEE,* Jiamiao Xu, Xinge You†, *Senior Member, IEEE,* Xiubao Jiang, Xiao-Yuan Jing, and Dacheng Tao, *Fellow, IEEE.*

I. K(MF)^{2}JMT ON VIDEO SEQUENCES WITH SHOT CHANGES AND POSSIBLE MODIFICATIONS

In this section, we provide tracking results of K(MF)^{2}JMT and five state-of-the-art trackers (i.e., MEEM [1], TGPR [2], Struck [3], SCM [4], ASLA [5]) on five video sequences (three are from OTB 2015, the remaining two are from VOT2015 benchmark) with shot changes or scene cuts. We also suggest two modifications to our current K(MF)^{2}JMT to alleviate the negative effects incurred by the these changes.

The first modification is to give different weights to different frames in the overall objective of K(MF)^{2}JMT (i.e., Eq. (1) in the main text). The motivation is intuitive: in the scenarios of shot changes or scene cuts, the temporal coherence (from previous frame) becomes weaker and the tracker needs to assign more weight to the most adjacent (or neighboring) frame to better capture the instantaneous information. The second modification is to incorporate a shot change detector (e.g., [6], [7]) into our K(MF)^{2}JMT, such that the system can automatically detect the shot changes. Once a shot change is confirmed, the system needs to re-detect or re-identify the location of the target. However, one should note that, there is no guarantee that the selected shot detector can reconcile with the given tracker. Moreover, the integration of shot detector will introduce more hyper-parameters.

The selected videos are *DragonBaby*, *BlurOwl*, *Soccer*, *Singer1* and *Singer3*. In the video *DragonBaby*, the shot change is caused by varying camera-subject distances, i.e., there is shot change from full shot to medium shot[1]. In the video *BlurOwl*, the shot change is caused by the sudden changes of camera point-of-view or angle. In the video *Soccer*, the shot change is caused by either the gradual changes of camera point-of-view or the varying camera-subject distances. In the videos *Singer1* and *Singer3*, the shot change is caused by (rapid) changes of both camera point-of-view and camera-subject distances.

We implement the first modification to validate its effectiveness due to its simplicity. Specifically, given M training frames in the overall objective, the weight in the current frame is A_0, then the weights in previous frames are decayed inversely proportional to the square of the distance from the current frame (i.e., the weight in the most adjacent frame is $A_0/4$, the weight in the second most adjacent frame is $A_0/9$, and the weight in the farthest frame is A_0/M^2). We term this modification K(MF)^{2}JMT-M1 and set $A_0 = 5$ in the following proof-of-concept experiment[2]. Fig. 1 plots the tracking results of our K(MF)^{2}JMT and K(MF)^{2}JMT-M1 as well as their five competitors. Table I summarizes the overlap precision (%) at threshold 0.5 for all competing trackers.

As can be seen, our basic K(MF)^{2}JMT performs favorably in these videos, but it may miss the target or overestimate the target size due to unconstrained shot changes. The simple modification can effectively alleviate the negative effects incurred by these changes, thus further improving the performance of K(MF)^{2}JMT. This result suggests that the precise utilization of temporal information (coupled with a careful weighting strategy) is preferred in (unconstrained) videos containing shot changes or scene cuts. At the same time, it also suggests the (possible) existence of the room for performance improvement with an advanced strategy to address shot changes. We leave the implementation of the second modification as future work.

REFERENCES

[1] J. Zhang, S. Ma, and S. Sclaroff, "Meem: robust tracking via multiple experts using entropy minimization," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 188–203.

[2] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with gaussian processes regression," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 188–203.

[3] S. Hare, A. Saffari, and P. H. Torr, "Struck: Structured output tracking with kernels," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 263–270.

[4] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1838–1845.

[5] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1822–1829.

*The first two authors contributed equally to this work and should be regarded as co-first authors.

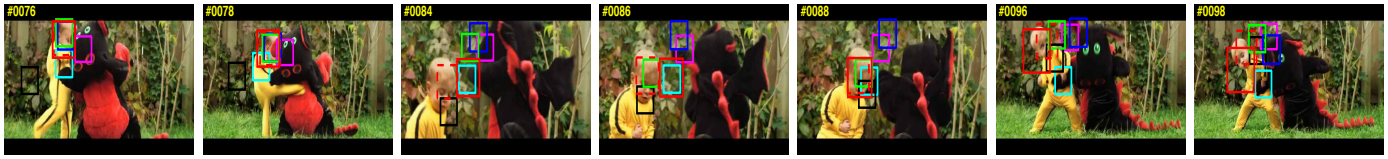
†Corresponding author.

[1]Please refer to [8] for definitions of full shot, medium shot, etc.

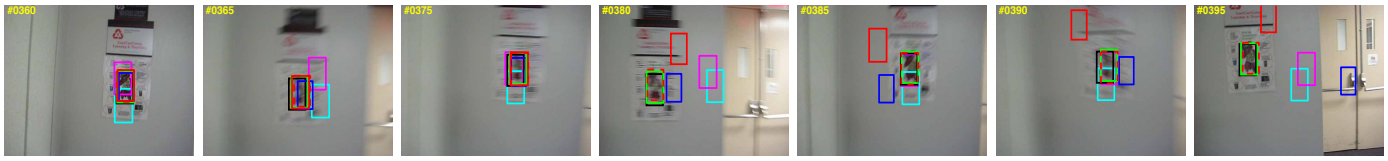
[2]The parameter A_0 is selected, from the range $[1, 10]$ with interval 1, as the one that achieves the highest mean success rate among all selected video sequences with scene cuts.

TABLE I
A COMPARISON OF K(MF)^{2}JMT AND K(MF)^{2}JMT-M1 WITH FIVE STATE-OF-THE-ART TRACKERS. FOR EACH TRACKER, THE OVERLAP PRECISION (%) AT THRESHOLD 0.5 IS PRESENTED. THE BEST TWO RESULTS ARE MARKED WITH RED AND BLUE RESPECTIVELY.

	MEEM	TGPR	STRUCK	SCM	ASLA	K(MF)^{2}JMT	K(MF)^{2}JMT-M1
DragonBaby	65.5	73.5	8.8	23.0	15.0	46.0	66.4
BlurOwl	98.6	51.2	98.6	21.6	17.6	55.9	90.2
Soccer	36.0	13.0	15.6	23.7	12.5	56.6	78.6
Singer1	25.1	22.8	29.9	100	100	93.7	98.6
Singer3	15.3	15.3	24.4	15.3	16.0	17.6	37.4
Mean	48.1	35.2	35.5	36.7	32.2	54.0	74.2

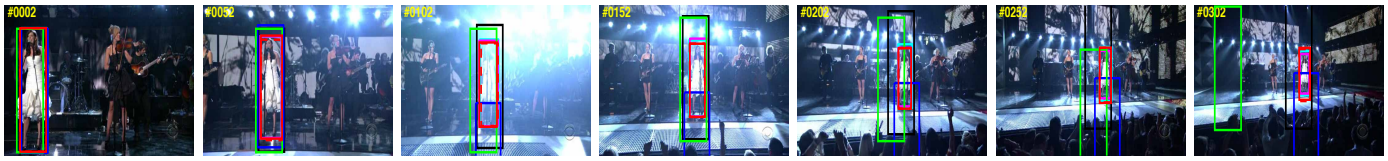


(a) Shot changes in *DragonBaby*: there are abrupt shot changes from full shot to medium shot (see frame 78 to frame 84) and from medium shot to full shot (see frame 88 to frame 94).

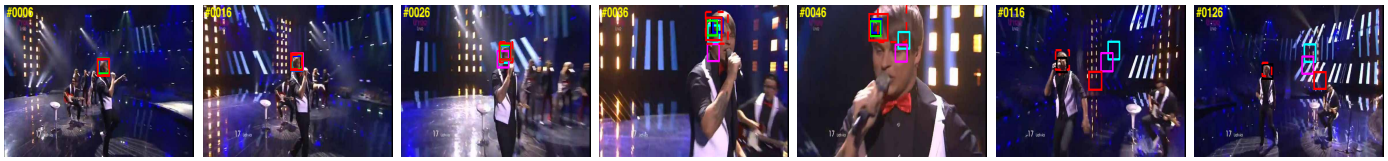


(b) Shot changes in *BlurOwl*: there are abrupt shot changes due to the sudden change of camera point-of-view (see the transition between frame 375 and frame 380 or the transition between frame 390 and frame 395).

(c) Shot changes in *Soccer*: there are gradual shot changes due to the changes of camera point-of-view (see frame 45 to frame 90 and frame 335 to frame 375) or incurred by varying camera-subject distances (see frame 200 to frame 230 and finally to frame 305).



(d) Shot changes in *Singer1*: there are shot changes due to the gradual changes of both camera point-of-view and camera-subject distances (see frame 2 to frame 102 and finally to frame 302).



(e) Shot changes in *Singer3*: there are shot changes due to the gradual (and rapid) changes of both camera point-of-view and camera-subject distances (e.g., frame 26 to frame 36). Our modification K(MF)^{2}JMT-M1 may underestimate the target size due to the rapid changes, but it still provides the most accurate estimation among others.

(f) Tracker legend

Fig. 1. A qualitative comparison of our method and its modification with five state-of-the-art trackers. Tracking results are shown on five videos contain scene cuts or shot transitions. *DragonBaby*, *BlurOwl* and *Soccer* are from OTB 2015, whereas *Singer1* and *Singer3* are from VOT2015 benchmark. The basic K(MF)^{2}JMT performs favorably in these videos. Our modification K(MF)^{2}JMT-M1 offers the best performance. (f) shows tracker legend.

[6] P. Bouthemy, M. Gelgon, and F. Ganansia, "A unified approach to shot change detection and camera motion characterization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1030–1044, 1999.
[7] M. Birinci and S. Kiranyaz, "A perceptual scheme for fully automatic video shot boundary detection," *Signal Processing: Image Communication*, vol. 29, no. 3, pp. 410–423, 2014.
[8] I. K. Sethi and N. V. Patel, "Statistical approach to scene change detection," in *Storage and Retrieval for Image and Video Databases III*, vol. 2420. International Society for Optics and Photonics, 1995, pp. 329–339.