Some Promising Ideas about Multi-instance Video Segmentation

Hao Zhao
hao.zhao@intel.com

Abstract

This paper presents our research on DAVIS 2017 video object segmentation challenge. (1) We present a baseline and an analysis into it, pointing to the challenges that DAVIS 2017 features (including but limited to multiple instances, small objects, bad illumination, and large poses). (2) We show optical flow extracted by a pre-trained network generally seems reasonable. (3) We show clearly that motions cues are complementary to appearance cues for instance edge detection. (4) We show qualitative results for several different FCN formulations to differentiate instances. We point to the fact that they are promising but far from applicable.

1. Baseline

Recently a method called one-shot video object segmentation (OSVOS) [3] treats video object segmentation as an one-sample semantic segmentation problem and achieves record-breaking performance on the DAVIS 2016 dataset [6]. Personally, We think it is very interesting as it points to a scientific fact that when over-fitted to a single sample a fully convolutional network could generalize to more samples with similar appearance. We follow the methodology of OSVOS in this research.

It seems that the architecture that OSVOS uses is quite out-dated, so we start with the recent PSPNet [9] architecture. We fine-tune the VOC-pretrained model on the first frame. We find a learning rate of $10e^{-6}$ and 2000 iterations to be good hyper-parameters. Training with 1000 iterations leads to obviously blurry edges even in simple sequences like *guitar-violin* or *orchid* in test-dev. We raise iterations up and 2000 iterations seems a saturated value as training meanIU no longer decreases (higher than 95% in all test-dev sequences). With 2000 iterations, we try several base learning rate and finally choose 10^{-6}. A small learning rate slows down training (which seems fundamental for rotoscoping yet actually we believe we are still far from real-world applications so this concern is not quite necessary) and a larger rate runs into obvious energy plateau (as a disclaimer, it always runs out finally according to our experience). So our conclusion is that training enough epoches is important but learning rate does not really matter if you are not developing real-time applications. Besides, all our experiments are done without DAVIS trainval representation learning.

This baseline entry scores a global score of 0.557 on test-dev. We experiment with several different techniques to improve this baseline yet got no positives results, so this report only shows some of promising ideas that may work statistically (in the future with proper modifications).

2. Look into the baseline

Let us start with a detailed analysis about this baseline entry. Please refer to Table 1 for a per-sequence accuracy report (of some typical sequences). *Aerobatics* is simple sequence, we think, with the only difficulty as the transparent material. The average IU of three objects is 83.4%. And as shown by Fig 1, the qualitative result is quite satisfactory although in some frames the transparent material does cause confusion. *Chamaleon* is another simple sequence. It features similar color distribution between foreground and background yet CNN handles this properly, scoring an IU of 90.1%. *Orchid* and *slack-line* are similar simple sequences which are not shown here. We think these results generally show that this formulation is good enough for sequences with large foreground area and single instance.

Performance gets significantly lower when it comes to multiple instances (e.g. *carousel* and *salsa* in test-dev. *choreography* and *dolphins* in test-challenge.). This is not surprising as this formulation is intrinsically same as semantic segmentation (a pixel-wise cross-entropy loss function) thus does not have the capability to address multiple instances. We report an average IU of 35.9%/29.3% on *carousel* and *salsa* respectively. Fig 1 demonstrates that predictions on these two sequences are nothing but mess.

Interestingly, *guitar-violin* scores an average IU of 88.8% although there are two people. We think the reason behind is that these two objects have quite different appearance (clothes and poses). Small objects is another challenge, typically shown by the *golf* and *monkeys-trees* sequence. In *golf*, the IU for the club is 4.0%. In *monkeys-trees*, the average IU is 22.4%. Generally speaking, We be-

Figure 1. This shows some qualitative results of the baseline entry. Left: aerobatics, aerobatics, carousel, monkeys-trees. Right: chameleon, guitar-violin, salsa, golf.

Table 1. Some per-sequence results on test-dev of the baseline entry.

name	aerobatics1	aerobatics2	aerobatics3	carousel1	carousel2	carousel3	carousel4
J	0.887	0.846	0.770	0.428	0.316	0.266	0.425
F	0.788	0.999	0.707	0.435	0.384	0.318	0.424

name	chamaleon1	deer1	deer2	giant-slalom1	giant-slasom2	giant-slalom3	golf1
J	0.901	0.156	0.324	0.528	0.392	0.645	0.796
F	0.929	0.184	0.500	0.528	0.392	0.836	0.758

name	golf2	golf3	guitar-violin1	guitar-violin2	guitar-violin3	guitar-violin4	monkeys-trees1
J	0.040	0.577	0.875	0.913	0.930	0.835	0.246
F	0.052	0.643	0.881	0.911	0.877	0.849	0.375

name	monkeys-trees2	salsa1	salsa2	salsa3	salsa4	salsa5	salsa6
J	0.201	0.262	0.622	0.534	0.106	0.309	0.039
F	0.304	0.422	0.636	0.627	0.253	0.467	0.134

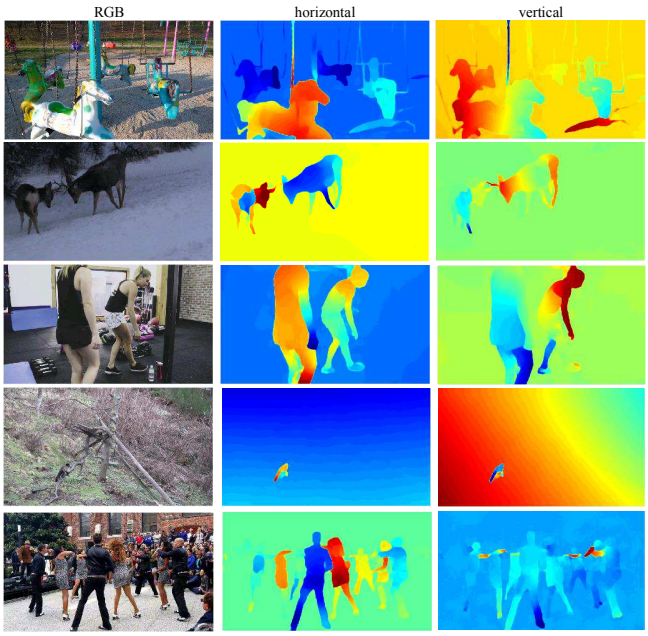

Figure 2. Optical flow quality visualization.

lieve this is solvable given an automatic zoom-in scheme is built (e.g. [7]). Multi-scale mechanism is a long-existing topic in semantic segmentation, and we find that a general multi-scale training/testing fashion (which Deeplab and PSPNet largely relies upon) is not better than solely using the original resolution.

To sum up, there are many challenges in DAVIS 2017, including but not limited to multiple instance, small objects, bad illumination (e.g. *deer*, *golf* and *people-sunset*), and large poses (e.g. *hoverboard* and *rollercoaster*). We think it is difficult to find a generic solution to solve them all, so we start our research on multiple instances which seems to be the most fundamental one.

3. Extracting Optical Flow

Optical flow seems to be a reasonable tool for solving the problem of multiple instances because: (1) motion cues may be useful for differentiating instances; (2) optical flow can be used to link instances between frames after instances are already segmented. We extract optical flow on DAVIS sequences with the pre-trained FlowNet [4]. We provide a visualization on several sequences (*carousel*, *deer*, *gym*, *monkeys-trees*, and *salsa*)in Fig 2. FlowNet is trained on synthetic datasets FlyingChair and Sintel, but we find the qualitative results are generally acceptable and we use them in all following experiments.

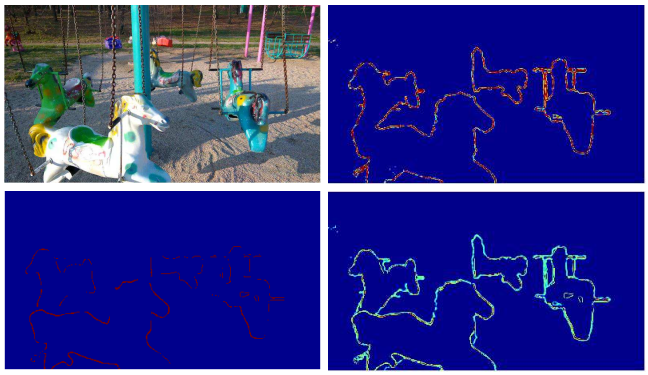

Figure 3. Top-left: RGB Input. Top-right: appearance edges generated by HE-PSPNet. Bottom-left: motion edges. Bottom-right: ad-hoc merge.

4. Better Edge with Motion Cues

We think motion cues and appearance cues are obviously complementary for instance edge detection. In order to extract appearance edges, We design a holistically-nested [8] PSPNet. Besides, informed by the recent work of DWT [2], we firstly train a semantic network (actually foreground network in this case) to mask out irrelevant regions in the pixel domain. We illustrate that motion cues are obviously complementary to appearance cues for extracting instance edges, with Fig 3. The motion edges are extracted by gradient operators on flow maps and masked by the foreground network. The final output is merged in an ad-hoc manner, thus we guess training HE-PSPNet with flow inputs will perform better. We believe video instance edge detection will definitely benefit from motion cues (if this task exists).

However, turning edges into segments is still an open question. The ucm [1][5] methodology is an alternative yet in our humble opinion, it is just too slow, heuristic, and not optimized for instance-level operations. So We have not tried ucm yet. Using FCNs to turn edges into category-specific instance masks may be an interesting research direction.

5. Other Formulations

Except for extracting instance edges, We try other formulations to differentiate instances. We discretize the distance transform map into four bins and train a model with classification (cross-entropy) loss. It seems that this can, to some extent, separate instances (shown in Fig 4 top two rows) but far from perfect. The regression loss on a dense distance transform map is also experimented with and the qualitative result is shown in Fig 4 third row. The logo is activated so We mask out irrelevant regions with the foreground network and the result is shown in Fig 4 fourth row. Anyway, although these ideas seem promising they have to

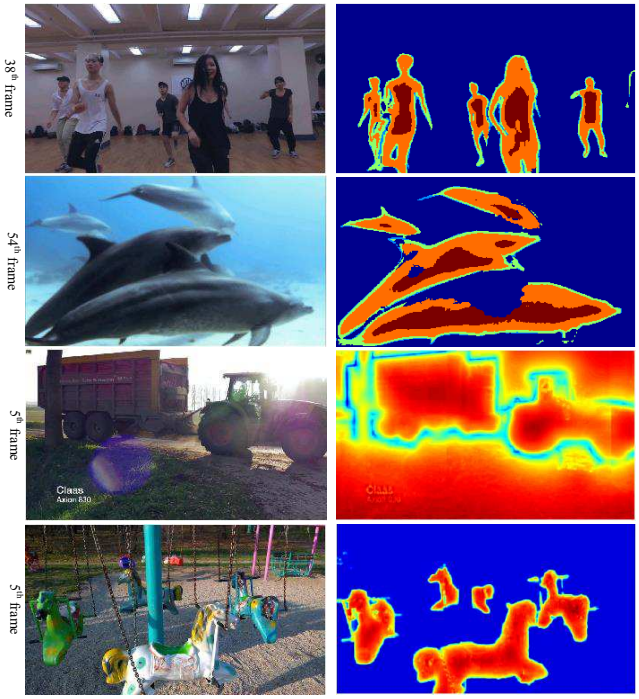

Figure 4. The qualitative results of other formulations.

be combined with down-streaming steps like instance generation/linking (in an ad-hoc manner). We give up these formulations to pursue cleaner ones after these experiments. Specifically, we build several spatio-temperal random field models and solve them with graphc-cuts and mean-field message passing. We also try to learn a deep policy with A3C to adjust their hyper-parameters. Yet these methods quantitatively render the performance lower.

6. Conclusion

This report is not well organized as we write it in a hurry (e.g. we insert some discussions about resolution augmentation into the second section and we don't have time to draw network architecture for the holistically-nested PSPNet). We will provide the baseline training code and the precomputed flow since we guess possibly some researchers in the community would be interested in them.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. In *PAMI 2011*. 3

[2] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In *CVPR 2017*. 3

[3] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR 2017*. 1

[4] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV 2015*. 3

[5] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool. Convolutional oriented boundaries. In *ECCV 2016*. 3

[6] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR 2016*. 1

[7] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV 2016*. 3

[8] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV 2015*. 3

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR 2017*. 1