First International Workshop on Video Segmentation - Panel Discussion

Thomas Brox[1], Fabio Galasso[2,3(✉)], Fuxin Li[4], James Matthew Rehg[4], and Bernt Schiele[2]

[1] University of Freiburg, Freiburg im Breisgau, Germany
[2] Max Planck Institute for Informatics, Saarbrücken, Germany
fabio.galasso@gmail.com
[3] OSRAM Corporate Research, Munich, Germany
[4] Georgia Institute of Technology, Atlanta, GA, USA

Abstract. Interest in video segmentation has grown significantly in recent years, resulting in a large body of works along with advances in both methods and datasets. Progress in video segmentation would enable new approaches to building 3D object models from video, understanding dynamic scenes, robot-object interaction and several other high-level vision tasks. The workshop brought together a broad and representative group of video segmentation researchers working on a wide range of topics. This paper summarizes the panel discussion at the workshop, which focused on three questions: (1) Why does video segmentation currently not meet the performance of image segmentation and what difficulties prevent it from leveraging motion? (2) Is video segmentation a stand-alone problem or should it rather be addressed in combination with recognition and reconstruction? (3) Which are the right video segmentation subtasks the field should focus on, and how can we measure progress?

Keywords: Video segmentation · Computer vision

1 The State of Video Segmentation

While early works on motion segmentation date back to the 1970s, video segmentation has received especially growing interest in computer vision for the last few years, as is witnessed by its increasing presence in mainstream journal and conference publications [1–3,5–18,20–25,28–31]. This interest has led to diverse definitions of the video segmentation problem: some researchers see it as the problem of separating foreground from background while taking into account a potentially moving camera [3,13,22,31]; some see it as identification of moving objects [6,15,20], some as a data simplification method that yields an over-complete supervoxel representation of a video [5,23,29,30], creates and ranks segmentation proposals [2,13,31], or computes hierarchical sets of coarse-to-fine video segmentations [11,21,30].

As a consequence of such diversity, different datasets have been proposed. These include Hopkins 155 on motion segmentation [26], GT-SegTrack (v1 [27]

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part IV, LNCS 8928, pp. 383–388, 2015.
DOI: 10.1007/978-3-319-16220-1_27

and v2 [16]), INRIA-video [15], Youtube segment annotation [25], the Freiburg-Berkeley Motion Segmentation Dataset (FBMS-59) [19], and VSB-100 [8]. The datasets come with different evaluation metrics and annotation standards, reflecting the diverse problem statements: supervoxels, video object proposals, motion segmentation, unsupervised general video segmentation.

2 The Panelists

- Michael Black, MPI for Intelligent Systems
- Irfan Essa, Georgia Institute of Technology
- Vittorio Ferrari, University of Edinburgh
- Cristian Sminchisescu, Lund University
- René Vidal, Johns Hopkins University
- Jue Wang, Adobe Systems

3 Three Open Questions to Initiate the Discussion

- The first open problem stems from a recent observation in Galasso et al. [8] that a relatively *simple* propagation of state-of-the-art image segments over time with a good optical flow method outperforms the tested video segmentation algorithms. Furthermore, almost all tested methods drop significantly in performance when the general video segmentation task (including also non-moving objects) is reduced to a subtask, where only moving objects are required to be segmented (motion subtask [8]). Common sense would expect the segmentation of moving objects to be better defined and *easier* than segmentation of static objects.
- Second, the panelists were asked for their opinion on whether video segmentation should be addressed as a standalone problem or studied in relation with recognition and reconstruction computer vision tasks.
- The third proposed discussion point regarded the particular tasks which video segmentation should address to meet the requirements of potential applications and to serve as intermediate problems that would likely advance progress. What is a good way to measure progress?

4 Panel Discussion

Michael Black suggested looking at the persistent properties of objects in a video, such as material properties of surfaces and object identity. He proposed to consider the recent contributions on intrinsic videos and albedo and to delve into the physical properties of objects in order to characterize them. In a short presentation before the panel discussion he had referred to the recent efforts of his group to introduce a segmentation benchmark based on the open movie Sintel [4], where labels are based on object identity and those parts of an object that differ in material. Michal Irani from the audience expressed doubts that

this would lead to a good definition of the problem. She points out that the annotation was per-frame, thus would not differ if the frames of the video were put in a different order. She compares this with an image that is interpreted just as a set of independent rows, undergoing row-based segmentation, row by row. Michael Black emphasized that the kind of annotation he mentioned *is* temporally coherent. Whether one would use motion segmentation to find such temporally coherent segmentations or some other way to localize the surfaces of the objects, in his opinion, does not change the definition of the problem. Shai Avidan from the audience proposed a stronger three-dimensional reasoning about objects, a suggestion that was later seconded by Cristian Sminchisescu. In contrast, Michael Black was sceptical on whether the reward of this would vanish due to the additional complexity which could introduce new problems.

Irfan Essa commented on the additional difficulties which video segmentation faces when compared to image segmentation. Temporal persistence of the provided segments, occlusions and disocclusions of objects over time especially as objects rotate in space, their appearance and disappearance, the varying size of regions over time are just a few examples he named. He recommended the use of metrics and observations which allow for progress in such tasks, looking beyond the frame-to-frame causality. He added that there is more to the definition of a segment, naming research on perceptual grouping and efforts to understand segments across scales alongside characteristics such as texture. In this respect, he believes that motion or the definition of a temporal tube, are probably not simplifying this complex problem.

The discussion on the task led to the question whether video segmentation is a stand-alone problem or whether it should be addressed alongside reconstruction and recognition tasks. Giving a first introductory statement on this question, Cristian Sminchisescu pointed out that certain problems could definitely be defined as stand-alone problems. Such problems include simple segmentation objectives enforcing continuation properties or forms of spatial layout loss, which could serve the definition of a fine-level detailed segmentation. On similar notes, motion segmentation may find justifications in the biological development of children, who first learn to distinguish to discern motion, and later Gestalt principles such as symmetry and continuity, before they understand the characteristics of simple objects. More generally, however, he believes that interaction with reconstruction and recognition might be essential, one such example being a 3D or 2.5D reconstruction to understand occlusion as opposed to simply tracking superpixels.

With respect to this interplay, Vittorio Ferrari introduced the term "Vision complete". In reference to terms like "NP-complete" and "AI complete", he uses this terminology for problems that require the whole vision problem to be solved before we will see satisfactory solutions. According to him, video segmentation will only be solved once also the other "vision" tasks are solved. He specifically underlined the interplay between segmentation and recognition, which builds upon the human capacity to segment objects from the background thanks to their prior knowledge on object appearance. In this respect, reconstruction may

come into play at a more mature stage of understanding of these problems. In his opinion, a 3D reconstruction of complex videos such as the Sintel movie will definitely come from such virtuous interplay.

Triggered from a comment from the audience, Cristian Sminchisescu said that it is desirable to be robust to different tasks, but he thinks there is a lesson that can be learned from biological systems, which rather aim for sufficiency rather than completeness. A video segmentation approach may not be required to work in all cases as long as it works reliably for the setting it is applied to. That going beyond a single task is desirable is agreed by Vittorio Ferrari.

As the discussion had turned to tasks, Jue Wang made a statement on the third suggested discussion theme: what to evaluate segmentation on. From his point of view, a number of tasks are currently relevant to industry, including video understanding, object extraction and video segmentation for composition. In particular, regarding composition, one important feature to benchmark is temporal consistency independent of accuracy. Supposedly, consistency becomes more difficult when both the object and background move. According to Jue Wang, it is desirable to have different sets of ground truth for different tasks. As an example, video segmentation for recognition could be tested on the base of the final recognition rate.

Picking up on the first question, René Vidal pointed out that judging image segmentation better than the video counterpart could as well be a problem of annotations and metrics, as both are prone to mistakes. One such example is the relevance of the pixel count in most metrics, which clearly favors larger objects. In his opinion, there should be research on metrics and a universal metric is not desirable. Evaluating tasks such as motion segmentation or high precision boundaries for medical imaging in isolation is meaningful as it helps make progress and understand the limits of that task in isolation, a desirable research question. According to him, there is value in addressing tasks both jointly and one-at-a-time.

A further point in the discussion concerned terminology. While terms such as image, motion and video segmentation should be used carefully in their own domain, there is agreement that these concepts can be intended as supersets, with the video segmentation one including the previous two.

Michal Irani suggested video compression as an additional valuable task since video segments should provide the elementary components to best describe the video. Another important related task is action recognition. Cristian Sminchisescu added that intending video segmentation as a layered process might lead to the necessity of different layers for different tasks.

As a final suggestion, Vittorio Ferrari proposed to evaluate video segmentation methods by a relative metric, where humans are asked which of two segmentations is better. The motivation for this is that humans are good at relative assessment compared to absolute ones. In the same the Turing test might not be the perfect indication of machine intelligence, he added, getting the best numbers on a video segmentation benchmark might not indicate the best practical performance.

References

1. Badrinarayanan, V., Budvytis, I., Cipolla, R.: Mixture of trees probabilistic graphical model for video segmentation. IJCV (2013)
2. Banica, D., Agape, A., Ion, A., Sminchisescu, C.: Video object segmentation by salient segment chain composition. In: International Conference on Computer Vision, IPGM Workshop (2013)
3. Bergh, M.V.D., Roig, G., Boix, X., Manen, S., Gool, L.V.: Online video seeds for temporal window objectness. In: ICCV (2013)
4. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012)
5. Chang, J., Wei, D., Fisher, J.W.: A video representation using temporal superpixels. In: CVPR (2013)
6. Dragon, R., Rosenhahn, B., Ostermann, J.: Multi-scale clustering of frame-to-frame correspondences for motion segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 445–458. Springer, Heidelberg (2012)
7. Fragkiadaki, K., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: CVPR (2012)
8. Galasso, F., Nagaraja, N.S., Cardenas, T.J., Brox, T., Schiele, B.: A unified video segmentation benchmark: Annotation, metrics and analysis. In: ICCV (2013)
9. Godec, M., Roth, P.M., Bischof, H.: Hough-based tracking of non-rigid objects. In: ICCV (2011)
10. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR (2010)
11. Jain, A., Chatterjee, S., Vidal, R.: Coarse-to-fine semantic video segmentation using supervoxel trees. In: ICCV (2013)
12. Lee, J., Kwak, S., Han, B., Choi, S.: Online video segmentation by bayesian split-merge clustering. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 856–869. Springer, Heidelberg (2012)
13. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV (2011)
14. Levinshtein, A., Sminchisescu, C., Dickinson, S.: Spatiotemporal closure. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 369–382. Springer, Heidelberg (2011)
15. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: Spatio-temporal video segmentation with long-range motion cues. In: CVPR (2011)
16. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV (2013)
17. Ma, T., Latecki, L.J.: Maximum weight cliques with mutex constraints for video object segmentation. In: CVPR (2012)
18. Maire, M., Yu, S.X.: Progressive multigrid eigensolvers for multiscale spectral segmentation. In: ICCV (2013)
19. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. PAMI (2014)
20. Ochs, P., Brox, T.: Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In: ICCV (2011)

21. Palou, G., Salembier, P.: Hierarchical video representation with trajectory binary partition tree. In: CVPR (2013)
22. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013)
23. Reso, M., Jachalsky, J., Rosenhahn, B., Ostermann, J.: Temporally consistent superpixels. In: ICCV (2013)
24. Sundaram, N., Keutzer, K.: Long term video segmentation through pixel level spectral clustering on gpus. In: ICCV Workshops (2011)
25. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: CVPR (2013)
26. Tron, R., Vidal, R.: A benchmark for the comparison of 3-D motion segmentation algorithms. In: CVPR (2007)
27. Tsai, D., Flagg, M., Rehg, J.M.: Motion coherent tracking with multi-label mrf optimization. In: BMVC (2010)
28. Vazquez-Reina, A., Avidan, S., Pfister, H., Miller, E.: Multiple hypothesis video segmentation from superpixel flows. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 268–281. Springer, Heidelberg (2010)
29. Xu, C., Corso, J.J.: Evaluation of super-voxel methods for early video processing. In: CVPR (2012)
30. Xu, C., Xiong, C., Corso, J.J.: Streaming hierarchical video segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 626–639. Springer, Heidelberg (2012)
31. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR (2013)