Supervoxel Approach in an Efficient Multiscale Object-Based Framework

Lucca S.P. Lacerda, Felipe D. Cunha, Silvio Jamil F. Guimarães

Image and Multimedia Data Science Laboratory (IMScience)
Computer Science Department (DCC) – Institute of Sciences Exacts and Informatics
Pontifical Catholic University of Minas Gerais (PUC Minas)
Belo Horizonte, MG, Brazil
luccasplacerda@gmail.com – {felipe,sjamil}pucminas.br

***Abstract*—The use of supervoxel segmentation has shown substantial improvement in video analysis because it can improve object delineation and reduce computer workload. In this work, we have used SICLE (Superpixel Through Iterative Clearcutting), which is an innovative graph-based superpixel framework that makes multi-scale segmentation by exploiting object information. For segmenting videos, we changed the graph creation step. The framework has exceeded state-of-the-art approaches, and its results precisely delineate the object of interest.**

I. Introduction

To analyze the segmentation of images and videos, normally, it is necessary to segment the objects from their background. This sort of study can be expensive to the computer whilst providing redundant and low-context information. A way of doing this more efficiently is by generating groups of several linked pixels or voxels ($i.e.$, superpixels or supervoxels) that have certain similar properties ($e.g.$, texture and color). Because of these properties, this type of segmentation is regularly applied, such as in exam detections.

In video supervoxel segmentation, four properties are ideally required [1]: (i) spatiotemporal boundaries coherence; (ii) computational efficiency; (iii) hierarchical spatiotemporal segmentation; and (iv) an arbitrary number of supervoxels. Except, there is yet to be a supervoxel segmentation algorithm with all of those. Some existing methods, like Graph-Based Supervoxel (GB) and MeanShift, struggle to determinate the relevance of the supervoxels alongside high segmentation accuracy. And methods like Hierarchical Graph-Based Supervoxel (GBH) and Segmentation by Weighted Aggregation (SWA) solve the problem presented above, but are too computationally costly.

In image superpixel segmentation, SICLE [2] has achieved impressive results in several data-sets. The Superpixel through Iterative Clearcutting (SICLE) is a generalization of two state-of-art methods that “starts off from oversampling and, through several iterations, generates superpixels from the seed set and remove a portion of irrelevant seeds to preserve the accurate object delineation from the previous iteration”, being seed a chosen pixel on the image.

The advantage of SICLE over other methods is that it is possible to make an analysis giving precise boundaries whilst having low computational cost. Thereby, trying to create a technique that can meet the requirements of different vision tasks, in this project, SICLE was upgraded to video segmentation, named after this as VI-SICLE1. The main goals are to be able to represent video as a graph and, through that, analyze if it keeps spatiotemporal coherence; and to use Image Foresting Transform ($i.e.$, IFT) to reduce the seeds to an exact number of supervoxels in each frame.

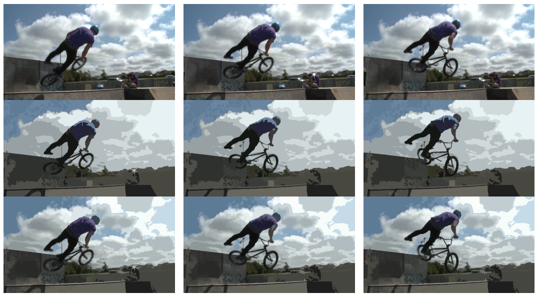

Fig. 1. Illustrated example of video segmentation using VI-SICLE changing the seed sampling. Video extracted from the SecTrackV2 data-set. The original frames are in the first row. Then, results for (i) 20 supervoxels and (ii) 200 supervoxels are in the following rows. Each resulting region is colored by its mean color.

This paper is organized as follows. In Section 2, the concepts used in VI-SICLE are explained. In Section 3, the methodology for the proposed segmentation approach is described. In Section 4, the results obtained are compared with other methods. Finally, Section 5 is the closure of this paper with suggestions for future work.

II. SICLE

SICLE operates by seed oversampling and repeating connectivity-based superpixel delineation and object-based seed removal until reaching the desired number of superpixels. And, since SICLE is the convergence of a classic and an object-based methods frameworks, it is essential to explain them briefly.

Classic methods, like Iterative Spanning Forest (ISF), use pixel clustering to generate superpixels in a short span of time, but they fail to ensure the desired number of superpixels

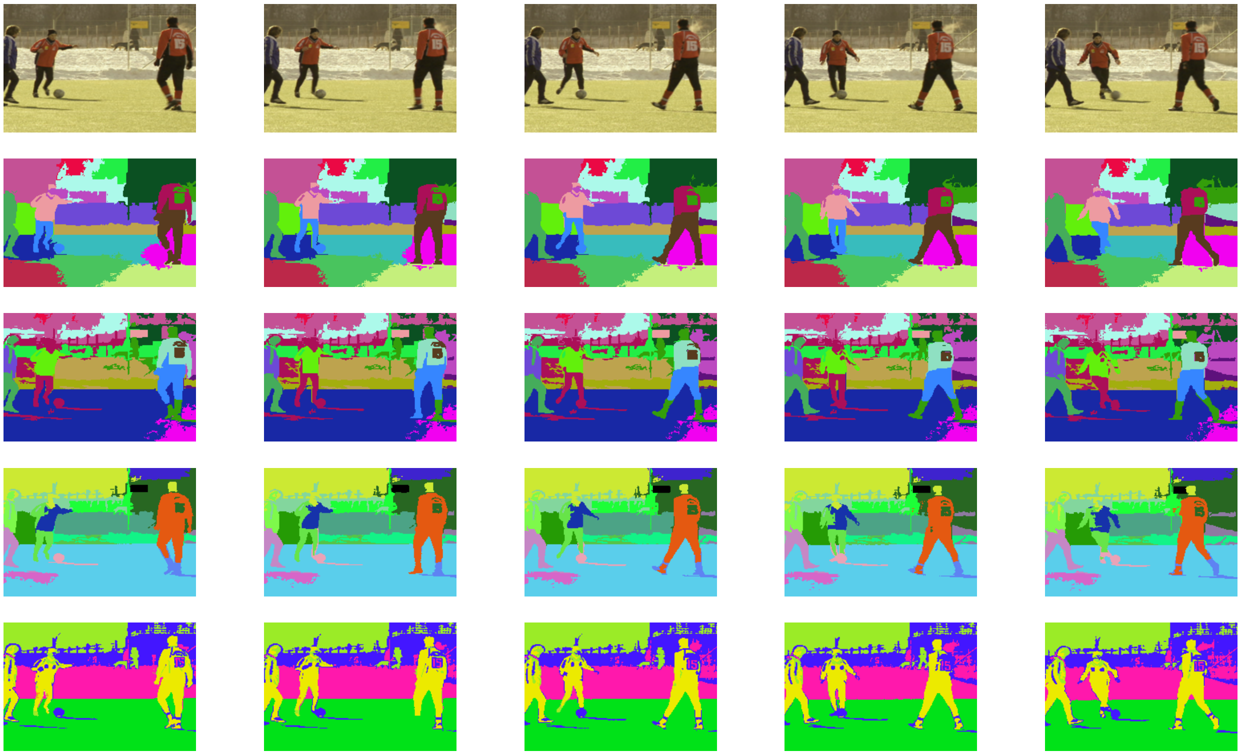

Fig. 2. Example extracted from Chen data-set. The first row is the original frames. From top to bottom, the following rows are results with 20 supervoxels obtained from GB, GBH, ISF2SVX-GRID-DYN, and VI-SICLE.

and have only moderate delineation. On the other hand, some object-based ones — such as Object-based ISF (OISF) — can control the superpixel displacement and morphology. Still, they are slow and highly dependable on the saliency map quality.

Considering that, an Object-based Dynamic Iterative Spanning Forest (ODISF) was created — using a method of (i) seed oversampling, (ii) superpixel generation, and (iii) object-based seed removal — to be more accurate and faster than the aforementioned frameworks and have a minimum influence of saliency errors [3]. Finally, upgrading the seed removal method of ODISF, SICLE was developed to achieve adequate delineation for all objects tested in Belém's [2] work, irrespective of map saliency errors.

III. Methodology

This section is going to detail the approach to the proposed framework for video, based on SICLE. The proposal follows a four-step methodology: (i) graph creation; (ii) seed oversampling; (iii) supervoxel generation; and (iv) seed removal criterion3, which is very similar to the used on SICLE.

A. Graph Creation

Differently from 2D and 3D images, the presence of the same object in-between frames imposes a major challenge for generating temporally coherent supervoxels. Therefore, it is recommended that the arcs and their respective arc-costs should reflect such conditions. In this project, we operate on a video $\text{V} = (\mathcal{V}, \boldsymbol{I})$ whose graph $\text{G} = (\mathcal{J}, \mathcal{H}, \boldsymbol{I})$ is modeled as a single volume of vertices, and the outgoing arcs $(s, t) \in \mathcal{H}$ of a vertex s, for any $t \in \mathcal{J}$ which $s \neq t$, are defined, for instance, by an adjacency relation ($e.g.$, 26-adjacency) [4]. So, the main difference between VI-SICLE and the original SICLE framework is that it can compare with its adjacents image-frames to keep the coherence of the supervoxels throughout the video. Both, SICLE and VI-SICLE, have the computational cost related to the graph formation, so for the segmentation the cost will be the same.

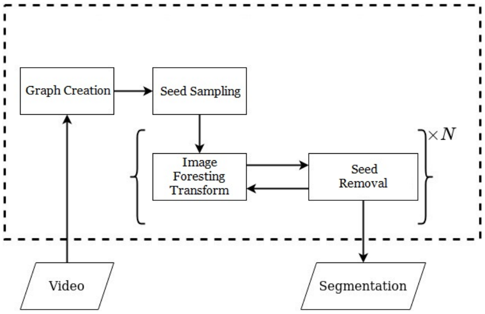

Fig. 3. Diagram of the proposed segmentation framework.

B. Seed Oversampling

Even though both seed-based and object-based algorithms aim to be precise with their sampling, the initial number of supervoxels (N_0) should be close to the desired number(N_f) because it leads to a more accurate delineation, the seed-based methods favor, paradoxically, sampling fast but imprecisely, and the object-based ones have a more precise but slower sampling algorithm. Therefore, considering the algorithm's efficiency primarily, it is counter-productive to attend to its precision.

As opposed to that, trying to have better results, by favoring recall over precision, VI-SICLE presents oversampling ($N_0 \gg N_f$). For the reason that high amounts of seeds have the potential to reach the boundary areas, it's reasonable to remove the irrelevant ones as a group, since nearby seeds with similar features will have similar relevance and will stay together. So, it is possible to assume that oversampling surpasses the drawbacks of using fast but imprecise sampling strategies. Thus, the main goal is to guarantee the presence of all the boundary seeds while removing the irrelevant ones throughout the iterations to ensure N_f supervoxels in the last iteration. And unlike traditional methods, this work has random seed oversampling (RND) because of its implementation simpleness, especially when there is an uneven area to be segmented.

C. Supervoxel Generation

In this project, supervoxels are generated by the seed-restricted version of the IFT framework. In trying to promote stability, it is better to consider the seed's features over its tree's mean vector since their features are immutable throughout the execution of the IFT. Beyond that, given that supervoxels (superpixels) minimize dissimilarity, it is arguable that the tree's features may resemble its seed's. For this reason, both arc-cost functions may present similar delineations, especially in object borders. Therefore, VI-SICLE uses of a root-based function, henceforth named ROOT [2].

D. Seed Removal Criterion

It is necessary to remove $N_0 - N_f$ seeds for generating the desired number of supervoxels in the final IFT execution. VI-SICLE establishes if a seed is relevant or not, based on its particular features and temporal coherence. It uses the information generated from the previous IFT execution, at each iteration, deciding through competition and supervoxel features which seeds are of relevance. In the VI-SICLE approach, seeds with the highest relevance are consistently identified.

It is anticipated that the first iteration will have plenty regions containing many equally relevant seeds and several irrelevant seeds. Over the iterations, the impact will be minimal when removing the least important seeds, due to the sample size. And as the number of seeds decreases, the more ”equal importance” they become; hence, the less relevant ones will be defined in relation to their pairs. From there, to avoid supervoxel instability and delineation errors, it is necessary to remove smaller quantities of seeds per iteration.

Akin to SICLE, the criterion to select the most relevant seeds is based on the characteristics of its supervoxels and the closeness to the object in question. In order to delimit the object without losing its boundaries, since the background tends to be bigger, but more uniform than the object, a solution is to evaluate the contrast between them. Another problem is that, just selecting the regions of high-contrast, though, might not be enough, once it can indicate noises and well-defined borders, and low-contrast ones may indicate voxels within objects or poorly-defined borders.

This way, it is necessary to use two new criterion that combine the size and the contrasts models. Both should consider large supervoxels. While one prioritizes minimum contrast amongst their adjacents, the other aims for maximum contrast. But that is not all. Since the object of interest is the main goal on the segmentation, to have a more accurate delineation, more supervoxels must be around (i.e., within or near) this object. So, the probable object location may help the choice of what seeds are not close by the object of interest.

IV. Results

Based on Jerônimo [4] analysis, it was possible to compare VI-SICLE results with other state-of-art segmentation frameworks, as demonstrated in Figures 2. This approach proposes a video version of SICLE using RND and ROOT (iv), and it is compared with (i) GB, (ii) GBH, and (iii) ISF2SVX-GRID-DYN.

The VI-SICLE was tested with different parameters and the best result was the RND and ROOT version. In this version, the results were more detailed and spending less supervoxels with the background.

In Figures 2 e 4, it is possible to observe the comparison of different frameworks in a single video. In general, VI-SICLE is able to, in contrast with GB, GBH, and ISF2SVX-GRID-DYN, be much more precise in creating larger supervoxels in the areas without objects of interest, mainly the immovable objects in the scene ($e.g.$, the snow and street signs).

When comparing the results, the disparity between this proposal's clarity in relation to the rest of the frameworks shows that the others generate too many small supervoxels in non-interest areas. At the same time, this one still created big supervoxels for the aforementioned areas. Finally, it is feasible to infer with these results that, while SICLE has proved to be as efficient as the ISF2SVX-GRID-DYN with small amounts of supervoxels, it was able to perform extremely better with a large sample.

V. Conclusion

In this work, developed as an undergraduate work, we have studied the use of a multi-scale segmentation for supervoxel generation. The proposed strategy was inspired by the “SICLE”, which is a superpixel framework. It oversamples and, with sequential iterations, creates supervoxels from the seed set, removing a part of the irrelevant seeds to preserve the precise delineation of the object from the previous iteration. Then, without dropping out of that calculus, re-run the framework for another image frame, measuring the temporal coherence.

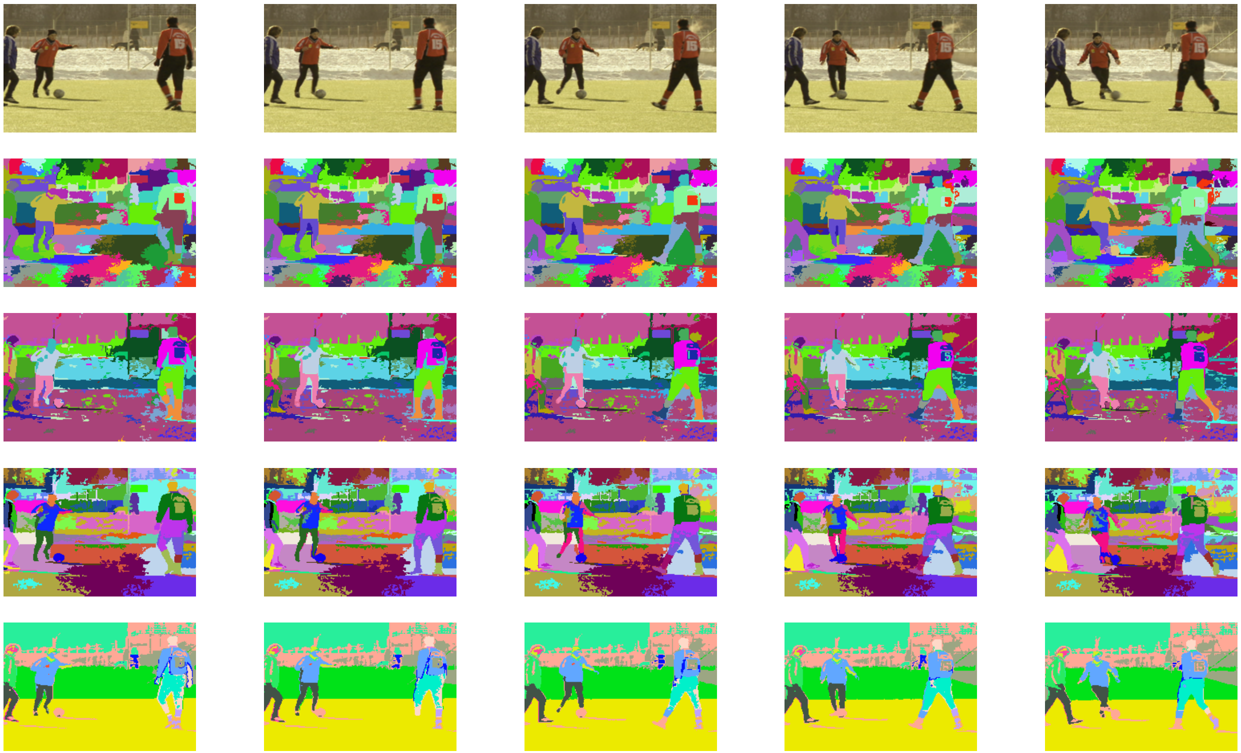

Fig. 4. Example extracted from Chen data-set. In the first row is the original frames, the following rows, from top to bottom, are results with 100 supervoxels obtained from GB, GBH, ISF2SVX-GRID-DYN, and VI-SICLE.

The obtained results, in terms of qualitative analysis, show good accuracy, especially in terms of delineation and color description (by its supervoxels). And the results also show the spatiotemporal coherence with different number of exact supervoxels, in such way that, since it ignores considerably the non-relevant areas, it would be possible to try out in more complex videos ($i.e.$ medical and botanical exams).

For future works, we would like to use this strategy for video object delienation by using a saliency object map as a prior. Moreover, we will quantitatively compare the proposed method to the state-of-the-art.

ACKNOWLEDGMENT

The authors thank the Pontifícia Universidade Católica de Minas Gerais – PUC-Minas, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES – (Grant COFECUB 88887.191730/2018-00, Grant PROAP 88887.842889/2023-00 – PUC/MG and Finance Code 001), the Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Grants 407242/2021-0, 306573/2022-9) and Fundação de Apoio à Pesquisa do Estado de Minas Gerais – FAPEMIG (Grant APQ-01079-23).

REFERENCES

[1] B. Wang, Y. Chen, W. Liu, J. Qin, Y. Du, G. Han, and S. He, "Real-time hierarchical supervoxel segmentation via a minimum spanning tree," *IEEE Trans. Image Process.*, vol. 29, pp. 9665–9677, 2020. [Online]. Available: https://doi.org/10.1109/TIP.2020.3030502

[2] F. Belém, B. Perret, J. Cousty, S. J. F. Guimarães, and A. X. Falcão, "Efficient multiscale object-based superpixel framework," *CoRR*, vol. abs/2204.03533, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.03533

[3] F. de Castro Belém, B. Perret, J. Cousty, S. J. F. Guimarães, and A. X. Falcão, "Towards a simple and efficient object-based superpixel delineation framework," in *34th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2021, Gramado, Rio Grande do Sul, Brazil, October 18-22, 2021.* IEEE, 2021, pp. 346–353. [Online]. Available: https://doi.org/10.1109/SIBGRAPI54419.2021.00054

[4] C. S. J. de Almeida, F. Belém, S. A. Carneiro, Z. K. G. do Patrocínio, L. Najman, A. X. Falcão, and S. J. F. Guimarães, "Graph-based supervoxel computation from iterative spanning forest," in *Discrete Geometry and Mathematical Morphology - First International Joint Conference, DGMM 2021, Uppsala, Sweden, May 24-27, 2021, Proceedings*, ser. Lecture Notes in Computer Science, J. Lindblad, F. Malmberg, and N. Sladoje, Eds., vol. 12708. Springer, 2021, pp. 404–415. [Online]. Available: https://doi.org/10.1007/978-3-030-76657-3_29