EVALUATION ON THE COMPACTNESS OF SUPERVOXELS

Ran Yi[†] *Yong-Jin Liu*[†] *Yu-Kun Lai*[⋆]

[†] Tsinghua University, China
[⋆] Cardiff University, UK

ABSTRACT

Supervoxels are perceptually meaningful atomic spatiotemporal regions in videos, which has great potential to reduce the computational complexity of downstream video applications. Many methods have been proposed for generating supervoxels. To effectively evaluate these methods, a novel supervoxel library and benchmark called LIBSVX with seven collected metrics was recently established. In this paper, we propose a new *compactness* metric which measures the shape regularity of supervoxels and is served as a necessary complement to the existing metrics. To demonstrate its necessity, we first explore the relations between the new metric and existing ones. Correlation analysis shows that the new metric has a weak correlation with (i.e., nearly independent of) existing metrics, and so reflects a new characteristic of supervoxel quality. Second, we investigate two real-world video applications. Experimental results show that the new metric can effectively predict some important application performance, while most existing metrics cannot do so.

***Index Terms*—** Supervoxel, compactness, video segmentation, metric evaluation

1. INTRODUCTION

Supervoxels are perceptually meaningful atomic spatiotemporal regions in videos, which are obtained by grouping similar voxels. Here *similarity* is defined in terms of coherence in both appearance and motion in a video. Instead of voxels, using supervoxels as basic elements has great potential to reduce the complexity of downstream video applications, e.g., foreground object segmentation [1] and spatiotemporal closures in videos [2], etc.

Many methods have been proposed to generate supervoxels with different characteristics (e.g., [3–10]). To effectively evaluate these methods, a novel supervoxel library and benchmark called LIBSVX was recently established [11]. LIBSVX includes six video datasets with a variety of content types and dense human annotations as ground truth. Seven widely used metrics are also collected in LIBSVX for evaluating the performance of different supervoxels, which are briefly summarized below.

Existing metrics. 3D under-segmentation error (UE3D), 3D segmentation accuracy (SA3D) and boundary recall distance (BRD) are three standard metrics for measuring supervoxels' adherence to object boundaries [10–12]. UE3D and SA3D measure the tightness of supervoxels that overlap with ground-truth segmentation. BRD measures how well the ground-truth boundaries are correctly retrieved by the supervoxel boundaries. Smaller UE3D and BRD values, and higher SA3D values indicate better supervoxels. Label consistency (LC) [10] measures how well supervoxels track parts of objects and can only be evaluated on datasets with ground-truth optical flow; therefore, we do not compare LC in this paper. The evaluation of UE3D, SA3D, BRD and LC relies on dense human annotations in videos. Three human-independent metrics, including explained variation (EV) [13], mean size variation (MSV) [10] and temporal extent (TEX) [6, 14], are also widely used. EV measures the color variations in supervoxels and a large EV indicates that the color in each supervoxel is close to homogeneity. MSV and TEX measure the size variation and average temporal extent of all supervoxels in a video.

Our contributions. None of the existing metrics consider the shape regularity of supervoxels. By replacing voxels by supervoxels, many real-world video applications (e.g., [1, 2]) construct a spatiotemporal supervoxel graph and minimize some energy functions. The shape regularity of supervoxels has a direct influence on the complexity of this spatiotemporal supervoxel graph, and therefore, directly affects the performance in these applications. In this paper, we propose a new *compactness* metric (CP) with two possible formulas to measure the goodness of supervoxels. A compact supervoxel with a high CP value has a regular shape and smooth boundary. Two contributions are made in this paper. First, correlation analysis is presented, showing that the CP metric has a weak correlation with (i.e., nearly independent of) existing metrics, and so reflects a new characteristic of supervoxel quality. Second, two real-world video applications are selected, and experimental results on them show that the CP metric can effectively predict some important application performance, while most existing metrics cannot do so. These

Y.J. Liu is the corresponding author. This work was supported by the Natural Science Foundation of China (61661130156, 61725204), BNRist and the Royal Society-Newton Advanced Fellowship (NA150431).

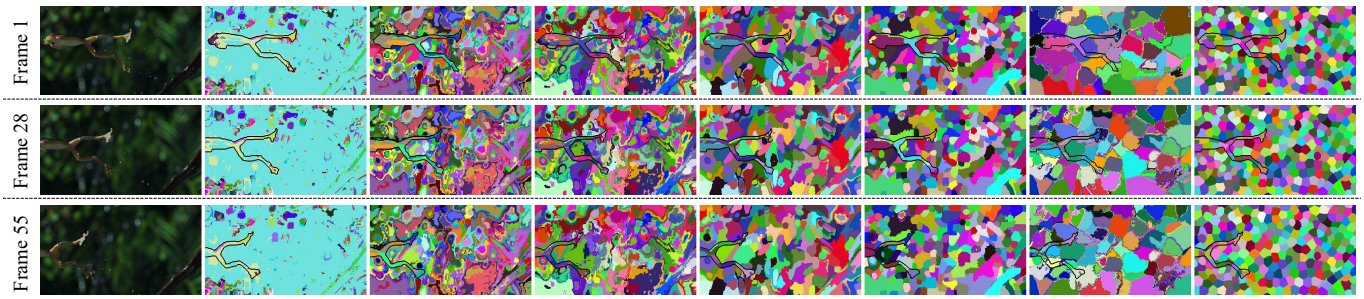


Fig. 1. Qualitative results of seven representative supervoxels on a video clip from the SegTrack v2 dataset [15]: GB [4], GBH [5], streamGBH [6], SWA [7], MeanShift [8], NCut [9], and TSP [10]. Supervoxels are illustrated by clipping them on each frame and a color indicates a supervoxel. The results clearly show that TSP has the best compactness and GB has the worst compactness. The values of both metrics CP_1 and CP_2 are also presented, which coincide with the qualitative results.

two contributions reveal that the new proposed CP metric is a necessary complement to existing metrics.

2. COMPACTNESS METRICS FOR SUPERVOXELS

In this section, we propose two compactness metrics which measure the shape regularity of supervoxels. The weak correlation between them and existing metrics is verified in Section 3. Their ability to predict some important application performance on two selected video applications is presented in Section 4.

2.1. Metric CP_1

Our first compactness metric CP_1 makes use of the 3-dimensional isoperimetric inequality:

$$Area(\Omega) \geq 3Vol(\Omega)^{\frac{2}{3}} Vol(B_1)^{\frac{1}{3}} \tag{1}$$

where $\Omega \subset \mathbb{R}^3$ is a connected region, B_1 is a unit ball, $Area(\Omega)$ and $Vol(\Omega)$ are bounding surface area and volume of Ω respectively. In formula (1), the equality holds when Ω is a ball.

The relation[1] between isoperimetric quotient and shape regularity can be explained by the well known physical phenomenon: When the volume of water in a drop is fixed, the surface tension will force the drop into a smooth shape (i.e., a round sphere which is most regular) by minimizing the surface area of the drop.

Let $S = \{s_1, s_2, \cdots, s_K\}$ be a given set of K supervoxels which over-segments a video clip. We define

$$CP_1(S) = \sum_{s_i \in S} \frac{|s_i|}{N} Q_1(s_i), \tag{2}$$

where

$$Q_1(s_i) = \frac{6\sqrt{\pi} Vol(s_i)}{Area(s_i)^{\frac{3}{2}}}, \tag{3}$$

[1]For a visual illustration, see http://demonstrations.wolfram.com/IsoperimetricInequalityForPolygons/.

$|s_i|$ is the number of voxels in s_i, N is the number of voxels in the entire video, and $Q_1(s_i)$ is the isoperimetric quotient for supervoxel s_i. The coefficient $\frac{|s_i|}{N}$ in (2) makes each supervoxel contribute to the compactness metric adaptively by its own size. The value of the metric CP_1 has a normalized range $[0, 1]$. The larger this value is, the more regular the shape of supervoxels is.

2.2. Metric CP_2

Our second compactness metric CP_2 is based on the same observation as used in isoperimetric inequality: Given a fixed value of bounding surface area, a larger number of voxels in a supervoxel (equally the larger volume of a supervoxel) indicates a higher compactness value. However, CP_2 makes use of a ratio between the number $|B(s_i)|$ of boundary voxels $B(s_i)$ and the number $|s_i|$ of total voxels in a supervoxel s_i, which is much simpler to evaluate than CP_1:

$$CP_2(S) = 1 - \sum_{s_i \in S} \frac{|s_i|}{N} Q_2(s_i), \tag{4}$$

where

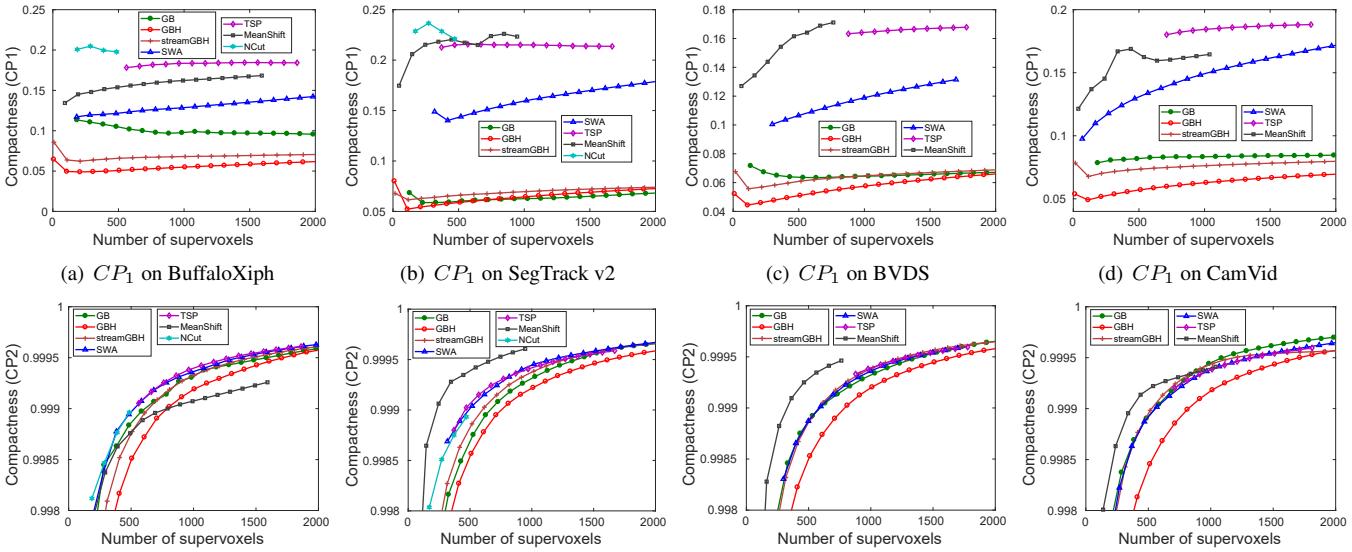
$$Q_2(s_i) = \frac{|B(s_i)|}{|s_i|} \tag{5}$$

The value of the metric CP_2 also has a normalized range $[0, 1]$, where

- 0 is reached when all voxels in each s_i are boundary voxels, meaning that all s_i have a very curved, narrow-banded shape, and
- 1 is reached when the volume of regular supervoxel (such as spherical or cubic shapes) approaches infinity.

Similar to CP_1, a higher value of CP_2 means that supervoxels are more compact.

Both metrics of CP_1 and CP_2 can effectively distinguish the shape regularity of different supervoxels. Figure 1



(a) CP_1 on BuffaloXiph (b) CP_1 on SegTrack v2 (c) CP_1 on BVDS (d) CP_1 on CamVid

(e) CP_2 on BuffaloXiph (f) CP_2 on SegTrack v2 (g) CP_2 on BVDS (h) CP_2 on CamVid

Fig. 2. Comparison of the discriminative powers of CP_1 and CP_2 for seven representative supervoxels on four video datasets: BuffaloXiph [16], SegTrack v2 [15], BVDS [17, 18] and CamVid [19].

shows an example, in which from qualitative results of seven representative supervoxels, one can clearly distinguish that TSP [10] obviously has the best compactness and GB [4] has the worst compactness. The values of CP_1 and CP_2 coincide with this observation.

2.3. Comparison of CP_1 and CP_2

Although both CP_1 and CP_2 can indicate the shape regularity of supervoxels, they have different discriminative powers. As shown in Figure 1, the differences among CP_1 values is significantly larger than those among CP_2 values.

We further investigate the discriminative powers of CP_1 and CP_2 for the seven representative supervoxels (GB [4], GBH [5], streamGBH [6], SWA [7], MeanShift [8], NCut [9], and TSP [10]) on four video datasets, i.e., BuffaloXiph [16], SegTrack v2 [15], BVDS [17, 18] and CamVid [19], which have human-annotated groundtruth labels. The results are illustrated in Figure 2, which clearly show that CP_1 can better separate the compactness values of different supervoxels than CP_2, and thus, has a better discriminative power. In the remainder of this paper, we only evaluate CP_1.

3. CORRELATION WITH EXISTING METRICS

We introduce the compactness metric to emphasize an important property of supervoxels — shape regularity, which has not been systematically evaluated before. But compactness is not expected to be decisive for a good supervoxel algorithm by itself. We propose this new metric not to replace the existing metrics, but as a necessary complement to them. When two supervoxel algorithms have similar performance on existing metrics, we prefer the one with better compactness, because a method with higher compactness values generates more regularly-shaped supervoxels.

Table 1. Correlation analysis for compactness CP_1 and existing supervoxel metrics on the BuffaloXiph dataset (the number indicates the correlation coefficient). Similar performance is observed on the other three datasets SegTrack v2, BVDS and CamVid.

	UE3D	SA3D	BRD	EV	MSV	TEX
CP_1	-0.23	-0.08	0.18	0.14	-0.03	-0.52

To explore the relation between compactness CP_1 and existing metrics, we conduct a correlation analysis between them (see Table 1). The results indicate weak correlation between CP_1 and UE3D, SA3D, BRD, EV and MSV. The medium negative correlation between CP_1 and TEX is because the temporal extent of supervoxels affects their shape to some extent. Compactness is nearly independent from other metrics, which indicates its ability to reflect a new aspect of supervoxel quality.

4. APPLICATION PERFORMANCE PREDICATION

In many real-world video applications, the solution relies on minimizing an energy function defined on a spatiotemporal supervoxel graph in a video clip. The shape regularity of supervoxels has a direct influence on the complexity of this spatiotemporal supervoxel graph, and thus, affects the application performance. Dependent on different energy forms, the application performance affected by supervoxels' compactness can be either the running time (Section 4.1) or the accuracy (Section 4.2).

Table 2. Performance of foreground propagation task on two supervoxels GBH and TSP, averaged on Youtube-Objects Dataset [20] (Best results are shown in bold).

Method	Running time		Accuracy			
	Time (sec)	CP_1	UE3D	SA3D	BRD	F_1
GBH	126.62	0.0439	3.5781	**0.8929**	**1.8567**	**0.7409**
TSP	**99.28**	**0.1795**	**1.7952**	0.8705	3.7914	0.7232

4.1. Foreground propagation

Given the first frame with an annotated foreground object, Jain and Grauman [1] propose a novel method to propagate the foreground region through time, by using supervoxels to obtain long-term coherent estimates. A spatio-temporal graph was constructed based on supervoxels and optical flow, in which a Markov random field is developed with a well-defined energy function consisting of unary, pairwise and higher order potentials. The energy is then minimized by α-expansion and iteratively updating the likelihood functions using label estimates.

The accuracy of this energy minimization solution (evaluated by the F-measure F_1) depends on the over-segmentation accuracy of supervoxels, which can be indicated by the metrics UE3D, SA3D and BRD in a comprehensive way. On the other hand, supervoxels' compactness affects the time complexity of this solution. More compact supervoxels tend to construct a simpler spatio-temporal graph owing to simpler neighborhood relationships. The reduction of combinatorial complexity of the graph leads to the reduction of the computational cost of the energy function and processing time of energy minimization.

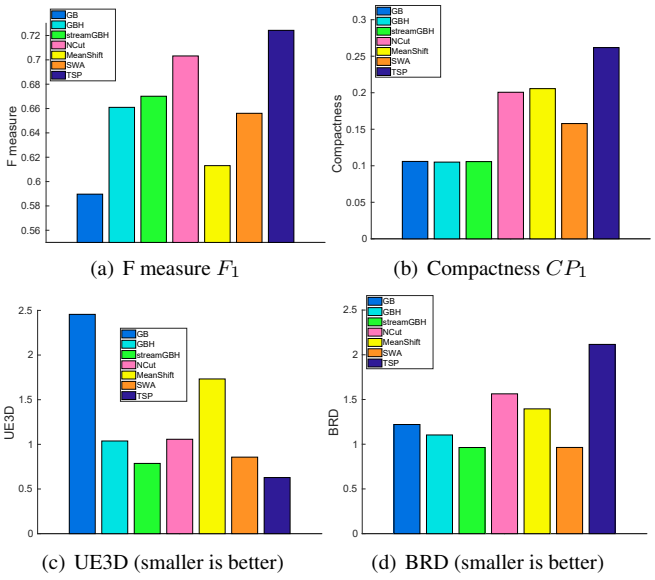
We verify this relation between compactness and processing time by experiments (see Table 2, tested on a PC with Intel Core E5-2683V3 and 256GB RAM). Given the same number of supervoxels, GBH result with non-compact supervoxels takes more time to minimize the energy function than compact TSP.

4.2. Optimal video closure

Levinshtein et al. [2] propose a novel spatiotemporal closure detection method to separate an object from background in a video clip. Based on the same spatio-temporal graph as summarized in Section 4.1, spatiotemporal closure detection is formulated as finding a subset of supervoxels that minimizes a spatiotemporal closure cost over the graph.

Minimizing this normalized cut exactly is NP-complete [22]. An approximation solution using parametric maxflow is applied in [2]. The accuracy of this approximation solution depends on the compactness of supervoxels: More compact supervoxels tend to have more compact subsets and hereby better results.

We verify this observation by experiments. For seven presentative supervoxels, their metric values of CP_1, UE3D and BRD, averaged on Stein et al.' dataset [21], are illustated in Figure 3, in which their performance in video closure (evaluated by F-measure F_1) is also presented. The results clearly show that TSP achieves the best performance simultaneously on average F-measure F_1 and compactness CP_1. Further experiments for investigating relations between F_1 and supervoxel metrics are conducted; see Table 3 for correlation coefficients. The results show that both UE3D and CP_1 have strong correlation with F_1. When supervoxels have similar performance on UE3D (e.g., GBH and NCut), the difference on CP_1 metric can further justify the goodness of different supervoxels (e.g. NCut with larger CP_1 also has higher F_1).



(a) F measure F_1 (b) Compactness CP_1

(c) UE3D (smaller is better) (d) BRD (smaller is better)

Fig. 3. The measures of CP_1, UE3D and BRD for seven representative supervoxels, and their performance (evaluated by F-measure F_1) in the video closure application [2], averaged over Stein et al.' dataset [21].

Table 3. Correlation analysis between F measure F_1 of video closure and supervoxel metrics.

	UE3D	BRD	SA3D	CP_1
F_1	-0.8061	0.2896	-0.0468	0.5385

5. CONCLUSION

In this paper, we propose a new metric of two possible forms CP_1 and CP_2 to measure the shape regularity of supervoxels. Their discriminative power is analyzed by comparing seven representative supervoxels on various datasets. We also investigate the relation between CP_1 and existing metrics, revealing that CP_1 reflects a new aspect of supervoxel quality. We further demonstrate the effect of compactness measure with two video applications, showing that CP_1 is a necessary complement to existing metrics.

6. REFERENCES

[1] Suyog Dutt Jain and Kristen Grauman, "Supervoxel-consistent foreground propagation in video," in *European Conference on Computer Vision*, 2014, pp. 656–671.

[2] Alex Levinshtein, Cristian Sminchisescu, and Sven Dickinson, "Optimal image and video closure by superpixel grouping," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 99–119, 2012.

[3] Ran Yi, Yong-Jin Liu, and Yu-Kun Lai, "Content-sensitive supervoxels via uniform tessellations on video manifolds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[5] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa, "Efficient hierarchical graph-based video segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2141–2148.

[6] Chenliang Xu, Caiming Xiong, and Jason J. Corso, "Streaming hierarchical video segmentation," in *European Conference on Computer Vision*, 2012, vol. VI, pp. 626–639.

[7] Jason J. Corso, Eitan Sharon, Shishir Dube, Suzie El-Saden, Usha Sinha, and Alan Yuille, "Efficient multilevel brain tumor segmentation with integrated Bayesian model classification," *IEEE Trans. Med. Imaging*, vol. 27, no. 5, pp. 629–640, 2008.

[8] Sylvain Paris and Fredo Durand, "A topological approach to hierarchical segmentation using mean shift," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[9] Charless C. Fowlkes, Serge J. Belongie, Fan R. K. Chung, and Jitendra Malik, "Spectral grouping using the nyström method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, 2004.

[10] Jason Chang, Donglai Wei, and John W. Fisher III, "A video representation using temporal superpixels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2051–2058.

[11] Chenliang Xu and Jason J. Corso, "LIBSVX: A supervoxel library and benchmark for early video processing," *Int. J. Comput. Vision*, vol. 119, no. 3, pp. 272–290, 2016.

[12] Alex Levinshtein, Adrian Stere, Kiriakos N. Kutulakos, David J. Fleet, Sven J. Dickinson, and Kaleem Siddiqi, "TurboPixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, 2009.

[13] Alastair Philip Moore, Simon Prince, Jonathan Warrell, Umar Mohammed, and Graham Jones, "Superpixel lattices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[14] Chenliang Xu and Jason J. Corso, "Evaluation of supervoxel methods for early video processing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1202–1209.

[15] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg, "Video segmentation by tracking many figure-ground segments," in *IEEE International Conference on Computer Vision*, 2013, pp. 2192–2199.

[16] Albert Y. C. Chen and Jason J. Corso, "Propagating multi-class pixel labels throughout video frames," in *Western New York Image Processing Workshop*, 2010.

[17] Patrik Sundberg, Thomas Brox, Michael Maire, Pablo Arbeláez, and Jitendra Malik, "Occlusion boundary detection and figure/ground assignment from optical flow," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2233–2240.

[18] Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jimenez Cárdenas, Tatiana Jimenez Brox, and Bernt Schiele, "A unified video segmentation benchmark: Annotation, metrics and analysis," in *IEEE International Conference on Computer Vision*, 2013, pp. 3527–3534.

[19] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *European Conference on Computer Vision*, 2008, pp. 44–57.

[20] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3282–3289.

[21] Andrew Stein, Derek Hoiem, and Martial Hebert, "Learning to find object boundaries using motion cues," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[22] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.