© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. This is the accepted version of: László Kopácsi, Áron Fóthi, Ádám Fodor, Ellák Somfai, András Lőrincz: Common Fate Based Episodic Segmentation by Combining Supervoxels with Deep Neural Networks. 2019 International Joint Conference on Neural Networks (IJCNN), 2019. DOI: 10.1109/IJCNN.2019.8851697

Common Fate Based Episodic Segmentation by Combining Supervoxels with Deep Neural Networks

László Kopácsi, Áron Fóthi, Ádám Fodor, Ellák Somfai†, and András Lőrincz
Department of Software Technology and Methodology
Eötvös Loránd University
Budapest, Hungary
† also with the *Institute for Solid State Physics and Optics*
Wigner Researh Centre for Physics of the Hungarian Academy of Sciences
Budapest, Hungary

***Abstract*—We estimated the contribution of different factors in segmentation tasks by means of deep neural networks. Results indicated that texture and optical flow have similar power, but they seem not to add up. In turn, we decided to study the '*Common Fate Principle*' of the 100 years gestaltism suggesting that elements that move together belong together. We developed a simple, fast, and efficient episodic segmentation method that – to some extent – resembles the 'how system' of the visual processing: we dropped every piece of information except motion, and started from pure optical flow estimations on 2D videos. For the sake of segmentation, we used a parallel and fast hierarchical supervoxel algorithm. We studied (i) grid topology in space and time, (ii) 2D grid in space and topology dictated by the optical flow in time, and (iii) added deep network based depth estimation from 2D images. We measure performances on episodic foreground-background segmentation task of the Davis benchmark videos. Results are competitive to state-of-the-art segmentation techniques.**

***Index Terms*—Deep networks, Gestalt principles, minimal spanning tree, supervoxel, optical flow**

I. Introduction

Flexible shape representation in space and time, in other words spatio-temporal episodic segmentation is critical for visual perception, prediction and interaction. We started from Gestalt principles or gestaltism. For reviews on this subject, the interested reader is referred to the literature; see, e.g., [1], [2] and the online Journal on Gestalt Theory[1]. Gestalt means pattern and Gestalt Theory is concerned with the hierarchy of patterns in space, time, and different sensory modalities. Here, we shall focus on the Common Fate Principle that deals with motion patterns.

Before going further, we note that there is a large number of segmentation algorithms and there are high quality solutions for solving the problems. It is intriguing, however, that the simplest episodic segmentation method – to our best knowledge – hasn't been tried before. We introduce it here and will detail related works in the next section.

The research has been primarily supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002). Áron F. and E.S. were supported by part grant EFOP-3.6.2-16-2017-00013 and by the ELTE Institutional Excellence Program (1783-3/2018/FEKUTSRAT) supported by the Hungarian Ministry of Human Capacities, respectively.

[1]http://www.gestalttheory.net/cms/

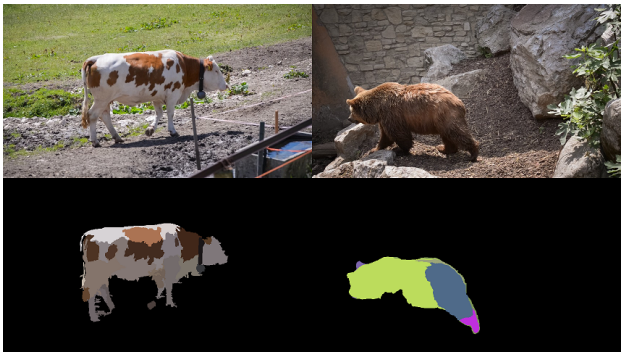

Fig. 1: Segmentation methods using Gestalt Principles. Top row: original images, bottom left: averaged RGB superpixel segments, bottom right: one frame from a randomly colored 2D optical flow supervoxel set. Note that Gestalt Principle 'similarity' (of texture) may be insufficient for joining parts that belong to the same object. Principle 'common fate' (captured by optical flow) cannot separate static or slow moving parts of an object from the background, like the legs on which the bear leans on and stays for a while. Eventually, one needs both optical flow and texture information and probably other Gestalt Principles for highly precise object segmentation. For the motivation of studying capabilities of the 'Common Fate Principle' alone, see text (Table I).

The Common Fate Principle proposes that parts that move together belong together. We rub off all unnecessary features to deal with the motion alone for segmentation. In turn, we shall assume the availability of *motion detectors* similarly to the visual motion perception in biological systems. We use optical flow for representing motion. In this study, we shall not use additional pixel-wise color or grayscale information.

For the sake of segmentation, we apply Borůvka's algorithm [3], a highly parallel and hierarchical supervoxel method that can be seen as a special kind of spreading activation that finds the minimal spanning tree of a graph. In our case, the graph could be the grid on frames with vertices (pixels) connected with corresponding pixels of the previous and the next frame. Alternatively, pixels of the frames belonging to different time

instants can be connected by means of the optical flow itself. We study both cases. We also add monocular depth estimation via a trained deep neural network [4].

Our contributions include (i) the introduction of a novel, but simple method for episodic segmentation motivated by Gestalt principles with competitive performance on a benchmark problem, (ii) changes to the topology of space according to motion, (iii) the application of deep networks as *adaptable sensors* for optical flow and depth estimation, and (iv) the studying of their effects. We also point to additional information pieces that should improve the performance of the method.[2]

The paper is organized as follows. Section II is about the related works. Section III details the methods that we apply. Section IV, V, and VI describe our results, discuss those results, and conclude in order.

II. Related Works

Segmentation of similar parts of the image is based on neighbor relations and image features. The first one determines a graph and the second one makes it a weighted one to be segmented. Graph-based image segmentation methods have been suggested as early as 2000 [5]–[7]. However, the methods suggested at that time are either restricted to 2D or are relatively slow, or both.

Superpixel methods have been concerned with the precision of the boundaries and the regularity of the tesselation. Researchers have suggested different regularization techniques, such as mean-shift [8], energy optimization [9], and boundary constraints [10]. An excellent overview as well as numerical comparisons of different methods using variations of the watershed algorithm, density based methods, methods using graphs and/or contour evaluations, path and clustering methods among others, can be found in [11].

Image processing requires considerable computational power and the speed of superpixel computations comes to question. Levinshtein et al. [12] suggested geometric flows for fast evaluations. A simple linear iterative algorithm (SLIC) [13] was put forth in 2012 and its GPU version [14] has become very fast.

Another direction of graph based algorithms utilizes minimum spanning trees computed bottom-up. The approach is attractive, since the bottom-up approach makes it highly parallelizable. The method is based on Borůvka's Algorithm [3] that we elaborate in the next section. Theoretical advances and inventions on parallel versions of the algorithm are flourishing, see, e.g., [15]–[17] and the cited references therein. This algorithm has been suggested for superpixel segmentation by Wei et al., very recently [18]. The method is of high speed. We extended it for supervoxel segmentation tasks according to the topology of pixels in subsequent frames and according to the topology of the optical flow of neighboring frame pairs, see later.

Supervoxels have been used for segmentation in 3D measurements, such as CT and MRI. For recent works on the subject, see, e.g., [19], [20]. In addition, video processing can also utilize supervoxel techniques [11], [21], [22]. Methods extend 2D techniques and exploit temporal features, including optical flow. A recent work [23] places optical flow into the focus of investigation and exploits motion saliency for improving spatio-temporal propagation. However, in all cases, optical flow is added to other features and – to our best knowledge – the 'naked' optical flow, i.e., the clean version of the Common Fate Principle hasn't been studied yet.

III. Methods

A. Optical flow

We used the novel deep learning PWC-Net method [24][3] for the estimation of the optical flow. This method exploits a pyramidal computation that regularizes the optical flow by diminishing the aperture problem through the pyramid based spatial context. It also uses a coarse-to-fine warping layer [25] for the CNN features to improve optical flow estimation: warped features contribute to flow estimation via forming the so called partial cost volumes, a concept borrowed from stereo matching; see, e.g., [26], [27] and the references therein.

B. Borůvka's Algorithm

A graph-based 2D superpixel segmentation algorithm has been suggested by Wei et al. in 2018 [18][4]. The method is based on building a hierarchy of merged regions, from which a segmentation into an arbitrary number of superpixels can be queried quickly. The order of the merges is determined by Borůvka's minimum spanning tree algorithm, where the most similar regions are merged first.

For a 2D image the vertices of the graph are the pixels, and the edges connect neighboring pixels in the pixel grid. Each vertex (pixel p) carries a feature vector f^p, for example the color components of the pixel, or the optical flow components belonging to that pixel, and alike. The weight of an edge is the ℓ_1 norm of the feature vector difference between the pixels and can be weighted. In the first iteration of Borůvka's algorithm, the smallest outgoing edge is considered from each vertex, which connect the vertices into small (but not necessarily size 2) trees. In the subsequent iterations the smallest outgoing edge from each tree is considered, which merge trees. After each merge the feature vector of the two trees (initially the pixels) is averaged, weighted by the size of the trees. The process concludes in $O(\log(V))$ iterations, where V is the number of vertices. The sequence of the merging edges is recorded (sorted by increasing weight within an iteration), enabling to obtain S superpixels (trees) by using the first $V-S$ edges.

In a sequence of video frames the pixels of the frames form a 3D structure and elements of the structure become voxels. A contiguous group of similar voxels, which has both spatial and temporal extent, is a supervoxel. We used two methods for computing supervoxels. In the first case, the topology of the

[2]Our implementation is available at https://github.com/lkopi/common-fate-segmentation.

[3]https://github.com/NVlabs/PWC-Net/tree/master/PyTorch

[4]http://faculty.ucmerced.edu/mhyang/pubs.html

graph was defined by neighboring pixels within each frame and the 'neighboring' pixels in time was also connected if they had the same spatial position. In the other case, the edges connecting pixels within a frame are the same as above, but the edges connecting pixels in neighboring frames follow the optical flow. The weights of the connections were determined by considering the absolute values of the differences of the optical flow components. In some experiments we also added the differences of depth estimations to the weights. The two components may have different multipliers. An edge connecting voxel v to any of its neighbors v' is denoted by $w(v, v')$ and assumes the following value

$$w(v, v') = \lambda_{\text{interframe}}^{v,v'} \left(\|f_{\text{OF}}^{v} - f_{\text{OF}}^{v'}\|_1 + \lambda_{\text{d}} \|f_{\text{d}}^{v} - f_{\text{d}}^{v'}\|_1 \right) .$$

where $\lambda_{\text{interframe}}^{v,v'}$ takes the value $\lambda_{\text{interframe}}$ when v and v' are in different frames, and 1 when they are in the same frame; $\|\cdot\|_1$ denotes the ℓ_1 norm. Subscripts OF and d denote the optical flow and depth component of the feature vector, respectively. One may optimize both prefactors: $\lambda_{\text{interframe}}$ and λ_{d}. Once the graph edges are determined, the same Borůvka-based algorithm is applied to the entire video to yield supervoxels.

C. Benchmark measures

We used the Davis video benchmark set of 2016, which has ground truth foreground segmentation for each frame [28]. There are different metrics to characterize and compare the quality of the superpixel and supervoxel algorithms, see, e.g., [21] and the cited references therein. We used two benchmark measures for quantifying the quality of our methods, we detail them below.

1) Region Similarity ($\mathcal{J}$): Region similarity is defined as the Jaccard index $\mathcal{J}$, also known as the Intersection over Union (IoU). IoU measures the similarity between the estimated $\mathcal{M}$ and the ground truth $\mathcal{G}$ foreground segmentation. As the name implies it is calculated the following way: $\mathcal{J} = \frac{|\mathcal{M} \cap \mathcal{G}|}{|\mathcal{M} \cup \mathcal{G}|}$.

2) Contour Accuracy ($\mathcal{F}$): Contour Accuracy is the so called F-measure. It is defined as $\mathcal{F} = \frac{2P_c R_c}{P_c + R_c}$, where P_c and R_c measures the contour-based precision and recall, respectively between the contour points of the proposed mask $c(\mathcal{M})$ and the ground truth mask $c(\mathcal{G})$. Pixel tolerance is set to the benchmark measure [28].

IV. Results

A. Evaluation of features

Motivation to study 'naked' optical flow systems from our exploratory work: we wanted to have crude estimates about the potential contributions of the different features for segmentation by training the ResNet50 deep CNN [29][5]. We used individual features and different combinations of features as inputs to the CNN (see Table I) and the ground truth segmentations as target outputs. The training parameters were as follows: learning rate=0.0001, decay=0.995, and the number of epochs was 300. Optical flow values were scaled between 0 and 255 by means of the sigmoid function $S(v) = 255 \cdot (1 + \exp(-0.02v))^{-1}$, where v is the x or y component of the optical flow (pixel difference between frames). Results are shown in Table I. The parameters above, as well as in the rest of this paper, are chosen as educated guesses and not as a result of hard optimization; therefore the same values are expected to work well on different video series.

CNN experiments			
	$\mathcal{J}$ mean	$\mathcal{F}$ mean	$\mathcal{J}\&\mathcal{F}$ mean
Depth (D)	55.6	57.5	56.5
RGB	53.8	54.4	54.1
Optical Flow (OF)	52.6	46.9	49.7
OF_{SV}	54.2	51.8	53.0
RGB+D	54.1	57.0	55.6
$\text{RGB+OF}_{\text{SV}}$	44.5	53.7	49.1
RGB+OF+D	52.6	58.4	55.5

TABLE I: Results of the supervised segmentation CNN experiments. Input images represented either depth, RGB, optical flow (OF), or their combinations. OF_{SV} denotes optical flow averaged over supervoxels: 16 supervoxels were calculated within a 5 frame sequence around each frame, with optical flow used both as features and connection topology in Borůvka's algorithm; the optical flow was averaged on the intersection of a supervoxel and the given frame.

According to Table I, OF alone produces good results, but combinations with OF sometimes give rise to lower performances. By contrast, similar degradation is not seen when depth is combined with the other features, although monocular depth estimations [4] may be imprecise.

B. Temporal supervoxel segmentation

To evaluate the performance of our method, we cut the Davis videos to shorter segments. The results depend on the length of the video segment (number of frames) and the number of supervoxels. We computed region similarities to the ground truth as follows. We selected those supervoxels as foreground supervoxels, which have over 50% overlap with the ground truth foreground for the whole video segment. The region similarities, contour accuracy and temporal stability were evaluated by comparing the union of foreground supervoxels to the ground truth.

First we show how the results depend on supervoxel number and video sequence length (see Fig. 2), where optical flow was used as feature vector and regular grid as connection topology for the supervoxel computation.

We calculated the average (i) for all videos and all frame series (left column of the figure) and (ii) for the 10 videos having lowest and highest background speeds (middle and right columns, respectively). Performance typically increases for larger number of supervoxels and shorter video segment, except for the fastest videos where for large supervoxel numbers decreasing the number of frames did not make a difference.

[5]https://github.com/GeorgeSeif/Semantic-Segmentation-Suite

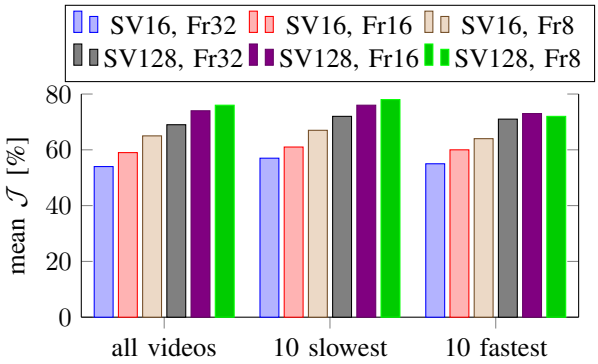

Fig. 2: Region similarities ($\mathcal{J}$ values) as a function of supervoxel numbers (SV#) and number of frames (Fr#). Here we used optical flow as feature vector and regular grid as connection topology in the supervoxel segmentation. Dependencies are similar, except for fast videos with lower frame numbers.

Next we fixed the number of supervoxels to 128, and investigated how the choice of the feature vector affects performance. Table II shows the averaged benchmark measures for sequences in all, the 10 slowest and the 10 fastest videos, and for different inputs and network topologies. The decay values represent temporal changes: the difference between the average over the first and last quartile of the sequence. Since the foreground voxels are selected using the entire sequence and not the first frame, negative decay values are not uncommon. In this setup the ideal decay values are the ones close to zero. We attribute the large negative region similarity $\mathcal{J}$ decay values for fast videos to the fact that a number of Davis videos depict approaching foreground, which grow in size, therefore the spacetime ground truth is dominated by the end of the video sequence.

Table II shows this simple method in itself is competitive, but it is not able to reach state-of-the art. According to the results, temporal topology determined by optical flow is more beneficial for fast videos than for slow ones. This is probably due to the error in the optical flow estimation. We note that the presence of depth in the feature vector made a number of benchmark results worse in the case of the slow videos.

The high values achieved by the supervoxel method highlights the importance of the information contained in optical flow to segmentation. Figures 1 and 3 demonstrate the relevance of the *Common Fate* approach and the need for information fusion. We have chosen the foreground supervoxels as those having over 50% overlap with the ground truth on the first frame of the sequences. Texture based segmentation separates parts that belong together (Fig. 1, left column) If part of the foreground moves similarly to the background, then the optical flow is unable to distinguish between them. The whole background might be integrated with one part of the foreground (Fig. 3).

Figure 4 is about another issue concerning the Common Fate Principle in the absence of textural information. If the object to be tracked is subject to occlusion and/or if the optical flow changes quickly, then the object may be present in one frame, but may be (partially) lost in another.



Fig. 3: 1st and 8th frame when using the superpixel combination which have over 50% overlap with the ground truth on the first frame. The 2D segments of each supervoxel were colored randomly. For details, see text.

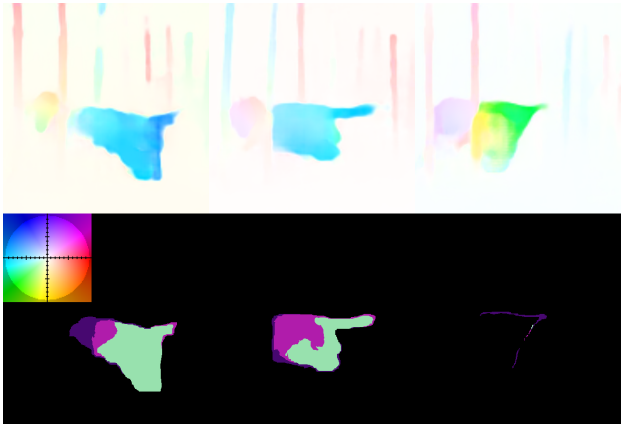

Fig. 4: Loss of mask: the initial mask was lost due to occlusion and sudden change of movement. Upper row: optical flow colored by direction and intensities. The color coding can be found at the upper left corner of the row below. Lower row: randomly colored 2D segments of the supervoxels.

Figure 5 depicts five frames from a high quality 16 frame series. The middle row shows a case with occlusion. The other rows depict occlusion free series.

V. Discussion

In a preliminary study (Table I), we found that segmentation based on OF alone gives rise to good results. However, deep learning based combinations with other features do not improve performances. In turn, 'naked' OF segmentation performance is to be investigated and optimized first, in order to introduce principled information fusion like in [30] afterwards.

This first step is similar to the motion sensor based 'how' processing channel in the brain. For recent results on this matter, see the full issue in the journal Cortex introduced Haan et al. [31]. To our best knowledge, such 'naked' OF based segmentation has not been studied before, although it seems relevant from the point of view of episodic representation is known to be necessary for declarative memory [32].

We investigated the potentials of optical flow for segmenting episodes. This approach that segments spatio-temporal regions based on their similarities and differences concerning motion

Videos	**all**				**10 slowest**				**10 fastest**			
Features	OF	OF, d	OF	OF, d	OF	OF, d	OF	OF, d	OF	OF, d	OF	OF, d
Topology	Grid	Grid	OF	OF	Grid	Grid	OF	OF	Grid	Grid	OF	OF
$\mathcal{J}\&\mathcal{F}$ mean ↑	75.8	75.3	**76.4**	75.0	77.0	76.1	**77.1**	74.7	68.7	69.9	70.7	**71.2**
$\mathcal{J}$ mean ↑	75.9	75.9	**76.4**	75.3	78.1	**78.2**	77.8	76.3	71.0	72.3	73.1	**73.7**
$\mathcal{J}$ recall ↑	93.0	92.8	**93.8**	90.5	**91.3**	88.8	**91.3**	83.8	89.4	90.8	90.0	**92.7**
$\mathcal{J}$ decay	-1.9	**-0.5**	-1.5	-1.1	-2.1	1.8	-2.2	**0.3**	-16.5	-13.9	-10.7	**-9.9**
$\mathcal{F}$ mean ↑	75.8	74.7	**76.5**	74.7	76.0	74.0	**76.4**	73.1	66.5	67.4	68.2	**68.7**
$\mathcal{F}$ recall ↑	88.8	87.0	**90.8**	87.5	**87.5**	81.3	86.3	78.8	81.8	82.7	**84.7**	84.2
$\mathcal{F}$ decay	-1.3	**-0.7**	-1.3	-1.5	-3.9	0.8	-3.1	**-0.7**	-1.2	**0.6**	4.4	3.7

TABLE II: Benchmark results for different inputs and temporal topologies. *'OF'*: optical flow, *'d'*: depth. *'Decay'* represents difference between the average over the first and last quartiles of the video sequence. Prefactors of Borůvka's method were optimized: $\lambda_d = 0.1$, and $\lambda_{\text{interframe}}$ were set to 1.6 for all cases except the last column where their values were equal to 1. Depth values were normalized between 0 and 255. Number of supervoxels: 128. Number of frames: 8 for *'all'* and for *'slow'* sequences, 32 for *'fast'* sequences.

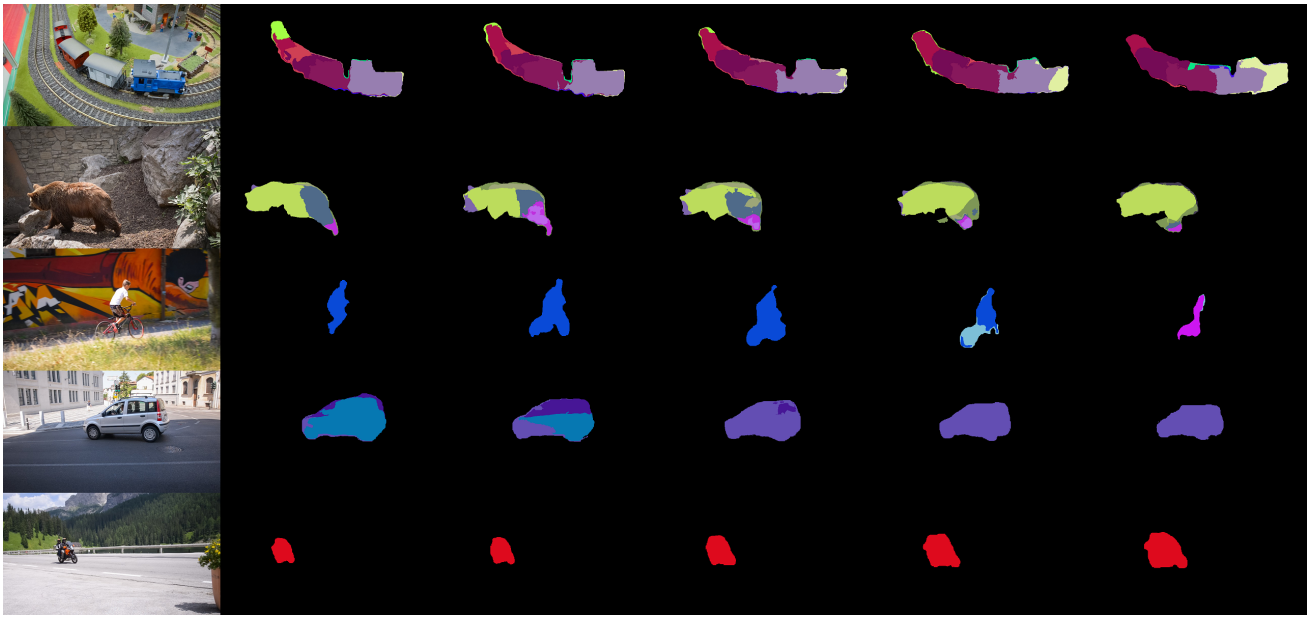

Fig. 5: High quality examples for Common Fate Principle based segmentation. Middle row: case where occlusion spoils the result. Columns in order from left to right: RGB image, supervoxel masks on the 1st, 4th, 8th, 12th and 16th frames.

coherence is the algorithmic manifestation of the *Common Fate Principle* of Gestalt Theory.

We used a simple and efficient graph based segmentation method (without customized training) that allowed us to change the topology from 3D geometrical discretization to 2D spatial and optical flow based temporal topology. We also added deep network based depth estimation on 2D images [4].

Results show that optical flow based information although has high potentials for segmentation when used alone, it may spoil the segmentation quality of RGB based methods. Reasons are uncertain and need further investigations. However, they might be due both to the noise in optical flow estimations and the occlusion based uncertainties. According to our studies, noise contribution of OF can be more destructive for performance in case of low speeds. Occlusion may spoil the results under all circumstances and especially for high speed motions (Table II). We found that adding information about the depth improves performance in all cases.

We note that in our work, we dealt with the coherences of spatio-temporal volumes instead of temporal propagation. In turn, we neglected causal relationships that have been used in other works, like the video propagation network [33]. Improvements are expected when adding causality related constraints and other pattern completion methods from the artillery of the segmentation literature [34], including novel supervised deep learning tools [35]. In addition, probabilistic inferences about the Common Fate of the parts cut by occlusion like in Fig. 4 could be improved by means of textural information. For a recent approach, see, e.g., [36] and the references therein.

According to Table II depth information can improve the results to some extent. Depth information, however, may be considered not as a feature, but as an additional dimension, increasing the number of dimensions to 4. It is left for future work to see if such 4 dimensional formulation improves

performance further or not.

As a further note, motions due to camera motion are relatively easy to separate, since the corresponding optical flow, apart from camera distortions, are uniform. In Fig. 3 the background color changes (from black to light green). This is due to the Borůvka algorithm: voxels compete for volume regions without causality constraints. Joining such large volumes calls for continuity estimations, being somewhat similar to techniques carefully worked out in 2 dimensions [37].

Our last note is that we use deep networks as adaptive sensors offering several advantages, such as the joint learning [38] and thus adaptation as well as sample collection for cross-network supervision by means of consistence seeking [39], [40]. The study of such additional training for performance improvements are left for future works.

VI. Conclusions

We have shown that optical flow based segmentation has strength by its own. We also found that optical flow based information may spoil texture based segmentation in both supervised and unsupervised settings, possibly due to the noise content of optical flow estimation as well as to occlusions.

Our results suggest that episodic segmentation can take advantage of the Common Fate Principle by using motion information alone. Here we studied properties related strictly to pattern coherence and left out information concerning causality, occlusions and disocclusions. The pure motion based information should gain considerable strength upon exploiting textural information. However, our supervised CNN based exploratory pre-studies as well as our results indicate that the brute force combination of temporal and textural features may not be optimal.

Acknowledgments

Áron Fóthi and László Kopácsi had equal contributions. We thank all members of the NIPG lab for valuable discussions.

Author Contributions

L.K. and A.L. conceived and designed the research, L.K., Ádám F., Áron F., and E.S. performed computational analyses.

References

[1] F. Sundqvist, "The Gestalt phenomena and archetypical rationalism," *Gestalt Theory*, vol. 29, p. 40, 2007.

[2] D. Todorovic, "Gestalt principles," *Scholarpedia*, vol. 3, p. 5345, 2008. revision #91314.

[3] J. Nešetřil, E. Milková, and H. Nešetřilová, "Otakar Borůvka on minimum spanning tree problem translation of both the 1926 papers, comments, history," *Discrete Mathematics*, vol. 233, no. 1-3, pp. 3–36, 2001.

[4] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2041–2050, 2018.

[5] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[6] X. Ren and J. Malik, "Learning a classification model for segmentation," in *International Conference on Computer Vision, 2012 IEEE Conference on*, vol. 1, pp. 10–17, IEEE, 2003.

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

[8] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[9] O. Veksler, Y. Boykov, and P. Mehrani, "Superpixels and supervoxels in an energy optimization framework," in *European Conference on Computer Vision*, pp. 211–224, Springer, 2010.

[10] Y. Zhang, X. Li, X. Gao, and C. Zhang, "A simple algorithm of superpixel segmentation with boundary constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 7, pp. 1502–1514, 2017.

[11] D. Stutz, A. Hermans, and B. Leibe, "Superpixels: An evaluation of the state-of-the-art," *Computer Vision and Image Understanding*, vol. 166, pp. 1–27, 2018.

[12] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2290–2297, 2009.

[13] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, *et al.*, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[14] C. Y. Ren, V. A. Prisacariu, and I. D. Reid, "gslicr: Slic superpixels at over 250hz," *arXiv preprint arXiv:1509.04232*, 2015.

[15] L. Dhulipala, G. E. Blelloch, and J. Shun, "Theoretically efficient parallel graph algorithms can be fast and scalable," *arXiv preprint arXiv:1805.05208*, 2018.

[16] M. Ghaffari and F. Kuhn, "Distributed MST and broadcast with fewer messages, and faster gossiping," in *32nd International Symposium on Distributed Computing (DISC 2018)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[17] R. Panja and S. Vadhiyar, "MND-MST: A multi-node multi-device parallel boruvka's mst algorithm," in *Proceedings of the 47th International Conference on Parallel Processing*, p. 20, ACM, 2018.

[18] X. Wei, Q. Yang, Y. Gong, N. Ahuja, and M.-H. Yang, "Superpixel hierarchy," *IEEE Transactions on Image Processing*, vol. 27, pp. 4838 – 4849, 2018.

[19] Y. Kong, Y. Deng, and Q. Dai, "Discriminative clustering and feature selection for brain MRI segmentation," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 573–577, 2015.

[20] L. Carvalho, A. Sobieranski, and A. von Wangenheim, "3d segmentation algorithms for computerized tomographic imaging: A systematic literature review," *Journal of Digital Imaging*, pp. 1–52, 2018.

[21] C. Xu and J. J. Corso, "Evaluation of super-voxel methods for early video processing," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1202–1209, IEEE, 2012.

[22] M. H. Saffar, M. Fayyaz, M. Sabokrou, and M. Fathy, "Semantic video segmentation: A review on recent approaches," *arXiv preprint arXiv:1806.06172*, 2018.

[23] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 786–802, 2018.

[24] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018.

[25] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conference on Computer Vision*, pp. 25–36, Springer, 2004.

[26] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2758–2766, 2015.

[27] J. Xu, R. Ranftl, and V. Koltun, "Accurate optical flow via direct cost volume processing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1289–1297, 2017.

[28] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 724–732, 2016.

[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, 2017.

[30] M. Patzold, R. H. Evangelio, and T. Sikora, "Counting people in crowded environments by fusion of shape and motion information," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pp. 157–164, IEEE, 2010.

[31] E. H. de Haan, S. R. Jackson, and T. Schenk, "Where are we now with 'What' and 'How'?," *Cortex*, vol. 98, pp. 1–7, 2018.

[32] E. Tulving, "Episodic memory: from mind to brain," *Annual review of psychology*, vol. 53, no. 1, pp. 1–25, 2002.

[33] V. Jampani, R. Gadde, and P. V. Gehler, "Video Propagation Networks," in *Computer Vision and Pattern Recognition*, pp. 3154–3164, 2017.

[34] D. Cremers, M. Rousson, and R. Deriche, "A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 195–215, 2007.

[35] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[36] Y.-C. Hsu, Z. Xu, Z. Kira, and J. Huang, "Learning to cluster for proposal-free instance segmentation," *arXiv preprint arXiv:1803.06459*, 2018.

[37] O. Ben-Shahar and S. Zucker, "General geometric good continuation: from taylor to laplace via level sets," *International journal of computer vision*, vol. 86, no. 1, p. 48, 2010.

[38] Y. Zou, Z. Luo, and J.-B. Huang, "DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency," in *European Conference on Computer Vision*, pp. 38–55, Springer, 2018.

[39] A. Lőrincz, M. Csákvári, Á. Fóthi, Z. Á. Milacski, A. Sárkány, and Z. Tősér, "Towards reasoning based representations: Deep Consistence Seeking Machine," *Cognitive Systems Research*, vol. 47, pp. 92–108, 2018.

[40] Z. Milacski, K. Faragó, A. Fóthi, V. Varga, and A. Lorincz, "Declarative description: The meeting point of artificial intelligence, deep neural networks, and human intelligence," in *XAI 2018, Proceedings of the 2nd Workshop on Explainable Artificial Intelligence*, pp. 97–103, 2018.