© 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: Ádám Fodor, Áron Fóthi, László Kopácsi, Ellák Somfai, András Lőrincz: Skeletonization Combined with Deep Neural Networks for Superpixel Temporal Propagation. 2019 International Joint Conference on Neural Networks (IJCNN), 2019. DOI: 10.1109/IJCNN.2019.8852391

Skeletonization Combined with Deep Neural Networks for Superpixel Temporal Propagation

Ádám Fodor, Áron Fóthi, László Kopácsi, 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*—Medial axis representation (a.k.a. shape skeleton) seems to be present in visual processing, but its relevance has remained unclear. Here, we show the potentials of the medial axis transformation in the temporal propagation of superpixels. We combine (i) state-of-the-art *deep neural network 'sensors'* for optical flow and for depth estimation and (ii) a superpixel algorithm with (iii) the medial axis transformation to obtain frame-to-frame propagation of visual objects. We study the precision of this deep learning facilitated superpixel temporal propagation. We discuss the advantages of the method compared to the temporal propagation of the superpixels themselves.**

***Index Terms*—Deep networks, medial axis, superpixel, optical flow, depth estimation, temporal propagation**

I. Introduction

Flexible shape representation in space and time is critical for visual perception. Blum [1] has suggested that medial axis transformation (a.k.a. medial points or shape skeletons) gives rise to a flexible and robust representation for bending and stretching. Since then, evidences have been found that visual processing indeed takes advantage of the medial axis.

Medial axis representation has been found in the brain in early visual processing areas as well as higher in the visual stream. In one experiment it was found that Gabor filters at the medial axis are more salient than at other points [2]. Lee et al. [3] found that medial axes are represented as early as the primary visual cortex and that the role of the transformation is figure-ground segregation. Lescoart and Biederman [4] suggests that different visual areas have access to the medial axis representation. Another study has shown that points of medial axis are also used when pointing onto a figure [5]. Intriguingly, skeletons are also influenced by the assumed history of the objects [6].

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

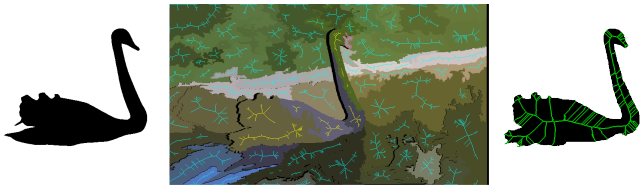


Fig. 1: **Multi-skeleton and single-skeleton methods.** Shown: 5th frame of the Blackswan video sample. (a) Ground truth mask. (b) Multi-skeleton method: thresholded medial axis transform is calculated for every superpixel of the 5th frame followed by the temporal propagation of every superpixel and skeleton. (c) Single-skeleton method: skeleton of the mask is computed and translated by means of the optical flow in order to label superpixels of the next frame.

These findings and the robustness of medial axis transformation motivate our choice that it can serve the temporal propagation of superpixel representation of the visual space. Superpixels are visual regions made of similar smaller units, such as smaller superpixels, or the pixels themselves. They form an efficient compression of the shape and can serve diverse interpretations, including figure-ground segregation, a.k.a. foreground-background segmentation.

Superpixel segmentation based on graph cuts has been in the forefront of image processing algorithms in the last 20 years [7]. Wei et al. [8] have demonstrated recently the high quality performance of Borůvka's minimal spanning tree (MST) algorithm in hierarchical superpixel computations. They already used deep neural networks for holistic edge detection [9]. This MST algorithm is advantageous, since it is highly parallelizable and thus can be very fast [10].

We extend their work to the temporal domain by means of (i) deep neural network based optical flow and depth estimation, and (ii) add the medial axis transformation for gaining robustness. We propose two temporal extensions: (a) 'multi-skeleton' method (Fig. 2(a)): where superpixels are propagated using their medial axis, and (b) 'single-skeleton' method (Fig. 2(b)): where only the skeleton of the mask is propagated.

Experiments are conducted on the Davis 2016 video database [11]. We found that the combination of the algorithms, i.e., skeletonization, superpixel computation, optical flow, and edge enhancements improve the quality of the propagation, which underlines the relevance of these methods in visual processing. The meticulous MST algorithm of Borůvka with robust skeletonization make a favorable combination for temporal propagation of visual information in time.[1]

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

II. Related Works

Superpixel segmentation of an image, joining contiguous sets of pixels with similar image features, has been suggested as a way to speed up image processing by reducing the number of independent image elements. Most superpixel algorithms fall into two broad families: they are based either on clustering or on graph methods. In graph-based methods, the pixel grid defines the graph, and the similarity of neighboring pixels set the weight of the edges [7], [12], [13].

Superpixel methods are useful if they follow object boundaries precisely, and are sufficiently regular. Various regularization techniques have been suggested, like energy optimization [14], boundary constraints [15], and mean-shift [16]. We refer the Reader to the recent review of Stutz et al. [17] for a detailed comparision of several state-of-the-art algorithms.

As superpixels were invented to speed up image processing, fast algorithms are especially valuable. Many efficient algorithms have been suggested, for example a simple linear iterative algorithm (SLIC) [18], which has a particularly fast GPU based implementation; and also the use of geometric flows can help fast processing [19].

A well established graph-based algorithm has been suggested recently in the context of segmentation, which is based on the minimum spanning tree of the image graph. The method uses Borůvka's minimum spanning tree algorithm [20], which is parallelizable and efficient, see [10], [21], [22]. The adaptation to compute superpixels has been proposed by Wei at al. [8], which we will explain in more detail in the next section. The method provides fast superpixel segmentation of 2D images, which enables us to extend it for video propagation.

III. Methods

Table I shows the integrated algorithmic components that we detail below:

Name of Algorithm	Type	References
Borůvka superpixel	graph cut	[20]
PWC-Net for optical flow estimation	deep network	[23]
MegaDepth	deep network	[24]
Holistically-Nested Edge Detection	deep network	[9]
Medial axis transformation	computer vision	[25]

TABLE I: Algorithmic components.

A. Borůvka's Algorithm

Wei et al. proposed a graph-based 2D superpixel segmentation algorithm in 2018 [8]. The algorithm merges image regions in a hierarchical way, enabling to segment the image into an arbitrary number of superpixels rapidly. The sequence of the region merges is given by Borůvka's minimum spanning tree algorithm.

To segment a 2D image, a graph is defined where the vertices are the pixels, and the edges connect neighboring pixels. A feature vector is assigned to each vertex, containing for example the color components (in some color space) or the depth of the pixel. Each edge has a weight, which is the ℓ_1 norm of the feature vector difference between the vertices. In the first iteration of the algorithm the lightest outgoing edges from the vertices are taken, which connect pixels into small clusters. In the next iterations the lightest outgoing edge from each cluster is taken, which yields a hierarchical structure of merged clusters. The sequence of the merges is recorded, where within an iteration they are sorted by increasing weight. Thus for V vertices we can obtain S superpixels by applying only the first $V - S$ merges.

To select which superpixels provide the best coverage of the image object, we use optical flow and medial axis transformation, detailed in the next sections.

B. Deep network for optical flow estimation

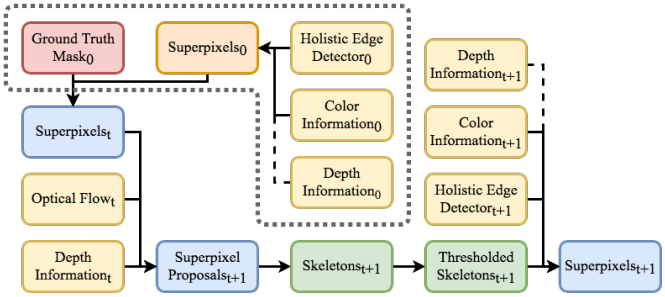
We obtain the optical flow information between subsequent frames by PWC-Net [23]. This supervised deep learning algorithm uses pyramidal processing to diminish the aperture problem and to regularize the optical flow. It applies the coarse-to-fine warping layer method for the CNN features [26], where warped features are used to construct partial cost volumes, which is then processed to obtain the optical flow. In this context cost volume is a concept borrowed from stereo matching [27], [28].

C. Medial axis transformation

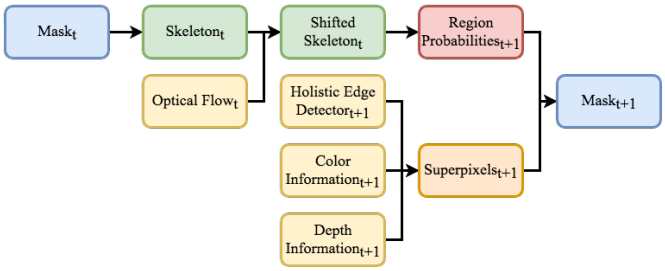
Medial axis transformation [1] is concerned with finite and connected regions in different dimensions, such as objects in 3D or their projections into 2D. A point within an object belongs to the medial axis if there is a circle in 2D, or sphere in 3D, which is centered at that point, has all of its points inside the object and has at least 2 points on the boundary of the object in 2D and 3 points in 3D. For example, the medial axis of a triangle is made of the angle bisectors, whereas the medial axis of a 3D sphere is a single point. The result of the medial axis transformation is sometimes called the skeleton of the object.

Medial axis transformation is sensitive to boundary irregularities: in principle, any convex irregularity may launch a medial axis line. Such sensitivity may be overcome by approximation and regularizing means, see, e.g., [29] leading to a smaller number of branches in the skeleton. In our case,

[1]Our implementation is available at https://github.com/fodorad/superpixel-skeletonization.



(a) 'Multi-skeleton' method. Algorithmic components of the initialization are shown in the area embraced by the dotted line.



(b) 'Single-skeleton' method.

Fig. 2: **Two methods for superpixel propagation.** For algorithmic details of the individual steps, see text.

imprecision of the optical flow around the boundary are of primary concern. We simply thresholded those parts of the medial axis that had small radius circles.

We used the implementation of the medial axis transformation from the scikit-image Python package.

D. Baseline I method for video propagation

As a baseline, we (i) computed the centroid, i.e., the geometric center of each superpixel on a frame, (ii) moved it according to the mean optical flow, and (iii) used the superpixel of this translated point on the next frame as the propagated part of the image. In the single pixel limit of the superpixel computation, this method corresponds to optical flow based pixel propagation.

E. Baseline II method for video propagation

As an improved baseline, we (i) computed the superpixel segmentation on frame t, (ii) translated the superpixels according to their mean optical flow, and (iii) marked those superpixels on frame $t+1$ which have more than 50% overlap with the translated superpixels of frame t. This method corresponds to mean optical flow based superpixel propagation.

F. 'Multi-skeleton' method for video propagation

We aim for video propagation, i.e., using the mask of an image object in one frame, we determine its mask on the next frame of the video. We propose two strategies: either using skeletons of every superpixel (described in this section), or working with a single skeleton of a larger mask (detailed in the next section).

In the first approach the skeleton of each propagated superpixel is fed as seed pixels to the superpixel segmentation of the next frame.

The first step starts from the superpixel segmentation of frame t. The superpixels are translated by their averaged optical flow. Then in the order of decreasing average depth, the labels of the superpixels are placed onto an initially empty image, where earlier (further away) superpixels can be covered by later ones. This results in a pixel grid containing regions of superpixel labels.

In the second step the medial axis transforms of contiguous identically labelled regions are calculated. We selected skeleton points by thresholding them according to their distance values. This way we eliminated those, which are close to the boundary. The threshold is determined by parameter $q \in (0, 1)$, which depends on the largest value $d_{\max}$ of skeleton points. All skeleton pixels with values at least $q \cdot d_{\max}$ are kept.

In the third step, a modified superpixel segmentation of frame $t+1$ takes place by means of the labeled skeleton points of the previous step. When preparing the pixel grid graph for the superpixel segmentation, edges connecting skeleton pixels with the same label are given a negative weight, so they are connected in the initial Borůvka iterations. During Borůvka's algorithm pixels inherit the label of the skeletons they belong to. In this process we ensure that pixel clusters originating from different labels never merge. This results in a superpixel segmentation, where the labels from frame t are carried over to frame $t+1$.

Propagation starts on frame 1 by selecting superpixels having more than 50% overlap with the ground truth mask. The propagated mask of an object on subsequent frames is obtained by taking the union of superpixels with identical labels.

G. 'Single-skeleton' method for video propagation

Our second strategy (Fig. 3) is to morph the mask using optical flow and medial axis transform, and refine with the help of superpixels.

As the first step the skeleton of the mask on frame t is determined by medial axis transform. For each pixel belonging to the skeleton, we also determine its distance from the mask edge, with the help of distance fields.

In the second step each skeleton pixel is translated by the local optical flow between frames t and $t+1$. Then a disk is drawn around each translated skeleton pixel with radius equal to its distance from the mask on frame t before translation (as calculated in the first step). We define the 'draft mask' as the union of the disks.

In the third step the draft mask is refined. A superpixel segmentation of frame $t+1$ is performed, and those superpixels are marked, which have more than 50% overlap with the draft mask. The mask of frame $t+1$ is then the union of the marked superpixels.

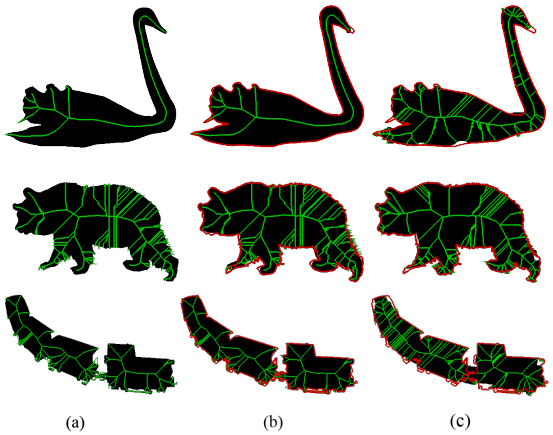


Fig. 3: **Steps of the single-skeleton method.** Blackswan, Bear and Train video samples. (a) Medial Axis Transform on the ground truth mask. (b) Draft mask: the skeleton is propagated pixel-wise by means of optical flow. Disks around each skeleton pixel are drawn. Radii for all disks equal to radii belonging to each medial axis point on the previous frame. Candidate mask on the present frame is the union of the disks. (c) Superpixels of present are marked 'foreground' if more than 50% of them overlap with the draft mask. This step is followed by skeletonization.

To propagate the mask of an object through the entire video, our algorithm takes the mask on the first frame as input, then repeats steps 1-3 between subsequent frame pairs.

H. Benchmark measures

To measure the performance of our method, we use the Davis video benchmark set of 2016, in which the ground truth mask of the foreground object is provided for each frame [11]. The following two quantities are measured:

Intersection over union ($\mathcal{J}$). Suppose for a given frame M and G are the set of pixels of our mask and the ground truth mask respectively, then the Jaccard index $\mathcal{J}$ is defined as the intersection over union $\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$. For $\mathcal{J}$ *mean* the average is taken over all frames, and $\mathcal{J}$ *decay* is defined as the difference between $\mathcal{J}$ averaged over the first and the last quartile of a video. Naturally one aims for high $\mathcal{J}$ mean and low $\mathcal{J}$ decay. Intersection over union is often referred to as *Jaccard index*, and it measures the similarity of M and G, i.e., how well the algorithm avoids mislabeled pixels.

Boundary accuracy ($\mathcal{F}$). In addition to the region based similarity measured by $\mathcal{J}$, we are also interested in how well the boundary of the mask is recovered. One can define the precision P and recall R between the boundary pixels of M and G, allowing for an imprecision of 8 pixels in case of 480p videos. As suggested in [11], we use a combination called F-measure: $\frac{2PR}{P+R}$. Similarly to the other case, $\mathcal{F}$ *mean* is averaged over entire video, and $\mathcal{F}$ *decay* is the difference between the first and last quartile's average.

IV. Results

We evaluated the baselines and both skeleton-based methods on the union of the training and validation subset of the Davis dataset, which contains 45 video samples. See Table II for results. For the fairly simple Baseline I approach, the best performance was achieved for 256 superpixels with considerable drops for both higher and lower superpixel numbers. This best performance was relatively poor; the method didn't pass the 10% IoU mean value. Baseline II (using superpixel overlaps) was better, but the really impressive improvement was achieved by introducing the medial axis transformation (almost 3 times better mean $\mathcal{J}\&\mathcal{F}$ values in case of the multi-skelton method). However when including the depth information as well in the feature vector of Borůvka's algorithm, the results did not improve.

Using only a single skeleton for the whole mask and propagate it with the estimated optical flow, competitive results can be reached, even without long-term feature descriptors, although occlusions and lagging superpixels remain a problem. However, since multiple solutions already exists which can overcome these, we did not address these issues in our paper. Our method can be further improved by introducing additional features, e.g. depth, and by thresholding the disks of the shifted skeleton points. Separately both of them do slightly enhance the overall outcome. The interrelation between these two modifications is unclear, since their combination (on the average) does not give rise to further improvement. On the other hand, results do improve for occlusion-free videos (Table III). By changing the color space from RGB to Lab the result were further improved.

Figure 5 depicts five frames from 5 sequences. The middle row shows a case where the presence of occlusion spoils the result. The other rows illustrate occlusion free series.

V. Discussion

We have proposed and studied superpixel propagation between video frames. Our efforts were motivated by some facts and assumptions listed below.

- [P1] Optical flow can be noisy and may spoil propagation.
- [P2] Borůvka's minimal spanning tree (MST) algorithm applied on pixel grids using color feature similarities is fast and precise [8].
- [P3] Depth estimation is relevant for correcting propagation errors.
- [P4] Edge enhancement greatly improve the results [8].

In our work, we studied the contributions of the different items listed above and neglected some algorithmic components being critical for benchmarks, such as methods for combining superpixels belonging to the same but occluded objects. In turn, we are far from the best results on the Davis dataset. On the other hand, the relevance and the capabilities of the novel skeleton assisted propagation method extended with state-of-the-art deep learning algorithms can be stated.

Naïve baseline methods can be improved by considerable margin with additional algorithmic components, such as depth

Method	Color space	HED	Number of superpixels	λ_d	q	p	$\mathcal{J}\&\mathcal{F}$ mean	$\mathcal{J}$ mean	$\mathcal{J}$ recall	$\mathcal{J}$ decay	$\mathcal{F}$ mean	$\mathcal{F}$ recall	$\mathcal{F}$ decay
Baseline I	RGB	✗	256	-	-	-	0.122	0.099	0.027	0.193	0.145	0.204	0.273
Baseline II	RGB	✗	1024	-	-	0.5	0.223	0.233	0.264	0.283	0.218	0.201	0.287
Baseline II	RGB	✓	1024	-	-	0.5	0.266	0.270	0.285	0.288	0.262	0.243	0.297
Multi-skeleton M1	RGB	✓	64	0	0.75	-	0.351	0.353	0.322	0.340	0.349	0.276	0.299
Multi-skeleton M2	RGB	✓	64	1.0	0.75	-	0.334	0.339	0.303	0.355	0.329	0.268	0.324
Single-skeleton S1	RGB	✓	512	0	0	0	0.494	0.509	0.557	0.341	0.478	0.478	0.323
Single-skeleton S2	RGB	✓	512	0.5	0	0	0.510	0.522	0.565	0.347	0.498	0.504	0.327
Single-skeleton S3	RGB	✓	1024	0	0	1	0.505	0.519	0.570	0.344	0.474	0.457	0.338
Single-skeleton S4	Lab	✓	512	0	0	0	0.534	0.548	**0.614**	0.309	0.519	0.554	0.301
Single-skeleton S5	Lab	✓	512	0.5	0	0	**0.540**	**0.552**	0.611	0.310	**0.527**	0.552	0.304
Single-skeleton S6	Lab	✓	512	0.5	0	1	0.528	0.543	0.604	0.313	0.512	**0.556**	0.295

TABLE II: Benchmark results for different methods: centroid and superpixel translational baselines and two skeleton-based algorithms for video propagation. Multiple hyperparameter configurations were evaluated during the experiments, only the best ones are presented from each group. All *'Number of superpixels'* were tested from 2^4 to 2^{11}. Depth prefactors ('λ_d') of the Borůvka's method were selected from the following options: [0, 0.1, 0.3, 0.5, 0.7, 1]. Parameter *'q'* sets the medial axis transform threshold. *'p'* threshold cuts the small values off from the draft mask. Tests were conducted with $p \in [0, 64]$. $\mathcal{J}$ and $\mathcal{F}$ are the Jaccard index and the Boundary accuracy (details in Section III-H).

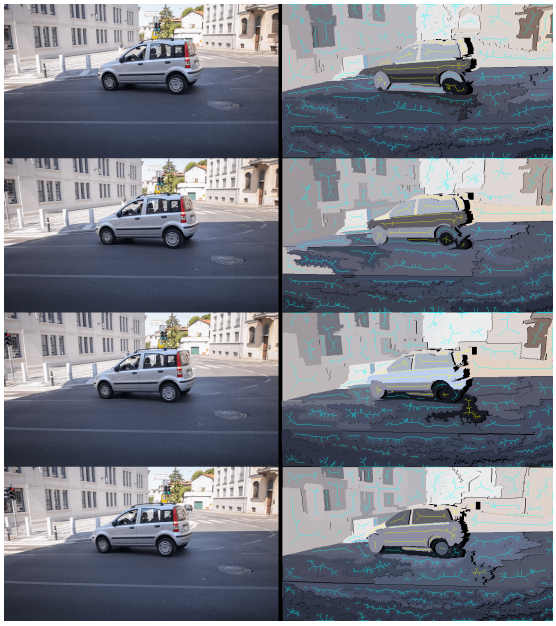


Fig. 4: **Limitation of the multi-skeleton method.** Rows: 3rd to 6th frames of the Car-shadow video. Left (right) column: original frames (corresponding superpixels filled with mean RGB values and skeletons). Blue (yellow) skeletons: corresponding superpixels belong to the background (foreground) according to the propagation algorithm. For the explanation, see text.

estimation, skeleton computation, thresholding and propagation detailed in the previous section.

In the case of the multi-skeleton algorithm, components of [P3]–[P4] improved the results by 8.5%, and 6.8%, respectively. Results in the single-skeleton case were much better, [P3] were 24.4%. The best results for the multi-skeleton case is 35%, whereas for the single skeleton algorithm, it is 54%. Performance is better for Lab than for RGB color descriptors.

Method	$\mathcal{J}\&\mathcal{F}$ mean	$\mathcal{J}$ mean	$\mathcal{J}$ recall	$\mathcal{J}$ decay	$\mathcal{F}$ mean	$\mathcal{F}$ recall	$\mathcal{F}$ decay
Multi-skeleton M1	0.521	0.548	0.606	0.353	0.494	0.498	0.319
Single-skeleton S1	0.674	0.707	0.855	0.225	0.641	0.810	0.251
Single-skeleton S2	0.678	0.712	0.854	0.196	0.643	0.792	0.225
Single-skeleton S3	0.716	0.750	0.914	0.168	0.682	0.874	0.208
Single-skeleton S4	0.711	0.739	0.903	0.182	0.683	0.867	0.207
Single-skeleton S5	0.736	0.756	0.912	0.143	0.716	0.915	0.157
Single-skeleton S6	**0.750**	**0.776**	**0.925**	0.106	**0.723**	**0.925**	0.126

TABLE III: Benchmark results in case of the two methods containing skeletonization. The problem of occlusion is not handled in the proposed method. Therefore, the video samples were filtered, and the results are measured using 15 videos without any or minimal presence of occlusion. The results show, that using proper skeletonization and motion estimation, the proposed method solves the problem of superpixel temporal propagation with reasonable precision.

This large difference between the multi-skeleton and single-skeleton methods is due to the weaknesses of the applied methods. Beyond its imprecision due to the aperture effect as well as other reasons, optical flow is spoiled when disocclusion occurs, that is when object parts are uncovered by moving objects in the foreground. There are methods to overcome such problems, like [30], being outside of the scope of the present study. Depth estimation in monocular videos is not sharp and that decreases precision for larger number of superpixels. Edges of superpixels are not smooth and the skeletons become 'hairy' (Fig. 3). Skeleton regularization methods [29] can be applied to overcome this bottleneck.

The multi-skeleton algorithm has a much lower performance measure in all combinations. This is due to the improper translation of the skeletons. Part of a foreground skeleton is torn off due to optical flow imprecision followed by superpixel computation. Such errors cumulate, propagate, and may increase by time. (Fig. 4). Propagation of these skeletons are not desired. However, the computation of the superpixels

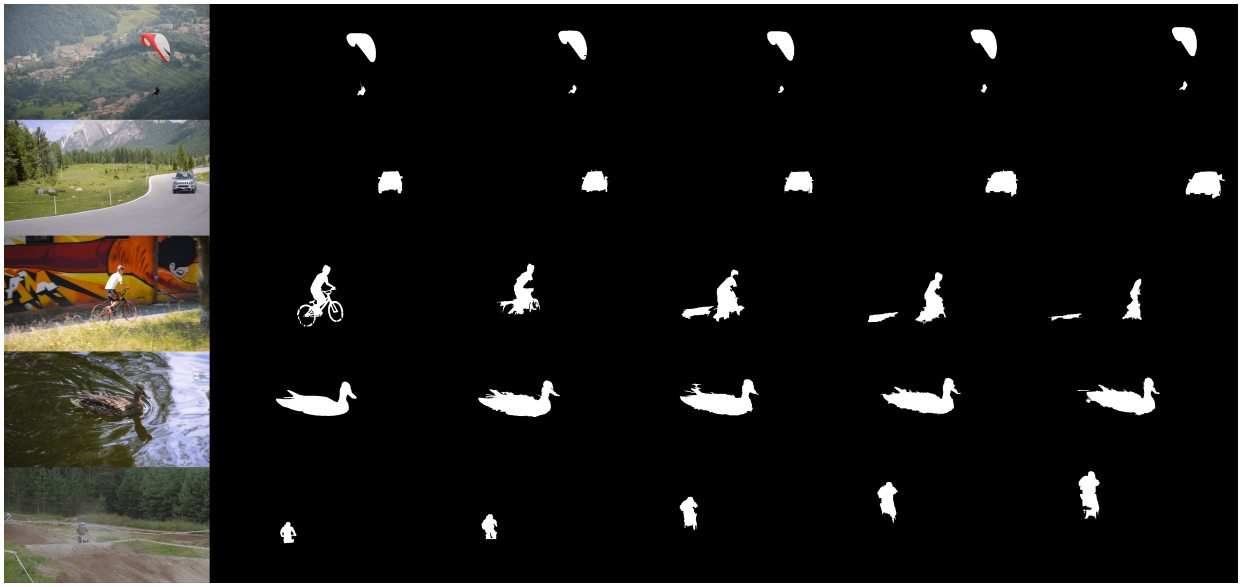


Fig. 5: **High quality examples of the single-skeleton method.** Columns in order from left to right: RGB image, proposed masks on the 1st, 4th, 8th, 12th and 16th frames. Consequences of to-be-treated occlusion are visible in the middle row. Motion coherence, if included, may eliminate torn parts in all rows.

by Borůvka's method is highly precise and one can take advantage of this feature by finding connections between the superpixels of the subsequent frames. We note that the hierarchical nature of Borůvka's method has the promise of having higher precision, i.e., lower size superpixels in case of uncertainties.

In our approach, we are using deep learning tools. This approach has a number of advantages:

[A1] Neural networks can be adapted. They can combine and exploit cost functions for improving the performance. For example, optical flow and depth estimations can jointly improve each other without additional annotation, like in the case of DF-Net [31].

[A2] Deep semantic networks concerning holistic recognition and recognition of the components can be used for probabilistic inferences about the full object, the missing parts of the full object and its behavioral characteristics, including the individual motion patterns [32].

[A3] In the case of semantic knowledge about the components, inconsistencies of the component can be found by rule-based systems, component labels can be fixed and, thus, new learning samples can be collected according to the method suggested in [33].

The problem of occlusion is not included into our study and that spoils our results. We should note, however, that for videos without occlusion, results increase by about 20%. Disocclusion is also a general problem and it should be considered, too. Furthermore, in case of superpixels, one may consider sophisticated texture descriptors that can go way beyond the strength of the Lab space. Such descriptors include histogram of oriented gradients [34], local directional ternary patterns [35], beyond traditional descriptors, such as BRIEF, BRISK, FERNS, ORB, SHIFT, SURF among the many others. For a comparison of some of these methods, see [36] and the cited references therein. The two independent pieces of information, such as motion coherence and texture similarity can be used to overcome occlusions in two complementary ways. (a) Similarity measures, data based Bayesian inference, and Conditional Random Field have been used with great success to overcome occlusions. Semantic information, if available, gives rise to considerable improvement [37]. Sophisticated deep learning methods can be applied in diverse ways, including spatio-temporal segmentation [38]. (b) One can take advantage dynamical equations based on the Newtonian mechanics and estimate occlusion delays between the disappearance and appearance of superpixels when they are occluded by objects by means of optical flow and depth estimations.

Here, our goal was to expand high precision Borůvka's MST method to the temporal domain by means of skeletons. The precision, the parallelizable nature and thus the high speed of the method makes it promising for real life applications in many domains upon the integration of occlusion aware extensions. This is left for future studies.

VI. Conclusions

We proposed a method to propagate visual objects in videos by combining state-of-the-art deep neural network image processing algorithms (calculating optical flow and estimating depth) with non-learned techniques (medial axis transform and superpixel segmentation). Our approach is inspired by biological visual processing both by using skeletons as a form of abstraction, and also by combining motion and depth cues which are processed at an early stage in natural vision systems.

Benchmark results on the Davis 2016 video data set, with both intersection over union and boundary accuracy exceeding 50%, show that our method is competitive with recent video propagation techniques. We expect that with future extensions, including the ability to bridge distances in space and time, which are crucial to handle partial and total occlusions, our method will reach or even exceed the state-of-the-art in video propagation.

Acknowledgment

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

Author Contributions

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

References

[1] H. Blum, "Biological shape and visual science (Part I)," *Journal of Theoretical Biology*, vol. 38, no. 2, pp. 205–287, 1973.

[2] I. Kovacs and B. Julesz, "Perceptual sensitivity maps within globally defined visual shapes," *Nature*, vol. 370, no. 6491, p. 644, 1994.

[3] T. S. Lee, D. Mumford, R. Romero, and V. A. Lamme, "The role of the primary visual cortex in higher level vision," *Vision Research*, vol. 38, no. 15-16, pp. 2429–2454, 1998.

[4] M. D. Lescroart and I. Biederman, "Cortical representation of medial axis structure," *Cerebral Cortex*, vol. 23, no. 3, pp. 629–637, 2012.

[5] C. Firestone and B. J. Scholl, ""Please Tap the Shape, Anywhere You Like" shape skeletons in human vision revealed by an exceedingly simple measure," *Psychological Science*, vol. 25, no. 2, pp. 377–386, 2014.

[6] P. Spröte, F. Schmidt, and R. W. Fleming, "Visual perception of shape altered by inferred causal history," *Scientific Reports*, vol. 6, p. 36245, 2016.

[7] 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.

[8] 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.

[9] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1395–1403, 2015.

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

[11] 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.

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Communications of the ACM*, vol. 27, no. 3, pp. 236–239, 1984.

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

[27] 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.

[28] 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.

[29] P. K. Saha, G. Borgefors, and G. S. di Baja, *Skeletonization: Theory, Methods and Applications*. Academic Press, 2017.

[30] 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.

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

[32] J. Stone, "Face recognition: When a nod is better than a wink," *Current Biology*, vol. 11, no. 16, pp. R663–R664, 2001.

[33] 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.

[34] E. Guo, L. Bai, Y. Zhang, and J. Han, "Vehicle detection based on superpixel and improved hog in aerial images," in *Proceedings of the International Conference on Image and Graphics*, pp. 362–373, Springer, 2017.

[35] A. Chahi, Y. Ruichek, R. Touahni, *et al.*, "Local directional ternary pattern: A new texture descriptor for texture classification," *Computer Vision and Image Understanding*, vol. 169, pp. 14–27, 2018.

[36] S. A. K. Tareen and Z. Saleem, "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK," in *Proceedings of the International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018*, pp. 1–10, IEEE, 2018.

[37] Z. Yang and L. S. Pun-Cheng, "Vehicle detection in intelligent transportation systems and its applications under varying environments: A review," *Image and Vision Computing*, vol. 69, pp. 143–154, 2018.

[38] L. Bao, B. Wu, and W. Liu, "Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5977–5986, 2018.