FAST LABEL PROPAGATION FOR REAL-TIME SUPERPIXELS FOR VIDEO CONTENT

Matthias Reso[⋆], *Jörn Jachalsky*[†], *Bodo Rosenhahn*[⋆], *Jörn Ostermann*[⋆]

[⋆]Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany
[†]Technicolor Research & Innovation, Hannover, Germany

ABSTRACT

Many recent superpixel algorithms for video content rely on dense optical flow vectors to propagate segmentation results from one frame to the next. In this paper, we assess the impact of the optical flow quality on the over-segmentation quality. Our evaluation shows that it is indispensable for videos with large object displacement and camera motion. But due to the high computational costs high-quality, dense optical flow is not suitable for real-time applications. Therefore, we propose a fast propagation scheme that is based on sparse feature tracking and mesh-based image warping. In a thorough evaluation, we compare our proposed scheme to the results of other state-of-the-art propagation methods using established benchmarks. The results show that our method speeds up the propagation process by a factor of 100 while producing a comparable segmentation quality.

***Index Terms*—** Superpixel, supervoxel, optical flow

1. INTRODUCTION

In recent years, many video-based computer vision applications like video segmentation and tracking (e.g. [1, 2, 3]) have relied on a superpixel segmentation of the video frames to improve efficiency as well as the quality of the final results. The idea to reduce the amount of image primitives by grouping spatially-coherent pixels, which share similar low-level features as e.g. color or texture, into small segments of approximately homogeneous size and shape was introduced in [4]. Several approaches were published that transfer the idea of over-segmentation from still images to the video domain. These are supervoxel algorithms (e.g. [5, 6]) and temporally coherent superpixel algorithms (e.g. [7, 8, 9, 10, 11]).

A majority of these approaches utilize information obtained from the optical flow in order to initialize new frames to be processed. This is especially beneficial for sequences with fast and rapid motion resulting in a better segmentation quality. The drawback is the often high computational effort of optical flow algorithms. As supervoxel or temporally coherent superpixel algorithms are usually utilized as a preprocessing step the flow computation can be a non-negligible part of the overall processing costs and thus a potential impediment for the wide acceptance and utilization of these algorithms. Reducing the processing costs of the flow generation without impairing the over-segmentation quality is therefore crucial. Hence, we propose a fast label propagation scheme utilizing sparse feature tracking in combination with Delaunay triangulation and image warping to initialize new frames as it is shown in Figure 1. The key **contributions** of this paper are:

- the evaluation of the impact of the optical flow information on the segmentation results as well as
- a new fast label propagation method based on sparse features and a mesh-based warping scheme suited for real-time processing.

The remainder of this paper is structured as follows. In Section 2, we discuss and evaluate the impact of optical flow information on the final segmentation quality. In Section 3, we describe our fast label propagation scheme and compare it to state-of-the-art propagation methods using the established benchmarks in Section 4. Finally, Section 5 concludes our paper.

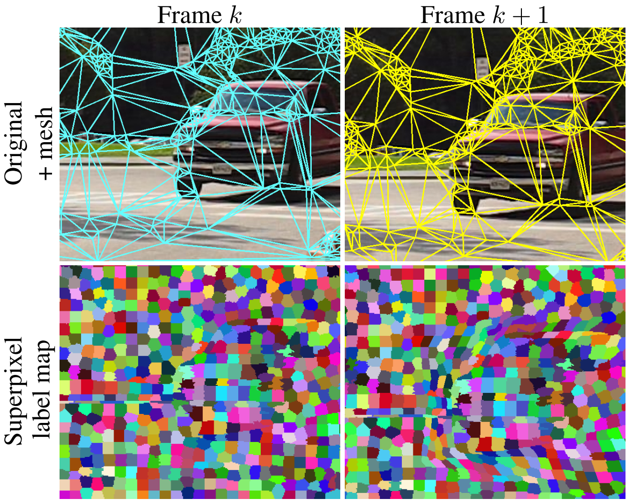

Fig. 1. Sparse features are tracked between frame k and frame $k+1$. A mesh is generated using a Delaunay triangulation. The mesh is laid over the superpixel label map of frame k and the content of each triangle is warped to the frame $k+1$ by performing an affine transformation.

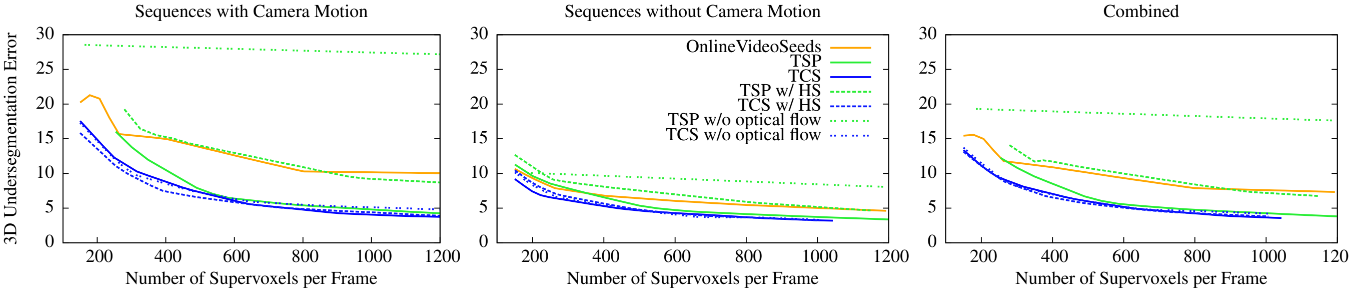

Fig. 2. The 3D undersegmentation error for two classes of sequences (with and without camera motion) and their combination. In case of camera motion the error is lower for approaches using optical flow (TSP, TCS, w/ HS) than for approaches that don't (OnlineVideoSeeds, w/o optical flow). The usage of an optical flow method with more moderate computational costs like Horn & Schunck [12] (w/ HS) doubles the error in some cases when compared to the results with high-quality optical flow of [13].

2. IMPACT OF OPTICAL FLOW

In this section, we evaluate the impact of the optical flow information on the final segmentation quality. Here, we will concentrate on superpixel algorithms for video content. So far, all of the temporally coherent superpixel approaches process a video either on frame-level [7, 8, 9] or on sets of adjacent frames [10, 11]. Thereby, new frames are initialized using intermediate or final segmentation results of the latest frame. Moreover, to capture rapid object movement as well as camera motion the approaches described in [7, 8, 10, 11] use dense optical flow to propagate label information into the new frame. Omitting this optical flow information might lead to superpixels that are not able to follow fast moving objects resulting in a non-consistent or even incorrect temporally coherent superpixel segmentation.

To analyze and evaluate the impact of the optical flow on the superpixel segmentation quality we use benchmark results[1] for the latest superpixel algorithms for video content. Figure 2 depicts the 3D undersegmentation error over the number of supervoxels in order to provide a good overview of the segmentation quality at one glance. Graphs with a benchmark metric plotted over the average number of superpixel per frame as proposed by [8] can be found in Section 4. It should be noted that for the second type of diagrams the average temporal length also needs to be taken into consideration. Otherwise a superpixelation of each frame individually could produce an equivalent undersegmentation error without providing any temporal consistency at all.

For our analysis we used the dataset proposed in [16] and split the eight sequences into two classes. The first class includes sequences with camera motion, whereas the second class comprises those sequences with a fixed camera and only rigid and non-rigid object motion. The corresponding results are plotted in Figure 2 in the left and center diagram while the right diagram shows their combination, i.e. the result for the complete dataset.

The diagrams include results for the recent state-of-the-art temporally coherent superpixel algorithms *OnlineVideoSeeds* [9], TSP [8] and TCS [11]. The latter two use dense optical flow of high quality to propagate the superpixel labels onto new frames to be processed. While [8] filters the optical flow using an approach similar to a bilateral filter in order to produce a smoother flow, [11] uses the pure optical backward flow to obtain the label information for each pixel of the new frame. In order to assess the importance of the optical flow, we also present results for two modified versions of TSP and TCS. One with optical flow calculated using Horn & Schunck's computationally less costly method [12] (denoted *w/ HS*) and one with completely deactivated optical flow (denoted *w/o optical flow*). Thereby, the latter case copies the superpixel segmentation from the latest frame to initialize the new frame. These results are directly comparable to the results of *OnlineVideoSeeds* [9] that does not utilize optical flow information and new frames are initialized by copying the label information from one of the higher block hierarchies of the previous frame.

The left diagram of Figure 2 shows that for the sequences with camera motion the undersegmentation error of the approaches using optical flow is lower than for approaches without optical flow. An exception is the TCS algorithms which seems to be quite robust against the different inputs for the propagation mechanism. For the sequences without camera motion the results of all approaches are close together. Only the modified version of [8] without optical flow usage produces a significantly larger segmentation error. But the difference to the unmodified version is much smaller for the sequences without camera motion than for those with camera motion. We conclude from these results that the usage of the optical flow information can be very beneficial, especially for video sequences with camera motion. This clearly indicates that flow information should be used but somehow the high computational costs for providing it has to be reduced. Therefore, it is essential to replace the calculation of the high-quality, dense optical flow with a rather lightweight propagation scheme without degrading the quality of the final over-segmentation.

[1] Benchmark routines used in this paper are provided by [14] and [15].

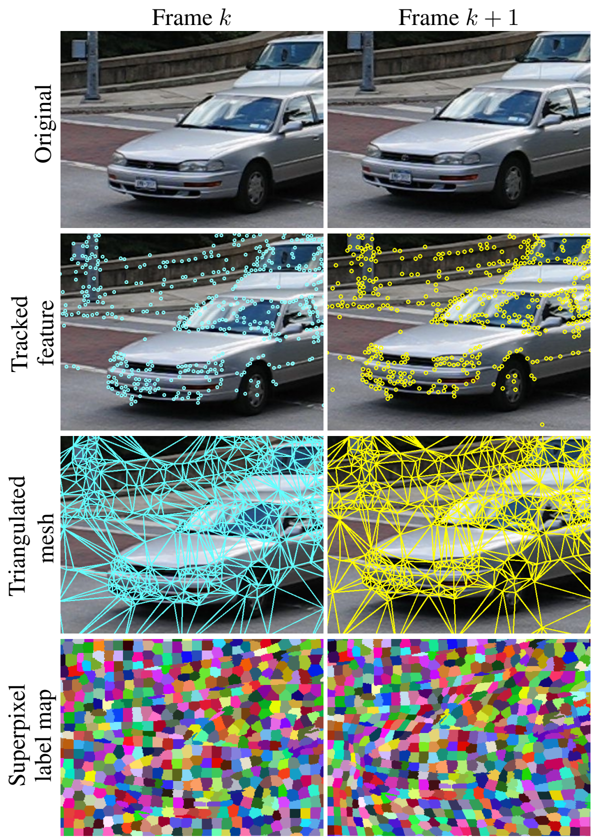

Fig. 3. From top to bottom row: Original frame (cropped). Sparse features found in frame $k+1$ are tracked back into frame k. A mesh is obtained from triangulation of the feature points and deformed by the movement of the tracked features. The superpixel label map of frame k is warped by an affine transformation according to the deformation of the mesh and is used as initialization for frame $k+1$.

3. FAST LABEL PROPAGATION

Our idea for a fast label propagation is inspired by the work presented in [17] and is visualized in Figure 3 for two sample video frames k and $k+1$ (see first row). Instead of calculating a dense optical flow as done in e.g. [7, 8, 10, 11] we only track a set of sparse features between the current frame k and the next frame $k+1$ whose superpixel label map needs to be initialized. The features are calculated for frame $k+1$ using a Harris corner detector [18]. We use the method of [19] to select "good" features and track them back to frame k using a Kanade-Lucas-Tomasi (KLT) feature tracker [20] (see Figure 3 second row). Outliers are removed by the cluster filter proposed in [17]. Using a Delaunay triangulation a mesh can be generated from the features of frame $k+1$ (Figure 3 third row, right). Subsequently, the mesh is warped (backward) onto the superpixel label map of frame k (see Figure 3 third row, left) using the information provided by the KLT feature tracker. Under the assumption of a piece-wise planar surface in each triangle we use an affine transformation to warp the superpixel labels inside each triangle (forward) from frame k onto frame k+1 as shown in Figure 4. The transformation matrix $\mathbf{T}_{i,k+1}$ in homogeneous coordinates for each triangle i between frame k+1 and k is determined using the three tracked feature points of a triangle.

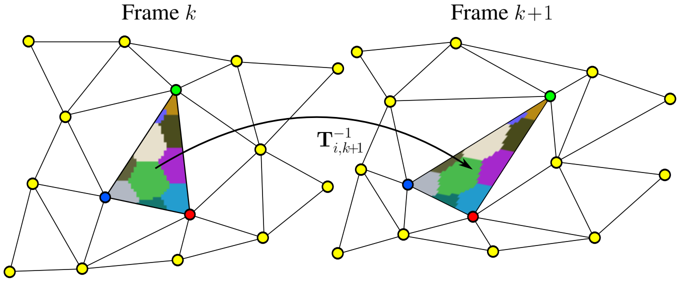

Fig. 4. Warping of superpixel labels covered by a triangle.

$$\mathbf{T}_{i,k+1} = \begin{bmatrix} t_{1,i} & t_{3,i} & t_{5,i} \\ t_{2,i} & t_{4,i} & t_{6,i} \\ 0 & 0 & 1 \end{bmatrix}_{k+1} \qquad (1)$$

The Matrix elements $t_{1,i}$ to $t_{4,i}$ determine the rotation, shearing, and scaling, whereas the elements $t_{5,i}$ and $t_{6,i}$ denote the translation. Using this transformation matrix of the triangle the homogeneous coordinates of each pixel $(x, y, 1)^T_{k+1}$ in frame $k+1$ can be transformed into coordinates $(\tilde{x}, \tilde{y}, 1)^T_k$ of frame k.

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{bmatrix}_k = \mathbf{T}_{i,k+1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}_{k+1} \qquad (2)$$

The coordinates are clipped to the nearest valid pixel position and then used to look up the label in the superpixel label map of frame k (see Figure 3 fourth row, left). The generated label map for frame $k+1$ is depicted in Figure 3 (forth row, right). To ensure that each pixel is covered by the mesh we force four features to be located at the corners of the frame and four at the middle of each frame border.

Occasionally after the warping some pixels are split-off from the main mass of a superpixel due to the transformation. As the spatial coherency of the superpixels has to be ensured these fractions are identified and assigned to a directly neighbored superpixel. As this step is also necessary if a dense optical flow is used it does not introduce any additional computational overhead.

4. EXPERIMENTS

In this section, we evaluate the quantitative performance of the proposed fast label propagation method. Therefore, we integrated our proposed method into the frameworks of [8] as well as [11] and compared the results to the original methods.

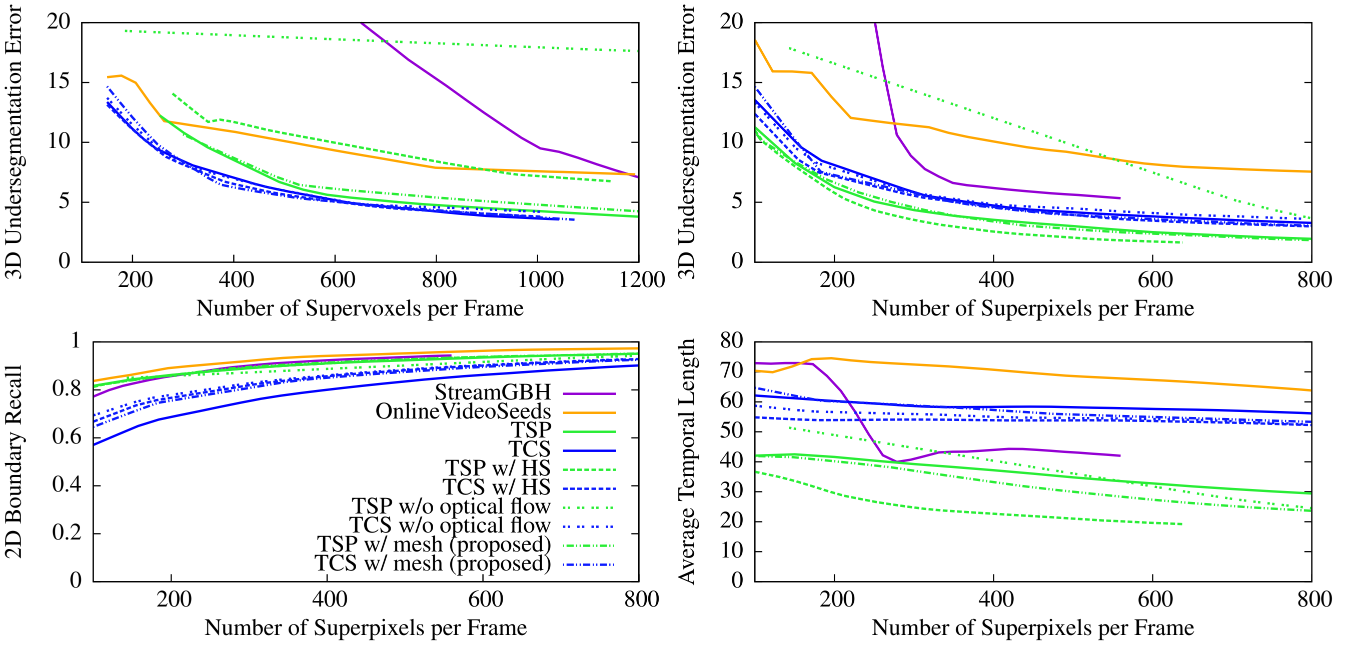

Fig. 5. Benchmark results for different temporally coherent superpixel algorithms using various superpixel label propagation approaches. Note that the 3D undersegmentation error is plotted over the number of supervoxels per sequence as well as over the number of superpixels per frame. To assess the spatio-temporal segmentation quality the latter has to be evaluated together with the average temporal length.

As a baseline we show again the results of the original methods and their modifications from Section 2. Additionally, we include the latest state-of-the-art supervoxel method from [6] for comparison. For the results we again used the eight sequences of [16] and set the default parameters as given by the authors. Figure 5 shows the 3D undersegmentation error and the average temporal length over the number of superpixels per frame as well as the 3D undersegmentation error plotted over the number of supervoxels. Note the remarks about plotting results over supervoxels and superpixels given in Section 2. Additionally, the 2D boundary recall is shown as a measure of the segmentation quality per frame. It can be seen that our proposed mesh based propagation method produces a comparable segmentation error while the average temporal length is only slightly decreased. While the 2D boundary recall stays the same for the framework of [8] the recall is improved if our propagation method is integrated into the algorithm of [11].

To evaluate the performance improvements in terms of computational cost we measured the average runtime of the dense optical flow based label propagation and the mesh based propagation. For a fair comparison we exclude the bilateral filter stage described in [8] and thus only have to consider the computation of the utilized dense optical flow [13]. For Horn & Schunck we used [21]. The comparison to our single threaded implementation was performed on an Intel Core i7-3770K @ 3.50GHz. Table 1 shows timing results solely for the superpixel label propagation task excluding segmentation. The results show that our method is 100 times faster than the originally proposed methods while creating nearly the same segmentation quality as shown in Figure 5.

	Avg.time/frame	
Method used in [8] and [11]	814.9	ms
Horn & Schunck [12] using [21]	114.3	ms
Proposed method	6.1	ms

Table 1. Average runtime needed to propagate a superpixel label map onto a new frame.

5. CONCLUSION

In this paper, we have shown that the utilization of optical flow to initialize the superpixel label maps of new frames is beneficial for video sequences with camera motion or object movement. We have also presented a new and fast superpixel label propagation method that uses sparse feature tracking combined with image warping techniques for the initialization of new frames. Our experiments indicate that using our method in a superpixel algorithm for video content a comparable segmentation quality can be achieved while speeding up the initialization process by a factor of 100.

We thank the authors of [17] for the fruitful discussion and advice.

6. REFERENCES

[1] Shu Wang, Huchuan Lu, Fan Yang, and Ming-Hsuan Yang, "Superpixel tracking," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1323–1330.

[2] Fabio Galasso, Roberto Cipolla, and Bernt Schiele, "Video segmentation with superpixels," in *Asian Conference on Computer Vision (ACCV)*, 2013, pp. 760–774.

[3] Anestis Papazoglou and Vittorio Ferrari, "Fast object segmentation in unconstrained video," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1777–1784.

[4] Xiaofeng Ren and Jitendra Malik, "Learning a classification model for segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2003, pp. 10–17.

[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 (CVPR)*, 2010, pp. 2141–2148.

[6] Chenliang Xu, Caiming Xiong, and Jason J. Corso, "Streaming Hierarchical Video Segmentation," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 626–639.

[7] Alex Levinshtein, Cristian Sminchisescu, and Sven Dickinson, "Spatiotemporal Closure," in *Asian Conference on Computer Vision (ACCV)*, 2010, pp. 369–382.

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

[9] Michael Van den Bergh, Gemma Roig, Xavier Boix, Santiago Manen, and Luc Van Gool, "Online Video SEEDS for Temporal Window Objectness," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 377–384.

[10] Matthias Reso, Jörn Jachalsky, Bodo Rosenhahn, and Jörn Ostermann, "Temporally Consistent Superpixels," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 385–392.

[11] Matthias Reso, Jörn Jachalsky, Bodo Rosenhahn, and Jörn Ostermann, "Superpixels for Video Content Using a Contour-based EM Optimization," in *Asian Conference on Computer Vision (ACCV)*, 2014.

[12] Berthold K.P. Horn and Brian G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[13] Ce Liu, *Beyond pixels: exploring new representations and applications for motion analysis*, Ph.D. thesis, Massachusetts Institute of Technology, 2009.

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

[15] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 5, pp. 898–916, 2011.

[16] Albert Y. C. Chen and Jason J. Corso, "Propagating multi-class pixel labels throughout video frames," in *Western New York Image Processing Workshop (WNYIPW)*, 2010, pp. 14–17.

[17] Marco Munderloh, Holger Meuel, and Jörn Ostermann, "Mesh-based global motion compensation for robust mosaicking and detection of moving objects in aerial surveillance," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1st Workshop of Aerial Video Processing (WAVP)*, 2011, pp. 1–6.

[18] Chris Harris and Mike Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988, vol. 15, pp. 147–151.

[19] Jianbo Shi and Carlo Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[20] Carlo Tomasi and Takeo Kanade, "Detection and tracking of point features," Tech. Rep. CMU-CS-91-132, 1991.

[21] Piotr Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)," http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html.