Evaluating Superpixels in Video: Metrics Beyond Figure-Ground Segmentation

Peer Neubert
peer.neubert@etit.tu-chemnitz.de
Peter Protzel
protzel@etit.tu-chemnitz.de
Chemnitz University of Technology
Germany

There exist almost as many superpixel segmentation algorithms as applications they can be used for. Figure 1 shows two example superpixel segmentations. So far, the choice of the right superpixel algorithm for the task at hand is based on their ability to resemble human-made ground truth segmentations (besides runtime and availability). We investigate the equally important question of how stable the segmentations are under image changes as they appear in video data (**stability-criteria**). Further we propose a new quality measure that evaluates how well the segmentation algorithms cover relevant image boundaries (**discontinuity-criteria**). Instead of relying on human-made annotations, that may be biased by semantic knowledge, we present a completely data-driven measure that inherently emphasizes the importance of image boundaries. In detail, we exploit ground truth optical flow data provided by two recently published datasets for evaluation of optical flow algorithms (KITTI [3] and Sintel[2]) to evaluate the stability- and discontinuity-citeria related to questions a) and b) in Figure 1. Both criteria are discussed, formalized and used to compare several existing superpixel algorithms with available open source implementations. For further evaluation of other algorithms, we provide the results, a Matlab implementation of the metrics and functions to interface the datasets on our website.[1]

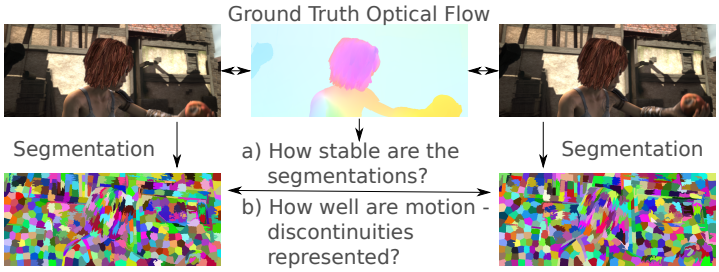

Figure 1: We propose to use ground truth optical flow fields to compare superpixel segmentation algorithms. On top left and right are two Sintel images with slight motion visualized by the optical flow field between them. Motion direction is coded by hue, saturations codes the motion magnitude. Beneath the images there are example superpixel segmentations (using *ERS*). While some object contours are visible, there seem to be a lot of spurious segment borders. In this work we provide metrics to answer questions a) and b).

1 A Metric for the Stability-Criteria

At first evaluate the stability of superpixel segmentations in image sequences or video. While superpixel borders at considerable image gradients may constantly be detected, oversegmentation algorithms tend to create lots of spurious segment borders that strongly vary under image changes. Even slight changes of the image, e.g. a small camera motion, can cause substantial changes of the produced segmentation. For some applications this might be irrelevant, while others could benefit from a superpixel algorithm with more stable segmentations. For evaluation of the stability we define the motion undersegmentation error (MUSE). The key idea is to segment two images I_1, I_2 showing the same scene before and after some changes (e.g. dynamic objects, camera motion, illumination changes), resulting in label imagess L_1, L_2. Then apply ground truth optical flow data F to compute L_1^F, a transformation of the segmentation L_1 of the first image into the view of the second image to make them comparable. Finally, use the undersegmentation error metric to evaluate how well segmentation L_1^F can be reconstructed by segments of segmentation L_2 and vice versa. Undersegmentation error is a repeatedly used measurement for comparing superpixel segmentations. We use the parameter free equation of [4]. To compare two segmentations L_1^F and L_2, and being N the total number of pixels, we define MUSE to be computed as follows:

$$MUSE = \frac{1}{N}\left[\sum_{a \in L_1^F}\left(\sum_{b \in L_2 : a \cap b \neq \emptyset} min(b_{in}, b_{out})\right)\right] \quad (1)$$

Each segment a of segmentation L_1^F is reconstructed with segments b of L_2 that overlap with a. MUSE accumulates the error that is introduced by b when reconstructing a either when b is included in the reconstruction or not. Since this is not a symmetric metric (the error diverges whether comparing L_1^F to L_2 or vice versa) we compute the average of both cases.

2 A Metric for the Discontinuity-Criteria

Superpixel segmentation algorithms are typically compared based on human annotated ground truth segmentations. Although there may exist multiple manual segmentations for each image (like in BSDS [1]), the ground truth data depends on the *semantic interpretations* of objects and their boundaries by humans. Instead of asking humans, what relevant image boundaries are, we propose to define them as image areas whith differently moving pixels in their neighborhood. This allows a direct computation of motion discontinuities as high gradients in a ground truth optical flow field. Dependent on the application, it is important to have a segment boundary near positions with high motion gradients (e.g. 3D reconstruction). To avoid arbitrarily chosen thresholds to separate important high gradients from ignored low gradients, we propose to use the following error measure: Given F, a ground truth optical flow field from an image I to another image, B the boundary image of a segmentation of image I, and $D(B)$ the distance transform of B containing for each pixel the distance to the nearest segment boundary, we define the Motion Discontinuity Error (MDE) as follows:

$$MDE = \frac{1}{\sum_i \sum_j \|\nabla F(i,j)\|_2} \sum_i \sum_j \|\nabla F(i,j)\|_2 \cdot D(B(i,j)) \quad (2)$$

In one sentence this is the Frobenius inner product of the optical flow gradient magnitude and the distance transform of the boundary image of the segmentation, divided by the sum of all gradients. A more intuitive formulation is to accumulate over all image pixels a penalty, which is the product of the strength of motion discontinuity at this pixel and its distance to the next segment border. Finally, the measure is normalized by the total amount of motion in the image.

3 Results of Superpixel Algorithm Comparison

We use the proposed criteria to compare several state of the art algorithms. With current algorithms, MDE and MUSE are somehow complementary measurements. Algorithms that perform well on one criteria often show problems with the other, in fact there is a lack of algorithms that produce stable segmentations and well resemble motion discontinuities. This opens space for further improvements and new superpixel segmentation algorithms. Based on the results on the present comparison and its previously published performance on figure-ground segmentations, the SLIC algorithm shows best balanced results.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33, 2011.

[2] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, 2012.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conf. on Comp. Vision a. Pattern Recog. (CVPR)*, 2012.

[4] P. Neubert and P. Protzel. Superpixel benchmark and comparison. In *Proc. Forum Bildverarbeitung*, 2012.

[1] http://www.tu-chemnitz.de/etit/proaut/forschung/superpixel.html