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

Abstract

There exist almost as many superpixel segmentation algorithms as applications they can be used for. 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. Further we propose a new quality measure that evaluates how well the segmentation algorithms cover relevant image boundaries. 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. Our evaluation is based on two recently published datasets coming with ground truth optical flow fields. We discuss how these ground optical truth fields can be used to evaluate segmentation algorithms and compare several existing superpixel algorithms.

1 Introduction

Superpixels are the result of an image oversegmentation or - seen the other way around, a perceptual grouping of pixels. They have become key building blocks of many image processing and computer vision algorithms. They are used for object recognition [19], segmentation [13], multi-class object segmentation [24], depth estimation [25], body model estimation [17] and many other tasks. Inspired by this multitude of applications, a considerable number of superpixel segmentation algorithms has been proposed. State of the art in comparing and benchmarking them is to evaluate their capability to resemble human-made figure-ground segmentations. We do not question the importance of this capability, yet we want to identify other qualities a good superpixel segmentation algorithm should have.

At first we want to emphasize 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 or changes in lighting, 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. The accompanied video demonstrates what is meant by an unstable segmentation.

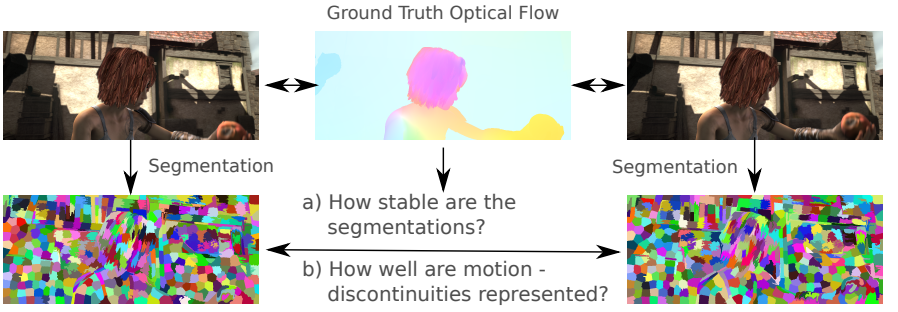

Figure 1: We propose to use ground truth optical flow fields to compare superpixel segmentation algorithms. On top left and right are two Sintel images with slight motion visualized by the optical flow field between them. Motion direction is coded by hue, saturations codes the motion magnitude, see [3] for details. 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).

Furthermore, there are some concerns regarding the human-made ground truth segmentations used for existing comparisons. Although there may exist multiple manual segmentations for each image (like in BSDS [2]), the ground truth data depends on the *semantic interpretations* of objects and their boundaries by humans. E.g., for 3D reconstruction it would be beneficial to have a more data-driven ground truth that is independent of semantic interpretation without relation to the actual 3D configuration. Imagine a superpixel based 3D reconstruction of a chair (similar to [1, 25]). Of course, separating the chair from the background is necessary. But what about separating the seat from the backrest? Or separating parts of a two-part backrest? Not all boundaries seem to be equally important. It remains unclear how this could be introduced in a manual annotation of large datasets.

In the remainder of this paper, we propose two completely data-driven metrics that can be used to evaluate and compare superpixel segmentation algorithms. In detail, we exploit ground truth optical flow data provided by two recently published datasets for evaluation of optical flow algorithms (KITTI [8] and Sintel[4]) to evaluate the following two criteria related to questions a) and b) in Figure 1:

Stability-Criteria Does the segmentation algorithm find the same regions or object boundaries independent of changes in the image?

Discontinuity-Criteria How well are motion discontinuities in the image sequence represented by the algorithms segment boundaries? E.g. the motion gradient between a moving foreground object and the background or in the interior of a non-rigid object.

We want to emphasize that we do not want to replace an evaluation of superpixel segmentation algorithms based on their ability to recover objects. But we want to provide an additional tool to advance the choice of suitable algorithms for the task at hand, i.e. for the use on image sequences and video. Section 3 gives details on the metrics we use to evaluate the above criteria, section 4 gives results of several superpixel segmentation algorithms on the two metrics and datasets. Details on the optical flow datasets can be found in the appendix. The algorithms' results, Matlab implementations of the error metrics and functions to interface the publicly available datasets are provided on our website. [1]

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

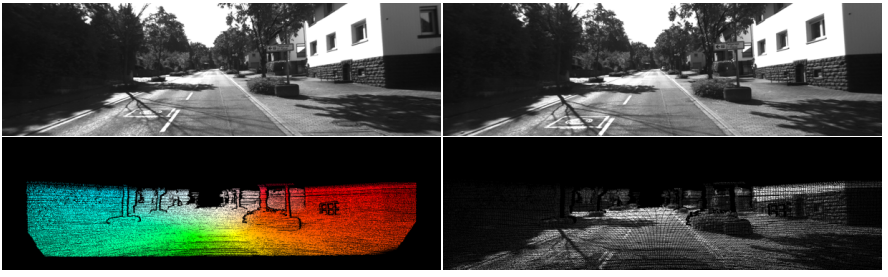

Figure 2: The top row shows two example images of the KITTI dataset. The flow field (bottom-left) illustrates the forward motion of the car mounted camera. Pixels with invalid optical flow are black. The bottom-right image shows the result of applying the flow field on the first image: Large parts of the view of the second image are recovered with pixels of the first image. However the many missing parts emphasize the advantages of synthetic datasets with much denser ground truth flow fields.

2 Related Work

To the best of our knowledge, this is the first attempt to evaluate superpixel segmentation algorithms based on optical flow data. State of the art in comparing and benchmarking superpixel segmentation algorithms is to evaluate their capability to recover human figure-ground segmentations. E.g. the Berkeley Segmentation Dataset and Benchmark [2] is a commonly used comparison framework for segmentation algorithms including superpixel segmentations [1]. It consists of 500 manually segmented images where humans were asked to outline object boundaries in the images for ground truth segmentations. Although there exist multiple manual segmentations for each image, the ground truth data depends on the semantic interpretations of objects and their boundaries by the human annotators. Other resources for manually annotated segmentations are e.g. the PASCAL VOC challenge [6] and the MSRC [21] dataset. To overcome the dependency on manual ground truth segmentations, Moore et. al [16] propose to use Explained Variation. This error metric describes the proportion of image variation that is explained if all pixelvalues within a superpixel were replaced with the superpixel mean color. Although they established a human independent metric, Explained Variation has the drawback of penalizing segments with consistent texture with large pixel variance. Koniusz and Mikolajczyk [10] measure the robustness of superpixel segmentations indirectly by the repeatability of features extracted from the segments. Their evaluation depends on the additional processing step of feature extraction and is based on a small dataset consisting of 48 images. Xu and Corso [23] evaluate several supervoxel methods based on video data. A supervoxel is the video equivalent to a superpixel. It covers a subset of the spatio-temporal lattice composed by the concatenated video frames such that all supervoxels of a video comprise the full lattice and are pairwise disjoint. They evaluate how well supervoxels cover human video segmentations based on combined spatio-temporal error metrics and Explained Variation[16]. Although the range of ready to use supervoxel algorithms is rather limited, their application should be probed whenever superpixels are applied on video data. In our previous work [18], we presented a benchmark of various superpixel algorithms including an evaluation of the algorithms robustness towards affine image transformations. The regarded affine image transformations were manually chosen and synthetically applied on the images. Affine image transformations do not cover all relevant image transformations, e.g. they exclude changing lighting conditions, occlusions and image noise. To overcome the dependence on human ground truth segmentations biased

towards semantic interpretation and the limitations of manually transformed images we propose to use ground truth optical flow data to evaluate the performance on video data. We use two image sequence datasets with known ground truth optical flow: the real world KITTI dataset and the synthetic Sintel dataset. Details on both datasets are given in the appendix.

3 Two novel Criteria for Evaluating Superpixel Segmentations based on Ground Truth Optical Flow

Given an image pair I_1 and I_2 from an image sequence, the optical flow is the vector field describing the motion of each image point between I_1 and I_2. The following sections present two novel criteria to evaluate superpixel segmentations based on ground truth optical flow.

3.1 Measuring the Segmentation Stability

Considering the **Stability-Criteria** (whether the segmentation algorithm finds the same regions or object boundaries independent of changes in the image), the key idea is to segment two images showing the same scene before and after some changes (e.g. dynamic objects, camera motion, illumination changes) and then use ground truth optical flow data to transform the segmentation of the first image into the view of the second image to make them comparable. We propose to use the following procedure to measure the stability of a segmentation between two images I_1, I_2:

1. Segment both images, resulting in label images L_1, L_2. In a label image, all pixels belonging to the same segment have the same pixel value (see Figure 1 for examples).
2. Apply the given ground truth optical flow on the first segmentation L_1 to bring it into the second image, resulting in L_1^F. In other words, we transform the labels of segmentation L_1 to the pixels of image I_2 (similar to the transformation in Figure 2).
3. Use the undersegmentation error to evaluate how well segmentation L_1^F can be reconstructed by segments of segmentation L_2 and vice versa. In particular, we do not expect to have the exact same label at a pixel in L_2 and L_1^F but if two pixels are in the same segment in L_2 we want them to be in the same segment in L_1^F, too.

The undersegmentation error is a repeatedly used measurement for comparing superpixel segmentations. We use the parameter free equation of [13]. To compare two segmentations L_1^F and L_2, and being N the total number of pixels, we define the motion undersegmentation error (MUSE) to be computed as follows:

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

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. If b is included, then the introduced error is the number of pixels of b that are outside a (defined as b_{out}). Otherwise, if we do not include b, there is a gap in the reconstruction of a and the error is the number of pixels in this gap (b_{in}, the number of pixels that are in $a \cap b$). Pixels without valid flow information are ignored. 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. The resulting error is visualized in Figure 3 and a comparison of several algorithms based on MUSE is given in section 4.

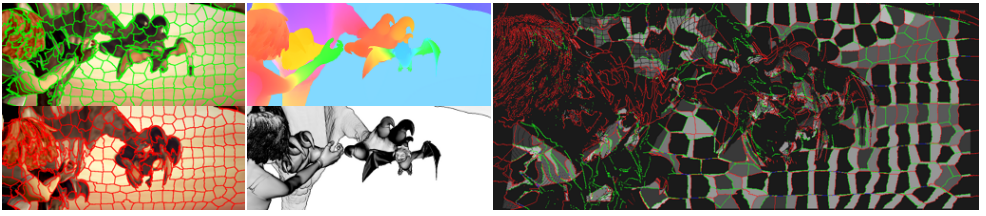

Figure 3: This figure complements the descriptions of the error metrics. The two left images show segmentations of subsequent images of a Sintel sequence with high amount of motion (using *vlSlic*). The flow field is visualized in the middle and together with its gradient beneath. Notice the high (dark) motion gradients inside objects. The image on the right is a visualization of the motion undersegmentation error (MUSE) (including segment boundaries): the gray level shows how often the pixel is counted in equation 1. The grid effect on homogeneous image regions is typical for algorithms with strong compactness constraints and can hardly be completely avoided. However, algorithms diverge significantly in the amount of border variation in homogeneous areas.

3.2 Measuring the Accordance with Motion Discontinuities

To evaluate the **Discontinuity-Criteria** (how well are motion discontinuities represented by the algorithms segment boundaries), again, we propose to use ground truth optical flow since motion discontinuities result in high gradients in the optical flow field. Figure 3 (bottom-mid) shows the gradient magnitude of the optical flow field between two subsequent images of the Sintel dataset. One can clearly see how high motion gradients (shown in dark color) appear at boundaries of moving objects, supplemented by smoother gradients inside objects. Dependent on the application, it is important to have a segment boundary near positions with high motion gradients. The intuitive argument is that the high motion gradient indicates objects or object parts that can move differently and thus should probably be handled individually in the application. This formulation includes objects as well as object parts, independently from a semantic interpretation by humans. Following this argumentation, there should be segment borders near high motion gradients to potentially handle differently moving parts individually. While the capability of a segmentation algorithm to generate the boundaries of moving objects is intuitively covered by a figure-ground segmentation evaluation, this is not the case for smooth gradients inside objects. Moreover it is unclear, at which of the smooth gradients there should be a segment border. 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:

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

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

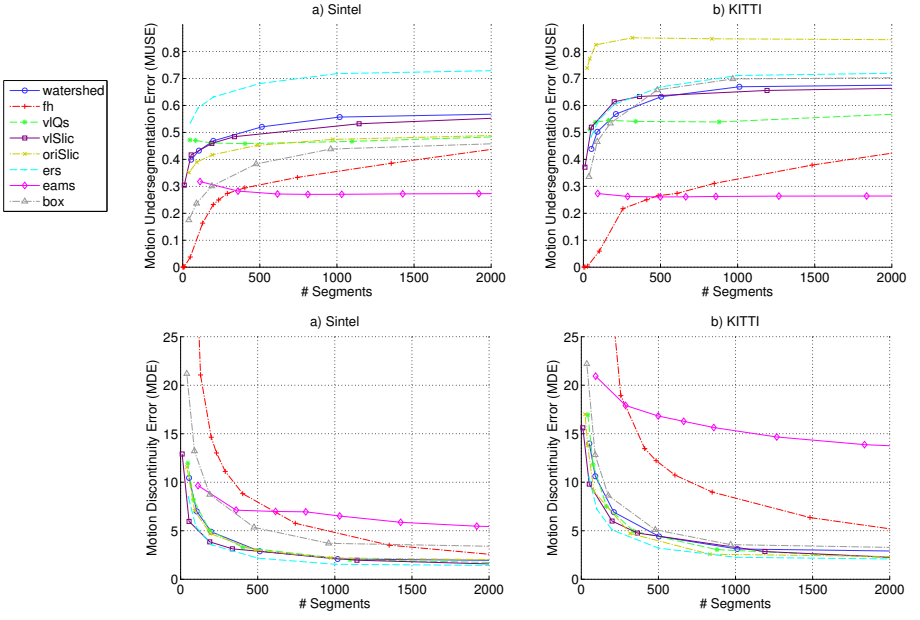

Figure 4: Results of the two metrics on several superpixel segmentation algorithms for both datasets. See text for details.

the image. We want to penalize if there is a motion discontinuity (a high gradient magnitude in the flow field) but no near segment border that can potentially separate the two differently moving image parts. However, this is a slightly optimistic measure, since we do not verify whether the nearest segment boundary really separates the two differently moving parts. Similar to measures like boundary recall, MDE favors segmentations with many boundary pixels (e.g. caused by strongly irregular segment borders) and should be used together with complementing measurements (e.g. undersegmentation error). In the following section we combine the two proposed criteria MUSE and MDE to evaluate several existing segmentation algorithms.

4 Results

4.1 Compared Superpixel Algorithms

Compared Algorithms are: Felzenszwalb-Huttenlocher Segmentation (*FH*) [7][2], Edge Augmented Mean Shift (*EAMS*) [5][12][3], Quickshift (*vlQS*) [23][4], Marker-Controlled Watershed (*WS*) [15][5], Entropy Rate Superpixel Segmentation (*ERS*) [11][6] and two implementations of Simple Linear Iterative Clustering [1] (*oriSLIC*[7] and *vlSLIC*[8]). The *oriSLIC* implementation does not strictly follow the description in [1] but incorporates some simplifications for speedup. For baseline comparison we simply divide the image into a regular grid (*BOX*). The list of superpixel algorithms is not exhaustive. Requirements for an algorithm to be used

[2] http://www.cs.brown.edu/~pff/segment/, $sigma = 1$, $minSize = 20$

[3] http://www.wisdom.weizmann.ac.il/~bagon/matlab.html, $h_S = 8$, $h_R = 8$

[4] http://www.vlfeat.org/, $ratio = 0.5$, $kernelsize = 2$

[5] OpenCV function with uniformly distributed markers http://opencv.willowgarage.com/wiki/

[6] http://www.umiacs.umd.edu/~{}mingyliu/research.html#ers

[7] http://ivrg.epfl.ch/supplementary_material/RK_SLICSuperpixels/index.html, $compactness = 10$

[8] http://www.vlfeat.org/, $regularizer = 1000$, $minRegionsize = 25$

here are an available open source implementation and reasonable runtime on images of size 1024×436 for application on videos (which excludes e.g. NC[20] and gPb-owt-ucm[2]).

4.2 Benchmark Results

The results are based on the KITTI and Sintel datasets described in the appendix. We evaluate the motion undersegmentation error (MUSE) and motion discontinuity error (MDE) criteria on both datasets independently and average over all images of each dataset. To evaluate the algorithms performance on varying numbers of segments, we run all algorithms with different parameter sets. For some algorithms the segment number is a direct parameter. For the others we varied the parameter with the largest impact on the segment number. Some algorithms have additional parameters, e.g. for balancing compactness and image gradient affinity of superpixels. Since we can not present results for all parameter settings, we use either the default parameters or these with best results on figure ground segmentations using the benchmark of [18].

For the **motion undersegmentation error**, we can not expect an oversegmentation algorithm to solely create stable segments, since they also split homogeneous image areas without image gradient support. The effects can be seen at the grid like areas on the visualization of the MUSE in Figure 3. However, we can expect algorithms to be better than the *BOX* segmentation where *all* segment boundaries lack image gradient support. Figure 4 shows the results of the comparison. The worse performance of *BOX* at the KITTI dataset also demonstrates the higher amount of systematic camera motion in KITTI images since the camera is mounted on a driving car. It is apparent that MUSE values increase with growing number of segments. This is contrary to the characteristic of the undersegmentation error when comparing a superpixel segmentation to a figure-ground segmentation (like in [18]). The intuitive reason is that when comparing two superpixel segmentations, an increasing number of segments in the segment set used for reconstruction is connected to an increased number of segments in the set that is reconstructed. Thus we do have smaller building blocks, but also want to build more filigree elements. The mean shift algorithms *EAMS* and *vlQS* can compensate this increase. Furthermore, we can distinguish three groups of algorithms: (1) Algorithms that strongly connect to image gradients and lack regulation of the segments size or distribution (like *FH* or *EAMS*) perform best. (2) In the middle, there is a large group of algorithms which perform similar. In majority, these are algorithms with strong compactness constraints. However, for Sintel this group performs worse than the baseline *BOX* algorithm. (3) Although *ERS* also contains a compactness constraint, its borders vary strongly in homogeneous image areas resulting in high motion undersegmentation error.

Since the average camera motion (that generates flow for all image pixels) increases in the KITTI dataset, results of most algorithms are worse. However, compared to the performance of the baseline *BOX* segmentation, almost all algorithms perform better. In particular *FH* and *vlQS* improve compared to the other algorithms. For KITTI, also the offset of *ERS* to the other algorithms decreases. While *oriSlic* and *vlSlic* implement the same algorithmic idea, *oriSlic* uses some simplifications (mainly for runtime reasons) that significantly influence the stability of segmentations in low gradient image regions. This causes much higher MUSE values on the KITTI dataset and make *vlSlic* preferable.

At **motion discontinuity error** superpixel algorithms with additional segment boundaries caused by compactness constraints perform superior since they decrease the average values of the distance transform of the boundary image. On the other hand algorithms that exclusively rely on image gradients (like *FH*) miss many motion discontinuities. The two SLIC implementations that already performed well on recovering object-ground segmen-

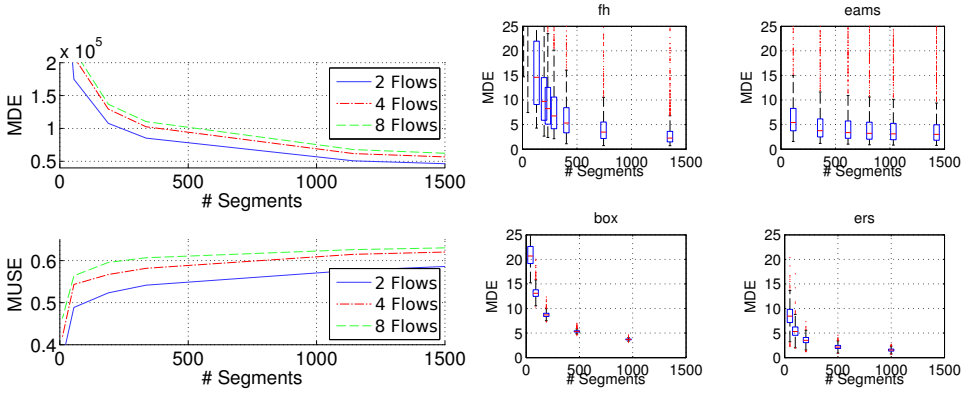

Figure 5: *(left)* Motion discontinuity error (MDE) and motion undersegmentation error (MUSE) of *vlSLIC* for concatenated motion fields of varying lengths. *(right)* Boxplots of motion discontinuity error on the Sintel dataset for some algorithms (whiskers at 1.5×IQR).

tations [13], also show good results on recovering motion discontinuities. The additional information about the statistics on MDE given by the boxplots in Figure 5 show that the high mean error of *eams* is mainly due to constantly high outlier rate. The median error is much lower. Other algorithms show statistics similar to *ERS*: median error is smaller than the mean and the number of outliers decreases with larger numbers of segments.

To evaluate the dependency of the algorithms performance on the amount of motion, Figure 5 shows exemplary results of one algorithm on concatenated flow fields (combined flow fields over multiple frames). All algorithms show similar behavior, both MUSE and MDE uniformly increase with growing sequence lengths.

In addition to the performance of the algorithms, we can state that most results are similar for both datasets. Reasons for variations can be the higher amount of camera motion in the KITTI dataset as well as the different color spaces and dynamic ranges. The overall comparison confirms the benefit of usage of a synthetic dataset.

5 Conclusions and Future Work

We proposed two novel metrics to exploit ground truth optical flow data for evaluating superpixel segmentations and compared several existing algorithms. The motion undersegmentation error (MUSE) was proposed to evaluate the stability of segmentations and used to identify different classes of algorithms. For larger number of superpixels the MUSE values rapidly increases and the overall stability of superpixel segmentations is questionable. Regarding video processing, the application of supervoxels should be probed if the ratio of framerate and image motion is sufficient.

We further proposed the Motion Discontinuity Error (MDE) to evaluate how well differently moving image parts are separated by the segmentation algorithms. MDE is a completely data-driven criteria to measure segmentation quality without bias towards human semantic interpretation. When comparing algorithms using these metrics, one should keep in mind that MDE preferes small superpixels (and irregular segment boundries) and MUSE favours large segments. 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 their previously published performance on figure-ground segmentations, the SLIC and Quickshift algorithms show best balanced results. However, the simplifications made in the original SLIC implementation *oriSLIC* significantly decrease the segment stability in homogeneous image regions. Thus we recommend the better balanced *vlSLIC* implementation from the VLFeat library. 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 (see section 1).

The results on the real world and the synthetic datasets comply in large parts. While we wait for more real world datasets appearing in the community, the synthetic dataset showed to be a rich source of ground truth information. Of future interest may also be a similar framework for comparing supervoxel algorithms. Critical parameter for supervoxel algorithms would be the ratios of framerate, segment size and amount of motion to have sufficient segment overlap between frames. I.e., some of the compared superpixel algorithms have direct supervoxel equivalents or could be extended in this direction. The cross comparison between superpixels and supervoxels would be interesting.

Appendix: Optical Flow Datasets

There exist several datasets for evaluation of optical flow algorithms. Our benchmark is based on the KITTI [8] and the Sintel [4] datasets. While the Middlebury dataset [3] is an established optical flow dataset, the amount of data with public ground truth is limited to eight sequences, each with up to eight frames. Recently, several large scale real world datasets for evaluation of optical flow algorithms have been published: KITTI [8], HCI [14] and a collection on the Image Sequence Analysis Test Site (EISATS)[9]. Due to the amount of data and the provided ground truth we decided to use the KITTI dataset for our evaluation. Images source is a stereo camera mounted on a driving car, thus there is severe camera motion in subsequent images. We only consider the left images of the stereo pairs. Ground truth optical flow is available for the training subset of the original KITTI dataset, resulting in 194 gray level image pairs of size 1226×370 for our benchmark. Each image pair shows an individual street scene. The ground truth has been generated using a 3D laser scanner. Due to limited sensor range, occlusions and other restrictions in the ground truth computation, there is flow information for about 25 % of the image pixels (averaged over all image pairs). E.g., there is no flow information for sky pixels. Note that although there is camera data available for longer sequences (20 frames per sequence), no ground truth optical flow data is available for this extended dataset.

Beside the real world datasets, computer rendered images are a great data source due to their near perfect ground truth information. The MPI Sintel Dataset [4] is based on the open source animated short film Sintel produced by Ton Roosendaal and the Blender Foundation. It provides naturalistic video sequences and is designed to encourage research on long-range motion, motion blur, multi-frame analysis and non-rigid motion. Moreover, the motion statistics of the dataset showed to be realistic [4]. MPI Sintel constitutes the second dataset in our benchmark. It consists of 23 scenes, each with 20 to 50 color images of size 1024×436. These longer scenes allow to combine flow fields to sequences over multiple frames. Moreover, there is much denser ground truth flow data. I.e., there is even ground truth motion for pixels that are occluded in one of the two scenes. This is possible since the

[9] http://tinyurl.com/EISATS-flow

ground truth optical flow is directly extracted from the data used for rendering. From the different levels of rendered details, we use the most realistic *final* rendering.

Finally we want to make some remarks on the usage of the ground truth optical flow fields. We used inverse mapping with nearest neighbor interpolation to transform label images. Pixels for which there are no ground truth flow information are ignored in computation of the metrics. The high amount of invalid pixels in the real world datasets restricts the transformation of a sparse image like a boundary image. Therefore metrics based on transformed boundary images should be avoided. However, for concatenated flows, when handling the gradients around occlusions, we can not distinguish, which of them belong to a moving object and whose are introduced solely by the occlusion. Therefore we combine the boundary maps of the segmentations of the first and the last image of the concatenated sequence and use this as input for the distance transform (this is only relevant for the results of Figure 5).

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. on Pat. Anal. and Mach. Intel.*, 34, 2012.

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

[3] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *Int. J. Comput. Vision*, 92, 2011.

[4] 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.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pat. Anal. and Mach. Intel.*, 24, 2002.

[6] M. Everingham, L.v. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision*, 88(2), 2010.

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2), 2004.

[8] 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.

[9] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Trans. Graph.*, 24, 2005.

[10] P. Koniusz and K. Mikolajczyk. Segmentation based interest points and evaluation of unsupervised image segmentation methods. In *Brit. Mach. Vis. Conf. (BMVC)*, 2009.

[11] M.-Y. Liu, O. Tuzell, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *Proc. IEEE Conf. on Comp. Vision a. Pattern Recog. (CVPR)*, 2011.

[12] P. Meer and B. Georgescu. Edge detection with embedded confidence. *IEEE Trans. on Pat. Anal. and Mach. Intel.*, 23, 2001.

[13] P. Mehrani and O. Veksler. Saliency segmentation based on learning and graph cut refinement. In *Brit. Mach. Vis. Conf.(BMVC)*, 2010.

[14] S. Meister, B. Jähne, and D. Kondermann. Outdoor stereo camera system for the generation of real-world benchmark data sets. *Optical Engineering*, 51(02), 2012.

[15] F. Meyer. Color image segmentation. In *Int. Conf. Image Processing (ICIP)*, 1992.

[16] A. Moore, S. Prince, J. Warrell, U. Mohammed, and G. Jones. Superpixel Lattices. In *Proc. IEEE Conf. on Comp. Vision a. Pattern Recog. (CVPR)*, 2008.

[17] G. Mori. Guiding model search using segmentation. In *Int. Conf. Comp. Vision (ICCV)*, 2005.

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

[19] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *European Conf. on Computer Vision (ECCV)*, 2008.

[20] X. Ren and J. Malik. Learning a classification model for segmentation. In *Int. Conf. Comp. Vision (ICCV)*, 2003.

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vision*, 81(1), January 2009.

[22] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *European Conf. on Computer Vision (ECCV)*, 2008.

[23] C. Xu and J. Corso. Evaluation of super-voxel methods for early video processing. In *Proc. IEEE Conf. on Comp. Vision a. Pattern Recog. (CVPR)*, 2012.

[24] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. *Proc. IEEE Conf. on Comp. Vision a. Pattern Recog. (CVPR)*, 2010.

[25] C. Lawrence Zitnick and Sing Bing Kang. Stereo for image-based rendering using image over-segmentation. *Int. J. Comput. Vision*, 75(1), 2007.