BENCHMARKING SUPERPIXEL DESCRIPTORS

Peer Neubert and Peter Protzel

Technische Universität Chemnitz
{peer.neubert, peter.protzel}@etit.tu-chemnitz.de

ABSTRACT

Superpixels are useful intermediate representations for many computer vision tasks. While the segmentation step is well studied, the subsequent creation of meaningful descriptors lacks this foundation. Superpixels have similar properties like affine covariant regions (keypoints), but there are fundamental differences that led to a different set of commonly used descriptors. In this paper we work towards general insights on requirements and properties of superpixel descriptors as well as a framework for experimental comparison. More precisely, we want to answer the question: Given superpixels from different images, what can superpixel descriptors tell about the ground truth overlap of the segments in the world? We propose and discuss an evaluation methodology based on image sequences with ground truth optical flow. Further, we present results of several types of superpixel descriptors and discuss the influence of the used segmentation algorithm as well as the problem of visual ambiguity in oversegmentations.

Index Terms— superpixels, descriptors, benchmark

1. INTRODUCTION AND RELATED WORK

Many computer vision pipelines compute increasingly complex features and descriptions for them. Superpixel segmentations are a low or mid level representation that showed to be useful for tasks like semantic segmentation or generating object proposals for state of the art object detection systems. While the task of creating the oversegmentation of the image is well studied, the subsequent step of describing the resulting segments is much less investigated. Various combinations of superpixels and descriptors are used in a multitude of applications (we will give a short overview later on). However, there is a lack of a deeper understanding of properties of such descriptors. In this paper we propose an evaluation methodology to deal with the fundamental question: What can superpixel descriptors tell about the ground truth overlap of the segments in the world? Fig. 1 illustrates our approach: Given pairs of images for which we know the ground truth correspondences of each pixel (the optical flow), we can create superpixels, compute descriptors for them and use the knowledge about the ground truth optical flow to evaluate the descriptors. Based on this idea, we present the overlap criterion to investigate the above question and discuss its theoretical properties. Further, we show practical results on typical superpixel descriptors ranging from shape or colour features, over keypoint-like descriptors and bag of words (BOW) approaches to deep learning features. This paper also includes results of these descriptors on a more specific matching task, underlining the problem of visual ambiguity of oversegmentations. We further gain insights on the influence of the properties of the underlying segmentation algorithms, particularly the segmentation stability. For evaluation of further descriptors, we provide a Matlab toolbox on our website[1].

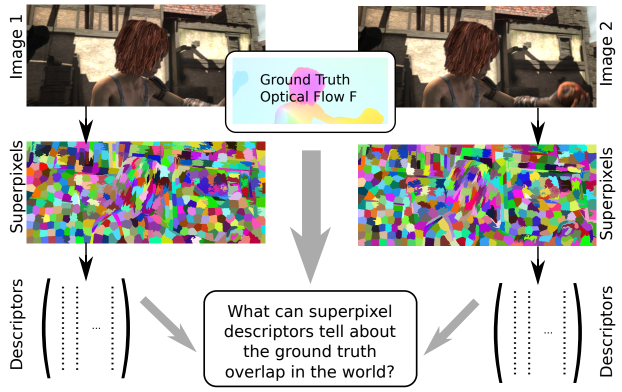
Image 1
Image 2
Ground Truth Optical Flow F
Superpixels
Superpixels
Descriptors
Descriptors
What can superpixel descriptors tell about the ground truth overlap in the world?

Fig. 1: Given superpixel segmentations of two sequence frames, we can use ground truth optical flow to evaluate superpixel descriptors.

While a complete overview of descriptors and their performance is beyond the scope of this paper, we want to point out some important steps of their development. Many superpixel descriptors are hand crafted combinations of large sets of features, sometimes in combination with dimension reduction techniques: e.g. a combination of 40 features: size, position, averages and standard deviations in multiple colour spaces, as well as responses to texture filters [1], an extension of this set to a combination of 83 features in [2], or more advanced contour and edge shape features as well as Lab colour and texton histograms in [3]. The authors of [4] extended the feature set by keypoint features (SIFT features and histograms of SIFT words). Recently deep learning features increased state of the art performance for several computer vision tasks. The available CNN features are trained for image classification, however, they also showed high discriminative power for image regions, e.g. [5]. While there exist various comparisons

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

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

of keypoint descriptors (e.g. [6]), there exists no such systematic evaluation of descriptors for superpixels. Keypoints and superpixels are somehow related: E.g. keypoint descriptors are used as superpixel descriptors [4]. Vice versa segmentations can be used as keypoint detectors [7] or spatial support for keypoint descriptors [8]. While keypoints (or affine covariant regions) and superpixels are both image parts, there are fundamental differences which raised the evolution of different types of descriptors for both. For superpixels, there is less a priori knowledge about their appearance. This contrasts e.g. corner keypoints, which are typically detected in the neighbourhood of gradients that can be used for their description. Since the superpixels cover the complete image (similar to dense SIFT), there is much higher visual ambiguity than for selected salient keypoints. Further, compared to affine covariant regions, the instabilities in the segmentation process produce much more varying regions. For these reasons, superpixel descriptors put somehow different requirements on their descriptors than salient keypoints. So far, the evaluation of superpixel descriptors has only been done as part of complete pipelines, e.g. [1,4,9]. In the following, we will present an evaluation methodology tailored to superpixel descriptors.

2. EVALUATION METHODOLOGY

2.1. Overlap Criterion

A major challenge when comparing superpixels is that they almost never show exactly the same part of the world. However, it is an intuitive demand that their descriptors should be the more similar the larger the fraction of the world is that is common to both segments. E.g. two segments showing completely different parts of the world should have a high descriptor distance, superpixels that share 70 % of common content should be more similar and segments that share 99 % should have very close descriptors. For evaluating superpixel descriptors we render the above question more precisely:

> *Given two arbitrarily shaped segments from two images that may show to some rate x the same part of the world, what is a good segment descriptor to evaluate this rate x?*

How can this property be concisely evaluated? Let us assume we have a large set of segment pairs for whose we can compute any descriptor we want to evaluate and for whose we have the ground truth overlap rate x. We propose to answer this question by running a set of binary classification tasks: How well can the descriptor distinguish whether there is ...

... any overlap between two segments?
... more than 5% overlap?
... more than 10% overlap?
$\vdots$
... more than 95% overlap?
... more than 99% overlap?

To evaluate the performance of a descriptor we run these 21 binary classification experiments and compute precision recall curves. However, this results in a vast number of curves when comparing several descriptors (21 curves for each compared descriptor). For a condensed illustration, we plot the maximum F1 score in each classification experiment over the overlap threshold of this experiment. The resulting curves can be seen in Fig. 4.

In detail, the overlap criterion is computed as follows: Given two images I_1, I_2 and their pixel wise associations F (the ground truth optical flow), such that $F(I_1) \sim I_2$. This means that $F(I_1)$ and I_2 are equivalent up to pixels that newly appear in I_2. Let there further be segmentations S_1, S_2 assigning each pixel of I_1, I_2 the unique label of the corresponding segment. We can now compute $F(S_1)$ by applying the ground truth optical flow F on these labels (i.e. move each pixel with its label according to its optical flow vector). This results in two segmentations S_2 and $F(S_1)$ from two different images that are now represented in the same image space of I_2. Now we can easily compute the rate of pixels p that are shared between each pair of segments $s_1 \in S_1$, $s_2 \in S_2$ as the "intersection over union" (IoU):

$$O_F(s_1, s_2) = \frac{|F(s_1) \cap s_2|}{|F(s_1) \cup s_2|} \tag{1}$$

This constitutes the overlap of two segments from different images given the ground truth optical flow between these images. For each classification problem, we can obtain test data with ground truth from $O_F(s_1, s_2)$ and classification can be done by simple thresholding. There are a couple of points to discuss on the overlap criterion:

Requirements on datasets Evaluating this metric requires ground truth knowledge of pixel wise associations between image pairs. It is hard to obtain such data, nevertheless there are such datasets, we use a real world and a synthetic dataset. Example images can be seen in Fig. 2 and 1. Details are given in Sec. 2.3.

Descriptor distances The overlap criterion evaluates descriptors together with a distances measure to compare them. Basis choices are Euclidean or cosine distance, however, more sophisticated distance measure could also be used. The proposed classification formulation is robust to changes of the distance measure (in contrast to e.g. evaluating the descriptor distances over the overlap directly).

Sample prior distribution The binary classification tasks are performed on pairs of segments of images related by ground truth optical flow. The vast majority of all possible pairs have no overlap. This results in a high skewness of the data: In case of 1000 segments per image, there are about one million negative samples for each positive sample. Moreover, the prior distribution is different for each binary classification task. The sensitivity of precison-recall curves and F-scores to these changes in the prior distribution would hide other effects: the F score would systematically decrease with growing overlap. Therefore we enforce constant prior distributions by resampling positive and negative samples for each classification task. A reasonable setup for a real application would be to find for each segment of one image the most similar

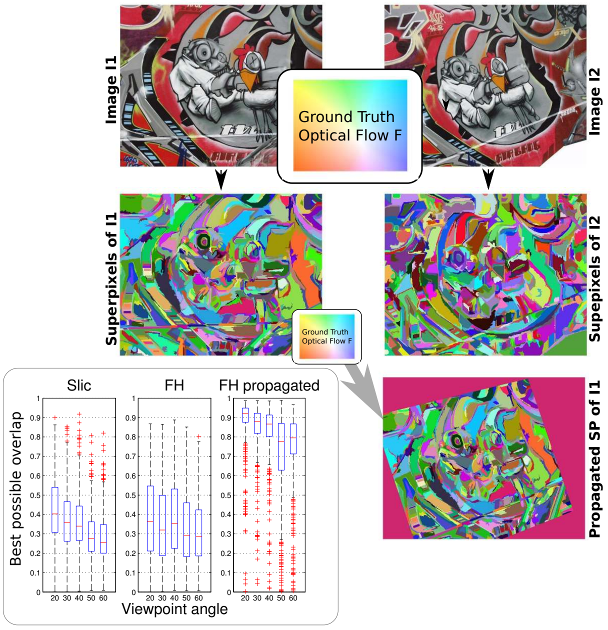

Fig. 2: Propagating superpixels scheme. The boxplots in the lower left show the best possible overlap for each segment for different viewpoint changes for individual and propagated segmentations.

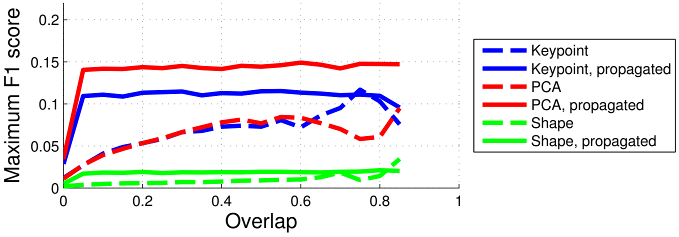

Fig. 3: Benefit when using the propagated segments for the dataset of Fig. 2 instead of independent segmentations for each images.

segment in another. This would result in about 1000 negative samples for each positive sample. Therefore we resample about one million segment pairs for each image pair keeping a prior distribution of 1:1000.

Stability of the segmentation Even the usage of a stable segmentation algorithm can not guarantee to obtain significantly overlapping superpixels between images. Therefore, we additionally evaluate single segmentations propagated to other images using the ground truth optical flow. Details can be found in Sec. 2.4.

2.2. Superpixel Matching Task

The overlap criterion evaluates a rather abstract performance for a wide range of possible tasks for a descriptor. We run a second set of experiments on a more concrete matching task: To meet the common methodology of evaluating region descriptors for keypoints, we also evaluate the ability of the superpixel descriptors to decide (classify) whether two segments are the same or not. Based on the overlap $O_F(s_1, s_2)$ we separate the set of superpixel pairs s_1, s_2 in three classes: segment pairs with zero overlap are non-matchings (they can become either true negatives or false positives), pairs with overlap larger than 70% are expected to be matched by the descriptor (they can become true positives or false negatives), and all other pairs are possible matchings (which can neither become false positives, nor false negatives). This classification problem can be evaluated by precision-recall curves with varying descriptor distance thresholds. Experimental results can be seen in Fig. 5.

Bias in case of visual ambiguity Since superpixel segmentations are an oversegmentation of the image, they tend to split a single object into several superpixels. Often these superpixels can not be separated visually. E.g. think of a segmentation of a cloudless blue sky. In particular superpixels with compactness constraints (like SLIC) will look very similar. Dependent on the dataset there may be a lot of segment pairs that are actually impossible to distinguish visually, but don't show exactly the same part of the world. Since matching of such segments may be application dependent, we consider two cases for the matching task: evaluating all segments or just evaluating the most salient segments which are expected to have low visual ambiguity. Saliency is the property of cues (or perceptions) to stand out from the surrounding. It is an important mechanism in bottom-up visual attention of the human visual system and basis for most keypoint detectors. In case of superpixels associated with descriptors, we define the saliency of the superpixel as the distance to the most similar other segment in this image. This formulation naturally incorporates the individual descriptor and its distance metric.

2.3. Datasets

The described evaluation methodology requires image pairs with pixel wise correspondences. A well known dataset is the affine covariant region dataset (**ACR**) [10]. It consists of eight image sequences each containing six images which are related by known homographies. The image sequences are designed for evaluation of the influence of viewpoint changes, rotation, zoom, blur, illumination changes and JPEG compression on keypoint detectors and descriptors. Due to space limitation, we present combined results of all sequences in this paper. We exclude the JPEG compression sequence since this is not a natural image change. Main drawback of the ACR dataset are the limited size and the non perfect ground truth (e.g. due to a not perfectly planar scene). Thus we also run experiments on all scenes of the **Sintel** dataset [11]. It consists of 23 scenes, each with 20 to 50 colour images from the open source animated short film Sintel. It provides naturalistic video sequences and is designed to encourage research on long-range motion, motion blur, multi-frame analysis and non-rigid motion. Sintel ships with almost perfect ground truth, i.e. there is even ground truth motion for pixels that are occluded in one of the two scenes.

2.4. Creating and Propagating Superpixels

We apply two commonly used oversegmentation algorithms to obtain about 1000 superpixels per image: **SLIC** [12] superpixels are regularly sized and shaped. **FH** [13] provides

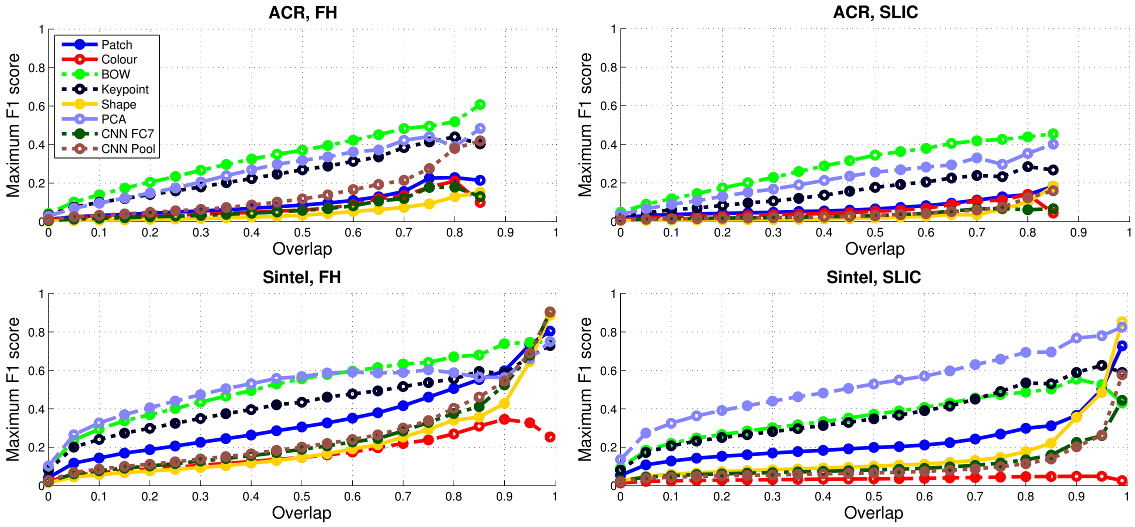

Fig. 4: Results on Overlap criterion for both datsets SLIC and FH segmentations. (Higher is better, best viewed in colour.)

an image oversegmentation into very differently sized and shaped segments that are well aligned with salient image gradients. These two algorithms produce apparently different segmentations, as can be seen in Fig. 1 and 2. Although these two algorithms showed to be robust against image changes [14], there are only few segment pairs with very high overlap in presence of severe image changes. The boxplots in Fig. 2 illustrate the maximum overlap that can be reached for each superpixel for the challenging "graf" image sequence of ACR (viewpoint changes up to 60°). While the average values for SLIC and FH are similar, FH provides more high overlap segments for large viewpoint changes. To identify effects caused by these properties of the segmentation algorithms, we created a test set with higher overlap (FH propagted). Instead of segmenting both images of the corresponding pair, we segment the first and apply the ground truth optical flow to propagate the segments to the image space of the second frame.

3. EXPERIMENTAL RESULTS

To identify properties of superpixel descriptors we compare examples of several often used region descriptor types with the proposed methodology: A native descriptor is a rescaled **Patch** of the superpixel. We rescale the bounding box around the superpixel to a 10×10 colour patch and blacken the parts that do not belong to the segment. **Colour** histograms are an often used descriptor for regions with varying shapes. There exist various colour spaces with different properties, e.g. RGB, CIE Lab or opponency colour space. Based on baseline experiments, we chose an RGB histogram with 30 bins, equally spread over the three channels. As descriptor for the segment **Shape**, we compute the mean and standard deviation over the shape context [15] of boundary pixels. For a **Keypoint** like descriptor, we compute a SURF descriptor for each superpixel (parts that do not belong to the segment are blackened). As texture descriptor, we use a 100 bin histogram of densely sampled SIFT words in a **BOW** scheme. The vocabulary was trained on the disjunct BSDS500 training data. We further evaluate a combination of colour, keypoint, shape and texture features by creating a **PCA** representation of the concatenated RGB, SURF, shape context and BOW features. We segment images and compute descriptors on the BSDS500 training data to obtain a 100 dimensional representation by PCA. Although they were learned for a fundamentally different task, we also evaluate **CNN Pool2** and **CNN FC7** features from CNN-M [16].

Overlap Criterion Fig. 4 shows results of the above superpixel descriptor types on the proposed overlap criterion. As a reminder, each point of theses curves illustrates how well the descriptor can decide whether superpixel pairs have at least the corresponding overlap or not. There are two general tendencies: 1. More sophisticated descriptors tend to perform better (except for CNN). 2. The higher the overlap, the easier it is to solve the classification problem. It seems to be very hard for the evaluated descriptors to decide whether there is a small overlap between segments or no overlap. However, this may be related to important generalisation capabilities in tasks like semantic label assignment. The larger the overlap, the more similar is the problem to the keypoint matching problem for which well studied descriptors exist. Fig. 3 shows that particularly the low overlap classification problem could benefit from more stable segmentations. The ACR dataset does not provide enough segments with valid ground truth overlap >85% for a robust evaluation. While the Sintel dataset was created in a fundamentally different way, the results are similar. However, the overall performance increased and ordering of some of the descriptors changed. Moreover, the Sintel dataset provides enough segment pairs with valid high ground truth overlap to see a massive improvement for most of the descriptors at near perfect overlaps.

Matching Task Fig. 5 evaluates the capabilities of the descriptors in the more specific matching task of Sec. 2.2. The top row plots are the results for matching all superpix-

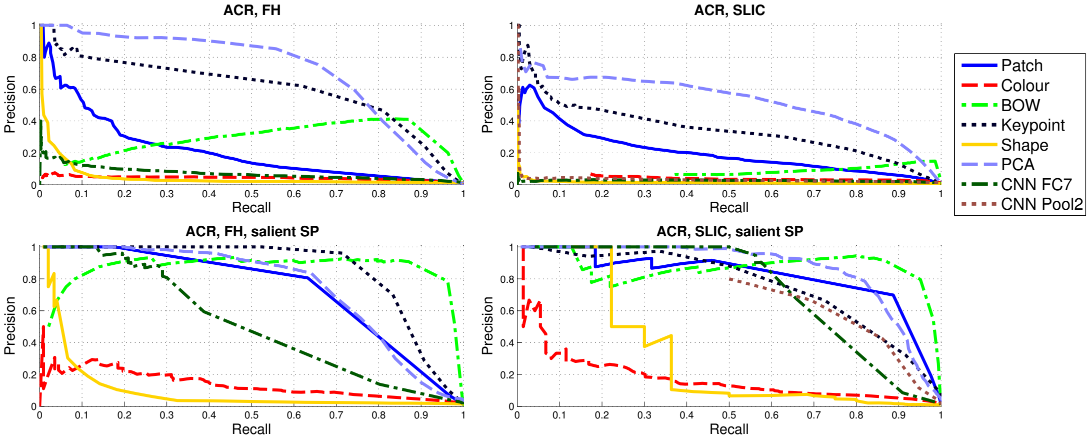

Fig. 5: Results for the matching task on the ACR dataset for SLIC and FH segmentations. (Top right is better, best viewed in colour)

els of each image pair. As expected, most descriptors suffer from the large visual ambiguity. This holds particularly for the more uniformly shaped SLIC segments and histogram based features that do not incorporate spatial arrangements. All descriptors benefit from the reduction of the problem to matching the 200 most salient superpixels per image as can be seen in the lower part of this figure. For typical keypoint-like applications one would wish high recall at 100% precision. To this end, the combination of the Keypoint descriptor with the irregularly shaped segments performs best (a combination similar to [7]). For other tasks that require high recall at considerable precison, dependent on the amount of visual ambiguity, BOW or PCA descriptors may be a good choice.

Some features like the used Shape or CNN features are sensitive to the different spatial supports of segment pairs with low overlap. However, one has to keep in mind that the used CNN features were trained for a fundamentally different task than describing such small structures. Considering the overall performance, from the evaluated descriptor types, the PCA descriptor (a combination of different features with dimensionality reduction) provides the most promising performance.

4. CONCLUSIONS

We discussed the problem of what superpixel descriptors can tell about ground truth overlap of segments, more precisely how large the part of the world is that is common to two segments. We used datasets with known ground truth optical flow in the proposed overlap criterion, a set of classification tasks, to answer this question and presented results on several types of descriptors. Obviously, the comparison presented here is not exhaustive but yielded some insights: It is particularly hard to distinguish segment pairs with small overlap from non overlapping ones and more stable segmentation algorithms could particularly help in these cases. Finally, we showed results on a more practical matching task and underlined the difficulties caused by visual ambiguity of oversegmentations. More detailed results and implementation of the evaluation methods are available from our website.

REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, “Matching words and pictures,” *J. Mach. Learn. Res.*, 2003.

[2] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, “Multi-class segmentation with relative location prior,” *Int. J. Comput. Vision*, vol. 80, no. 3, pp. 300–316, Dec. 2008.

[3] C. Gu, J. Lim, P. Arbelez, and J Malik, “Recognition using regions,” in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2009.

[4] J. Tighe and S. Lazebnik, “Superparsing: Scalable nonparametric image parsing with superpixels,” in *Proc. of European Conference on Computer Vision (ECCV)*, 2010.

[5] J. Girshick, R.and Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *CoRR*, vol. abs/1311.2524, 2013.

[6] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” in *Proc. of British Machine Vision Conference (BMVC)*, 2002.

[8] F. Navarro, M. Escudero-Violo, and J. Bescos, “Sp-sift: enhancing sift discrimination via super-pixel-based foreground-background segregation,” *Electronics Letters*, vol. 50, no. 4, pp. 272–274, Feb. 2014.

[9] X. Wang, M. Yang, S. Zhu, and Y. Lin, “Regionlets for generic object detection,” in *Proc. of Intern. Conf. on Computer Vision (ICCV)*, 2013.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, “A comparison of affine region detectors,” *Int. J. Comput. Vision*, vol. 65, pp. 43–72, 2005.

[11] 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.

[12] 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.*, vol. 34, 2012.

[13] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *Int. J. Comput. Vision*, vol. 59, no. 2, 2004.

[14] P. Neubert and P. Protzel, “Evaluating Superpixels in Video: Metrics Beyond Figure-Ground Segmentation.,” in *Proc. of British Machine Vision Conference (BMVC)*, 2013.

[15] S. Belongie and J. Malik, “Matching with shape contexts,” in *Content-based Access of Image and Video Libraries. Proc. of Ws. on*, 2000.

[16] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” in *British Machine Vision Conference*, 2014.