Patch Growing: Object segmentation using spatial coherence of local patches

Marc Masias, Albert Torrent, Xavier Lladó and Jordi Freixenet
Institute of Informatics and Applications, University of Girona, Girona, Spain

Abstract.
Object segmentation is a challenging and important problem in computer vision. The difficulties to obtain accurate segmentations using only the traditional Top-down or Bottom-up approaches have introduced new proposals based on the idea of combining them in order to obtain better results. In this paper we present a novel approach for object segmentation based on the following two steps: 1) oversegment the image in homogeneous regions using a Region Growing algorithm (Bottom-up), and 2) use prior knowledge about the object appearence (local patches and spatial coherence) from annotated images to validate and merge the regions that belong to the object (Top-down). Our experiments using different object classes from the well-known TUD and the Weizmann databases show that we are able to obtain good object segmentations from a generalistic segmentation method.

Keywords. Image Analysis, Image Segmentation, Top-down and Bottom-up Strategies

Introduction

Segmenting objects of interest in images is an important problem in computer vision and computer graphics. A classical approach for the object segmentation task is the Bottom-up strategy [1,2]. The idea is to segment a given image without knowing any prior knowledge about the object of interest we are looking for. However, the high variability among the regions of an object as well as the regions of the background, make difficult the process of merging the object regions. On the other hand, during the last years many works have also been presented in order to tackle the problem of object segmentation using a Top-down strategy [3,4]. These algorithms use a known object model (prior knowledge) to get the object segmentation in new images. For instance, following this strategy, Leibe and Schiele [3] propose to generate a codebook of object parts and to obtain object probabilities matching the image with this codebook. Although these Top-down methods obtain promising results, they still fail to provide accurate object segmentations, specially missing important details in the boundaries.

The difficulty of obtaining accurate object segmentations using only a Top-down or a Bottom-up approach has introduced the idea of combining both methods. For instance, Borenstein et. al [5] propose a combination of a Top-down and Bottom-up segmentation. On the Top-down approach they first obtain a segmentation by matching a set of templates from training images. Afterwards, they obtain the final segmentation from regions extracted during the Bottom-up approach. In a further work, Borenstein and Ma-

lik [6] extend this proposal introducing a multiscale approach in the Top-down process and therefore avoiding the restriction of knowing the object size. Similarly, Levin and Weiss [7] also propose to combine Top-down and Bottom-up approaches. However, they incorporate a low-level segmenter during the Top-down process which helps to segment objects without a defined shape. In a different way, Cao and Fei-Fei [8] propose to add spatial coherency from small parts of the objects (local patches) to the traditional bag-of-words approach. In a first step, they oversegment the image in homogenous regions using a graph based segmentation algorithm [9] and characterize these object regions using SIFT descriptors [10]. Afterwards, the regions are labeled by considering neighboring appearance of local descriptors, and imposing also spatial constraints.

Following the ideas of these recent works, in this paper we present a new approach for object segmentation combining Top-down and Bottom-up strategies. Our proposal is based on two different steps. First of all, we oversegment the image in homogeneous regions using a standard image segmentation method. Afterwards, we apply a Top-down approach which uses a set of annotated images to extract prior knowledge of the object using local appearence and spatial coherence information. The idea is that we validate the regions obtained from the Bottom-up method with the information from the annotated images. In order to evaluate our approach we use the well-known object TUD database[1] [11], which contains 100 images for the car class and 111 images for the cow class, and Weizmann database[2] [5], which contains 327 images for the horse class. The obtained results validate the performance of our method.

The rest of the paper is structured as follows. Section 1 describes the framework of our proposal. Section 2 shows the obtained results, discussing the parameter optimization, and comparing our results with the state-of-the-art approaches. Finally, the paper ends with discussions and conclusions, pointing out also the undergoing future work.

1. Our proposal for object segmentation

Figure 1 shows an overview of our proposed approach. The idea is to segment objects of an specific class combining a Bottom-up and a Top-down strategy given a new image with their detection provided by a bounding box (see top right image). The need of knowing the object detection would mean that users have to annotate the bounding box of the test images. However, while the more challenging problem of object segmentation is still far to be fully solved, different works have been presented during the last years for the object detection task providing very good results [12,13,14,15]. Therefore, we can assume that the object center and bounding box of the object are accurately identified with the detection process.

Our approach is divided in two main steps. First we apply a Bottom-up strategy to oversegment the bounding box images in homogeneous regions. More details about the Bottom-up segmentation are given in the following section. Afterwards, in a second step, a Top-down method is applied to determine the final segmentation of the object. Our Top-down approach is based on using a prior knowledge provided by a reference set of images with their ground truth annotations (top left corner). At this point, and starting from object center, all the regions are analyzed checking their spatial coherency with

[1] TUD database can be downloaded from: http://www.vision.ee.ethz.ch/~bleibe/data/datasets.html

[2] Weizmann horses database can be downloaded from: http://www.dam.brown.edu/people/eranb/databases

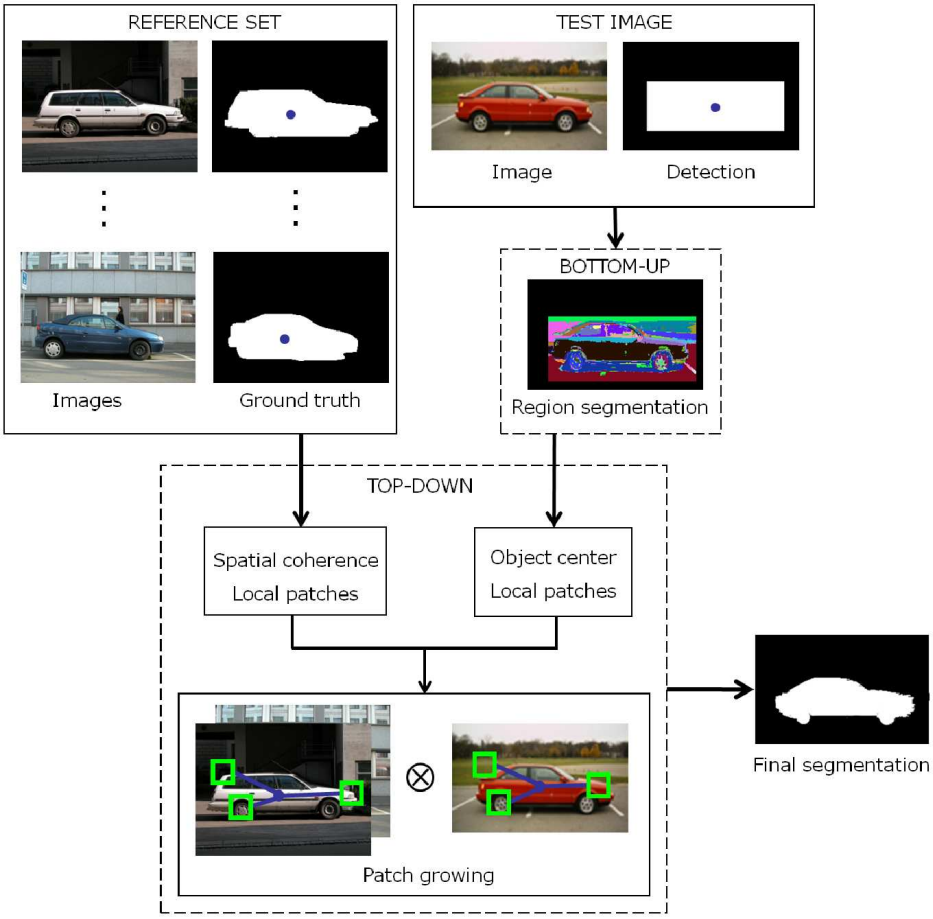

Figure 1. Graphical representation of our proposal. Given an image and its bounding box, we perform an oversegmentation using a Bottom-up approach. Afterwards, a Top-down strategy is used to validate the obtained regions using prior knowledge from a reference set of annotated images.

the reference set (patch growing). Finally, all the regions validated with the reference set are merged, obtaining the final segmentation. Section 1.2 describes in more details this Top-down strategy.

1.1. Region segmentation

In this first step we perform an oversegmentation of the bounding box in homogeneous regions using a Bottom-up strategy. We could use here several approaches for this purpose [16,2]. However, as one of our objective is to get a good final object segmentation from a simple general purpose segmentation method, we decided to use a region growing algorithm [17], which provides satisfactory results with a low computational cost. In order to perform our experiments with the region growing algorithm we evaluated the use of several features providing different initial oversegmented images. For instance, we tested the use of the RGB and HSV color space components, and texture descriptors extracted from coocurrence matrices [18] and Laws masks [19] methods. After an ex-

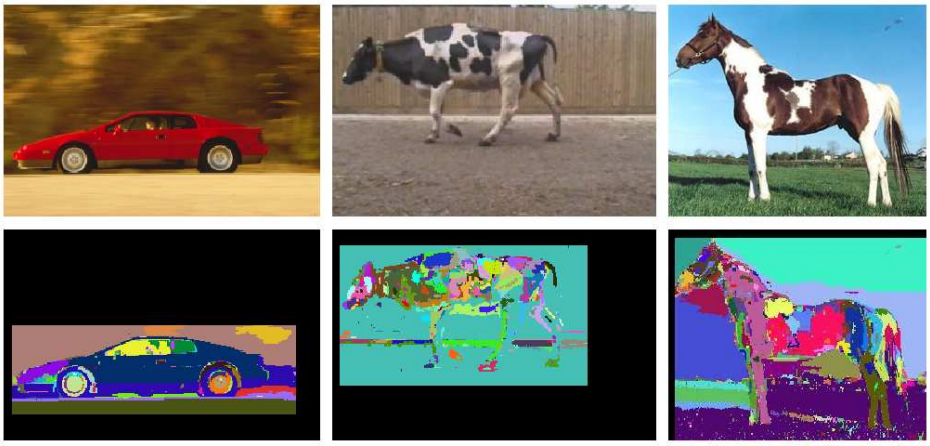

Figure 2. Some segmentation examples (images of a car, cow and horse) using the region growing algorithm.

haustive set of tests, we decided to use only RGB and HSV components, basically due to the fact that more complex combinations did not improve the quality of the segmentations, but increased severally the computational cost. Figure 2 illustrates some results of the region growing algorithm in images of three different object classes: car, cow and horse. Note that the region segmentation is limited by the location of the object and restricted to the bounding box area. In our proposal these bounding boxes comming from the detection step are increased by a factor of 10% to be sure of containing the whole object. Figure 2 illustrates the restriction of the area of interest, where the non interesting zones have not been segmented during the region growing segmentation.

Once an image has been segmented with the region growing algorithm, an adjacency graph of the spatial relationship between the objects is generated. This adjacency graph will be then used to analyze the regions in an optimal order during the patch growing step, described in the following section.

1.2. Patch growing algorithm

Once we have the bounding box images segmented in homogeneous regions we perform a patch growing algorithm to validate and select the regions that belong to the object. The patch growing algorithm starts from the region of the object center, and continues with their adjacent regions sorted by their proximity to the object center (adjacency graph). When a region is added to the object segmentation all their adjacent regions are then marked as possible parts of the object, and they will be processed. This process is repeated until all the regions have been analyzed. The most important part here is the criterion used to determine if a region belongs to the object or not. At this point is where the prior knowledge extracted from the ground truth images is used.

The idea is to compare the local appearence of all the segmented regions with local patches of the training images at the same relative position from the object center. First of all, some patches are extracted from each segmented region. These patches are dispersed randomly around a given region, with the restriction that they have to be centered on a pixel of the region. The number of patches extracted and their size depends on the region and the object size respectively. Afterwards, every patch is compared with all the

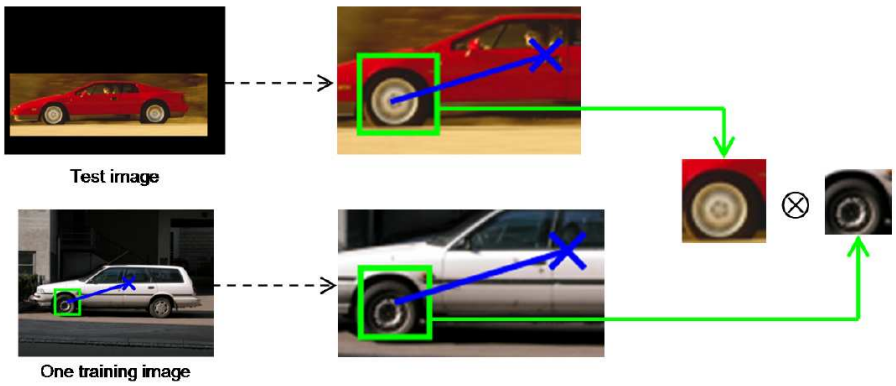

Figure 3. Correlation between a patch (template) of a test image region and the same patch opened in the relative position from the object center of a training image. Note that the second patch is smaller than the template.

training images using the normalized cross correlation [20]. Figure 3 illustrates this step in more detail, where a patch is extracted from the region of the test image and a template is extracted from the training image at the same relative position. Note that the patch opened in the training image is bigger, since we only know the approximate position where the patch should be, and we have to search for it in a bigger area. Finally, the two patches are compared using a normalized cross correlation in order to get their similarity.

From each patch we obtain a correlation with every training image. High values of correlation indicate that they are similar regions. We get the maximum correlation of each patch with all the training images. Finally we consider the region as part of the object if a certain number of patches have a high correlation with the training images, defined by a correlation threshold. Note that with this procedure is very difficult to deal with small regions, where we do not have enough local information (patches). In this situation, our algorithm do not take them into account during the patch growing, and at the end of this process, the small regions are validated looking at their segmented adjacent regions.

2. Experimental Results

The aim of our experimental results is twofold: 1) to evaluate the performance of our proposal compared with current state-of-the-art approaches, and 2) to demonstrate the validity of our approach. In order to provide a quantitative and qualitative evaluation of our proposal we have used three different object classes from two well-known databases: cars (side view) and cows (side view) from the TUD database [11] and horses (side view) from the Weizmann database [5]. These databases contain 100, 111 and 327 images for the car, cow and horse classes respectively, with their corresponding ground truth annotations. We then scaled the images so that per each specific object the corresponding bounding box had a similar size.

For testing the segmentations of each class, we decided to use a cross validation method. We divided all the images of one particular class in four sets, and we segmented each image set using the other three sets of images as the reference set. Our method works with four basic parameters. In this sense, we used an specific threshold for each

Table 1. Obtained results (percentage of well classified pixels and area overlap) for the three object classes: cars, cows and horses.

		Cars	Cows	Horses
Percentage	Mean	0.9238	0.9255	0.8630
	Std. deviation	0.0403	0.0671	0.1015
	Median	0.9298	0.9460	0.8878
Area overlap	Mean	0.6926	0.6520	0.6213
	Std. deviation	0.1503	0.1114	0.1421
	Median	0.7177	0.6815	0.6423

object class in the region growing process, empirically determined from the training images. Another important parameter is the threshold used in order to determine if a region is candidate to belong to the object from the correlation of the local patches. This parameter is constant through the object classes and has also been empirically determined. Finally, the number of patches extracted from each region and the size of those patches are calculated for each region depending on the size of the region and the size of the object respectively, being patches of around 30 pixels.

To evaluate the performance of our segmentation method we obtained some statistics from the segmented images. From every image we calculate the true and false positive fractions (TP and FP) and the true and false negatives fractions (FP and FN). With these values we are able to obtain meaningful statistics, like the percentage of pixels successfully classified. Although this is the most common quantitative measure used on the state-of-the-art approaches, these results can be too much optimistic since sometimes the object represents a small part of the image, having the background segmentation an important influence to the final percentage. For this reason we also used the area overlap measure, computed as: $Area\ Overlap = \frac{TP}{TP+FP+FN}$. This gives us a more robust measure than the percentage of pixels well classified. The global statistics obtained for each class are shown in Table 1, where we present the mean, standard deviation and median results.

We also compared our segmentation results for horses, cows and cars with those reported by recent segmentation approaches [4,21,5,6,7,8]. The aim was to provide a general trend of the performance of our segmentation approach with respect to different strategies. However, we want to clarify that not all these works used the same databases and number of images. For instance, for the cars TUD class we obtain a little worse results than those reported by Winn and Jojic [4]. For the cows class we obtain better performances than those reported by Levin and Weiss [7]. While for the horses class we achieve worse results than those reported in [4,6,21], but better results than the ones presented in [8] although in this work authors tested the images in inverted direction and under significant occlusions. To sum up, our segmentation approach provides competitive results with the current state-of-the-art.

Figure 4 illustrates some qualitative segmentation results obtained with our approach for the three tested classes. Columns (a) and (d) show the original image, columns (b) and (e) the obtained segmentation results overlayed in green, and columns (c) and (f) show the ground truth segmentations. Note that in general the segmentations are good although in some cases the results are not as accurate as it was expected. See for instance the last car of Figure 4. Observe also that the common problem in the cars class is on the wheels segmentation. Very often the tires are joined with the road during the region growing

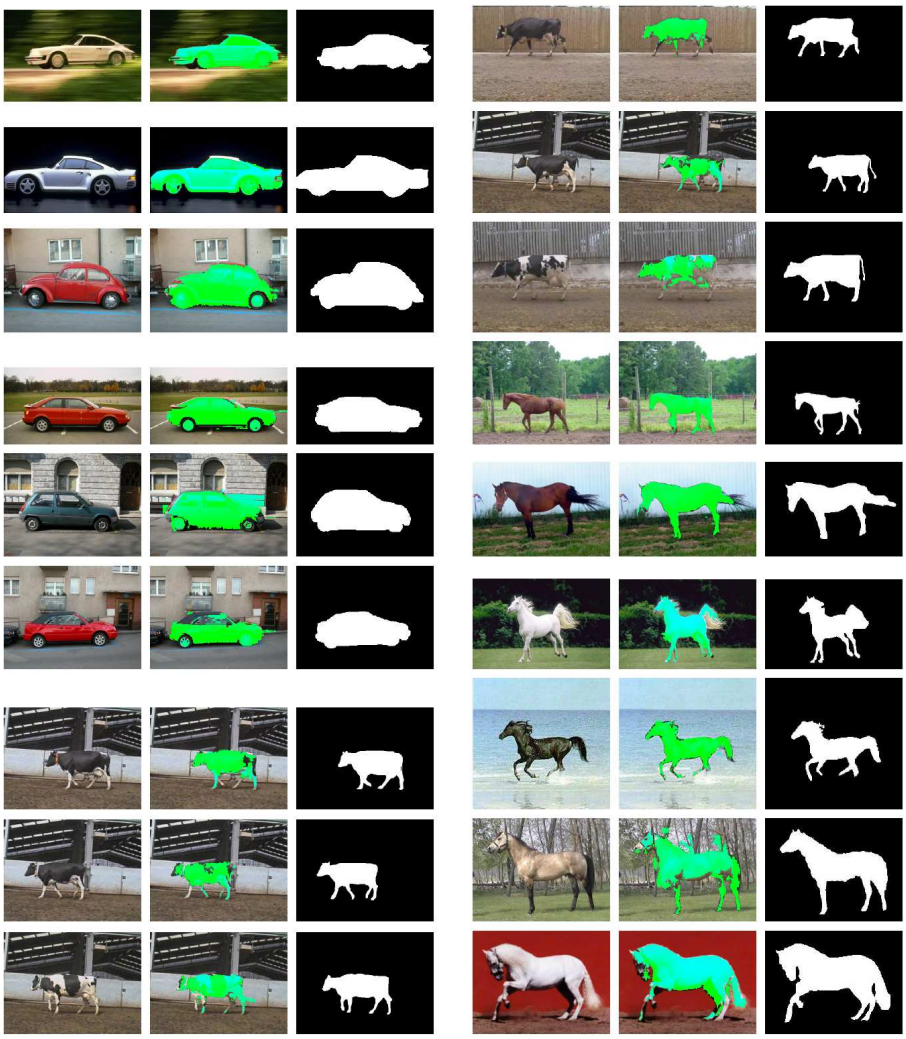

Figure 4. Object segmentation results: (a) and (d) original image, (b) and (e) segmentation results overlayed in green, and (c) and (f) ground truth.

process due to its darkness properties, so our patch growing approach cannot correctly recognize them. On the other hand, for the other classes, specially for the cows, the patch growing algorithm do not recognize the legs, since they are very thin and difficult to segment.

3. Conclusions

In this paper we have presented a new approach to perform single object segmentation combining Bottom-up and Top-down strategies. Our approach segment object classes in two steps: 1) oversegmentation of the image in homogeneous regions using the region

growing algorithm, and 2) patch growing algorithm to validate and merge the regions that belong to the object. This step is achieved using prior knowledge from annotated images: local patches and spatial coherency.

We evaluate our proposal using 3 different classes: cars and cows from the TUD database and horses from the Weizmann database. The obtained results demonstrate that our approach obtain good object segmentations. The experiments presented in this paper assume that the object detection (bounding box and object center) is known, although this could be accurately achieved by the detection approach proposed Murphy et al. [12].

This research work open a set of further research works that could improve our segmentation strategy. First of all, the use of more sophisticated region segmentation methods, such as mean shift [16] or normalized cuts [2], may produce better initial segmentations instead of a single oversegmented result. In this paper we have used the region growing for its simplicity, although more sophisticated techniques, such as the ones based in SIFT [10] or SURF [22], could be applied for this purpose without an increase of the computational cost. However, from our experiments, we have seen a gap for improvement on this aspect. Moreover, we could also use a set of initial segmentations and combine them to obtain a better final object segmentation. This has also been analyzed in [23,24]. Finally, we are currently working on segmenting more object classes from more complex databases like LabelMe [25].

Acknowledgements

This work has been supported by MEC grants AYA2007-68034-C03-(01/02/03) and TIN2007-60553. Massias is supported by a BTI-UdG 2008 scholarship and Torrent is supported by Spanish government scholarship AP2007-01934.

References

[1] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. Graph Cut Based Image Segmentation with Connectivity Priors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[2] Timothee Cour, Florence Benezit, and Jianbo Shi. Spectral segmentation with multiscale graph decomposition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1124–1131, 2005.

[3] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *Proc. British Machine Vision Conference*, pages 759–768, 2003.

[4] J. Winn and N. Jojic. Locus: learning object classes with unsupervised segmentation. In *Proc. International Conference on Computer Vision*, pages 756–763, 2005.

[5] Eran Borenstein, Eitan Sharon, and Shimon Ullman. Combining top-down and bottom-up segmentation. *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[6] E. Borenstein and J. Malik. Shape guided object segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 969–976, 2006.

[7] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *Proc. European Conference on Computer Vision*, 2006.

[8] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Proc. International Conference on Computer Vision*, pages 1–8, 2007.

[9] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[11] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[12] K. Murphy, A. Torralba, D. Eaton, and W. T. Freeman. Object detection and localization using local and global features. In *Sicily workshop on object recognition*, 2005.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[14] A. Shotton, J. andBlake and R. Cipolla. Multiscale categorical object recognition using contour fragments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1270–1281, 2008.

[15] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[16] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[17] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[18] Robert M. Haralick, K. Shanmugam, and Its'hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, 1973.

[19] K. I. Laws. Rapid texture identification. In *Conference on Image Processing for Missile Guidance*, volume 238, pages 367–380, 1980.

[20] J. P. Lewis. Fast normalized cross-correlation. In *Vision Interface*, pages 120–123, 1995.

[21] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj Cut. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 18–25, 2005.

[22] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[23] Zhuowen Tu and SongChun Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):657–673, 2002.

[24] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1605–1614, 2006.

[25] B. Russell, A. Torralba, K. Murphy, and W.T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.