Towards Fully Automatic Image Segmentation Evaluation

Lutz Goldmann[1], Tomasz Adamek[2], Peter Vajda[3], Mustafa Karaman[1], Roland Mörzinger[4], Eric Galmar[5], Thomas Sikora[1], Noel E. O'Connor[2], Thien Ha-Minh[3], Touradj Ebrahimi[3], Peter Schallauer[4], and Benoit Huet[5]

[1] Technical University Berlin (TUB), Berlin, Germany
[2] Dublin City University (DCU), Dublin, Ireland
[3] Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland
[4] Joanneum Research (JRS), Graz, Austria
[5] Institut Eurecom, Sophia-Antipolis, France

Abstract. Spatial region (image) segmentation is a fundamental step for many computer vision applications. Although many methods have been proposed, less work has been done in developing suitable evaluation methodologies for comparing different approaches. The main problem of general purpose segmentation evaluation is the dilemma between objectivity and generality. Recently, figure ground segmentation evaluation has been proposed to solve this problem by defining an unambiguous ground truth using the most salient foreground object. Although the annotation of a single foreground object is less complex than the annotation of all regions within an image, it is still quite time consuming, especially for videos. A novel framework incorporating background subtraction for automatic ground truth generation and different foreground evaluation measures is proposed, that allows to effectively and efficiently evaluate the performance of image segmentation approaches. The experiments show that the objective measures are comparable to the subjective assessment and that there is only a slight difference between manually annotated and automatically generated ground truth.

1 Introduction

Image segmentation is a fundamental step for many multimedia analysis steps, since it helps to understand and describe the structure of the data and identify relevant objects. A large number of different approaches has been proposed and several surveys [1] provide a comprehensive overview of this domain. Although image segmentation has been a very active research field only little work has been done concerning image segmentation evaluation [2]. Approaches for image segmentation evaluation can be categorized according a taxonomy depicted in figure 1.

Most of the feature based approaches, e.g. [3] rely on a ground truth that partitions the image into multiple disjoint regions similar to the output of most automatic image segmentation approaches. Unfortunately the ground truth of

J. Blanc-Talon et al. (Eds.): ACIVS 2008, LNCS 5259, pp. 566–577, 2008.

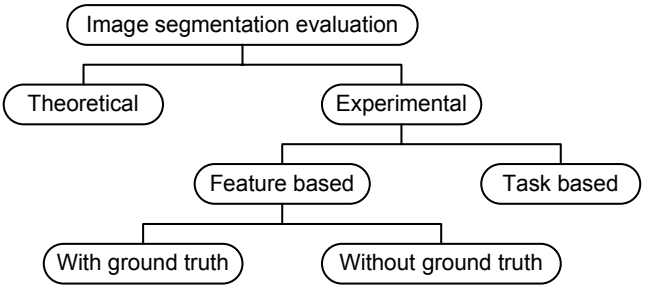


Fig. 1. Approaches for image segmentation evaluation [2]

this general purpose segmentation evaluation is quite ambiguous and may differ between different persons or applications. Recently figure ground segmentation evaluation [4] has been proposed to solve this issue, by considering only the most salient foreground object for the evaluation. In contrast to the general purpose segmentation evaluation the ground truth is well defined, which makes the evaluation more objective. Figure 2 illustrates the ground truth of both evaluation approaches for an example.

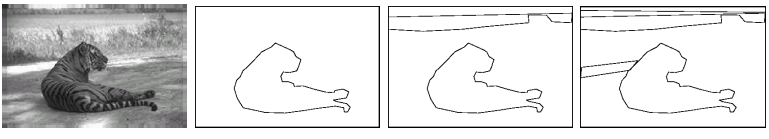


Fig. 2. Figure ground vs. general purpose segmentation [4]. From left to right: Original image, well-defined ground truth for figure ground segmentation evaluation, and two samples of the ambiguous ground truth for general purpose segmentation evaluation.

This paper describes a framework for fully automatic image segmentation evaluation, based on the figure ground methodology proposed by Ge et al. [4]. The goal is to provide an effective and efficient way to evaluate image segmentation approaches that provides objective measures close to subjective assessment without the time consuming manual annotation of suitable ground truth. A set of recent approaches for automatic image segmentation is compared against each other using the proposed framework. Both manually annotated and automatically generated ground truth are considered. Furthermore, different evaluation measures are used.

The structure of the paper is the following: section 2 describes the proposed evaluation framework focusing on the ground truth generation and the figure ground segmentation. Section 3 summarizes the different approaches for automatic image segmentation that have been considered for the comparison. The experimental results are presented in section 4. Section 5 draws conclusions and discusses future work.

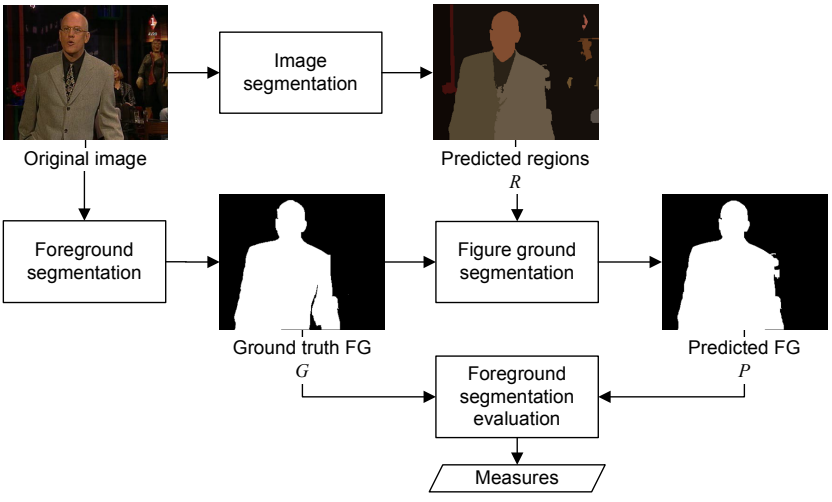


Fig. 3. Overview of the figure ground segmentation evaluation framework with individual steps and intermediate results

2 Evaluation Framework

Based on the goal to develop a fully automatic framework for figure ground image segmentation evaluation, the structure depicted in figure 3 has been derived. Based on videos obtained by a static camera, the ground truth foreground can either be annotated manually or extracted automatically using foreground segmentation techniques. The video frames are segmented into disjoint image regions using different image segmentation methods. The figure ground segmentation generates a predicted foreground mask P based on the predicted image regions R and the ground truth foreground mask G. Both foreground masks are given to the foreground segmentation evaluation, which measures the similarity or dissimilarity between them.

2.1 Foreground Segmentation

The pixel-wise annotation of foreground objects is a very time consuming process. Therefore, the idea is to use automatically generated ground truth for the evaluation, given that it provides qualitative results similar to that of manually annotated ground truth. For videos there are different approaches to segment each frame into foreground and background regions. While motion segmentation can be used for moving cameras, background subtraction is a suitable way to obtain foreground objects for static cameras.

In the work of Karaman et al. [5], selected state of the art background subtraction methods have been compared and it has been proven that most of the proposed systems in literature are developed for videos with specific environmental conditions. Furthermore, a method which combines the Gaussian colour

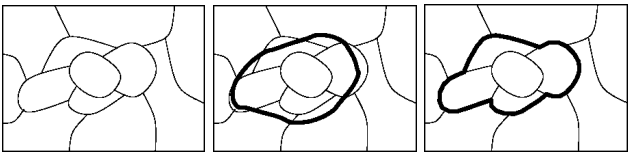


Fig. 4. Illustration of the figure ground segmentation [4]. From left to right: predicted image regions R, ground truth foreground G and predicted foreground P.

model (GCM) [6] and temporal information is proposed. The GCM is based on the measurement of object reflection and the description of colour invariants.

The foreground segmentation system consists of a classification and post processing stage. Multiple frames without moving foreground objects are used to train a pixel-wise background model by calculating the mean and standard deviation. Based on this model and the actual frame, channel wise difference images are computed and binarized using unimodal thresholding. The individual channel masks are combined by an OR operator into the preliminary foreground mask. The union of the final foreground mask of the last frame and the motion mask of the two previous frames is taken to define a region of interest (ROI). This mask is combined with the preliminary foreground mask using an AND operator to eliminate foreground pixels outside of the ROI. Morphological filtering (median, opening, closing) is applied to remove residual noise and obtain the final foreground mask.

2.2 Figure Ground Segmentation

Since most of the automatic image segmentation methods segment an image into a set of disjoint regions $R_1, R_2, \ldots, R_N$ with $R_i \cap R_j = \emptyset$ for $i \neq j$ and $\cup_{n=1}^{N} R_n = R$, the second strategy from Ge et al. [4] was adopted for the figure ground evaluation. The goal of this strategy is to find a subset of all the regions that corresponds to the ground truth foreground object. Figure 4 illustrates the idea.

The best matching subset of the regions is found by evaluating the overlap between each individual region R_i and the ground truth region G and merging matching regions into the predicted foreground region P, i.e.

$$P = \bigcup_{\forall i: c(R_i, G)=1} R_i \tag{1}$$

The original matching criteria by Ge et al. [4] is modified in order to merge only regions with a certain relative size (1%) and a minimum overlap (50%) with the ground truth regions, i.e.

$$c(R_i, G) = \begin{cases} 1 \text{ if } a_{R_i} > 0.5 \wedge a_{G_i} > 0.01 \\ 0 \end{cases} \tag{2}$$

with

$$a_{R_i} = \frac{|R_i \cap G|}{|R_i|}, \; a_{G_i} = \frac{|R_i \cap G|}{|G|} \tag{3}$$

By using the ground truth mask to identify the regions which are merged into the predicted mask, the figure ground evaluation provides an upper bound performance of the image segmentation assuming the use of a postprocessing step that provides the merged foreground region. This upper bound performance may not be achieved in real applications, where the ground truth is not available. Nevertheless, it provides a measure of how well an approach preserves the boundaries and regions that constitute the most salient object.

2.3 Evaluation Criteria

Based on the predicted foreground mask P from the figure ground segmentation and the ground truth foreground mask G obtained from manual annotation or automatic generation, evaluation methodologies for foreground segmentation can be used to measure the performance of the image segmentation. Beside the accuracy used by Ge et al. [4] measures from the widely used PETS metrics [7] were adopted to enable a more comprehensive comparison of the segmentation approaches.

Overall regions (OR). This is the number of disjoint regions returned by the image segmentation. A large number of regions is usually an indicator for over-segmentation, while a small number of regions may indicate under-segmentation.

Merged regions (MR). This is the number of regions merged together into the predicted foreground region, during the figure ground evaluation. This is related to the number of overall regions and depending on the ground truth.

Accuracy (ACC). The accuracy ACC used by Ge et al. [4] belongs to the category of pixel wise evaluation measures and describes the ratio between the overlapping area of the regions and the overall area of both regions.

$$ACC = \frac{|G \cap P|}{|G \cup P|} \tag{4}$$

The higher the accuracy the better is the segmentation quality.

Negative rate metric (NRM). The negative rate metric NRM [7] is based on the pixel-wise mismatches between the ground truth and prediction. It combines the false negative rate NR_{FN} and the false positive rate NR_{FP} into a single measure which is given as

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \tag{5}$$

with

$$NR_{FN} = \frac{N_{FN}}{N_{TP} + N_{FN}},\ NR_{FP} = \frac{N_{FP}}{N_{FP} + N_{TN}} \tag{6}$$

where N_{FN} and N_{FP} denote the number of false negative and false positive pixels, respectively. N_{TN} and N_{TP} are the number of true negatives and true positives. In contrast to the accuracy the segmentation quality is better for lower NRM.

Misclassification penalty metric (MPM). The misclassification penalty metric MPM [7] as shown in equation 7 evaluates the prediction against the ground truth on an object-by-object basis. Misclassified pixels are penalized by their distances from the ground truth object's border.

$$MPM = \frac{MP_{FN} + MP_{FP}}{2} \tag{7}$$

with

$$MP_{FN} = \frac{\sum_{i=1}^{N_{FN}} d_{FN}^{i}}{D},\ MP_{FP} = \frac{\sum_{j=1}^{N_{FP}} d_{FP}^{j}}{D} \tag{8}$$

where d_{FN}^{i} and d_{FP}^{j} are the distances of the i^{th} false negative and j^{th} false positive pixel from the contour of the ground truth segmentation. The normalization factor D is the sum over all the pixel-to-contour distances of the ground truth object. If an algorithm has a low MPM score, it is good at identifying an object's boundary.

3 Segmentation Methods

In contrast to Ge et al. [4] more recent image segmentation approaches have been considered for the comparison. While Ge et al. used a single variation of approaches published between 1999 and 2005, multiple variations of approaches proposed between 2005 and 2007 are considered here. The different methods are shortly described below.

3.1 Mean Shift Segmentation (MS)

This segmentation algorithm is an optimization of the mean shift algorithm for color segmentation in image sequences [8]. The mean shift is a technique for analysis of features spaces and has been proposed in 1997 by Comaniciu and Meer [9].

When used for color image segmentation, the image data is mapped into the features space, resulting in a cluster pattern, where each cluster corresponds to a significant feature in the image domain, namely a dominant color. The mean shift procedure locates these clusters by applying a search window in the feature space, which shifts towards the cluster's center. This procedure is repeated until all significant clusters have been extracted.

A drawback of the mean shift technique is its computational cost, especially when applied to image sequences. For speed up of processing of image sequences, we exploit the fact that subsequent frames are similar in terms of color content and propagate the cluster centers from frame to frame, with special treatment for case of quickly varying content. Using the cluster centers of previous frames as initial estimates significantly reduces the number of iteration until the algorithm converges. Another optimization used for reducing the runtime of this algorithms, is the moderate quantization of input data, so that there are fewer

feature vectors to be processed in both image domain and features space. Further optimizations enhance the temporal stability by removing border pixels and small regions.

3.2 Region Based Automatic Segmentation (RBAS)

The RBAS method [10] integrates several extensions to the well-known Recursive Shortest Spanning Tree (RSST) algorithm [11] corresponding to an extended color model and spatial configuration of regions and their geometric properties.

The original RSST algorithm starts by mapping the input image into a weighted graph [11], where the regions (initially pixels) form the nodes of the graph and the links between neighboring regions represent the merging cost, computed according a selected homogeneity criterion. At each iteration two regions connected by the least cost link are merged.

In the extended approach, the initial partition is obtained with the original color based homogeneity criterion [11]. In consecutive stages several additional homogeneity measures are utilized. Two additional colour homogeneity measures and four geometric features are used. When the total number of regions falls below a pre-defined value, each region's average color is replaced by the *Adaptive Distribution of Color Shades* (ADCS) representation. In this model each region contains a list of pairs of color/population. In addition to the above, over-segmentation of objects with slow gradual color changes is prevented by adopting the *Boundary Melting* approach which favors merging of regions with low magnitude of color gradient along their common boundary. Further evidence for merging is provided by syntactic features such as global and local shape complexity (boundary jaggedness), region adjacency, and total inclusion of a small region inside another.

3.3 Modified Recursive Shortest Spanning Tree (MRSST)

This is an extension of the RBAS approach [12]. There are two main differences between both methods. The first lies in the way the evidence provided by different features (colour and geometric properties) is fused. The second difference is the stopping criterion used in MRSST aimed at producing partitions containing the most salient objects present in the scene.

this approach, the merging order is based on evidence provided by all features fused using an integration framework based on *Dempster-Shafer* (DS) theory [13] which takes into account the reliability of different sources of information as well as the fact that certain measurements may not be precise (doubtful) or even "unknown" in some cases.

During the merging process all merges are recorded in the binary partition tree (BPT). A single partition that reflects meaningful image content is selected based on the evolution of the merging cost accumulated during the overall merging process. The accumulated merging cost measure measuring the total cost of all mergings performed to produce a certain number of regions is computed at each iteration. A suitable stopping threshold point is located by detecting a corner of

the merging cost curve. Once the threshold point is found, the final partition is obtained by deactivating the corresponding number of nodes from the BPT.

3.4 Spatio Temporal Video Segmentation (SEG2DT)

This method extends image segmentation to the spatiotemporal domain. Spatiotemporal methods usually take as input a single 3D pixel volume, generating an important computational cost and large memory load [14]. The framework proposed aims to reduce both problems, while still approaching the temporal coherency of 3D methods [15]. Instead of processing the entire video volume with computationally intensive clustering or optimization algorithms, the segmentation process is decomposed into several stages, using low-complexity graph merging procedures.

The method is based on a causal 2D+T scheme, i.e. the segmentation at a given time depends only of the previously segmented frames. Segmentation is initialized with an efficient graph merging algorithm that partitions the first frame into homogeneous components [16]. This step sets approximately the level of spatial details for the whole sequence. Then the graph components are propagated to the next frame. To this aim, an over-segmented partition is created in the new image, where possible region boundaries are excluded from merging using a contour map. The new components are merged with the previous segmentation according to a criterion that compares simultaneously their local and global properties. Finally, the segmentation of the new frame is completed by grouping the remaining small components until they reach a minimum size.

4 Experiments

Extensive experiments have been carried out in order to evaluate several aspects of the proposed framework and to compare the different segmentation approaches. The methods have been compared both subjectively and objectively using the three different measures described in section 2.3. Manually annotated as well as automatically generated ground truth have been used and compared to each other. For each of the approaches 7 different variations were used, ranging from under- to over-segmentation.

4.1 Database

In order to evaluate the stability of the segmentations for small changes in the visual content, a database consisting of 6 videos with 375 frames has been used. For each of the frames both manual and automatic ground truth masks were created. Figure 5 shows a sample for each of the videos along with the corresponding manual and automatic ground truth mask. It can be seen that the videos feature quite different visual characteristics, including different sizes of foreground objects, cluttered and uniform background and different colors and contrasts.

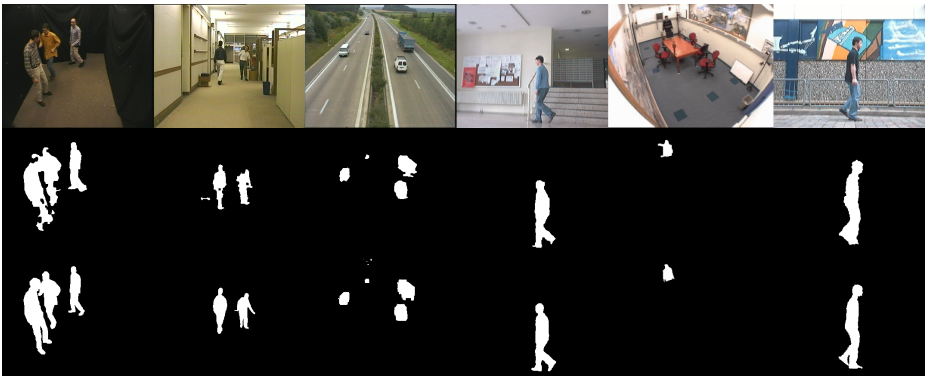


Fig. 5. Sample frames of the videos in the database along with automatically generated and manually annotated foreground masks

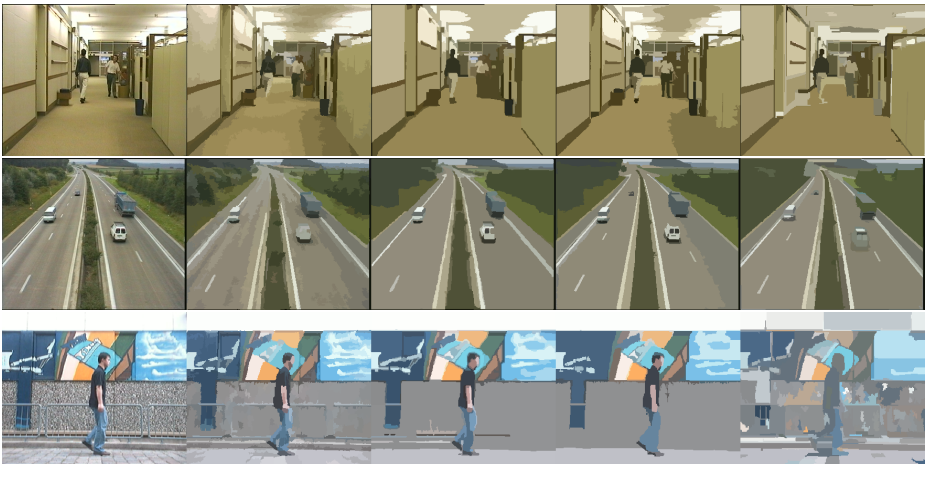
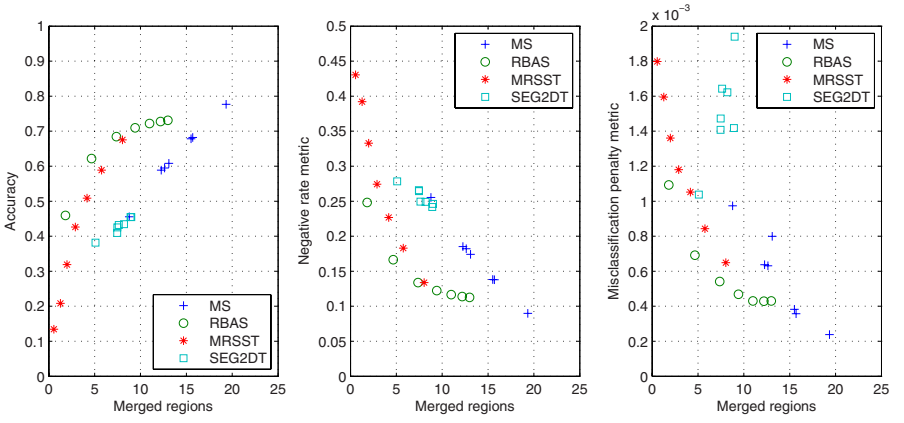


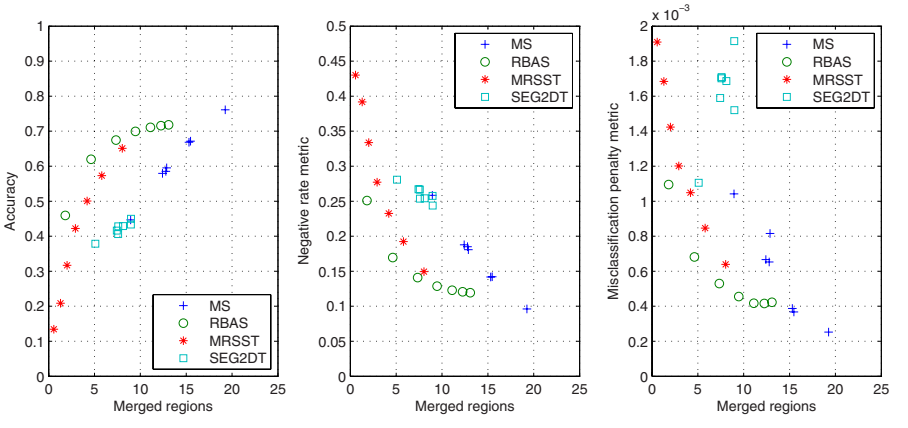
Fig. 6. Subjective comparison of the different image segmentation approaches (best variation) for the different videos. From left to right: Original image, MS, RBAS, MRSST, SEG2DT.

4.2 Results

Figure 6 presents a subjective comparison between the different image segmentation approaches by showing a representative sample for each of the videos. Beside the original video, the segmentation result for the best variation of each approach (MS, RBAS, MRSST, SEG2DT) is shown. It can be observed that the number of regions and the quality of the segmentation differ quite noticeably. MS, MRSST and RBAS provide comparable results, although the MS segments the image into much more regions than the other approaches. For some videos the SEG2DT approach creates some artificial regions.



(a) Manual ground truth



(b) Automatic ground truth

Fig. 7. Various measures vs. number of merged regions for different approaches/variations based on different ground truth

Since the number of regions (OR, MR) differ quite a lot, it is difficult to compare the performance measures with each other. Therefore, figure 7(a) plots the different measures (ACC, NRM, MPM) vs. the number of regions (MR) to show the tradeoff between under-/over- segmentation and performance. This type of plot allows to visually compare the different approaches/variations. As expected a better performance is achieved with a higher number of regions. Nevertheless, it is evident that different methods achieve a comparable performance with less regions than others. For approximately 10 merged regions the ranking of the different approaches is RBAS, MRSST, MS, SEG2DT. Furthermore, it can be perceived that the different measures (ACC, NRM, MPM) provide comparable results.

The next experiments are based on the automatically generated ground truth. Figure 7(b) provides the results for these experiments in the same way as figure

7(a) for the manual ground truth. It can be seen already by visual inspection that the results are quite similar which leads to the conclusion that even noticeable differences between the manual and the automatic ground truth do not influence the figure ground evaluation very much. More specifically the relative absolute differences between the manual and the automatic ground truth for the different measures vary between 1% and 3%.

5 Conclusions

This paper describes a novel framework for image segmentation evaluation that supports automatic evaluation and comparison of image segmentation approaches. The experiments indicate that figure ground evaluation offers an effective way for image segmentation evaluation with objective results comparable to the subjective assessment. Furthermore, it provides an efficient and unambiguous way to annotate ground truth manually. It is also shown that automatic ground truth generation is feasible for static camera videos by adopting background subtraction techniques. Although the ground truth differs noticeably, the evaluation results are only slightly affected.

While the measure vs. number of region plot provides a visual way for comparing different approaches/variations, it is not suitable for extracting a ranking automatically. Therefore, a relation between number of regions and the different performance measures needs to be defined or other measures considering the number of regions have to be developed. For more general results a larger database with a wide range of characteristics (colors, texture, objects, enviroments) is required.

Acknowledgements

This research was supported by the European Commission under contract FP6-027026-K-SPACE.

References

1. Cheng, H., Jiang, X.H., Sun, Y., Wang, J.L.: Color image segmentation: Advances & prospects. Pattern Recognition 34 (2001)
2. Jiang, X., Marti, C., Irniger, C., Bunke, H.: Distance measures for image segmentation evaluation. EURASIP Journal on Applied Signal Processing 1 (2006)
3. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: International Conference on Computer Vision (2001)
4. Ge, F., Wang, S., Liu, T.: Image segmentation evaluation from the perspective of salient object extraction. In: Conference on Computer Vision and Pattern Recognition (2006)
5. Karaman, M., Goldmann, L., Sikora, T.: A new segmentation approach using gaussian color model and temporal information. In: Visual Communications and Image Processing (2006)

6. Geusebroek, J.M., van den Boomgaard, R., Smeulders, A.W.M., Geerts, H.: Color invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence (2001)
7. Young, D.P., Ferryman, J.M.: Pets metrics: On-line performance evaluation service. In: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS) (2005)
8. Bailer, W., Schallauer, P., Haraldson, H.B., Rehatschek, H.: Optimized mean shift algorithm for color segmentation in image sequences. In: SPIE Image and Video Communications and Processing (2005)
9. Comaniciu, D., Meer, P.: Robust analysis of feature spaces: color image segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR 1997), p. 750 (1997)
10. Adamek, T., O'Connor, N.E., Murphy, N.: Region-based segmentation of images using syntactic visual features. In: International Workshop on Image Analysis for Multimedia Interactive Services (2005)
11. Alatan, A., Onural, L., Wollborn, M., Mech, R., Tuncel, E., Sikora, T.: Image sequence analysis for emerging interactive multimedia services - the European COST 211 Framework. IEEE Transactions on Circuits and Systems for Video Technology 8 (1998)
12. Adamek, T., O'Connor, N.: Using dempster-shafer theory to fuse multiple information sources in region-based segmentation. In: International Conference on Image Processing (2007)
13. Smets, P., Mamdami, E., Dubois, D., Prade, H.: Non-Standard Logics for Automated Reasoning. Academic Press, Harcourt Brace Jovanovich Publisher (1988) ISBN 0126495203
14. DeMenthon, D., Doermann, D.: Video retrieval using spatio-temporal descriptors. In: ACM International Conference on Multimedia (2003)
15. Galmar, E., Huet, B.: Graph-based spatio-temporal region extraction. In: International Conference for Image Analysis and Recognition (2006)
16. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59 (2004)