A Survey on Outdoor Scene Image Segmentation

Elizabeth Sama Sam
Department of Computer Science and Engineering
Karunya University

A. Kethsy Prabhavathy
Department of Computer Science and Engineering
Karunya University

J. Devi Shree, PhD.
Department of Computer Science and Engineering
CIT, Coimbatore

ABSTRACT
Image segmentation is the process of partitioning an image into multiple parts, so that each part or each region corresponds to an object or area of interest that is more significant and easier to analyze. Several general-purpose algorithms and techniques have been developed for image segmentation. This paper describes the different segmentation techniques used to achieve outdoor scene image segmentation. Unlike other surveys that only describe and compare qualitatively different approaches, this survey deals with a real quantitative comparison of the F-measure.

Keywords
Image segmentation, structured object, unstructured object, superpixels.

1. INTRODUCTION
Digital Image Processing involves using a computer to change the nature of a digital image [13]. Image processing deals with changing the nature of an image to either develop its pictorial information for human interpretation or to make it more suitable for autonomous machine perception.

Image segmentation is one of the most vital precursors for image processing based applications and has an essential impact on the whole performance of the developed systems. Image segmentation is to divide an image into number of regions so that each region gives information about an object or area of interest [15]. The scene objects are classified as shown in Fig 1.

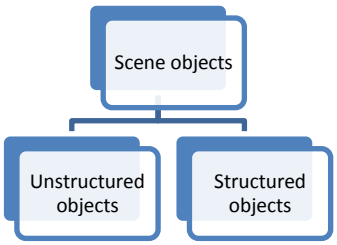

Fig 1: Classification of scene objects

The unstructured objects are the sky, roads, trees, grass etc. and the structured objects are the people, buildings, cars etc. In most of the images, the structured objects are the foreground objects composed of multiple parts and the unstructured objects are the background objects having nearly homogeneous surfaces. It is difficult to segment an outdoor scene image as it is composed of both the structured and unstructured objects. Segmenting the structured object is the ultimate challenge as it is poised with numerous parts with each part having different surface characteristics. Image segmentation can be performed using one of the following approaches:

1. Top down approach
2. Bottom up approach

Top down approach uses prior knowledge about an object such as its shape, color or texture to guide the segmentation. Top–down approach follows the theory that the image contains a particular object or can be categorized as a particular type of scene. Further tasks are performed to verify the existence of a hypothesized object. The complexity in this approach stems from the large unevenness in the shape and appearance of objects within a given class [2].

Bottom up approach, the image is first segmented into regions and then the image regions that correspond to a single object are identified. The pixels are grouped according to the grey level or texture uniformity of the image regions, as well as smoothness and continuity of bounding contours. The complexity in this approach is that an object may be segmented into numerous regions some of which may occlude with the background.

2. IMAGE SEGMENTATION TECHNIQUES
The different types of image segmentation techniques for outdoor scenes can be classified as follows:

- Graph based approaches
- Region based image segmentation
- Multiclass image segmentation
- Boundary detection approach
- Image segmentation based on perceptual organization
- Hybrid approach

2.1 Graph based approaches
The graph based image segmentation approach defines the boundaries between regions by measuring the dissimilarity between the neighboring pixels. Each pixel is equivalent to a node in the graph. Weights on each edge determine the dissimilarity between pixels. Shi and Malik [4] proposed the normalized cut criterion that removes the trivial solutions of cutting small sets of isolated nodes in the graph. Ncut method organizes nodes into groups so that within the group the similarity is high and/or between the groups the similarity is low. This method is relatively robust and can be recursively applied to get more than two clusters. Each time the subgraph is partitioned that has the maximum number of nodes (random selection for tie breaking). When the bound on the number of clusters is reached or Ncut value exceeds some threshold, the process terminates. The normalized cut algorithm can be explained as follows: First the image is represented as a weighted graph G=(V,E) and then compute the weight of each edge. Also summarize the values D and W where D is a N*N diagonal matrix and W is a N*N symmetrical matrix. Next solve the $(D - W)y = \lambda Dy$ for the eigenvector with the second smallest eigenvalue. Finally the graph is bipartitioned using the entries of the eigenvector. However the normalized cut criterion is an NP-hard computational problem. Shi and Malik developed methods for computing the minimum normalized cut, but even this could not tackle the error in these estimates.

The computations are hard to compute and take much time to complete and also this approach works well only for relatively small images. Other problems are the high storage requirement and this approach is bias towards partitioning into equal segments. Fig 2 illustrates the image segmentation using Normalized cut approach.

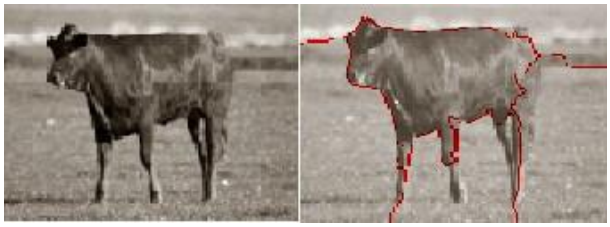

Fig 2: Image Segmentation using Normalized Cut approach: (a) Input Image (b) Segmentation of the image.

Felzenszwalb and Huttenlocher [3] proposed an efficient graph-based generic image segmentation algorithm. Without first performing a filtering step, this method works directly on the data points in feature space and uses a deviation on single linkage clustering. The success of this method is principally due to the adaptive thresholding. A minimum spanning tree of the data points is first generated (using Kruskal's algorithm) to perform traditional single linkage clustering, from which any edges with length greater than a given hard threshold are detached. The joined components become the clusters in the segmentation. More specifically, let G = (V,E) be a (fully connected) graph, with m edges and n vertices. Each vertex is represented as a pixel x in the feature space. The final segmentation will be a cluster of data points. This method has high inference and can incorporate data likelihoods and priors.

2.2 Region based image segmentation

Region-based techniques make use of common patterns in intensity values within a cluster of neighboring pixels. The cluster is referred to as the region, and the goal of the region based image segmentation algorithm is to group regions according to their anatomical or functional roles**.** Micusik *et al.* [6] segmented the semantic street scenes into coherent regions simultaneously categorizing each region as one of the predefined categories representing an object or background class. A small blob based superpixels is used for segmentation and it exploits a visual vocabulary tree as an image representation. The goal of the semantic labeling of street scenes is to automatically annotating different regions by labels of commonly encountered object and object categories. In this technique, instead of modeling co-occurences of class labels here the spatial co-occurrences between visual words of neighboring superpixels are evaluated. Pantofaru et al. [8] proposed a method suggesting that integrating the information from multiple image segmentations can offer a more robust basis for object class recognition and object segmentation than for single image segmentation. It relies on two basic principles First the same segmentation region containing group of pixels in multiple segmentations should be consistently classified and the pixel groups can be efficiently classified using multiple image segmentation. Fig 3 shows the object recognition using multiple image segmentation.

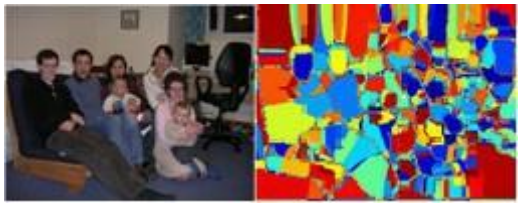

Fig 3: An example of Intersections of Regions (Iofrs) adapted from [8]

Gould et al. [9] uses the semantic region labels (e.g. road, sky, building etc.) and coherent geometric placement (orientation and location with respect to horizon) to decompose the scene into regions. An image I is decomposed into an unknown number (K) of geometrically and semantically consistent regions by iteratively optimizing an energy function that measures the quality at hand. Classifying large regions rather than individual pixels can compute more robust features and reduce inference complexity. Multiple over-segmentations allow refining region boundaries and making large moves in energy space. The context can be easily captured using a pair wise term between adjacent regions. This approach provides a foundation to integrate many other vision tasks (e.g. 3D reconstruction and object detection). The only disadvantage is its reliance on large amount of hand labeled data.

2.3 Multiclass image segmentation

Multi-class image segmentation uses one of a number of classes (e.g., road, sky, water, etc) for labeling every pixel in an image. Many state-of-the-art methods first over-segment the image into superpixels (or small coherent regions) and classify each region since classifying every pixel can be computationally expensive. Shotton et al. [5] described the need to label each pixel in the image with one of a set of predefined object class labels. In this approach, Shotton assigned a class label to a pixel based on a joint appearance, shape and context model. The aim of this approach is, given an image, the system should be capable of automatically partitioning it into semantically meaningful regions each labeled with a specific object class. For this a discriminative model for object class is learned incorporating texture, layout and context information efficiently. The learned model is then used for automatic visual understanding and semantic segmentation of images. This technique can model very long range contextual relationship extending over half the size of the image. The high level description of this approach is as follows:

- Learn classifier based on relative texture locations for each class.
- Refine classification with Conditional Random Field (CRF).
- Improve classification with additional pixel information.

The TextonBoost system has been applied in several exciting new areas like AutoCollage, Semantic Photo Synthesis, Interactive Semantic Segmentation and Interactive Image Editing. The primary limitation is the performance of the texture-layout potentials learned by boosting. The classification cost grows sub-linearly with the number of classes due to the use of Joint Boosting although training time increases quadratically. When moving to more classes, the simple ontological model is used where each pixel is assigned only one class label. This can lead to semantic confusions. The detection of objects at smaller scales is sometimes poor with the current system. In [7], Shotton et al. proposed the use of semantic texton forests for fast classification. The aim of

this approach is to perform simultaneous segmentation and recognition of objects in the images. The two applications of STFs are

- Image categorization
- Semantic segmentation

According to a learned binary function of the feature vector, the decision tree recursively branches left or right until a leaf node l is reached. The random decision tree training can be done as follows:

- Take a random subset of training data.
- Generate the random features f.
- Generate the random threshold t.
- Split the data into left I_l and right I_r.
- Repeat for each side.

The training time speeds up and also over-fitting reduces when decision trees trained on small random subsets of data are used. As only a small portion of the tree is traversed for each data point, the trees are fast to learn and evaluate. The total computation time using STF is 605ms whereas the same using TextonBoost is 6000ms. The disadvantages are low resolution classification, the segmentation forest operates at patches and so the test time inference is dependent on amount of training. It must iterate through all the trees in the forest at test time. High supervision is needed for segmenting forests. Gould et al. [10] proposed a superpixel-based conditional random field to learn the relative location offsets of categories. Unlike object recognition methods that intend to find a particular object, multi-class image segmentation methods are intended at concurrent multi-class object recognition and attempt to classify all pixels in an image. For each superpixel region, this method first extracts appearance (color and texture), geometry and location features. Then boosted classifiers are learnt over these features for each region class. Finally, a CRF or logistic model is learned using the output of the boosted classifiers as features. It does not distinguish between objects at different scales. Fig 4 shows the example results of simultaneous object class recognition and segmentation algorithm.

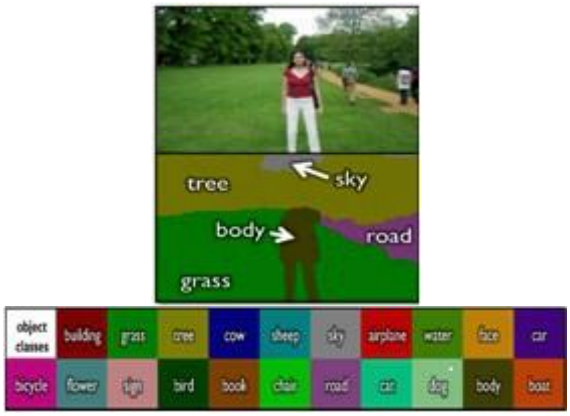

Fig 4: Example results of simultaneous Object Class Recognition and Segmentation algorithm. Adapted from [5]

2.4 Boundary detection approaches

A boundary is a contour in the image plane that represents dissimilar pixels between the neighboring objects. Dollar et al. [14] designed their boundary detection algorithm based on a large number of generic features calculated over a large image patch. In this algorithm, the context information is provided by a large aperture. The algorithm selects and combines a set of features out of a pool in the learning stage with tens of thousands of generic, efficient Haar wavelets in order to learn a discriminative model. True probabilities are output in this method whereas other edge detection methods either output a soft value based on edge strength or a binary value (which is not a true probability). When making a decision, this method combines low-level, mid-level and context information across different scales. Learning edge probability can be done by the classification framework used which is an extended Probabilistic Boosting Tree (PBT) that combines the bootstrapping procedure directly into the tree formation while properly maintaining priors. This approach uses tens of thousands of very simple features considered over a much larger region. The main advantage of such an approach is that the human effort is minimized; as an alternative the work is shifted to the classification algorithm. This approach is highly adaptive and scalable. Hoeim et al. [12] estimated occlusion boundaries based on both 2-D perceptual cues and 3-D cues such as surface orientation and depth estimates. The boundaries and depth ordering of prominent objects are recovered in sufficient detail to provide an accurate sense of depth. This strategy is to simultaneously reason about the regions and boundaries in the image and the 3D surfaces of the scene using learned model. The learned model identifies boundaries based on a wide variety of features: color, position, and alignment of regions; strength and length of boundaries; 3D surface orientation estimates and depth estimates. Develop the segmentation gradually by iteratively computing cues over the current segmentation and using them with the learned models to combine regions that are likely to be part of the same object. Each iteration consists of three steps based on the image and the current segmentation: (1) compute cues; (2) assign confidences to boundaries and regions; and (3) remove weak boundaries, forming larger regions for the next segmentation. Fig 5 shows the example result of supervised learning of edges and object boundaries.

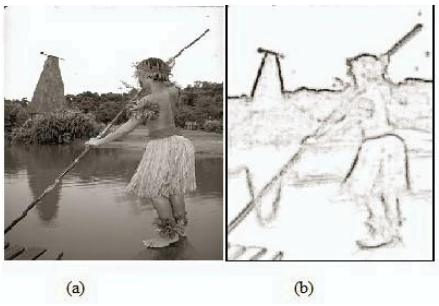

Fig 5: Example result of Supervised Learning of Edges and Object Boundaries. a) Gray scale image from Berkeley Dataset. (b) BEL result. Adapted from [14]

2.5 Image Segmentation based on perceptual organization

Perceptual organization refers to a basic capability of the human visual system to obtain relevant groupings and structures from an image without having prior knowledge of the contents of the image. The Gestalt psychologists summarized some underlying principles (e.g., proximity, similarity, continuity, symmetry, etc.) that lead to human perceptual grouping. Maire et al. [11] suggests a state-of-the-art solution for the problems related to finding contours (segmentation curves), and finding junction (points joined by multiple contours). The contours are found by combining the local and global features. The local cues are combined in a

multi-scale oriented signal including brightness, color and texture gradients. The global information is considered to be in the first 9 generalized eigenvectors, from which a signal is extracted with Gaussian directional derivatives at multiple orientations. The local and global information are then linearly combined, resulting in a globalized probability of boundary, which claims the top spot in the standard Berkeley segmentation benchmark. The procedure for contour detection can be briefly stated as:

- Collect Data Set of Human segmented images.
- Learn the local boundary model for combining brightness, color and texture.
- Global framework to capture closure, continuity.
- Detect and localize junctions.
- Integrate the low, mid and high-level information for grouping and figure-ground segmentation.

Chang et al. [1] explores detecting object boundaries in outdoor scene images based on some general properties of the real world objects such as perceptual organization without depending on a priori knowledge of the object. It is well accepted that segmentation and recognition should not be separated and should be treated as an interleaving procedure. This method basically follows this scheme and requires identifying some background objects as a starting point. Compared to the large number of structured object classes, there are only a few common background objects in outdoor scenes. These background objects have low visual variety and hence can be reliably recognized. After background objects are identified, it is roughly know where the structured objects are and delimit perceptual organization in certain areas of an image. For many objects with polygonal shapes, such as the major object classes appearing in the streets (e.g., buildings, vehicles, signs, people, etc.) and many other objects, this method can piece the whole object or the main portions of the objects together without requiring recognition of the individual object parts. In other words, for these object classes, this method provides a way to separate segmentation and recognition. Fig 6 illustrates the result of implementation of outdoor scene image segmentation based on background recognition and perceptual organization.

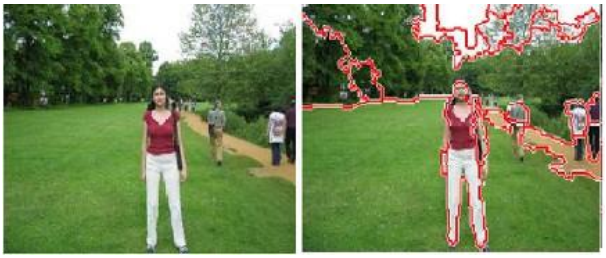

Fig 6: Example result of outdoor scene image segmentation using background recognition and perceptual organization. Adapted from [1].

2.6 Hybrid approach

Hybrid techniques combine methods of both the top down and bottom up segmentation approaches. Borestein et al.[2] integrates the top down and bottom up segmentation techniques and constructs a classification map K(x,y). In the top-down approach, the requirement is to make K as close as possible to the initial top down classification map T. The bottom up constraint requires K to match the image structure, so that pixels within the homogeneous image regions, as defined by the bottom up process are likely be segmented together into either the figure or background part of the image. The advantage of this approach is that it provides a reliable confidence map indicating the regions of residual ambiguity with no additional computation cost. The problem is that the top down and bottom up approaches may conflict with each other.

3. COMPARISON USING F-MEASURE

The precision-recall framework is recommended by BSDS for the boundary-based measurement. The tradeoff between accuracy and noise is captured by the precision-recall curve. Precision is the fraction of detections that are true boundaries, whereas recall is the fraction of true boundaries that are detected. Thus, precision is the probability that the segmentation algorithm's signal is valid, and recall is the probability that the ground truth data is detected. These two quantities can be combined in a single quality measure, i.e., F-measure, defined as the weighted harmonic mean of precision and recall. The comparison of F-measure of various methods is shown in Table 1 and its graphical representation is as depicted in Fig. 7.

Table 1.F-measure of various techniques

Methods	F-measure
Chang et al.[1]	0.72
Borenstein et al.[2]	0.68
Shotton et al.[7]	0.67
P.Dollar et al. [14]	0.64
Gould et al.[10]	0.66
Maire et al.[11]	0.65

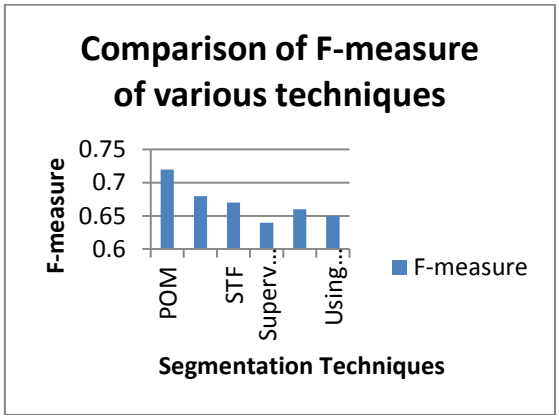

Fig 7: Graphical representation of F-measure of various segmentation techniques

4. CONCLUSION

To summarize, a comprehensive survey highlighting different outdoor scene image segmentation techniques have been presented. The techniques discussed were graph based approaches, region based image segmentation, multiclass image segmentation, boundary detection approach, image segmentation based on perceptual organization and hybrid approach. The various techniques have been compared using the F-measure which is defined as the weighted harmonic mean of precision and recall. The future work can include improving the outdoor scene image segmentation using

background recognition and perceptual organization by solving the problem caused by strong reflection and over-segmentation.

.

5. REFERENCES

[1] C. Cheng, A. Koschan, D. L. Page, and M. A. Abidi, "Outdoor scene image segmentation based on background recognition and perceptual organization," in *Proc. IEEE Trans*, vol.21, no.3,pp. 1007–1019, March 2012.

[2] E. Borenstein and E. Sharon, "Combining top-down and bottom-up segmentation," in *Proc. IEEEWorkshop Perceptual Org. Comput. Vis.,CVPR*, 2004, pp. 46–53.

[3] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep.2004.

[4] J. B. Shi and J. Malik, "Normalized cuts and image segmentation", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[5] 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. Vis.*,vol. 81, no. 1, pp. 2–23, Jan. 2009.

[6] B. Micusik and J. Kosecka, "Semantic segmentation of street scenes by superpixel co-occurrence and 3-D geometry," in *Proc. IEEE Workshop VOEC*, 2009.

[7] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE CVPR*, 2008, pp. 1–8.

[8] C. Pantofaru, C. Schmid, and M. Hebert, "Object recognition by integrating multiple image segmentations," in *Proc. ECCV*, 2008, pp.481–494.

[9] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. IEEE ICCV*, 2009, pp. 1–8.

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

[11] M.Maire, P. Arbelaez, C. C. Fowlkes, and J.Malik, "Using contours to detect and localize junctions in natural images," in *Proc. IEEE CVPR*, 2008, pp. 1–8.

[12] D. Hoiem, A. N. Stein, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from a single image," in *Proc. IEEE ICCV*, 2007, pp. 1–8.

[13] Solomon C.J. and Breckon T.P., *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*, Wiley-Blackwell,2010.

[14] P. Dollar, Z. W. Tu, and S. Belongie, "Supervised learning of edges and object boundaries," in *Proc. IEEE CVPR*, 2006, vol. 2, pp. 1964–1971.

[15] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *Proc. BMVC*, 2007.