

D - Clutter: Building object model library from unsupervised segmentation of cluttered scenes

Gowri Somanath¹, Rohith MV¹, Dmetris Metaxas², Chandra Kambhamettu¹
¹VIMS Lab, Department of Computer Science, University of Delaware, USA.
²CBIM Center, Department of Computer Science, Rutgers University, USA.
{somanath,rohithmv,chandra}@cis.udel.edu, dnm@cs.rutgers.edu

Abstract

Autonomous systems which learn and utilize a limited visual vocabulary have wide spread applications. Enabling such systems to segment a set of cluttered scenes into objects is a challenging vision problem owing to the non-homogeneous texture of objects and the random configurations of multiple objects in each scene. We present a solution to the following question: given a collection of images where each object appears in one or more images and multiple objects occur in each image, how best can we extract the boundaries of the different objects? The algorithm is presented with a set of stereo images, with one stereo pair per scene. The novelty of our work is the use of both color/texture and structure to refine previously determined object boundaries to achieve segmentation consistent with each of the input scenes presented. The algorithm populates an object library, which consists of a 3D model per object. Since an object is characterized both by texture and structure, for most purposes this representation is both complete and concise.

1. Introduction

Design of algorithms to autonomously segment objects (class/category) given a set of example images has been the focus of many research works recently [1,2,3,4,5]. Human perception of an object is based on prior knowledge of the target object or surrounding [6]. Hence, it is imperative to define ‘object’ before we proceed with designing systems to detect/learn objects from visual observations. For the purpose of this work, an object is a set of 3D points appearing in a consistent configuration over all examples. The problem we attack is as follows: given a collection of images where each object appears in one or more images, how best can we extract the boundaries of the different objects? The fact that the objects do not have uniform color/texture and that they appear (possibly partially occluded) in cluttered scenes makes the problem challenging. We present a solution to the following problem: In an unsupervised setup



Figure 1: Illustration of example input scenes and expected output. Top: Input scenes. Bottom: Some expected models

(unknown boundaries/models), given a set of stereo images of cluttered scenes, (1) segment all objects seen (2) group multiple instances and (3) build 3D model of each object. The novelty of our work is that our algorithm learns from each new input and refines already known boundaries. The result is hence the net knowledge gained, represented as 3D models of the objects. We illustrate the idea through an example in Figure 1. Given the two scenes, some of the objects we would like to extract are shown. Notice that the model of the triangular object has the information obtained from both images (the black and green sides). Figure 2 shows two stereo inputs and the results we obtain using our algorithm. More results are shown in Figure 9.

Our motivation for using depth and 3D information (from stereo in our case; in general, multiple views may be used) has come from both segmentation and object representation phases. Human cognitive studies like those done by Palmer [25] have shown that depth perception play an important role in segmentation and grouping. From the systems perspective, an object can be characterized in essence by its texture and its structure/shape. Images provide texture/color information while 3D models retain both texture and shape/structure information, hence providing a more complete and concise representation. Matas et al.[7] indicate that as the object



(a) Input stereo pair 1



(b) Input stereo pair 2



(c) Some of the objects extracted. Left to right: Peanuts box, Nestle box, House model, Stuff toy, Minute Maid carton, Ice-cream cups box.

Figure 2: Results using input stereo pairs 1 & 2

library size increases, searching by one to one comparison is impractical. They discuss a sub-linear indexing scheme using trees formed from salient texture patches. Using 3D shape descriptors can further improve such indexing schemes. Intuitively, we would not like to compare a box with irregular or sphere shaped objects. Some results [12,23,24] show that knowing the 3D model aids object recognition. Creating CAD models or individual object models from multiple views for a large number of objects is impractical. Training a system with multiple views of each object would require many images. Hence a system to build 3D models of many objects simultaneously from a relatively smaller collection of images is desirable. We present an algorithm which builds a library of 3D models of objects (seen in one or more scenes) consistent with all the examples provided. The overview of our method is shown in Figure 3. For each new stereo input, we use structure from motion technique [8] using a few robust matches to obtain the camera matrices. Dense matches are obtained by interpolating the disparity obtained using SDM [9] based on image segmentation. Tentative object boundaries are obtained by simultaneous depth and color segmentation. The interleaved use of depth and color for segmentation/grouping is similar to theories proposed in cognitive sciences [25]. Each tentative object is compared to objects already in the library and a prediction is made by projecting the model onto the current scene. Object boundaries are refined using the model and current scene. Refined and new objects are added to the model library.

1.1. Background and related work

Object (class) segmentation and discovery algorithms are mostly based on Gestalt's principles of similarity, proximity and continuity. Previous works [1,2,5,6] have been driven by image segmentation i.e. Gestalt's principles applied to color/texture. We extend the notion to 3D, similar to theories proposed by Palmer [25]. Real world objects as illustrated in Figure 1 are not composed of single color/texture. Therefore, we use stereo images as our input. Depth and color together provide initial object boundaries. The use of depth for segmentation has been previously applied for foreground-background separation [10,11]. Stereo/multiple view images of individual objects have been used by [12, 23] to build models, but each input contains only the object being trained (placed in a known simple background). Other methods for object segmentation include [13] which uses Layered Pictorial Structures learnt from video sequences; [14] which performs segmentation based on key point features learnt a priori and [15] which extracts segments from stereo video sequences. Our work is aimed towards solving the problem for the case of cluttered scenes and in an unsupervised fashion. Russell et al. [1] presented an algorithm to discover object classes from a large image collection. It may appear that we can apply their approach to the problem we seek to solve. The results from applying their multiple segmentations algorithm [1] are shown in figure 6. We observe that relying on image (color/texture) segmentation alone to provide correct object boundary leads to less than satisfactory results. More recently, [5]

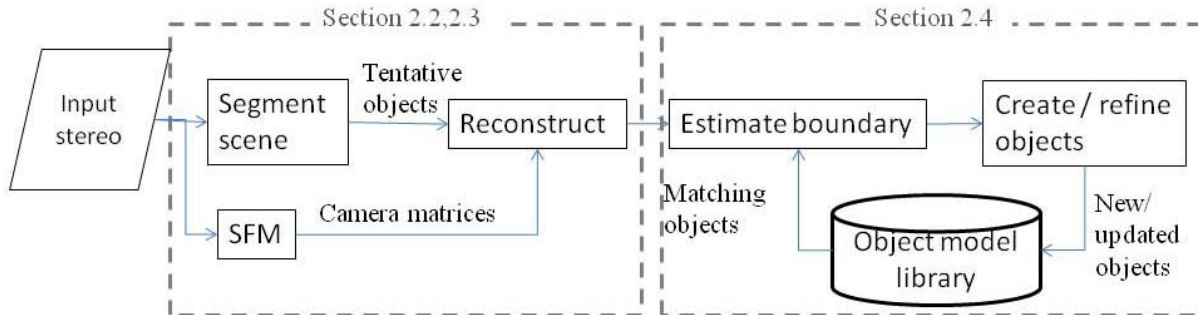


Figure 3: Overview of our algorithm

used motion cues to obtain boundary estimates, followed by matting to achieve whole object segmentation. In general, the previous works stated above perform segmentation on a per-image or per scene basis. In contrast we seek to achieve segmentation (object boundary) consistent with each of the input scenes presented to the algorithm.

2. The Algorithm

The algorithm is presented with a set of stereo image pairs. In general, each pair is taken with an arbitrary stereo configuration and may not be calibrated. The algorithm consumes each pair sequentially to update the 3D models and objects in the library. The final library is only dependent on the examples provided and not the order in which they are processed.

2.1. Object segmentation per scene/stereo pair

The goal of this step is to obtain tentative object

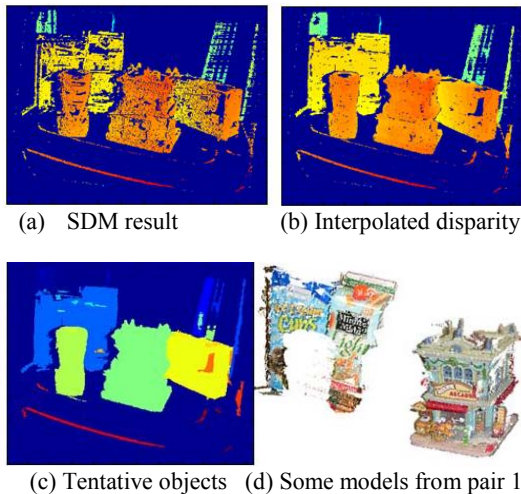


Figure 4 : Disparity map and some objects from pair 1 alone.

segmentation for the given stereo pair. We use color and depth information to obtain these segments. To this end we desire a dense matching and disparity map. Since we do not know the calibration for the stereo configuration, we use the uncalibrated rectification method of [16]. Stratified dense matching (SDM) [9] is used to obtain an initial disparity map. SDM does not provide a dense map, hence we linearly interpolate the disparity to obtain a denser object model. The interpolation is done within each color segment to ensure that we preserve object boundaries. Color segments are obtained using mean shift segmentation [17] of the left image. Intuitively, objects are formed from segments which are ‘smoothly’ connected in 3D. This idea is used to segment the disparity map into tentative objects using mean shift clustering. The results from this step for input pair 1 (see Figure 2) are shown in Figure 4(a)-(c). The Minute maid carton and Ice-cream cups box are grouped as same object since they are placed flush against each other. At this step the algorithm has no knowledge to separate the two models.

2.2. Model building

Using the structure from motion technique outlined in [8] we obtain the camera projection matrices. Using the dense matches provided by the disparity map, 3D reconstruction of the objects is done using the triangulation method described in [22]. Figure 4(d) shows the 3D models of two objects from input pair 1. For the first input pair, these models are directly added to the object model library since no other information is present to refine them. For subsequent pairs the following steps are performed.

2.3. Library refinement

For each of the tentative objects in the current stereo pair we search the library to check if parts of the object have been seen before (from prior input). SIFT features [18] are used to find correspondences. The model must then be projected onto the given scene. Since the camera

projection matrix is different between the stereo pairs, the reconstructed points are related by a collineation. The model obtained from the previous step is up to a uniform scale, hence given the corresponding 3D points (inferred from SIFT matches) we can estimate the collineation relating the two point clouds. This is done using quaternions [19].

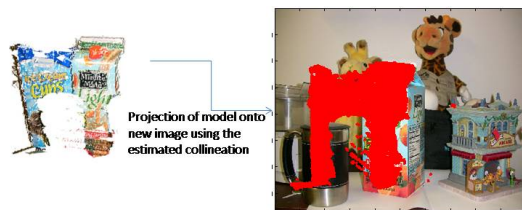
Once the model is transformed to current projective space, we use the camera projection matrix to project the model onto the image (Figure 5a). At this stage the following cases may hold:

- (1) The library model has points which form a subset of the object points seen in the current image.
- (2) Some of the points from the library model are occluded in the current scene.
- (3) Some points in the library model were falsely associated with the object and hence not found in the current scene.

Once the library model has been projected onto the current scene, it is possible to check for the above cases. Model points matching those in the scene are retained. If a model point is occluded in the current scene, then we cannot make a decision yet so we retain the point (red parts of Figure 5b). If neither of the above hold, then it is case (3), and such points are separated to form a new object (green parts of Figure 5b). Using the above reasoning the Minute Maid carton and Ice-cream cups box is separated (Figure 5b). The two house models (from current pair and the one in Figure 4d) were merged to form the model shown in Figure 2. We repeat the above checks now with the current image model and library scene reversed. This ensures we do not accumulate information inconsistent with the previous scenes.

3. Experiments and Results

Figure 2 shows some of the objects extracted from the collection of two scenes. The images are each of size 1024 X 768. The disparity range was approximately -100 to +100. The point clouds obtained for the objects are of the order of 40-100 thousand points. On seeing just input pair 1, the minute maid carton and the ice cream cups box were initially labeled as one object since the boxes were placed flush against each other. The second input scene (pair 2) shows the same minute maid carton in a different pose and without the ice cream cups box at its side. The library model of the matching object projected onto the current scene (Figure 5) is the expected span of that object. The refining steps discussed in section 2.3 then separate parts of the cups box which are determined to be absent in the current scene. The other parts of the minute maid carton seen earlier are merged with the new model obtained from pair 2. The parts of the minute maid carton not seen in



(a) Projection of model onto current image (here pair 2) shown in red over the left image of pair 2.



(b) Left: Estimation of maximum overlap. Red – possible occlusion. Green – parts to split to form new object. Right: The new object split from old model (ice cream cups box) and the merged model of minute maid carton.

Figure 3: Refining a model by transforming to current projective space, projection onto scene and verification.

scene 1 are those which were occluded (self and other objects) and hence are retained as part of minute maid model. Results using an input set of five pairs are shown in Figure 9.

We next provide qualitative and quantitative comparison of our algorithm with multiple segmentation approach [1] and normalized cuts [20]. The source codes for the above methods were obtained from the author web pages. Our method outputs segmentation for the whole input sequence, while the above methods perform segmentation on a per image basis. Therefore, our object segmentation mask is determined by projecting the final object model onto the image. We generated the ground truth by manually labeling the pixels. To provide quantitative comparison we use the measure employed by the Semantic Robot Vision Challenge (SRVC) [21]. The score is the ratio of area (number of pixels) of intersection and area of union, of the binary masks generated from ground truth and the algorithm. This score lies in the range 0 to 1, with 1 indicating perfect segmentation. It takes into account the parts correctly labeled, the parts not included and those falsely attached to the object. Since object discovery would mostly be followed by object description/learning for recognition, we must penalize both incomplete and incorrect grouping.

The object discovery technique presented in [1] uses multiple image segmentations with the assumption that an object is segmented correctly in at least one of the

segmentations. We show the results from the above algorithm on the two scenes in Figure 6. The algorithm requires the number of topics (here objects) and number of (image color) segmentations to be done per image. We used 14 topics and the multiple segmentations were obtained by varying number of segments from 3 to 25 in steps of 2. Top segments for each topic were recovered. Only the non-repeating segments are shown in Figure 6.

The normalized cuts algorithm [20] requires the number of segments as input. We found that this number did not correspond directly to the number of objects in the scene. The results with different number of segments are shown in Figure 7. For the quantitative comparison we use the segment which provided highest (best) score for the object.

Figure 8 shows the qualitative and quantitative comparison for three of the objects. Our algorithm performs better in all the three cases. Also in the first two cases the scores are above 0.9.



Figure 6: Results from object discovery using multiple segmentations [1]



(a) Right to left: Number of segments = 6, 8 & 9.

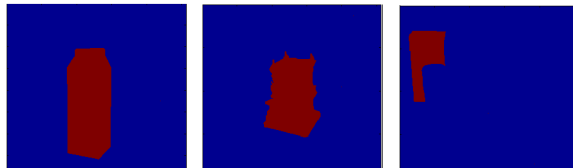


(b) Right to left: Number of segments = 7, 8 & 10.

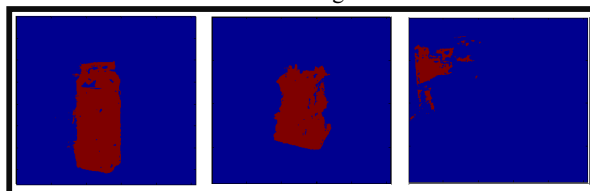
Figure 7: Results from using Normalized cuts (NCuts) [20]



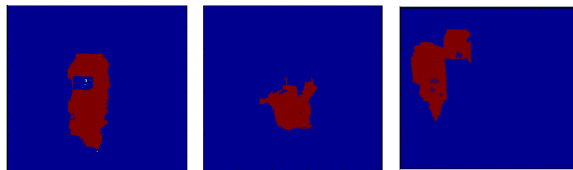
Ground truth masks



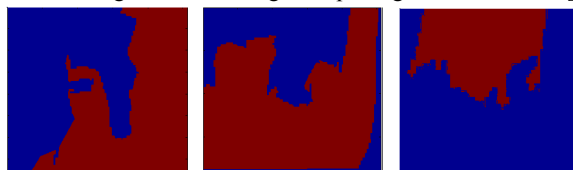
Masks from our segmentations



Masks for segmentation using Normalized Cuts [20]



Masks for segmentation using multiple segmentation method [1]



Scores (=area of intersection / area of union)

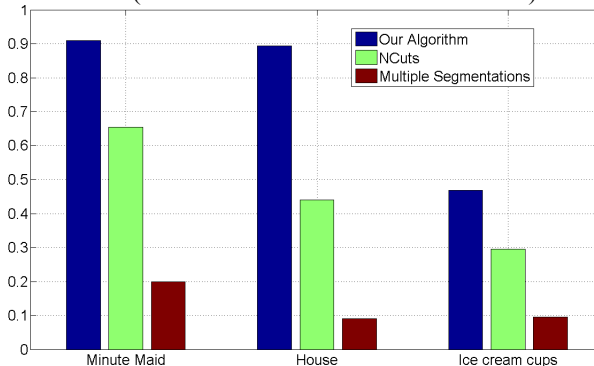


Figure 8: Qualitative & quantitative comparison of results for 3 objects (Minute maid, House model and Ice-creams cups box)



(a) Right images from input pairs and some of the objects extracted by our algorithm.



(b) Some montages for objects discovered using the multiple segmentation algorithm [1]

Figure 9: More results using 5 stereo pairs. (a) Results using our algorithm (b) objects discovered using [1].

4. Discussion and Future direction

We have demonstrated a novel scheme for object discovery from cluttered stereo images. The use of stereo was motivated from many problems common in object segmentation, learning and recognition. The bane of segmentation of objects in images is that, object and texture/color boundaries cannot be distinguished. Gestalt's principles of continuity and proximity are applied to the 3D world where the object resides. We have shown that this provides very good initial object boundaries. Our scheme takes a novel step in object discovery research by providing segmentation consistent with all the input examples. This is done by refining the objects with each new input scene. Quantitative comparisons with other object discovery and image segmentation techniques showed that our algorithm performed significantly better. The other motivation for choosing 3D representation is the fact that it describes the object completely i.e both texture and structure. Many papers have demonstrated the use of 3D models in improving the performance of object recognition algorithms. These algorithms relied on available CAD models or reconstructions from multiple views of an object. These algorithms demonstrated the advantage in terms of pose recovery. We are currently studying techniques to improve searching of large databases/object libraries using both texture and structure cues.

References

- [1] B. C. Russell, A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* 2006.
- [2] J. Sivic, B. Russell, A. A. Efros, A. Zisserman, W. T. Freeman, Discovering Objects and their Location in Images *International Conference on Computer Vision (ICCV)*, Beijing, China, Oct. 2005.
- [3] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. *International Conference on Computer Vision (ICCV)*, Oct 2005.
- [4] J. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation *International Conference on Computer Vision (ICCV)*, 2005.
- [5] A. Stein, T. Stepleton, and M. Hebert, Towards Unsupervised Whole-Object Segmentation: Combining Automated Matting with Boundary Detection, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2008.
- [6] M. Bravo and H. Farid. Object segmentation by topdown processes. *Visual Cognition*, Volume 10(4), pages 471–491, 2003.
- [7] S. Obdrzálek and J. Matas. Sub-linear indexing for large scale object recognition. *Proceedings of the British Machine Vision Conference*, Volume 1, pages 1-10, 2005
- [8] Noah Snavely, Steven M. Seitz and Richard Szeliski, Modeling the World from Internet Photo Collections *Int. J. Computer Vision*, 80(2), pages 189-210, 2008.
- [9] R. r. Jana Kostkov. Stratified dense matching for stereopsis in complex scenes. *British Machine Vision Conference*, pages 339–348, 2003.
- [10] J.-H. Ahn, K. Kim, and H. Byun. Robust object segmentation using graph cut with object and background seed estimation. *Intl Conference on Pattern Recognition (ICPR)*, pages 361–364, 2006.
- [11] K. Y. Lee S H. Object segmentation in stereo images using cooperative stochastic diffusion. *Proceedings of the IEICE General Conference*, page 239, 2001.
- [12] A. Kushal and J. Ponce. Modeling 3d objects from stereo views and recognizing them in photographs, *European Conference on Computer Vision (ECCV)*, pages 563–574, 2006.
- [13] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, 2005.
- [14] H. Arora, N. Loeff, D. A. Forsyth, and N. Ahuja. Unsupervised segmentation of objects using efficient learning, *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [15] K. N. Anastasios Doulamis, Nikolaos Doulamis and S. Kollias. Unsupervised semantic object segmentation of stereoscopic video sequences. *ICIIS '99: Proceedings of the 1999 International Conference on Information Intelligence and Systems*, page 527, 1999.
- [16] A. Fusiello and L. Irsara. Quasi-euclidean Uncalibrated Epipolar Rectification. *International Conference on Pattern Recognition (ICPR)*, 2008.
- [17] P. M. D. Comaniciu. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Analysis Machine Intelligence*, pages 603–619, 2002.
- [18] D. G. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, volume 60, pages 91–110, 2004.
- [19] B. K. P. Horn, Closed-form solution of absolute orientation using unit quaternions, *J. Opt. Soc. Am. A* 4 pages, 629-642, 1987
- [20] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 731–743, 1997.
- [21] <http://www.semantic-robot-vision-challenge.org/>
- [22] Hartley, Rand Zisserman, A, Multiple View Geometry in Computer Vision, *Cambridge University Press*, 2004.
- [23] Vittorio Ferrari, Tinne Tuytelaars, Luc Van Gool, Integrating Multiple Model Views for Object Recognition, *IEEE Computer Vision and Pattern Recognition* Washington, USA, June 2004.
- [24] A. Johnson and M. Hebert, Using spin images for efficient object recognition in cluttered 3D scenes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 21, No. 5, pages 433 - 449. 1999.
- [25] Palmer, Stephen E., Vision Science: Photons to Phenomenology. *Cambridge: The MIT Press* 1999.