Over- and Under-Segmentation Evaluation based on the Segmentation Covering Measure

Jose Sigut
University of La Laguna
Department of Computer Science
Faculty of Physics
Spain (38200), La Laguna, Tenerife
sigut@isaatc.ull.es

Francisco Fumero
University of La Laguna
Department of Computer Science
Faculty of Physics
Spain (38200), La Laguna, Tenerife
franfumero@isaatc.ull.es

Omar Nuñez
University of La Laguna
Department of Computer Science
Faculty of Physics
Spain (38200), La Laguna, Tenerife
omar@isaatc.ull.es

ABSTRACT

Very few measures intended for evaluating the quality of image segmentations account separately for over- and under-segmentation. This distinction is highly desirable in practice because in many applications under-segmentation is considered as a much serious issue than over-segmentation. In this paper, a new approach to this problem is presented as a decomposition of the Segmentation Covering measure into two contributions, one due to over-segmentation and the other one to under-segmentation. Our proposal has been tested on the output of state-of-the-art segmentation algorithms using the Berkeley image database. The results obtained are comparable to those provided by similar evaluation methods allowing a clear separation between over- and under-segmentation effects.

Keywords

Image segmentation, segmentation evaluation, over-segmentation, under-segmentation.

1. INTRODUCTION

Image segmentation plays a major role in a broad range of computer vision applications. Therefore, there is a strong need for objective measures of the quality of a segmentation algorithm on an image or set of images. The most usual way to accomplish this task is by comparing the segmentation at hand with a set of manually-segmented reference images which are often referred as gold standard or ground truth. In recent years there has been a great effort to provide adequate evaluation measures and image databases which have been used as gold standards for different applications [Mar01a] [Unn07a]. However, hardly any of these measures accounts explicitly for over- and under-segmentation. This distinction is highly desirable in practice because in many applications under-segmentation is considered as a much serious problem than over-segmentation since it is usually easier to merge segments to obtain bigger ones than splitting large regions to recover the true segments.

The Segmentation Covering measure has been proven to be a good choice for evaluating segmentation performance [Arb11a]. We will show that under mutual refinement this measure can be written as the contribution of two terms, one of them dealing with over-segmentation and the other one with under-segmentation. An extension to the more general case of arbitrary overlapping regions is also provided. The proposed evaluation method has been tested on the output of three state-of-the-art segmentation algorithms and compared with other evaluation measures using the well-known Berkeley image database [Mar01a].

The rest of the paper is organized as follows. Section II is about related approaches to deal separately with over- and under-segmentation. Section III describes the Segmentation Covering measure and the proposed evaluation method which is derived from it. The experimental results are shown and discussed in section IV. Section V is devoted to the conclusions.

2. RELATED WORK

As far as we know, there are few approaches which account separately for over- and under-segmentation as compared to global evaluation measures. Cardoso and Corte-Real [Car05a] introduce the concept of partition distance $d_{sym}(G,S)$ between a reference segmentation G and the segmentation under study S as a symmetric measure and propose to use an asymmetric version $d_{asy-ov}(G,S)$ for the case of applications where over-segmentation is not an issue. An analogous asymmetric measure $d_{asy-un}(S,G)$ is proposed for the case of under-segmentation.

The information based distance VI proposed by Meila [Meiqq07a] is one of the most popular evaluation measures and is given by

$$VI(S,G) = H(S) + H(G) - 2I(S,G) \tag{1}$$

Where H and I, respectively, represent the entropies and mutual information between S and G.

Meila shows that VI can be written as the sum of two conditional entropies

$$VI(S,G) = H(S|G) + H(G|S) \tag{2}$$

Where the conditional entropies $H(S|G)$ and $H(G|S)$ are identified by Gong and Shi [Gon11a] as over- and under-segmentation metrics, respectively.

Other researchers have focused only on the under-segmentation error. Levinshtein et al [Lev09a] compute this error by means of

$$U_{E-TP} = \frac{1}{numG} \sum_{R_i \in G} \frac{\left[\sum_{S_j \in S : S_j \cap R_i \neq \emptyset} |S_j|\right] - |R_i|}{|R_i|} \tag{3}$$

Where $numG$ is the number of regions in G, R_i denotes any region belonging to G and S_j denotes any region belonging to S. The main disadvantage of using (3) is that it tends to overestimate the amount of under-segmentation because of the inclusion in the calculation of large regions in S with very little overlap. In order to avoid this, Achanta et al [Ach12a] suggest a similar error measure but restricting the overlap to be at least a certain percentage of the segment size as it is expressed in

$$U_{E-Slic} = \frac{1}{N} \sum_{R_i \in G} \sum_{S_j \in S : |S_j \cap R_i| > B} |S_j| \tag{4}$$

Where N is the image size and B is the specified percentage which is set by the authors to 5%.

Protzel and Neubert [Pro12a] propose an alternative under-segmentation measure which overcomes the need for additional parameters. They define the under-segmentation error as

$$U_e = \frac{1}{N} \sum_{R_i \in G} \sum_{S_j \in S : S_j \cap R_i \neq \emptyset} min(S_{jin}, S_{jout}) \tag{5}$$

Where S_{jin} is the portion of S_j inside R_i and S_{jout} is the portion of S_j outside R_i

3. SEGMENTATION COVERING AND PROPOSED MEASURES

The classic overlap measure between two regions R and R' is given by:

$$O(R,R') = \frac{|R \cap R'|}{|R \cup R'|} \tag{6}$$

The Segmentation Covering measure introduced by Arbelaez et al [Arb09a] can be seen as a generalization of (6) to multiple regions so that the covering of a reference segmentation G by a segmentation S is defined as

$$SC(G,S) = \frac{1}{N} \sum_{R_i \in G} |R_i| maxO(R_i, S_j)_{S_j \in S} \tag{7}$$

The definition in (7) can be extended to a family of ground truth segmentations $\{G_i\}$ by first covering each G_i separately with S, and then averaging over them. It can also be analogously defined the covering of S by $\{G_i\}$ but in what follows we will assume that the segmentation covering is calculated as in (7).

Let us consider the ideal case of mutual refinement between the ground truth segmentation and the segmented image. G is said to be a mutual refinement of S if the intersection of every region R_i of G with every region S_j of S is either empty or equal to any of them. From the definition, it is easy to see that if G is a mutual refinement of S, then S is a mutual refinement of G. Figure 1 shows a trivial example of mutual refinement between two images. Under this assumption, it can be shown that each term in the summation in (7) will contribute to the final covering with either over-segmentation or under-segmentation.

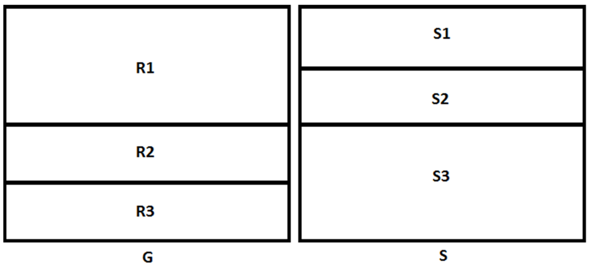

Figure 1. Example of mutual refinement between G and S

In the case of over-segmentation, according to Figure 2, it is clear that

$$O(R_i, S_j)_{S_j \in S} = \frac{|S_j|}{|R_i|} \tag{8}$$

Therefore, the whole contribution can be simply written as

$$|R_i| maxO(R_i, S_j)_{S_j \in S} = max|S_j| \tag{9}$$

In the case of under-segmentation there must be at least two regions of G, R_1 and R_2, contained in a region of S, as shown in Figure 2. It is clear that in this situation the overlap is already maximum so that

$$|R_i| maxO\left(R_i, S_j\right)_{S_j \in S} = \frac{|R_i|^2}{|S_j|}, i = 1, 2 \quad (10)$$

By adding the two terms, we obtain:

$$\sum_{i=1,2} |R_i| maxO\left(R_i, S_j\right)_{S_j \in S} = \frac{|R_1|^2 + |R_2|^2}{|S_j|} \quad (11)$$

The expression in (11) can be easily generalized to an arbitrary number of regions. From the exposed above, (7) can be written as

$$SC(G, S) = SC_{ov} + SC_{un} \quad (12)$$

Where SC_{ov} and SC_{un} are defined respectively as the over- and under-segmentation contributions to the Segmentation Covering and given by

$$SC_{ov}(G, S) = \frac{1}{N} \sum_{R_i \in G} max|S_j|_{S_j \subseteq R_i} \quad (13)$$

$$SC_{un}(G, S) = \frac{1}{N} \sum_{S_j \in S} \frac{\sum_{R_i \subset S_j} |R_i|^2}{|S_j|} \quad (14)$$

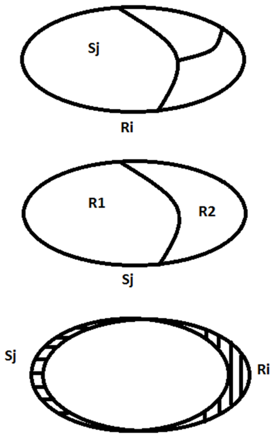

Figure 2. The top image shows over-segmentation under the mutual refinement assumption, the image in the middle shows under-segmentation, and the bottom image shows a more realistic situation of arbitrary overlap

According to (13), in the case of perfect overlap, i.e. $G = S$, $SC_{ov}(G, S) = SC(G, S)$. On the other hand, if the assumption of mutual refinement is not met, as it is usually the case, the expressions in (13) and (14) are not adequate to compute over- and under-segmentation. Figure 2 shows an example of a more realistic scenario of overlap between two segments. Each region is mostly contained in the other one but not completely so it is not clear how over- and under-segmentation should be measured in such a situation.

Our proposal consists of setting a threshold parameter to determine which contribution to the covering in (7) should be considered as either over- or under-segmentation. More concretely, given a region belonging to the ground truth R_i, a segment S_j will be seen to contribute to over-segmentation in that region as long as

$$|R_i \cup S_j| \leq |R_i| + \gamma |R_i| \quad (15)$$

So that the amount of pixels outside R_i to be considered as over- or under-segmentation is controlled by the γ parameter. If every segment S_j which overlaps with a region R_i satisfies (15), the contribution to over-segmentation will be equal to the covering itself for that region. The under_segmentation contribution can be simply defined as the difference between the covering and over_segmentation values. Thus, we can write

$$SC_{ov}(G, S) = \frac{1}{N} \sum_{R_i \in G} |R_i| maxO\left(R_i, S_j\right)_{S_j : |R_i \cup S_j| \leq |R_i| + \gamma |R_i|} \quad (16)$$

$$SC_{un}(G, S) = SC(G, S) - SC_{ov}(G, S) \quad (17)$$

By setting γ=0, (16) and (17) become equivalent to (13) and (14) under the assumption of mutual refinement. SC_{ov} and SC_{un} can be either used as absolute measures as they appear in (16) and (17) or as relative measures given by

$$SC_{ovrel} = \frac{SC_{ov}}{SC},\ SC_{unrel} = \frac{SC_{un}}{SC} \quad (18)$$

As it will be shown in the next section, the relative measures provide a convenient means of evaluating over- and under-segmentation.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Some experiments have been carried out to show the performance of the proposed measures with respect to other measures mentioned in section II. First of all, we will focus on how to set the γ parameter in (15). In general, γ can be set to any positive or zero value depending on the application at hand but in this section we propose a more neutral procedure independent of any particular application or segmentation algorithm.

The proposed procedure is based solely on the reference segmentations provided by the Berkeley image database. Each of the 500 images has an associated ground truth consisting of between 4 and 9 hand-labeled images. The average segmentation covering among these reference images has been computed as well as the average SC_{ov} over them for different values of γ. Figure 3 shows the results of the computation sorted by the average covering value in ascending order, i.e. the agreement among humans for the different images according to this evaluation measure. For the sake of clarity, only the part of the curve with a covering value above 0.9 is shown, corresponding to those ground truth for which there is a strong agreement among subjects. Under these circumstances, very little under-segmentation can be expected and the values of SC_{ov} should be very close to the covering values. According to Figure 3, in order to comply with this requirement, the value chosen for γ should be above 0.25, otherwise it turns out to be too sensitive to small deviations from perfect overlap. For this reason, in all our experiments the value of γ was set to 0.25.

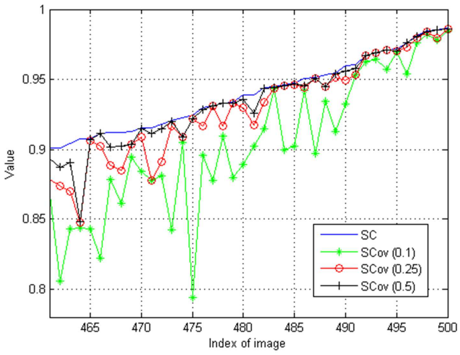

Figure 3. SC and SCov for different values of γ

For the purpose of performance comparison, the asymmetric distances $d_{asy-ov}(G,S)$ and $d_{asy-un}(S,G)$, the conditional entropies $H(S|G)$ and $H(G|S)$, and the under-segmentation error U_e have been selected. The evaluation measures have been tested on the output of three state-of-the-art segmentation methods: the OWT-UCM [Arb11a], the Mean-Shift algorithm [Com02a], and the Efficient Graph segmentation method [Fel04a]. The OWT-UCM has only one threshold parameter to be set which was varied in the range 0<level<1. The Mean-Shift algorithm has three free parameters: color range hr, spatial range hs, and minimum region size minsizeMS. It is well known that the most influential one is hr and for this reason we have set the two others to constant values hs=25, minsizeMS=10, and varied hr in the range 1<hr<30. The Efficient Graph segmentation method has also three parameters and as it happens with the Mean-Shift algorithm, one of them is more influential than the others. Following [Pen13a], we have set the alpha and minimum region size parameters to constant values: alpha=0.5, minsizeEG=10, and let the K parameter vary in the range 100<K<3000. It is very important to remark that the ranges for the parameters of the different methods have been chosen to provide segmentations at varying granularities, from strong over-segmentation with a lot of small regions to strong under-segmentation with very few segments or even just one.

Figures 4, 5, 6, 7, 8 and 9 show the values of the selected over- and under-segmentation evaluation measures averaged over the 500 images of the Berkeley database for the three segmentation algorithms at the specified parameters. The curves corresponding to the conditional entropies have been scaled to the range [0, 1] using the bounds provided in [Gon11a], $log2(N) - H\{G\}$ for the over-segmentation entropy and $H\{G\}$ for the under-segmentation entropy ($H\{G\}$ being the entropy of G and N defined as in (4)), so that they can be more easily compared to the other measures.

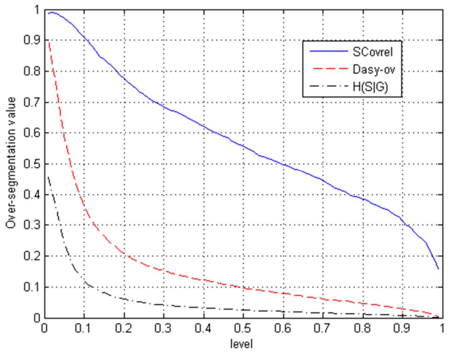

Figure 4. Average over-segmentation values for OWT-UCM

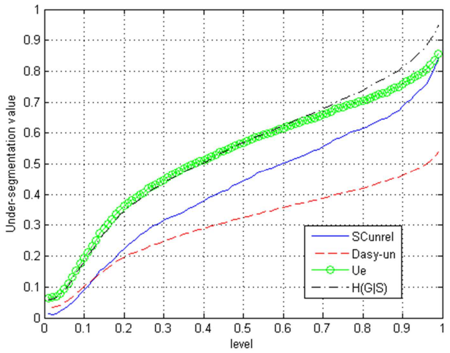

Figure 5. Average under-segmentation values for OWT-UCM

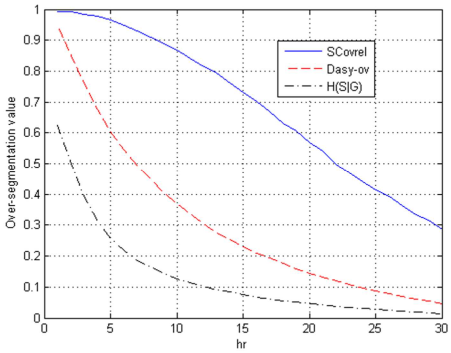

Figure 6. Average over-segmentation values for Mean Shift

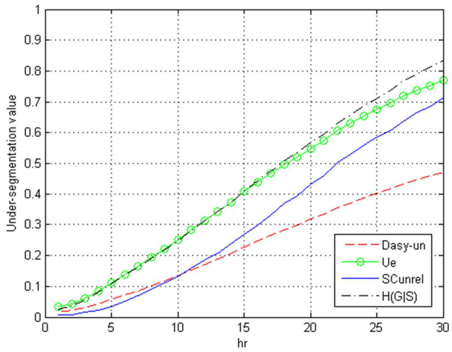

Figure 7. Average under-segmentation values for Mean Shift

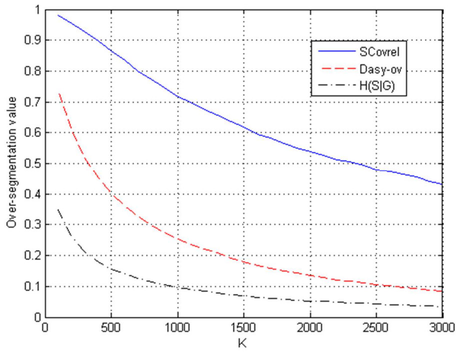

Figure 8. Average over-segmentation values for Efficient Graph

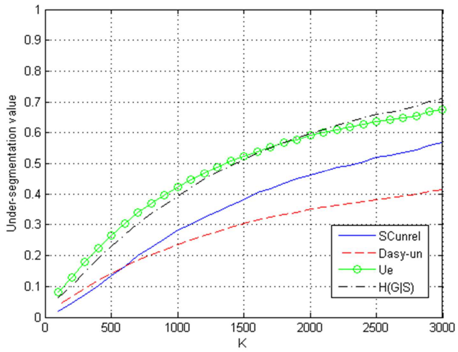

Figure 9. Average under-segmentation values for Efficient Graph

It is difficult to establish a fair comparison among the results obtained for the different measures since they are strongly dependent on how these measures should be interpreted according to their definition. In any case, a high dynamic range to clearly distinguish over-segmentation from under-segmentation seems a reasonable requirement.

In what respects to average over-segmentation, $H(S|G)$ and D_{asy-ov} decrease faster than SC_{ovrel}. As it was already pointed out, SC_{ovrel} measures the relative amount of over-segmentation and takes high values for any over-segmented image including the perfect overlap (good segmentation) as an extreme case. For this reason, it should be considered in conjunction with the covering value itself SC so that it can be correctly interpreted. The dynamic range of $H(S|G)$ is lower than the other two measures, in particular for the UCM-OWT algorithm where the rate of change in the granularity of the segmentations is higher than in the other two algorithms.

Concerning average under-segmentation, the behavior of $H(G|S)$ is very similar to U_e showing a high dynamic range. SC_{unrel} provides also a high dynamic range. Particularly remarkable are the values obtained for the different measures at the upper bound of the interval in the OWT-UCM algorithm where segmentations with only one region are common. Despite this extreme under-segmentation, the average value for D_{asy-un} is only around 0.5 (half the scale).

Table 1 shows the values of the different evaluation measures for certain images at different levels of granularity (OWT-UCM algorithm computed at levels 0.05, 0.5 and 0.9). The images are shown in Figure 10 in the appendix together with the corresponding ground truth. The results are, in general, in accordance with the average curves. The proposed measures clearly separate the over- and under-segmentation effects as it can be seen, for example, in image 6. The image is over-segmented for level=0.05 and consequently SC_{ovrel}=1 as opposed to what happens for level=0.9 where there is only one segment so that SC_{unrel}=1. For level=0.5, the tiger and part of the prey are still correctly segmented but some large parts of the image are not, leading to moderate under-segmentation and this is reflected in the values of SC_{ovrel}=0.29 and SC_{unrel}=0.71.

5. CONCLUSIONS AND FUTURE WORK

Two new evaluation measures have been proposed for dealing separately with over- and under-segmentation. They have been obtained as a decomposition of the Segmentation Covering measure in two contributions. The results of the experiments carried out have been satisfactory showing a good agreement between the values taken by the proposed measures and what should be clearly considered as over- or under-segmentation. It seems that this approach could be also used as a global segmentation evaluation methodology and this is the aim of our future work.

	Evaluation measures							
	SC	SC_{ovrel}	SC_{unrel}	$D_{asy\text{-}ov}$	$D_{asy\text{-}un}$	$H(S\|G)$	$H(G\|S)$	U_e
1	0.53	1	0	0.47	0	2.46	0.03	0.01
	0.99	1	0	0	0	0.03	0.03	0.01
	0.99	1	0	0	0	0.03	0.03	0.01
2	0.42	1	0	0.57	0.05	3.17	0.23	0.10
	0.40	0.63	0.37	0.15	0.46	0.59	1.87	0.70
	0.16	0.34	0.66	0.03	0.77	0.11	3.59	0.94
3	0.30	0.80	0.20	0.64	0.09	3.50	0.33	0.18
	0.43	0.54	0.46	0.03	0.48	0.13	1.81	0.76
	0.18	0	1	0	0.75	0	2.75	1
4	0.62	0.88	0.12	0.30	0.10	1.81	0.42	0.21
	0.23	0.23	0.77	0.05	0.63	0.27	3.06	0.90
	0.08	0.11	0.89	0.02	0.83	0.12	4.08	0.98
5	0.42	1	0	0.58	0.04	2.99	0.17	0.08
	0.70	0.32	0.68	0.08	0.20	0.25	0.79	0.40
	0.35	0	1	0	0.48	0	1.78	0.97
6	0.22	1	0	0.78	0.02	5.18	0.09	0.04
	0.51	0.29	0.71	0.05	0.38	0.16	1.32	0.76
	0.33	0	1	0	0.53	0	1.93	1
7	0.69	1	0	0.30	0.03	2.07	0.15	0.06
	0.84	1	0	0.12	0.04	0.42	0.27	0.08
	0.59	0	1	0	0.28	0	0.91	0.56

Table 1. Evaluation measures calculated for the segmented images in Figure 10. There are three values for each measure corresponding to levels 0.05, 0.5 and 0.9, from top to bottom in that order. The image index is shown on the left

6. ACKNOWLEDGMENTS

This research has been supported by an FPU grant from the Spanish Government (AP2012-6800), and has also been partially funded by the Project PI12/02301 from the Fondo de Investigación Sanitaria, Instituto Carlos III (co-financed with Feder funds).

7. REFERENCES

[Ach12a] Achanta, R., Shaji, A., Smith, K., et al., SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34, no. 11, pp. 2274–2282, 2012.

[Arb09a] Arbelaez, P., Maire, M., Fowlkes, C., et al., From contours to regions: An empirical evaluation. In IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, 2009, pp. 2294–2301.

[Arb11a] Arbelaez, P., Maire, M., Fowlkes, C., et al., Contour Detection and Hierarchical Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33, no. 5, pp. 898–916, 2011.

[Car05a] Cardoso, J., and Corte-Real, L., Toward a generic evaluation of image segmentation. IEEE Transactions on Image Processing, 14, no. 11, pp. 1773–1782, 2005.

[Com02a] Comaniciu, D., and Meer, P., Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, no. 5, pp. 603–619, 2002.

[Fel04a] Felzenszwalb, P.F., and Huttenlocher, D.P., Efficient Graph-Based Image Segmentation. International Journal of Computer Vision, 59, no. 2, pp. 167–181, 2004.

[Gon11a] Gong, H., and Shi, J., Conditional entropies as over-segmentation and under-segmentation metrics for multi-part image segmentation. Technical Reports (CIS), 2011.

[Lev09a] Levinshtein, A., Stere, A., Kutulakos, K., et al., TurboPixels: Fast Superpixels Using Geometric Flows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, no. 12, pp. 2290–2297, 2009.

[Mar01a] Martin, D., Fowlkes, C., Tal, D., et al., A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings, 2001, vol. 2, pp. 416–423 vol.2.

[Meiqq07a] Meila, M., Comparing clusterings - an information based distance. Journal of Multivariate Analysis, 98, no. 5, pp. 873–895, 2007.

[Pen13a] Peng, B., Zhang, L., and Zhang, D., A survey of graph theoretical approaches to image segmentation. Pattern Recognition, 46, no. 3, pp. 1020–1038, 2013.

[Pro12a] Protzel, P., and Neubert, P., Superpixel Benchmark and Comparison. Proc. of Forum Bildverarbeitung, 2012.

[Unn07a] Unnikrishnan, R., Pantofaru, C., and Hebert, M., Toward Objective Evaluation of Image Segmentation Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, no. 6, pp. 929–944, 2007.

8. APPENDIX

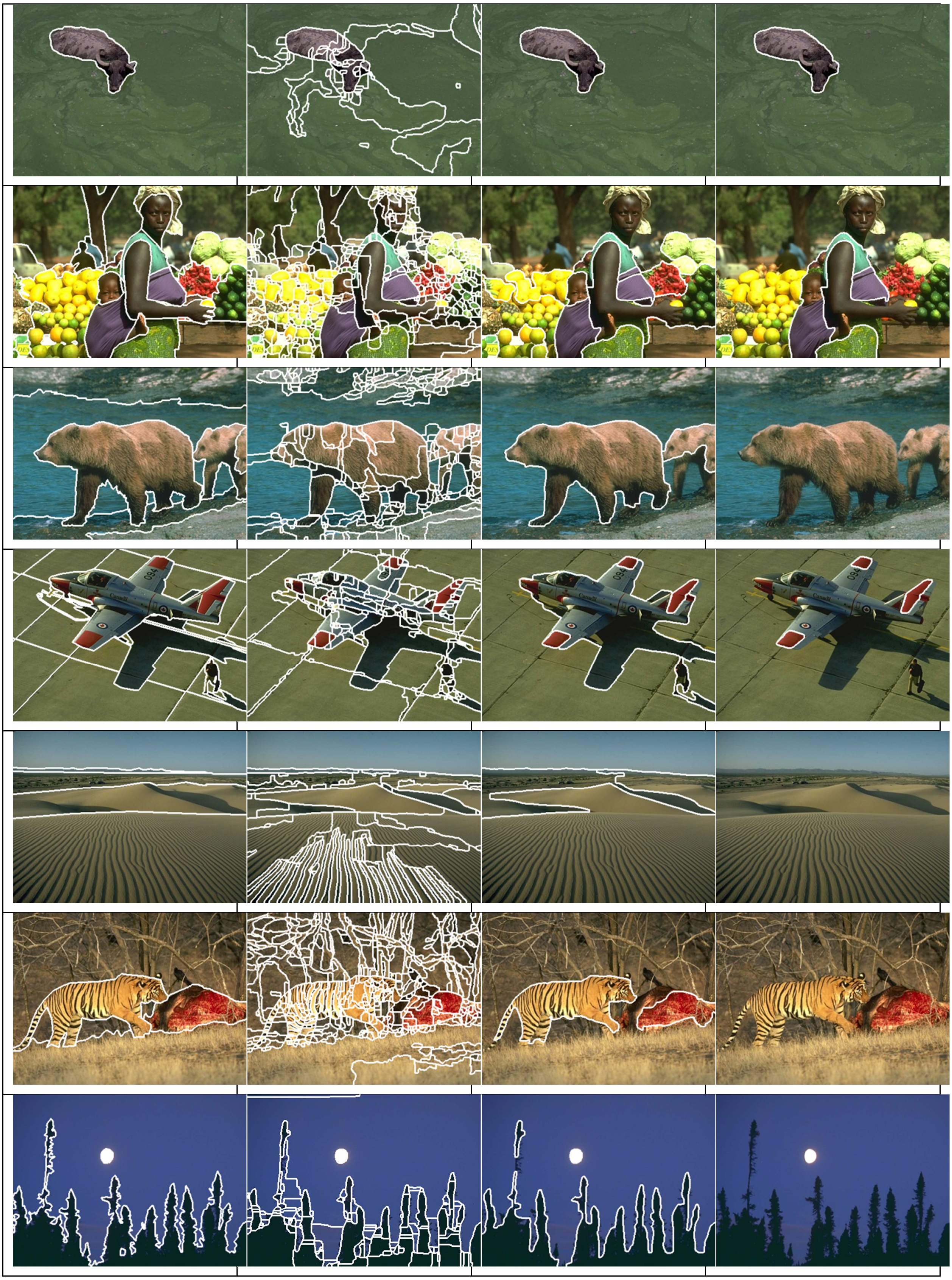

Figure 10. OWT-UCM segmentations at 0.05, 0.5 and 0.9, left to right. Reference images in first column