Evaluation of Morphological Hierarchies
for Supervised Segmentation

Benjamin Perret[1(✉)], Jean Cousty[1], Jean Carlo Rivera Ura[1],
and Silvio Jamil F. Guimarães[1,2]

[1] Université Paris Est, LIGM, ESIEE Paris, France
{benjamin.perret,jean.cousty}@esiee.fr
[2] PUC Minas - ICEI - DCC - VIPLAB, Minas Gerais, Brazil
sjamil@pucminas.br

Abstract. We propose a quantitative evaluation of morphological hierarchies (quasi-flat zones, constraint connectivity, watersheds, observation scale) in a novel framework based on the marked segmentation problem. We created a set of automatically generated markers for the one object image datasets of Grabcut and Weizmann. In order to evaluate the hierarchies, we applied the same segmentation strategy by combining several parameters and markers. Our results, which shows important differences among the considered hierarchies, give clues to understand the behaviour of each method in order to choose the best one for a given application. The code and the marker datasets are available online.

Keywords: Hierarchy · Supervised segmentation · Morphology

1 Introduction

We propose an application driven comparison of several partition hierarchies proposed in the mathematical morphology community: quasi-flat zones hierarchy [1,2], various watershed hierarchies [3,4,5] (by altitude, area, volume, and dynamics), constrained connectivity hierarchy [6], hierarchical observation scale [7]. Compared to the evaluation strategy proposed by Arbelaez et al. [8] which searches for the best segmentation compared to a ground truth segmentation without any prior on the number of regions, we propose to focus on the single object supervised segmentation relying on two marker images: the foreground marker indicates pixels that must be in the segmented object while the background marker gives a set of pixels that is not in the object. Thus, rather than searching if one can find a segmentation that resembles a human segmentation of the whole scene, we evaluate: 1) if a hierarchy contains a set of regions that matches a given object of the scene, and 2) how difficult it is to find this set.

The problem of the automated evaluation of supervised and interactive segmentation algorithms has recently received increasing attention [9,10,11,12]. Following the idea of [12], and in order to perform an objective and quantitative

evaluation, we have automatically generated markers from two publicly available one object image datasets: Grabcut [13] and Weizmann [14]. For each image of the databases, we created a set of several foreground and background markers from the ground truth segmentations. Those markers were designed in order to represent different difficulty levels and do not necessarily aim to reproduce possible user interactions. Then, we have selected a simple marked segmentation strategy which consists in searching for the largest regions of the hierarchy which intersect the foreground marker and does not touch the background marker. Our tests also evaluate the importance of several parameters as the adjacency relation and the dissimilarity measure between pixels.

The contributions of this paper are the following. We propose a novel evaluation framework for partition hierarchies relying on the marker based segmentation problem. The test images are taken from public datasets and the generated markers are available online at http://perso.esiee.fr/~perretb/markerdb/. This framework is applied to several morphological hierarchies and allows us to draw some conclusions on relevance of the evaluated hierarchies. A demonstration website where users can segment their own images using the evaluated hierarchies is also available at http://perso.esiee.fr/~perretb/ISeg/.

2   Hierarchies

We give a short description of the hierarchies that will be evaluated and the reader can refer to the cited articles in order to get formal definitions or additional information. A hierarchy of partitions is a sequence of partitions such that each partition is a refinement of the previous partition in the sequence. It is usually represented as a tree or a dendrogram and can be visualized as a saliency map, which is a contour map in which the grey level represents the strength of the contour: *i.e.*, its level of disappearance in the hierarchy.
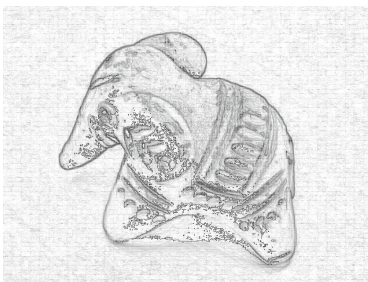
The links that exist among most of the presented hierarchies and efficient algorithms to construct them are described in [15,16].

The Quasi-Flat Zones (QFZ) hierarchy (Fig. 1(b)) is a classical structure that is constructed by considering the connected components of the level sets of the dissimilarity function [1,2]. More precisely, we say that two adjacent pixels of an image are λ-connected if there dissimilarity is lower than or equal to a value λ. For a given λ in $\mathbb{R}$, the equivalence classes of the relation "is λ-connected" form the λ-partition of the image into its λ-connected components also called λ-flat zones. The set of all λ-partitions for every λ in $\mathbb{R}$ forms the QFZ hierarchy.
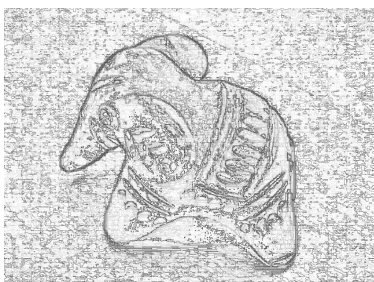
Constrained connectivity (CC) hierarchy (Fig. 1(c)) is a filtered version of the QFZ hierarchy [6]. It is constructed by adding additional constraints to the definition of the connectivity. In this work we consider only the global range constraint, which limit the maximal dissimilarity between two pixels of a same connected component. This idea was introduced in order to prevent the chaining effect that may appear in the QFZ hierarchy.
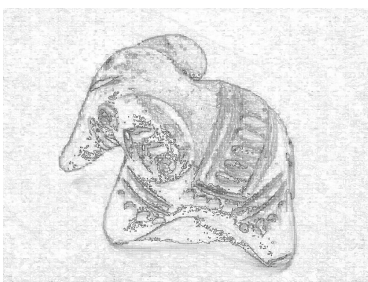
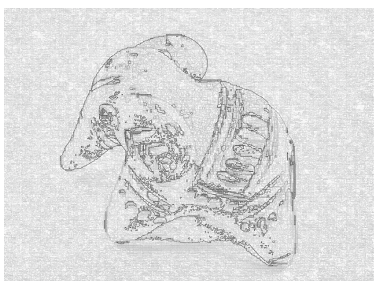(a) Elefant from Grabcut DB   (b) Quasi-flat zones
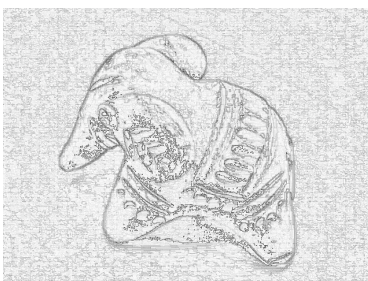
(c) Constrained connectivity   (d) Watershed Altitude

(e) Watershed Area   (f) Watershed Volume

(g) Watershed Dynamics   (h) Observation scale

Fig. 1. Examples of saliency maps for each method

Watershed (WS) hierarchies (Figs. 1(d), 1(e), 1(f), and 1(g)) are constructed by considering the watershed segmentation of an image that is iteratively flooded under the control of an attribute [3,4,5,17]. For example, the watershed segmentations of the area closings of size k of an image for every positive integer k form the WS hierarchy by area of the image. In this article, we consider 4 possible attributes: altitude (WSAlt), dynamics (WSDyn), area (WSArea), and volume (WSVol).

Observation scale hierarchy (Fig. 1(h)) is a hierarchical version [7] of Felzenswalb et al. segmentation algorithm [18]. This approach relies on a predicate that measures the evidence for a boundary between two regions using scale and contrast information.

3   Supervised Segmentation Algorithm

In this evaluation we have chosen to use the procedure described in [19] that constructs a two classes segmentation from a hierarchy and two non-empty markers: one for the background and one for the object of interests. Its principle is to identify the object as the largest regions of the hierarchy that intersect the object marker but does not touch the background marker. This result can be computed efficiently in two passes on the hierarchy (real time interaction). In the first pass, we browse the hierarchy from the leaves to the root in order to determine for each node if it intersects each marker. The second pass, where the tree is browsed from the root to the leaves, determines the final class of each node following this rule:

- if the node intersects the background then its final label is *background*;
- else, if the node intersects the object marker then its final label is *object*;
- else the node has the same label as its parent.

Finally, as this procedure tends to produce segmentations with a lot of holes in some hierarchies, we also consider a post-processing where the holes of the segmented object that do not contain a pixel of the background marker are filled.

4   Database and Marker Generation

We consider two publicly available datasets focused on the single object segmentation problem: Grabcut [13] and Weizmann [14]. Grabcut and Weizmann datasets are composed of 50 and of 100 colour images respectively. Each image contains at least one relatively large object which is identified in a ground-truth segmentation. In the Grabcut dataset, the ground truths are stored as tri-maps which identify the object, the background, and mixed pixels (which are excluded from the computation of the scores). In the Weizmann dataset, each image comes with 3 human 2 classes segmentations and the final ground truth is determined with a majority vote (a pixel is classified as object if at least two humans have classified it as object).
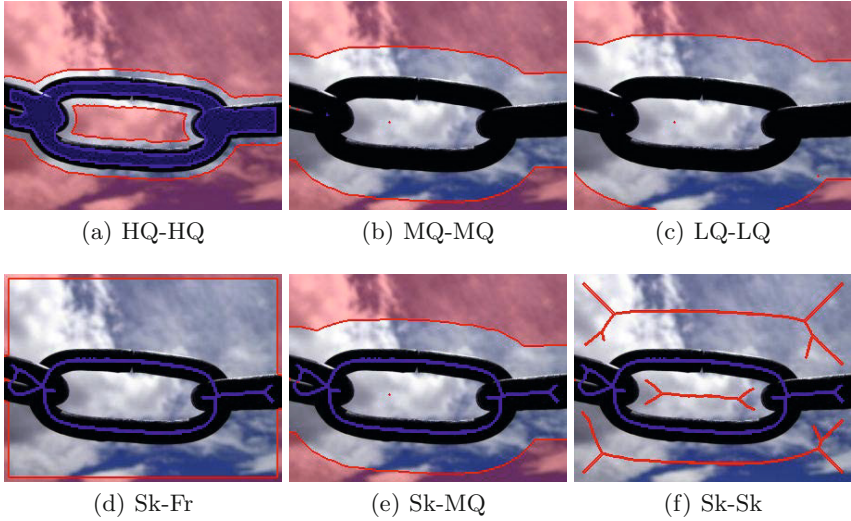
(a) HQ-HQ   (b) MQ-MQ   (c) LQ-LQ

(d) Sk-Fr   (e) Sk-MQ   (f) Sk-Sk

Fig. 2. Different combinations of markers. The combination of markers is indicated in the caption of each sub-figure in the form Background Marker-Object Marker. In each figure the background and foreground marker are respectively depicted in red and blue.

In order to perform a fair and objective evaluation of the different hierarchies we propose several automatic strategies to generate object and background markers from the ground truth. Our main idea here is not to reproduce the interactive segmentation process experienced by a real user but rather to obtain markers representing either various difficulty levels using erosions of various sizes, or markers that resembles to human generated markers using skeletonization. The generated markers are the following (see Fig. 2):

- – High Quality (HQ) marker: erosion by a ball of radius 7 pixels;
- – Medium Quality (MQ) marker: erosion by a ball of radius 30 pixels;
- – Low Quality (LQ) marker: erosion by a ball of radius 45 pixels;
- – Skeleton (Sk): morphological skeleton given by [20]; and
- – Frame (Fr): frame of the image minus the object ground truth if the object touches the frame (background only). Using the frame as the background marker is nearly equivalent to having no background marker in the sense that it does not depend of the ground truth or the image.

If a connected component is completely deleted by the erosion then a single point located in the ultimate erosion of this connected component is added to the marker.

In the following, the combination of the background marker MB and the foreground marker MF is denoted MB-MF (for example, HQ-MQ stands for the combination of a high quality marker for the background and a medium quality marker for the foreground). Among all the possible combinations of markers, we chose to concentrate on the following ones:

- HQ-HQ, MQ-MQ, LQ-LQ represent a sequence of increasing difficulty as the markers get smaller. Nevertheless, all those combinations are symmetric in the sense that the correct segmentation is roughly at equal distance from the foreground to the background marker;
- Sk-Sk, MQ-Sk, Fr-Sk: here the foreground marker is always the skeleton of the ground-truth, while the background marker gets further and further from the ground-truth.

The performance of each segmentation result is evaluated with the F-measure

$$F = \frac{2.Recall.Precision}{Recall + Precision} \tag{1}$$

with

$$Recall = \frac{TP}{TP + FP} \quad , \text{ and } Precision = \frac{TP}{TP + FN} \tag{2}$$

where TP, FP, and FN stand respectively for the number of True Positives, False Positives, and False Negatives pixels.

5   Results and Discussions

The overall results, combining the scores obtained with the 6 marker combinations on the two datasets (there are thus $6 \times 150 = 900$ measurements for each method), are presented in Fig. 3 using box-and-whisker plots. In these experiments, we considered a 4-adjacency relation with a Lab gradient (that is the Euclidean distance in the L*a*b* colour space) for the dissimilarity measure. We can see that the WSArea and WSVol globally achieve the best performance with very similar results. QFZ and CC are at the bottom of the ranking: while this result is not surprising for QFZ, which is the most basic method, it suggests that the global range constraint in CC, which is supposed to remove spurious regions from the hierarchy, has nearly no effect in this application. The theoretical similarity between QFZ and WSAlt [15] is confirmed by the experiments which show similar results for the two methods. WSDyn provides a small improvement compared to QFZ. Surprisingly, WSVol which can been seen as a combination between WSArea and WSDyn does no seem to take advantage of the information given by the depth measure (similar to the dynamics) compared to WSArea. Finally, OS achieves slightly better results than QFZ, but it does not seem that the segmentation strategy is able to take advantage of the area regularization provided by the method as it remains far from WSArea and WSVol.

The results per marker combination are presented in Fig. 5. The ranking remains the same in the symmetric cases (first row). Nevertheless, we can observe a large gap between the results obtained with low quality and medium quality markers (average increase of 0.1 on the median f-measure). In cases more similar to user interactions implying the skeleton as the foreground marker, we can see that all methods are relatively robust to the quality of the background marker. Nevertheless when the frame is used as the background marker, the combination is strongly asymmetric, and the results of WSVol and especially WSArea become
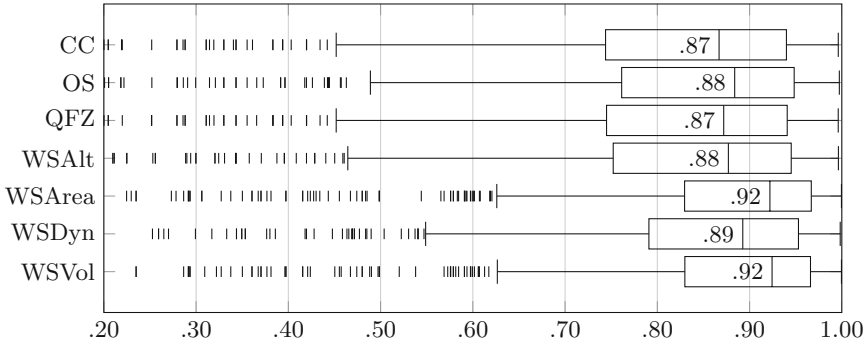
Fig. 3. Results obtained with each hierarchy for all combinations of markers. For each method, we see: 1) the median F-measure (central bar), 2) the first and third quartile (extremities of the box), 3) the lowest datum still within 1.5 inter quartile range (difference between the third and first quartile) of the lower quartile, and the highest datum still within 1.5 inter quartile range of the upper quartile range (left and right extremities), and 4) the outliers (individual points).

less reliable (lower first quartile). This last effect is understandable as the area regularization used in WSArea tends to produce regions of homogeneous sizes which does not always reflect the content of the images. Fig. 4 shows the evolution of the segmentations with respect to the marker combination for the method WSVol on a sample image.

Concerning the influence of the adjacency relation, Fig. 6 presents the result obtained for two hierarchies QFZ and WSArea using a 4- or a 8-adjacency relation (with a Lab gradient and the 6 combinations of markers). We observe

Fig. 4. Segmentation examples using WSVol for the 6 marker combinations. In each figure the background and foreground marker are respectively depicted in red and blue while the contour of the segmentation result is in green.
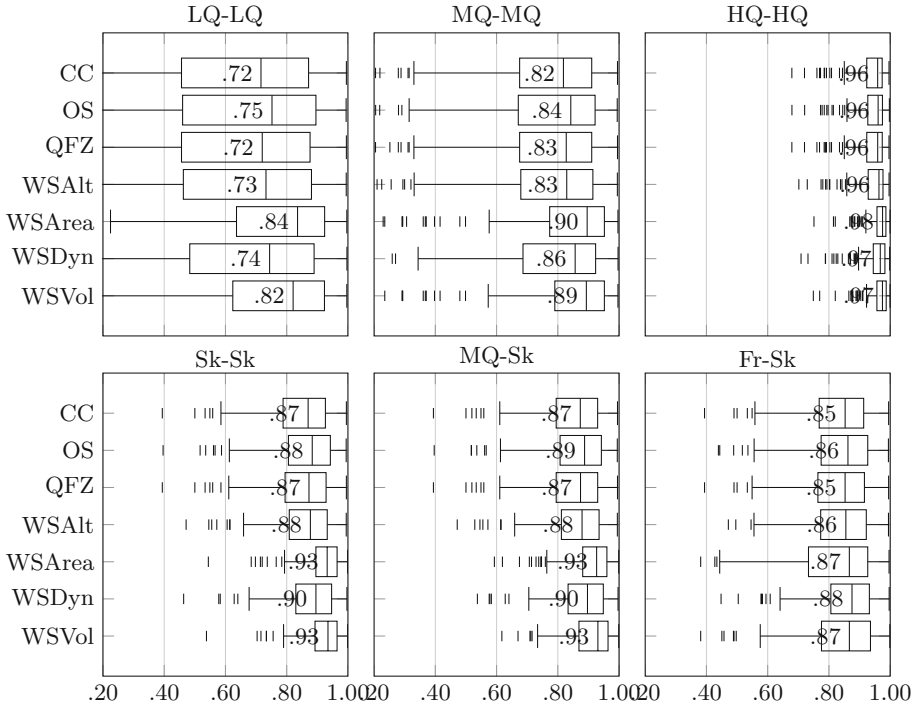
Fig. 5. Results obtained with each hierarchy and each combination of markers. See description of Fig. 3.

that there is nearly always a small gain when using a 8-adjacency instead of a 4-adjacency relation (see Fig. 6). Thus, due to the limited computational overhead, it seems a good idea to favour 8-adjacency over 4-adjacency relation in such segmentation applications.

The importance of the dissimilarity measure is illustrated in Fig. 7. We compare the results obtained with QFZ and WSArea with 3 different dissimilarity measures: 1) the absolute difference of the luminance (grey-scale image computed as the average of the RGB channels), 2) a city-block distance in the RGB space, and 3) an Euclidean distance in the Lab space. While the results are greatly improved by more complex dissimilarity measures for QFZ, the effect is mostly negligible for WSArea.

In Fig. 8, we measure the effect of filling the holes in detected object as a post processing of the segmentation process. As expected, the hierarchies that do not use a form of size regularization criterion are very sensitive to this post-processing, meaning that there is a lot of small contrasted regions (noise, specular, or textures) that lies close to the root in such hierarchies: those regions are thus more probably assigned to the background.
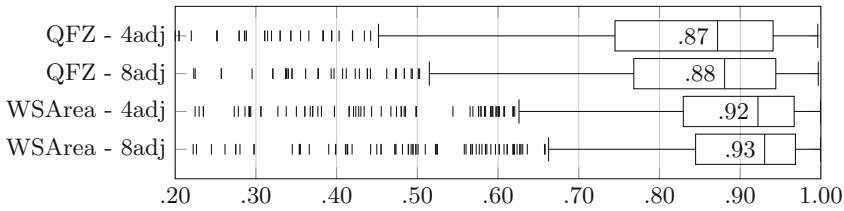
Fig. 6. Comparison between 4- and 8-adjacency relations. See description of Fig. 3.
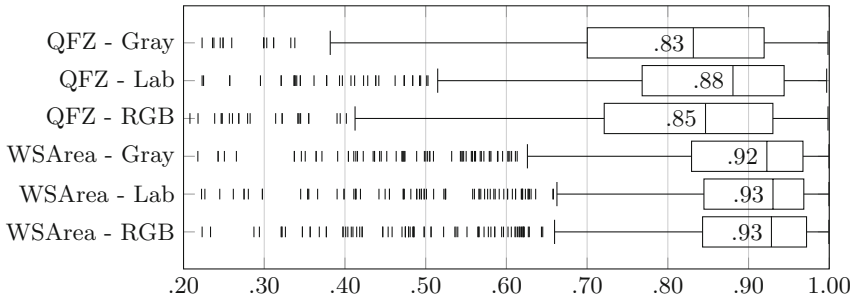

Fig. 7. Comparison between dissimilarity measures. See description of Fig. 3.
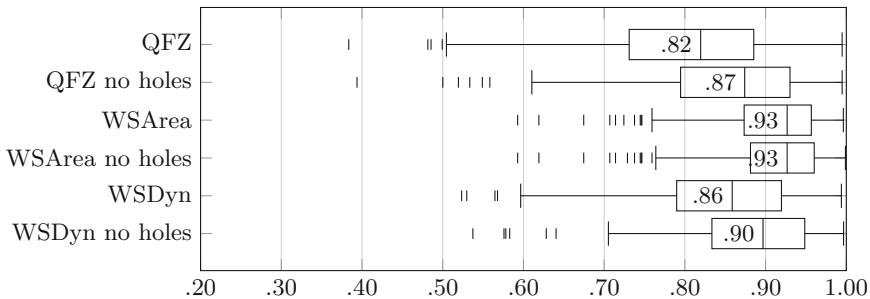

Fig. 8. Effect of filling the holes as a post processing. See description of Fig. 3.

One can also note that there is an important difference of difficulty between Grabut and Weizmann datasets. In Fig. 9, which compiles the overall results (Fig. 3) per dataset (Lab gradient, and 4-adjacency relation), we can observe a significant difference in the median score and even a larger difference in the first quartile. Indeed a visual inspection of the results shows that Weizmann dataset, contrarily to Grabcut dataset, contains images of objects with either very low contrast or with very large scale textures that all the considered methods have difficulties to segment correctly.

Finally, Fig. 10 shows several examples of segmentations.

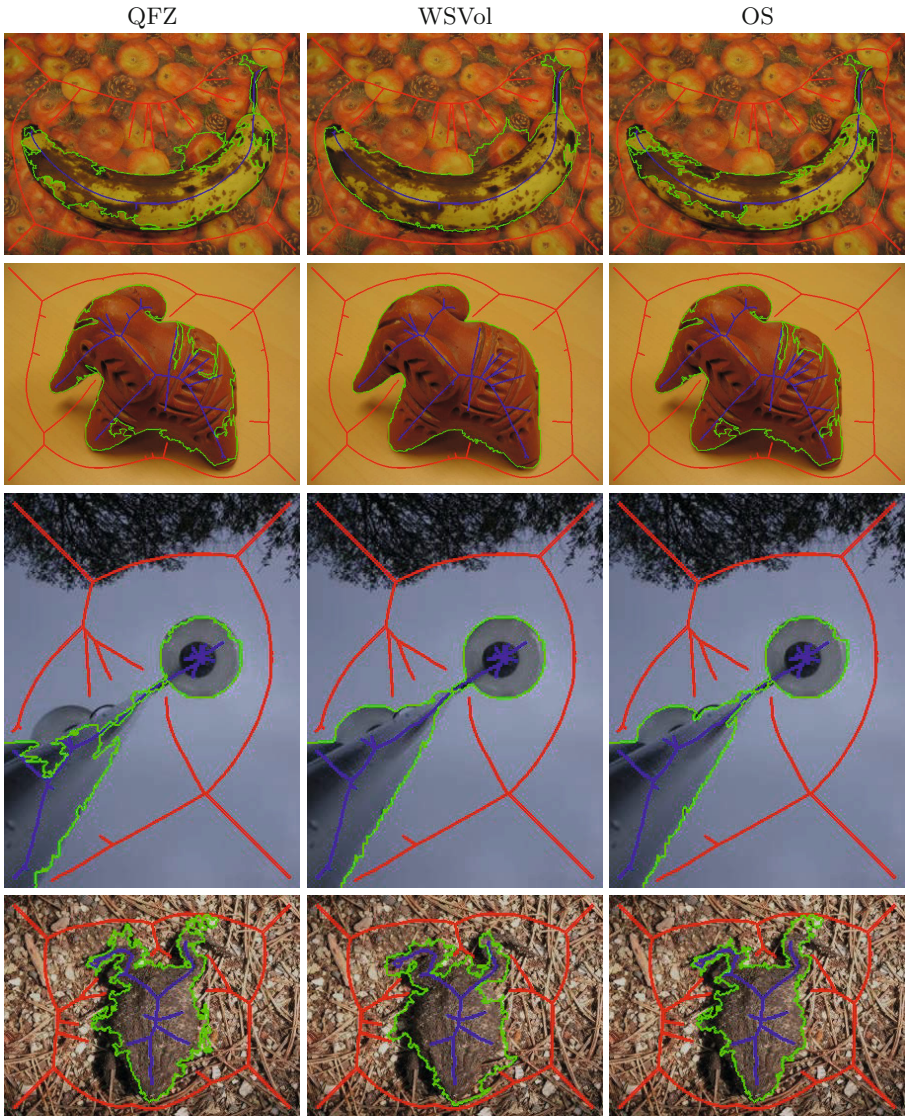**Fig. 9.** Comparison between Grabcut and Weizmann datasets. See description of Fig. 3.

Fig. 10. Examples of segmentation results for QFZ, WSVol, and OS. See description of Fig. 4.

6 Conclusion

We performed a systematic and automatic comparison of several morphological hierarchies in the context of the supervised segmentation task. We have designed a framework in order to automatically generates markers from the ground-truth segmentations. The segmentation strategy, which extracts the object from the hierarchy and the markers, is voluntarily simple in order to keep the intelligence in the design of the hierarchy.

Our results suggest that the hierarchical watersheds based on area and volume attributes are generally the best choice among morphological hierarchies for such task. The only limiting case is when the information about the background marker is very weak: in this case, that is nearly the one marker segmentation problem with the adopted segmentation strategy, it is best suited to use a hierarchical watershed by dynamics.

It is of course difficult to say if these results can be generalized to other vision tasks. Nevertheless, we believe that our test framework is a good indicator to see if a given hierarchy is well suited to represent objects in natural scenes and how easy it is to extract such object from it. Moreover, as shown in [15,16], constructing a hierarchical watershed from the QFZ hierarchy is only a linear (with respect to the number of pixels) time post-processing and it may significantly improve results.

The marker datasets are available at http://perso.esiee.fr/~perretb/markerdb/. We have also designed a web site http://perso.esiee.fr/~perretb/ISeg/ that implements the segmentation strategy with an interactive interface and that enables to test the hierarchies on custom images. The site software is implemented in JavaScript and runs entirely in the client web browser: the code to construct the hierarchies and to perform the segmentation is available.

In future works, we plan to test if we obtain similar results in other evaluation frameworks as the one proposed by Arbelaez et al. [8] or more recently by Pont-Tuset et al. [21]. Another question we will have to investigate is also the comparison to non morphological hierarchies and state of the art supervised segmentation methods. Moreover, all those results show that none of the presented methods is perfect and the construction of better hierarchies is an open issue.

References

1. Nagao, M., Matsuyama, T., Ikeda, Y.: Region extraction and shape analysis in aerial photographs. CGIP 10(3), 195–223 (1979)
2. Meyer, F., Maragos, P.: Morphological scale-space representation with levelings. In: Nielsen, M., Johansen, P., Fogh Olsen, O., Weickert, J. (eds.) Scale-Space 1999. LNCS, vol. 1682, pp. 187–198. Springer, Heidelberg (1999)
3. Beucher, S.: Watershed, hierarchical segmentation and waterfall algorithm. In: ISMM, pp. 69–76 (1994)
4. Najman, L., Schmitt, M.: Geodesic saliency of watershed contours and hierarchical segmentation. PAMI 18(12), 1163–1173 (1996)
5. Meyer, F.: The dynamics of minima and contours. In: ISMM, pp. 329–336 (1996)

6. Soille, P.: Constrained connectivity for hierarchical image partitioning and simplification. PAMI 30(7), 1132–1145 (2008)
7. Guimarães, S.J.F., Cousty, J., Kenmochi, Y., Najman, L.: A hierarchical image segmentation algorithm based on an observation scale. In: Gimel'farb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR & SPR 2012. LNCS, vol. 7626, pp. 116–125. Springer, Heidelberg (2012)
8. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI 33(5), 898–916 (2011)
9. Moschidis, E., Graham, J.: A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction. In: IEEE ISBI, pp. 928–931 (2010)
10. McGuinness, K., O'Connor, N.E.: Toward automated evaluation of interactive segmentation. CVIU 115(6), 868–884 (2011)
11. Klava, B., Hirata, N.: A model for simulating user interaction in hierarchical segmentation. In: ICIP (2014)
12. Zhao, Y., Nie, X., Duan, Y., Huang, Y., Luo, S.: A benchmark for interactive image segmentation algorithms. In: IEEE Workshop on POV, pp. 33–38 (2011)
13. Blake, A., Rother, C., Brown, M., Perez, P., Torr, P.: Interactive image segmentation using an adaptive GMMRF model. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 428–441. Springer, Heidelberg (2004)
14. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: IEEE CVPR (2007)
15. Cousty, J., Najman, L., Perret, B.: Constructive links between some morphological hierarchies on edge-weighted graphs. In: Hendriks, C.L.L., Borgefors, G., Strand, R. (eds.) ISMM 2013. LNCS, vol. 7883, pp. 86–97. Springer, Heidelberg (2013)
16. Najman, L., Cousty, J., Perret, B.: Playing with kruskal: Algorithms for morphological trees in edge-weighted graphs. In: Hendriks, C.L.L., Borgefors, G., Strand, R. (eds.) ISMM 2013. LNCS, vol. 7883, pp. 135–146. Springer, Heidelberg (2013)
17. Cousty, J., Najman, L.: Incremental algorithm for hierarchical minimum spanning forests and saliency of watershed cuts. In: Soille, P., Pesaresi, M., Ouzounis, G.K. (eds.) ISMM 2011. LNCS, vol. 6671, pp. 272–283. Springer, Heidelberg (2011)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV 59(2), 167–181 (2004)
19. Salembier, P., Garrido, L.: Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. TIP 9(4), 561–576 (2000)
20. Chaussard, J., Couprie, M., Talbot, H.: Robust skeletonization using the discrete lambda-medial axis. PRL 32(9), 1384–1394 (2011)
21. Pont-Tuset, J., Marques, F.: Measures and meta-measures for the supervised evaluation of image segmentation. In: CVPR (2013)