Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation

Xiang Li
x.li-29@sms.ed.ac.uk
Ben Aldridge
ben.aldridge.ed.ac.uk
Jonathan Rees
jonathanlrees.mac.com
Robert Fisher
rbf@inf.ed.ac.uk

University of Edinburgh
Edinburgh, UK

Abstract

Having ground truth is critical for evaluating segmentation algorithms and finding the ground truth remains a hard problem. In this paper, three methods to estimate the ground truth for skin lesion segmentation using multiple manual results collected from different experts are proposed and compared. We also analyze the manual segmentations and discuss how to use them more effectively. We conclude that a voting policy produces a slightly better ground truth than the other two optimization based approaches. We propose that a better ground truth should take into account different styles of segmentations.

1 Introduction

Segmentation evaluation can be categorized into two groups: supervised and unsupervised evaluation, depending on whether the method utilizes *a priori* knowledge[3, 7]. Here we are only concerned with supervised evaluation which is widely used in medical image research. It computes the difference between the ground truth and a segmentation result using a given evaluation metric. Much effort is spent on the design of the metrics[1, 7]. However, there is the interesting question of how to obtain the ground truth against which the metrics are calculated. This is always a difficult issue to tackle and there have been few investigations of it. The most common method is to use an expert's manual segmentation and declare that as the ground truth [5]. A single expert's segmentation is likely to be subject to that expert's bias, hence it is proposed to make several manual segmentations for one image by different people[7] and the ground truth is derived from these results. For example, Yuan et al.[8] used the average contour of three dermatologists as the ground truth; we previously [2] considered the ground truth as that agreed by at least half of the experts. However, it is worth questioning whether these simple ways of combining multiple segmentations produce a good quality ground truth; are there more appropriate ways to provide the ground truth?

This article is the first 1) to propose and compare three different ways to derive the ground truth and 2) to categorize the manual segmentations into different groups.

2 Methods for ground truth estimation

Some notations used in the paper are as following:
$Manual_{ij}(x)$: the manual segmentation of the i^{th} image drawn by the j^{th} of J experts at pixel x
$GT_i(x)$: the estimated ground truth of the i^{th} image at pixel x
I: the number of images; J: the number of manual results
$\mathbf{P}(\Omega)$: the partition of the image Ω into N regions: $\{\Omega_n\}_{n=1}^{N}$, $\bigcup_{n=1}^{N}\Omega_n \equiv \Omega$, Ω denotes the image domain, N is the number of regions (N = 2 for binary-value images).
Both the manual results and the ground truth are represented as binary-valued image. The foreground has value 1 and the background has value 0. We propose the three methods:

Voting policy
Finding the ground truth based on multiple reference segmentations can be considered as a labeling problem. The most intuitive way of solving such problems is to use a voting policy (or label voting [4]). A voting threshold k is used to determine the classification of each pixel. The threshold is normally defined as $k = \frac{J+1}{2}$ and a pixel belongs to the foreground if and only if at least k people vote for it as the foreground. The binary-valued ground truth is defined as:
$GT_i(x) = \begin{cases} 1 & \text{if } \sum_{j=1}^{J} Manual_{ij}(x) \geq k; \\ 0 & \text{otherwise.} \end{cases}$

Variation Based Method
The second approach minimizes the average variation between the GT and manual results. This is equivalent to minimizing the average area of the non-overlap region between GT_i and $Manual_{ij}$. Hence, the energy function is, $E_i = \sum_{j=1}^{J}\sum_{n=1}^{N}\{\sum_{x_k\in\Omega_n}[GT_i(x_k) - Manual_{ij}(x_k)]^2\}$.

Maximal *a posteriori* probability based method
The third method is based on statistical theory. The probabilistic formulation estimates the ground truth as a process of finding an optimal partition $\mathbf{P}(\Omega)$ of the image domain. It maximizes the *a posteriori* probability $p(\mathbf{P}(\Omega))$ based on a set of manual results. Simply speaking, the ground truth should be the segmentation that makes all the manual results most probable. As a result, the *a posteriori* probability function has the form:

$$p(M_{i\{1,\ldots,J\}}|\mathbf{P}) = p(Manual_{i\{1,\ldots,J\}}|\Omega_1,\Omega_2,\ldots,\Omega_N) = \prod_{n=1}^{N} p_{in}(Manual_{i\{1,\ldots,J\}}|\Omega_n) = \prod_{n=1}^{N}\prod_{x\in\Omega_n} p_{in}(x). \quad (1)$$

Here, p_{in} is defined as the probability of a pixel selected as region n by J manual results for the i^{th} image: $p_{in}(x) = \frac{1}{J}\sum_{j=1}^{J} Manual_{ij}(x)$. This model assumes that 1) the medical experts derive their segmentations of the same image independently from one another and 2) the segmentation at each pixel is independent. The same assumption appears in STAPLE [6].

3 Experiments on ground truth estimation

Our goal is to estimate and compare the ground truth using the 3 criteria different approaches described in the section 2. The 50 test images we used are randomly selected from our lesion data-base. Their manual segmentations are obtained by 8 dermatologists from the Dermatology department of the University of Edinburgh who directly draw the lesion boundary on the colour image displayed in Adobe Photoshop CS3 using a Wacom Clintiq 12WX Interactive pen tablet.

To evaluate and compare the ground truth derived from different approaches, a quantitative metric XOR that measures the difference between the ground truth and the manual

results is used. For the i^{th} lesion data ($i = 1,\ldots,50$), the corresponding average XOR_i measure is: $XOR_i = \frac{1}{J}\sum_{j=1}^{J} \frac{Area(GT_i \oplus Manual_{ij})}{Area(GT_i + Manual_{ij})}$, ranging from 0 (best) to 1 (worst). $\bigoplus$ denotes exclusive-OR and gives the pixels for which GT_i and $Manual_{ij}$ disagree; $+$ means union. The smaller the XOR, the closer the ground truth is to the manual results.

3.1 The best voting threshold

For the voting method, it is interesting to find out whether the voting threshold $k = \frac{J+1}{2}$ is the best option. Hence, we compute the GT using different threshold values k for different numbers of manual results (J). The XOR measure (mean±standard deviation) comparing the GT against its corresponding manual results is shown in the left of Table 1 (the smallest XOR measures are highlighted in red). It shows that the best estimation of the ground truth is determined when using the voting method with $k = \frac{J+1}{2}$. Also, the XOR decreases when the reducing number of the manual results, which reflects the reduced variation among the dermatologists.

XOR measure ($\times 100$)								
	Voting Threshold (k)				**Methods**			
Manual(J)	3	4	5	6	Voting	Prob	Diff	STAPLE [6]
8	6.70 ± 3.90	6.17 ± 3.62	6.24 ± 3.80	6.92 ± 4.29	6.17 ± 3.62	6.20 ± 3.59	6.20 ± 3.57	6.38 ± 3.76
7	5.46 ± 4.13	5.19 ± 3.87	5.59 ± 4.16	6.82 ± 4.96	5.19 ± 3.87	5.20 ± 3.85	5.21 ± 3.87	6.23 ± 3.69
6	4.59 ± 4.27	4.66 ± 4.39	5.56 ± 5.17		4.59 ± 4.27	4.59 ± 4.26	4.60 ± 4.23	6.39 ± 3.95
5	3.52 ± 3.89	4.03 ± 4.48			3.52 ± 3.89	3.52 ± 3.89	3.52 ± 3.89	6.18 ± 3.61

Table 1: Left: Average segmentation error rates and their standard deviations; Right: Comparison between different methods

3.2 The best ground truth estimation method

We compare the ground truth computed by different approaches using the same evaluation metric XOR. The results are shown in Table 1 (right). According to the XOR measure, the voting method gives the smallest XOR compared to the other two estimation methods. However, considering the range of values in the table, there is no fundamental difference between the three methods. We also compare STAPLE [6] to our 3 algorithms and conclude that its ground truth is worse under the XOR criterion. However, STAPLE optimizes a different criterion so this comparison is not quite fair. We also implemented another dissimilarity measure called Pratt's Figure Of Merit (FOM) which stood out in comparison with five other supervised evaluation criteria for segmentation results and proved to be most effective in a comparison study conducted by Chabrier et al. [1]. It corresponds to an empirical contour distance between the ground truth and the manual results. The additional test results confirm the conclusion obtained by XOR measure.

There are big variations between the manual results given by different people for the same data. This can be explained by both a difference in the segmentation policies, as well as randomness. Take the lesion segmentation problem for example: some dermatologists only draw the boundary along the lesion edge, while others extend the lesion region a little bit more onto the adjacent skin region. This can be considered as a segmentation policy difference. In addition, there are different opinions on the importance of finding the exact lesion boundary. This leads to different attitudes when people perform the manual segmentation. For some of them, locating a general lesion region is necessary for a good diagnosis. Hence, they pay less effort to the exact edge details; while others might pay a

great deal of attention to drawing a very precise pixel-by-pixel boundary. Given the aim of comparing computer-based segmentations against the ground truth, it is more reasonable to use the ground truth which has the more accurate boundary. Therefore, we question if it is appropriate to treat all manual segmentation results equally rather than, for example, using a weighting policy according to their performances. For instance, STAPLE [6] treats each manual segmentation differently according to their performance parameters estimated using EM algorithm. But first, we need to prove that there does exist different segmentation styles. We hypothesize that there are two patterns of manual results. Segmentations that have finer details along the boundary should be comparatively more detailed, while less careful segmentations tend to have a more compact lesion region. In this context, we categorize the manual results into two patterns (detailed *vs* compact) based on the compactness measurement defined as the ratio of the area of a circle (the most compact shape) having the same perimeter to the area of the shape, $compactness_j = \frac{perimeter_j^2}{4\pi \times area_j}$. For each manual segmentation, a compactness value is assigned. There are J manual results from different humans as $Compactness(Manual_{ij}), i = 1,\ldots,N, j = 1,\ldots,J$. Based on this value, J manual resources could be categorized into two patterns by $kmeans(k = 2)$.

3.3 Experiments

For 30 randomly selected test images, one dermatologist repeated the manual segmentation for 5 times on the images of the same lesion. Two trials were on the original orientation, while the other three are rotated clockwise by 90, 180, 270 degrees, respectively. As a result, we obtain 5 manual segmentations for each lesion image. The comparison results are shown in Table 2. The first row demonstrates the comparison result between the 2 non-

Measures ($\times 100$)		**XOR**	**FOM [1]**
Intra	No rotation (2 samples)	6.33	15.66
	Rotation (4 samples)	5.80	16.67
Inter	Other dermatologist (7 samples)	8.07	12.39

Table 2: Intra and Inter comparison

rotated segmentations from the same person. The second row compares the results drawn by the same person but on 4 images rotated every 90 degrees. They can be considered as the intra-person comparison since they are given by the same person and they reflect the randomness measure. The third row is the comparison results between different people. As it can be seen, the intra-differences are relatively small compared to the inter-difference. Hence, we hypothesize that the segmentation policy is the main factor that influences the segmentation rather than the randomness and slightly different segmentation policies lead to slightly different segmentations.

We find the pattern of the manual results by analyzing the compactness values of all the manual segmentations (50×8). For each image, the compactness of the 8 manual segmentations is calculated and categorized into two groups by *kmeans* and assigned with a class label (e.g., 1 for compact, 2 for detailed). Each dermatologist has a corresponding class vector recording how compactly they draw the lesion boundary over the 50 lesions. The mean and the standard deviation of the class label over the 50 lesions are shown in Table 3 (left), as well as the counts of the compact segmentation for each dermatologist.

The table shows 1) the dermatologists are reasonably consistent according to the standard deviation value. This means each dermatologist obeys the same rule when doing the manual segmentation. 2) There exist two patterns of segmentations according to the obvious

	Compactness				**Performance(STAPLE [6])**			
Doctor	counts for compact (out of 50)	mean group label	std	groups	precision		specificity	
1	26	1.48	0.50	detailed	0.9379	small	0.9890	big
2	37	1.26	0.44	compact	0.9578	big	0.9647	small
3	10	1.80	0.40	detailed	0.8417	small	0.9904	big
4	24	1.52	0.50	detailed	0.9095	small	0.9924	big
5	47	1.06	0.24	compact	0.9466	big	0.9794	small
6	35	1.32	0.47	compact	0.9437	big	0.9597	small
7	43	1.16	0.37	compact	0.9620	big	0.9821	small
8	41	1.18	0.39	compact	0.9220	small	0.9828	small

Table 3: Patterns of detailed versus compact segmentations

difference of the mean compactness. To get an idea of how well-separated the resulting clusters are, the silhouette values for each person using the cluster indices output from *kmeans* are calculated. The silhouette is a measure showing how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. The average value for the detailed group is 0.69 and 0.86 for the compact group. As can be seen, both clusters 'detailed' and 'compact' have measures significantly above 0, so the hypothesis of two segmentation patterns is confirmed. The above results are echoed by the performance parameter of each doctor from the STAPLE algorithm [6], as shown in Table 3 (right). The ones giving 'compact' segmentations normally have bigger precision (percentage of unhealthy skin area that is identified as lesion) and smaller specificity(percentage of healthy skin that is identified as skin) as they tend to include more tissue into the lesion area. The performance parameters (precision and specificity) are categorized into 'big' and 'small' groups using *kmeans*.

4 Conclusion

Based on the experiments with the manual segmentation results for lesion images, we conclude:
1 - computing the ground truth with the voting policy method is simple and effective and produces slightly better results compared to two other approaches based on optimization, although there is no significant difference between the three methods.
2 - It is reasonable to use $k = (J+1)/2$ as the voting threshold.
3 - There are generally two clusters of manual segmentations due to different segmentation policies. Hence, it would be reasonable to treat each cluster differently when computing the ground truth. In the future, we plan to investigate how to exploit this observation to produce better ground truth.
4 - We have also compared STAPLE [6] to our 3 algorithms and concluded that its ground truth is worse under the XOR criterion. However, STAPLE optimizes a different criterion and weights segmentations depending on the estimated performance level, so this comparison is not quite fair. In another paper, we will present results that demonstrate an improvement on STAPLE on a common criterion.
5 - The independence assumption of individual experts of method 2 needs further verification. Pixel label independence should be reconsidered in eqn 1, e.g., by introducing Markov random field modeling the relationship between each pixel and its neighbors.

References

[1] S. Chabrier, H. Laurent, B. Emile, C. Rosenberger, and P. Marche. A comparative study of supervised evaluation criteria for image segmentation. In *Proceedings of the 4th European Signal Processing Conference*, pages 1143–1146, 2004.

[2] Xiang Li, Ben Aldridge, Lucia Ballerini, Bob Fisher, and Rees Jonathan. Depth improves skin lesion segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention*, volume 2, pages 1101–1107, 2009.

[3] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. Still image objective segmentation evaluation using ground truth. In *Proceedings of the Fifth COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*, 2003.

[4] Torsten Rohlfing and Calvin R. Maurer, Jr. Shape-based averaging for combination of multiple segmentations. In *Proc. Medical Image Computing and Computer-Assisted Intervention*, volume 3750, pages 838–845, 2005.

[5] Ning Situ, Xiaojing Yuan, G. Zouridakis, and N. Mullani. Automatic segmentation of skin lesion images using evolutionary strategy. In *IEEE International Conference on Image Processing*, volume 6, pages 277–280, 2007.

[6] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[7] Luren Yang, Fritz Albregtsen, Tor Lønnestad, and Per Grøttum. A supervised approach to the evaluation of image segmentation methods. In *Proc. 6th Int. Conf. on Computer Analysis of Images and Patterns*, volume 970/1995, pages 759 – 765. Springer, 1995.

[8] Xiaojing Yuan, Ning Situ, and George Zouridakis. A narrow band graph partitioning method for skin lesion segmentation. *Pattern Recognition*, 42(6):1017–1028, 2009.