ESTIMATING THE GROUND TRUTH FROM MULTIPLE INDIVIDUAL SEGMENTATIONS INCORPORATING PRIOR PATTERN ANALYSIS WITH APPLICATION TO SKIN LESION SEGMENTATION

Xiang Li, Ben Aldridge, Robert Fisher, Jonathan Rees

University of Edinburgh, UK

ABSTRACT

Having ground truth is critical for evaluating segmentation algorithms and estimating the ground truth from a collection of manual segmentations remains a hard problem. A proper estimation approach should take into account and compensate for the inter-rater variation. In this paper, we conduct an analysis of manual segmentations in order to have a better understanding of the pattern of the variation and investigate whether incorporating such pattern information will improve the ground truth estimation. We propose a level-set based approach that solves the ground truth estimation in a probabilistic formulation. The prior pattern information is incorporated into the estimation model by adding a specially designed term in the energy function. Experiments on both synthetic and real data show that this prior information helps to find a more accurate estimate of the ground truth.

Index Terms— Segmentation evaluation, ground truth

1. INTRODUCTION

Segmentation is the first step of the computer-based skin lesion diagnosis algorithms and its accuracy is of crucial importance for the subsequent analysis. The evaluation of numerous computer-based skin lesion segmentation methods becomes necessary. Having ground truth (**GT**) is critical for the supervised evaluation, which considers the accuracy of the segmentation result as the degree to which the result corresponds to **GT** through evaluation metrics. Unfortunately, **GT** normally does not exist in practice and must be estimated as a compromise within a group of raters [3]. However, the inter-rater segmentations show a significant disagreement according to the rater's subjective criteria in placing the boundary [3, 5]. Hence, the question is raised as *how to compensate for this inter-rater variability*. To date, the most appropriate strategy to combine such segmentations is unclear and it has become a popular research topic in itself [7].

The **STAPLE** algorithm proposed by Warfield *et al.* [7] is so far one of the most referenced approaches in the field. **STAPLE** treated decision fusion as a maximum-likelihood problem and solved it using the expectation-maximization(EM) algorithm that guaranteed convergence, but not necessarily global optimality. **STAPLE** gave the quantitive estimate of the performance level parameters of raters in terms of the sensitivity and specificity and, based on which, it could output a probabilistic estimation of the **GT** simultaneously. However, Langerak *et al.* and Klein *et al.* highlighted that the performance of **STAPLE** was application dependent [6, 4]. It failed when the performances of the raters varied greatly. This can be explained by the fact that, even though fusing results in a weighted way, **STAPLE** takes into account all raters. A bad rater can contaminate the overall result, especially when an inappropriate initialization is allocated. In this context, the authors in [6] proposed a simplified **STAPLE** variant. This variant iteratively selected the optimal segmentation results based on image similarity measures and abandoned the ones with poor quality due to the wrong registration result. The final result was a combination of the optimal segmentations in a weighted Majority Vote procedure. The selection step helped to deal with the large variability problem encountered in their application and hence produced better results than **STAPLE**. Their approach required a large number of manual results because of their abandonment strategy and several parameters needed to be tuned in the iteration step, like the number of segmentations to be discarded and the similarity threshold. Moreover, the algorithm had no guarantee of convergence. It does however give us a hint as to whether or not a prior study of the segmentations would help. Hence, we hypothesize that a proper estimation of **GT** should take into account the segmentation bias pattern, which can later serve as $a\ priori$ information that guides the weighting of raters and drives the **GT** closer to the truth. This strategy has not been attempted in the related literature as far as the authors are aware.

In this paper, we conduct a pattern analysis of manual segmentations and then investigate whether incorporating such pattern information will improve the ground truth estimation. We represent the ground truth estimation as an optimization issue and propose a level-set based approach that maximizing the *a posteriori* probability (**MAP**) function. The performance of this method will be evaluated on both synthetic and real data. Some notations in the paper are listed as following:

$D_{ij}(x)$: the manual segmentation of the i^{th} image drawn by the j^{th} expert at pixel x

$T_i(x)$: the estimated ground truth of the i^{th} image at pixel x. $T_i(x) \in \{0, 1\}$ and [1: lesion, 0: normal skin]

I: the number of images; J: the number of raters

$\mathbf{P}(\Omega)$: the partition of the image Ω into N distinct regions: $\{\Omega_n\}_{n=1}^{N}, \cup_{n=1}^{N}\Omega_n \equiv \Omega$ and $\Omega_i \cap \Omega_j = \emptyset, \forall_{i\neq j} i, j$. N is the number of regions.

2. LESION MANUAL SEGMENTATION PATTERN ANALYSIS

We present a comprehensive assessment using carefully validated segmentations of 50 lesion images. The lesion boundaries 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 independently. We then convert the results into binary-valued images. To our knowledge, ground truth estimation for lesion segmentation analysis has not been studied on such a comparably large data set.

Visual inspection reveals the existence of both intra-rater and inter-rater variations when segmenting the same lesion, but the latter is more significant than the former according to a previous study [5]. Hence, we assume that the inter-rater variation is the main factor that differentiates the segmentations from different raters and should be compensated for during the ground truth estimation procedure. In order to account for the inter-rater variation, the authors in [8] considered the existence of two bias patterns as underestimation and overestimation and compensated for it through the estimated bias parameters. However, in lesion segmentation, using these two patterns is not enough, since the rater would either trace the lesion boundary exactly, or overestimate the boundary to an extent, but no underestimation exists. The estimated ground truth would still have some overestimation since all the results make a contribution to the ground truth. We question whether or not some other factors can help the **GT** to converge to a more accurate lesion boundary.

According to observation, lesion manual segmentations vary because of different rater's segmentation policy. For some of them, locating a general lesion region is necessary for a good diagnosis. Hence, they give 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. In this context, we assume that there are two patterns of manual results: **detailed** segmentations that have finer details along the boundary and less careful **compact** segmentations that tend to have a more compact lesion region. We categorize all the manual results drawn by 8 raters into two patterns based on two measurements: **Compactness measurement** ($CM = \frac{perimeter^2}{4\pi \times area}$) and **Fractal Dimension** ($FD = \frac{\log(N(s))}{\log(N)}$, N denotes the number of squares covering the image field), using the *k-means* clustering approach. For the purpose of comparison, both **CM** and **FD** values are normalized across 50 test images. The scatter plot of these two values is shown in *Fig.* 1. The results for a detailed style rater are highlighted in red; while the ones for a compact style rater are highlighted in green. *Fig.* 1 clearly illuminates a separation between two segmentation patterns (detailed v.s. compact). Furthermore, each rater keeps a consistent segmentation style over the 50 images. Hence, this pattern clustering result can be considered as useful prior information with potential value in estimating the ground truth. In our application, we consider a good quality lesion segmentation as one which has a small average distance from the true boundary. In this context, the **detailed** segmentation outperforms the **compact** segmentation and should be considered as more important.

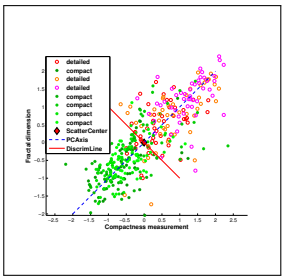


Fig. 1: The scatter plot of **FD** and **CM**.

3. GROUND TRUTH ESTIMATION METHODS

We treat the ground truth estimation as an optimization issue and propose a level-set based approach. The main advantages of using a level-set framework are: 1) the force that drives the evolution of the level set function has a physical interpretation, 2) it enables us to directly incorporate prior segmentation pattern information into the estimation framework by adding a specially designed term in the energy function E.

(A) Maximize the *a posteriori* (MAP) probability based method (*LSML*)

Our **MAP** based formulation estimates the **GT** as a process of finding the most possible partition $\mathbf{P}(\Omega)$ of the image domain under the observation of a set of manual results:

$$p(\mathbf{P}|D_{i\{1,2,\ldots,J\}}) = \prod_{n=1}^{2} p(\Omega_n|D_{i\{1,2,\ldots,J\}}) \quad (1)$$

$$= \prod_{n=1}^{2}\prod_{x\in\Omega_n} p(T(x)|D_{i\{1,2,\ldots,J\}}(x)). \quad (2)$$

$p(T(x)|D_{i\{1,2,\ldots,J\}}(x))$ is the conditional probability of the pixel x belongs to region $T(x)$ and it has the format:

$$p(T(x)=1|D_{i\{1,2,\ldots,J\}}(x)) = \frac{p(T(x)=1, D_{i\{1,2,\ldots,J\}}(x))}{p(D_{i\{1,2,\ldots,J\}}(x))} \quad (3)$$

$$= \frac{a}{a+b} = W. \quad (4)$$

$$p(T(x)=0|D_{i\{1,2,\ldots,J\}}(x)) = 1 - W = V. \quad (5)$$

where a and b are defined under the assumption that the raters perform the segmentation independently:

$$a = p(D_{i\{1,2,\ldots,J\}}(x)|T(x)=1)p(T(x)=1) \quad (6)$$

$$= \prod_{j=1}^{J} p(D_{ij}(x)|T(x)=1)p(T(x)=1) \quad (7)$$

$$b = p(D_{i\{1,2,\ldots,J\}}(x)|T(x)=0)p(T(x)=0) \quad (8)$$

$$= \prod_{j=1}^{J} p(D_{ij}(x)|T(x)=0)p(T(x)=0) \quad (9)$$

W and V are the joint conditional probability that pixel x belongs to the lesion and skin, respectively. The definition of the likelihood function for an individual rater is inspired by **STAPLE** and upholds the idea that the contribution of each rater to **GT** estimation differs based upon their performance (in terms of sensitivity ($sent_j$) and specificity ($spef_j$)):

$$p(D_{ij}(x)|T(x)=1) = sent_j \times sign(D_{ij}(x)) + (1-sent_j) \times (1 - sign(D_{ij})) \quad (10)$$

$$p(D_{ij}(x)|T(x)=0) = spef_j \times sign(1 - D_{ij}(x)) + (1-spef_j) \times (sign(D_{ij})). \quad (11)$$

The prior information term $p(T(x)$ is determined solely by the labeling results at x. Maximizing the *a posteriori* probability in Eq. 2 is equivalent to minimize its negative logarithm as

$$E = -\sum_{n} \sum_{x \in \Omega_n} \log p(T(x)|D_{i\{1,2,\ldots,J\}}(x)) \quad (12)$$

$$= -\{ \sum_{x \in \Omega_{lesion}} \log(W) + \sum_{x \in \Omega_{lesion}} \log(V) \} \quad (13)$$

The level-set representation of the above energy function can be expressed as

$$E(\phi) = -\int_{x \in \Omega} H(\phi(x)) \log(W) + (1 - H(\phi(x))) \log(V) dx, \quad (14)$$

where, $H(\phi)$ denotes the heaviside step function:$H(\phi) \equiv H(\phi(x)) = \begin{cases} 1 & \phi(x) \geq 0, x \in \Omega_{lesion} \\ 0 & \phi(x) < 0, x \in \Omega_{skin} \end{cases}$. The contour evolution equation is obtained by maximizing the energy functional using a gradient descent of the embedding function ϕ:

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E(\phi)}{\partial \phi} = \delta(\phi) \left(\log \frac{W}{V} \right). \quad (15)$$

$\delta(\phi) = \frac{dH(\phi)}{d\phi}$ is the Dirac delta function. It has value 1 at the lesion boundary and 0 elsewhere. The values of W and V keep updating iteratively until the boundary evolves to the location where the probabilities of that pixel belonging to the lesion and the skin are identical. If the conditional probability of pixel x being lesion is larger than skin, there is a positive force proportional to $log(W/V)$ driving the boundary to move towards the skin direction and vice versa.

(B) Maximize the *a posteriori* probability based method incorporating the segmentation pattern information (*LSMLP*)

As discussed in section 2, given the aim of comparing computer-based segmentations against the **GT**, it is reasonable to generate a **GT** that has a more accurate boundary. We remark that the **detailed** segmentations suit this requirement better. Hence, we introduce a Shape Prior Model (denoted as SPM) that is built upon the **detailed** manual segmentations using the *Majority Vote Rule* [5]. A shape prior based term aiming at minimizing the distance between the estimated T_i and SPM is formalized as:

$$E_{shape} = \int_{\Omega} [T_i(x) - SPM(x)]^2 dx \quad (16)$$

$$= \int_{\Omega} H(\phi)[1 - SPM(x)]^2 + [1 - H(\phi)][0 - SPM(x)]^2 dx \quad (17)$$

We add this term to the energy function of the **LSML** in Eq. 14 and lead to a new energy function as:

$$E_{LSMLP} = E_{LSML} + E_{shape}. \quad (18)$$

Minimizing the above energy function derives the boundary evolution equation as:

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E(\phi)}{\partial \phi} = \delta(\phi) \left(\log \frac{W}{V} + \gamma \times (2 \times SPM(x) - 1) \right). \quad (19)$$

Here, γ weights the importance of the shape prior energy (we set it to 0.4 in our experiment). In such a way, we give prominence to the detailed segmentations and reduce the impact of the compact segmentations. However, the above level-set formulations are based on two key assumptions: 1) Each rater independently perform the lesion segmentation job. 2) There is no spatial correlation between pixels. The second assumption can be relaxed by incorporating a Markov random field model as stated in [8], but it is out of the scope of our work.

4. EXPERIMENTS AND RESULTS

In this section, we will compare the proposed approaches against two popular ground truth estimation methods: the *Majority Voting Rule* (**MV**) and **STAPLE**, based on both synthetic and real lesion data. In order to compare different methods, we choose to use both *XOR* [2] and *FOM* [1] as our comparison tools. **XOR** measures the spatial-region-based dissimilarity between the real ground truth GT and the estimated ground truth EGT. It is defined as $XOR = \frac{Area(\mathbf{EGT} \bigoplus \mathbf{GT})}{Area(\mathbf{EGT}+\mathbf{GT})}$, with a range from 0 (best) to 1 (worst). $\bigoplus$ denotes exclusive-OR; $+$ means union. The smaller the XOR, the closer the ground truth is to the manual results. **FOM** (Pratt's Figure Of Merit) is a dissimilarity measure that is often used to compare the performance of edge detection algorithms [1]. It corresponds to an empirical contour distance between the GT and EGT in the form of $FOM(CoT, CoD) = \frac{1}{max\{length(CoT), length(CoD)\}} \cdot \sum_{k=1}^{length(CoT)} \frac{1}{1+d^2(k)}$, where CoT and CoD are the boundary representations of the GT and EGT. $d(k)$ is the Euclidean distance between the k^{th} pixel of CoT and the nearest pixel of CoD. It ranges from 0 (worst) to 1 (best).

Comparison on Synthetic Data

In order to compare the estimated **GT**s derived from different approaches, we generate synthetic data that simulates the two patterns of manual segmentations. The data is derived by using a selected computer segmentation as the ground truth that is represented as a level set function as ϕ. Developing the synthetic data is compiled as the evolution of ϕ. The force that drives the evolution of the level set function takes into account both systematic and random errors. The formulation of this process is as following:

$$\frac{\partial \phi}{\partial t} = Norm \times F = Norm \times (Random + \nu div \frac{\nabla \phi}{|\nabla \phi|}). \quad (20)$$

$Norm$ is the normal to the curve and can be determined directly from the level set function as $Norm = -\frac{\nabla \phi}{|\nabla \phi|}$. F is the force comprised of two terms: 1) The $Random$ term simulates randomness errors. A uniformly distributed pseudorandom value ranging from -1 to 1 is assigned to it. 2) The second term is a regularization term related to the smoothness of the evolving contour. ν denotes the weight. A larger weight is used to simulate compact segmentation; while a smaller

	Methods			
Metrics	**MV**	**STAPLE**	**LSML**	**LSMLP**
XOR (%)	3.8409	3.7212	3.2733	2.1615
FOM (%)	8.9026	10.6596	13.1484	26.7412
Sensitivity	1.0000	1.0000	1.0000	1.0000
Specificity	0.9709	0.9719	0.9754	0.9839

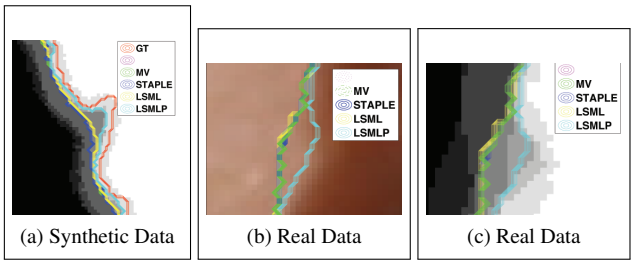
Table 1: Performance of Ground Truth estimation methods.

weight is used for detailed segmentation. Moreover, overestimation is simulated using a morphological operation: dilation. The scale of the dilation structure differs between **detailed** and **compact**, smaller for the former and larger for the latter.

Our synthetic data are comprised of 4 **detailed** and 4 **compact** segmentations. The **GT**s estimated from the synthetic data are displayed in *Fig.* 2a. The background grey image is generated by aggregating individual rater binary segmentations and it provides a visual representation of rater agreement. The comparison results using *XOR, FOM, sensitivity and specificity* metrics are demonstrated in Table. 1. These results show: 1) The **GT** estimated using **LSMLP** is the closest to the real ground truth. The improvement is significant compared to the other approaches according to all four metrics. The method outperforms the others mainly because it produces finer boundary details, especially at the locations where two groups of segmentations have big differences, such as those shown in *Fig.* 2a. 2) **LSML** produces the second best result and **STAPLE** comes the third, though there is no significant difference between them. **MV** is the worst because it only produces a comprise that minimizes the average discrepancy between estimated ground truth and the manual results without taking into account individual performances.

Comparison on Real Lesion Data

We also apply the approaches on the real lesion data, examples of which are shown in *Fig.* 2b and *Fig.* 2c. The same



(a) Synthetic Data (b) Real Data (c) Real Data

Fig. 2: Test on both synthetic and real data

conclusion holds. The **LSMLP** outperforms the others at the locations where the boundary is non-convex and missed in the **compact** segmentations. Note, for fair comparison on both synthetic and real data, the three iteration-based approaches (**STAPLE**, **LSML** and **LSMLP**) are initialized with the same setting: an initial circular boundary covering the lesion.

5. CONCLUSION AND FUTURE WORK

Having ground truth is critical for evaluating segmentation algorithms and finding the ground truth remains a hard problem. A good **GT** estimation algorithm should take into account inter-rater variability that appear in manual segmentations. Little research has analyzed the patterns of the manual segmentation results and we are the first group to study this subject. We found that the manual segmentations of lesion differed mainly because of the rater's segmentation policies and could be categorized into two groups: **detailed** and **compact**. Using the categorization result as prior information, we introduce a shape prior model that is built upon the **detailed** segmentations. We then treat ground truth estimation as an optimization problem and solve it under a level-set framework based on a **MAP** formulation. The rater's pattern is incorporated by adding an energy term related to a shape prior model into **LSML** and results in **LSMLP**. Experiments on both synthetic and real lesion data reveal that **LSMLP** outperforms all the other methods that do not consider the prior information, followed by **LSML** and **STAPLE**.

Future work will concentrate on: 1) generate the prior shape model in a more comprehensive way, *e.g.,*based on principal components analysis (PCA). 2) relax the assumption that pixels have a spatial independence by introducing a Markov random field model as stated in [7]. 3) extend this work into multiple phase applications, *e.g.,*multiple lesions in one image.

References

[1] S. Chabrier a d H. Laurent, B. Emile, C. Rosenberger, and P. Marche. A comparative study of supervised evaluation criteria for image segmentation. *14th European Signal Processing Conference (EUSIPCO 2006)*, pages 1143–1146, 2006.

[2] M. Emre Celebi, Hitoshi Iyatomi, Joseph M. Malters, Gerald Schaefer, William V. Stoecker, and James M. Grichnik. An improved objective evaluation measure for border detection in dermoscopy images. *Skin Research and Technology*, 15(4):444–450, 2009.

[3] Joel G, Schmid-Saugeon P, Guggisberg D, Cerottini JP, Braun R, Krischer J, Saurat JH, and Murat K. Validation of segmentation techniques for digital dermoscopy. *Skin Research and Technology*, 8(4):240–249, 2002.

[4] Stefan Klein, Marius Staring, and Josien P. W. Pluim. Automatic segmentation of the prostate in 3d mr images by atlas matching using localized mutual information. *Medical Physics*, 35(4):1407–1417, 2008.

[5] Xiang Li, Ben Aldridge, Jonathan Rees, and Robert Fisher. Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation. *Medical Image Understanding and Analysis (MIUA)*, 1(1):101–106, 2010.

[6] T.R.Langerak, U.A. van der Heide, I.M. Lips, A.N.T.J. Kotte, M. van Vulpen, and J.P.W. Pluim. Label fusion using performance estimation with iterative label selection. *IEEE International Symposium on Biomedical Imaging ISBI 2009*, pages 1186 – 1189, 2009.

[7] 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.

[8] Simon K Warfield, Kelly H Zou, and William M Wells. Validation of image segmentation by estimating rater bias and variance. *Phil. Trans. R. Soc. A*, 366(1874):2361–2375, July 2008.