Restorable synthesis: average synthetic segmentation converges to a polygon approximation of an object contour in medical images

Shuyue Guan, Ravi K. Samala, Seyed M. M. Kahaki, Weijie Chen
U.S. Food and Drug Administration
Center for Devices and Radiological Health
Office of Science and Engineering Laboratories
Division of Imaging, Diagnostics, and Software Reliability
Silver Spring, Maryland, U.S.
{shuyue.guan, ravi.samala, seyed.kahaki, weijie.chen}@fda.hhs.gov

***Abstract*—Synthesis of segmentation contours is useful in evaluating truthing methods, *i.e.*, the establishment of a segmentation reference standard by combining multiple segmentation results (*e.g.*, by multiple experts). In contrast to a real-world application where the ground truth is often not available, the ground truth of objects is defined in synthetic data. Contours with combinations of segmentation errors, as compared to the defined ground truth, can be synthesized. A desired property of segmentation contour synthesis for evaluating truthing methods, which we call the restorability property, is that the average of multiple segmentation contours can converge to the truth contour. This property is desired because such a dataset can serve as a benchmark for evaluating if commonly used truthing methods have bias. We developed a segmentation contour synthesis tool that has the restorability property and conducted simulation studies to validate this tool.**

***Index Terms*—synthetic segmentation, segmentation synthesis, medical image segmentation, restorable segmentation, restorability**

I. Introduction

Medical image segmentation is commonly employed to determine the boundaries of anatomical structures in medical images, such as organs or lesions. This technique has numerous clinical uses, including extracting features for diagnostic purposes, designing treatment plans in radiation therapy, and tracking tumor growth in response to treatment, among others. Despite the rapid development of advanced artificial intelligence and machine learning (AI/ML) algorithms for medical image segmentation [1], there is a lack of consensus on evaluation methods for image segmentation. Many metrics for evaluating segmentation performance have been proposed in the literature and guidelines are needed for selecting the most appropriate metrics for a particular clinical task [2], [3]. There are many truthing methods, *i.e.*, the establishment of a segmentation reference standard usually by combining multiple segmentation results (*e.g.*, by multiple experts) [4], [5], and the assessment and comparison of these truthing methods need more research.

The ground truth segmentation of medical images is useful for investigating the characteristics of performance metrics, guiding the choice of metrics, comparing and aiding the selection of truthing methods. However, the ground truth is not available for real-world images, but can be defined in synthetic data. Prior efforts [3], [6] have utilized simple geometric contours such as circles in synthetic data and have been shown to be effective in demonstrating certain characteristics of performance metrics. However, simulation of simple contours do not capture the complexity and variability in anatomical structures like organs and lesions. The use of manually adjusted segmentations is often cost prohibitive. While synthesizing segmentation contours using deep learning techniques appears to be an alternative solution, it would need large amount of training data and it can be difficult to generate contours with specific type of segmentation errors. Hence, the aim of this study is to devise a tool that generates synthetic segmentation derived from anatomies identified in actual medical images with known reference boundaries. The goal of this tool is to aid in the evaluation of AI/ML segmentation metrics in medical imaging.

Our previous works developed a *Medical Image Segmentation Synthesis tool* (MISS-tool) [7] to generate synthetic segmentation contours with defined truth masks based on objects in real-world medical images. The MISS-tool allows users to customize segmentation errors by configurable parameters. Emulated segmentation by using MISS-tool are used to inform the selection of performance metrics for medical image segmentation evaluation [8]. The synthetic segmentation contours are generated from truth masks by setting specified parameters to simulate certain types of segmentation errors. For one truth mask, tool parameters can be varied to create an arbitrary number of synthetic segmentation contours; however, the average of those contours are not guaranteed to converge to the truth contour.

A desired property of segmentation contour synthesis for

evaluating truthing methods is that the average of multiple segmentation contours do converge to the truth contour. This property is desired because such a dataset can serve as a benchmark for evaluating if commonly used truthing methods have bias, *i.e.*, the reference standard for a truthing method is systematically over- or under-segmentation of the truth. Therefore, in this paper, we proposed a method that generates restorable synthetic segmentation contours that converge to the truth contour.

II. Method

The principle idea of our approach follows the law of large numbers which describes that the average of observations of a Gaussian random variable converges to its mean. For example, if we synthesize (simulate) measurements of the largest diameter of a lesion, we can set the truth of diameter as μ and synthesize the i-th measurement as $l_i = \mu + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is zero mean Gaussian noise. The average of l_i converges to the truth μ. However, the synthesis of contours analogous to this simple idea is not trivial. We show below that our approach can restore the polygon approximation of the truth contour using synthesize contours.

A. Restorable synthesis

We represent the truth contour in a 2D digital image with the Cartesian coordinates of N points on the contour $P_i(x_i, y_i), i = 1, 2, .., N$. We first consider adding Gaussian noise to the coordinates of one point. For a point $P_1(x_1, y_1)$, it is adjusted to a new location $P_1^1(x_1 + \epsilon_1, y_1 + \epsilon_2)$ after adding Gaussian noise of zero mean: $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \sigma_i^2)$. The zero mean of the Gaussian distribution ensures that the expectation of new coordinates is the original coordinate. Obviously, the location of P_1^1 follows a 2-D Gaussian distribution with a mean at the original location P_1. We consider an image (I_1) containing one-pixel object with value of 1 and background pixels' value of 0. We generate n synthetic images $(I_1^1, I_1^2, I_1^3, \cdots I_1^n)$ by adding Gaussian noise to the coordinates of that pixel point (Figure 1 upper left). In the average image of the n synthetic images $(\frac{1}{n}\sum_i I_1^i)$, the pixel with the maximum value is expected to be located at the location of the point object in the original image (I_1). Figure 1 illustrates that the original location of the point object is the most probable location for the point object in the synthetic images and the maximum value of the average image is located in the original point.

If we add Gaussian noise to the locations of all points on a contour, it may break the contour into unconnected pieces. To overcome this issue, we select some key points based on a rule on the contour and connect them into a polygon approximation of the ground truth contour. We then synthesize segmentation contours by adding Gaussian noise to the locations of the key points, and then reconnect them using straight lines. To avoid the crossing of lines in the synthesized contour, the moving distance of the key points (which is controlled by the σ of the Gaussian noise) needs to be restricted.

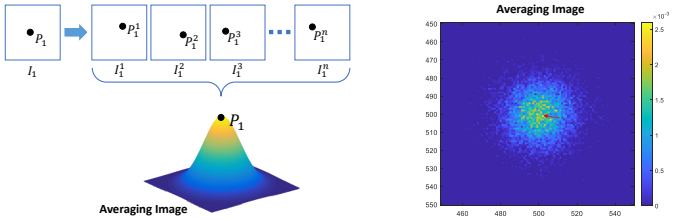

Fig. 1: Synthesis of images with a point object. Upper left shows the original and synthesized images. Lower left shows the probability distribution for the location of the point object in the synthetic images. The plot on the right shows the average image with the expected point indicated with an arrow.

In summary, our method to generate synthetic segmentation is:

1) Select key points on the original contour (Figure 2-1).
2) The polygon approximation of the truth mask is created by connecting the selected key points using straight lines and filling the closed area (Figure 2-2)
3) Add Gaussian noise to the coordinates of the key points with limited σ_t (Figure 2-3)
4) The synthetic segmentation is generated by connecting the modified key points using straight lines and filling the closed area (Figure 2-4)

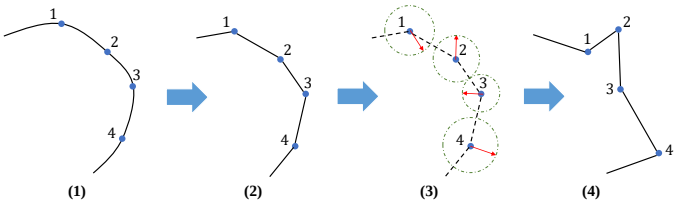

Fig. 2: Illustration of our method to synthesize segmentation contours that converge to a polygon approximation of the original truth contour.We only show a part of the contour; all shapes are closed in real cases. (1) Truth mask with selected key points; (2) Polygon approximation; (3) Adding Gaussian noise to the key points. The green circles show the probable locations of each key point, which is define here as $3\sigma_t$; (4) Synthetic segmentation contour.

B. Restoration of polygon

An important property of the contour synthesis method as described above is that the average of the synthesized contours asymptotically converge to the polygon approximation. Specifically, we represent the synthesized segmentation contour as a binary mask with pixel value of 1 for the object and 0 for the background. Then applying a threshold of 0.5 on the the average of infinite number of such images would be the polygon contour. This is shown in Figure 3. Note that the maximum value on the average image is 1, but unlike the single-point image situation where the maximum value corresponding to the original truth location, a threshold of 0.5 must be used to

converge to the polygon contour. This is because any (pixel) point on the polygon contour has 0.5 probability of being included/excluded by the synthetic segmentation. The value of the pixel inside the synthetic segmentation is 1, and outside the synthetic segmentation is 0. Hence, the expectation of the value of a pixel in the polygon contour should be equal to or greater than $1 \times 0.5 + 0 \times 0.5 = 0.5$. That explains why pixels on the boundary of the polygon contour has value of 0.5 in the average image of synthetic segmentation masks.

III. Experiment and Result

The critical characteristic of our proposed method is that averaging the synthetic segmentation contours generated from a defined polygon contour can approximately reproduce the polygon contour. As mentioned in the Introduction section, this would provide a benchmark dataset for evaluating truthing methods. Thus, in the experiments, we investigate how well the polygon contour can be restored by averaging the synthetic contours. We also investigated the diversity of synthetic segmentation contours because the method controls the amount of variation to avoid crossing of the polygon sides. In this work, we used the Dice index [9] as a segmentation performance metric.

In the Method Section II-A, we summarized the major steps of our method to generate synthetic segmentations. Here, we describe the methods for key points selection on the original contour to generate a polygon approximation and for specifying the extent of the Gaussian noise (the σ_i parameter) for each key point.

A. Parameter selection

In this study, the key points (*i.e.*, the vertices of the polygon) are selected sequentially on the contour of interest by a constant interval. Specifically, the gap between two key points has the same number of pixels on the contour with the gap between the first and last selected key points potentially being different from others. Thus, the number of key points (k) depends on the interval/gap (g) and total number of pixels on the contour (N); that is: $k = [N/g]+1$. The minimum number of key points is three to define a polygon. This requires the gap: $g \leq \lfloor (N-1)/2 \rfloor$. A large gap value is not recommended because it may render the polygon to be far off the original contour.

The σ_t parameter that defines the Gaussian noise to be added to the locations of the key points (*i.e.*, polygon vertices) depends on the Euclidean distance between two key points. For a key point P_t, its two neighboring key points are P_{t-1} and P_{t+1}. The Gaussian noise added to P_t is:

$$\sigma_t = \frac{w}{3} \lfloor \min\{d(P_{t-1}, P_t), d(P_t, P_{t+1})\} \rfloor$$

where $w \in [0, 1]$ is the weight of σ and d is the distance in pixels along the contour. Dividing by 3 is because the radius of the point changing area is about $3\sigma_t$ (covering 99.7% of the Gaussian distribution).

The extent of variation of the polygon vertices to generate a synthetic segmentation is determined by the values of gap (g) and weight (w). Since the gap is set to be identical for a contour, the larger gap allows for larger σ_t values. We vary the w parameter between 0 and 1.

B. Materials

In this preliminary study, the original contour (the blue contour in Figure 3A) is a lung nodule segmentation from the LIDC-IDRI dataset [10], which consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with radiologist-annotated lesions.

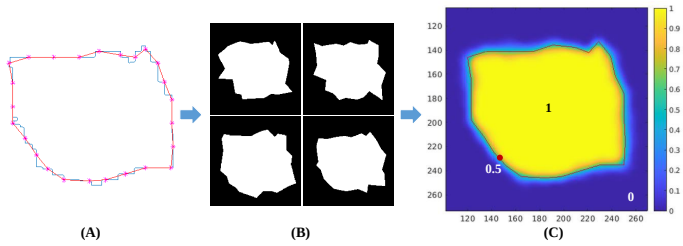

Fig. 3: Synthetic segmentation contours from a lung nodule segmentation and the convergence to the polygon approximation. In (A), the blue contour is the lung nodule segmentation by a radiologist, and the red contour is the polygon approximation. (B) shows four examples of synthetic segmentations generated from the polygon mask. (C) is the average image through 1000 synthetic segmentations. The red contour in (C) is a close aprpximation of the same as the polygon contour in A, which is identified with pixel value threshold of 0.5.

C. Restorability

The first experiment is to show the restorability, *i.e.*, the ability to reproduce the initial polygon contour by averaging the synthetic segmentation masks. From the lung nodule mask, we created a polygon approximation by using the gap: $g = 20$. Since the contour of the lung nodule contains 500 pixels, 25 key points were selected as the polygon vertices (red star-marks in Figure 3A). We then used the weight $w = 1$ and generated five groups of synthetic segmentations including 1k (1000), 2k, 3k, 5k, and 10k images respectively (Figure 3B).

As shown in the fourth column of Table I, the Dice indexes between the average of the synthetic masks and the initial polygon mask are greater than 99.5% for all groups. Figure 3C shows the result for group 1 with 1k synthetic masks. Other columns of Table I show that the mean of Dice indexes between synthetic segmentations and the initial polygon mask is stable when the number of generated synthetic segmentations is over 1000.

D. Variability

It is important that the synthetic segmentations have certain level of variability to mimic the real-world situations. Here we examine how the two parameters gap (g) and weight (w) affect the synthetic contour's variability. We generated eight groups of synthetic segmentation by setting $g = 20, 10$ and $w = 1.0, 0.8, 0.5, 0.3$. Each group includes 1000 images.

TABLE I: The second column is the mean of Dice indexes for comparing synthetic segmentation masks with the initial polygon mask. The third column is the standard deviation (std) of these Dice indexes. The fourth column is the Dice indexes between the initial polygon mask and the reproduced mask by averaging synthetic segmentations.

#	Dice%(mean)	Dice%(std)	**Dice%(Restored)**
1k	94.5376	0.8733	**99.5840**
2k	94.5086	0.9150	**99.6461**
3k	94.5202	0.8886	**99.6542**
5k	94.5171	0.8894	**99.6583**
10k	94.5154	0.8992	**99.6624**

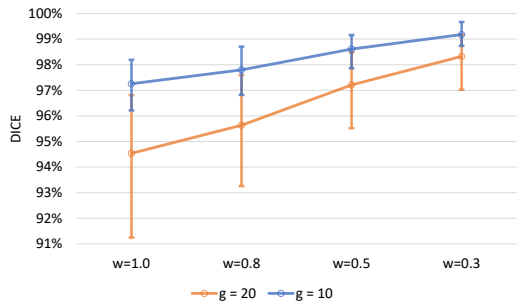

Fig. 4: Dice between the initial polygon mask and synthetic segmentations generated by different gap (g) and weight (w) parameters. The top and bottom of the bars are the maximum and minimum of Dice indexes to show the range of Dice indexes through 1000 synthetic segmentations.

Figure 4 shows that a larger gap (g) results in a lower mean Dice and a greater range (variability) of synthetic segmentations. And weight (w) can also control the variability. Specifically, a smaller weight results in a higher mean Dice and a narrower range (variability) of synthetic segmentations. Thus, the variability of generated synthetic segmentation is managed by these two parameters for a given mask.

IV. Discussion

One limitation of our method is that we set the gap parameter as a constant for a given case when creating the polygon approximation of an object contour in a medical image. Our future work will include investigating an alternate approach by setting the gap parameter to be adaptive to the spatial frequency of the contour, *i.e.*, a larger gap value for a flat portion of the contour and a smaller one for a sharp corner.

To evaluate truthing methods, we can apply selected truthing methods to a set of restorable synthetic segmentations generated from the polygon approximation of a truth mask, then compare fusion results from truthing methods with the polygon approximation. The better performing truthing method should yield an estimate of the truth segmentation that has higher accuracy (*e.g.*, greater Dice score) as assessed by the polygon approximation and/or use fewer segmentations to reach that accuracy. Another approach in a future work could be application of image-to-image models like CycleGAN to transform the synthetic masks to nodule images as an augmentation for training models with better generalizability.

V. Conclusions

In this paper, we proposed a method that can generate synthetic segmentation contours from an initial polygon mask, which approximates the actual object contour in a medical image, with the property that averaging these synthetic segmentations can restore the polygon approximation. Using the Dice index, we verified that the synthetic segmentations generated using our approach converge to the polygon mask. We also showed that he synthesis can generate certain amount of variability in the synthetic segmentation contours. This approach can allow for the creation of a benchmark dataset that includes synthetic segmentation contours with the property that their average converges to the polygon truth mask. Such a dataset would be useful for evaluating truthing methods in medical image segmentation.

Acknowledgment

The authors would like to thank Nicholas Petrick, Ph.D. for providing helpful comments on the manuscript. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright.

References

[1] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[2] A. A. Taha, A. Hanbury, and O. A. J. d. Toro, "A formal method for selecting evaluation metrics for image segmentation," in *2014 IEEE International Conference on Image Processing (ICIP)*, Conference Proceedings, pp. 932–936.

[3] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, pp. 1–28, 2015.

[4] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE transactions on medical imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[5] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: Is the problem solved?" *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 541–552, 2017.

[6] H. Kim, J. I. Monroe, S. Lo, M. Yao, P. M. Harari, M. Machtay, and J. W. Sohn, "Quantitative evaluation of image segmentation incorporating medical consideration functions," *Medical physics*, vol. 42, no. 6Part1, pp. 3013–3023, 2015.

[7] S. Guan, R. K. Samala, A. Arab, and W. Chen, "Miss-tool: medical image segmentation synthesis tool to emulate segmentation errors," in *Medical Imaging 2023: Computer-Aided Diagnosis*, vol. 12465. SPIE, 2023, pp. 273–281.

[8] S. Guan, R. K. Samala, and W. Chen, "Informing selection of performance metrics for medical image segmentation evaluation using configurable synthetic errors," in *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 2022, pp. 1–8.

[9] A. Carass, S. Roy, A. Gherman, J. C. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel *et al.*, "Evaluating white matter lesion segmentations with refined sørensen-dice analysis," *Scientific reports*, vol. 10, no. 1, p. 8242, 2020.

[10] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, L. P. Clarke *et al.*, "Data from lidc-idri. the cancer imaging archive," *DOI http://doi. org/10.7937 K*, vol. 9, no. 7, 2015.