

Qualitative Analysis of Skull Stripping Accuracy for MRI Brain Images

Shafaf Ibrahim, Noor Elaiza Abdul Khalid, Mazani Manaf
and Mohd Ezane Aziz

Abstract Skull stripping isolates brain from the non-brain tissues. It supplies major significance in medical and image processing fields. Nevertheless, the manual process of skull stripping is challenging due to the complexity of images, time consuming and prone to human errors. This paper proposes a qualitative analysis of skull stripping accuracy for Magnetic Resonance Imaging (MRI) brain images. Skull stripping of eighty MRI images is performed using Seed-Based Region Growing (SBRG). The skull stripped images are then presented to three experienced radiologists for visual qualitative evaluation. The level of accuracy is divided into five categories of “over delineation”, “less delineation”, “slightly over delineation”, “slightly less delineation” and “correct delineation”. Primitive statistical methods are calculated to examine the skull stripping performances. In another note, Fleiss Kappa statistical analysis is used to measure the agreement among radiologists. The qualitative performances analysis proved that the SBRG is an effective technique for skull stripping.

Keywords Qualitative analysis • Skull stripping • Seed-based region growing • Medical imaging • Magnetic resonance imaging

S. Ibrahim (✉) · N. E. A. Khalid · M. Manaf
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, 40450 Selangor, Malaysia
e-mail: shafaf_ibrahim@yahoo.com

N. E. A. Khalid
e-mail: elaiza@uitm.edu.my

M. Manaf
e-mail: mazani@uitm.edu.my

M. E. Aziz
Department of Radiology, Health Campus, Universiti Sains Malaysia,
Kubang Kerian 16150 Kelantan, Malaysia
e-mail: drezane@kb.usm.my

1 Introduction

Medical imaging modalities such as X-ray, Computed Tomography (CT) scan, Ultrasound and Magnetic Resonance Imaging (MRI) has allowed non invasive insight into human internal organs. It has made it possible to visualize and observe various organs and cells structures, their function, detect abnormalities or dysfunction as well as assist in pathologic diagnosis [1]. The brain which is one of the most complex, least accessible and prone to complex abnormalities can be expressed in variety of complexity scales [2] is the primary beneficiary of these medical imaging techniques. Deeper understanding of the brain anatomical structures plays crucial role in improving brain lesions and diseases detection [3].

Skull stripping is an important pre-processing step for the analysis of neuroimaging data and MRI images [4, 5, 6]. It refers to the process of delineation and removal non-cerebral tissue region such as skull, scalp and meninges from the brain soft tissues [7]. The accuracy in skull stripping process affects the efficiency in detecting tumor, pre-surgical planning, cortical surface reconstruction and brain morphometry [8], and has been considered as an essential step for brain segmentation [9]. Removal of the skull region reduces the chances of misclassifying diseased tissues [10]. The process of skull stripping is poses some challenge due to the complexity of the human brain, variability in the parameters of Magnetic Resonance (MR) scanners and individual characteristics [11]. Poor quality and low contrast images also contribute to difficulties in segmenting the images precisely [10].

From the reviews done, it is presumed that accurate and reliable quantification of the skull stripping outcomes is one of the biggest challenges in the medical imaging domain [4]. Up until now, only a few evaluation criteria have been proposed to quantify the quality of skull stripping outcomes [6]. The common standard used for validating skull stripping is manual delineation which acted as a ground truth where the skull stripping outcomes is compared [12]. Manual delineation which still considered as gold standard [13] is a tedious task, time consuming and subjective due to inter and intra-expert variability [14].

A main issue is that obtaining these validation data and comparison metrics for skull stripping are difficult tasks due to the lack of reliable ground truth [15]. Thus, even if a rich set of manual delineations are available, they may not reflect the ground truth and the true gold standard may need to be estimated [16]. In addition, the subjectivity of human decisions could also introduce inaccuracies and inconsistencies [6].

Thus, this research investigates the accuracy of the proposed techniques, Seed-Based Region Growing (SBRG) segmentation results through a qualitative evaluation of three experienced radiologists. The non-cerebral tissue region are delineated, segmented and removed using SBRG. Then the resulting images are presented to the radiologist for performances assessment. The proposed qualitative evaluation technique is expected to offer a new way of skull stripping evaluation in MRI brain images.

The organization of the rest of this paper is as follows: [Sect. 2](#) presents our materials and methods, including the overview of SBRG methods and descriptions

of the qualitative evaluation method proposed. The results and discussions are discussed in Sect. 3. Finally, we present our conclusion in Sect. 4

2 Materials and Methods

Eighty axial sequence of Fluid Attenuated Inversion Recovery (FLAIR)-MRI of brain normal and abnormal slices were acquired from the Hospital Sungai Buloh, Selangor, Malaysia. The MRI brain images criteria are limited to adult male and female (with their age ranging between 20 and 60 years).

2.1 The Seed-Based Region Growing Algorithm

The skull stripping process is performed using a Seed-Based Region Growing (SBRG) algorithm [17, 18], which developed using a Borland C++ Builder 6.0. SBRG is very attractive especially for semantic object extraction as well as image applications. Furthermore, SBRG algorithm is observed to be successfully implemented in various applications of medical images [18].

The process of SBRG begins by selecting a seed pixel which is located within the area of delineation. This seed grows iteratively into neighboring pixels of window size 3×3 pixels to produce a region with similar mean values. The mean value, M for the $M \times M$ neighborhood is calculated as in (1).

$$\text{Mean } (M) = \frac{\sum \text{grey level pixels value in } M \times M \text{ neighborhood}}{\sum \text{number of pixels in } M \times M \text{ neighborhood}} \quad (1)$$

For every growth from the seed pixel to one of its neighbors, the calculated mean value, M and the grey level of the particular neighbor, G_j is compared using (2).

$$|G_j - M| < T \quad (2)$$

If the absolute difference of the two pixels is less than a pre-defined threshold, T the neighbor pixel will be included into the growing region. The predefined threshold, T is set to 10. The mean value is updated constant while the growing process is recursively iterated until no neighboring pixels are found.

2.2 Qualitative Evaluation Method

Unsupervised qualitative evaluation method is employed for the skull stripping accuracy evaluation. A group of three experienced radiologists is requested to visually analyze the accuracy of 80 skull stripping images produced by the SBRG.

The accuracy level of skull stripping assessment is divided into five categories which are less delineation, slightly less delineation, correct delineation, slightly over delineation, and over delineation as elaborated in Table 1.

Table 1 Accuracy level of skull stripping assessment

Accuracy level	Weightage	Description	Visual indicator
Over delineation	1	>30 % of brain tissue cut	include elimination of cerebral cortex
Less delineation	2	>30 % of residual skull	include elimination of skin, skull and dura mater
Slightly over delineation	3	<30 % of brain tissue cut	include elimination of pia mater
Slightly less delineation	4	<30 % of residual skull	include elimination of skin, skull, dura mater and arachnoid mater
Correct delineation	5	all non-cerebral tissue region are removed	include elimination of skin, skull, dura mater, arachnoid mater and subarachnoid space

Table 2 Significance of fleiss kappa value

Kappa value	Significance
< 0	Poor agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost Perfect agreement

Based on the assessment conducted, the performances of skull stripping are then evaluated. Each level of accuracy mentioned is assigned to a weightage based on its significance. The weightage values are significant as it will be used further in the qualitative statistical analysis.

The reliability of agreement among the radiologists is observed. It is used to monitor the consistencies among the radiologists in analyzing the skull stripping analysis. A statistical analysis method known as Fleiss Kappa is employed. Fleiss Kappa is defined as a useful statistical measure for assessing the reliability of agreement between a number of raters when assigning categorical ratings to a number of items or classifying items [19].

Finally, the significance of agreement between the raters is identified based on the Fleiss Kappa values calculated. Richard and Gary [20] in their study summarized that the significance of agreement of Fleiss Kappa can be divided into several categories according to its range values as tabulated in Table 2.

3 Results and Discussion

The accuracy of skull stripping among the radiologists is measured by observing the mode value for level of accuracy rated by each radiologist. From the overall analysis conducted, the percentage of accuracy is calculated using (3):

$$\%Accuracy = \frac{Rated\ Weightage}{Best\ Weightage\ Value * No.\ of\ Data\ Images} \tag{3}$$

Table 3 Skull stripping accuracies among radiologists

	Radiologist 1	Radiologist 2	Radiologist 3
Mode	(5)	(5)	(5)
% Accuracy	97	95.3	92.5
Standard Deviation		0.287	

Table 4 Qualitative performances review for radiologists

RADIOLOGISTS	Weightage				
	1	2	3	4	5
No. of Occurrence	0	0	10	39	191
% of Occurrence	0	0	4.2	16.3	79.6

The variation results among the radiologists are then evaluated using standard deviation. All modes, percentage of accuracy and standard deviation for radiologists produced are tabulated in Table 3.

From Table 3, it can be monitored that the mode values for all radiologists return the value of 5 (correct delineation) level of accuracy. Moreover, the percentage of accuracy for each radiologist is noted to show good and consistent performance as it produced 97, 95.3 and 92.5 % for Radiologist 1, Radiologist 2 and Radiologist 3 respectively. The overall standard deviation value among the radiologists is seen to be at a low rate of 0.287, which verifies a strong consistency of agreement among the radiologists.

Table 4 tabulates the break review of qualitative performances for radiologists where the total occurrence for each weightage value is counted. The percentage of occurrence is evaluated using (4):

$$\% \text{ of Occurrence} = \frac{\text{No. of Weightage Occurrence}}{\text{No. of Raters} * \text{No. of Data Images}} \tag{4}$$

From the Table 4, it is noticeable that radiologists return a highest total occurrence of weightage 5 (correct delineation) which is 191 numbers of occurrence. The weightage 4 (slightly less delineation) is also cannot be underestimated as they produced a good numbers of occurrence too which is 39. The total occurrence is then followed by weightage 3 (slightly over delineation) which is 10 numbers of occurrences. No occurrence of weightage 1 (over delineation) and weightage 2 (less delineation) are reported.

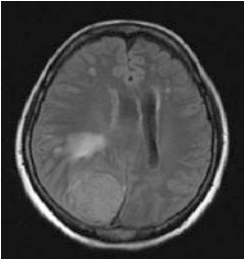
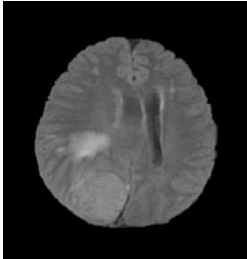
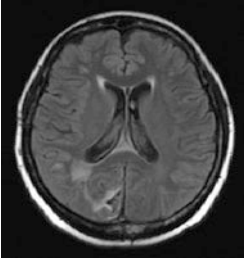
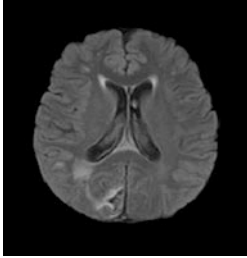
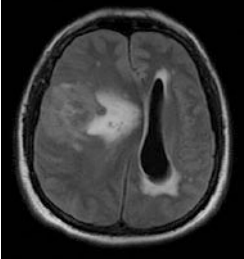
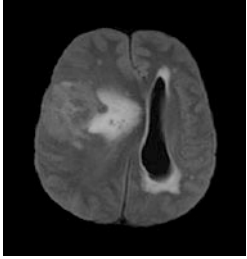
Next, the reliability of agreement among the radiologists in analyzing the skull stripping performances and its significance is identified using Fleiss Kappa Analysis as tabulated in Table 5.

Referring to Table 5, the Kappa value for the radiologists is found to be fairly high at 0.686 which is categorized as substantial agreement. The significance is considerably good for qualitative performances analysis. Thus, the overall qualitative performances analysis in the study revealed that: (1) the overall

Table 5 Significance of fleiss kappa analysis among radiologists

Total P_i	\bar{P}	\bar{P}_e	Kappa	Significance
59	0.738	0.164	0.686	Substantial Agreement

Table 6 Samples of correct delineation of skull stripping

No.	Original MRI Brain Image	Correct Skull Stripping Delineation
1		
2		
3		

performances of SBRG returns “correct delineation” level of accuracy outcome, which proved that the SBRG skull stripped images are significantly capable to be used further in various medical applications processing (2) the SBRG is an effective technique for skull stripping (3) the substantial agreement among the radiologists in reliability of agreement significances proved that the number of raters involved in the study are appropriate for the skull stripping qualitative assessment. (4) the proposed qualitative evaluation method of skull stripping may offer a new way of skull stripping evaluation in MRI brain images.

Table 6 tabulates the samples of correct delineation of skull stripping quantified by radiologists.

4 Conclusion

This research investigates the qualitative performances of skull stripping accuracy for Fluid Attenuated Inversion Recovery (FLAIR)-Magnetic Resonance Imaging (MRI) brain images. The segmentation technique of Seed-based Region Growing (SBRG) is implemented to strip the brain skull region. The skull stripped images are then visually analyzed by a group of three experienced radiologists which return “correct delineation” accuracy level for overall accuracy outcome. Therefore, based on the qualitative analysis performed, it can be concluded that SBRG is an effective method for skull stripping purpose, whereas the proposed qualitative evaluation method of skull stripping may present an innovative method of skull stripping evaluation in MRI brain images.

Acknowledgments Thousand thanks to Hospital Sungai Buloh for their full cooperation during the collection of MRI brain images. Special thanks to Dr Mohd Ezane Aziz, Dr Win Mar Jalaluddin and Dr Nik Munirah Nik Mahdi, the radiologists involved in the qualitative analysis. Finally, thanks to Research Management Institute (RMI), UiTM and financial support from ERGS-grant (600-RMI/ST/ERGS 5/3/(6/2011)) under the Department of Higher Education, Malaysia.

References

1. Isaac, N.B.: Handbook of medical imaging: processing and analysis, Academic, New York (2000)
2. Bullmore, E., Sporn, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198 (2009)
3. Ibrahim, S., Khalid, N.E.A., Manaf, M., Ngah, U.K.: Particle swarm optimization vs seed-based region growing: brain abnormalities segmentation. *Int. J. Artif. Intell.* **7**(A11), 174–188 (2011)
4. Zhuang, A.H., Valentino, D.J., Toga, A.W.: Skull-stripping magnetic resonance brain images using a model-based level set. *NeuroImage* **32**, 79–92 (2006)
5. Roslan, R., Jamil, N., Mahmud, R.: Skull stripping magnetic resonance images brain images region growing versus mathematical morphology. *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.* **3**, 150–158 (2011)
6. Notestine, C.F., Ozyurt, I.B., Clark, C.P., Morris, S., Grethe, A.B., Bondi, M.W., Jernigan, T.L., Fischl, B., Segonne, F., Shattuck, D.W., Leathy, R.M., Rex, D.E., Toga, A.W., Zou, K.H., Birm, M., Brown, G.: Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction and slice location, *Human Brain Mapp.* **27**(2), 99–113 (2006)
7. Eskildsen, S.F., Coupe, P., Fonov, V., Manjon, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Ostergaard, L.R., Collins, D. L.: BEaST: brain extraction based on nonlocal segmentation technique. *Neuroimage* **59**(3), 2362–2373 (2012)
8. Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B.: A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**(3), 1060–75 (2004)
9. Ishak, N.F., Logeswaran, R., Tan, W.H.: Artifact and noise stripping on low-field brain MRI. *Int. J. Biology Biomed. Eng.* **2**(2), 59–68 (2008)
10. Shen, S., Snadham, W., Granat, M, Sterr A.: MRI Fuzzy Segmentation of brain tissue using neighborhood attraction with neural network optimization. *IEEE Trans. Inf. Technol. Biomed.* **9**(3), 459–467 (2005)

11. Park, J.G., Lee, C.: Skull stripping based on region growing for magnetic resonance brain images. *NeuroImage*. **47**(4), 1394–1407 (2009)
12. Rex, D.E., Shattuck, D.W., Woods, R.P., Narr, K.L., Luders, E., Rehm, K., Stolzner S.E., Rottenberg, D.A., Toga, A.W.: A meta-algorithm for brain extraction in MRI. *NeuroImage* **23**, 625–627 (2004)
13. Souplet, J.C., Lebrun, C., Clavelou, P., Camu, W., Chanalet, S., Ayache, N., Malandain, G.: A Comparative Study of Atrophy Measurements in Multiple Sclerosis. In: 17th Franco-UK scientific meeting of ARSEP (Francois Lhermitte Conferences), (2008)
14. Gouttard, S., Styner, M., Joshi, S., Smith, R.G., Hazlett, H.C., Gerig, G.: Subcortical structure segmentation using probabilistic atlas priors. In: Proceedings of the SPIE, vol. 6512, pp. 65122 J (2007)
15. Lin, X., Qiu, T., Nicolier, F., Ruan, S.: Automatic hippocampus segmentation from brain MRI images. *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.* **2**, 1–10 (2010)
16. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation, *IEEE Trans. Med. Imaging* **23**(7), 903–921 (2004)
17. Hai, O.T., Ngah, U.K., Khalid, N.E.A., Venkatachalam, P.A.: 2000(b), Mammographic calcification clusters using the region growing technique. In: New millennium international conference on pattern recognition, image processing and robot vision, TATI, Malaysia, pp. 157–163 May 2000
18. Mat-Isa, N.A., Mashor, M.Y., Othman, N.H.: Seeded region growing features extraction algorithm; its potential use in improving screening for cervical cancer. *Int. J. Comput. Internet Manage.* **13**(1), 61–70 Jan–April (2005)
19. Sim, J., Wright, C.C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. Therapy* **85**(3), 257–268 (2005)
20. Richard, L.J., Gary, G.K.: The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174 (1977)