REVIEW ARTICLE

대한의학영상정보학회지 2017;23:14-20

KOREAN SOCIETY OF IMAGING INFORMATICS IN MEDICINE

3차원 의료 영상 분할 평가 지표에 관한 고찰

김장우[1], 김종효[1,2,3]

[1]서울대학교 융합과학기술대학원 융합과학부 방사선융합의생명전공, [2]서울대학교 차세대융합기술연구원 의료 IT융합기술 연구센터, [3]서울대학교병원 영상의학과

Review of Evaluation Metrics for 3D Medical Image Segmentation

Jangwoo Kim[1], Jong Hyo Kim[1,2,3]

[1]Program in Biomedical Radiation Sciences, Department of Transdisciplinary Studies, Graduate School of Convergence Science and Technology, Seoul National University; [2]Center for Medical-IT Convergence Technology Research, Advanced Institutes of Convergence Technology, Seoul National University; [3]Department of Radiology, Seoul National University Hospital, Seoul, Korea

Background: Although the research on the automatic medical image segmentation method is active, the research on the evaluation of the image segmentation result is insufficient. It can't be possible to study the accurate and unbiased segmentation method without a correct evaluation of automatic segmentation result. Thus, establishment of the evaluation metrics should be prioritized. In this context, this study reviews and summarizes the 20 evaluation metrics and six classification groups suggested by Taha et al in 2015.
Materials and Methods: A total of 20 image segmentation evaluation metrics are classified into six groups as follows. The first group is the overlap-based image segmentation metrics. Sensitivity, specificity, false positive rate, false negative rate, F-measure, Dice similarity coefficient, Jaccard index and global consistency error belong to this group. The second group is a volume-based metric and has volume similarity. The third group is an information theory-based metric that has mutual information and variation of information. The fourth group is the probability based metrics composed of Interclass correlation, probability distance, Cohens kappa, and Area under ROC curve. The fifth group is a distance-based evaluation metric and reflects the spatial position of the division result. The last group is based on paircounting involving Rand index and Adjusted Rand Index.
Results: It is a reasonable to evaluate the performance of the automatic image segmentation method with minimum proper evaluation metrics. Therefore, we suggested evaluation metrics suitable for a given image segmentation problem.
Conclusion: In this study, we supplemented the lack of description of Taha's paper and simplified it briefly. Through this review, we can perform quantitative evaluation of new image segmentation method, and use it as a basic study to analyze the segmentation problems and improve the performance of new methods to be studied later.

Key Words: Medical image segmentation evaluation; Evaluation metricsmetric selection

서 론

영상 분할(image segmentation)이란 주어진 디지털 영상을 하나 혹은 여러 개의 픽셀 집합으로 분류하는 과정으로 주어진 영상을 보다 의미 있고 해석 목적에 적합한 형태로 단순화 혹은 변환하는 것을 의미한다[1]. 의료 분야에서의 영상 분할(medical image segmentation)은 다양한 의료 영상 장비로부터 얻은 영상을 의학 자료로써 활용이 가능한 형태로 가공하는 것을 의미한다. 의료 영상 분할은 방사선 치료를 위한 정상 장기 분할, 종양 추적을 위한 종양 감지 등, 필요한 의학 목적에 맞게 다양한 형태로 활용되고 있다[2].

의료 영상은 치료, 진단, 임상 연구를 위한 단층촬영(Computed Tomography, CT)과 자기 공명(Magnetic resonance) 영상이 주를 이루며 기술의 발전과 의료 영상 시스템 구축으로 자료의 양이 폭발적으로 증가하고 있다[3]. 의료 영상 정보가 많지 않던 시기에는

교신저자: 김종효
서울대학교 융합과학기술대학원 융합과학부 방사선융합의생명전공, 서울시 관악구 관악로1, 18동 2층
Tel: +82-2-2072-3677, **Fax:** +82-31-888-9148, **E-mail:** kimjhyo@snu.ac.kr
Received: January 25, 2018 / Accepted: March 6, 2018 / Published: May 16, 2018

전문가가 직접 영상 분할 작업을 수행할 수 있으나 영상 정보의 폭발적인 증가로 인해 그 한계에 부딪혔다. 이로 인해 자동 분할 방법에 대한 필요성이 대두되었고 다양한 방법이 고안되었다. 자동 영상 분할 방법에는 전통적인 특성 기반(feature-based) 방법과 텍스처 기반(texture-based) 방법이 있다[4]. 특성 기반 방법은 주어진 영상의 grey-level 값을 직접 이용하는 방식으로 히스토그램 진폭 기반 방법, 경계 기반 방법, 부위(region) 기반 방법이 있다. 텍스처 기반 방법은 주어진 자료에서 유의미한 특성 값을 추출하여 활용하는 방식으로 통계학 기반 접근법, 구조 기반 접근법, 스펙트럼(spectral) 기반 접근법이 있다. 이 외에도 모델 기반 방법과 Atlas 기반 방법이 있으며, 최근에는 신경망(neural network) 기반 방법이 새로이 고안되었다. 특히 신경망 기반 방법은 자동 영상 분할에 탁월한 성능을 보여 활발히 연구가 진행되고 있으며 영상 분할 외에도 의료 영상 전 분야에 걸쳐 응용되고 있다[5].

자동 영상 분할 방법에 대한 활발한 연구와 달리 자동 분할 결과 평가에 대한 연구는 상대적으로 미비하다. 일반적으로 의료 영상 자동 분할 결과는 오랜 경험을 가진 전문가가 직접 분할을 수행하고 검증한 Gold standard (경우에 따라 reference, ground truth라고 칭하기도 함)와 자동 영상 분할 방법을 통해 얻은 결과의 비교를 통해 평가한다. 이때 분할 결과의 유사도를 어떻게 정량적으로 나타낼지가 중요하나 기존 대다수의 연구에서는 Dice similarity coefficient만을 사용하는 경우가 많다. 하지만 하나의 영상 분할 평가 지표만으로는 그 한계가 있음이 여러 연구를 통해 밝혀지며, Dice similarity coefficient 외에도 Jaccard index, Rand index 등, 다양한 평가지표가 제안되었다[6-9]. 따라서 본 연구에서는 영상 분할 방법 연구에 대한 선행 연구로서 다양한 평가지표에 대해 정리한 논문들을 살펴보고 이에 대해 논해보고자 한다.

현재까지 개발된 자동 영상 분할 방법은 올바른 분할 결과로 가정하는 Gold standard와 완전히 일치하지 않고 오차가 존재하는데 이를 분할 오차(segmentation error)라 한다. 분할 오차는 크게 분할 객체 개수, 분할 객체 크기, 분할 객체 경계, 분할 객체 내부 공백 4가지로 나눌 수 있다[10]. 분할 오차를 정량적으로 나타내기 위해 다양한 평가지표를 사용하며 평가지표의 필수조건을 다음과 같이 3가지로 나눌 수 있다[11]. 두 비교 결과의 유사도를 나타내는 정확도(accuracy), 항상 일정한 결과를 도출해내는 반복성(repeatability), 계산 시간이 길지 않는 효율성(efficiency). 의료 산업계 내에 표준 평가 지표는 확립되어 있지 않은 실정이나 2015년 Taha et al. [12]이 의료 영상 분할 평가 지표를 총 망라하여 "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool"을 발표하였다. 해당 논문에서는 의료 영상 분할 결과 평가 연구에 쓰이는 지표 중 가장 많이 쓰이는 20개의 지표를 선정하였으며, 해당 평가 지표의 정의가 하나 이상일 경우 통합하여 설명하였다. 또한 다양한 의료 영상 문제에 적합한 평가 지표를 경험적으로 제시하였다. 하지만 많은 내용을 포괄하다보니 설명이 부족한 부분이 있으며 잘 정리되어 있진 않다. 따라서 해당 논문을 검토 및 정리하였다. 기 논문과 동일하게 총 20개의 평가 지표의 정의(definition), 의미, 장점 및 한계점에 대해 제시하고 6개의 항목으로 분류한 결과를 소개하였다. 또한 해당 논문의 설명이 부족한 부분은 보충하여 설명하였다.

내용은 다음과 같이 구성하였다. 재료 및 방법에서는 6개의 분류 항목별 평가 지표를 설명하였고 결과에서는 주어진 영상에 적합한 평가지표 선정 방법을 보다 단순화하여 제시하였다. 결론에서는 이를 정리 및 요약하였다. 본 고찰을 통해 새로운 영상 분할 방법의 정량적인 평가를 수행할 수 있으며, 이를 통해 새로운 방법의 문제점을 분석하여 영상 분할 방법 성능 개선에 활용하고자 한다.

재료 및 방법

본 연구에서는 총 20개의 영상 분할 평가지표를 overlap 기반, 부피 기반, 정보이론 기반(Information theoretic based), 확률 기반, 공간상 거리 기반, paircounting 기반 그룹으로 총 6종류로 분류하였다. 이와 같이 분류한 이유는 주어진 의료 영상 분할 문제에 대해 적합한 최소한의 평가지표만을 쉽게 선택하기 위함이다.

도메인 정의

주어진 영상의 각 지점을 $X=\{x_1, x_2, \cdots, x_n\}$라 하자. 이때 $|X|=\omega\times h\times d=n$은 총 부피로, w는 너비, h는 높이, d는 깊이를 의미한다. Gold standard의 분할 영역은 다음과 같이 정의한다. $S_g=\{S_g^1, S_g^2\}$, 하위첨자 g는 gold standard를 의미하고 상위첨자 1과 2는 각각 분할 관심 영역과 관심 영역 외의 배경을 의미한다(단 주어진 계층[class]을 관심 영역과 배경 영역 두 개의 계층만이 존재한다고 가정함). 이와 동일하게 자동 분할 결과는 $S_t=\{S_t^1, S_t^2\}$라 정의한다. 또한 분할 할당 함수($f_g^i(x)$)는 gold standard에서는 $f_g^i(x)=1\ if\ x\in(0,1)$, $f_g^i(s)=0\ if\ x\notin(0,1)$로 정의하고, 자동 분할 결과에서는 $f_t^i(x)=1\ if\ x\in(0,1)$, $f_t^i(s)=0\ if\ x\notin(0,1)$ 라 정의한다. 즉 영상의 한 지점이 분할 관심 영역일 확률과 분할 관심 영역이 아닐 확률의 합은 1이 되도록 분할 할당 함수를 설정한다고 볼 수 있다.

Overlap 기반 평가

Overlap 기반 평가 지표 그룹에 속하는 지표는 총 8가지로 민감도(sensitivity), 특이도(specificity), 위양성률(false positive rate), 위음성률(false negative rate), F-Measure (FMS), Dice similarity coefficient (DICE), Jaccard index (JAC), global consistency error (GCE)이다. Overlap 기반 평가지표로 위의 8가지 평가지표를 분류한 것은 오차행렬(confusion matrix)의 구성요소인 true positive (TP), false positive (FP), true negative (TN), false negative (FN)로

부터 도출되었기 때문이다. 오차행렬의 구성요소 각각을 앞서 정의한 도메인에서 나타내면 다음과 같다.

$$\mathrm{TP} = \sum_{r=1}^{n} \min\left(f_t^1(x_r), f_g^1(x_r)\right) \quad (1)$$

$$\mathrm{FP} = \sum_{r=1}^{n} \max\left(f_t^1(x_r) - f_g^1(x_r), 0\right) \quad (2)$$

$$\mathrm{TN} = \sum_{r=1}^{n} \min\left(f_t^2(x_r), f_g^2(x_r)\right) \quad (3)$$

$$\mathrm{FN} = \sum_{r=1}^{n} \max\left(f_t^2(x_r) - f_g^2(x_r), 0\right) \quad (4)$$

민감도와 특이도는 각각 gold standard의 관심영역을 자동 분할 결과가 관심영역으로 할당하고 배경 영역은 배경으로 할당했을 확률을 의미하는 직관적인 평가 지표이다. 하지만 분할 영역이 작은 경우에 오차를 더 크게 반영하여 분할 영역 크기에 민감하다는 단점을 가지고 있어 영상 분할 결과 평가에는 잘 쓰이지 않는다. 민감도와 특이도의 정의는 다음과 같다.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (6)$$

이와 유사하게 위양성률과 위음성률도 계산 가능하며, 이 두 평가지표는 민감도와 특이도와 상응하므로 민감도와 특이도와 동시에 평가지표로 쓰이진 않는다.

$$\text{False positive rate} = \frac{FP}{FP+TN} = 1 - specificity \quad (7)$$

$$\text{False negative rate} = \frac{FN}{FN+TP} = 1 - sensitivity \quad (8)$$

F-measure (FMS)란 Rijsbergen 유효성 측정 척도의 특수한 경우로 F_β-measure가 정확한 표현이다. FMS_β는 정보의 복원정도를 나타내는 지표로 정확도(precision)와 민감도의 상호작용을 나타낸다.

$$\text{Precision} = \text{PPV} = \frac{TP}{TP+FP} \quad (9)$$

$$FMS_\beta = \frac{(\beta^2+1)*PPV*TPR}{\beta^2*PPV+TPR} \quad (10)$$

FMS_β에서 β의 값이 1일 때의 값을 주로 사용하며 이는 Dice similarity coefficient와 같다. 즉 정확도와 민감도가 같을 때 FMS_β의 값은 Dice similarity coefficient와 같다.

Dice similarity coefficient는 영상 분할 평가에 쓰이는 가장 대표적인 지표이다. DICE는 두 영상 분할 결과를 직접 비교하며 그 유사도를 나타내며 정의는 다음과 같다.

$$\mathrm{DICE} = \frac{2|s_g^1 \cap s_t^1|}{|s_g^1|+|s_t^1|} = \frac{2TP}{2TP+FP+FN} \quad (11)$$

Jaccard index (JAC)는 두 분할 결과의 합을 교차 값으로 나눈 것으로 DICE와 상응한다. 따라서 DICE와 JAC를 동시에 평가지표로 사용하는 것은 의미가 없다. Jaccard index의 정의는 다음과 같다.

$$\mathrm{JAC} = \frac{|s_g^1 \cap s_t^1|}{|s_g^1 \cup s_t^1|} = \frac{DICE}{2-DICE} \quad (12)$$

Global consistency error (GCE)는 두 영상 분할 결과 간의 오차를 측정하는 지표이다. 임의의 지점 x를 포함한 분할 결과 S에 속하는 모든 복셀의 집합을 $R(S, x)$라 하자. 이 때 두 분할 결과의 차이를 다음과 같이 정의한다.

$$\mathrm{E}(S_t, S_g, \mathrm{x}) = \frac{|R(S_t,x)/R(S_g,x)|}{|R(S_t,x)|} \quad (13)$$

이를 이용하여 모든 복셀에서의 오차 평균을 계산한 것이 GCE이며 다음과 같다.

$$\begin{aligned}&\mathrm{GCE}(\mathrm{S_t}, \mathrm{S_g})\\ &= \frac{1}{n}\min\left\{\sum_i^n E(S_t, S_g, x_i), \sum_i^n E(S_g, S_t, x_i)\right\}\\ &= \frac{1}{n}\left\{\frac{FN(FN+2TP)}{TP+FN} + \frac{FP(FP+2TN)}{TN+FP},\right.\\ &\left.\frac{FP(FP+2TP)}{TP+FP} + \frac{FN(FN+2TN)}{TN+FN}\right\}\end{aligned} \quad (14)$$

오차를 측정해야만 하는 경우를 제외하면 자주 사용되진 않는다.

NCC부피 기반 평가 지표

부피 기반 평가 지표에는 체적 유사도(volumetric similarity)가 있다. 체적 유사도는 두 분할 결과각각의 분할 객체 부피를 비교하는 방식이다. 주어진 영상에서 부피를 추정하는 방법은 다양하나 본 연구에서는 두 대상의 절대 부피 차이를 비교 대상 부피로 나눈 방식을 기준으로 한다. 정의는 다음과 같다.

$$\mathrm{VS} = 1 - \frac{\left||s_t^1|-|s_g^1|\right|}{|s_t^1|+|s_g^1|} = 1 - \frac{|FN-FP|}{2TP+FP+FN} \quad (15)$$

체적 유사도는 오직 분할 결과의 부피만을 고려하므로 반드시 비교 대상 결과 간의 정렬이 선행되어야만 한다. 제한된 환경에서 측정한 결과를 분석하는 의료 영상 문제에서는 결과 간 정렬이 비교적 쉽게 이루어지므로 앞서 언급한 단점에도 불구하고 활용 가치가 높다고 볼 수 있다. 부피 기반 평가지표는 Overlap 기반 지표와 동일하게 오차행렬의 구성요소만으로 표현이 가능하나 Overlap 분류와 달리 두 분할 간 중복 구간을 고려하지 않기 때문에 부피 기반 평가 지표로 별도 분류하였다.

정보 이론 기반 평가 지표

정보 이론 기반 평가 지표에는 상호 정보량(mutual information), 정보 변화량(variation of information)이 있다. 상호 정보량 지표는 하나의 변수(variable)가 다른 변수에 대해 갖고 있는 정보의 총량

을 의미하며, 영상 분할 결과에서는 개별 지점의 값이 아닌 분할 결과 구간을 기반으로 유사도를 측정하는 것을 의미한다. 상호 정보량을 정의하기 위해서는 먼저 주변 엔트로피(marginal entropy)와 조인트 엔트로피(joint entropy)를 정의해야 한다. 주변 엔트로피와 조인트 엔트로피는

$$\mathrm{H(S)} = -\sum_i p(S^i)\log(p(S^i)) \tag{16}$$

$$\mathrm{H}(S_1, S_2) = -\sum_{ij} p\left(S_1^i, S_2^i\right)\log\left(p\left(S_1^i, S_2^i\right)\right) \tag{17}$$

Where (18)

$$\mathrm{p}\left(S_g^1\right) = \frac{TP + FN}{n}$$
$$\mathrm{p}\left(S_g^2\right) = \frac{TN + FN}{n}$$
$$\mathrm{p}\left(S_t^1\right) = \frac{TP + FP}{n}$$
$$\mathrm{p}\left(S_g^2\right) = \frac{TN + FP}{n}$$

and

$$\mathrm{p}(S_1^1, S_2^1) = \frac{TP}{n}$$
$$\mathrm{p}(S_1^1, S_2^2) = \frac{FN}{n}$$
$$\mathrm{p}(S_1^2, S_2^1) = \frac{FP}{n}$$
$$\mathrm{p}(S_1^2, S_2^2) = \frac{TN}{n}$$

와 같이 정의한다. 이를 이용하여 상호 정보량은 아래와 같이 정의한다.

$$\mathrm{MI}\left(S_g, S_t\right) = \mathrm{H}\left(S_g\right) + \mathrm{H}(S_t) - \mathrm{H}(S_g, S_t) \tag{19}$$

이와 유사하게 정보 변화량은 하나의 변수가 변할 때 다른 변수가 변화하는 양을 나타낸다. 정의는 다음과 같으며 상호 정보량이 정보 변화량보다 영상 분할 결과 평가에 좀 더 자주 사용된다.

$$\mathrm{VOI}\left(S_g, S_t\right) = \mathrm{H}\left(S_g\right) + \mathrm{H}(S_t) - 2MI(S_g, S_t) \tag{20}$$

확률 기반 평가 지표

확률 기반 평가 지표는 분할 결과의 중복 영역에 위치한 복셀로부터 계산된 통계적 수치 값을 의미한다. 확률 기반 평가 지표는 계층 간 상관도(interclass correlation, ICC), 확률 거리(Probabilistic Distance, PBD), Cohens kappa, AUC (Area under ROC curve)가 있다. 계층 간 상관도는 관찰 쌍 간의 상관관계를 나타내는 지표이다. 즉, 분할 결과 간의 일치도(conformity)를 측정하는 지표이다. 일반적으로 계층 간 상관도는 다음과 같이 정의한다.

$$\mathrm{ICC} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\epsilon^2} \tag{21}$$

이때 σ_s는 분할 결과 간의 표준편차를 의미하고, σ_ϵ는 두 분할 결과의 차이에 의한 한 점의 표준편차를 나타낸다. 이를 앞서 정의한 도메인을 적용하면 다음과 같다.

$$\mathrm{ICC} = \frac{MS_b - MS_w}{MS_b + (k-1)MS_w} \tag{22}$$

Where (23)

$$\mathrm{MS}_b = \frac{2}{n-1}\sum_x (m(x) - \mu)^2$$
$$\mathrm{MS}_w = \frac{1}{n}\sum_x (f_g(x) - m(x))^2 + (f_t(x) - m(x))^2$$

주어진 식에서 k는 계층의 수로, 주어진 문제에서는 2가 된다. μ는 두 분할 결과 각각의 평균을 평균한 값이고 m(x) = (f_g (x)+f_t (x))/2로 주어진 복셀 x의 평균값이다. MS_b는 분할 결과 간의 평균의 제곱을, MS_w는 하나의 분할 내에서의 평균의 제곱을 의미한다.

확률 거리는 Gerig et al. [14]이 제안한 지표로 두 분할 간 거리를 측정한다.

$$\mathrm{PBD}(S_g, S_t) = \frac{\sum_x |f_g(x) - f_t(x)|}{2\sum_x f_g(x) f_t(x)} \tag{24}$$

Cohen Kappa 상수는 주어진 두 표본 간에 일치(agreement)를 나타내는 지표이다. 의도치 않게 임의로 발생하는 문제까지 고려하므로 다른 지표에 비해 보다 일정한 결과를 낼 수 있다는 장점이 있다. Cohen Kappa 값은 다음과 같이 정의한다.

$$\mathrm{KAP} = \frac{f_a - f_c}{N - f_c} \tag{25}$$

Where (26)

$$f_a = \mathrm{TP} + \mathrm{TN}$$
$$f_c = \frac{(TN + FN)(TN + FP) + (FP + TP)(FN + TP)}{N}$$

Receiver operating characteristics (ROC) 곡선이란 한 축에는 TPR을 다른 축은 FPR로 그린 그래프이다. AUC는 ROC 곡선에 아래가 차지하는 면적으로 1에 가까울수록 두 분할 결과가 유사하다고 볼 수 있다. AUC는 정확도(accuracy)를 나타내는 지표로 방사선 의료 진단 분야에서 자주 사용되며 그 값은 다음과 같다.

$$\mathrm{AUC} = 1 - \frac{FPR + FNR}{2} \tag{27}$$

공간상 거리 기반 평가 지표

공간상 거리 기반 평가 지표는 두 결과 간의 다른 정도를 평가하는 지표로 다른 항목의 지표들과 달리 분할된 객체 간의 공간상 거리를 고려한다. 특히 이러한 특징은 Dice similarity coeffiecient로 대표되는 overlap 기반 평가지표에서는 overlap 구간(false positives와 false negatives로 판정된 구간)이 아닌 복셀은 고려하지 못하기 때문에 더욱 중요하다고 볼 수 있다. 공간 기반 평가 지표에는 Haus-

dorff 거리, 평균 Hausdorff 거리, Mahalanobis 거리가 있다. Hausdorff 거리는 다음과 같이 정의한다.

$$\mathrm{HD(A,B)} = \max_{a\in A}\max_{b\in B}(h(A,B), h(B,A)) \tag{28}$$

이때 거리는 유클리디안 공간상에서의 거리를 나타내며, HD (A, B)를 계산하는 알고리즘은 다양하나 본 연구에서는 효율적인 계산이 가능한 nearly-linear time 방식을 택하였다[13]. 일반적으로 Hausdorff 거리 지표는 이상치(outlier)에 민감하게 반응하여 평가 지표로는 적합하지 않다고 여겨졌으나 분위수 정규화(quantile normalization)와 같은 정규화 과정을 거치면 평가지표로 활용이 가능하다.

평균 Hausdorff 거리는 분할 대상 점들의 Hausdorff 거리의 평균값을 의미한다. Hausdorff 거리와는 달리 안정적이고 이상치에 덜 민감하다. 정의는 다음과 같다.

$$\mathrm{AVD(A,B)} = \max(\mathrm{d(A,B)}, \mathrm{d(B,A)})$$

$$\text{where } \mathrm{d(A,B)} = \frac{1}{N}\sum_{a\in A}\min_{b\in B}\|a-b\| \tag{29}$$

Mahalanobis 거리는 Hausdorff 거리와는 달리 두 결과 점 간의 상관도(correlation)까지도 고려한다. 이로 인해 계산식에 공분산 행렬을 포함하게 된다. Mahalanobis 거리는 특정 군(cluster)과 하나의 점 사이의 거리를 계산하는 것으로, 군의 평균이 되는 지점과의 거리를 군의 표준편차로 나눈 값이다.

임의의 두 점 x, y에 대한 Mahalanobis 거리는

$$\mathrm{MHD(A,B)} = \sqrt{(x-y)^T S^{-1}(x-y)} \tag{30}$$

으로 S는 공분산 행렬을, T는 행렬의 전치행렬을 의미한다. 이를 두 분할 결과 집합 X,Y에 대해 표현하면

$$\mathrm{MHD(X,Y)} = \sqrt{(\mu_x-\mu_y)^T S^{-1}(\mu_x-\mu_y)} \tag{31}$$

으로 μ_x와 μ_y는 각각 결과 집합의 평균을 의미한다. 이때 공분산 행렬 S는 다음과 같다.

$$\mathrm{S} = \frac{n_1 S_1 + n_2 S_2}{n_1+n_2} \tag{32}$$

이때 n_1과 n_2는 각각 해당 집합에 속하는 복셀의 수이다.

Paircounting 기반 평가 지표

Pair-counting 기반 평가 지표에는 Rand index와 Adjusted Rand index가 있다. 해당 지표를 살펴보기에 앞서, overlap 기반 평가 지표를 정의하기 위해 오차행렬의 4요소를 정의했듯이 먼저 4개 상수 a, b, c, d를 정의해야 한다. 비교하고자 하는 두 분할 결과의 집합을 X라 할때, X × X에서 나타날 수 있는 튜플(tuple)의 집합을 P라 하자(이때 가능한 경우의 수는 $\binom{n}{2} = \frac{n(n-1)}{2}$이다). 이때 각각의 튜플은 주어진 4개의 그룹 중 하나에 속한다. 첫 번째 그룹은 두 점 x_i와 x_j가 두 분할 결과 S_g와 S_t에서 각각 동일한 분류에 속하는 것으로 분류되는 경우로 이를 a라고 하자. 두 번째 그룹은 두 점 x_i와 x_j가 S_g에서는 동일한 분류에 속하나 S_t에서는 다른 분류에 속하는 것으로 분류되는 경우이다. 이를 b라 하자. 세 번째 그룹은 두 점 x_i와 x_j가 S_t에서는 동일한 분류에 속하나 S_g에서는 다른 분류에 속하는 것으로 분류되는 경우로 이를 c라 하자. 마지막 그룹은 두 점 x_i와 x_j가 두 분할 결과 S_g와 S_t에서 모두 다른 분류에 속하는 것으로 분류되는 경우로 이를 d라 하자. 앞서 overlap based 평가 지표와 동일하게 2개의 분할 집합만이 존재한다고 가정하자. 그러면 이때

$$\mathrm{a} = \frac{1}{2}[TP(TP-1) + FP(FP-1) + TN(TN-1) + FN(FN-1)] \tag{33}$$

$$\mathrm{b} = \frac{1}{2}[(TP+FN)^2 + (TN+FP)^2 - (TP^2+TN^2+FP^2+FN^2)] \tag{34}$$

$$\mathrm{c} = \frac{1}{2}[(TP+FP)^2 + (TN+FN)^2 - (TP^2+TN^2+FP^2+FN^2)] \tag{35}$$

$$\mathrm{d} = \frac{\mathrm{n(n-1)}}{2} - (\mathrm{a}+\mathrm{b}+\mathrm{c}) \tag{36}$$

로 나타낼 수 있다. 이를 바탕으로 Rand index를 정의하면 다음과 같다.

$$\mathrm{RI}(S_g, S_t) = \frac{a+b}{a+b+c+d} \tag{37}$$

Rand index는 두 군집(cluster) 간의 유사도를 측정하기 위해 고안된 지표로서 할당된 집합의 특정 값에 무관하다는 특징을 가지고 있어 분할 외에 군집 문제에도 자주 사용되는 지표이다. 하지만 Rand index는 우연히 발생하는 경우에 대한 보정이 이루어지지 않아 이를 보안한 adjusted Rand index가 제안되었다. 그 정의는 다음과 같다.

$$\mathrm{ARI} = \frac{2(ad-bc)}{c^2+b^2+2ad+(a+d)(c+b)} \tag{38}$$

결 과

위와 같이 분류하고 정의한 20개의 평가 지표를 모든 영상 분할 문제에 적용할 필요는 없다. 즉, 주어진 영상 분할 평가 목적에 적합한 평가 지표만을 적용하여 자동 영상 분할 결과의 성능을 평가하는 것이 효율적인 방법이다. 따라서 몇 가지 상황을 가정하고 그에 적합한 평가 지표를 Table 1에 제시하였다. 해당 Table은 기 논문(Taha et al., 2015)의 Table 5를 요약 및 재편집하여 정리한 것으로 제시한 상황 이외의 경우에 대해서는 기 논문을 참조하길 권한다. 중요한 것은 각 평가 지표의 장단점과 한계를 인지하고 상호보완이 가능한 여러 개의 평가지표를 사용해야한다는 점이다. 즉, 단일 평

Table 1. Recommendations for selecting appropriate metrics according to evaluation task

		Case				
		1	2	3	4	5
Overlap	Sensitivity					
	Specificity					
	FPR				v	
	FNR					
	FMS					v
	DICE					v
	JAC					v
	GCE					
Volume	VS					v
Information theory	MI				v	v
	VOI					v
Probabilistic	ICC					
	PBD					
	KAP					v
	AUC					v
Distance	HD	v	v	v		
	AVD	v	v	v		v
	MHD		v			v
Pair-counting	RI					
	ARI					

가 지표만으로는 영상 분할 결과의 단편적인 부분을 전체로 확대 해석하거나 평가 지표에 반영되지 않는 부분은 놓칠 수 있다. 따라서 다양한 평가 지표를 통해 결론을 도출해야 영상 분할 결과를 올바르게 해석할 수 있다.

Case 1: 자동 분할 결과의 경계 정확도(accuracy)가 가장 중요한 경우에는 false positives와 false negatives의 공간상 위치까지 고려해야 하므로 공간 기반 평가 지표를 사용하는 것이 적합하다. 특히 평균 Hausdorff 거리가 가장 적합하다. 결과의 부피만을 고려하는 체적 유사도를 평가 지표로 포함시키는 것은 지양해야 한다.

Case 2: 분할 구간이 주위 배경에 비해 너무 작은 경우에는(전체 크기 대비 5% 이하의 분할 구간만이 존재하는 경우) 분할 오차가 분할 객체의 크기와 비례하기 때문에 오차의 절대치를 너무 작게 평가할 수 있다. 따라서 이와 같은 문제가 생길 수 있는 overlap 기반 평가 지표보다는 공간 기반 평가 지표로 분할 결과를 평가해야 한다.

Case 3: 분할 경계가 복잡한 경우에는 분할 객체 내 점들을 대표하는 통계치보다는 경계점 각각의 값과 위치가 중요하다. 따라서 이러한 경우에는 각 점들의 위치를 잘 반영할 수 있는 Hausdorff 거리와 평균 Hausdorff 거리가 적합하다. 단, 이상치의 영향을 받아 결과가 왜곡되지 않도록 분위수 정규화를 거친 Hausdorff 거리를 쓰거나 평균 Hausdorff 거리를 사용해야 한다.

Case 4: 경우에 따라 false positives 혹은 false negatives를 포함하더라도 모든 true 구간을 놓쳐서는 안 되는 경우가 있다. 이러한 경우에는 true 구간에 대한 보정이 이루어지는 상호 정보량이나 위양성률을 평가지표로 삼는 것이 적합하다.

Case 5: 데이터의 이상치가 존재하는 경우에는 다양한 지표를 적용하는 것이 가능하나 이상치에 민감한 Hausdorff 거리를 평가지표로 삼아서는 안된다.

고찰 및 결론

본 연구에서는 3차원 의료 영상 분할 평가 지표를 망라한 2015년 Taha et al. 연구팀의 "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool"을 검토 및 정리하였다. 해당 논문에서는 총 20개의 영상 분할 평가 지표를 6개 항목으로 분류 및 정의하였고 주어진 영상 분할 문제에 적합한 평가 지표를 제시하였다. 6가지 항목에는 overlap 기반, 부피 기반, 정보이론 기반, 확률 기반, 공간상 거리 기반, paircounting 기반 그룹이 있으며 각각 장단점 및 한계점이 존재한다. 따라서 실제 의료 영상 분할 결과 평가에는 하나가 아닌 둘 이상의 평가 지표를 조합하여 평가하여야 자동 영상 분할 기법의 성능을 올바르게 평가할 수 있다. 본 연구에서는 기 논문의 설명이 부족한 부분은 보충하였고 보다 단순화하여 간략히 정리하였다. 본 고찰을 통해 새로운 영상 분할 방법의 정량적인 평가를 수행할 수 있으며, 이를 통해 추후 연구할 새로운 방법의 문제점을 분석하고 성능을 개선하는 데 활용하고자 한다.

감사의 글

본 연구는 미래창조과학부의 방사선바이오의료 기술개발사업으로 지원된 연구결과임(0581-20170022, 적응적 방사선종양치료의 정밀도 및 효율 향상을 위한 딥러닝 기반 지능형 종양추적 컨투어링 기술 개발).

REFERENCES

1. Haralick, Robert M, Linda G. Shapiro. Image segmentation techniques. Computer Vision, Graphics, and Image Processing 1985; 29: 100-132.
2. Pham, Dzung L, Chenyang Xu, Jerry L. Prince. Current methods in medical image segmentation. Annual Review of Biomedical Engineering 2000; 2: 315-337.
3. Hess, Erik P, et al. Trends in computed tomography utilization rates: a longitudinal practice-based study. Journal of Patient Safety 2014; 10: 52-58.
4. Sharma, Neeraj, Lalit M. Aggarwal. Automated medical image segmentation techniques. Journal of Medical Physics/Association of Medical Physicists of India 2010; 35: 3.
5. Litjens, Geert, et al. A survey on deep learning in medical image analysis. arXiv Preprint arXiv:1702.05747 (2017).

6. Chalana, Vikram, Yongmin Kim. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Transactions on Medical Imaging 1997; 16: 642-652.
7. Saha, Punam K., Jayaram K. Udupa, Dewey Odhner. Scale-based fuzzy connected image segmentation: theory, algorithms, and validation. Computer Vision and Image Understanding 2000; 77: 145-174.
8. Taha, Abdel Aziz, Allan Hanbury, Oscar A. Jimenez del Toro. A formal method for selecting evaluation metrics for image segmentation. Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014.
9. Buckley, Chris, Ellen M. Voorhees. Evaluating evaluation measure stability. ACM SIGIR Forum. Vol. 51. No. 2. ACM, 2017.
10. Shi, Ran, King Ngi Ngan, Songnan Li. The objective evaluation of image object segmentation quality. International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham, 2013.
11. Fenster, Araon, Bernard Chiu. Evaluation of segmentation algorithms for medical imaging. Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. IEEE, 2005.
12. Taha, Abdel Aziz, Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging 2015; 15: 29.
13. Taha, Abdel Aziz, Allan Hanbury. An efficient algorithm for calculating the exact Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 2015; 37: 2153-2163.
14. Gerig, Guido, Matthieu Jomier, Miranda Chakos. Valmet: A new validation tool for assessing and improving 3D object segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001. Springer Berlin/Heidelberg, 2001.

초록

의료 영상 분할은 의료 영상 처리의 핵심적인 과정으로 의료 영상 전문가가 주어진 의료 영상을 하나 혹은 여러 개의 픽셀 집합으로 분류하여 해부학적 구조 기반의 객체로 할당하는 것을 의미한다. 자료의 수가 많지 않았던 과거에는 의료 영상 전문가가 영상 분할을 수행하였으나, 의료 장비의 성능 향상과 보급률 증가 및 환자 수 증가로 인하여 의료 영상 수는 폭발적으로 증가하였고 의료 영상 전문가가 모든 영상 분할 작업을 수행하기에는 한계에 이르렀다. 이를 해결하기 위해 다양한 자동 분할 방법이 제시되었고 활발히 연구 중에 있으나 영상 분할 결과 평가에 대한 연구는 미비한 실정이다. 편향(bias)되지 않고 정확한 자동 영상 분할 방법의 개발하기 위해서 올바른 영상 분할 지표의 확립이 우선시되어야한다. 일반적인 의료 영상 분할 결과의 평가는 올바르게 분할이 되었다고 판단되는 Gold standard와 의료 영상 분할 방법을 통해 얻은 분할 결과를 비교하여 그 성능을 평가하는 방식으로 이루어진다. 이때 성능은 의료 영상 분할 방법을 통해 얻은 결과가 Gold standard와 얼마나 유사한가를 수치적으로 나타낸다. 두 결과의 유사도를 나타내기 위해서 기존 연구는 주로 Dice coefficient 값만을 통해 평가하였으나 단일 평가지표만으로 성능을 평가하기에는 그 한계가 있다. 따라서 본 연구에서는 의료 영상 분할 방법 평가에 대한 고찰을 위해 평가 지표를 총 망라한 "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool"을 검토하였다. 해당 논문에 설명이 부족한 부분은 보충하여 정리하였다. 기 논문과 동일하게 총 20개의 평가 지표의 정의, 의미, 장점 및 한계점에 대해 제시하고 6개의 항목으로 분류하였다. 또한 주어진 의료 영상 문제에 적합한 평가지표를 제시하였다. 이를 통해 추후 연구할 새로운 영상 분할 방법의 정량적인 평가 및 문제점을 분석하여 영상 분할 방법 성능 개선에 활용하고자 한다.