Medical Image Analysis 60 (2020) 101601

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

LinSEM: Linearizing segmentation evaluation metrics for medical images

Jieyu Li[a,b], Jayaram K. Udupa[b,*], Yubing Tong[b], Lisheng Wang[a], Drew A. Torigian[b]

[a] Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, 800 Dongchuan RD, Shanghai 200240, China

[b] Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 602 Goddard Building, 3710 Hamilton Walk, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 14 March 2019
Revised 6 August 2019
Accepted 7 November 2019
Available online 9 November 2019

Key words:
Medical image segmentation
Evaluation metrics
Acceptability score
Linear relationship

ABSTRACT

Numerous algorithms are available for segmenting medical images. Empirical discrepancy metrics are commonly used in measuring the similarity or difference between segmentations by algorithms and "true" segmentations. However, one issue with the commonly used metrics is that the same metric value often represents different levels of "clinical acceptability" for different objects depending on their size, shape, and complexity of form. An ideal segmentation evaluation metric should be able to reflect degrees of acceptability directly from metric values and be able to show the same acceptability meaning by the same metric value for objects of different shape, size, and form. Intuitively, metrics which have a linear relationship with degree of acceptability will satisfy these conditions of the ideal metric. This issue has not been addressed in the medical image segmentation literature. In this paper, we propose a method called *LinSEM* for linearizing commonly used segmentation evaluation metrics based on corresponding degrees of acceptability evaluated by an expert in a reader study.

LinSEM consists of two main parts: (a) estimating the relationship between metric values and degrees of acceptability separately for each considered metric and object, and (b) linearizing any given metric value corresponding to a given segmentation of an object based on the estimated relationship. Since algorithmic segmentations do not usually cover the full range of variability of acceptability, we create a set (S_S) of simulated segmentations for each object that guarantee such coverage by using image transformations applied to a set (S_T) of true segmentations of the object. We then conduct a reader study wherein the reader assigns an acceptability score (*AS*) for each sample in S_S, expressing the acceptability of the sample on a 1 to 5 scale. Then the metric-*AS* relationship is constructed for the object by using an estimation method. With the idea that the ideal metric should be linear with respect to acceptability, we can then linearize the metric value of any segmentation sample of the object from a set (S_A) of actual segmentations to its linearized value by using the constructed metric-acceptability relationship curve.

Experiments are conducted involving three metrics – Dice coefficient (*DC*), Jaccard index (*JI*), and Hausdorff Distance (*HD*) – on five objects: skin outer boundary of the head and neck (cervico-thoracic) body region superior to the shoulders, right parotid gland, mandible, cervical esophagus, and heart. Actual segmentations (S_A) of these objects are generated via our Automatic Anatomy Recognition (AAR) method. Our results indicate that, generally, *JI* has a more linear relationship with acceptability before linearization than other metrics. LinSEM achieves significantly improved uniformity of meaning post-linearization across all tested objects and metrics, except in a few cases where the departure from linearity was insignificant. This improvement is generally the largest for *DC* and *HD* reaching 8–25% for many tested cases. Although some objects (such as right parotid gland and esophagus for *DC* and *JI*) are close in their meaning between themselves before linearization, they are distant in this meaning from other objects but are brought close to other objects after linearization. This suggests the importance of performing linearization considering all objects in a body region and body-wide.

* Corresponding author.
E-mail address: jay@pennmedicine.upenn.edu (J.K. Udupa).

https://doi.org/10.1016/j.media.2019.101601
1361-8415/

1. Introduction

1.1. Background

Image segmentation is the process of recognizing and delineating objects in images. Literature on general image segmentation dates back to the early 1960s (Doyle, 1962; Narasimhan and Fornango, 1963). Principles for medical image segmentation began to appear from the late 1970s (Herman et al., 1979; Liu, 1977) with the routine availability of computed tomography (CT) images. Approaches to medical image segmentation can be classified broadly into two groups: purely image-based or *PI-approaches* and prior-knowledge-based or *PK-approaches*. PI-approaches make segmentation decisions based entirely on information derived from the given image (Baxter et al., 2017; Beucher, 1992; Boykov et al., 2001; Falcao et al., 1998; Kasset al., 1987; Malladi et al., 1995; Mumford and Shah, 1989; Pope et al., 1984; Udupa and Samarasekera, 1996). They predate PK-approaches and continue to seek new frontiers. In PK-approaches (Ashburner and Friston, 2009; Christensen et al., 1994; Chu et al., 2013; Cootes et al., 1995; Drozdzal et al., 2018; Gee et al., 1993; Li et al., 2014; Moeskops et al., 2016; Oda et al., 2018; Pizer et al., 2003; Shen et al., 2011; Shi et al., 2017; Staib and Duncan, 1992; Udupa et al., 2014; Zhang et al., 2015), known object shape, image appearance, and relation information over a subject population are first codified (learned) and then utilized on a given image to bring constraints into the segmentation process. They evolved precisely to overcome failure of PI-approaches in situations such as lack of definable object boundaries in the image, variable object boundary characteristics, and image artifacts, and also simply to increase level of automation. Among PK-approaches, three distinct classes of methods can be identified – *model-based* (Cootes et al., 1995; Pizer et al., 2003; Shen et al., 2011; Staib and Duncan, 1992; Udupa et al., 2014), *atlas-based* (Ashburner and Friston, 2009; Christensen et al., 1994; Chu et al., 2013; Gee et al., 1993; Shi et al., 2017), and *deep-learning (DL)-based* (Drozdzal et al., 2018; Li et al., 2014; Moeskops et al., 2016; Oda et al., 2018; Zhang et al., 2015). The division between model- and atlas-based groups is somewhat arbitrary and a matter of semantics. In fact, DL networks are also often referred to as "models." Segmentation is crucial in radiological practice since accurate delineation of tissues and organs provides solid means for disease diagnosis, staging, treatment planning and guidance, and treatment response assessment and prediction.

In clinical practice, "degree of acceptability" subjectively evaluated by experts based on clinical knowledge and practical concerns, is perhaps the most meaningful metric to evaluate goodness and usefulness of segmentations. However, it is impractical to employ reader studies for technical bench testing of every algorithm at the developmental phase. As such, it is more realistic to use objective computational metrics to evaluate segmentations. Empirical discrepancy metrics (Zhang, 1996, 2001) are commonly used in measuring the similarity or difference between segmentations by algorithms and "true" segmentations which are often referred to as ground truth. However, one rather serious issue with these metrics, whether for technical bench testing or end clinical evaluation in an application, is that the same metric value often represents different levels of clinical acceptability for different objects depending on their size, shape, and complexity of form. For example, a Dice coefficient value of 0.8 for a large non-sparse blob-like object such as liver may imply good, and not outstanding, quality of segmentation, whereas for a thin and narrow spatially sparse object such as esophagus, this value represents excellent quality. This is mainly due to the fact that small deviations in segmentation cause much larger changes in the Dice coefficient value for sparse objects than for large non-sparse objects.

An ideal segmentation evaluation metric should: (a) be able to reflect degrees of acceptability directly from metric values; (b) be able to show the same acceptability meaning by the same metric value for objects of different shape, size, and form; and (c) be easily calculated for a large set of segmentations. Intuitively, metrics which have a linear relationship with the degree of acceptability will satisfy these conditions of the ideal metric. In this paper, we propose a method called *LinSEM* for linearizing commonly used segmentation evaluation metrics based on corresponding degrees of acceptability evaluated by an expert. In this way, linearized metrics will have close-to-linear relationships with acceptability and therefore the same (or similar) acceptability meaning for different objects.

1.2. Related work

There are two main categories of segmentation evaluation metrics: region-based and boundary-based. Region-based metrics compare regions occupied by segmentations by algorithms and their corresponding ground truth. Fractioned values are calculated among area or volume of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) regions. TP and TN stand for correctly segmented object and background regions, respectively, and FP and FN represent wrongly segmented object and background regions, respectively. Commonly used region-based metrics include Dice coefficient (*DC*) (Dice, 1945), Jaccard index (*JI*) (Jaccard, 1901), and separately expressed volume fractions TPVF, TNVF, FPVF, and FNVF for both binary and fuzzy segmentations (Udupa et al., 2006). Boundary-based metrics express the difference between boundaries of segmentations by algorithm and ground truth. Common boundary-based metrics include Hausdorff distance (*HD*) (Huttenlocher et al., 1993), average symmetric surface distance (ASD) (Lamecker et al., 2004), and root mean squared distance (RMSD) (Detmer et al., 1990), which are all different descriptions of some statistic of the distance between the two boundaries. These metrics are sometimes simultaneously reported to show the effectiveness of algorithms (Baiker et al., 2010; Chen et al., 2012; Dou et al., 2017; Linguraru et al., 2012; Wolz et al., 2013). Some evaluations combine scores from different metrics. For example, a composite metric created by combining two region-based metrics with three boundary-based metrics (ASD, RMSD, and *HD*) is described in Heimann et al. (2009). Scores from these five metrics are used in Lopez-Molina et al. (2013), Schmid et al. (2011), Tomoshige et al. (2014), and the average scores are calculated as a balanced form of segmentation evaluation in Ruskó et al. (2009).

The above commonly-used basic metrics all have their drawbacks. Whereas boundary-based metrics are not precise in expressing the segmentation quality of objects of complex shape, region-based metrics always emphasize the importance of some traits/measures (such as under segmentation or FNs) and weaken others (such as over segmentation or FPs). Several improved metrics have been created to mitigate certain concerns in practice. A metric designed to detect and measure a wider range of segmentation errors which may be overlooked by common metrics is described in Yeghiazaryan and Voiculescu (2018). It combines region-based and boundary-based metrics, by estimating region-based measures in the neighborhood of the boundaries of ground truth and segmentations by algorithms. The works in Kim et al. (2012, 2015) combine metrics with a medical consideration function, which considers regions inside and outside the object boundary as having different medical importance and so calculates bidirectional boundary distance. Ref. Cappabianco et al. (2017) noticed the fact that large FN implies small TP, and, since FP has no relationship with TP, commonly-used region-based metrics, such as *DC* and *JI*, portray the influence of FN and FP differently. The au-

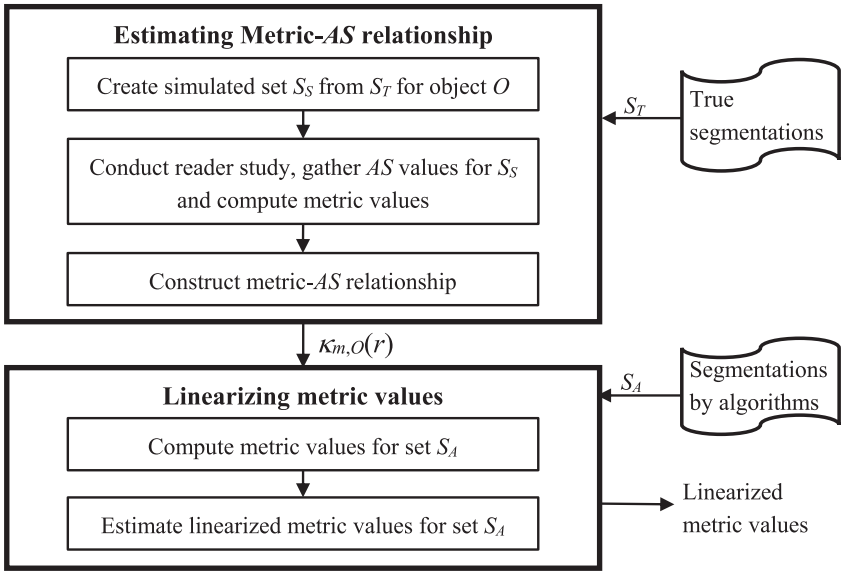

Fig. 1. A schematic representation of the LinSEM method.

thors proposed a metric that is balanced with respect to FP and FN variations.

Although improvements are made in these proposals, the new metrics all focus only on segmentation compositions and none of them concerned the problem of metrics having different acceptability meaning for different objects. In this work, we propose the LinSEM method to address this problem by studying the relationship between a metric and segmentation acceptability in an object-dependent manner. After linearization, metrics for different objects will have more similar acceptability meaning than the original metrics.

1.3. Outline of approach

The proposed LinSEM method[1] is depicted in Fig. 1 and is described in detail in Section 2. In this method, we first estimate (model) the actual relationship between each metric and its degree of acceptability for each object O via a reader study. The relationship is estimated based on a set of simulated segmentations (S_S) of O, created from a set of true segmentations (S_T) so as to cover various degrees of segmentation qualities in S_S from excellent to unacceptable for O. To this end, we design a sequence of operations to mimic deviations between true segmentations and actual segmentations of O by using morphological and image algebraic operations. These operations are applied to true segmentations S_T of O to create S_S. A reader study is then conducted wherein the reader assigns an acceptability score (abbreviated as AS) $a(w)$ for each segmentation sample w in S_S, expressing the degree of acceptability of w in the subjective opinion of the expert on a 1 to 5 scale. Metric values $m(w)$ are also calculated for these segmentations for each metric of interest.

We estimate a *probabilistic acceptability score* $a_P(r)$ for each metric value r. Then, the relationship between the metric and AS of each considered object is constructed from pairs of metric values and their corresponding $a_P(r)$ by sequentially linking these pairs in a piece-wise linear manner. With the idea that the ideal metric should be linear with respect to acceptability, we can then linearize any given metric value of the object under consideration to its linearized value. Correction factors $\kappa_{m,O}(r)$ are estimated for O which indicate how a given metric value r of O resulting from any algorithmic segmentation should be corrected for it to be linearized by using the metric-AS relationship curve for O. We then devise a method to transform the metric values to linearized values based on this estimated correction factors and test on a set S_A of segmentations of O created by an actual segmentation algorithm.

Section 3 describes experiments conducted using three metrics (*DC*, *JI*, and *HD*) and five anatomic objects defined in computed tomography (CT) images of the head and neck (H&N) and thoracic body regions of cancer patients undergoing radiation therapy. Over 3000 slices in total from 100 3D segmentation samples are involved in our reader experiments. The segmentation samples in S_A are obtained via the AAR method (Udupa et al., 2014; Wu et al., 2019). In Section 4, we evaluate the effectiveness of LinSEM in three ways: (i) by assessing the similarity of acceptability among different objects for the same metric value before and after linearization; (ii) by assessing the deviation of the object's acceptability scores from the ideal values before versus after linearization for each object; and (iii) for each object and for each theoretical acceptability value, the closeness of the metric value achieved by linearization to the value corresponding to the ideal curve. Our conclusions, gaps remaining in this work, and avenues for potential improvements are discussed in Section 5.

2. Method

Notations:

O: An anatomical object.

S_A, S_S, S_T: Respectively, a set of actual segmentations via algorithms, a set of simulated segmentations, and a set of true segmentations used for simulation.

$a(w), m(w), m_l(w)$: Respectively, acceptability score, metric value, and linearized metric value associated with a segmentation sample w.

$a_P(r)$: Probabilistic acceptability score estimated for metric value r.

$\kappa_{m,O}(r)$: Correction factor for metric m at its value r for object O.

$G(m, O, W)$: Plot of $\{(m(w), a(w))\}$ for metric m and object O, determined from segmentation set W.

[1] Although very different, LinSEM is reminiscent of intensity standardization methods developed in the 1990's to handle MR image intensity non-standardness (Nyul and Udupa, 1999).

$g_{m,O,W}(r)$: Estimated function describing *AS*-metric relationship of metric m for object O determined from segmentation set W.

$G_l(m,O,W,Q)$: Plot of $\{(m_l(w), a(w))\}$ for metric m and object O, determined from segmentation set Q, where linearization is based on the segmentation set W.

$h_{m,O,W,Q}(r)$: Linearized function describing *AS*-metric relationship of metric m for object O, determined from segmentation set Q, where linearization is based on the segmentation set W.

$\psi(O_1,O_2,m,r), \psi_L(\cdot), \psi_g(\cdot)$: Respectively, semantic similarity of metric m at its value r between objects O_1 and O_2 before (ψ) and after (ψ_L) linearization and gain (ψ_g) in linearization.

$\rho(O,m,r)$, $\rho_L(.)$, $\rho_g(.)$: Respectively, closeness of the acceptability of object O for metric m at its value r with ideal *AS* before (ρ) and after (ρ_L) linearization and gain (ρ_g) in linearization.

$\gamma(O,m,r)$, $\gamma_L(.)$, $\gamma_g(.)$: Respectively, closeness of the metric value of object O for metric m with ideal metric value before (γ) and after (γ_L) linearization and gain (γ_g) in linearization.

Our intent is that the linearization model needs to be developed only once for any object O (such as liver) for a given metric such as *DC*. Subsequently, for any given segmentation of O in any given image by any algorithm, it should be possible to apply the linearization correction for that metric to this segmentation to obtain the linearized value of the metric. In other words, we assume (see further comments in Section 5) that the linearization process would depend only on the metric and the object. For this to be valid, a standard definition of O should be adopted, which implies that the body region housing O should also be unambiguously defined. To make this point clear, consider O to be cervical esophagus. For this object to be anatomically defined consistently in any image of any subject, the H&N body region in which it is housed should be first clearly defined, especially regarding its superior and inferior axial boundary plane locations. Otherwise, this object may vary in its very definition from case to case due to its varying extents in the cranio-caudal direction. Similarly, what is included in the anatomic object named O and what is excluded should also be clearly specified. For example, when O = liver, including or excluding the hepatic portal system (at least its major vessels) in the definition of O would make a significant difference in the complexity of the shape of O which may influence the linearization process. Therefore, as in our previous work on automatic anatomy recognition (Udupa et al., 2014; Wu et al., 2019), we assume that a standardized definition of each body region and each object considered in it is available for the LinSEM process.

The main idea of LinSEM is illustrated schematically in Fig. 2 where two different objects O_1, and O_2 are shown to have different *DC-AS* curves, and two *DC* values – d_1 for O_1, and d_2 for O_2 – both correspond to the same acceptability score A. Alternatively, the same *DC* value M may also indicate different acceptability meaning for the two objects as illustrated in the figure. We take the *DC-AS* curves as reference, and after linearization, we would like the same *AS* value, for both considered objects, to correspond to the same *DC* value. That is, the two *DC-AS* curves should be linearized to the ideal curve (diagonal line). So, as shown in the figure, d_1 for O_1 and d_2 for O_2 will be both linearized to M, which is the DC value with an *AS* of A on the ideal curve. Unfortunately, metric-*AS* relationships based on empirical *AS* values determined from reader studies do not present as smooth curves or even functions, and are generally 2D graphs or plots (see Fig. 3). So, first we need to estimate a function that fits this 2D graph, which can then be used to linearize the relationship. Consequently, LinSEM is composed of two main parts: (i) estimating relationships between metric and acceptability for all considered metrics and objects, and (ii) linearizing metric values in given segmentation samples. These parts are described in Sections 2.1 and 2.2, respectively. Throughout, we assume that, there is an object O and a segmentation evaluation metric m (which is one of *DC*, *JI*, and *HD* in this paper) under consideration. Even when these entities are not mentioned explicitly, the reference to a specific object O and metric m is to be understood.

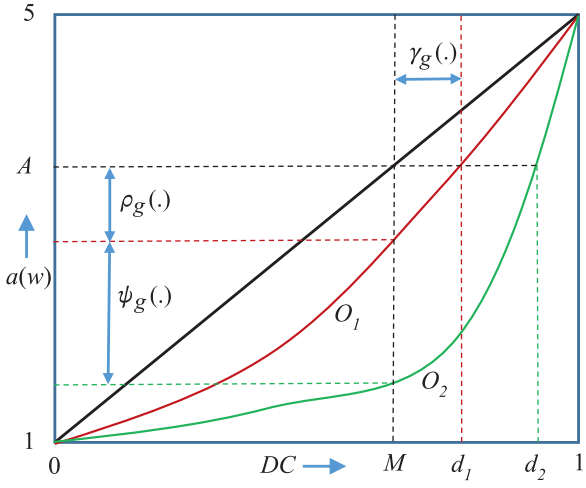

Fig. 2. Hypothetical *DC-AS* curves for two objects O_1 and O_2 are illustrated where the same acceptability score A corresponds to different DC values d_1 and d_2. The goal of LinSEM is to map these values as closely as possible to the ideal value M. $\psi_g(.)$, $\gamma_g(.)$, $\rho_g(.)$ are three measures employed to evaluate the effectiveness of LinSEM, which will be described in Section 2.3.

2.1. Estimating metric-AS relationship

2.1.1. Generating set S_S for object O by simulating segmentations

To obtain metric-*AS* relationship, we need segmentation samples with qualities covering the full spectrum from excellent to unacceptable. Segmentations output by algorithms usually do not cover the whole range of qualities. For example, some well-defined and non-sparse objects such as skin outer boundary in a body region are easy to segment by algorithms, their samples will have $AS = 4$ or 5 and will not include cases of $AS = 1$ or 2. Conversely, sparse objects such as esophagus which are difficult to segment rarely cover cases with $AS = 5$. For creating a segmentation set with diverse degrees of quality and mimicking segmentations by different algorithms with different quality behavior and enough samples, we create a set of simulated segmentations, denoted S_S. The simulation process is composed of three steps:

Step 1: Collect a set of images which appear radiologically near-normal for the body region of interest and create the ground truth segmentations of O for these images[2] following the definitions of O. These segmentations will be denoted by set S_T.

Our idea is to design sequences of morphological and image algebraic operations which when applied to segmentations in S_T would create S_S. We decided to perform these operations in a 2-dimensional manner within the xy-plane of the axial images for several reasons (see Section 5 for further comments). First, from the human reader's perspective, because of the mode of slice visualization used for close and detailed scrutiny in radiological tasks, we decided it is best to generate the deviations also in a 2D manner. Second, for the same reason, it is easier for the reader to judge the quality of a segmentation more consistently on the individual slices than to examine all slices and then to judge the quality as a single score for the whole 3D volume. Third, a true

[2] LinSEM is applicable to any set of images and not just near-normal. We believe that it is better to understand the metric-acceptability relationship first on near-normal objects before applying to objects with abnormal or distorted shapes.

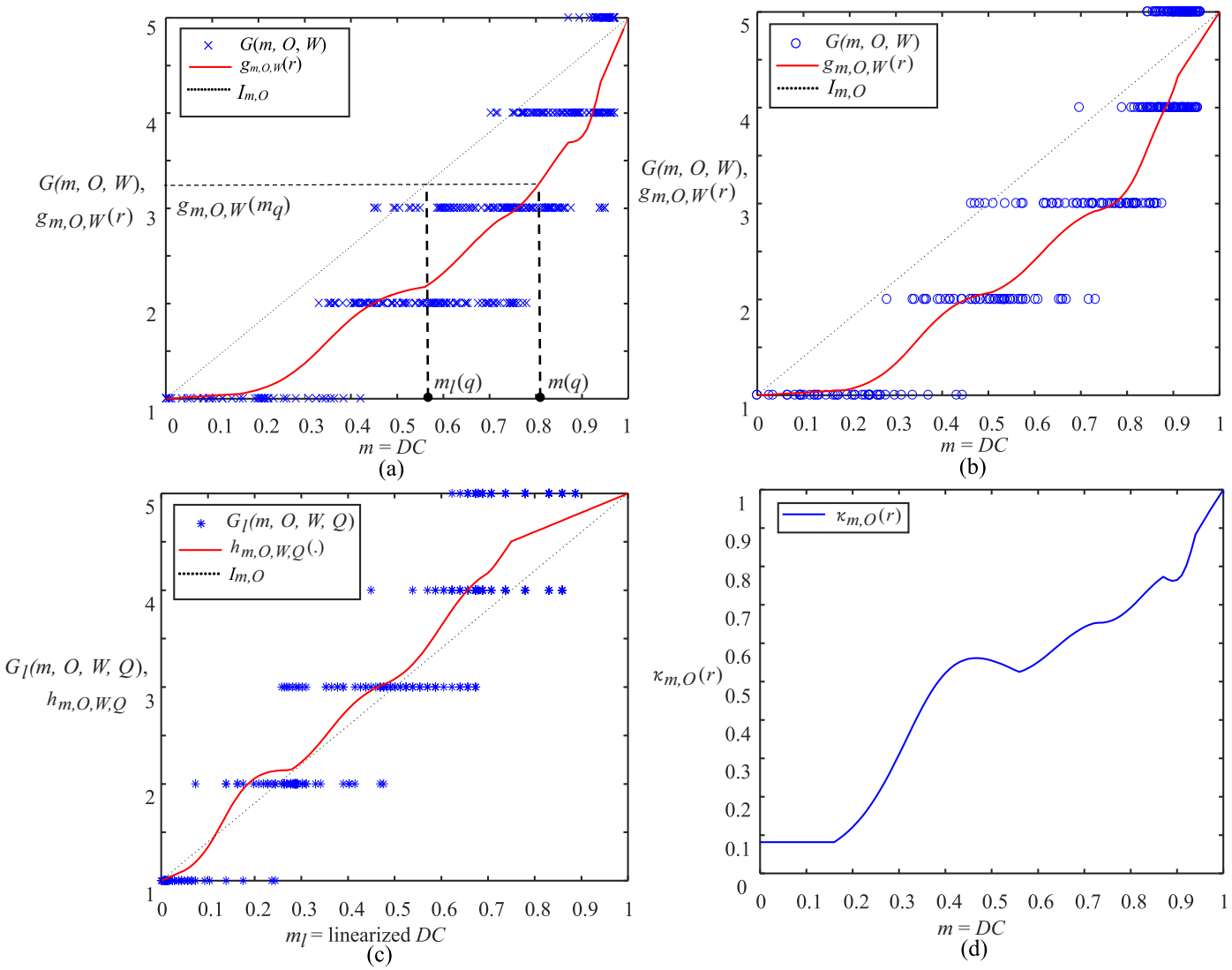

Fig. 3. Illustration of the process of estimating metric-*AS* relationship for the object O = Mandible (Mnd). (a) The blue marks denote the plot $G(m, O, W)$ of raw metric-value-empirical-*AS*-value pairs for metric $m = DC$ and the simulated set $W = S_S$. The smooth (red) curve represents the estimated function $g_{m,O,W}(r)$. (b) Similar to (a) but for $W = S_A$. (c) The plot $G_l(m, O, W, Q)$ of linearized metric-*AS* pairs (blue marks) for $W = S_S$ and $Q = S_A$ and the fitted linearized function (red) $h_{m,O,W,Q}(.)$. (d) The correction factor $\kappa_{m,O}(r)$ for the samples in $Q = S_A$ estimated by using the fitted curve $g_{m,O,W}(r)$ where $W = S_S$ (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

3D reader study would involve overall many more slices than a 2D reader study and would quickly become very time-consuming and impractical. Finally, since most acquired images do not possess isotropic resolution, we did not want the 3D simulation process to introduce its own vagaries that may treat the z-dimension (orthogonal to the xy-plane) differently from the other two dimensions.

Each sequence of operations we designed is composed of shift (S), dilation (D), and erosion (E) operations. Each of these operations may be performed in x or y or both directions. The magnitude of the operation is expressed in *strides* which in turn is expressed in number of pixels. A sequence is composed of a set of basic operations. The basic operations are expressed as:

$$\begin{aligned} &\pm x - S - n, \quad \pm y - S - n, \quad \pm x \pm y - S - n, \\ &x - D - n, \quad y - D - n, \quad xy - D - n, \\ &x - E - n, \quad y - E - n, \quad xy - E - n. \end{aligned} \tag{1}$$

where n denotes the number of strides, $\pm x - S - n$ denotes two operations – shift in the $+x$ or -x direction, and other operations involving S are similarly defined. x- D-n denotes symmetric dilation in the x-direction, and other operations involving D are similarly defined. x-E-n denotes symmetric erosion in the x-direction; other operations involving E are similarly defined. For example, -x+y-S-2 with a stride = 3 pixels denotes a shift in the -x direction by 2 strides (= 6 pixels) followed by a shift in the +y-direction by 2 strides (= 6 pixels). These basic operations are combined to create sequences. Example:

$$xy - D - 2 \rightarrow +x - S - 3. \tag{2}$$

This sequence consists of an initial dilation by 2 strides in the x- and y-directions, followed by a shift by 3 strides in the $+x$-direction. We express the deviation of a segmentation sample w in S_S from its ground truth counterpart w_T in S_T resulting by applying a sequence to w_T by the maximum number δ of pixels of deviation. In the above example in Eq. (2), if a stride is 3 pixels, then the resulting sample w will have a deviation of $\delta = 15$ pixels. We have designed a set of sequences as shown in Table 1 (Section 3) which we employ to simulate segmentations with very small to large and realistic deviations.

Step 2: The stride values utilized are estimated in an object-specific manner according to the thickness of the object sample.

Table 1
Sequences employed to simulate segmentations and their associated deviations δ expressed in stride size.

Sequence	δ	Sequence	δ
+xy-*S*-1	1	xy-*D*-6 → +*x*-y-*S*-10	16
xy-*E*-1 → -x-*S*-1	2	xy-*E*-6 → -*x*++*y*-*S*-11	17
xy-*E*-1 → +xy-*S*-2	3	xy-*E*-8 → +*x*-*S*-10	18
xy-*D*-2 → +*y*-*S*-2	4	xy-*D*-5 → -y-*S*-14	19
xy-*E*-2 → +*x*-*S*-3	5	xy-*D*-4 → +xy-*S*-16	20
xy-*D*-2 → -y-*S*-4	6	xy-*E*-7 → +*y*-*S*-14	21
xy-*D*-3 → -x-*S*-4	7	xy-*D*-8 → +*x*-*S*-14	22
xy-*E*-3 → +*x*-y-*S*-5	8	xy-*D*-6 → -xy-*S*-17	23
xy-*D*-5 → -*x*++*y*-*S*-4	9	xy-*E*-5 → +xy-*S*-19	24
xy-*E*-4 → -xy-*S*-6	10	xy-*D*-6 → -*x*++*y*-*S*-19	25
xy-*D*-5 → +xy-*S*-6	11	xy-*E*-5 → +*x*-y-*S*-21	26
xy-*D*-4 → +*y*-*S*-8	12	xy-*E*-4 → -y-*S*-23	27
xy-*E*-6 → +*x*-*S*-7	13	xy-*D*-3 → -x-*S*-25	28
xy-*D*-7 → -y-*S*-7	14	xy-*D*-5 → -*x*++*y*-*S*-24	29
xy-*E*-5 → -x-*S*-10	15	xy-*E*-4 → +*x*-y-*S*-26	30

The reason for not designing deviations in units of pixels or millimeters is that, for objects of different sizes and different shape, the same magnitude of deviation in pixels may not result in a similar change in quality. For illustration, a small object with a thickness of $2b$ pixels may disappear after symmetric erosion by b pixels, and a large object with a thickness of $10b$ pixels will still be $8b$ pixels thick after erosion, which may not constitute a significant change in quality.

Since erosion is the limiting operation determining the disappearance or degeneration of an object, we set a limit defined by a parameter θ to denote the fraction of the thickness of an object to which we allow to diminish (by erosion). Let T_{min} be the minimum thickness of O (in pixels) over all its samples in S_T, let t_m be a sample of S_T with minimum thickness, and let n_{max} be the maximum number of strides allowed for erosion. Our idea is that the number of strides n selected for morphological operation in any sequence in Table 1 should be less than n_{max}. The size of a stride (in terms of number of pixels) permitted for sample t_m will be

$$e_m = \frac{\theta T_{\min}}{2n_{\max}}. \tag{3}$$

If $e_m < 1$, it implies that the stride size is less than 1 pixel for t_m, and so this triggers (shape-based) interpolation (Raya and Udupa, 1990) to be performed on all samples of S_T. Interpolation is done in such a manner that the pixel size p_s of t_m (in mm) is changed to p_o after interpolation and the new thickness of t_m (in pixels) becomes T_o.

$$p_o = p_s \times e_m,$$
$$T_o = \frac{T_{\min}}{e_m}. \tag{4}$$

For other samples of S_T, their pixel size and thickness also change per factor e_m as in Eq. (4), and the stride size for each sample is calculated as in Eq. (3). Note that due to the manner in which the stride size and interpolation factor are determined, the new stride size after interpolation for t_m becomes 1 and the thickness (in mm) of all samples remains the same. If $e_m \geq 1$, then there is no need for interpolation.

Step 3: Apply the sequences as per calculated strides to all samples of S_T to create the simulated segmentation set S_S.

We also generated segmentations of O output by our AAR algorithms (Wu et al., 2019) for a set of patient images (which are different from the near-normal data sets used for generating S_S and S_T) to compose set S_A. We also created ground truth segmentations of O for these data sets following our standardized object definitions so that the different metric values for the samples in S_A can be calculated. Set S_A will be used for testing the linearization process of LinSEM.

2.1.2. Reader study to determine acceptability score AS

In our reader study, a radiologist (co-author DAT) with 22 years of experience in various radiological tasks involving image analysis determined the acceptability *AS* of segmentations. The reader examined each slice of a segmentation, which was displayed as an overlay on to the corresponding CT slice image, and assigned an *AS* value to each slice on a 1 to 5 scale, with 1 denoting unacceptable or poor segmentation and 5 representing excellent segmentation. The reader was blinded to the ground truth segmentations, and thus, acceptability scores were assigned based only on clinical knowledge and not influenced by the comparison with ground truth. The reader study was conducted on both sets S_S and S_A.

The standard for *AS* assignment is hard to express in formulation because *AS* values show comprehensive concerns of the expert, which encapsulate the size, shape, anatomical relationship of objects, and clinical importance. As a result, the standard for *AS* may be object (and application) dependent, which further suggests that it may not be possible to evaluate qualities of segmentations simply and easily by computational metrics.

We will use the following notations for simplifying our description. Let $m(.)$ denote one of the metrics *DC, JI,* and *HD*, and for any segmentation sample w, $m(w)$ will denote the value of that metric for w. Let $a(w)$ denote the acceptability score assigned to w in the reader study. Since we will perform experiments involving both S_S and S_A, we will use $W \in \{S_S, S_A\}$ to denote the set under consideration. For $W \in \{S_S, S_A\}$ and $m \in \{DC, JI, HD\}$, the reader study generates a 2D graph or plot which we will denote by $G(m, O, W) = \{(m(w), a(w))\}$. Fig. 3a and b show an example of $G(DC, O, S_S)$ and $G(DC, O, S_A)$, respectively, where the object O is Mandible.

2.1.3. Constructing metric-AS relationship

The metric-*AS* curve to be constructed (modeled) is intended to show the relationship between metric values and acceptability scores as a function for the considered object. However, there is a challenge arising from the fact that the empirical *AS* values are discrete and the computed metric values are continuous, resulting in $G(m, O, W)$ being a 2D graph as illustrated in Fig. 3. Notably many different metric values may be assigned the same *AS* value. Conversely, segmentation samples with the same metric value may be assigned different *AS* values according to clinical factors which may not be adequately reflected by metric values. Thus, metric-*AS* curves do not directly emerge from $G(m, O, W)$, although we can intuitively understand the rough tendency of the metric-*AS* relationship from such plots. See Fig. 3.

We overcome this ambiguity by estimating a *probabilistic acceptability score*, denoted $a_P(r)$, via the concept of Mahalanobis distance (Mahalanobis, 1936) determined for each possible metric value r over the whole range of the considered metric $m(.)$. Mahalanobis distance is a measure of the distance from a point to a probability distribution. The metric values (as random variables) corresponding to each *AS* value have their own distribution. For a metric value r, we measure its Mahalanobis distance to the metric-value distribution corresponding to each discrete *AS* value. A small value of this distance implies higher probability and a large value suggests lower probability that a segmentation sample with metric value r should be assigned this *AS* value. Mahalanobis distance values are calculated for each *AS* value and are taken as weight factors in the estimation of $a_P(r)$. The resulting measure $a_P(r)$ can assume any real acceptability value in the range [1,5]. Since 5 discrete levels are the finest resolution usually employed in reader studies to specify a grade for the phenomenon under observation, *AS* values assigned by the reader have to be integers in {1, 2, 3, 4, 5}. The estimated probabilistic acceptability score $a_P(r)$ for each r, however, is in the continuous range [1, 5].

Let $D_M(r, i)$ denote the Mahalanobis distance of a specific metric value r to the distribution p_i of metric values corresponding to

$AS=i$ and let $\mu(i)$ and $\sigma(i)$ denote the mean and standard deviation of this distribution p_i. $a_P(r)$ is estimated by the weighted average of the AS values, where the weight given to an AS value is the reciprocal of the exponential of $D_M(r, i)$ to reflect the fact that larger distance should indicate lower probability. In this way, continuous and probabilistically estimated acceptability scores $a_P(r)$ result for every possible metric value as expressed in Eq. (5)[3], and each distinct metric value r is represented by exactly one $a_P(r)$ value.

$$a_P(r) = \begin{cases} 1 + \frac{r}{\mu(1)}(H(\mu(1)) - 1), & \text{if } r<\mu(1), \\ H(\mu(5)) + \frac{r-\mu(5)}{1-\mu(5)}(5 - H(\mu(5))), & \text{if } r>\mu(5), \\ H(r), & \text{otherwise,} \end{cases}$$

$$\text{where} \quad H(r) = \frac{\sum_{i=1}^{5} i \times \exp(-D_M(r,i))}{\sum_{i=1}^{5} \exp(-D_M(r,i))},$$

$$\text{and } D_M(r,i) = \frac{|r-\mu(i)|}{\sigma(i)}. \tag{5}$$

For extreme cases where metric value r is 0 or 1, we take them as the most unacceptable or the best possible cases and directly assign them $a_P(0)=1$ or $a_P(1)=5$, respectively. For cases where r is in the range $[0, \mu(1)]$ or $[\mu(5), 1]$, we consider them as having a linear relationship from point (0, 1) to $(\mu(1), H(\mu(1)))$ or $(\mu(5), H(\mu(5)))$ to (1, 5) on the plot, respectively. For implementation, we discretize the metric value range, and for each discretized metric value r, estimate $a_P(r)$. In estimating $a_P(r)$, we exclude those r values for each i for which $D_M(r, i) > D_{max}$, the idea being that a large Mahalanobis distance value indicates an outlier and potentially highly improbable AS value. The metric-acceptability relationship curve as a function, denoted by $g_{m,O,W}(.)$, is then created by piecewise linear linking of the discrete $(r, a_P(r))$ pairs. $g_{m,O,W}(r)$ is then defined for any real value of r in [0, 1]. In Fig. 3a and b, we demonstrate $g_{m,O,W}(r)$ for $W=S_S$ and $W=S_A$, respectively, where the object is Mandible and $m=DC$. Notably the curves seem to aptly express the underlying plots.

Three metrics are considered in this work: Dice coefficient (DC), Jaccard index (JI), and a normalized version of Hausdorff Distance (HD_N). All metrics are computed for 2D segmentations on slices since our reader study assigning acceptability scores is carried out on slices. DC and JI are commonly-used metrics (Eq. (6)). When considering Hausdorff Distance (HD), there are two issues: (i) Unlike DC and JI which are fractions lying in [0, 1], HD is not a ratio (hence not normalized) and is measured in physical units. Hence, its worst possible value (maximum distance from true boundary) has no easily definable bound although the best possible value is 0. (ii) Minute false positives in segmentations, such as isolated pixels or small clusters of pixels that lie far away from the true object which do not influence AS, may pose a challenge for normalizing HD. To overcome these issues, we use a median version, instead of maximum, for HD and normalize HD to arrive at HD_N as described below.

To normalize HD for an object O, we use the maximum value HD_{M2} of HD among all samples of O for which $AS=2$ as a normalizing factor. If HD of a segmentation on a slice is greater than HD_{M2}, we may infer that this segmentation is of really unacceptable quality and the HD_N value should be set to 1. HD_N is calculated as in Eq. (7) where HD_{M2} should be determined separately for each considered object O. For a segmentation sample, large DC and JI (~ 1), and small HD_N (~ 0) mean good quality, and small DC and JI (~ 0), and large HD_N (~ 1) suggest unacceptable quality. So when conducting linearization on HD_N, a slight modification should be made to Eq. (5) where $\mu(1)$ and $\mu(5)$ interchange their roles for cases where $r < \mu(5)$ or $r > \mu(1)$. In Eq. (6) and (7), $w \in W$ denotes the segmentation sample to be assessed ($W \in \{S_S, S_A\}$), w_T denotes the corresponding true segmentation, and $\beta(w)$ and $\beta(w_T)$ denote the boundaries of samples w and w_T, respectively.

$$DC(w, w_T) = \frac{2 \times TP(w, w_T)}{2 \times TP(w, w_T) + FP(w, w_T) + FN(w, w_T)},$$

$$JI(w, w_T) = \frac{TP(w, w_T)}{TP(w, w_T) + FP(w, w_T) + FN(w, w_T)}, \tag{6}$$

$$HD(w, w_T) = median\left(\left\{ \inf_{y \in \beta(w_T)} \{d(x,y) | x \in \beta(w)\} \right\} \cup \left\{ \inf_{x \in \beta(w)} \{d(x,y) | y \in \beta(w_T)\} \right\} \right), \tag{7}$$

$$HD_N(w, w_T) = \min\left(\frac{HD(w, w_T)}{HD_{M2}(w, w_T)}, 1 \right).$$

2.2. Linearizing metric values

Function $g_{m,O,W}(r)$ can be used to linearize the value of m for any given segmentation sample q of O as follows. Let the ideal metric-acceptability curve (diagonal line) be denoted by $I_{m,O}(r)$. The linearized metric value $m_l(q)$ of q is then obtained by simply projecting the point $(m(q), a(q))$ on the curve of $g_{m,O,W}(r)$ on to the ideal curve and reading off the corresponding metric value $m_l(q)$ as shown in Fig. 3a. Thus,

$$m_l(q) = I^{-1}_{m,O}(g_{m,O,W}(m(q))). \tag{8}$$

In particular, we can use $g_{m,O,W}(r)$ to linearize metric values of segmentation samples of O coming from another set $Q \neq W$. For example, we may create $g_{m,O,W}(r)$ from $W=S_S$ and then use this to linearize $Q=S_A$. The resulting pairs $(m_l(q), a(q))$ of linearized metric-values $m_l(q)$ and acceptability scores $a(q)$ (assigned in a reader study) again constitute a 2D graph (and not necessarily a function), which will be denoted by $G_l(m, O, W, Q)$. On $G_l(m, O, W, Q)$, we may again use the above fitting method to determine a function that will portray the "linearized curve" for the samples in Q. We will denote this function by $h_{m,O,W,Q}(.)$. Fig. 3c illustrates the plot $G_l(m, O, W, Q)$ and the linearized curve $h_{m,O,W,Q}(.)$ for $W=S_S$ and $Q=S_A$. Note that when $Q=W$, $G_l(m, O, W, Q)$ will represent a plot where the marks in Fig. 3a are all shifted (nonlinearly) to align closely around the diagonal, and the resulting fitted curve $h_{m,O,W,Q}(.)$ will be mostly a diagonal line, within computational approximations. Comparing fitted curves $g_{m,O,Q}(r)$ (Fig. 3b) and $h_{m,O,W,Q}(r)$ (Fig. 3c) for $W=S_S$ and $Q=S_A$ and the associated plots, it is clear that the distribution of samples is better centered around the ideal curve, and the fitted curve after linearization is closer to the ideal curve than before linearization.

Note that the ideal curve is different for metrics which are based on similarity versus dissimilarity. After all metrics are normalized from their original range to [0, 1], for metrics DC and JI which evaluate similarity of segmentations and their ground truth, the ideal curve is the linear line from point (0, 1) to (1, 5), indicating that low similarity means unacceptable quality and high similarity implies excellent quality. For metrics which evaluate the deviation between segmentations and their ground truth, such as HD_N (and other metrics like False Positive and False Negative Volume Fractions not considered in this paper), the ideal curve is the linear line from point (0, 5) to (1, 1) indicating that low deviation means excellent quality and high deviation suggests unacceptable quality.

To make the linearization process convenient to use, object-specific correction factors $\kappa_{m,O}(r)$ are computed for each metric m which indicate how a given value r of m should be corrected multiplicatively for it to be linearized by using the metric-AS relationship curve $g_{m,O,W}(r)$ for the object under consideration. The

[3] These equations are fashioned for DC and JI. For HD, changes are made along similar lines.

correction factor is given by (see Fig. 3a)

$$\kappa_{m,O}(r) = \frac{m_l(q)}{m(q)} \tag{9}$$

For any test segmentation sample t of an object with metric value $m(t)$, its corresponding linearized metric $m_l(t)$ value is then given simply by the product $m(t) \times k_{m,O}(m(t))$. For the illustration in Fig. 3, the $\kappa_{m,O}(r)$ curve is displayed in Fig. 3d for the samples of object Mandible in the set $Q = S_A$ and $m = DC$, where the fitted curve was estimated from the set $W = S_S$.

The LinSEM methodology as a whole has three parameters: θ denoting the fraction of the minimum thickness of an object we allow to diminish after erosion, n_{max} representing the maximum number of strides allowed for erosion, and a threshold D_{max} on Mahalanobis distance $D_M(.)$ that is used for detecting outliers. The first two parameters are associated with the method of simulating segmentations, and the third parameter relates to the method of curve fitting.

2.3. Evaluating the effectiveness of LinSEM

Since LinSEM aims to harmonize acceptability-meaning among different objects, to evaluate its effectiveness, we check whether metric values have more similar meaning among different objects after linearization. We collect another set of segmentations as a test set, for which we will check whether differences of metric-*AS* curves among considered objects are narrowed after linearization. The evaluation process for each given metric m comprises of four steps:

Step 1: Collect a set of (test) segmentations S_A produced by an algorithm based on image data sets from a set of different subjects for each of a set of different objects. For each such object, create simulated segmentations S_S based on the set S_T of true segmentations coming from image sets of subjects different from those whose data sets yielded set S_A. In other words, for each object, sets S_A and S_S constitute completely disjoint sets of subjects and hence image and object samples.

Step 2: Conduct reader studies for the samples in sets S_A and S_S for each object.

Step 3: From $G(m, O, S_A)$ and $G(m, O, S_S)$ for each considered object O, determine the fitted curves $g_{m,O,S_A}(r)$ and $g_{m,O,S_S}(r)$ showing the variability of (probabilistic) acceptability as a function of m for S_A and S_S, respectively, before linearization. Estimate curve $h_{m,O,S_S,S_A}(r)$ after linearization of the metric values of the samples in S_A by using $g_{m,O,S_S}(r)$ for each object O.

Step 4: If curves $h_{m,O,S_S,S_A}(r)$ for different objects are more similarly distributed compared to curves $g_{m,O,S_A}(r)$ for these objects, and if they are closer to the ideal curve, we may conclude that the linearized metric, compared with the original metric, has more similar acceptability meaning among objects, and the LinSEM method is effective.

We employ three types of evaluation measures, denoted by ψ, ρ, and γ, to assess the distribution of the linearized-metric-*AS* curves $h_{m,O,S_S,S_A}(r)$ for different objects. For two objects $O_1 \neq O_2$ and any given value $r \in [0, 1]$ of metric m or its linearized version m_l, we define the *semantic dissimilarity* in m between O_1 and O_2 prior to linearization (ψ) and post-linearization (ψ_L) by

$$\begin{aligned} \psi(O_1, O_2, m, r) &= \left| g_{m,O_1,S_A}(r) - g_{m,O_2,S_A}(r) \right| \\ \psi_L(O_1, O_2, m, r) &= \left| h_{m,O_1,S_S,S_A}(r) - h_{m,O_2,S_S,S_A}(r) \right|, \end{aligned} \tag{10}$$

and the *gain in sematic similarity* by

$$\psi_g(O_1, O_2, m, r) = \psi(O_1, O_2, m, r) - \psi_L(O_1, O_2, m, r) \tag{11}$$

We expect $\psi_g(.) > 0$ or $\psi(O_1, O_2, m, r) > \psi_L(O_1, O_2, m, r)$ for most $r \in [0, 1]$, or the mean value of $\psi_g(.)$ over all r to be positive.

The second measure $\rho_g(.)$ analogously describes the gain in closeness of the acceptability score to the ideal value from pre-linearization to post-linearization. We define the *closeness* of acceptability to the ideal value prior to (ρ) and post-linearization (ρ_L) and the *gain in closeness* due to linearization by

$$\begin{aligned} \rho(O, m, r) &= |g_{m,O,S_A}(r) - I_{m,O}(r)|, \\ \rho_L(O, m, r) &= |h_{m,O,S_S,S_A}(r) - I_{m,O}(r)|, \\ \rho_g(O, m, r) &= \rho(O, m, r) - \rho_L(O, m, r). \end{aligned} \tag{12}$$

Again, we expect $\rho_g(.) > 0$ or $\rho(O, m, r) > \rho_L(O, m, r)$ for most $r \in [0, 1]$, or the mean value of $\rho_g(.)$ over all r to be positive.

The third measure $\gamma_g(.)$ is similar to $\rho_g(.)$ but describes the gain in closeness of the metric value to the metric value on the ideal curve. We define the *closeness* of the metric value to the ideal value prior to (γ) and post-linearization (γ_L) and the *gain in closeness* due to linearization by

$$\begin{aligned} \gamma(O, m, r) &= |r - I^{-1}_{m,O}(g_{m,O,S_A}(r))|, \\ \gamma_L(O, m, r) &= |h^{-1}_{m,O,S_S,S_A}(g_{m,O,S_A}(r)) - I^{-1}_{m,O}(g_{m,O,S_A}(r))|, \\ \gamma_g(O, m, r) &= \gamma(O, m, r) - \gamma_L(O, m, r). \end{aligned} \tag{13}$$

We expect $\gamma_g(.) > 0$ or $\gamma(O, m, r) > \gamma_L(O, m, r)$ for most $r \in [0, 1]$, or the mean value of $\gamma_g(.)$ over all r to be positive. See Fig. 2 for a pictorial depiction of the meaning of these three measures.

3. Experiments

3.1. Data sets

This retrospective study was conducted following approval from the Institutional Review Board at the Hospital of the University of Pennsylvania along with a Health Insurance Portability and Accountability Act waiver. Experiments are conducted on CT images of two body regions, H&N and Thorax. The following five objects as defined in Wu et al. (2019) are considered: the outer skin boundary of the H&N (cervico-thoracic) body region superior to the shoulders (CtSkn-h), right parotid gland (RPG), mandible (Mnd), cervical esophagus (CtEs), and heart (Hrt). The full name and the acronym for these objects are listed in Table 2 for ready reference. The first four objects are from the H&N region and the 5th object is from the thoracic region. The objects have been selected to represent a mix of different shapes and sizes. CtEs is a thin and narrow spatially sparse object. CtSkn-h, RPG, and Hrt are non-sparse blob-like objects. Mnd is a hybrid between these two types. Furthermore, CtSkn-h and Hrt are large objects with large thickness, and RPG, Mnd, and CtEs have relatively low thickness.

The set S_T of true segmentations employed for generating S_S was chosen from images of subjects wherein the shape of the object O considered was not affected significantly due to an abnormality for making sure that we are dealing with roughly the same shape in the samples of O contained in S_T (see further comments in Section 5). Since CT scans of H&N and Thorax regions are commonly separately acquired, it is hard to find images of these two body regions from the same subjects, and object samples for the two regions come from different subjects, although object samples for the same body region are selected from images of the same subjects. The set S_T of true segmentations was created by strictly following our body region and object definitions. The set S_A of actual segmentations is derived from the output of AAR methods (Udupa et al., 2014; Wu et al., 2019). The pixel size and slice spacing of the CT data sets which produced S_A were 1–1.6 mm and 1.5–3 mm, respectively. Since these data sets pertained to cancer patients undergoing radiation therapy, they contained various degrees of pathology.

Following the method of Section 2.1, we designed 30 sequences, as listed in Table 1, with deviations δ within 30 strides. Simu-

Table 2
Number of slices for the five objects considered in our reader study.

Data sets	Hrt (Heart)	RPG (Right parotid gland)	Mnd (Mandible)	CtEs (cervical esophagus)	CtSkn-h (Cervico-thoracic skin outer boundary – superior part)	Total
S_S	280	255	434	356	539	1864
S_A	271	311	353	269	509	1713
Total	551	566	787	625	1048	3577

lated segmentations for each object are created by applying these sequences to the object samples in S_T. The structuring elements for erosion and dilation both consist of the pixel plus its 4-adjacent neighbors in the 3×3 neighborhood. We set $n_{max} = 20$ which means a maximum of 20 strides are allowed for the erosion operation, and θ is set to 0.7 which means, after symmetric erosion by 20 strides, the original thickness t (in mm) will be reduced to $0.7t$. The stride size in millimeter will then be $0.7t/40$. When determining stride size in pixels, we should first find out the minimum thickness of an object among all its samples, and then check if e_m calculated by Eq. (3) is greater than 1 to decide if samples of the object need to be interpolated. If $e_m < 1$ and interpolation is needed, the original thickness T (in pixels) will be enlarged to T/e_m, and stride size in pixels will be $e = 0.7T/(40e_m)$. If $e_m \geq 1$, the stride size in pixels will be $e = 0.7T/40$. In this way, the interpolation factor e_m is the same for all samples of the object but the stride sizes are different. The resulting deviation will be $\delta \times e$ pixels based on the deviation for the designed sequence and the stride size of the object sample. For the convenience of calculation and avoiding introducing extra interpolation, the floor integer $\lfloor e \rfloor$ is selected as the stride size in generating the samples of S_S.

Since reader studies are time-consuming and expensive, we conducted them on S_S samples generated from 10 sequences and 10 S_A samples generated from 10 subjects of each body region via AAR algorithm. The selected sequences covered deviations from small to large, and samples of objects from the same body region are subjected to the same set of sequences. The reader study is thus conducted on 20 3D segmentation samples per object, or 100 3D object samples in total. For a given object, the object samples in S_S and S_A are shuffled, and while performing the study, the reader is blinded to the set (S_S or S_A) from which the data set originated and to the actual sequence used and the magnitude of the deviations. For reader visualization, the boundary contours of the object derived from the corresponding segmentation are displayed as an overlay on the corresponding CT slices of the data set. As mentioned earlier, the true segmentations are not available to the reader so as to keep decisions on scoring acceptability independent of the ground truth. The number of slices for each object and each segmentation set involved in the reader experiment is summarized in Table 2. Our experiment involved 3577 slices in total, and for each of them the reader assigned an acceptability score on a 1–5 scale.

3.2. Experiments

The LinSEM methodology as a whole has three parameters – θ, n_{max}, and D_{max}. As mentioned above, we set $\theta = 0.7$ and $n_{max} = 20$. For estimating acceptability-metric relationship (Eq. (5)), we set $D_{max} = 2$, which implies that about 95% of all samples will be considered in the linearization process if the metric values for a given acceptability score are normally distributed. These parameters are fixed once for all in the whole LinSEM process.

Metric-acceptability curves $g_{m,O,W}(r)$ are estimated for each metric m and each object O and separately based on sets $W = S_S$ and $W = S_A$. For computations involving Eqs. (6) and (7) and for fitting the curve, we discretize the metric value range [0, 1] into 100 equal intervals at increments of 0.01 which results in a total of 101 discrete values (including the end values 0 and 1). The metric-acceptability curve $g_{m,O,S_S}(r)$ derived from S_S is used as the metric-*AS* relationship to linearize metric values of samples from S_A. Curves $g_{m,O,S_A}(r)$ and $h_{m,O,S_S,S_A}(r)$ show the metric-*AS* relationship for set S_A before and after linearization. For S_A, if the deviation of curves $h_{m,O,S_S,S_A}(r)$ from the ideal curve and/or the difference among the curves for different objects is smaller than those of curves $g_{m,O,S_A}(r)$, the effectiveness of the LinSEM method is demonstrated.

Similarly, to determine how realistic our simulations are, we performed the above experiment reversing the roles of S_S and S_A.

To quantitatively assess the performance of LinSEM, we analyze the mean and standard deviation of ψ_g, ρ_g, and γ_g over all samples of S_A where linearization is performed based on S_S. Since the closeness of the linearized curves to the ideal curves also matter for each object, we also examine $\rho(.)$ and $\rho_L(.)$ (Eq. (12)) over all samples of S_A. We conduct a similar analysis over all samples of S_S where linearization is performed based on S_A.

4. Results and discussion

4.1. Image examples

In Fig. 4, we display sample images chosen from S_S for different levels of deviation (δ) where the matching images from S_T and closely matching sample images from S_A are also shown as well as the expert-assigned *AS*. The deviations observed in S_A from corresponding true segmentations can be well simulated by S_S with designed sequences, and more potential variations which have not been collected in the current S_A set can also be simulated by designing different sequences for deviation. Fig. 5 demonstrates several examples of different object samples where metric values achieved significantly improved similarity of meaning. In Fig. 5a, S_A samples of different objects with widely different *DC* values are assigned the same *AS* via reader study, and their resulting linearized *DC* (*LDC*) values are more similar to reflect the same acceptability meaning. Fig. 5b and c give two examples where same *DC* values for different objects correspond to same *AS* and the resulting *LDC* values, although different, maintain the same meaning after linearization.

4.2. Metric-acceptability curves

Curves $g_{m,O,S_A}(r)$ and $h_{m,O,S_S,S_A}(r)$ are portrayed in Figs. 6–8 for $m = DC$, *JI*, and HD_N, respectively, for set S_A. We make the following observations from these plots. (i) Compared with the original curves $g_{m,O,S_A}(r)$, linearized curves $h_{m,O,S_S,S_A}(r)$ distribute more compactly and closer to the ideal curve for all objects. Understandably, the degree of compactness achieved seems less for HD_N than for the other two metrics. (ii) As we pointed out previously, we cannot collect segmentations with diverse quality from the S_A set. An obvious case is CtSkn-h, where almost all collected samples are of good quality and are assigned $AS = 4$ or 5. That is why set S_S is needed to estimate metric-*AS* relationship and the reason for linear connection from ($\mu(4)$, $a_P(\mu(4))$) to (0, 1) (or (1, 1)). (iii) The maximum improvement seems to be in the curve for Hrt for *DC*.

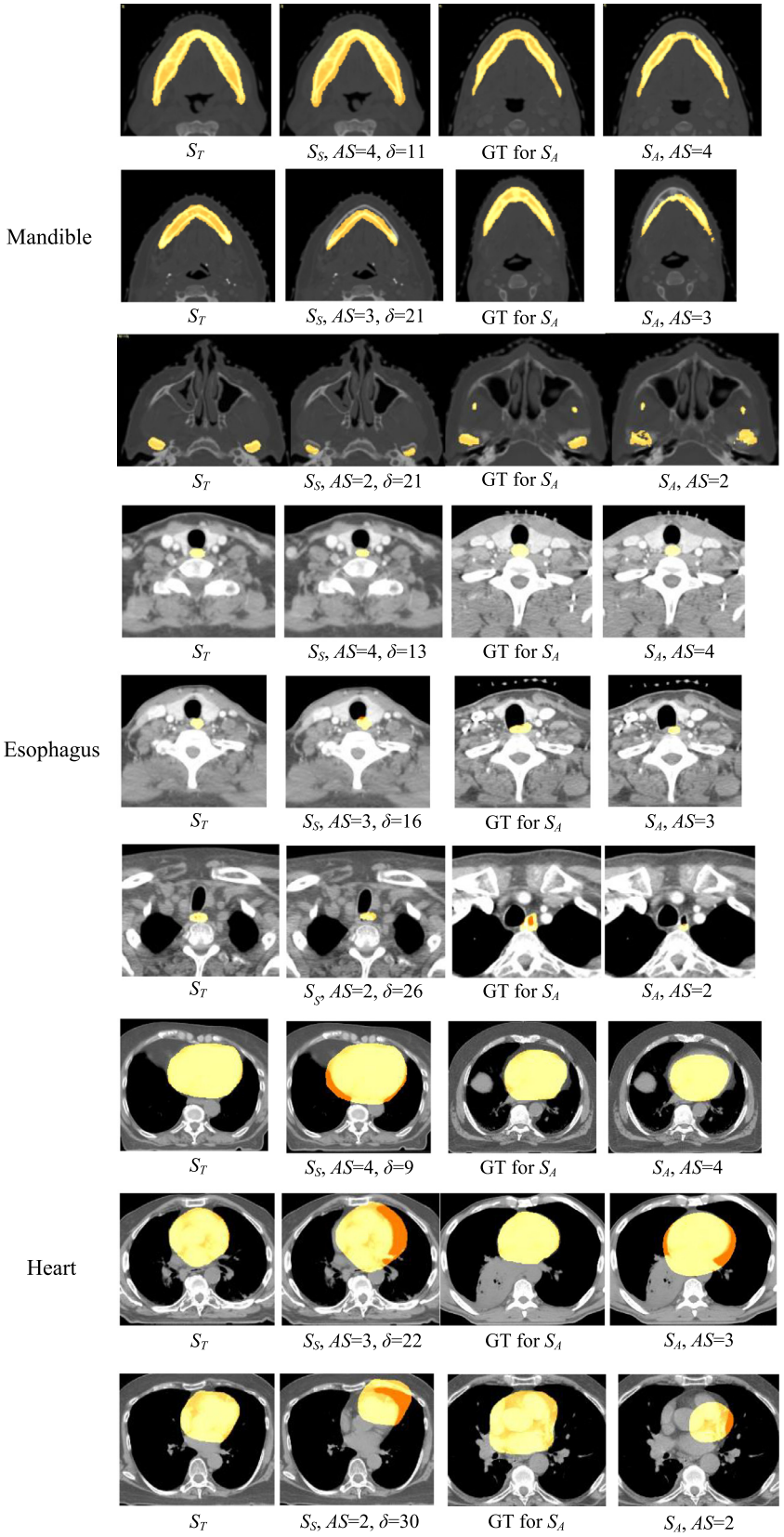

Fig. 4. Image samples from sets S_T (1st column), S_S (2nd column), S_A (4th column), and ground truth (3nd column) corresponding to S_A. For S_S, three different levels of deviations are shown (in different rows) together with the corresponding image from S_T and a closely matching sample from S_A with its ground truth. The assigned acceptability scores (AS) and designed deviation (δ) for samples of S_S are also shown. Examples displayed are for objects Mnd, CtEs, and Hrt.

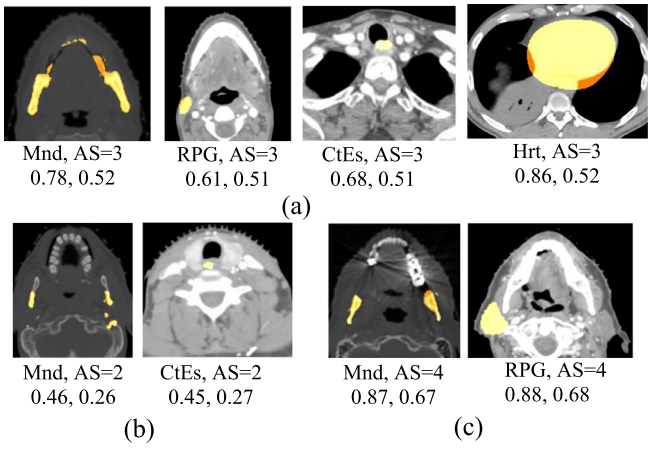

Fig. 5. Image samples from set S_A of objects Mnd, RPG, CtEs, and Hrt. Their associated *DC* values before (1st value) and after (2nd value) linearization are also shown.

CtEs and RPG are both diminutive or sparse objects and have similar *DC*-meaning before linearization, which is maintained after linearization while bringing them closer to the ideal line. Interestingly, for small *DC* (up to 0.5), Mnd, CtEs, and RPG have similar meaning and behavior before linearization. (iv) Understandably, *DC* and *JI* behave similarly and quite differently from HD_N, although *JI* seems to produce curves that are closer to the ideal line compared to *DC*, suggesting that *JI*'s behavior is more linear than that of *DC* even before linearization. After linearization, *DC* and *JI* curves seem to distribute very similarly.

Analogous to Figs. 6–8, we created curves $g_{m,O,S_S}(r)$ and $h_{m,O,S_A,S_S}(r)$ showing before and after linearization of metric values of samples from S_S based on linearization mapping estimated from S_A. Since the trends of these curves are very similar to those shown in Figs. 6–8, we have included only the curves for *DC* as an example in Fig. 9.

Our set S_S contains samples with *AS* of 1–5, except two cases – CtEs (a challenging object to segment) with $AS=1$ and CtSkn-h (an easy object to segment) with $AS=5$. S_S and its associated *AS* values demonstrate that large deviations seem to be more acceptable for segmentations of small sparse objects and even small deviations are less tolerated for large blob-like non-sparse objects. Another phenomenon to notice is the discrete steps in the CtSkn-h curves of Fig. 9, where the metric value range corresponding to each discrete *AS* value is more clear-cut than in other smaller objects and the ambiguity of samples with the same metric value assigned with different *AS* values is minimal. From Fig. 9 and similar curves for *JI* and HD_N, we may conclude that the simulation method is effective and needed for the linearization process.

Table 3
Mean (1st value) and sd (2nd value) of $\psi_g(.)$ over all samples of S_A for *DC* where the linearization mapping was estimated based on S_S.

	RPG	Mnd	CtEs	CtSkn-h
Hrt	0.440	0.285	0.522	0.478
	0.446	0.408	0.554	0.317
RPG		−0.114	−0.038	0.142
		0.198	0.138	0.306
Mnd			0.003	0.004
			0.233	0.327
CtEs				0.092
				0.339

Table 4
Mean (1st value) and sd (2nd value) of $\psi_g(.)$ over all samples of S_A for *JI* where the linearization mapping was estimated based on S_S.

	RPG	Mnd	CtEs	CtSkn-h
Hrt	0.606	0.270	0.571	0.203
	0.407	0.237	0.464	0.321
RPG		−0.077	−0.056	0.164
		0.230	0.152	0.209
Mnd			0.069	−0.092
			0.236	0.235
CtEs				0.241
				0.181

4.3. Quantitative evaluation

We list the mean and standard deviation of ψ_g for *DC*, *JI*, and HD_N in Tables 3–5, respectively, and of ρ_g, ρ_L, γ_g, and γ_L for all three metrics in Table 6. Recall from Eqs. (10)–(13) that, in these results, the linearization transformation was estimated based on S_S and applied to the samples in S_A. Since ψ_g and ρ_g express gain in similarity of acceptability, their range will be [−4, 4]. On the other hand, γ_g describes the similarity of metric values achieved for the same acceptability value, and so its range will be [−1, 1]. In both

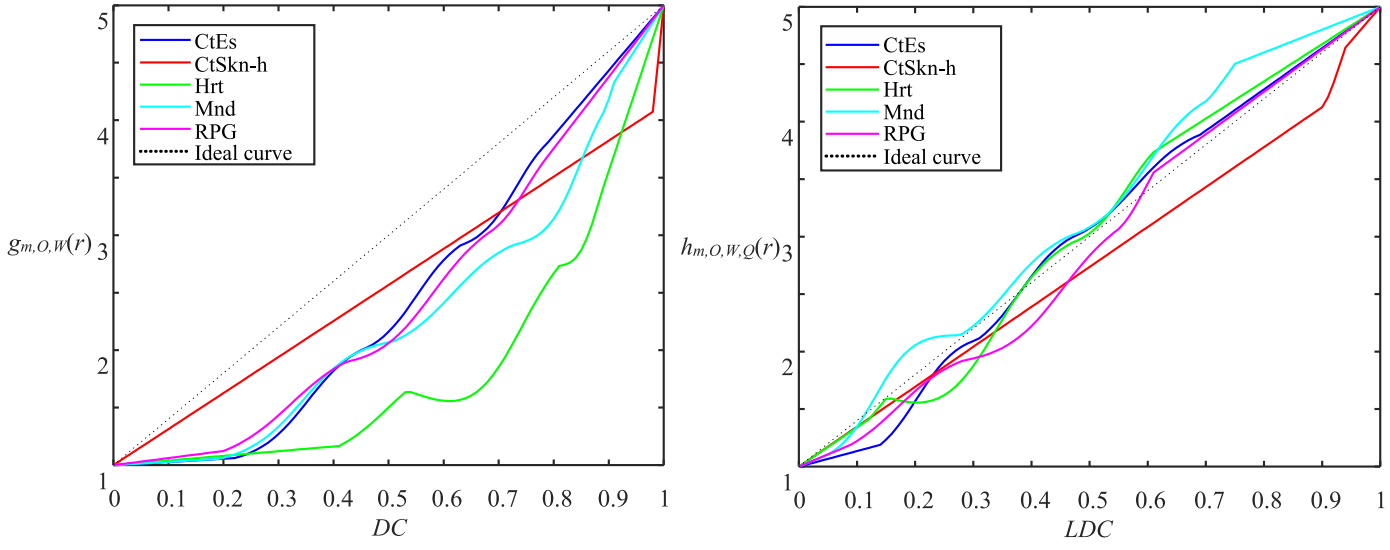

Fig. 6. Curves $g_{m,O,S_A}(r)$ (left) and $h_{m,O,S_S,S_A}(r)$ (right) for the five objects for set S_A for $m=DC$.

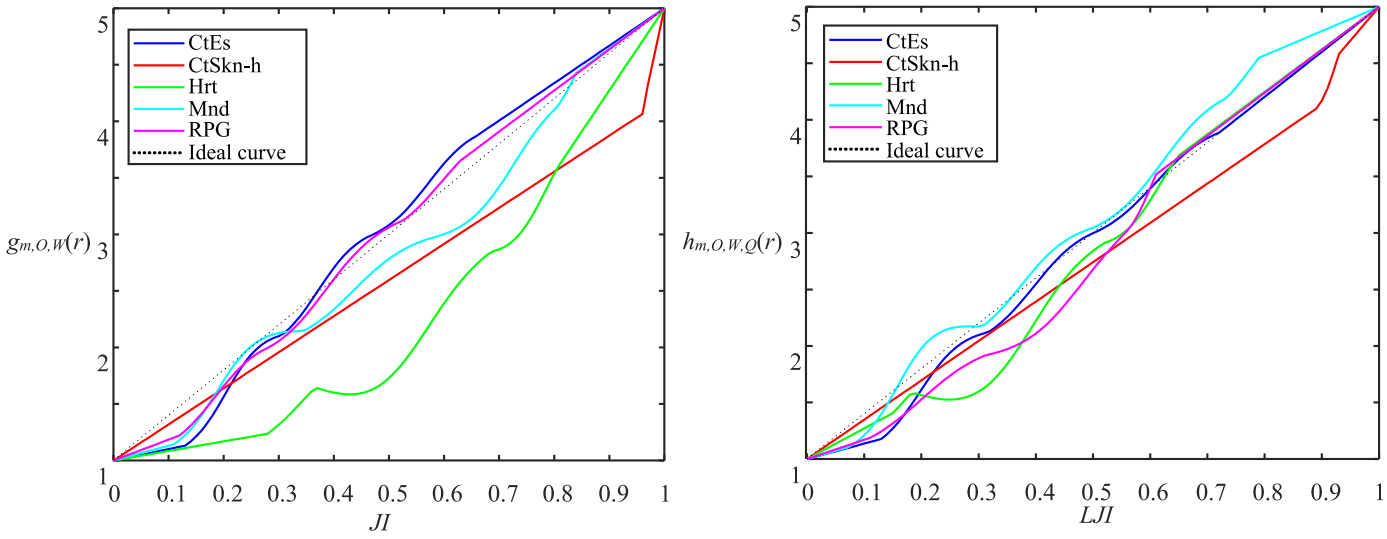

Fig. 7. Curves $g_{m,O,S_A}(r)$ (left) and $h_{m,O,S_S,S_A}(r)$ (right) for the five objects for set S_A for $m = JI$.

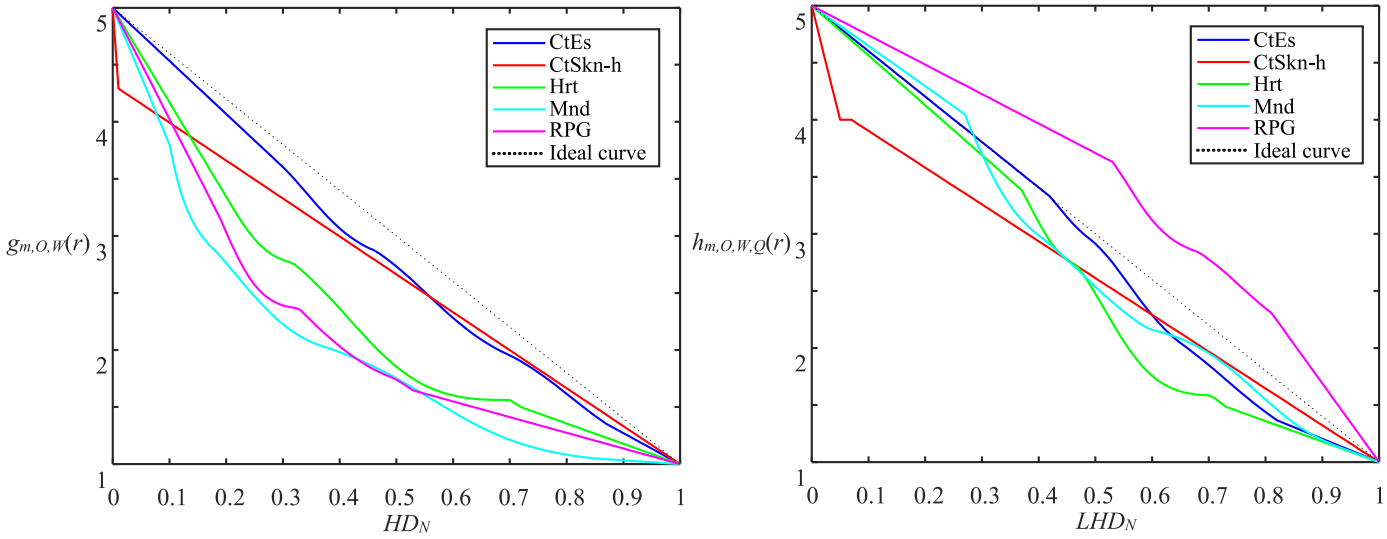

Fig. 8. Curves $g_{m,O,S_A}(r)$ (left) and $h_{m,O,S_S,S_A}(r)$ (right) for the five objects for set S_A for $m = HD_N$.

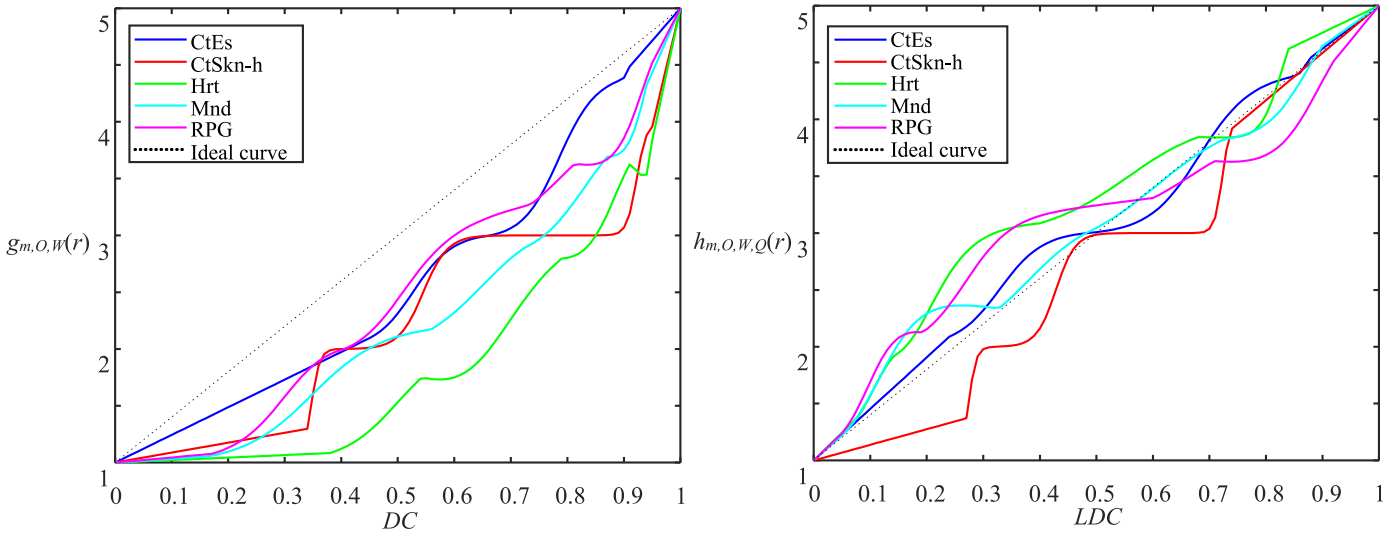

Fig. 9. Curves $g_{m,O,S_S}(r)$ (left) and $h_{m,O,S_A,S_S}(r)$ (right) for the five objects for set S_S for $m = DC$.

Table 5
Mean (1st value) and sd (2nd value) of $\psi_g(.)$ over all samples of S_A for HD_N where the linearization mapping was estimated based on S_S.

	RPG	Mnd	CtEs	CtSkn-h
Hrt	−0.562	0.141	0.306	0.083
	0.485	0.224	0.231	0.321
RPG		−0.457	0.118	−0.198
		0.444	0.455	0.209
Mnd			0.653	0.436
			0.363	0.440
CtEs				−0.146
				0.144

Table 7
Mean (1st value) and sd (2nd value) of $\psi_g(.)$ over all samples of S_S for *DC* where the linearization mapping was estimated based on S_A.

	RPG	Mnd	CtEs	CtSkn-h
Hrt	0.396	0.166	0.446	−0.121
	0.395	0.233	0.330	0.468
RPG		0.019	−0.080	−0.272
		0.265	0.190	0.440
Mnd			0.190	−0.150
			0.181	0.418
CtEs				0.038
				0.536

cases, a +ve value suggests improvement due to LinSEM and a -ve value implies deterioration.

We make the following observations from Tables 3–6 which are also borne out by the acceptability curves. (i) Of the 30 pairwise estimations of gain in similarity of semantic meaning ψ_g (among objects) over all metrics, 21 of them are positive (of which 19 are statistically significant, $P < 0.05$) and 9 of them are negative (of which all are statistically significant, $P < 0.05$). Since some pairs of objects may be similar even before linearization (such as RPG and CtEs for *DC* and *JI* as noted earlier), we do not expect for them to show significant $\psi_g > 0$ values post-linearization. In fact, they may show a small -ve value. (ii) More importantly, for such and most objects, we expect their curve to move closer to the ideal line after linearization (meaning $\rho_g > 0$) since this would guarantee that all objects would behave similarly in their metric meaning. From the 15 pairs of ρ_g mean values for the metric and its linearized version over all objects and metrics, we see that this is indeed the case except for two cases – RPG for *JI* and CtSkn-h for HD_N. The gain ρ_g (i.e., how much the curve is moved closer to the ideal curve after linearization) is large (~0.7–1 acceptability score units, or 18–25% improvement) for Hrt-*DC* and Mnd-HD_N, intermediate (~0.3–0.6, or 8–15%) for many cases (RPG-*DC*, Mnd-*DC*, CtEs-*DC*, Hrt-*JI*, Hrt-HD_N, and RPG-HD_N), and small (~1–5%) for the rest of the cases. (iii) Gain γ_g in metric similarity of meaning shows large positive values (0.1–0.26 on a [0, 1] range, or 10–26%) for Hrt-*DC*, RPG-*DC*, Mnd-*DC*, CtEs-*DC*, Hrt-*JI*, RPG-HD_N, and Mnd-HD_N, and <10% for the other cases with only two negative values (−1% to −2% for RPG-*JI* and CtSkn-h-HD_N). γ_L values show how close the curves of linearized metrics are to the ideal line, where the values are 0.01–0.06 in cases of *DC* and *JI* and 0.03–0.1 in cases of HD_N. Again, from the γ_L values and the curves, it is clear that in a majority of the cases, the actual metric values across objects are moved close to the ideal line. And only two cases have statistically significantly negative γ_g values.

Since the trend in the results of linearizing metrics of samples from S_S based on S_A were similar to those listed in Tables 3–5, we show results for S_S only for ψ_g in Table 7 for *DC*. Among 10 pairwise estimations of gain in similarity of semantic meaning ψ_g, 6 of them are positive (of which 4 are statistically significant, $P < 0.05$) and 4 of them are statistically significantly negative. Because of less variability in set S_A, especially of CtSkn-h, the metric-*AS* relationship is not completely fitted by the samples of S_A but partly by estimation due to linear connection, so semantic meaning for S_S have not improved as well as for S_A. (This is in the spirit of the justification provided earlier for estimating the linearization transformation based on set S_S and then applying it to set S_A.) However, comparing among curves in Fig. 9, we can also tell that curves of linearized *DC* distribute more closely along the ideal curve.

4.4. Gaps and challenges

There are several gaps in this investigation and further challenges to be addressed. First, limited by the cost of running the reader study, we decided to perform the LinSEM process on a 2D slice basis rather than in a true 3D fashion. Although there may be differences using the 2D versus 3D approach at the simulation stage and in the linearization process, we believe these differences are small and inconsequential. We admit however that this needs to be proven. The 3D approach has two serious drawbacks which hindered us in pursuing this approach – the reader-study cost due to a substantially increased number of "slices to read", and a disconnection in the reader's ability between reading 2D slices while having to score acceptability three-dimensionally. From our experience, we believe that this may result in less reliable acceptability scores than from 2D experiments.

Second, we indirectly assumed that the meaning of *AS* as determined by one expert is sufficient for the LinSEM process. Obviously, there may be differences in how experts score which may also vary for different applications. Considering multiple readers is directly feasible within our linearization method by generalizing Mahalanobis distance from a single variable to a multi-variate version or by pooling data from all readers. For different applications, application experts should perform the reader study to make sure that application-specific concerns are expressed in the scores.

Table 6
Mean (1st value) and sd (2nd value) of ρ_g, ρ_L, γ_g, and γ_L for all three metrics over all samples of S_A where the linearization mapping was estimated based on S_S.

	DC				JI				HD_N			
	ρ_g	ρ_L	γ_g	γ_L	ρ_g	ρ_L	γ_g	γ_L	ρ_g	ρ_L	γ_g	γ_L
Hrt	1.011	0.139	0.265	0.022	0.522	0.179	0.125	0.051	0.363	0.317	0.069	0.101
	0.554	0.109	0.138	0.030	0.338	0.180	0.059	0.053	0.352	0.273	0.059	0.067
RPG	0.422	0.139	0.106	0.034	−0.094	0.185	−0.020	0.043	0.436	0.405	0.135	0.075
	0.232	0.106	0.050	0.026	0.153	0.156	0.034	0.038	0.363	0.208	0.088	0.051
Mnd	0.515	0.188	0.141	0.035	0.042	0.141	0.011	0.035	0.750	0.231	0.191	0.054
	0.265	0.135	0.086	0.029	0.168	0.103	0.050	0.024	0.433	0.156	0.086	0.036
CtEs	0.421	0.117	0.107	0.028	0.072	0.074	0.020	0.016	0.054	0.133	0.009	0.038
	0.282	0.085	0.061	0.021	0.076	0.092	0.019	0.022	0.143	0.147	0.035	0.037
CtSkn-h	0.192	0.227	0.047	0.057	0.162	0.221	0.040	0.056	−0.036	0.366	−0.013	0.095
	0.166	0.142	0.028	0.034	0.139	0.138	0.023	0.034	0.077	0.218	0.010	0.056

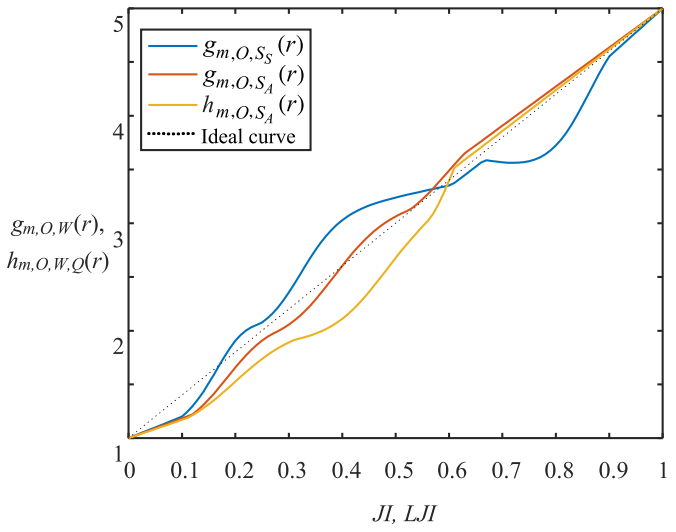

Fig. 10. An example for illustrating the fact that the difference between curves $g_{m,O,S_S}(r)$ and $g_{m,O,S_A}(r)$ can be larger than the difference between $g_{m,O,S_A}(r)$ and the ideal line.

Third, the most difficult challenge is how to handle cases of objects distorted by surgery or pathology. This takes us back to the issue of object definition. What is actually being segmented in these cases becomes crucial. If the goal is still the segmentation of the object outer boundary and if its shape is roughly the same as that of typical object samples, then it will not matter from LinSEM's perspective even if the object contains extensive pathology. However, if the object shape is severely distorted and the segmentation method cannot recover the original shape or if that is not the goal, then LinSEM's performance will be affected.

Finally, a question arises as to why not perform curve fitting in Section 2 using methods other than the proposed probabilistic approach based on Mahalanobis distance. Our early efforts, such as directly fitting from raw metric values and *AS* pairs of samples to polynomial curves, did not yield meaningful and similarly explainable results for different objects. This is the reason that we developed the proposed method. It is also conceivable that deep learning networks can be designed to perform this regression in more sophisticated ways, which we are currently examining.

We noticed that although the patterns of curves $g_{m,O,S_S}(r)$ and $g_{m,O,S_A}(r)$ obtained from S_S and S_A were similar, there were differences in distributions p_i (see Section 2.1) between the two cases. The main culprit is the lack of full coverage of segmentation quality in the case of set S_A as we already mentioned. For example, since object CtSkn-h is usually easy to segment, its samples in S_A will have $AS=4$ or 5 and will not include cases with $AS=1$ or 2. Conversely, sparse objects such as CtEs rarely cover cases with $AS=5$. This causes the distributions pertaining to S_S and S_A to differ and, we believe, the deterioration of linearity we encountered in our experiments. We expect the difference among curves due to this difference in distribution between S_S and S_A to be smaller than the actual difference in curves among objects. We observed that violation of this expected behavior can lead to deterioration of linearity. An example is shown in Fig. 10 for RPG for the case of linearizing *JI*. Notice that, for $m=JI$, the difference between curves $g_{m,O,S_S}(r)$ and $g_{m,O,S_A}(r)$can be larger than the difference between $g_{m,O,S_A}(r)$ and the ideal line for some metric and acceptability values.

Computational considerations: LinSEM was implemented on a computer with the following specifications: 6-core Intel i7-7800X CPU 3.5 GHz with 64 GB RAM and running the Linux operating system. Computational time for curve fitting for each object based on each dataset (S_A or S_S) is less than 0.2 s in MATLAB R2018b. Subsequent linearization is instantaneous.

5. Concluding remarks

In this paper, we introduced a new concept (LinSEM) of linearizing segmentation evaluation metrics for achieving uniformity of meaning across different anatomic objects based on corresponding degrees of expert-scored acceptability. We designed a set of sequences of basic image operations to be applied to true segmentations to mimic the full spectrum of deviations potentially observable in actual segmentations by varied algorithms. We performed a reader study on simulated segmentations (S_S) wherein an expert determines an acceptability score *AS* for each study. The rationale for and advantage of employing simulations are that they can cover the full spectrum of overall quality distribution much better and by design within a smaller population of samples than actual segmentations which typically cover a partial range of acceptability and may also require a larger sample size to have a proper coverage within the restricted range. Also, for some large, well-defined objects, even very large sample sets of actual segmentations may not capture the needed full range of variations. Thus, the cost associated with reader studies can be considerably reduced via simulated segmentations. Based on *AS*, we estimate object- and metric-dependent metric-*AS* relationships via the concept of probabilistic acceptability scores by employing the Mahalanobis distance over a discretized set of metric values covering the full domain of the metric. The relationships determined by using S_S are taken as calibration reference to linearize the metric for each object on actual segmentations (S_A). We conducted experiments on five anatomic objects (cervical esophagus (CtEs), cervical skin outer boundary (CtSkn-h), heart (Hrt), mandible (Mnd), and right parotid gland (RPG)) utilizing three most commonly-used metrics (*DC, JI*, and *HD*) to assess the improvement brought about by LinSEM in the uniformity of metric meaning across objects.

We summarize our conclusions as follows. (i) Generally, *JI* seems to have a more linear relationship with acceptability before actual linearization than other metrics. (ii) LinSEM achieves significantly improved uniformity of meaning post-linearization across all tested objects and metrics, except in a few cases where the departure from linearity was insignificant before linearization. This improvement, expressing how close metric-to-acceptability relationship has been brought to the ideal curve, is generally the largest for *DC* and *HD* reaching 8–25% for many tested cases. (iii) Although some objects (such as RPG and CtEs for *DC* and *JI*) are close in their meaning between themselves before linearization, they are distant in this meaning from other objects. This emphasizes the importance of bringing all objects individually close to the ideal curve to realize uniformity of meaning across all objects. This in turn suggests that, eventually, linearization must be performed considering all objects in a body region, and preferably, all objects body-wide. (iv) Our results suggest that the proposed method of simulating segmentations may be a practical way of addressing the dual challenges of keeping the set of segmentations to be dealt with manageable and minimizing the cost of conducting reader studies. (v) Although we used image data sets from CT from H&N and thorax body regions, the LinSEM process is applicable as is to other image modalities and body regions as long as sets S_A and S_T are available for a set of objects for the body region of interest.

Medical practice relies heavily on graded or categorical scoring systems for assessing various phenomena such as health status and disease stage (for example, BI-RADS D'Orsi et al., 2013, PI-RADS Turkbey et al., 2019, etc.). These systems are body-region-, object-, disease-, and application-specific, and have been arrived at through standardized guidelines for scoring. For wide-spread use of any

method such as LinSEM, standardized guidelines will become necessary for acceptability scoring in order to reduce intra- and inter-reader variability. We are in the process of conducting a multi-center study for acceptability scoring in the two body regions considered in this paper for the application of auto-contouring organs at risk for radiation therapy planning (Wu et al., 2019). Currently this application is perhaps the largest consumer of segmentation tools and tools for clinically meaningful evaluation.

In this paper, we focused on anatomical objects which have known prior shape. To apply LinSEM to objects of irregular shape such as tumors and pathological regions, they need to be first categorized into groups (Cao et al., 2016) based on their geometric attributes (such as spherical, ovoid, polygonal, smooth, lobulated, spiculated) and morphological attributes (such as extensive, small). Then the linearization process can be studied by group. This clearly requires much further work.

Declaration of Competing Interest

There is no any conflict of interest and this is the solo submission to Medical Image Analysis.

Acknowledgements

The research reported here is supported by a DHHS grant R42CA199735. Jieyu Li's training at the Medical Image Processing Group was supported partly by China Scholarship Council.

References

Ashburner, J., Friston, K.J., 2009. Computing average shaped tissue probability templates. Neuroimage 45 (2), 333–341. doi:10.1016/j.neuroimage.2008.12.008.

Baiker, M., Milles, J., Dijkstra, J., Henning, T.D., Weber, A.W., Que, I., Kaijzel, E.L., Löwik, C.W., Reiber, J.H., Lelieveldt, B.P., 2010. Atlas-based whole-body segmentation of mice from low-contrast micro-CT data. Med. Image Anal. 14 (6), 723–737. doi:10.1016/j.media.2010.04.008.

Baxter, J.S.H., Gibson, E., Eagleson, R., Peters, T.M., 2017. The semiotics of medical image segmentation. Med. Image Anal. 44, 54–71. doi:10.1016/j.media.2017.11.007.

Beucher, S., 1992. The watershed transformation applied to image segmentation. In: Proceedings of the 10th Pfefferkorn Conference on Signal and Image Processing in Microscopy and Microanalysis, pp. 299–314.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23 (11), 1222–1239. doi:10.1109/34.969114.

Cao, L., Udupa, J.K., Odhner, D., Huang, L., Tong, Y., Torigian, D.A., 2016. A general approach liver lesion segmentation in CT images. In: Proceedings of SPIE doi:10.1117/12.2217778, 9786: 978623-1 – 978623-7.

Cappabianco, F.A.M., de Miranda, P.A.V., Udupa, J.K., 2017. A critical analysis of the methods of evaluating MRI brain segmentation algorithms. In: Proceedings of the IEEE International Conference in Image Processing, pp. 3894–3898. doi:10.1109/ICIP.2017.8297012.

Chen, X., Udupa, J.K., Bagci, U., Zhuge, Y., Yao, J., 2012. Medical image segmentation by combining graph cuts and oriented active appearance models. IEEE Trans. Image Process. 21 (4), 2035–2046. doi:10.1109/TIP.2012.2186306.

Christensen, G.E., Rabbitt, R.D., Miller, M.I., 1994. 3D brain mapping using a deformable neuroanatomy. Phys. Med. Biol. 39, 609–618. doi:10.1088/0031-9155/39/3/022.

Chu, C., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Hayashi, Y., Wolz, R., Rueckert, D., Mori, K., 2013. Multi-organ Segmentation from 3D Abdominal CT Images Using Patient-Specific Weighted-Probabilistic Atlas. SPIE Medical Imaging, SPIE 86693Y-86691-86693Y-86697 doi:10.1117/12.2007601.

Cootes, T.F., Taylor, C.J., Cooper, D.H., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61 (1), 38–59. doi:10.1006/cviu.1995.1004.

Detmer, P.R., Bashein, G., Martin, R.W., 1990. Matched filter identification of left-ventricular endocardial borders in transesophageal echocardiograms. IEEE Trans. Med. Imaging 9 (4), 396–404. doi:10.1109/42.61755.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302. doi:10.2307/1932409.

D'Orsi, C.J., Sickles, E.A., Mendelson, E.B., Morris, E.A., et al., 2013. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. American College of Radiology, Reston, VA.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.A., 2017. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54. doi:10.1016/j.media.2017.05.001.

Doyle, W., 1962. Operations useful for similarity-invariant pattern recognition. J. ACM 9, 259–267. doi:10.1145/321119.321123.

Drozdzal, M., Chartrand, G., Vorontsov, E., Shakeri, M., Jorio, L.D., Tang, A., Romero, A., Bengio, Y., Pal, C., Kadoury, S., 2018. Learning normalized inputs for iterative estimation in medical image segmentation. Med. Image Anal. 44, 1–13. doi:10.1016/j.media.2017.11.005.

Falcao, A.X., Udupa, J.K., Samarasekera, S., Sharma, S., Hirsch, B.E., Lotufo, R.D.A., 1998. User-steered image segmentation paradigms: live wire and live lane. Graph. Models Image Process 60 (4), 233–260. doi:10.1006/gmip.1998.0475.

Gee, J.C., Reivich, M., Bajcsy, R., 1993. Elastically deforming 3D atlas to match anatomical brain images. J. Comput. Assist. Tomogr. 17, 225–236. doi:10.1097/00004728-199303000-00011.

Heimann, T., Van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans. Med. Imaging 28 (8), 1251–1265. doi:10.1109/TMI.2009.2013851.

Herman, G.T., Srihari, S., Udupa, J.K., 1979. Detection of changing boundaries in two- and three-dimensions. In: Badler, N.I., Aggarwal, J.K. (Eds.), Proceedings of the Workshop on Time Varying Imagery. University of Pennsylvania, Philadelphia, Pennsylvania, pp. 14–16.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15 (9), 850–863. doi:10.1109/34.232073.

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des alpes et des Jura. Bull Soc. Vaudoise Sci. Nat. 37, 547–579.

Kass, M., Witkin, A., Terzopoulos, D., 1987. Snakes: active contour models. Int. J. Comput. Vis. 1 (4), 321–331. doi:10.1007/BF00133570.

Kim, H., Monroe, J.I., Lo, S., Yao, M., Harari, P.M., Machtay, M., Sohn, J.W., 2015. Quantitative evaluation of image segmentation incorporating medical consideration functions. Med. Phys. 42 (6), 3013–3023. doi:10.1118/1.4921067.

Kim, H.S., Park, S.B., Lo, S.S., Monroe, J.I., Sohn, J.W., 2012. Bidirectional local distance measure for comparing segmentations. Med. Phys. 39 (11), 6779–6790. doi:10.1118/1.4754802.

Lamecker, H., Lange, T., Seebass, M., 2004. Segmentation of the Liver Using a 3D Statistical Shape Model. Technical Report, Zuse Institute, Berlin.

Li, R., Zhang, W., Suk, H.I., Wang, L., Li, J., Shen, D., Ji., S., 2014. Deep Learning Based Imaging Data Completion For Improved Brain Disease Diagnosis. MICCAI, Springer, Cham, pp. 305–312. doi:10.1007/978-3-319-10443-0_39.

Linguraru, M.G., Pura, J.A., Pamulapati, V., Summers, R.M., 2012. Statistical 4D graphs for multi-organ abdominal segmentation from multiphase ct. Med. Image Anal. 16 (4), 904–914. doi:10.1016/j.media.2012.02.001.

Liu, H.K., 1977. Two- and three-dimensional boundary detection. Comput. Graph. Image Process 6, 123–134. doi:10.1016/S0146-664X(77)80008-7.

Lopez-Molina, C., De Baets, B., Bustince, H., 2013. Quantitative error measures for edge detection. Pattern Recognit. 46 (4), 1125–1139. doi:10.1016/j.patcog.2012.10.027.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. In: Proceedings of the National Institute of Sciences of India, 2, pp. 49–55.

Malladi, R., Sethian, J.A., Vemuri, B.C., 1995. Shape modeling with front propagation: a level set approach. IEEE Trans. Pattern Anal. Mach. Intell. 17, 158–175. doi:10.1109/34.368173.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J.N.L., Išgum, I., 2016. Automatic segmentation of brain Mr images with a convolutional neural network. IEEE Trans. Med. Imaging 35 (5), 1252–1262. doi:10.1109/TMI.2016.2548501.

Mumford, D., Shah, J., 1989. Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. 42 (5), 577–685. doi:10.1002/cpa.3160420503.

Narasimhan, R., Fornango, J.P., 1963. Some further experiments in the parallel processing of pictures. IEEE Trans. Electronic Comput. 6, 748–750. doi:10.1109/PGEC.1964.263936.

Nyul, L.G., Udupa, J.K., 1999. On standardizing the Mr image intensity scale. Magn. Reson. Med. 42 (6), 1072–1081 https://doi.org/10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

Oda, H., Roth, H.R., Bhatia, K.K., Oda, M., Kitasaka, T., Iwano, S., Homma, H., Takabatake, H., Mori, M., Natori, H., Schnabel, J.A., Mori, K., 2018. Dense volumetric detection and segmentation of mediastinal lymph nodes in chest CT images. In: Proceedings of SPIE Medical Imaging Conference 10575 doi:10.1117/12.2287066, 1057502-1 – 1057502-6.

Pizer, S.M., Fletcher, P.T., Joshi, S., Thall, A., Chen, J.Z., Fridman, Y., Fritsch, D.S., Gash, A.G., Glotzer, J.M., Jiroutek, M.R., Lu, C.L., Muller, K.E., Tracton, G., Yushkevich, P., Chaney, E.L., 2003. Deformable m-reps for 3D medical image segmentation. Int. J. Comput. Vis. 55 (2–3), 85–106. doi:10.1023/A:1026313132218.

Pope, D.L., Parker, D.L., Gustafson, D.E., Clayton, P.D., 1984. Dynamic search algorithm in left ventricular border recognition and analysis of coronary arteries. IEEE Proc. Comput. Cardiology 9, 71–75.

Raya, S.P., Udupa, J.K., 1990. Shape-based interpolation of multidimensional objects. IEEE Trans. Med. Imaging 9 (1), 32–42. doi:10.1109/42.52980.

Ruskó, L., Bekes, G., Fidrich, M., 2009. Automatic segmentation of the liver from multi-and single-phase contrast-enhanced ct images. Med. Image Anal. 13 (6), 871–882. doi:10.1016/j.media.2009.07.009.

Schmid, J., Kim, J., Magnenat-Thalmann, N., 2011. Robust statistical shape models for mri bone segmentation in presence of small field of view. Med. Image Anal. 15 (1), 155–168. doi:10.1016/j.media.2010.09.001.

Shen, T., Li, H., Huang, X., 2011. Active volume models for medical image segmentation. IEEE Trans. Med. Imaging 30 (3), 774–791. doi:10.1109/TMI.2010.2094623.

Shi, C., Cheng, Y., Wang, J., Wang, Y., Mori, K., Tamura, S., 2017. Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic patho-

logical organ segmentation. Med. Image Anal. 38, 30–49. doi:10.1016/j.media.2017.02.008.

Staib, L.H., Duncan, J.S., 1992. Boundary finding with parametrically deformable models. IEEE Trans. Pattern Anal. Mach. Intell. 14, 1061–1075. doi:10.1109/34.166621.

Tomoshige, S., Oost, E., Shimizu, A., Watanabe, H., Nawano, S., 2014. A conditional statistical shape model with integrated error estimation of the conditions; application to liver segmentation in non-contrast CT images. Med. Image Anal. 18 (1), 130–143. doi:10.1016/j.media.2013.10.003.

Turkbey, B., Rosenkrantz, A.B., Haider, M.A., Padhani, A.R., Villeirs, G., Macura, K.J., Tempany, C.M., Choyke, P.L., Cornud, F., Margolis, D.J., Thoeny, H.C., Verma, S., Barentsz, J., Weinreb, J.C., 2019. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. Eur. Urol. doi:10.1016/j.eururo.2019.02.033.

Udupa, J.K., Leblanc, V.R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L.M., Hirsch, B.E., Woodburn, J., 2006. A framework for evaluating image segmentation algorithms. Comput. Med. Imaging Graph. 30 (2), 75–87. doi:10.1016/j.compmedimag.2005.12.001.

Udupa, J.K., Odhner, D., Zhao, L., Tong, Y., Matsumoto, M.M., Ciesielski, K.C., Falcao, A.X., Vaideeswaran, P., Ciesielski, V., Saboury, B., Mohammadianrasanani, S., 2014. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. Med. Image Anal. 18 (5), 752–771. doi:10.1016/j.media.2014.04.003.

Udupa, J.K., Samarasekera, S., 1996. Fuzzy connectedness and object definition: theory, algorithms, and applications in image segmentation. Graph. Models Image Process 58 (3), 246–261. doi:10.1006/gmip.1996.0021.

Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2013. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Trans. Med. Imaging 32 (9), 1723–1730. doi:10.1109/TMI.2013.2265805.

Wu, X., Udupa, J.K., Tong, Y., Odhner, D., Pednekar, G.V., Simone II, C.B., McLaughlin, D., Apinorasethkul, C., Lukens, J., Mihailidis, C., Shammo, G., James, P., Tiwari, A., Wojtowicz, L., Camaratta, J., Torigian, D.A., 2019. AAR-RT – A system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. Med. Image Anal. 54, 45–62. doi:10.1016/j.media.2019.01.008.

Yeghiazaryan, V., Voiculescu, I., 2018. Family of boundary overlap metrics for the evaluation of medical image segmentation. J. Med. Imaging 5 (1). doi:10.1117/1.JMI.5.1.015006, 015006.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D., 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. Neuroimage 108, 214–224. doi:10.1016/j.neuroimage.2014.12.061.

Zhang, Y.J., 1996. A survey on evaluation methods for image segmentation. Pattern Recognit 29 (8), 1335–1346. doi:10.1016/0031-3203(95)00169-7.

Zhang, Y.J., 2001. A review of recent evaluation methods for image segmentation. In: Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat. No. 01EX467), 1. IEEE, pp. 148–151. doi:10.1109/ISSPA.2001.949797.