Orientation-sensitive overlap measures for the validation of medical image segmentations

Tasos Papastylianou[1], Erica Dall' Armellina[2], and Vicente Grau[1]

[1] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2] Acute Vascular Imaging Centre, Radcliffe Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford, UK
tasos.papastylianou@kellogg.ox.ac.uk,
WWW home page: http://tpapastylianou.com

Abstract. Validation is a key concept in the development and assessment of medical image segmentation algorithms. However, the proliferation of modern, non-deterministic segmentation algorithms has not been met by an equivalent improvement in validation strategies. In this paper, we briefly examine the state of the art in validation, and propose an improved validation method for non-deterministic segmentations, showing that it improves validation precision and accuracy on both synthetic and clinical sets, compared to more traditional (but still widely used) methods and state of the art.

Keywords: validation, segmentation, fuzzy, probabilistic, t-norms

1 Introduction

It is often noted in the segmentation literature, that while research on newer segmentation methods abounds, corresponding research on appropriate evaluation methods tends to lag behind by comparison [1, 2]. This is particularly the case with medical images, which suffer inherent difficulties in terms of validation, such as: limited datasets; clinical ambiguity over the ground truth itself; difficulty and relative unreliability of clinical gold standards (which usually defaults to expert-led manual contour delineation); and variability in agreed segmentation protocols and clinical evaluation measures [1]. Furthermore, many of the latest approaches in segmentation have been increasingly non-deterministic, or "fuzzy"[3] in nature [3]; this is of particular importance in medical image segmentation, due to the presence of the Partial Volume Effect (PVE)[4]. However, appropriate validation methods that take fuzziness specifically into account are rarely considered, despite the fact that gold standards are also becoming increasingly fuzzy (e.g. expert delineations at higher resolutions; consensus voting [3]). On the contrary, most segmentation papers approach fuzziness as a validation nuisance instead, as they tend to rely on more conventional binary validation methods, established from early segmentation literature, and work around the 'problem' of fuzziness by thresholding pixels at a (sometimes arbitrary) threshold, so as to produce the binary sets required for traditional validation.

[3] We use the term "fuzzy" here in a broad sense, i.e. all methods assigning non-discrete labels, of which modern probabilistic segmentation methods are a strict subset.

State of the art in the validation of fuzzy / probabilistic segmentations: There is a multitude of validation approaches and distance / similarity metrics; Deza and Deza's The Encyclopedia of Distances [5] alone, spans over 300 pages of such metrics from various fields. In the medical image segmentation literature, traditional binary methods like the Dice and Tanimoto coefficients seem to be by far the most popular overlap-based metrics [8, 7], even when explicitly validating inherently non-binary segmentation / gold-standard pairs.

While there are a few cases where thresholding *could* be deemed appropriate (e.g. when fuzziness does not have a straightforward interpretation, therefore a simplifying assumption is made that all output pixels are *pure*) we would argue that in most cases, the thresholding approach is still used mostly by convention, or at most out of a need for consistency and comparison with older literature, rather than because it is an appropriate method for fuzzy sets. This occurs at the cost of discarding valuable information, particularly in the case where fuzziness essentially denotes a PVE.

Yi *et al.* [6] addressed this issue by treating PVE pixels as a separate binary class denoting 'boundary' pixels. While they demonstrated that this approach led to higher scores on validation compared to thresholding, there was no discussion as to whether this genuinely produces a more accurate, precise, and reliable result; furthermore, it is still wasteful of information contained in pixel fuzziness, since all degrees of fuzziness at the boundary would be treated as a single label.

Chang *et al.* [7] proposed a framework for extending traditional validation coefficients for fuzzy inputs, by taking the intersection of two non-zero pixels to be equal to the complement of their discrepancy. In other words, two pixels of equal fuzzy value are given an intersection value of 1. However, this is a rather limited interpretation of fuzziness, which is not consistent with PVE or geometric interpretations of fuzziness, as we will show later. Furthermore, the authors did not formally assess validation performance against their binary counterparts.

Crum *et al.* [8] proposed a fuzzy extension of the Tanimoto coefficient, demonstrating increased validation accuracy. They used a specific pair of Intersection and Union operators derived from classical fuzzy set theory, and compared against the traditional thresholding approach. They assessed their operator using fuzzy segmentations and gold standards derived from a synthetic 'petal' set, whose ground-truth validation could be reproduced analytically.

Main contributions: Our work expands on the theoretical framework put in place by Crum *et al.* by examining the geometric significance of the particular fuzzy intersection and union operators used. Armed with this insight, we proceed to: **1.** establish absolute validation operator bounds, outside of which, pixel and geometry semantics do not apply. **2.** show that thresholding tends to violate these bounds, leading to reduced validation precision and accuracy, and rendering it unreliable and unsuitable for the assessment and comparison of non-deterministic segmentations. **3.** propose a novel fuzzy intersection and union operator defined within these bounds, which takes into account the orientation of fuzzy pixels at object boundaries, and show that this improves validation precision and accuracy on both synthetic and real clinical sets.

2 Background theory and motivation

2.1 Fuzziness and probability in medical image segmentation

What does it mean for a pixel to be fuzzy? In classic segmentation literature, a segmentation (SG) mask — and similarly, a gold-standard (GS) mask — is a binary image (i.e. pixels take values in the set {0,1}) of the same resolution as the input image, where the values denote the abscence or presence in the pixel, of the tissue of interest. In a fuzzy SG mask, pixels can instead take any value in the interval [0,1]. The underlying semantics of such a value are open to interpretation; however, perhaps the most intuitive interpretation is that of a mapping from the fuzzy value, to the extent to which a pixel is occupied by the tissue in question. For example, in the simplest case of a linear mapping, a pixel with a fuzzy value of 0.56, could be interpreted as consisting of the tissue in question by 56%, and 44% background. Figure 1 demonstrates this graphically.

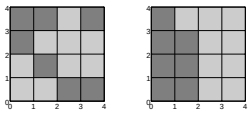

Fig. 1: Two fuzzy pixels with the same fuzzy value of 0.5625, but different underlying configurations (i.e. tissue distribution inside the pixel). The pixel's fuzzy value is the average of its constituent subpixels. Note the pixel on the right is more homogeneous.

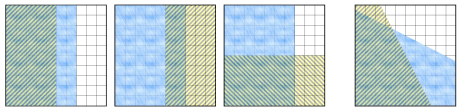

Fig. 2: Edge Pixels: In all cases, GS (thin blue lines) covers 70% of the pixel, and SG (coarse yellow lines) covers 50%. From Left to right: GS and SG have the same orientation; GS and SG have opposite orientations; GS and SG have perpendicular directionality; GS and SG exhibit arbitrary orientations. Their intersections have pixel coverages of 50%, 20%, 35%, and 46.2% respectively.

T-Norms and T-Conorms: Intersection and Union operations on fuzzy inputs: Triangular Norms and Triangular Co-norms, or T-Norms ($\mathrm{T_N}$) and T-Conorms ($\mathrm{T_C}$), are generalisations of the Intersection and Union operators (denoted $\bigcap$ and $\bigcup$ respectively) in Fuzzy Set Theory [9]. A $\mathrm{T_N}$ takes two fuzzy inputs and returns a fuzzy output, is commutative ($A \cap B \equiv B \cap A$), associative ($A \cap (B \cap C) \equiv (A \cap B) \cap C$), monotonically nondecreasing with respect to increasing inputs, and treats 0 and 1 as null and unit elements respectively ($A \cap 0 \equiv 0,\ A \cap 1 \equiv A$). A $\mathrm{T_C}$ is similar to a $\mathrm{T_N}$, except the null and unit elements are reversed (i.e. $A \cup 0 = A,\ A \cup 1 = 1$). $\mathrm{T_N}$ and $\mathrm{T_C}$ are dual operations, connected (like in the binary case) via De Morgan's Law: $A \cup B \equiv \neg(\neg A \cap \neg B)$, where $\neg$ represents complementation, i.e. $\neg(A) = 1 - A$ in this context.

2.2 Motivation: Defining theoretical limits for valid intersections

We shall now make special mention of two specific $\mathrm{T_N}$ / $\mathrm{T_C}$ pairs:

	T − Norm	**T − Conorm**
Gödel :	$A \cap_G B = min(A, B)$	$A \cup_G B = max(A, B)$
Łukasiewicz :	$A \cap_L B = max(0, A + B - 1)$	$A \cup_L B = min(1, A + B)$

In this section, we show that these T_Ns are of special relevance in the context of pixels: when the fuzzy inputs are SG and GS masks denoting underlying tissue composition, these T_Ns have the following properties:

Theorem 1: The Gödel T_N ($\cap_G$) represents the most *optimal* $\bigcap$ possible for that pixel. *Proof:* We remind ourselves that fuzziness here is defined in the frequentist sense, i.e. as the number of subpixels labelled true, over the total number of subpixels N comprising the pixel. Optimality occurs when the set of all true subpixels in one mask is a strict subset of the set of all true subpixels in the other mask; therefore the $\bigcap$ of the two sets is equivalent to the set with the least elements. ■

Theorem 2: The Łukasiewicz T_N ($\cap_L$) represents the most *pessimal* $\bigcap$ possible. *Proof:* For any two sets of true subpixels A and B such that $|A| + |B| = N$, the most pessimistic scenario occurs when A and B are mutually exclusive (i.e. $\bigcap = 0$). Decreasing the number of true subpixels in either A or B still results in $\bigcap = 0$. Increasing either input (to, say, $\widetilde{A}$ and $\widetilde{B}$) will result in a necessary overlap equal to the number of extra true subpixels introduced from either set, i.e. $(\widetilde{A}-A)+(\widetilde{B}-B) = \widetilde{A}+\widetilde{B}-A-B = \widetilde{A}+\widetilde{B}-1$. ■

Therefore, these T_Ns represent theoretical $\bigcap$ bounds for fuzzy pixels. Any validation, which was obtained via $\bigcap$ outputs outside these theoretical bounds, should be considered theoretically infeasible, and therefore unreliable.

We immediately note that the traditional thresholding approach and the Chang approach *can lead to unreliable validations* (e.g. for SG and GS fuzzy inputs of 0.6, both methods lead to an out-of-bounds $\bigcap$ output of 1).

2.3 Boundary pixel validation – The case for a Directed T_N ($\cap_D$)

Pixels at object boundaries commonly exhibit PVE, which manifests in their corresponding SG masks as fuzziness. Such mask pixels can be thought of as homogeneous fuzzy pixels (see fig 1) with a particular orientation. At its simplest, we can portray such a pixel as divided by a straight line parallel to the object boundary at that point, splitting it into foreground and background regions. We define as the fuzzy orientation for that pixel the outward direction perpendicular to this straight line (in other words, the negative ideal local image-gradient).

It is easy to see that optimal overlap between an SG and GS mask pixel occurs when their corresponding orientations are congruent; similarly, pessimal overlap occurs when they are completely incongruent. In other words, $\cap_G$ and $\cap_L$ in the context of boundary pixels, correspond to absolute angle differences in orientation of 0° and 180° respectively. Fig 2 demonstrates this visually for the particular case of a 2D square pixel. It should be clear that for any absolute orientation angle difference between 0° and 180°, there exists a suitable $\bigcap$ operator which returns a value between the most optimal (i.e. $\cap_G$) and most pessimal (i.e. $\cap_L$) value, and which decreases monotonically between these two limits as this absolute angle difference increases within that range. Furthermore, as fig 2 suggests, for a particular pixel of known shape and dimensionality, this can be calculated exactly in an analytical fashion. We define such an operator as a *Directed* T_N ($\cap_D$), and its dual as a *Directed* T_C ($\cup_D$), and distinguish between generalised and exact versions as above.

We can now define suitable fuzzy validation operators, in a similar fashion to Crum *et al.*[8], by substituting binary $\bigcap$ and $\bigcup$ operations with fuzzy ones in the definitions of the validation operators used; here, by way of example, we will focus on the Tanimoto Coefficient, defined as $Tanimoto(SG, GS) = \frac{\sum[SG \bigcap GS]}{\sum[SG \bigcup GS]}$; however any validation operator which can be defined in terms of set operations can be extended to fuzzy set theory this way.

3 Methods and Results

3.1 Exact and Generalised Directed T_Ns

An Exact Directed T-Norm : As mentioned above, a $\cap_D$ can be calculated *exactly*, if we assume the exact shape / dimensionality of a particular pixel, to be known and relevant to the problem. In other words, given a fuzzy value and orientation, the shape of the pixel dictates the exact manner in which tissue is distributed within it. An $\bigcap$ operation between a SG and a GS pixel, whose fuzzy values and orientations are both known, can therefore be calculated exactly in an analytical manner, but the result is specific to that particular pixel shape. Fig 3 shows the profile of such an Exact $\cap_D$ for the specific case of a 2D pixel and a range of fuzzy inputs and angle differences (the algorithmic steps for this particular Exact $\cap_D$ calculation are beyond the scope of this paper, but an Octave / Matlab implementation is available on request).

A Generalised Directed T-Norm: Clearly, while an exact $\cap_D$ should be more accurate, this dependence on the exact pixel shape adds an extra layer of complexity, which may be undesirable. The biggest benefit of the $\cap_D$ is its property of outputting a suitable value between the theoretical limits imposed by $\cap_L$ and $\cap_G$; therefore, any function that adheres to this description, should be able to provide most of the benefits afforded by an Exact $\cap_D$, regardless of pixel shape / dimensions, and potentially with very little extra computational overhead compared to standard T_Ns.

For the purposes of this paper, we chose a sinusoidal function: $A \cap_D B = \left(\frac{1+cos\ \theta}{2}\right) G + \left(\frac{1-cos\ \theta}{2}\right) L$, where $G = A \cap_G B$, $L = A \cap_L B$, and θ signifies the discrepancy angle between SG and GS front orientations. Our particular choice of function here aimed to provide a good fit to the Exact version specified above (see fig 3), while also being generalisable to pixels of higher dimensions. However, we reiterate that this is simply one of many valid formulations, chosen as proof of concept; in theory, any valid formulation (i.e. monotonically decreasing for increasing discrepancy in angle) should prove more accurate than $\cap_L$ or $\cap_G$ alone which is already considered state of the art.

3.2 Demonstration on Synthetic and Clinical Sets

Synthetic example: The synthetic 'Petal' set introduced in Crum *et al.*[8] was replicated, to obtain a 'high resolution' binary image (100×100), like the one shown in fig 4. The *Ground Truth validation* (GTv), i.e. the latent truth, was obtained by rotating one copy of the petal image (acting as the SG mask), onto another, stationary petal image acting as the GS mask, and calculating a normal Tanimoto coefficient via the $\bigcap$ and $\bigcup$ of the two masks, at various angles of rotation. At each rotation angle, fuzzy SG and GS masks (25×25

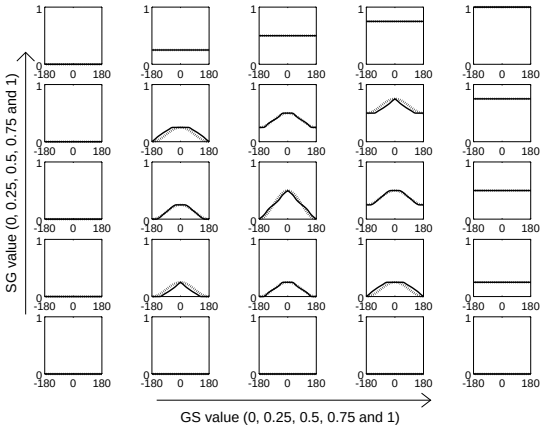

Fig. 3: The Exact (solid line) and Generalised (dashed line) $\cap_D$ described in section 3.1, for a range of fuzzy inputs. The x-axis of each subplot corresponds to the angle difference between the SG and GS pixels, and the y-axis to the corresponding fuzzy intersection output.

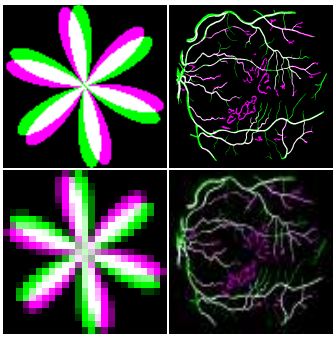

Fig. 4: Fused masks from the petal and clinical sets used. Top row: GTv (high-resolution) sets, bottom row: fuzzy (low-resolution) sets. SG shown as violet, GS as green throughout; colour-fusion (i.e. intersection) produces white colour.

resolution) were also produced from their corresponding GT masks, using simple 4×4 block-averaging (i.e., each 4×4 block in the high-resolution masks became a single pixel in the fuzzy, low-resolution masks). For each angle, we obtained a validation output for each of the following methods: traditional (i.e. binary validation post-thresholding), Yi [6], Chang [7], Crum [8] (i.e. the Gödel T_N), Łukasiewicz T_N, and finally the Exact and Generalised Directed operators. Fig 5a shows the difference between each method and the GTv at each angle.

Clinical example: The STARE (STructured Analysis of the REtina) Project [10] provides a clinical dataset of 20 images of human retinae, freely available online. For each image, it also provides a triplet of binary masks (700×605): two manual delineations of retinal blood vessels from two medical experts and one automated method. One of the manual sets was treated as the GS mask, and the other two were treated as SG masks (human rater vs computer algorithm); fig 4 shows an example of the automated SG against the GS. Similar to the petal set, fuzzy versions were produced using various degrees of block-averaging, and validated using the same array of methods. Fig 5b shows the validation accuracy and precision profiles of each method over the whole dataset for the case of a human rater (the algorithm-based results were very similar) at 4×4 block averaging. Larger blocks resulted in less accurate / precise curves (not shown here), but interrelationships between methods were preserved. Fig 5c compares human rater vs automated method accuracy as assessed by each validation operator.

4 Discussion and Conclusions

There are several interesting points to note with regard to the datasets and T_Ns: **1.** Pixels at the object boundary are generally more likely to be fuzzy than pixels at the core (indeed, true core pixels should be deterministic); for segmentations

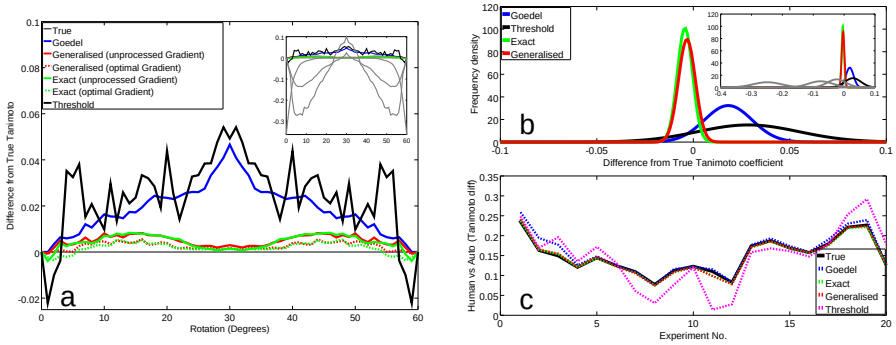

Fig. 5: a) Difference between the GTv and our proposed methods, the standard 0.5 threshold approach, and state of the art (Crum), plotted against different angles of rotation for the rotating petal set. b) Distribution of the differences between fuzzy and latent validations over 20 retinal sets (human rater only), represented as gaussian curves. The insets are 'zoomed-out' views, showing in gray the remaining methods ($\cap_L$, Yi and Chang) which were far less accurate to depict at this scale. c) Difference between human rater and computer algorithm SG validation for the above methods.

of large organ structures, one generally expects to see a relatively high overall-reported validation value, even with less sophisticated segmentation algorithms, due to the disproportionately large contribution of core pixels. More importantly, one might also expect less variability in validation values for the same reason, even between algorithms of variable quality. However, a segmentation algorithm's *true* quality / superiority compared to other algorithms, generally boils down to its superior performance on exactly such boundary pixels; it is therefore even more important in such objects that appropriate validation methods are chosen that ensure accurate and precise discriminating ability at boundary pixels as well as core ones. Nevertheless, for demonstration purposes, both the synthetic and clinical sets used in this paper, involved relatively thin structures, i.e. a high "boundary-to-core" pixel ratio. This choice was intentional, such that the relationship between fuzziness, choice of validation operator, and validation precision / accuracy could be demonstrated more clearly; this also explains why the GTv itself between the two experts was of a fairly low value (with a mean of the order of 0.6). **2.** The Generalised and Exact $\cap_D$ seem to perform equally well; oddly enough, the Exact $\cap_D$ seems to be slightly less accurate for the clinical dataset, but it is more precise. **3.** Both are much more accurate and precise than all the other methods investigated, and even more so when a more appropriate gradient response is used. **4.** Decreasing the resolution affects all methods negatively, but $\cap_D$ still performs much better. **5.** Out of all the methods in the literature, Crum's approach (i.e. $\cap_G$) is the next most accurate / precise overall; the $\cap_L$ operator, while very precise at times of bad overlap, seems to be completely off at times of good overlap. **6.** The threshold approach seems very unreliable, at least as demonstrated through these datasets: the theoretical maximum established by $\cap_G$ is violated consistently (fig 5a); it tends to be inaccurate and imprecise (fig 5b); and as a result, its use for comparing segmentation algorithms can lead to false conclusions: e.g. in fig 5c, experiment 11, the algorithm is judged to be

almost as good as the human rater (validation difference of 1%), but in fact, the GTv difference is as high as 10%. **7.** The Yi and Chang algorithms also violate this boundary. Furthermore they both seem to be overpessimistic at times of good overlap, and overoptimistic at times of bad overlap.

Conclusion: Validation is one of the most crucial steps in ensuring the quality and reliability of segmentation algorithms; however, the quality and reliability of validation algorithms *themselves* has not attracted much attention in the literature, despite the fact that many state-of-the-art segmentation algorithms are already deployed in clinical practice for diagnostic and prognostic purposes. We have shown in this paper that in the presence of non-deterministic sets, conventional validation approaches can lead to conclusions that are inaccurate, imprecise, and theoretically unsound, and we have proposed appropriate alternatives which we have demonstrated to be accurate, precise, and robust in their theoretical underpinnings. Further work is needed to evaluate and understand the advantages of the proposed validation framework on more, specific clinical scenarios, including segmentations of large organs and small lesions.

Acknowledgements

TP is supported by the RCUK Digital Economy Programme (grant EP/G036861/1: Oxford Centre for Doctoral Training in Healthcare Innovation). ED acknowledges the BHF intermediate clinical research fellow grant (FS/13/71/30378) and the NIHR BRC. VG is supported by a BBSRC grant (BB/I012117/1), an EPSRC grant (EP/J013250/1) and by BHF New Horizon Grant NH/13/30238.

References

1. Udupa, J. K. *et al.* A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics*, 2006, 30, 75-87
2. Zhang, Y. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 1996, 29, 1335 - 1346
3. Weisenfeld, N. I. & Warfield, S. K. SoftSTAPLE: Truth and performance-level estimation from probabilistic segmentations. *IEEE International Symposium on Biomedical Imaging*, 2011, 441-446
4. Ballester, M. À. G.; Zisserman, A. P. & Brady, M. Estimation of the partial volume effect in MRI. *Medical Image Analysis*, 2002, 6, 389 - 405
5. Deza, M. M. & Deza, E. *Encyclopedia of distances*, 2nd ed. Springer, 2013
6. Yi, Z. *et al.* Discriminative, Semantic Segmentation of Brain Tissue in MR Images. *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, 2009, 5762, 558-565
7. Chang, H.H. *et al.* Performance measure characterization for evaluating neuroimage segmentation algorithms. *NeuroImage*, 2009, 47, 122 - 135
8. Crum, W. R.; Camara, O. & Hill, D. L. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 2006, 25, 1451-1461
9. Bloch, I. & Maitre, H. Fuzzy mathematical morphologies: a comparative study. *Pattern Recognition*, 1995, 28, 1341-1387
10. Hoover, A.; Kouznetsova, V. & Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 2000, 19, 203-210