Christian Janorschke*, Jingyang Xie, Xinyu Lu, and Floris Ernst

About the Comparison and Evaluation of Real-Valued Multimodal Medical Point Clouds

https://doi.org/10.1515/cdbme-2025-0174

Abstract: Two- or three- dimensional point clouds are common representations of data in the medical context, especially in medical imaging and segmentation. When assigning real-valued coordinates and transformations - e.g., for registration with an additional imaging modality - metrics like the Dice similarity coefficient, which rely on set operations on discrete point sets, have to be approximated. We applied different calculation methods for multiple metrics on a synthetic dataset of basic shapes, combined with modifications like transformations or occlusions. The evaluation resulted in a wide range of values for all metrics, depending on the modification, calculation method, and parametrization. This variability suggests both a potential clinical risk and the opportunity for selectively reporting more favourable results in publications. Therefore, and considering the influence of point spacing and distribution, we recommend reporting multiple metrics from diverse categories in the context of real-valued point clouds.

Keywords: Point cloud evaluation metrics, real-valued point cloud comparison, 3D segmentation evaluation, multimodal medical imaging.

1 Introduction

Medical imaging plays a critical role in diagnostics and treatment planning. In complex cases, multimodal imaging is required either to acquire complementary information available in different modalities or to overcome physical limitations - for example in ultrasound (US)-guided interventions following a computed tomography (CT) scan. Typically, information transfer between different modalities is based on image registration. Due to different underlying coordinate systems, this process introduces sub-voxel displacements, resulting in real-valued coordinates for information like landmarks or segmentations. Additionally, differences in imaging modalities can lead to partial views of anatomical structures and variation in point density and spacing due to modality resolution.

Evaluation metrics such as the Dice similarity coefficient (DSC) [1] are typically defined on a discrete voxel grid. These metrics generally assume uniform sampling and equal cardinality. Thus, custom strategies or approximations are necessary when working with real-valued multimodal medical point clouds. This introduces undesired variability in metric outputs, which can affect medical risks, for example in case of radiation therapy errors due to misaligned registrations.

Considering known metric-related pitfalls in image analysis [2], this work aims to highlight the challenges of comparing real-valued point clouds and to provide guidance on the selection and computation of suitable evaluation metrics.

***Corresponding author: Christian Janorschke,** Institute of Robotics and Cognitive Systems, University of Lübeck, Ratzeburger Allee 160, Lübeck, Germany, e-mail: ch.janorschke@uni-luebeck.de

Jingyang Xie, Xinyu Lu, Floris Ernst, Institute of Robotics and Cognitive Systems, University of Lübeck, Lübeck, Germany

2 Methods

The literature offers a wide range of metrics for point cloud comparison [3]. In this work, we focus on the Hausdorff distance (HD) and its 95th percentile (HD95) as distance-based, DSC as overlap-based, volumetric similarity (VS) as volume-based and distance in center of mass (CoM) as localisation-based metric and compute different calculation methods over a synthetic dataset.

2.1 Hausdorff distance

The HD describes the maximum distance between one of the points $a \in A$ of one point cloud A to the closest point $b \in B$ of the other point cloud B [4]:

$$\mathrm{HD}(A,B) = \max(h(A,B), h(B,A)) \tag{1}$$

with $h(A,B)$ as the directed HD, or unidirectional HD:

$$h(A,B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{2}$$

Consequently, it is highly susceptible to outliers which is why the calculation of a specific quantile is recommended in applications like medical segmentations, where outliers are common and practically unavoidable. A common practice is the HD95 which accounts for a 5 % outlier ratio [5].

A simple solution for dealing with one partial point cloud C compared to one complete point cloud D is to reduce the HD to the unidirectional HD ($h(C,D)$), eliminating the high distance of points without correspondence in the incomplete point cloud but also not accounting for possible points from D with high distance but still belonging to the partial structure.

2.2 Dice similarity coefficient

The DSC [6] is the most common metric for medical segmentations [1] and falls into the category of overlay-based metrics:

$$\mathrm{DSC} = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

In the context of discrete points, the intersection only applies to identical points, whereas this definition fails in the context of real-valued point clouds. Therefore, a discrete version of the DSC can be defined with the intersection being defined as:

$$(A \cap B)_t := \{a \in A \mid \min_{b \in B} ||a - b|| \leq t\} \tag{4}$$

by defining an overlap radius t within which points are matched. On a unit-spaced grid, reasonable values for t include 1 (adjacent voxels) or $\sqrt{3}$ (diagonal neighbours). This approach does not compensate for unequal point counts. Another way of calculating the DSC is the volumetric formulation with the volumes V_A and V_B:

$$\mathrm{DSC} = \frac{2\,(V_A \cap V_B)}{V_A + V_B} \tag{5}$$

In this work, volume approximation is performed with alpha shape triangulation of the point clouds [7]. This introduces the alpha radius α as parameter. High values for α will result in filled non-convex shapes, while a small α creates jagged boundary regions and occlusions within the volume.

2.3 Volumetric similarity

The definition of VS varies in the literature. Based on [3], we use the following definition:

$$\mathrm{VS} = 1 - \frac{||A| - |B||}{|A| + |B|} \tag{6}$$

Since the definition is based on the number of points per point cloud, VS can be calculated for real-valued data. Similar to the DSC, volume reconstruction algorithms like triangulation can be used for a volumetric calculation for uneven distributions. Alternatively, the volume reconstruction can be used to resample both point clouds over a uniform grid.

2.4 Center of mass

The CoM can be calculated as the mean coordinate for point clouds when assuming consistent density. In case of reconstructions of the volumes, the analytic definition as balance point can be applied for calculation or the point clouds can be resampled over a uniform grid. The Euclidean distance between the CoM of ground truth and test data quantifies the global displacement between both datasets.

2.5 Dataset

The evaluation of the different methods, metrics, and approximation techniques requires a dataset that reflects a range of geometric and structural characteristics. Therefore, we created ten ground truth point clouds ranging from a basic cube to more complex structures to incorporate multiple influencing factors such as rotational or axial symmetry and extreme surface area-to-volume ratios. They are shown in Fig. 1. The point clouds are generated in within limits of 0 mm and 100 mm for all axes in discrete steps of 2 mm in all directions. All computations were carried out in MATLAB (version 9.13.0, The MathWorks, Inc., Natick, MA, USA) [8].
Adding Gaussian noise and applying these modifications results in a total of 100 real-valued datasets for evaluation:

1. Just Gaussian noise with standard deviation of 1.5 mm
2. Translation of 2 mm in direction of x-axis
3. Translation of 20 mm in direction of x-axis
4. Translation of 10 mm in direction of z-axis
5. Rotation by 5 degree around center axis in z direction
6. 2 mm x-axis translation + 5 degree rotation around z-axis
7. Randomized downsampling by 50 %
8. Directed uneven distribution of points (downsampling proportional to x-axis from 20-100 %)
9. Occlusion of sphere (at $(60,30,30)^{\mathrm{T}}$, $r = 30$)
10. Cropping points with $y > 70$

The transformations are parametrized based on size and spacing of the ground truths. In the context of multimodal medical imaging, the differences in technology affect point spacing, density and distribution. Thus, we included randomised downsampling and an uneven distribution. The occlusion represents an image artefact and cropping simulates a partial view. Fig. 2 shows the cropped bucket-shaped point cloud alongside the reconstruction for two different alpha radii, generated via MATLAB's *alphashape* function [8]. Corresponding volumes were determined with the *volume* function for alpha shape objects.

3 Results

Tab. 1 presents results for the DSC calculation. Discretization was performed using two distance thresholds (2 mm and 3.5 mm), and alpha shape triangulation was applied with two different alpha radii (α = 5 and α = 10). Bi- and unidirectional HD95, the difference in CoM and the VS are listed in Tab. 2. CoM and VS were computed using three approaches: directly from the original point cloud, from resampled point clouds derived from the alpha shape reconstruction and from the enclosed volume of the alpha shape reconstruction. An al-

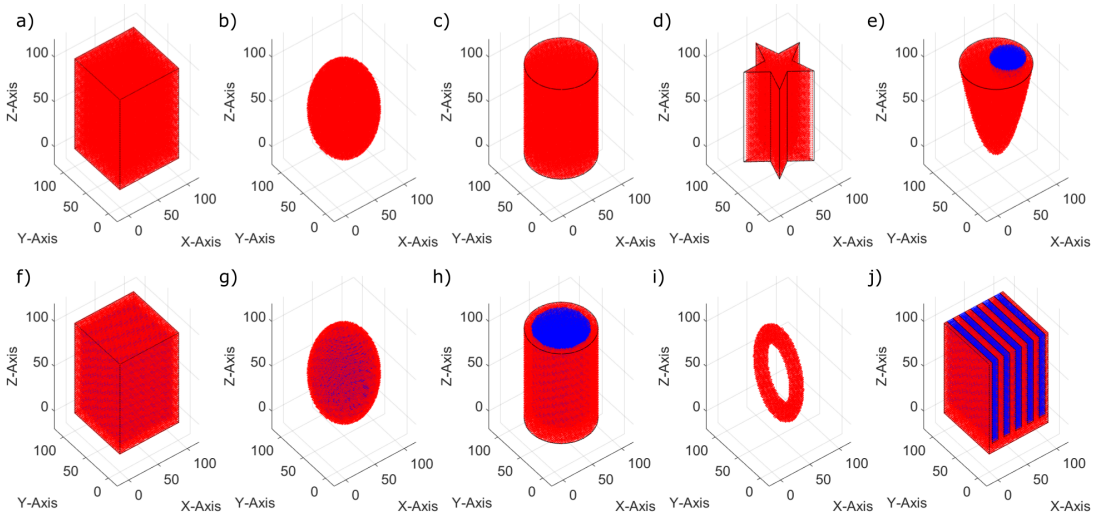

Fig. 1: Ground truth point clouds: a) cube b) sphere c) cylinder d) pentagrammic prism e) paraboloid with spherical occlusion f) hollow cube g) hollow sphere h) bucket i) torus j) cooling unit. Red: point cloud, blue: occlusions and gaps for visualization purposes

Tab. 1: Dice scores from discretization and alpha shape triangulation for ten point clouds and their real-valued, modified versions.

Dice score:	Discrete t = 2 mm				Discrete t = 3.5 mm				Alpha shapes α = 5				Alpha shapes α = 10			
Test data	**Min**	**Max**	**Mean**	σ	**Min**	**Max**	**Mean**	σ	**Min**	**Max**	**Mean**	σ	**Min**	**Max**	**Mean**	σ
Gauss. n.	**.81**	**.96**	**.91**	**.05**	**.99**	**1.0**	**1.0**	**.00**	**.52**	**.75**	**.69**	**.07**	**.68**	**.96**	**.85**	**.11**
Transl. x1	**.75**	**.95**	**.87**	**.08**	**.96**	**1.0**	**.99**	**.01**	**.44**	**.75**	**.66**	**.10**	**.67**	**.95**	**.84**	**.12**
Transl. x2	**.00**	**.78**	**.50**	**.27**	**.04**	**.84**	**.58**	**.27**	**.00**	**.61**	**.36**	**.22**	**.00**	**.78**	**.47**	**.27**
Transl. z	**.28**	**.88**	**.71**	**.19**	**.53**	**.92**	**.83**	**.13**	**.11**	**.68**	**.52**	**.18**	**.24**	**.87**	**.67**	**.21**
Rotation	**.74**	**.96**	**.87**	**.08**	**.93**	**1.0**	**.98**	**.02**	**.46**	**.75**	**.64**	**.11**	**.63**	**.96**	**.82**	**.13**
Tr. + Rot.	**.53**	**.92**	**.77**	**.14**	**.82**	**.98**	**.91**	**.06**	**.22**	**.71**	**.56**	**.16**	**.49**	**.92**	**.74**	**.16**
Downs.	**.54**	**.64**	**.61**	**.04**	**.66**	**.67**	**.66**	**.00**	**.13**	**.26**	**.23**	**.04**	**.76**	**.97**	**.89**	**.08**
Uneven dst.	**.61**	**.72**	**.68**	**.04**	**.74**	**.75**	**.75**	**.00**	**.25**	**.41**	**.36**	**.05**	**.73**	**.97**	**.88**	**.09**
Occlusion	**.77**	**.91**	**.84**	**.05**	**.86**	**.99**	**.92**	**.04**	**.46**	**.69**	**.61**	**.07**	**.65**	**.88**	**.78**	**.09**
Cropping	**.65**	**.88**	**.76**	**.08**	**.77**	**.94**	**.83**	**.05**	**.39**	**.67**	**.55**	**.09**	**.56**	**.85**	**.73**	**.11**

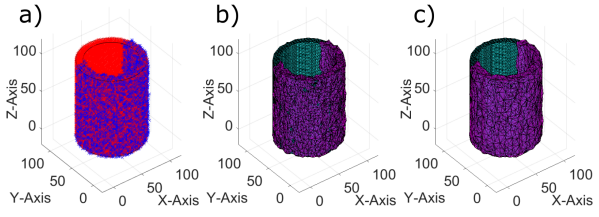

Fig. 2: Test data generation and reconstruction for the bucket shape: a) ground truth (red) and test data (blue), b) alpha shape reconstruction $\alpha = 5$, c) alpha shape reconstruction $\alpha = 10$

pha radius of 10 was selected, as it yielded DSC results more consistent with those obtained through discrete evaluation.

4 Discussion

While some of the ground truth point clouds can be categorized as academic examples, the inclusion of, for instance, the shape inspired by a cooling unit emphasizes the influence of the physical shape on specific metrics. By design, a translation of specifically the gap length leads to a perfect overlap of all but the final cooling fins, resulting in a higher DSC than smaller translations. This illustrates that a higher DSC does not necessarily imply a closer match to the ground truth. Similarly, rotations have negligible impact on all metrics when applied to nearly rotationally symmetric shapes. This is evident from the maximum DSC values for rotated point clouds which match those with added Gaussian noise. In medical applications, such as evaluating the helix-shaped contraction of a left ventricle, this behaviour becomes critical: rotational components may be under- or overestimated if the evaluation relies solely on segmented point clouds. The difference in CoM behaves analogously under rotation, showing minimal displacement when the rotation is centered around the CoM itself.
Discretization of the DSC highlights the importance of the threshold. The greater threshold resulted in a mean increase of 0.09 in DSC, despite both thresholds being derived from the

Tab. 2: 95th percentile Hausdorff Distance, bi- (HD95) and unidirectional (U-HD95), difference in center of mass based on points (CoM), resampling (CoM-RS) or approximated balance point (CoM-α), all in mm. Volumetric similarity based on points (VS), resampling (VS-RS) and approximated volumes (VS-α). Reconstruction was performed with alpha shape triangulation with α = 10 mm.

Metric:	HD95		U-HD95		CoM		CoM-RS		CoM-α		VS		VS-RS		VS-α	
Test data	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Gauss. n.	1.8	0.1	1.8	0.4	0.0	0.0	0.4	0.3	0.4	0.4	1	0	.85	.10	.85	.10
Transl. x1	2.1	0.3	2.2	0.7	2.0	0.0	2.1	0.2	2.1	0.2	1	0	.85	.10	.85	.10
Transl. x2	15.7	2.1	16.8	2.2	20.0	0.0	20.1	0.2	20.1	0.2	1	0	.86	.09	.85	.10
Transl. z	6.3	1.5	7.2	1.6	10.0	0.0	10.1	0.4	10.2	0.5	1	0	.87	.10	.85	.10
Rotation	2.2	0.6	2.4	1.0	0.0	0.0	0.4	0.3	0.5	0.4	1	0	.85	.10	.85	.10
Tr. + Rot.	3.5	1.2	4.1	1.5	5.0	0.0	5.1	0.1	5.1	0.2	1	0	.86	.10	.85	.10
Downs.	2.3	0.1	1.8	0.4	0.2	0.1	0.4	0.2	0.5	0.4	.67	.00	.88	.09	.88	.09
Uneven dst.	2.3	0.2	1.8	0.4	9.0	4.5	1.4	1.0	1.4	1.0	.75	.00	.88	.09	.88	.09
Occlusion	5.4	3.7	1.8	0.4	5.0	2.8	5.2	2.9	5.3	2.9	.93	.04	.91	.11	.91	.11
Cropping	19.0	8.8	1.8	0.4	13.7	5.7	13.7	5.7	13.7	5.7	.84	.05	.92	.05	.92	.04

grid spacing. Reconstruction approaches compensate for non-uniform distributions. Still, results depend on the quality of reconstruction. Tab. 1 shows that $\alpha = 10$ produces results closer to the discrete calculation which may be an indicating factor for a better fitting parameter. Additional volume caused by the triangulation of the Gaussian noise at the surfaces resulted in about 15 % deviation between reconstruction-based VS and point-wise calculation. Nonetheless, the results are consistent, even for non-uniform distributions.

In scenarios involving cropped volumes - such as physically limited US scans registered to complete CT volumes - only unidirectional metrics remain unaffected. While this approach also addresses differences in point count and distribution, it inherently overlooks missing structures in the test data.

In the context of real-valued multimodal medical point clouds, additional factors must be considered compared to conventional metric selection [3]. While the choice of metric remains influenced by sensitivity to translations, rotations, scaling, and deformations, the imaging modality and its resolution introduce further variability, such as differences in point spacing, density, and distribution, as well as partial views or artefacts. Volume reconstruction techniques perform most consistently for the DSC, but the results should be validated through qualitative assessment of reconstruction quality and supplemented with metrics from different categories.

5 Conclusion

The experimental evaluation yielded a wide range of values for all metrics across identical datasets. In the context of real-valued multimodal medical data, factors like geometry, spacing, or point distribution must be taken into account. Therefore, selecting appropriate evaluation metrics and justifying the computation method and parameter choices are essential to mitigate risk and ensure unbiased reporting. Since strong performance in a single metric does not necessarily reflect closer alignment with the ground truth, we recommend reporting multiple metrics from diverse categories.

Author Statement

CJ, JX and XL have received founding from EU Grant Agreement No. 945119 outside of the presented work. All authors state no conflict of interest.

References

[1] Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 2028;9:5217

[2] Reinke, A., Tizabi, M., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D.,Kavur, A., et al. Understanding metric-related pitfalls in image analysis validation. Nat Methods 2024;21:182–194

[3] Taha, A., Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29

[4] Huttenlocher, D., Klanderman, A., Rucklidge, W. Comparing Images Using the Hausdorff Distance. IEEE Trans Pattern Anal Mach Intell. 1993;15:850–63

[5] Reinke, A., Tizabi, M., Sudre, C., Eisenmann, M., Rädsch, T., Baumgartner, M., et al. Limitations of Image Processing Metrics: A Picture Story. arXiv preprint arXiv:2104.05642, 2021

[6] Dice, L. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297-302

[7] Edelsbrunner H, Mücke E. Three-dimensional alpha shapes. ACM Transactions on Graphics. 1994;13(1):43–72

[8] The MathWorks Inc. (2022). MATLAB version: 9.13.0 (R2022b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com