Evaluation of Evaluation Metrics

Grimm Lorenz and Rohrbach Simon

***Abstract*—In this paper, we have a look at selected evaluation metrics. We performed a series of tests in order to evaluate the metrics responses to certain influences and to determine, which evaluation metrics are the most suited, for what kind of application.**

I. Introduction

IN the field of medical image analysis a large multitude of evaluation metrics exists, which causes one question: Which one to chose? In this paper, we aim to answer this questions, based both on literature and our own measurements we will determine the most optimal evaluation metrics.

II. Materials and Methods

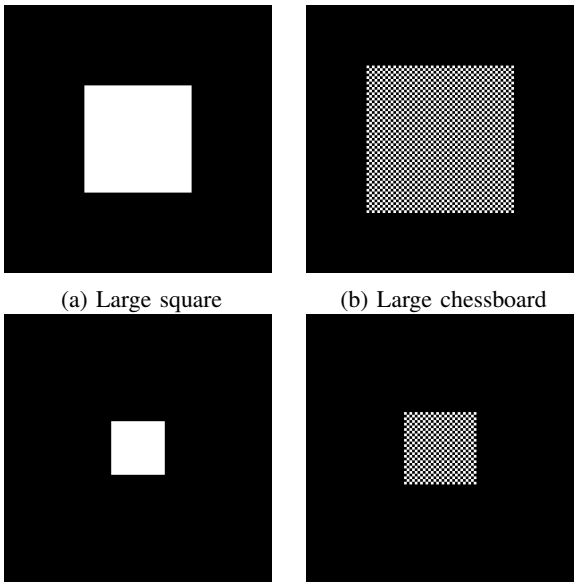
The aim was to test the influence of different factors and modifications on a given feature of an image. In order to investigate this, we decided to create a simple template image (Figure: 1a) in shape of a square, as to limit the shape bias and to consider possible influences caused by a complex shape. For this purpose, we define a black and white image, such as black pixel represent background and white pixels represent the feature the metric should consider.

The investigation of multiple features was not done, as these multiple features can just be turned into multiple individual segmentations, each with only one feature, making the other features part of the background. With this, any number of features can be analyzed. If a total statement of all features is required, the individual answers can be averaged and weighted according to importance.

From prior knowledge, we knew that certain metrics, such as the DICE, react sensitively to small feature density and small amounts of the feature. In this context, small density refers to the features being only present as small clusters and small amounts of feature refer to the fact, that the background takes up the majority of the image.

To investigate the influence of the density, we created a second image(Figure: 1b) apart from the square. A chessboard which has a feature area roughly equivalent to the one in the square image, making the main difference between these two images the feature density.

To investigate the influence of the amount of feature in the image, we made a second set of images (Figure: 1c and Figure: 1d), with each image containing one fourth of the amount of features present in the corresponding bigger images.



(a) Large square (b) Large chessboard

(c) Small square (d) Small chessboard

Fig. 1: The total white area is identical for both the large and the small shapes respectively. Horizontal comparison alows to evaluate density, vertical comparison alows to evaluate size.

For each of those four images, we imposed four different types of modifications, each performed with small steps, as large deviations between a created segmentation and a given ground truth can already be determined by the naked eye and thus do not require an evaluation metric.

1) Planar shift: The images were shifted by 1 through 5 pixels in steps of 1 pixel. This will be the main basis for comparison, as it turned out to be the most linear and consistent type of modification.
2) Upscale and downscale: The images were scaled by 1 through 5 percent in steps of 1 percent. This was done both for upscaling and downscaling. Upscaling changes the segment by adding false positive (FP) values, while downscaling changes the segment by adding false negative (FN) values.
3) Blurring the edges: The edges of the features inside the images were blurred, by exchanging random pixels in a defined distance within the image. This distance was varied by 1 through 5 pixels in steps of 1 pixel. This changes the border of the segment clusters in a different way than the scaling does.
4) Rotation: The images were rotated by 1 through 5 degrees in steps of 1 degree. Works as an approach but the interpolation makes changes inconsistent for small angles.

III. Results

To present our results in a clear manner, we will classify them by the types of metrics we looked at. The results are presented as they are and the appropriate conclusions are drawn from the measurements, as well as from literature, for each metric respectively. The consequence of these results are later on discussed in more detail in chapter IV: Discussion and their essence are synthesized in chapter V: Conclusion.

A. Overlap based evaluation metrics

Overlap based evaluation metrics are the most intuitive type of metric to understand. The segmentation and the ground truth are overlayed and each pixel in the segmentation is compared to the corresponding pixel in the ground truth. Depending on the two values the pixel position gets one of four states assigned.

1) True Positive: If the segmentation defines the pixel as belonging to the feature and the ground truth says the pixel belongs to the feature, the pixel is classified as True Positive (TP).
2) False Positive: If the segmentation defines the pixel as belonging to the feature and the ground truth says the pixel belongs to the background, the pixel is classified as False Positive (FP).
3) True Negative: If the segmentation defines the pixel as belonging to the background and the ground truth says the pixel belongs to the background, the pixel is classified as True Negative (TN). Metrics considering the true negatives have a strong dependence on the amount of background in the image and with that a sensitivity to the feature size.
4) False Negative: If the segmentation defines the pixel as belonging to the background and the ground truth says the pixel belongs to the feature, the pixel is classified as False Negative (FN).

After each pixel has been assigned to a cardinality, they are summed up and their results are stored in the so-called confusion matrix.

True Positive	False Positive
False Negative	True Negative

TABLE I: Confusion Matrix

The most important overlap based metric, we are going to talk about is the F- Measure. The F-Measure takes the sensitivity (TPR) and the precision (PPV) into consideration, which both are based on the confusion matrix, and weights them according to a factor β. The DICE is the special case of the F-Measure for which β is equal to 1.

$$TPR = \frac{TP}{TP + TN} \quad (1) \qquad PPV = \frac{TP}{TP + FP} \quad (2)$$

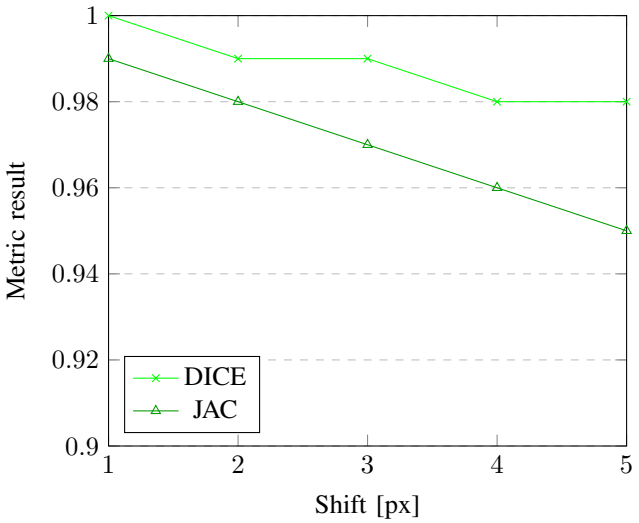
$$FMS_\beta = \frac{(\beta^2 + 1) \cdot PPV \cdot TPR}{\beta^2 \cdot PPV + TPR} \quad (3)$$

The DICE ist the most commonly used evaluation metric. Depending on the circumstance one could desire a metric such as the Jaccard index (JAC), which is based on the DICE but reacts more sensitive to changes.

$$JAC = \frac{DICE}{2 - DICE} \quad (4)$$

To show the relationship between the DICE and the Jaccard index we compare them directly. Below their reaction to the large square is shown.

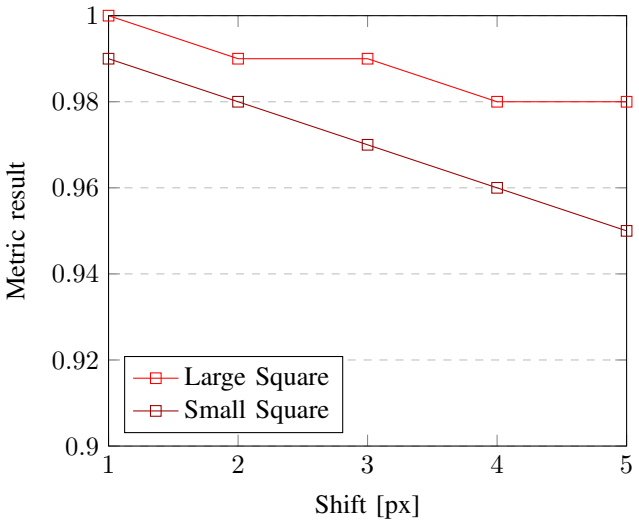
Comparison of DICE and JAC [Large square]



We can see, that the Jaccard index reacts stronger to the same change than the DICE does and this is independent of the type of change, as this behavior could be observed in all our measurements.

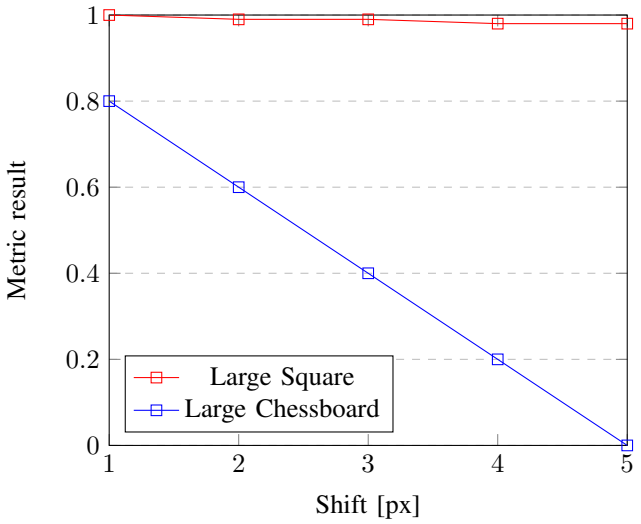
For further tests, we limit the representation to the DICE as it is the most used of the two metrics and they depend mathematically on each other. We investigated then further the influence of feature size and density for the DICE, which yielded the following results confirming our expectations:

Size comparison for DICE [Square]



Similiar to how the Jaccard index reacts more sensitive than the DICE on an identical image, the DICE reacts more sensitive on smaller features, as this size comparison shows.

Density comparison for DICE [Large]



This measurement clearly shows the extreme dependance of the DICE on high feature density, as a small change between the segmentation and the ground truth, already changes the result of the DICE drastically.

B. Probabilistic evaluation metrics

In contrast to the overlap based evaluation metrics, the ones based on a probability measure the correlation between the two segments, rather than just evaluating how good the areas overlap. Correlation is a statistical relationship, that determines to how close the relationship between two segments to a linear relation is.

The first probabilistic evaluation metric we looked at was the Interclass Correlation (ICC) given by the following formula, with σ_S as the variance caused by the differences between the segmentation and the ground truth and σ_ϵ as the variance caused by differences between the points of the segmentation and the ground truth.

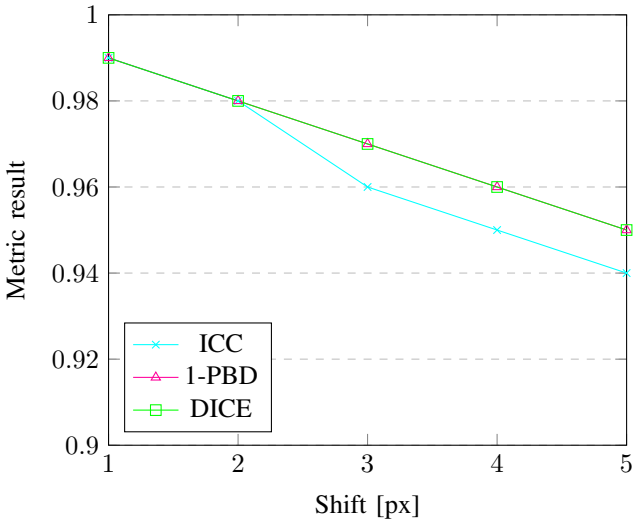
$$ICC = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_\epsilon^2} \tag{5}$$

The second probabilistic evaluation metric we investigated was the Probabilistic Distance (PBD) given by the formula below, with P_S and P_G are the probability distributions of the segmentation and the ground truth respectively and P_SG being their joint probability distribution.

$$PBD(S,G) = \frac{\int |P_S - P_G|}{2 \int P_{SG}} \tag{6}$$

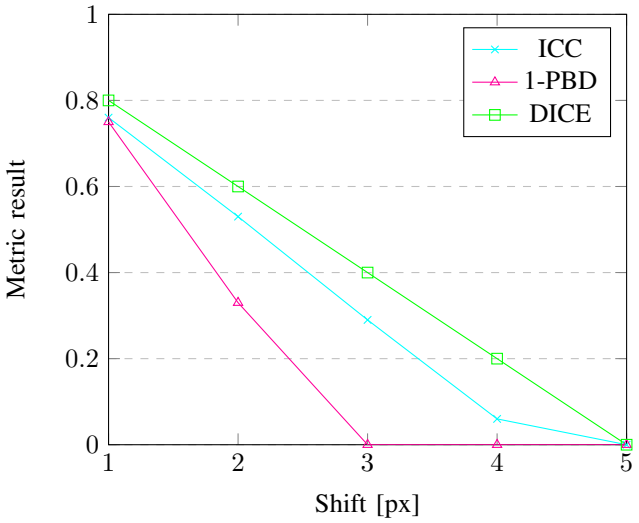
Both of these metrics displayed the same sensibility as the overlap based metrics to the size and even stronger sensibility to the density. This is shown in the two figures below, as the interclass correlation and the probabilistic distance are plotted against the DICE for comparison.

Comparison of ICC, PBD and DICE [Large Square]



This figure shows the reaction of the three metrics to the small square, the reaction to the big square is analogous to the small one, except smaller (See Size comparison for DICE). From the measurements, we were able to deduct, that both the interclass correlation and the probabilistic distance react about equal to the DICE, which makes them redundant metric choices in this perspective.

Density comparison of ICC, PBD and DICE [Large]



As it can be derived from this figure, the probabilistic metrics react even stronger to changes in the large chessboard than the dice, which makes them even more sensitive to low densities.

The Cohen Kappa Coefficient(KAP) measures how good the agreement between two given segments is. A big advantage of this metric compared to the others is, that it takes changes caused by chance into consideration, however, at the same time, it still suffers from the high sensitivity to size and density, even though a bit less than the others. It considers the agreement between the segmentation and the ground truth P_a and the hypothetical probability, that the two agree by chance P_c, as given in the formula below.

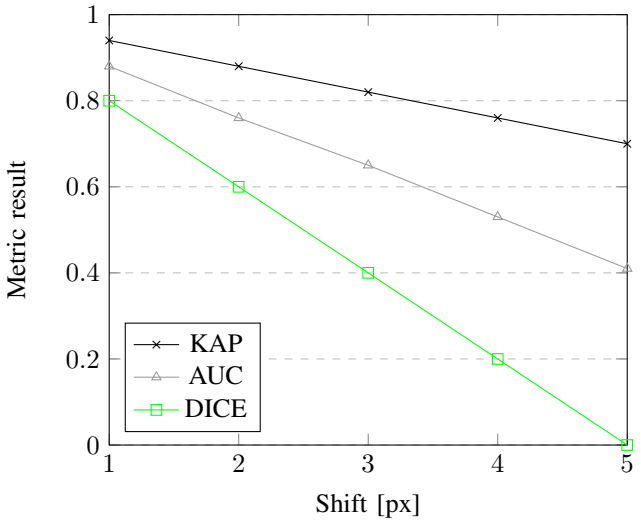
$$KAP = \frac{P_a - P_c}{1 - P_c} \tag{7}$$

The Area Under Curve (AUC) considers the fallout (FPR) and the complementarily (FNR), which causes it to suffer from the same problems as the other metrics mentioned up until now.

$$AUC = 1 - \frac{FPR + FNR}{2} \tag{8}$$

The Cohen Kappa coefficient and the area under the curve are compared to the other probabilistic metrics less sensitive to density as the following figure will show. Even though the sensitivity falls short of the DICE it is still rather high. In our measurements, it further turned out that the reaction to size was only minimal jet still detectable.

Density comparison of KAP, AUC and DICE [Large]



C. Information theoretic based evaluation metrics

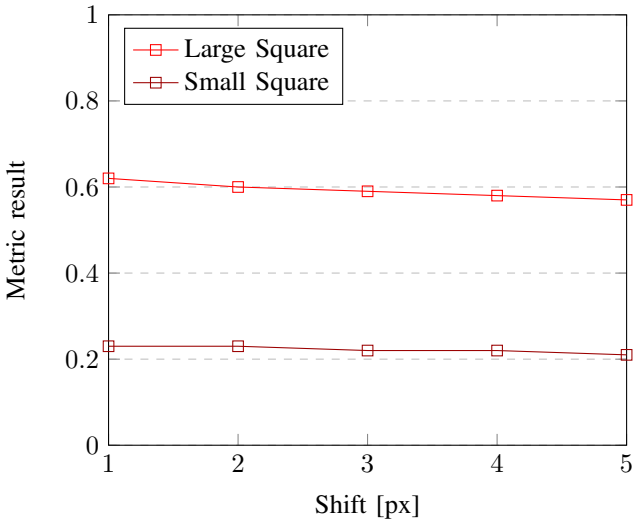
The Mutual Information (MI) measures the amount of information the segmentation contains about the ground truth and vise versa, or with other words, how similar the two segments are. The formula below is simplified as H(S) corresponds to the entropy of S. The big advantage of this is, that the calculation is region based and not pixel based.

$$MI = H(S_g) + H(S_t) - H(S_g, S_t) \tag{9}$$

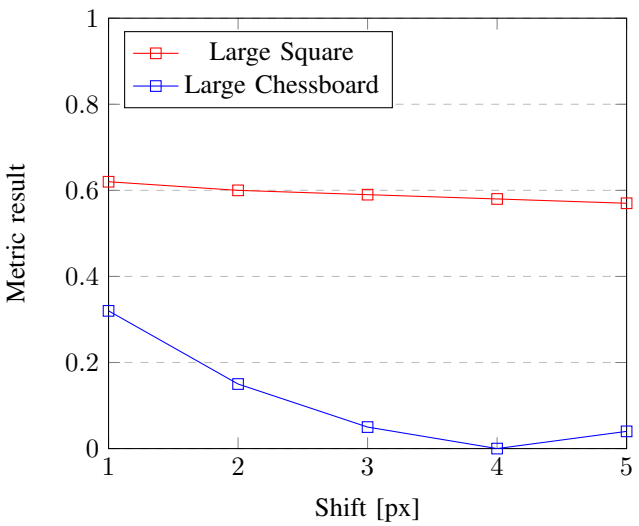
$$VOI = 2H(S_g) + 2H(S_t) - 3H(S_g, S_t) \tag{10}$$

The Variation of Information (VOI) is a modification of the MI, as it effectively measures the change of information between the segments (loss or gain). Since the variation of information is derived from the mutual information, we will primarily consider the mutual information in our representation going forward.

Size comparison for MI [Square]



Density comparison for MI [Large]



Our measurements have shown, that both these metrics react strongly to changes in size and density, however, they react only slightly to modifications of the image. This can be seen in the two figures above.

D. Pair counting based evaluation metrics

The Rand Index (RI) measures the similarity between two clusters, one taken from the segmentation and one from the ground truth. The comparison is mathematically related to the accuracy, but the important difference is, that the Rand index is not label based, so it can even be used when no labels are in place. In the equation given below (a+b) can be considered the total number of agreements and (c+d) can be considered the total number of disagreements between the segmentation and the ground truth.

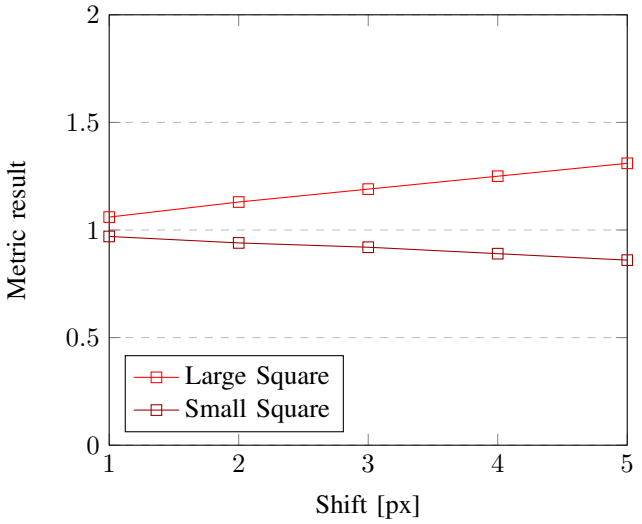
$$RI = \frac{a + b}{a + b + c + d} \tag{11}$$

The modification of the RI is the Adjusted Rand Index (ARI), which corrects the RI for influences caused by chance.

$$ARI = \frac{2(ad - bc)}{c^2 + b^2 + 2ad + (a + d)(c + b)} \tag{12}$$

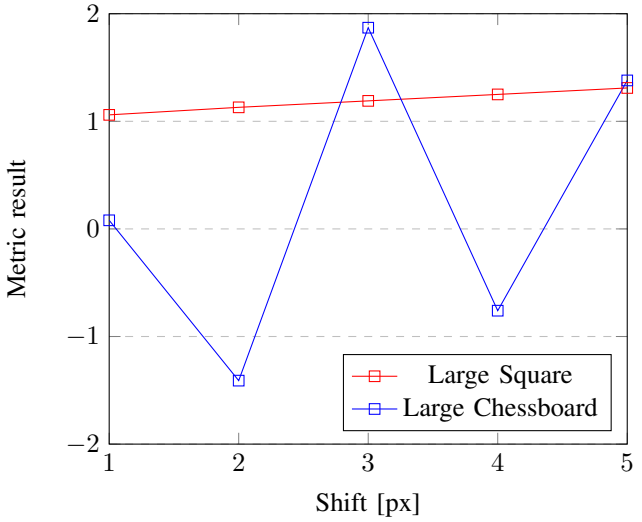
The Rand index showed only a small reaction to changes in size and density, the adjusted Rand index, on the other hand, showed stronger reactions.

Size comparison for ARI [Square]



This figure shows a dependance size, as it was the case for all metrics until this point.

Density comparison for ARI [Large]



In this figure a very strong dependence on the density can be seen, as the small changes within the chessboard image cause extreme variation in the adjusted Rand index.

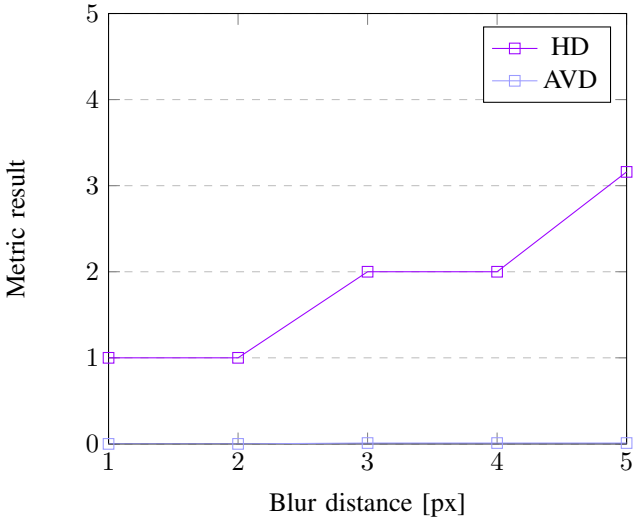
E. Spacial distance based evaluation metrics

The Hausdorff Distance (HD) measures the distance between two given subsets. The distance then is defined as the difference between the two closest points and the two points the most far away between the two subsets. This has the effect, that outliers are strongly taken into consideration, an effect, that no metric until now has displayed. Depending on the application, this can be sought after, for example, to empathize on the boundaries, but mostly this is a hindrance. Our measurements clearly showed this behavior, as changes to the boundary had a tremendous effect, while other changes did almost nothing.

The Average Hausdorff Distance (AVD), calculates all distances in the subsets and averages over them, to account for the outliers. This makes it more stable and an overall better choice than the regular Hausdorff distance, if the outlier sensitivity is not desired.

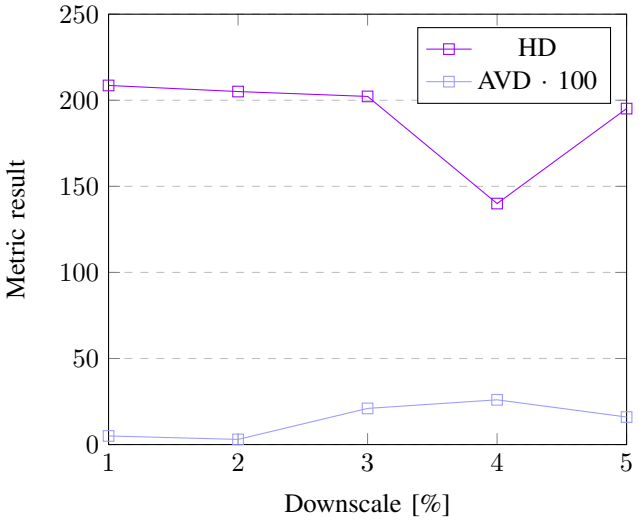
Slight changes, such as rotation or shifting the image, as it was the most effective modification throughout all other metrics, showed almost no reaction, as the Hausdorff distance detected no big change. Where reaction could be observed was if changes happened primarily to the boundary of the image, such as blurring the edges and scaling the image.

Comparison of HD and AVD [Large Square]



By blurring the edges slightly the Hausdorff distance reacts, whereas the averaged Hausdorff distance detects only a minimal modification. The blur as modification adds to the same extend new FP and FN and this happens equally around the square so that these changes overall seem to negate each other. If they would not, something like in the figure below can be observed.

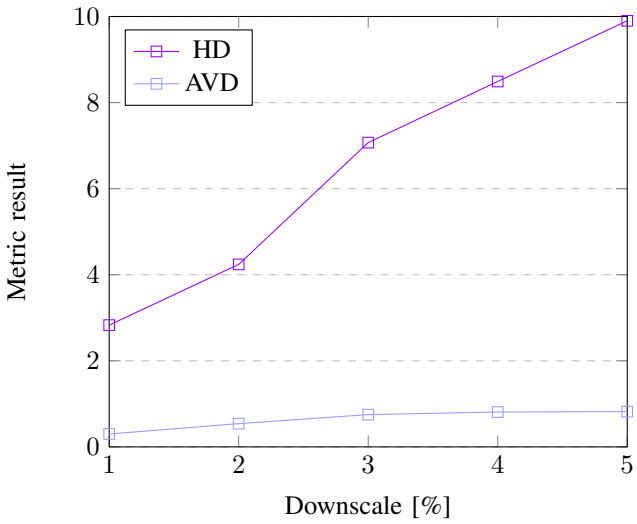
Comparison of HD and AVD [Large Square]



Special emphasis is to be given to the scale of this figure compared to the figure above this one. The downscale modification adds one layer of FN all around the square, which

adds up to an extremely high value for the Hausdorff distance. Here becomes the difference between the Hausdorff distance and the averaged Hausdorff distance evident, as the later had to be scaled a hundredfold just to be visible in the plot.

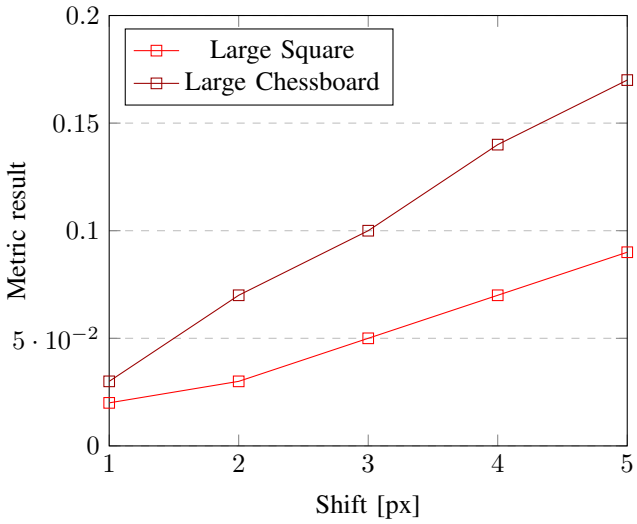
Comparison of HD and AVD [Large Chessboard]



Apart from the dependence of the boundary of the Hausdorff distance, both metrics seem to be uninfluenced by size or density, as this figure shows quite a similar behavior to the square.

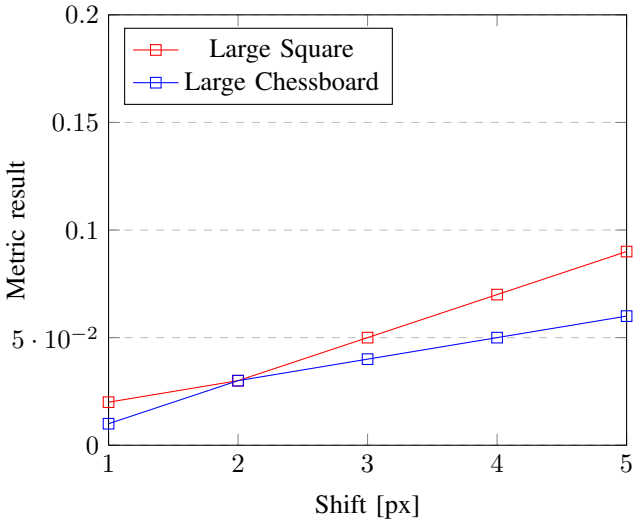
The other spacial distance based metric we looked at, was the Mahalanobis Distance (MHD). While it works similar to the Hausdorff distance it also considers the spacial position of the false negatives (FN) and the false positives (FP), not just the spacial position the true positives (TP) and/or the true negatives (TN) like most other spacial metrics do. In essence, this makes the Mahalanobis Distance one of the best-structured evaluation metrics.

Size comparison for MHD [Square]



The influence of the size is quite small compared to what other metrics have shown.

Density comparison for MHD [Large]



In this figure, it becomes evident that the Mahalanobis distance is not at all influenced by the density of the feature.

F. Volume based evaluation metrics

Since all our tests were performed on two-dimensional images, we could not deliver results for the volume based evaluation metrics, however, it is the metric of choice, if the general alignment of the whole volume is the most important requirement, as every other metric does not provide such a functionality.

IV. Discussion

Our measurements have shown, that almost all metrics have a medium to a strong dependency on the size and density of the feature present in the image. However this does not automatically exclude these metrics from being used, one just has to be aware of their weaknesses.

1) Overlap based evaluation metrics: If the general alignment of the segmentation and the ground truth can be guaranteed, then the overlap based metrics can be utilized to detect even the smallest deviations between the segmentation and the ground truth. This makes them perfect evaluation metrics to polish the last bit of the algorithm used to match the segmentation on the ground truth. This means, however, that these metrics highly depend on the existence of a ground truth for testing. Furthermore, these metrics are all stable against outliers and boundary effects.
2) Probabilistic evaluation metrics: These metrics are about equal in dependency on the feature size, as the overlap based metrics are, but they react quite differently to the density. As one part of the probabilistic metrics reacts more, the other part reacts less than the overlap based metrics (here Cohens Kappa coefficient excels over the others as it considers chance in its computations). In the end the results of these metrics are comparable to the overlap based metrics, however, they need much more computational effort, which might have to be considered, if whole body volumes would be evaluated. These metrics as well are stable against outliers and boundary effects. They also work best if the general alignment is given and on the premise that a ground truth is given.
3) Information theoretic based evaluation metrics: The entropy as a means to measure correlation works fine, but is also quite calculation intensive, especially compared to the computational effort of the overlap based metrics. The reaction to size and density are as it is for the overlap and probabilistic metrics as well. The main advantage of these metrics over the others is that they can be used to compare consecutive segmentations to determine changes in between segmentations. This is especially useful if no real ground truth exists.
4) Pair counting based evaluation metrics: While the adjusted Rand index also takes the chance into consideration, the sensitivity to size and especially the sensitivity to density are quite bad. One might consider using these metrics, as they require less computational effort than the information theoretic based and the probabilistic metrics, while still considering more relations than the overlay based metrics. This is possible since these metrics only considers clusters and not the whole segmentation.
5) Spacial distance based evaluation metrics: These metrics deliver by far the best overall results and can be used in almost every case mentioned above, without a problem. The only exception being the Hausdorff distance, which reacts sensitively to outliers. The averaged Hausdorff distance is capable of identifying every type of modification we imposed, while still being independent of size and density, which makes it a great all-around metric. The only problem with the averaged Hausdorff distance is, that it can not guarantee good alignment of the segmentation and the ground truth. The Mahalanobis distance is the only tested metric, capable of securely determine correct alignment of the segmentation and the ground truth. This ability is independent of size and density of the features. However, the Mahalanobis distance has difficulties if the boundaries become too complex.

V. Conclusion

Based on our measurements, which support the findings of other literature such as the paper published by A.Taha and A.Hanbury, with the goal to segment brain scans, we recommend to use a combination of the Mahalanobis distance, averaged Hausdorff distance and the DICE.

The Mahalanobis distance is the only metric that guarantees correct alignment of the segmentation and the ground truth, while the averaged Hausdorff distance, determines the detailed alignment. The averaged Hausdorff distance is superior to the DICE, as the DICE is strongly influenced by feature size and density. The reason why the DICE is still used is that under the right conditions it is absolutely viable and also it is probably the most referenced metric in publications, so it offers a value to compare to the literature.

Appendix A
as png

Database: Images with modifications

Appendix B
as excel/ods

Measurements and corresponding data evaluation

References

[1] A. Taha and A. Hanbury, *Metrics for evaluating 3D medical image segmentation:analysis, selection and tool*. Vienna, Austria: BioMed Central, 2015.