5
METRICS

This chapter explains the quality metrics considered for evaluation of the segmentation results. A total of seven metrics have been chosen. These metrics include many ways to compare a machine segmentation (S) with a corresponding ground truth (GT).

5.1.
Hoover metric (H)

Proposed by Hoover et al. (1996), it compares a segmentation outcome S with its ground truth GT. Let N_S be the number of segments in S and N_{GT} the number of segments in the GT. Then, the number of pixels in each machine segment C_i (where $i = 1, \dots, N_S$) is called O_i. Similarly, the number of pixels in each ground truth segment C_j (where $j = 1, \dots, N_{GT}$) is called O_j. Let $O_{ij} = C_i \cap C_j$ be the number of pixels in both segments C_i and C_j. Thus, if there is no overlap between the two segments, $O_{ij} = 0$, while if there is a complete overlap, $O_{ij} = O_i = O_j$. According to this metric, a segment could be classified into five different instances: **correct detection**, **over-segmentation**, **under-segmentation**, **missed** and **noise** (cf. Hoover et al. (1996)).

As demonstrated by Jiang et al. (2006), it is only necessary to consider **correct detections** in our analysis. Then, a **correct detection** is defined as follows (Hoover et al., 1996):

*A pair of regions C_j, in the GT image, and C_i, in the segmentation S, are classified as an instance of **correct detection** if:*

1. *$O_{ij} \geq T \times O_i$ (at least T percent of the pixels in region C_i in the segmentation S are marked as pixels in region C_j in the GT image), and*
2. *$O_{ij} \geq T \times O_j$ (at least T percent of the pixels in region C_j in the GT image are marked as pixels in region C_i in the segmentation S).*

It must be understood the term *region* as a segment from a segmentation S or ground truth image. Then, the final index will be defined as:

$$H = 1 - \frac{CD}{N_{GT}} \quad \textbf{(27)}$$

where CD is the number of correct detections between all C_j in the GT image and C_i in the segmentation S, and N_{GT} is the number of regions in the GT image. Noticed the necessity of a threshold T to define what is taken as a correct detection or not. In the experiments, this threshold was set to 0.8 according to the study done by Jiang, where this value showed to be enough sensitive to distortions.

5.2. Area Fit Index (AFI)

Based on the Area Fit Index (AFI), proposed by Lucieer (2004), a global metric is defined as follows:

$$AFI = \frac{1}{N_{GT}} \sum_{k=1}^{N_{GT}} \frac{A_k - A_{l.i._k}}{A_k} \quad \textbf{(28)}$$

where $A_{k.}$ is the area, in pixels, of a reference segment C_k in the GT image and $A_{l.i._k}$ is the area, in pixels, of the segment, in the segmentation S, with the largest intersection with the reference region C_k. N_{GT} is the number of segments in the GT image and is used to normalize the expression. Noticed from eq. (28) that:

- A perfect overlap occurs when $AFI = 0$.
- The reference segment will be over-segmented if the overlap is less than 100% and AFI is greater than 0.0.
- The reference segment will be under-segmented if overlap is 100% and AFI is less than 0.0.
- In some situations, the overlap could be less than 100% and AFI could be less than 0.0. Then, the reference segment is over-

segmented; however, the segment with the largest intersection is larger than the reference.

In other words, the numerator in eq. (28) represents the pixels belonging to the GT and not to the segmentation S.

5.3. Shape Index (SI)

The Shape Index (SI) comes from landscape ecology and addresses the polygon's form or the shape conformity between the segmentation outcome and the reference segments. It is defined as the ratio between the area (A) and the perimeter (ρ) of a segment C_k:

$$SI_k = \frac{\rho_k}{4\sqrt{A_k}} \quad \textbf{(29)}$$

where ρ_k is the perimeter of the reference segment C_k and A_k is its area.

Given a segment C_j in the GT image, let's call C_i the segment, from the segmentation S, with the largest intersection with C_j. Then, the SI measure is defined as in (Neubert & Meinel, 2003):

$$SI = \frac{1}{N_{GT}} \sum_{j=1}^{N_{GT}} \left(\frac{\rho_j}{4\sqrt{A_j}} - \frac{\rho_i}{4\sqrt{A_i}} \right) \quad \textbf{(30)}$$

where N_{GT} is the number of segments in the GT image, ρ_i and ρ_j are the perimeters of the segments i and j, and A_i and A_j are their respective areas.

5.4. Rand Index (RI)

Rand Index (RI) is a clustering evaluation measure originally defined by Rand (1971). As image segmentation can be interpreted as a clustering of pixels into groups with certain similarity, this measure can be applied to measure the quality of the segmentation.

Let $I = \{p_1, \dots, p_N\}$ be the set of pixels of the original image and consider the set of all pairs of pixels $\mathcal{P} = \{(p_i, p_j) \in I \times I | i < j\}$. Given two partitions of I, C_i in the segmentation S and C_j in the GT image, then, $\mathcal{P}$ is divided into four different sets, depending on where a pair (p_i, p_j) of pixels falls (Pont-Tuset & Marques, 2013):

- $\mathcal{P}_{11}$: in the same segment both in C_i and C_j.
- $\mathcal{P}_{10}$: in the same segment in C_i but different in C_j.
- $\mathcal{P}_{01}$: in the same segment in $C_j$but different in C_i.
- $\mathcal{P}_{00}$: in different segments both in C_i and C_j.

Then, the RI will be defined as follows:

$$RI(C_i, C_j) = \frac{|\mathcal{P}_{00}| + |\mathcal{P}_{11}|}{|\mathcal{P}|} \quad \textbf{(31)}$$

where $|\ |$ is the cardinal operator, i.e. for an image with N pixels, $|\mathcal{P}|$ will be equal to $N(N-1)/2$. In other words, RI counts the pairs of pixels that have coherent labels for the two partitions being compared, with respect to the number of possible pairs of pixels. A good segmentation will have a high value of RI. For that reason, and for implementation purposes, we took in our experiments $(1 - RI)$ as the actual metric. The segments C_i and C_j, from a segmentation S and the GT image respectively, to be compared, are those that have the largest intersection between them.

5.5. Precision-Recall (F)

Let´s take as an example the Figure 12. There are two main segments delineated, one from GT image (green) and one from segmentation S with the largest intersection with it (red).

Then, there are four regions that are easily differentiated. The first one is the intersection between the two segments and is called True Positives (tp). It represents the pixels from the GT segment that are also in the segment from segmentation S. The second one is called False Positives (fp) and represents the pixels from the segment of the segmentation S that do not belong to the GT

segment. Later, the third region is called False Negatives (fn) and represents the pixels from the GT segment that do not belong to the segment of the segmentation S. Finally, the remainder region outside the union of GT and S segments is called True Negatives (tn). This region is not delineated in Figure 12. Based on the aforementioned definitions, Precision (P) and Recall (R) are defined by the eq. (32).

$$P = \frac{tp}{tp + fp} \quad R = \frac{tp}{tp + fn} \tag{32}$$

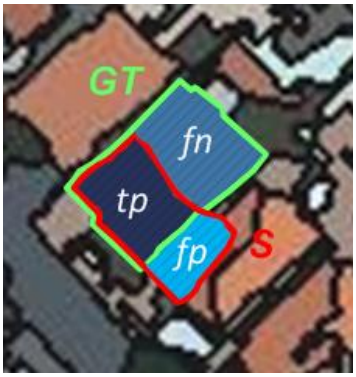


Figure 12: Precision and Recall definition. A segment of the segmentation outcome appears in red and its respective reference in green. There are four main regions, tp represents the true positives, fn the false negatives, fp the false positives and all pixels outside the union of the green and red regions are considered true negatives (tn).

Actually, the F measure was used for segmentation evaluation because it captures the trade-off as the weighted harmonic mean of P and R (Van Rijsbergen, 1979):

$$F = \frac{1}{\alpha P^{-1} + (1 - \alpha) R^{-1}} \tag{33}$$

$$F = \frac{PR}{\alpha R + (1 - \alpha) P} \tag{34}$$

where α defines a relative cost between P and R to focus the attention at a specific measure. In the experiments, this parameter was set to 0.5 as done by Martin (2003). A good segmentation will have a high value of F, for that reason, $(1 - F)$

was used as the actual metric. Similar to the aforementioned metrics, the segments from the segmentation S and the GT image for the calculation of the metric, are those whose have the largest intersection between them. Noticed the following, a low value of this metric will not guarantee a good segmentation outcome. As it measures a trade-off between Precision and Recall, it is necessary to consider these values too and made a decision based on what is more important, a high Precision or a high Recall.

5.6. Precision and Recall plots

As part of the calculation of F measure, the values of Precision (P) and Recall (R) are important too for further analysis. Let's use the same terminology introduced in the previous section for true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) (cf. Van Rijsbergen (1979)).

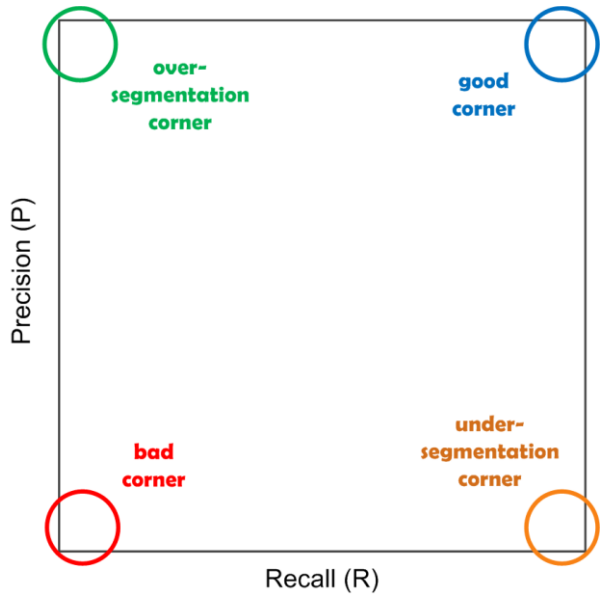
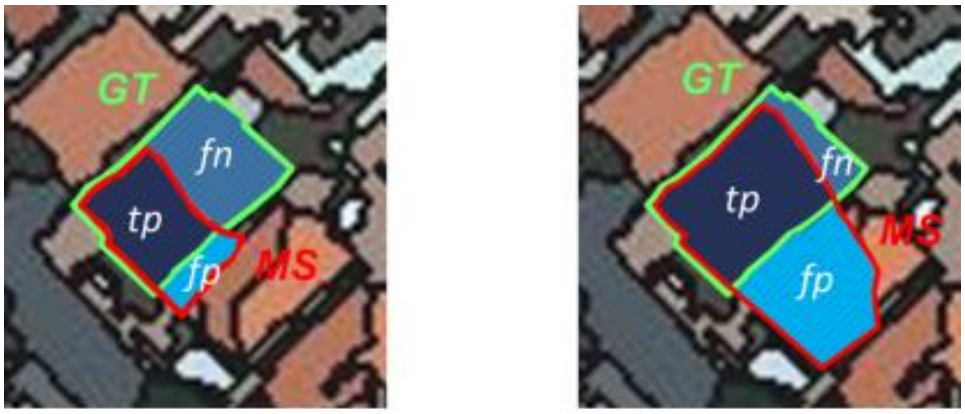


Figure 13: Interpretation of areas (corners) in a Precision-Recall plot.

There are four corners to take into consideration in this plot area: top-left, top-right, bottom-left and bottom-right (as shown in Figure 13). The top-right and bottom-left corners represent the "good corner" and the "bad corner" respectively. It is called "good corner" because the best results lie here and "bad corner" for the opposite reason. The top-left corner is called the "over-segmentation corner". In this region of the plot, the number of false positives (fn) are greater than the false

negatives (fp), which implies an over-segmentation. The bottom right corner is called the "under-segmentation corner". On this corner, the number of false positives (fp) are greater than the number of false negatives (fn), corresponding to under-segmented results (see Figure 14). Generally, over-segmentation is preferred over under-segmentation because it is easier to join regions to form a large segment than to split a region into smaller segments for further analysis.



$fn \gg fp$ $fp \gg fn$

Figure 14: Explanation of over-segmentation and under-segmentation corners of a Precision-Recall plot. The result of the segmentation (S) and its ground truth (GT) are represented by the red and green contours respectively. In the first case (left), there are more fn than fp, which favor the over-segmentation. In the second case (right), the number of fp is greater than fn, which is related to an under-segmentation.

5.7. Segmentation Covering ($\mathcal{C}$)

The concept of *overlap* between two segments, C_i in a segmentation S and C_j in a GT image, is defined in Arbeláez et al. (2011) as follows:

$$\mathcal{O}(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \quad \textbf{(35)}$$

Then, the covering of a segment C_j by a segment C_i is:

$$\mathcal{C}(C_i \to C_j) = \frac{1}{\sum N_{GT}} \sum_{C_t \in GT} |C_t| \cdot \max_{C_i \in S} \mathcal{O}(C_i, C_t) \quad \textbf{(36)}$$

where $\sum N_{GT}$ is the total number of pixels in the GT image. Noticed that this metric only considers the maximum overlap between each region $C_t \in GT$ and a region C_i in the segmentation S. As a good segmentation will have a high value of $\mathcal{C}$, $(1 - \mathcal{C})$ was used in the experiments for implementation purposes.

5.8. Reference Bounded Segments Booster ($RBSB$)

Proposed by Feitosa et al. (2006), the Reference Bounded Segments Booster ($RBSB$) metric corresponds to the ratio between the non-intersection area and the reference's area.

Let's take the aforementioned definitions from Figure 12. Then, $RBSB$ is defined as follows:

$$RBSB = \frac{1}{N_{GT}} \sum_{t=1}^{N_{GT}} \frac{fn_t + fp_t}{fn_t + tp_t} \quad \textbf{(37)}$$

where t represents a segment from GT and N_{GT} is the number of regions in the GT image. $RBSB$ will be 0.0 when the segmentation S fits perfectly the GT image; otherwise, it will be more than 0.0. Similar to the aforementioned metrics, the segments from the segmentation S and the GT image for the calculation of the metric are those that have the largest intersection between them.

Finally, Table 1 presents a summary of the aforementioned metrics.

Metric / Reference	Equation	Description
Hoover Index (H) (Hoover et al., 1996)	$H = 1 - \frac{CD}{N_{GT}}$ where CD are correct detections and N the total number of segments in the GT.	measures the number of correct detections based on the percentage of overlapping between segmentation and reference.
Area-Fit-Index (AFI) (Lucieer, 2004)	$AFI = \frac{1}{N_{GT}} \sum_{k=1}^{N_{GT}} \frac{A_k - A_{l.i._k}}{A_k}$ where A is the area and $l.i.$ represents the largest intersection.	addresses over-/under-segmentation by analyzing the overlapping area between segmentation and reference.
Shape Index (SI) (Neubert & Meinel, 2003)	$SI = \frac{1}{N_{GT}} \sum_{j=1}^{N_{GT}} \left(\frac{\rho_j}{4\sqrt{A_j}} - \frac{\rho_i}{4\sqrt{A_i}} \right)$ where A is the area and P is the perimeter.	addresses the shape conformity between segmentation and reference regions.
Rand Index (RI) (Rand, 1971)	$RI = \frac{\#\ of\ agreements}{\#\ total\ of\ pairs\ of\ pixels}$	measures the ratio between pair of pixels that were correctly classified and the total pairs of pixels.
Precision-Recall (F) (Van Rijsbergen, 1979)	$P = \frac{tp}{tp + fp} \quad R = \frac{tp}{tp + fn}$ $F = \frac{PR}{\alpha R + (1 - \alpha)P}$ where tp, fp and fn are true positives, false positives and false negatives respectively.	measures the trade-off between Precision and Recall considering segmentation as a classification process.
Segmentation Covering ($\mathcal{C}$) (Arbeláez et al., 2011)	$\mathcal{O}(C_i, C_j) = \frac{\lvert C_i \cap C_j \rvert}{\lvert C_i \cup C_j \rvert}$ $\mathcal{C}(C_i \to C_j) = \frac{1}{N} \sum_{C_t \in GT} \lvert C_t \rvert \cdot \max_{C_i \in MS} \mathcal{O}(C_i, C_t)$ where $\mathcal{O}$ is the overlapping between two segments.	measures the number of pixels of the intersection of two segments.
Reference Bounded Segments Booster ($RBSB$) (Feitosa et al., 2006)	$RBSB = \frac{1}{N_{GT}} \sum_{t=1}^{N} \frac{fn_t + fp_t}{fn_t + tp_t}$ where fn and fp are false negatives and false positives respectively.	measures the ratio between the number of pixels outside the intersection of two segments with the area of the reference.

Table 1: Metrics for segmentation evaluation selected for this study.