Revisiting the Coco Panoptic Metric to Enable Visual and Qualitative Analysis of Historical Map Instance Segmentation

Joseph Chazalon and Edwin Carlinet

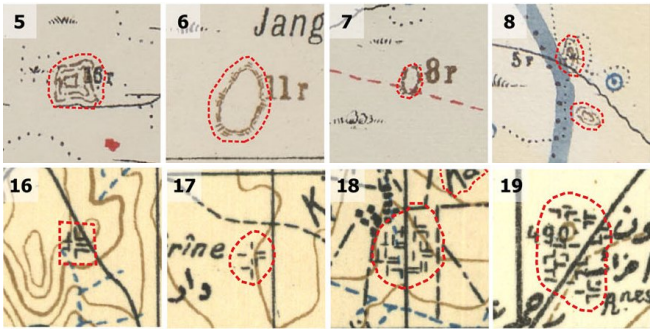
EPITA Research & Development Laboratory (LRDE)
14-16 rue Voltaire, 94270, Le Kremlin-Bicêtre, FRANCE
{joseph.chazalon,edwin.carlinet}@lrde.epita.fr

Abstract. Segmentation is an important task. It is so important that there exist tens of metrics trying to score and rank segmentation systems. It is so important that each topic has its own metric because their problem is too specific. Does it? What are the fundamental differences with the ZoneMap metric used for page segmentation, the COCO Panoptic metric used in computer vision and metrics used to rank hierarchical segmentations? In this paper, while assessing segmentation accuracy for historical maps, we explain, compare and demystify some the most used segmentation evaluation protocols. In particular, we focus on an alternative view of the COCO Panoptic metric as a classification evaluation; we show its soundness and propose extensions with more "shape-oriented" metrics. Beyond a quantitative metric, this paper aims also at providing qualitative measures through *precision-recall maps* that enable visualizing the success and the failures of a segmentation method.

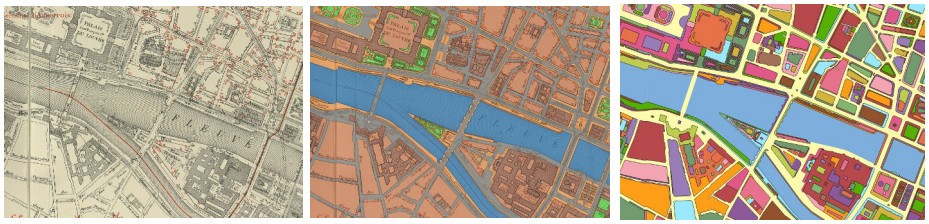
Keywords: Evaluation · Historical Map · Panoptic segmentation

1 Introduction

The massive digitization of historical data by the national institutions have lead to a huge volume of data to analyze. These data are a source of fundamental knowledge for historians and archaeological research. In this paper, we focus on the processing of a specific type of documents: the historical maps. Historical maps are very rich resources that can identify archaeological sites [15], can track the social evolution of places [6], can help study the urban mobility by analyzing historical changes performed on a road system [14]. To exploit effectively these data in geographical applications, one need to extract the interesting features from the map automatically. The automation is required as the time needed to extract them by hand is not tractable. The process, *i.e.* the image processing pipeline required to exploit the information from the maps is application dependent but most of them require the following items: *1.* spotting the elements of interest in the map such as texts, legends, geometric patterns; *2.* segmentation and classification of each of these elements; *3.* georeferencing the extracted elements. The automation of historical map processing has been studied for a long time. For instance, in [1], the authors apply well-known morphological approaches for



(a) Sparse Instance segmentation



(b) Dense Semantic segmentation (c) Panoptic segmentation

Fig. 1: Instance segmentation vs Panoptic segmentation. In (a), the instance segmentation of the mounds (1st row) of the Sol map and the ruins (2nd row) on the French Levant maps (images from [10]). Each object is described by its region (usually a binary mask) and its class. In (b), the semantic segmentation of a 1925 urban topographic map: building blocks (orange), parks (green), roads (gray) and rivers (blue); each pixel is associated to a class label. In (c) the panoptic segmentation of the same 1925 urban topographic map. A pixel-wise annotation that combines object instances (in false colors) and classes [5].

segmenting maps. With the advances of the deep learning techniques over the last decade, these techniques have been further developed and are getting more and more efficient. In [5], the authors combine the morphological methods from [1] with deep edge detectors and [10] rely on the U-Net deep-network to spot and segment features on the map.

In this context, we are interested in **evaluating and comparing different approaches for historical map processing**.

Segmentation is itself very content and application specific, even when considering just maps. Indeed, one may expect from a segmentation to have strong guaranties such as being a total partition of the space or just delimiting interesting areas. Other may require the segmentation to have thin separators between regions... In fig. 1 (a) and (b), we show two classes of segmentation tasks related to historical map processing: object detection (or instance segmentation) and semantic segmentation. Recently, it has been suggested in [11] that both types of segmentation should be joint in a unique framework and should be evaluated with the same metrics. The work of [11] has resulted in the creation of a new

segmentation task, namely the *panoptic segmentation*, that encompasses both the semantic and the instance segmentation. Moreover, they have proposed a parameter-less metric, the *panoptic quality*, that renders the quality of the segmentation and is becoming a standard for the evaluation.

In this paper, we aim at providing a set of tools that enables to visualize, compare and score a (panoptic) segmentation in the context of a challenge [4]. The task consists in providing an instance segmentation of the building blocks of the maps. In others words, there are several classes in the groundtruth but only the *block* class is of interest. Each pixel must be assigned to an instance id (of class *block*) or to the *void* class (with a unique instance). The blocks are **closed** 4-connected shapes and should be separated by 8-connected boundaries (pixels of class *void*). The boundary requirement is task specific as it enables vectorizing and georeferencing the blocks afterward. The set of *blocks* and *void* segments forms a partition of the image.

The contributions of this paper lies in three points that supply simple and fast tools for visualizing, evaluating and comparing segmentation methods. First, we propose another point of view of the COCO Panoptic metric where it has solid foundations in terms of prediction theory and is closely related to the well-known precision, recall and Area Under the Curve. It follows a quantitative evaluation of the segmentation systems and a ranking of the systems with respect to a shape pairing score. Second, we provide a meaningful way to visualize the segmentation results in a qualitative way. We introduce the *precision* and *recall* maps based the metrics previously defined, that highlights the locations where a system succeeds and where it fails. Third, we provide an insight about how to extend these metrics by studying the algebraic requirements of the metrics in bipartite graph formalism. It follows that the panoptic metric can be easily customized with another metric that may make more sense in some specific domains like document processing.

The paper is organized as follows. First, we have a short review of the evaluation protocols for image segmentation and their differences in section 2 with an emphasis on the COCO Panoptic score. In section 3, we demystify, explain and extend the COCO metric within a bipartite graph framework and propose a qualitative evaluation of a segmentation through the *prediction* and *recall* maps. In section 4, we compare our approach with other usual segmentation scoring and highlight their pitfalls. Last, we conclude in section 5.

2 State of the art

Here we focus on assessing which existing approaches are suitable for evaluating the quality of dense instance segmentation (of map images). Classification (in the sense of semantic segmentation) is not our priority here, as it can be handled quite easily on top of an instance segmentation evaluation framework. Hence, the challenge consists in combining two indicators: a measure of detection performance and a measure of segmentation quality. Among the methods currently used for object detection, the accuracy of the segmentation usually is of little importance. An approximate location is usually sufficient to count a detection

as successful. Indeed, most of the approach reported in a recent survey [12] show that an IoU of 0.5 is often used, which is far from being acceptable from a segmentation point of view. Pure segmentation metrics, on the other hand, often focus on pixel classification and do not consider shapes as consistent objects. The ability to accurately delineate objects in an image therefore requires dedicated metrics which can be either *contour-based*, ie., measuring the quality of the detection of the boundary of each shape, or *region-based*, i.e., measuring the quality of the matching and coverage of each shape [2]. Contour-based approach often rely on an estimation of the fraction of pixels accurately detected, which does not incur any performance regarding actual shape detection because the removal of a single pixel may prevent a shape from being detected. Region-based approaches combine a matching stage between expected and detected shapes, followed by the computation of a segmentation quality.

Several metrics have been proposed, both for natural images and document images, with differences in mind. In both cases, the matching step is performed by measuring a spatial consistency using a coverage indicator, i.e., the intersection of two shapes normalized by the area of either one or the union of the shapes. For natural images, evaluation metrics used to assume that objects to detect were sparse and non-overlapping. As a result, segmentation quality used to be reported as a rough indication of the distribution of matching scores: average of maximum pair-wise IoUs [2] or average of IoUs over 0.5 to consider only one-to-one matches [7]. Such information is usually insufficient in the context of document processing, especially when it comes to manual annotation: a measure of error costs is necessary to assess whether a system will provide a gain over manual work, and a simple global accuracy measure is not sufficient.

Early document segmentation metrics [17] focused on classifying error cases to enable the identification of correct, missed, false alarms, over- and under-segmentation cases, but relied on many thresholds and did not provide a normalized scoring; it was possible to rank systems but not to know whether their performance was acceptable or not. Further work like DetEval [20] proposed an improved formulation with a normalized score leading to precision and recall indicators blending detection and segmentation measures, with the possibility to include over- and under-segmentation costs. However, this approach requires the calibration of several thresholds and setting a minimal segmentation quality is hard to tune. The ZoneMap [9] metric improved DetEval to enable the support of overlapping shapes thanks to a greedy matching strategy to identify matching shapes, but did not make the usage cost of a system easier to assess as the metric is not normalized.

The F_{ob} [16] also performs a greedy strategy while reporting normalized precision, recall and F-measures while detecting under- and over-segmentation, but depends on two thresholds which prevent an easy calibration. The COCO Panoptic metric [11], finally, combines most of the features we are looking for: a single threshold on the minimum IoU to consider shapes as matching, a measure based on a F-score combining detection performance and segmentation accuracy, and can be evaluated class-wise if needed. This metric is both a solid and simple foundation to build our evaluation protocol on.

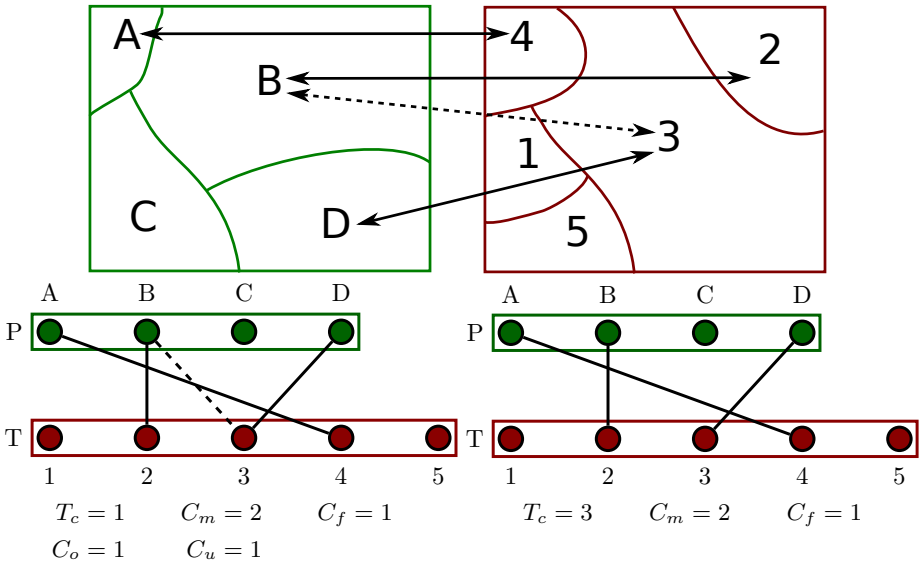


Fig. 2: Bipartite matching graph

3 Extending and visualizing the COCO metrics

3.1 Metric on pairings and bipartite graph

While aiming at providing a measure of the quality of two segmentations, our measure relies on a metric on bipartite graphs. Let $G = (P, T, E)$ denote a bipartite graph of a pairing relation between partitions P (the *prediction*) and T (the *target*). Each node of P (resp. T) actually represents a component of the predicted (resp. ground truth) segmentation. This graph is said to be a *matching* when no edge share the same endpoints (in section 3.3, we provide a way to build such a graph from two segmentations). In other words, a *matching* is a 1-1 relation between components of P and T. Note that this relation does not need to be total (or serial), some nodes of P or T may not have any match in the other set. Figure 2 shows an example of such a bipartite graph and a matching. In [17,9], the authors deduce several metrics from these graphs:

Number of matched components (T_c) The number of nodes (in P or T) that have a one-to-one match.

Number of missed components (type-I C_m) is the number of nodes in T that do not match any component in P, i.e., the number of components of the groundtruth that our segmentation has not been able to recover.

Number of false alarms (type-II C_f) is the number of nodes in P that do not match any component in T, i.e., the number of components in our segmentation that are false detection.

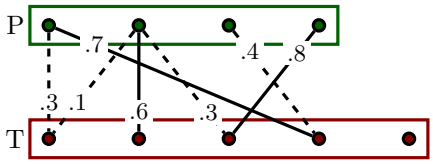


Fig. 3: Bipartite graph G with edges weighted by a matching score. Plain edges have weights above $\alpha = 0.5$ and form a one-to-one relation. Dashed edges have scores below α and exhibit a many-to-many relation.

Number of over-segmented components - Splits (C_o) is the number of nodes in T that match more than one component in P, *i.e.,* involved in a many-to-one association. We note T_{split} the subset of T involved in this relation.

Number of under-segmented components - Merges (C_u) is the number of nodes in P that match more than one component in T, *i.e.,* involved in one-to-many association. We note P_{merge} the subset of P involved in this relation.

It is worth noticing that counting the numbers of over/under-segmented components does not make sense on *matchings* as they are non-null on non-one-to-one relations only. **The type of relation actually characterizes of the features needed to assess the quality of the pairing**. If the relation is always one-to-one, an evaluation protocol would rely on T_c, C_m, C_f only. It does not mean that over-segmentations or under-segmentations are not possible in protocols handling only one-to-one relation but rather than they are converted to C_m or C_f errors in these frameworks. Therefore, *matchings* can be likened to a binary classification where matched components are *true positives*, missed components are *false negatives* and false alarms are *false positives*. It follows the definitions of the *recall* and the *precision* as:

$$precision = \frac{T_c}{T_c + C_f} = \frac{T_c}{|P|} \tag{1}$$

$$recall = \frac{T_c}{T_c + C_m} = \frac{T_c}{|T|} \tag{2}$$

$$F\text{-}score = \frac{T_c}{T_c + \frac{1}{2}(C_m + C_f)} = \frac{2.T_c}{|P| + |T|} \tag{3}$$

In the context of a segmentation evaluation, the *recall* stands for the ability to recover the components of the ground truth, while the *precision* stands for the ability to match the components of our segmentation with those of the ground truth.

3.2 Metrics on weighted pairings

We are now focusing on the case where edges are weighted by a *matching score* (the higher, the better). These scores are typically base on shape matching met-

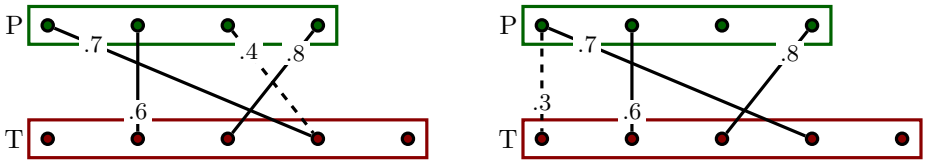


Fig. 4: Precision $G_{\text{precision}}$ (left) and recall G_{recall} (right) sub-graphs of G.

rics like the DICE, the IoU, the Hausdorff distance.... We note $w(X, Y)$ the similarity score between any two components X and Y normalized in $[0, 1]$.

Moreover, the similarity score must have the following property:

$$\exists\, \alpha \in [0, 1] \text{ s.t. } \{(X, Y) \mid w(X, Y) > \alpha\} \text{ is a bijection.} \tag{4}$$

In other words, there exists a threshold α for which the subset of edges whose weights are above α forms a one-to-one relation in $P \times T$. It follows that by considering any subgraph of G formed by edges with weights greater than $\alpha < t \leq 1$, we have a one-to-one relation for which we can compute the *precision* and *recall* with eq. (3).

If the scoring function does not feature the property eq. (4), it is still possible to extract a bipartite graph using a maximum weighted bipartite matching algorithm as in [19] or any greedy algorithm [16].

Qualitative assessment: the recall and precision maps. We propose the *precision* and *recall* maps that provide a *meaningful* way to *visualize and summarize* **locally** the performance of the segmentation.

The *precision* map reports how well a component of the *predicted* segmentation matches with the *ground truth*. It is built upon the subgraph of G with the best incident edge of each node of the prediction. Then, the pixels of the components of the prediction partition are rendered with the corresponding incident edge value. On the other hand, the *recall* map reports how well the ground truth has been match by the prediction. It is built upon the subgraph of G with the best incident edge of each node of the ground truth. The selected edge weights are then rendered pixel-wise just as before.

$$G_{\text{precision}} = (P, T, E_p) \text{ with } E_p = \{(X, Y), X \in P, Y = \arg\max_{Y} w(X, Y)\}$$
$$G_{\text{recall}} = (P, T, E_r) \text{ with } E_r = \{(X, Y), Y \in T, X = \arg\max_{X} w(X, Y)\}$$

The precision and recall graphs are depicted in fig. 4 while precision and recall maps deduced from them are illustrated on fig. 5. The interpretation of the color of a component depends on the "section" it is located (green when $> \alpha$, red when $< \alpha$). The green section reports the minimum score that a pairing must have so that a component is counted as a *matched component* (T_c). Below this score, the two components of the pairing would be considered as a *missed component* ($T_m+ = 1$) and a *false alarm* ($T_f+ = 1$). Components in the red section have to

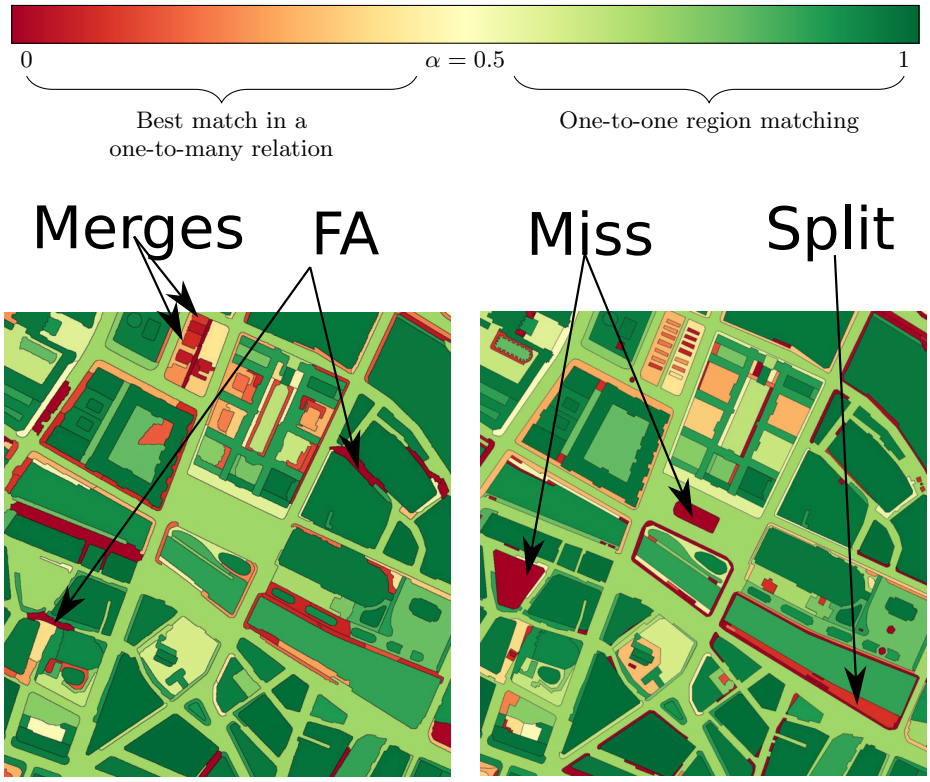


Fig. 5: Precision (left) - recall (right) maps.

be interpreted quite differently. Those components will never be part of *match* as they are included in a one-to-many relation in G. Therefore, they are always counted as *missed components* (if a red region the recall map) and as *false alarms* (if a red region of the precision map).

Quantitative assessment: precision, recall and f-measures curves on matchings. When assessing the segmentation quality, most protocols (especially those used in detection in computer vision) start with defining a minimum overlap score to define the *matched components*. It makes sense to have such threshold as a mean to dismiss regions that are not good enough to be "usable" (*usable* meaning is application dependent). This is equivalent to taking the pairing graph G, removing edges below a given score, getting a one-to-one relation, and computing the precision/recall/F-measure as in section 3.1. Afterward, edges values are not considered anymore.

Instead of relying on an arguably given overlap threshold and forgetting the edges weights, one can compute the precision/recall/F-score as a function of the edges weights. Indeed, we have seen that for any threshold t, $\alpha < t \leq 1$, the sub-graph is a one-to-one relation. When t increases, the number of matched

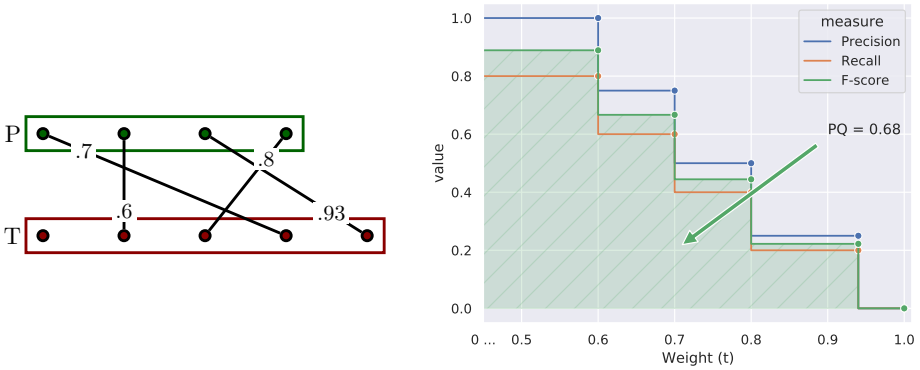


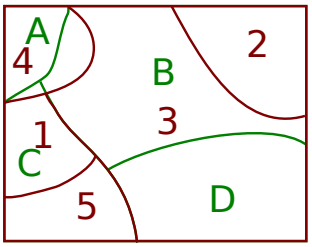
Fig. 6: Precision/recall/F-score curves from a one-to-one weighted pairing and the Panoptic quality got as the AUC of the F-score.

components $T_c(t)$ decreases, and the number of missed components $C_m(t)$ and false alarms $C_f(t)$ increases. It allows to plot the precision, recall and f-score curves as a function of the matching score threshold t as shown on fig. 6.

Quantitative assessment: the Panoptic Quality score. While providing a visual insight of the performance of a segmentation method when being more and more strict on the segmentation quality, the previous curves does not allow comparing and rank a set of methods. We need a single, simple and informative metric for this purpose. When applied on predictors, the Area Under the Curve (AUC) of the PR (or the ROC) curves is a widely used metric that summarizes how well a predictor perform for the whole space of prediction thresholds [8]. Here, the AUC of the F-score curve reflects how well the segmentation performs, both in-term of recall and precision, without committing to a particular threshold on the matching quality.

$$\mathrm{NPQ} = (1-\alpha)^{-1} \int_{t=\alpha}^{1} \textit{f-score}(t) \tag{5}$$

The *Normalized Panoptic Quality* depicted in eq. (5) is named after the *Panoptic Quality (PQ)* metric described in [11]. It is an unacknowledged rewriting of the Area under the F-score Curve as shown with eq. (8) (with $\alpha = 0$). Interpreted this way, the *PQ* metric matches the well-established practices in prediction evaluation, more specifically, averaging the performance over all the thresholds. Also, splitting the formula into the terms *segmentation quality* and *recognition quality* (as in [11]) has a straightforward interpretation with this formalism. The *recognition quality* is related to the initial number of edges in the graph (one-to-one matches) and corresponds to the f-score of the most permissive segmentation. On the other hand, the *segmentation quality* integrates the weights of the edges and renders the loss of matched regions while being less and less permissive with the region matching quality.



$X \cap Y$	1	2	3	4	5	$\lvert Y \rvert$
A				11		11
B		15	31	12		58
C	20			3	22	45
D			29			29
$\lvert X \rvert$	20	15	40	36	22	

Fig. 7: Fast computation of pairing scores based on the 2D-histogram of the image label pairs *(prediction, ground truth)*. Left: label-pair maps, right: 2d-histogram with marginal sums.

$$\mathrm{PQ} = \int_{t=0}^{1} \textit{f-score}(t) \tag{6}$$

$$= \textit{f-score}(\alpha) . \int_{0}^{1} \frac{T_c(t)}{T_c(\alpha)} \tag{7}$$

$$= \underbrace{\textit{f-score}(\alpha)}_{\text{recognition quality}} \cdot \underbrace{\sum_{(X,Y)\in E} \frac{w(X,Y)}{T_c(\alpha)}}_{\text{segmentation quality}} \tag{8}$$

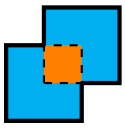
The *Normalized Panoptic Quality* ensures that the score is distributed in [0-1] and removes a constant offset of the PQ due to the parameter α which is minimal matching quality (the "usable" region quality threshold). This normalization enables comparing PQ scores with several "minimum matching quality".

3.3 Pairing strategies

In the previous section, we have defined the algebraic requirements of a weighted bipartite graph so that we can compute a parameter-less measure of the quality of a segmentation. In this section, we provide hints to build a graph out of a segmentation.

Symmetric pairing metrics We first need a shape matching metric that renders the matching of two regions. In Mathematical Morphology, distances between any set of points have been largely studied [18,1,3]. While the Jaccard distance (one minus the Jaccard index) and the Hausdorff distance are some well-known examples of such distances, more advanced metrics based on skeletons and median sets allow for a more relevant analysis of the shape families.

The *Intersection over Union (IoU)* (or *Jaccard index*) defined in eq. (10) has many advantages and is a *de facto* standard to compute similarity between two regions [2,7]. Whereas advanced morphological distances are computationally



$$\underline{\text{Cov}}_B(A) = \frac{|A \cap B|}{|\mathbf{B}|} \tag{9}$$

$$\underline{\text{Jaccard}}(A, B) = \frac{|A \cap B|}{|\mathbf{A} \cup \mathbf{B}|} \tag{10}$$

$$\underline{\text{Dice}}(A, B) = \frac{|A \cap B|}{|\mathbf{A}| + |\mathbf{B}|} \tag{11}$$

$$\underline{\text{Force}}(A, B) = \text{Cov}_A(B)^2 + \text{Cov}_B(A)^2 \tag{12}$$

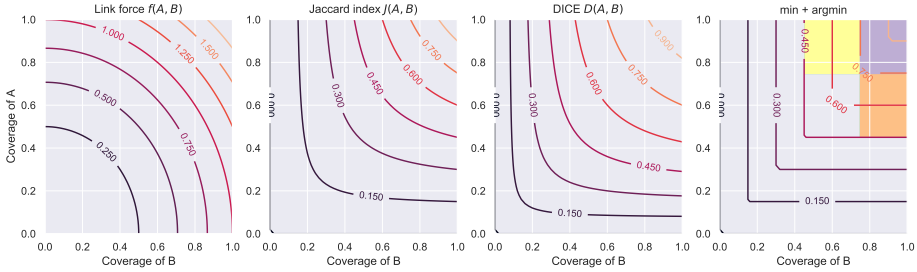


Fig. 8: Standard shape similarity for scoring edges. Link force [9], IoU (*Jaccard*), F1-score (*DICE*), and $\min(\text{Cov}_A, \text{Cov}_B)$ with edge categorization [16] (the color zone defines the category of the edge).

expensive, the IoU is fast to compute between any pair of regions of a segmentation as it is based the matrix of the intersections (see fig. 7). Second, it features the property required in section 3.2 as shown in [11]. Indeed, the authors have proven that *each ground-truth region can have at most one corresponding predicted region with IoU strictly greater than 0.5 and vice-verse*. Other popular pairing metrics includes the DICE coefficient D (or F1 score) which is closely related to the Jaccard index and so ensures a one-to-one pairing when the $D > \frac{2}{3}$. In the ZoneMap metric used for page segmentation[9], the regions' coverage are combined with a euclidean distance. All these metrics are related and can be expressed as a function of the coverage of the groundtruth's shapes by the prediction and the coverage of the prediction's shapes by the ground truth as shown fig. 8. The Zonemap pairing score differs as it allows good score as soon as one of the coverage score is high. Roughly speaking, *min*, the DICE and the IoU are "logical ands" while the Zonemap score is a "logical or". In [16], the score is also associated with a category depending on its "zone" — it will be further detailed in section 4.

The question when choosing a pairing function is "what is a match?" In our opinion, a *match* has a practical definition. Two objects/regions should be considered as *matching* if there is no need for human manual intervention to correct the result produced by a system. In the case of a segmentation, it means that the predicted region should perfectly fit with the ground truth and vice verse. As a consequence, the IoU of the DICE score should be favored.

(a) Precision. Left: IoU, right: $\mathrm{Cov}_Y(X)$ (b) Recall. Left: IoU, right: $\mathrm{Cov}_X(Y)$

Fig. 9: Comparison of the precision/recall values when edges are valued with IoU vs marginal coverage scores. Marginal coverages are clearly optimistic and score some of the under/over-segmented areas as correct.

Different pairing and edge weighting functions. Most evaluation framework use the same scoring function for pairing, *i.e.* building the association graph, and *evaluating*, *i.e.* getting a score out-of-the graph. A notable exception is the ZoneMap score [9], where the second phase of the ZoneMap calculation deals with the computation of an error combining a surface error and a classification error. With this graph formalism, nothing prevents using two different scoring functions to build and weight edges. In particular, it is quite reasonable to build the associations based on a fast-to-compute and sensible metric that exhibits the properties in eq. (4) (like the IoU) and keep only edges that provide a matching. On this new graph, we can either keep the original edge weight and compute the statistics described in section 3.1 — this yields the COCO Panoptic metric — or weight edges with another function more sensible for the application. Since this metric is only computed between regions which are one-to-one match, it does not matter if it is more computationally expensive. In the context of map segmentation and block partitioning, the 95% Hausdorff has shown to be closer to the human perception of distance between shapes as it penalizes non-regular boundaries. Yet, once the graph re-weighted (with normalized distances), the evaluation protocol described in section 3.1 becomes available.

4 Comparison with other metrics and limitations

Non-symmetric metrics for precision-recall In [16], the authors have an alternate definition of the precision, recall, F-measure based on partition coverage and refinement (eq. (9)). Explained in a bipartite graph framework, given two shapes X and Y from the predicted segmentation P and the groundtruth T, an edge is added to E if $\min(\mathrm{Cov}_X(Y), \mathrm{Cov}_Y(X)) > \gamma_1$ and $\max(\mathrm{Cov}_X(Y), \mathrm{Cov}_Y(X)) > \gamma_2$ (with $\gamma_1 < \gamma_2$) where γ_1 represents a *candidate* threshold and γ_2 a *match* threshold. These thresholds define the *match* category areas as shown in fig. 8. When both $\mathrm{Cov}_X(Y)$, $\mathrm{Cov}_Y(X)$ pass the match threshold, we have an *object match* (simply a *match* in our terminology). If only one of them passes the *match* threshold, we get a *partial* match that is either a *merge* or a *split* depending on which of $\mathrm{Cov}_X(Y)$ and $\mathrm{Cov}_Y(X)$ is higher. It follows definitions of the precision/recall based on the these classes. They have been rewritten to highlight the relations with the Panoptic metric.

$$precision[16] = \underbrace{precision(1)}_{\text{P match rate}} + \underbrace{\beta.\frac{C_u}{|P|}}_{\text{under-segmentation rate}} + \underbrace{\sum_{(x,y)\in E} \frac{\text{Cov}_Y(x)}{|P|}}_{\text{fragmentation quality}} \quad (13)$$

$$recall[16] = \underbrace{recall(2)}_{\text{T match rate}} + \underbrace{\beta.\frac{C_o}{|T|}}_{\text{over-segmentation rate}} + \underbrace{\sum_{(X,Y)\in E} \frac{\text{Cov}_X(Y)}{|T|}}_{\text{fragmentation quality}} \quad (14)$$

These metrics are actually over-scoring the quality of the segmentation. Indeed, *recall* and *precision* are related to measuring "correct" things, *i.e.*, *matches*. Here the precision is maximal whenever $T \sqsubseteq P$ where $\sqsubseteq$ denotes the partition refinement and the recall is maximal whenever $P \sqsubseteq T$. However, neither an over-segmentation, nor an under-segmentation are good as they may have no "correct" regions (regions that would need no human correction). This statement is motivated by the experiment in fig. 9 where the precision and recall maps are valued with the IoU, $\text{Cov}_Y(X)$ and $\text{Cov}_Y(X)$ and shows that some over/under-segmented regions appear (and are counted) as "correct".

Nevertheless, the metrics are sensible to measure the quality of an over-/under-segmentation. For instance, it may be used to measure the ability to recover a correct segmentation by only merging regions. The *recall* and *precision* from [16] should not be used for assessing a fully-automated system and might rather be named "Coarse Segmentation Score" and "Fine Segmentation Score". These metrics also make sense when evaluating hierarchies of segmentation to allow over- and under-segmented regions as soon as the boundaries match [13].

Comparison with DETVAL and ZoneMap document-oriented metrics. The DETVAL metric [20] shares many features with [16]. The pairing function between two regions depends on the marginal coverage $\text{Cov}_X(Y)$ and $\text{Cov}_Y(X)$ that must pass two detection quality thresholds. It also uses a fragmentation score in the precision/recall that allows to consider a one-to-many and many-to-one (splits and merges) as partial match. For document processing, it may make sense to consider such situation where it is "better have an over/under segmentation than nothing" since documents are structured hierarchically. In [20], precision/recall are defined as follows (again rewritten to highlight the relations with the Panoptic metric):

$$precision[20] = \underbrace{precision(1)}_{\text{P match rate}} + \underbrace{\sum_{X\in E_{merge}} \frac{fs(X)}{|P|}}_{\text{fragmentation quality}} \quad (15)$$

$$recall[20] = \underbrace{recall(2)}_{\text{T match rate}} + \underbrace{\sum_{Y\in E_{split}} \frac{fs(Y)}{|P|}}_{\text{fragmentation quality}} \quad (16)$$

where $fs(X) = (1 + \ln(degree(X)))^{-1}$ is the fragmentation score that decreases as the number of incident edges in X increases. At the end, the drawbacks of this metric are: 1. it considers *fragmentation* as an acceptable error, 2. it has two detection quality thresholds (vs one for the PQ), 3, it does not average the quality of the segmentation once a match has been accepted.

The ZoneMap[9] is generalization of the DETVAL protocol that supports zone overlapping. As a consequence, it takes account of *splits* and *merges* by decomposing partial overlaps in sub-zones that are scored and contribute to the final score. Contrary to DETVAL, the score penalizes all "imperfect" matches since the pixels that are not in the intersection participate in the error. ZoneMap also introduces a classification error that is of prime interest as part of a panoptic segmentation where instances are associated to classes. However, it has some drawbacks for our application. 1. There are (almost) no one-to-one association when computing the graph with all non-zero edges over two map partitions, we need to filter out edges with a detection quality threshold. 2. The error is based on the global surface and not the instance surface, *i.e.* imprecise small regions are not as important as imprecise big regions. 3. the score, not normalized between [0-1], is hard to interpret.

5 Limitations, conclusions and perspectives

The COCO Panoptic metric shines for the segmentation task: it is fast, simple and intelligible. We have shown that it is theoretically sound as it can be seen as classical prediction evaluation for which a strong background exists and is widely accepted. Also, expressed in a bipartite graph framework, it is extendable as we can separate the metric used for pairing regions (that needs strong properties and has to be fast to compute) and the one assessing the segmentation quality. As a matter of reproducible research, the metric implementation and a python package are available as a tool set on Github[1]. We have also highlighted the similarities and differences with some other metrics used in the document and natural image segmentation. Especially, considering an over-/under-segmentation as "half"-success is debatable for map segmentation, but may make sense for page segmentation. We believe that a generalization of the COCO Panoptic metric will lead to a unification of the evaluation segmentation protocol with just an application-dependent customization. Once the actual limitations are addressed, we will be able to quantify the differences between a unified-COCO Panoptic metric and task-specific metrics on various dataset. The following table summarizes the current limitations and the domain accessible once addressed:

Limitation	Application — Domain
1. Region Confidence Score	A. Evaluation of hierarchies [13]
2. Region overlapping	B. Page/Multi-layer map segmentation
3. Over/under-segmentation scoring	A + B

[1] https://github.com/icdar21-mapseg/icdar21-mapseg-eval

Acknowledgements
This work was partially funded by the French National Research Agency (ANR): Project SoDuCo, grant ANR-18-CE38-0013.

References

1. Angulo, J., Meyer, F.: Morphological exploration of shape spaces. In: International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing. pp. 226–237 (2009)
2. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 898–916 (2010)
3. Charpiat, G., Faugeras, O., Keriven, R., Maurel, P.: Distance-based shape statistics. In: IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. vol. 5, pp. V–V (2006)
4. Chazalon, J., Carlinet, E., Chen, Y., Perret, J., Duménieu, B., Mallet, C., Géraud, T., Nguyen, V., Nguyen, N., Baloun, J., Lenc, L., Král, P.: Icdar 2021 competition on historical map segmentation. In: IEEE International Conference on Document Analysis and Recognition (Sep 2021), https://icdar21-mapseg.github.io/, to appear
5. Chen, Y., Carlinet, E., Chazalon, J., Mallet, C., Duménieu, B., Perret, J.: Combining deep learning and mathematical morphology for historical map segmentation. DGMM (2021), (to appear)
6. Dietzel, C., Herold, M., Hemphill, J.J., Clarke, K.C.: Spatio-temporal dynamics in california's central valley: Empirical links to urban theory. International Journal of Geographical Information Science **19**(2), 175–195 (2005)
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
8. Flach, P.A., Hernández-Orallo, J., Ramirez, C.F.: A coherent interpretation of auc as a measure of aggregated classification performance. In: ICML (2011)
9. Galibert, O., Kahn, J., Oparin, I.: The zonemap metric for page segmentation and area classification in scanned documents. In: IEEE International Conference on Image Processing. pp. 2594–2598 (2014)
10. Garcia-Molsosa, A., Orengo, H.A., Lawrence, D., Philip, G., Hopper, K., Petrie, C.A.: Potential of deep learning segmentation for the extraction of archaeological features from historical map series. Archaeological Prospection (2021)
11. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9404–9413 (2019)
12. Padilla, R., Passos, W.L., Dias, T.L.B., Netto, S.L., da Silva, E.A.B.: A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics **10**(3) (2021)
13. Perret, B., Cousty, J., Guimaraes, S.J.F., Maia, D.S.: Evaluation of Hierarchical Watersheds. IEEE Transactions on Image Processing **27**(4), 1676–1688 (Apr 2018)
14. Perret, J., Gribaudi, M., Barthelemy, M.: Roads and cities of 18th century france. Scientific data **2**(1), 1–7 (2015)
15. Petrie, C.A., Orengo, H.A., Green, A.S., Walker, J.R., et al.: Mapping archaeology while mapping an empire: Using historical maps to reconstruct ancient settlement landscapes in modern india and pakistan. Geosciences **9**(1), 11 (2019)

16. Pont-Tuset, J., Marques, F.: Supervised evaluation of image segmentation and object proposal techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(7), 1465–1478 (2015)
17. Shafait, F., Keysers, D., Breuel, T.: Performance evaluation and benchmarking of six-page segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(6), 941–954 (2008)
18. Vidal, J., Crespo, J.: Sets matching in binary images using mathematical morphology. In: 2008 International Conference of the Chilean Computer Science Society. pp. 110–115 (2008)
19. West, D.B., et al.: Introduction to graph theory, vol. 2 (2001)
20. Wolf, C., Jolion, J.M.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. International Journal on Document Analysis and Recognition **8**(4), 280–296 (2006)