THE ZONEMAP METRIC FOR PAGE SEGMENTATION AND AREA CLASSIFICATION IN SCANNED DOCUMENTS

Olivier Galibert, Juliette Kahn and Ilya Oparin

LNE, Laboratiore national de métrologie et d'essais
National Metrology and Testing Laboratory
Trappes, France
Email: firstname.secondname@lne.fr

ABSTRACT

A novel metric for the detection and classification of different areas in scanned documents is presented in this paper. This metric, ZoneMap, aims at evaluating both page segmentation and zone classification. Moreover, for the segmentation subtask, it handles the superposition of overlapping zones. These characteristics allow to evaluate systems in a coherent way using a single ZoneMap metric. Weights assigned to different parameters add flexibility and allow for fine-tuning of the metric in order to reflect the specificity of a particular applicative context. ZoneMap was experimented in the Maurdor evaluation campaigns where it is used as a primary metric for page segmentation and area classification. Evaluation results show that ZoneMap provides additional ways to assess system performance and analyze the results. ZoneMap is implemented in a publicly available LNE *maurdor-eval* evaluation toolkit that is distributed under the GPL license.

Index Terms— Page segmentation, area classification, metric, evaluation

1. INTRODUCTION

Scanned documents processing is an important issue for information retrieval. The goal of the Maurdor evaluation campaigns is to evaluate the ability of existing technologies to extract relevant information in scanned documents.

Maurdor is based on a processing chain in which five separate tasks are evaluated. Each task corresponds to a particular function and contributes to a complete processing of a scanned document [1, 2]. This paper is concerned with the evaluation of the first task, which is page segmentation and zone (area, region) classification. The aim of this task is to evaluate the ability of existing systems to identify various areas in a document (table, text, image...) and specify their position, thus partitioning document images into distinct and homogeneous semantic areas.

One distinct characteristic of the Maurdor challenge is that different areas may overlap. For instance, a table area may define a set of other areas, including text and graph areas (logo, signature, etc.) that may appear both next to and in the background of the text. Another specificity is that both segmentation and classification are evaluated jointly.

We propose a new metric called ZoneMap to evaluate this task. As compared to existing metrics, ZoneMap makes it possible to:

- Take into account overlapping zones;
- Integrate evaluation of segmentation and classification in a single metric;
- Assign weights to errors of different types, if needed to evaluate a particular task;
- Obtain consistent metric behavior in the whole range of values.

These properties make ZoneMap a useful and flexible tool for jointly evaluating segmentation and classification of zones for scanned documents in a real-world scenarios.

2. RELATED WORKS

The task of page segmentation has existed for several decades and a number of metrics and evaluation schemes were proposed [3, 4, 5, 6]. However, they were not sufficient for evaluating the task of document segmentation and classification in Maurdor campaigns. For example, the metric implemented in the DetEval toolkit does not handle zone superposition [7]. The PSET toolkit [8] is designed to evaluate errors for text zones only.

The so-called Jaccard index takes into account the per-class surface but not the actual decomposition in zones. Its calculation is based on the area assigned to a class in the reference R and the area assigned to this class in the hypothesis H. For each zone class i the Jaccard index J_i is:

$$J_i = \frac{|H_i \cap R_i|}{|H_i \cup R_i|} \tag{1}$$

The document score J_{doc} is defined as

$$J_{doc} = \frac{\sum_{i=0}^{N}(|H_i \cup R_i|J_i)}{\sum_{i=0}^{N}|H_i \cup R_i|} = \frac{\sum_{i=0}^{N}|H_i \cap R_i|}{\sum_{i=0}^{N}|H_i \cup R_i|} \tag{2}$$

A metric based on the Jaccard index is used in ICDAR competitions [9]. This ICDAR metric [10] was compared to

The Maurdor evaluation campaigns (www.maurdor-campaigns.org) are part of the Maurdor project, managed by Cassidian and funded by DGA. The authors are thankful to colleagues from A2iA, ELDA, IRISA and LITIS for their valuable input during long discussions concerned with the way ZoneMap may be refined and improved.

ZoneMap on four different databases in [11]. It was found that for the task of handwritten text line detection ZoneMap and ICDAR are correlated but ZoneMap provides greater detail on error types. At the same time ZoneMap was shown to have an higher correlation with the recognition error rate.

3. ZONEMAP

ZoneMap is the extension and generalization of the metrics proposed in the DetEval [8] evaluation tools that enables taking into account superposition of zones as it can appear, for example, in tables or crossing-outs.

3.1. Mapping

The first phase of calculation of ZoneMap is called mapping. Zones in the document are grouped according to four possible configurations (*Match*, *Miss*, *False Alarm* and *Split*). The groups are constructed in a hierarchical way based on the coverage rate between hypothesis and reference zones.

For a zone H in the hypothesis and a zone R in the reference, the coverage rate of H by R (and, symmetrically, of R by H) is defined as:

$$C_{H,R} = \frac{Area(H \cap R)}{AreaH} \tag{3}$$

where $Area$ corresponds to the number of black pixels present in the zone in question. The larger coverage rate is, the stronger the link between the hypothesis and the reference zone. This link is characterized by the link force f defined as

$$f_{R,H} = C^2_{R,H} + C^2_{H,R} \tag{4}$$

The force of each link between each hypothesis and reference zone is calculated and non-zero links are kept.

Grouping hypothesis and reference zones is performed incrementally. First, each zone is considered as a separate group. The links between hypothesis and reference zones are sorted in descending order according to their forces and the corresponding zones are grouped incrementally. If adding a new zone to a group leads to the situation where a group contains multiple hypotheses and references zones, such a zone is not added to the group in question. Thus, one-to-one or one-to-many correspondences are allowed but not many-to-many. After all the links were examined, the different groups and their configurations are established. The process of mapping is illustrated by Figure 1.

The second phase of the ZoneMap calculation deals with calculating the error, E_i, attributed to each group i in order to calculate the global error rate, $E_{ZoneMap}$:

$$E_{ZoneMap} = \frac{\sum_{i=1}^{N} E_i}{Area(R)} \tag{5}$$

Each of E_i is a linear interpolation of two error rates

$$E = (1 - \alpha_c)E_S + \alpha_c E_C \tag{6}$$

Here E_S is the surface error that characterizes quality of zone segmentation and E_C is the zone classification error, $\alpha_c \in [0; 1]$ is the weight given to the classification error. The calculation of E_S and E_C depends on the group configuration. Possible configurations are presented in the sections below.

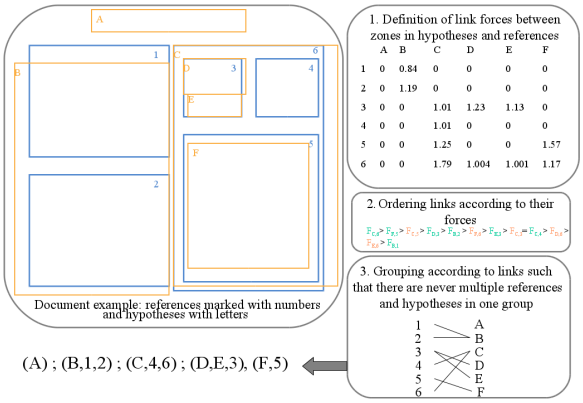

Fig. 1. Mapping process for calculating ZoneMap

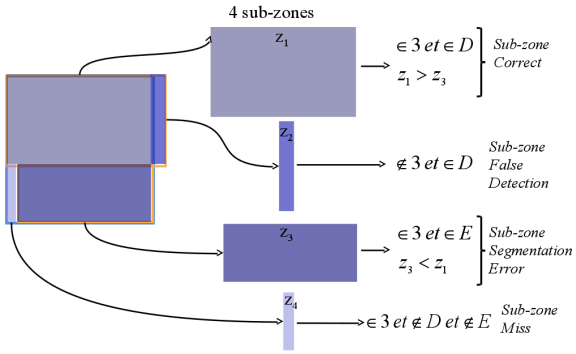

Fig. 2. Example of segmentation in sub-zones for the configuration *Split*, group (D,E,3)

3.2. Configurations

False Alarm. Configuration *False Alarm* (Group (A) in Figure 1) corresponds to the situation when one hypothesis zone h_i was detected but there is no corresponding reference zone. In this case $E_S = Area(h_i)$ and $E_C = E_S$.

Miss. Configuration *Miss* corresponds to the case when a reference zone r_i was not grouped with any hypothesis zone. In this case $E_S = Area(r_i)$ and $E_C = E_S$.

Match. In configuration *Match* a hypothesis zone h_i has one corresponding reference zone r_j. It is possible that two zones do not have a perfect overlap and thus the surface error is calculated as a size of non-overlapped regions:

$$E_S = Area(h_i \cup r_j - h_i \cap r_j) \tag{7}$$

The classification error corresponds to the difference between zone types in the hypothesis and the reference (between 0 and 1, 0 corresponding to the type match) multiplied by the common area plus the surface error:

$$E_C = d(t_H, t_R)Area(h_i \cap r_j) + E_S \tag{8}$$

Split. Configuration *Split* corresponds to the case when one reference zone R corresponds to at least two hypothesis zones $H1$ and $H2$ (group (D,E,3) in Figure 1). These zones are not bound to have a perfect overlap, as shown in Figure 2.

In order to calculate the error associated with a group under the *Split* configuration, the area of a group G is decomposed in a set $E_Z = \{Z_1 \dots Z_m\}$ of m sub-zones. Each sub-zone z respects the following rules:

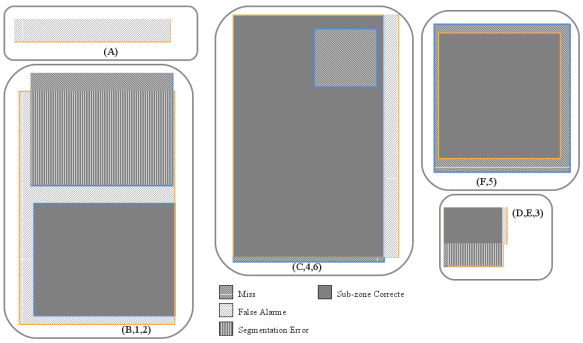

Fig. 3. Types of zones and sub-zones for each group.

- Two sub-zones cannot overlap, that is a pixel assigned to one sub-zone cannot also belong to another one.
- All the pixels of a group must be assigned to sub-zones.
- A sub-zone, in its turn, may follow one of the configurations: *Miss Error*, *False Detection Error*, *Segmentation Error* or *Correct*, as shown in Figure 2.

If a sub-zone covers only a reference zone but not a hypothesis one it is considered as a *Miss Error*. If a sub-zone covers one or several hypothesis zones without covering a reference zone it corresponds to a *False Detection Error*. In these cases the error $Er(z)$ is equal to the area of z: $Er_z = Area(z)$. A sub-zone covering one or several hypothesis zones and a reference zone can be either *Correct* or *Segmentation Error*.

If the sub-zone in question is the largest in the set of sub-zones E_Z which cover both the hypothesis and the reference, it is considered *Correct* and *Segmentation Error* otherwise. The error $Er(z)$ is defined as

$$Er(z) = (1 - \alpha_c)E_S(z) + \alpha_c E_C(z) \quad (9)$$

where $E_C(z) = [|h_z| - 1 + \min_{h \in h(z)} d(t_h, t_r)] \times Area(z)$

For a sub-zone *Correct* $E_S(z) = 0$ and and for a sub-zone *Segmentation Error*

$$E_S(z) = Area(z) \times \alpha_{MS} \times |h(z)| \quad (10)$$

α_{MS} is the coefficient of Merge/Split that permits to change the weight that is assigned to this type of error. If $\alpha_{MS} = 0$ then segmenting a zone is not an error.

The error E for a group in the *Split* configuration is the sum of zone errors $Er(z)$.

Merge. Configuration *Merge* corresponds to the case when one hypothesis zone corresponds to at least two reference zones. Reference zones do not always perfectly cover the hypothesis. The method that is used is the same as in the *Split* configuration but inverting hypothesis and reference.

Figure 3 illustrates the segmentation in zones and sub-zones for each group in the initial example in Figure 1.

3.3. ZoneMap Behavioral Stability

An important feature of ZoneMap is its behavioral stability in its whole range. We start from an hypothesis H with ZoneMap score $z = \frac{E}{R}$ (E = Error amount, R = Reference area) and Jaccard $j = \frac{I}{U}$ (I=Intersection area, U = Union area). We create a new hypothesis H' which adds exactly δ more correct pixels and the same number of incorrect pixels. The new ZoneMap score is:

$$z' = \frac{E - \delta + \delta}{R} = z \quad (11)$$

That shows that the ZoneMap score does not change when the hypothesis changes in a balanced way. Under the same condition, using Taylor series on $\frac{\delta}{U}$, Jaccard becomes:

$$j' = \frac{I + \delta}{U + \delta} = \frac{j + \frac{\delta}{U}}{1 + \frac{\delta}{U}} = j + (1 - j)\frac{\delta}{U} + O((\frac{\delta}{U})^2) \quad (12)$$

So not only the Jaccard score increases when the hypothesis is changed in a balanced way but the amount of the increment varies depending on the original system performance. That makes measuring the impact of small changes in the system at the development phase much harder.

4. METRIC VALIDATION

4.1. Data

The Maurdor evaluation corpus consists of documents of different types (handwritten and printed). This corpus was created by ELDA (www.elda.org) and will be distributed through the ELRA (www.elra.info) catalogue under fair licensing conditions after the Maurdor campaigns is finished. Document zones have different nature and after segmentation are classified into following categories: writing (text); photographic image; line drawing; graphics; table; separator; damaged/undefined area. A graphic area, in its turn, may belong to one of the following sub-types: logo, diagram or figure, stamp, signature, form field (comb field or sequence of identical boxes), underlined form field, line drawing. The documents are either in French, Arabic or English.

Altogether 5002 documents were used in the first Maurdor campaign with 3002 documents in the training set and 1000 documents in the development and test sets. Participants were allowed to use the development set without any restrictions.

4.2. Submissions and Scoring

Four different participants took part in the evaluation of area segmentation and classification in the first Maurdor evaluation campaign. According to the rules of the campaign the identities of participants can not be undisclosed and all the primary submissions are referenced as *S1*, *S2*, *S3* and *S5* [1].

The system of weights used in ZoneMap enables two evaluation scenarios. In the first one, equal weights (classification and segmentation error weights of 0.5 and other weights equal to 1) are used. This way all types of errors are considered equally important. This is the default configuration of ZoneMap allowing for a straightforward comparison of results across different evaluation campaigns. This scenario is used in the first Maurdor campaign reported in this paper. The second scenario consists in tailoring ZoneMap to better meet the goals of a particular evaluation campaign. For example, different weights can be assigned to different types of zones,

[1]The participants are assigned the same indices in the papers concerned with the Maurdor campaign and as there were no submission from *S4* to the task of zone segmentation and classification this index is skipped.

Table 1. Scores ZoneMap and Jaccard for primary systems submitted to the first Maurdor evaluation campaign

System	ZoneMap			Jaccard
	$\alpha_c = 0$	$\alpha_c = 0.5$	$\alpha_c = 1$	
S1	90.0	107.1	124.1	0.150
S2	60.1	75.9	91.8	**0.315**
S3	**31.2**	**57.3**	83.4	0.190
S5	52.2	62.4	**72.7**	0.287

reflecting their importance in the evaluation context. In the second Maurdor campaign text zones have weight 1, zones of images, tables, logos, signatures etc. are all weighted at 0.5, while the importance of noise regions is dropped down by assigning them a very small weight of 0.01. Different weights can also be assigned to confusions between different zone types. While for the Maurdor campaigns all the confusions were equally important, it is possible to state that, e.g. the error of classifying a graphic zone as an image is less serious than confusing it with text.

4.3. Results

The results of the first Maurdor evaluation campaign (primary systems) are presented in Table 1. ZoneMap results are presented for different decomposition/classification importance ratios. $\alpha_c = 0$ corresponds to the case when no classification error is taken into account (ZoneMap score is entirely based on decomposition errors) and vice versa for $\alpha_c = 1$. The row $\alpha_c = 0.5$ correponds to the primary setup according to which errors in segmentation and area classification are considered equally important. A ZoneMap score can exceed 100 if a submission includes a large number of false alarms.

First, it can be seen from Table 1 that performance both in terms of ZoneMap and Jaccard differs a lot across different submissions. Second, different submissions have the best ZoneMap and Jaccard scores (highlighted in bold). Moreover, the submission having the best ZoneMap score is far behind the best in terms of Jaccard and vice versa. This points out to the fact that ZoneMap and Jaccard metrics take account of different properties of the submitted systems. Most importantly, ZoneMap takes better account of area segmentation. The different configurations obtained by the different submissions are presented in Table 2.

A high ZoneMap score of the *S1* can be explained by a tendency of this system to generate many small zones that are often false alarms or splits. *S2* has the best ZoneMap score. Despite also having a rather a high number of false alarms as compared to *S3* and *S4*, it is characterized by a high number of matches and low amount of misses. *S3* has 40% less matches and almost three times more misses than *S2* - and is still the best in terms of the Jaccard score. *S4* has the Jaccard score close to that of *S3* and exhibits similar patterns for matches, false alarms and misses.

The ZoneMap score for all systems depending on the decomposition/classification importance ratio (α_c in Table 1) is presented in Figure 4. As ZoneMap score is a linear combination of segmentation and classification errors it changes linearly according to the classification error weight α_c. As different systems exhibit different behavior, an operating point may be chosen depending on the aims of a particular evaluation (0.5 for the Maurdor evaluations). One can see that the submission *S3* is the best in zone decomposition (see the leftmost part of the Figure 4 corresponding to $\alpha_c = 0$) while *S5* is better in zone classification as for higher values of α_c it obtains better scores than *S3*. The possibility of performing such an analysis and focusing on particular operating points is an advantage of ZoneMap over Jaccard as the latter does not take decomposition error into account.

Table 2. Different configurations used to calculate ZoneMap for different systems

System	Total	Match	Merge	Split	FA	Miss
S1	50 145	7 855	3 226	10 122	21 236	7 706
S2	30 625	8 852	4 710	5 025	2 324	9 714
S3	32 846	13 034	4 784	4 225	6 851	3 552
S5	26 418	8 233	4 534	4 231	2 246	7 174

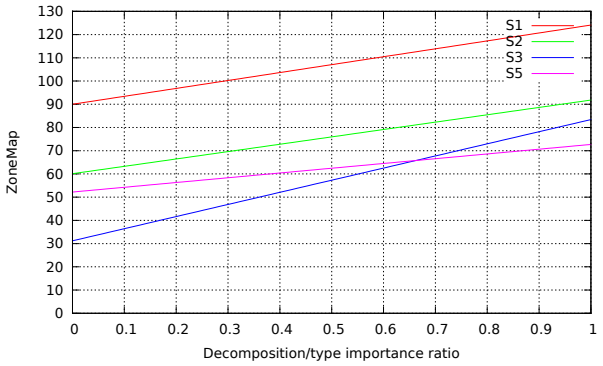

Fig. 4. ZoneMap score for all systems depending on the decomposition/classification importance ratio

5. CONCLUSION

ZoneMap, a new metric for area segmentation and classification in scanned documents was described in this paper. ZoneMap incorporates several important features that make it useful to assess system quality and interpret the results.

ZoneMap was experimented within the Maurdor evaluation campaigns and compared to a well-known metric based on the Jaccard score. The experimental results show that different systems participating in the evaluation campaign use different strategies and, as a result, exhibit different performance in terms of ZoneMap and Jaccard, demonstrating complementarity of both metrics. The major benefit of ZoneMap as compared to Jaccard is that it takes decomposition of detected zones into account and exhibits stable performance under small changes in all its range. ZoneMap also provides a flexible system of weights that can easily be used to target specific issues or user needs. For example, it is possible to give more importance to segmentation errors over errors in zone classification and vice versa. ZoneMap is implemented in the LNE *maurdor-eval* evaluation toolkit that is freely distributed under the GPL license. It makes ZoneMap a useful tool for system development and evaluation.

6. REFERENCES

[1] S. Brunessaux, P. Giroux, B. Grilheres, M. Manta, M. Bodin, K. Choukri, O. Galibert and J. Kahn, "The Maurdor project - Improving automatic processing of digital documents", Accepted to the 11th IAPR International Workshop on Document Analysis Systems (DAS), 2014.

[2] I. Oparin, J. Kahn and O. Galibert, "First Maurdor 2013 Evaluation Campaign in Scanned Document Image Processing", Accepted to the 39th IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'14, 2014.

[3] S. Mao and T. Kanungo, "Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3), 2001, pp. 242-256.

[4] J. Kanai, S. V. Rice, T. A. Nartker, and G. Nagy, "Automated evaluation of OCR zoning", IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 1995. pp. 86-90.

[5] B.A. Yanikoglu and L. Vincent, "Pink Panther: A complete environment for ground-truthing and benchmarking document page segmentation", Pattern Recognition 31, 1998, pp. 1191-1204.

[6] J. Liang, I. T. Phillips, and R.M. Haralick, "Performance evaluation of document layout analysis algorithms on the UW data set", Proc. of SPIE Conference on Document Recognition, vol. 3027, 1997, pp. 149-160.

[7] S. Mao and T. Kanungo, "Architecture of PSET: a page segmentation evaluation toolkit", International Journal of Document Analysis and Recognition (IJDAR), 4(3):205-217, 2002.

[8] Ch. Wolf and J-M. Jolion, "Object count/Area Graphs for the Evaluation of Object Detection and Segmentation Algorithms", International Journal on Document Analysis and Recognition (IJDAR), 8(4):280-296, 2006.

[9] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICDAR2009 handwriting segmentation contest", International Journal on Document Analysis and Recognition, vol. 14, no. 1, pp. 2533, 2010.

[10] I. Phillips and A. Chhabra, "Empirical performance evaluation of graphics recognition systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 9, pp. 849870, 1999.

[11] B. Moysset and C. Kermorvant, "On the evaluation of handwritten text line detection algorithms", Proc. of International Conference of Document Analysis and Recognition ICDAR'13, 2013.