# ICDAR2015 Competition on Smartphone Document Capture and OCR *(SmartDoc)*

JC. Burie*§, J. Chazalon*‡§, M. Coustaty*§, S. Eskenazi*§, M.M. Luqman*§,
M. Mehri*§, N. Nayef*§, JM. Ogier*§, S. Prum*§ and M. Rusiñol*†‡
*L3i Laboratory, University of La Rochelle, France.
†Computer Vision Center, Autonoma Univeristy of Barcelona, Spain.
‡Challenge-1 §Challenge-2
*Author names are ordered alphabetically.*

***Abstract*—Smartphones are enabling new ways of capture, hence arises the need for seamless and reliable acquisition and digitization of documents, in order to convert them to editable, searchable and a more human-readable format. Current state-of-the-art works lack databases and baseline benchmarks for digitizing mobile captured documents. We have organized a competition for mobile document capture and OCR in order to address this issue. The competition is structured into two independent challenges: smartphone document capture, and smartphone OCR. This report describes the datasets for both challenges along with their ground truth, details the performance evaluation protocols which we used, and presents the final results of the participating methods. In total, we received 13 submissions: 8 for challenge-1, and 5 for challenge-2.**

## I. Introduction

As smartphones are replacing personal scanners, the need arises for reliable digitization solutions to achieve the goal of converting captured textual information to editable and searchable format. A document image captured by a smartphone poses challenges to the digitization process such as: background removal (document segmentation), perspective correction, lighting normalization, focus or motion blur [1]. As current state-of-the-art works lack databases and baseline benchmarks for digitizing mobile captured documents, we have organized a competition for smartphone document capture and OCR in order to address this issue.

This competition aims at evaluating two steps of the digitization process of document images captured by smartphones under realistic conditions. Two new datasets are released for carrying out such evaluation. There are two independent challenges in this competition: "SMARTPHONE DOCUMENT CAPTURE" and "SMARTPHONE OCR". The competition with its datasets and evaluation protocols has an important impact on the decision process regarding the use or the development of OCR systems for mobile captured documents.

Many competitions have been proposed over the past ten years to detect and recognize text in documents (ICDAR 2003, 2005 and 2013 Robust Reading Competitions), and on skew estimation and page dewarping (ICDAR 2011 Page Dewarping and ICDAR 2013 Document Image Skew Estimation Contests). This competition – titled ICDAR15 SMARTPHONE DOCUMENT CAPTURE AND OCR COMPETITION (**SmartDoc**) – addresses a crucial lack in the past competitions by bringing together the two challenges of *smartphone capture* and *text recognition* using smartphones into a real scenario of document digitization in mobile environment.

To the best of our knowledge, there is no existing publicly available dataset comprising of documents captured in real conditions with various photometric and geometric distortions. The datasets generated for this SmartDoc competition have been made publicly available (along with their ground truth), and we hope that they will serve the scientific community for many years to come for benchmarking digitization methods.

The competition was run in open mode, meaning that the participants submitted results of their systems on the test set, and not their executables. We have relied on the scientific integrity of the authors to follow the rules of the competition. The authors were free to participate in one or both challenges. They were allowed to make multiple submissions for the same task. In total, we received 13 submissions, 8 submissions for challenge-1, and 5 submissions for challenge-2.

## II. Challenge-1: Smartphone Document Capture

This challenge is focused on efficiently detecting and segmenting page outlines within frames acquired with mobile devices. The input is a set of video clips containing a document from a predefined set, and the expected output contains quadrilateral coordinates of page outlines per each frame.

### A. Dataset

To build our dataset, we took six different document types coming from public databases and we chose five document images per class. We have chosen the different types so that they cover different document layout schemes and contents (either completely textual or having a high graphical content). In particular, we have taken data-sheet documents and patent documents retrieved from the Ghega dataset [2], title-pages from medical scientific papers from the MARG dataset [3], color magazine pages from the PRIMA layout analysis dataset [4], american tax forms from the NIST Tax Forms Dataset (SPDB2) [5], and finally typewritten letters from the Tobacco800 document image database [6]. We removed some small noise and margins from the original document images and finally rescaled them to all have the same size and fit an A4 paper format.

Each of these documents were printed using a color laserjet and we proceeded to capture them using a Google Nexus 7

tablet. We recorded small video clips of around 10 seconds for each of the 30 documents in five different background scenarios. The videos were recorded using Full HD 1920x1080 resolution at variable frame-rate. Since we captured the videos by hand-holding and moving the tablet, the video frames present realistic distortions such as focus and motion blur, perspective, change of illumination and even partial occlusions of the document pages. Summarizing, the database consists of 150 video clips comprising near 25.000 frames. We ground-truthed this collection by semi-automatically annotating the quadrilateral coordinates of the document position for each frame in the collection (see details in [7]).

### B. Evaluation Protocol

To assess the performance of the methods, we used the Jaccard index measure [8] that summarizes the ability of the different methods at correctly segmenting page outlines while also incorporating penalties for methods that do not detect the presence of a document object in some frames.

Using the document size and its coordinates in each frame, we start by transforming the coordinates of the quadrilaterals of a submitted method $S$ and of the ground-truth $G$ to undo the perspective transform and obtain the corrected quadrilaterals $S'$ and $G'$. Such transform makes all the evaluation measures comparable within the document referential. For each frame $f$, we compute the Jaccard index (JI) that measures the goodness of overlapping of the corrected quadrilaterals as follows:

$$JI(f) = \frac{area(G' \cap S')}{area(G' \cup S')}$$

where $G' \cap S'$ defines the polygon resulting as the intersection of the detected and ground-truth document quadrilaterals and $G' \cup S'$ the polygon of their union. The overall score for each method will be the average of the frame score, for all the frames in the test dataset.

### C. Participant methods

***A2iA run 1****: C. Kermorvant, A. Semenov, S. Sashov and V. Anisimov* from *A2iA St. Petesburg and Paris and Teklia Paris*. Their method starts with a Canny edge detector in the RGB space followed by an interpolation of the detected contours by Bezier curves. Some contours are selected based on their contrast and then quadrangles are selected depending on their squareness. If such steps fail to detect a valid quadrangle, a set of similar steps are applied to a denoised binary version of the input image.

***A2iA run 2****:* The second method from A2iA is the same of the first run without the low contrast contour removal.

***ISPL-CVML****: S. Heo, H.I. Koo and N.I. Cho* from *Seoul National University and Ajou University*. Their method starts by applying the Line Segment Dectector (LSD) presented in [9] to down-sampled images. Document boundaries are then generated by selecting two horizontal and vertical segments that minimize a cost function exploiting color and edge features. The final document boundaries are refined in the original high resolution image.

***LRDE****: E. Carlinet and T. Géraud* from *EPITA Research and Development Laboratory*. Their method relies on a hierarchical representation of the image named Tree of Shapes. In each frame of the video, an energy on the tree is computed in order to select the shape that looks the most like a papersheet. The energy involves two terms measuring how the shape fits a quadrilateral and if it has sub-contents like lines or images. Two Trees of Shapes are computed on the $L$ and $b*$ components of the frame (converted in the $La*b*$ space). Shapes having the highest energies in both trees are retained as candidate objects and the location of the detection in the previous frame is used to finally select the right shape among the candidate components.

***NetEase****: P. Li, Y. Niw and X. Li* from *NetEase*. Their method starts by extracting line segments by the LSD method [9], such segments are then grouped and quadrangles are formed by selecting two horizontal and vertical segment groups. The final quadrangle is selected based on its aspect-ratio, area and inner angles.

***RPPDI-UPE****: B.L.D. Bezerra, L. Leal and A. Junior* from *University of Pernambuco and Document Solutions*. Their method starts by using the HSV color space and filtering the hue channel in order to make the document pages stand from the background. Morphological operations followed by a Canny edge detector and a Hough transform yields a set of candidate polygons. Such polygons are then filtered according to their shape and position.

***SEECS-NUST****: S.A. Siddiqui, F. Asad, A.H. Khan and F. Shafait* from *School of Electrical Engineering and Computer Science and National University of Science and Technology*. Their method applies a Canny edge detection on the gray-level image to get a first estimate of the document position. A subsequent analysis of the different color channels is used to determine in which channel there is a higher contrast between document and background followed by a probabilistic Hough Transform to obtain the accurate document segmentation.

***SmartEngines****: A. Zhukovsky, D. Nikolaev, V. Arlazarov, V. Postnikov, D. Polevoy, N. Skoryukina, T. Chernov, J. Shemiakina, A. Mukovozov, I. Konovalenko and M. Povolotsky* from *Moscow Institute for Physics and Technologies, National University of Science and Technology, Institute for Systems Analysis, of Russian Academy of Sciences and Institute for Information Transmission Problems of Russian Academy of Sciences*. Their method starts with a segment extraction step by means of the LSD algorithm [9] followed by a graph construction of segments. A quadrangle selection is done on such graph after applying several size and angle filters. The final candidate quadrangle is selected by fitting a motion model by using a Kalman filter powered by an inter-frame matching strategy of local descriptors based on SURF and BRIEF.

### D. Results and Analysis

Table I summarizes the global results for challenge-1. The system proposed by the LRDE team exhibits the best performance for the task. It not only provides a higher average result quality, but also a narrower confidence interval which demonstrate a stable behavior. Systems proposed by ISPL-CVML and SmartEngines teams also perform very well, with good performance under difficult conditions like very

TABLE I. GLOBAL RESULTS FOR CHALLENGE-1.

| Ranking | Method | Jaccard Index | Confidence Interval |
|---|---|---|---|
| 1 | LRDE | 0.9716 | $[0.9710, 0.9721]$ |
| 2 | ISPL-CVML | 0.9658 | $[0.9649, 0.9667]$ |
| 3 | SmartEngines | 0.9548 | $[0.9533, 0.9562]$ |
| 4 | NetEase | 0.8820 | $[0.8790, 0.8850]$ |
| 5 | A2iA run 2 | 0.8090 | $[0.8049, 0.8132]$ |
| 6 | A2iA run 1 | 0.7788 | $[0.7745, 0.7831]$ |
| 7 | RPPDI-UPE | 0.7408 | $[0.7359, 0.7456]$ |
| 7 | SEECS-NUST | 0.7393 | $[0.7353, 0.7432]$ |

light background *(background04)* and strong occlusions *(background05)*, as it can be seen in Table II. As the 95% confidence intervals presented in Table I are overlapping for the methods proposed by the RPPDI-UPE and the SEECS-NUST teams, they are tied in seventh position.

Among the main strengths of the leading methods, we identified those 3 common stages. First, those methods rely on a very robust shape or line extraction process, which enables the formation of valid shape candidates. Then, shape filtering and selection is performed. Finally, results from previous frames are used to improve the final decision.

Regarding the average performance of all the methods on the dataset, several interesting observations can be noted. First, the study of the performance against capture conditions (the different "backgrounds") reveals that low light conditions *(background03)* are not really challenging, while light backgrounds (*background02* and *background04*) are more difficult. Severe occlusions *(background05)* are very challenging. Second, the study of the performance against document classes reveals that documents with many lines (the tables of the *tax* forms) are more challenging, as well as documents with rich bloc content like *magazine* pages.

## III. CHALLENGE-2: SMARTPHONE OCR

The goal of this challenge is to extract the textual content from document images which are captured by mobile phones. The images are taken under varying capture conditions (perspective angles, light and blur). This causes geometric and photometric distortions that hinder the OCR process. The task required from participants is to provide transcriptions of the textual content of the captured document images using an OCR system. A sample dataset was provided to participants to aid the training process.

### A. Dataset

The dataset has 12100 document images captured from 50 different paper documents with real content from wiki-books and cooking recipes from Internet. 15 documents are used to create the 3630 images of the sample set, and 35 documents are used to create the 8470 images of the test set.

All documents contain single column text printed with multiple scales, fonts, font-faces and colors. Original content is mostly English, however we have randomly replaced some words, sentences and paragraphs with random text generated by *lorem* and also by other dictionary words. Random text permits a character-level performance evaluation in the context of out-of-dictionary words.

At least 240 different images are captured per document, those captures are taken using representative values of different distortions (see "capture parameters" below). For each image, the information about the document and about capture conditions is stored for evaluation purposes.

*1) Fixed Capture Parameters:*

- Background: one colored, clear contrast with documents to facilitate the page border detection process
- Document: Fixed size (A4) and orientation
- Smartphone camera: No flash

*2) Variable Capture Parameters:*

- Smartphone: 2 mobiles: Samsung Galaxy S4 (camera: 13MP), Nokia Lumia 920 (camera: 8.7MP)
- Light: During day time: 2 lights, 1 light, no lights
- Blur: 6 values of out-of-focus (captured only with the Nokia Lumia 920 mobile)
- Perspective 1: Longitudinal incidence angle (mobile rotation around Y-axis): 4 values around the parallel position with a discrete step of $5\,\mathrm{deg}$
- Perspective 2: Lateral incidence angle (mobile rotation around X-axis): 3 values around the parallel position with a discrete step of $5\,\mathrm{deg}$
- Distance between camera and document: 35cm, 40cm

### B. Evaluation Protocol

The global ranking of participating methods is based on average character accuracy. Prior to computing character accuracy, the text results from participant methods and ground truth text are normalized in order to have consistent character encoding and remove illegal characters.

*1) Input and Output Formats:* The images of the dataset (both sample and test images) are stored as .jpg files. The ground truth contains a text file in UTF-8 encoding corresponding to each image file. As output for the test images, the participants are required to submit a text file corresponding to each test image.

Only a small subset of the Unicode character set is accepted. This subset is defined to cover most of the frequently used characters of Latin-script languages. The characters present in the dataset have been selected from the Windows CP 1252 char-set. We added ligature support to cope with the output produced by common OCR systems.

*2) Text Normalization:* Both the results and ground truth characters are normalized to avoid Unicode composition ambiguity, superscript characters, variations in quotes and hyphens, and splits of some compound characters. The Unicode Normalization Form Compatibility Composition was first used to decompose characters and re-compose them by canonical equivalence. We also performed manual transformation to limit the number of OCR errors on similar characters (quotes and hyphens). This procedure allows correct comparison between ground truth and the results of participants methods. Both normalization and evaluation programs are publicly available at the following address: https://github.com/SmartDOC-MOC

TABLE II. AVERAGE PERFORMANCE PER BACKGROUND AND DOCUMENT CLASS, FOR EACH METHOD OF CHALLENGE-1.

| | Background | | | | | Document class | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | 01 | 02 | 03 | 04 | 05 | datasheet | letter | magazine | paper | patent | tax |
| A2iA-1 | 0.9724 | 0.8006 | 0.9117 | 0.6352 | 0.1890 | 0.8245 | 0.8005 | 0.7026 | 0.8555 | 0.7774 | 0.7073 |
| A2iA-2 | 0.9597 | 0.8063 | 0.9118 | 0.8264 | 0.1892 | 0.8538 | 0.8250 | 0.7577 | 0.8821 | 0.8060 | 0.7240 |
| ISPL-CVML | 0.9870 | 0.9652 | 0.9846 | 0.9766 | 0.8555 | **0.9761** | 0.9691 | 0.9558 | **0.9719** | 0.9586 | 0.9626 |
| LRDE | 0.9869 | 0.9775 | 0.9889 | **0.9837** | **0.8613** | 0.9758 | **0.9718** | **0.9707** | 0.9715 | **0.9698** | **0.9696** |
| NetEase | 0.9624 | 0.9552 | 0.9621 | 0.9511 | 0.2218 | 0.8950 | 0.8666 | 0.8958 | 0.8798 | 0.8723 | 0.8817 |
| SEECS-NUST | 0.8875 | 0.8264 | 0.7832 | 0.7811 | 0.0113 | 0.7745 | 0.8035 | 0.7292 | 0.7186 | 0.7470 | 0.6552 |
| RPPDI-UPE | 0.8274 | 0.9104 | 0.9697 | 0.3649 | 0.2163 | 0.6606 | 0.7126 | 0.8232 | 0.7547 | 0.7191 | 0.7803 |
| SmartEngines | **0.9885** | **0.9833** | **0.9897** | 0.9785 | 0.6884 | 0.9671 | 0.9498 | 0.9438 | 0.9596 | 0.9562 | 0.9517 |
| All | 0.9465 | 0.9031 | 0.9377 | 0.8122 | 0.4041 | 0.8659 | 0.8624 | 0.8474 | 0.8742 | 0.8508 | 0.8290 |

*3) Evaluation Metrics:* The character accuracy per document image is computed using the UNLV-ISRI *accuracy* tool [10] with UTF-8 encoding support. This tool computes character and word accuracy and provides 95% confidence intervals for these values. The average character accuracy across the test set is used for ranking participants methods.

### C. Participants methods

***A2iA**: C. Kermorvant et al. from Teklia & A2iA, France*: This method preprocesses the images using their second method submitted to challenge-1, dewarps the images and extracts the text lines using projection profiles. Then, an LSTM recurrent neural network is trained to recognize the binary text-lines.

***CartPerk**: D. Kumar from CartPerk Technologies, India*: This method uses the blue background to extract and dewarp the document. The local contrast is computed, and the resulting image is then binarized with a local threshold in a 64x64 window. Finally Tesseract processes the binary image.

***CCC**: M. Soheili et al. from DFKI, Germany and T. Modares University, Iran*: This method uses the background color to detect and dewarp the document. The image is then binarized to extract lines, words and subwords. Those are then clustered incrementally across all the corpus. A 1D LSTM is trained on both sharp and blurry gray-scale text-lines for recognizing subwords. Clusters of subwords are labeled by majority voting.

***Digiform**: G. Kragoz from Kocaeli University, Turkey*: This method applies a strong blur followed by a canny filter to detect the corners of a document. The image is then dewarped and remapped to 300dpi. The image is binarized with an adaptive threshold. Finereader performs the OCR step.

***LRDE**: E. Carlinet and T. Graud from EPITA's LRDE, France*: This method uses the corners of the largest centered component to dewarp the document. The document is then binarized based on a morphological thick gradient and a morphological Laplacian. Finereader performs the OCR step.

### D. Results and Analysis

The participants submitted their results in the form of text files after performing the transcription process on the images of the test set. Evaluating the results has been carried out according to what we described in the previous subsections. Firstly, we present the overall average recognition accuracy of the participating methods in Table III to show the global ranking of the methods. Note that we have also included the performance of Abbyy Finereader Engine 11 – without any pre- or post-processing – as a baseline reference method.

TABLE III. RESULTS AND RANKING OF THE PARTICIPANTS METHODS OF CHALLENGE-2 IN TERMS OF AVERAGE CHARACTER ACCURACY.

| Ranking | Method | Character Accuracy (%) | Confidence Interval (%) | # Errors per Page |
|---|---|---|---|---|
| 1 | CCC | 99.93 | [99.92,99.93] | 2 |
| 2 | LRDE | 95.85 | [95.56,96.14] | 120 |
| 3 | Digiform | 95.33 | [94.98,95.68] | 135 |
| 4 | A2iA | 93.84 | [93.54,94.15] | 178 |
| 5 | CartPerk | 91.19 | [90.60,91.79] | 254 |
| 6 | Finereader | 87.61 | [87.20,88.02] | 357 |

The best two performances have been achieved by the two methods "CCC" and "LRDE" respectively, where the "CCC" method performed the OCR process on the test set by considering it as a whole book, hence taking a great advantage of having the same document repeated in many different captures. The "LRDE" method – and also the rest of the submitted methods – processed the images of the test set one by one as it should be in a realistic scenario.

Next, we present the distortion-wise performance evaluation. We have 3 main types of distortions: perspective distortions due to camera position with respect to the document, blur due to unfocused camera and light distortions due to the degree of available light during the capture. Combinations of these distortions are present in the dataset images in different levels ranging from none, low to high levels. Figures 1, 2 and 3 show the performance of the methods with respect to different degrees of each of the mentioned distortions.

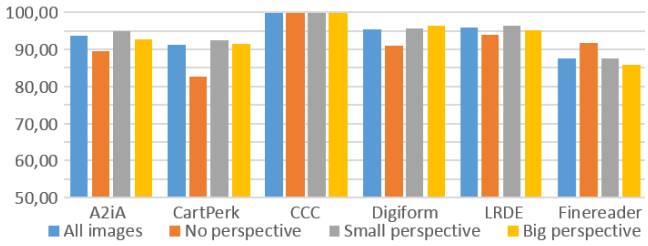


Fig. 1. Character recognition accuracy on different degrees of perspective distortions.

The "CCC" method is insensitive to the perspective distortions. The 4 other methods ("A2iA", "CartPerk", "Digiform" and "LRDE") provide recognition accuracy on no-perspective images lower than images with perspective distortion. This

could be due to inappropriate preprocessing (dewarping, perspective correction) steps, and/or that their methods are trained on images with perspective distortion. Method "CCC" is insen-

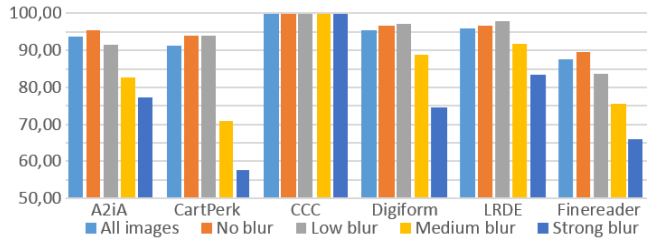


Fig. 2. Character recognition accuracy on different levels of blur distortion.

sitive to blur distortion while the other methods are sensitive to it, specially with medium and high blur.

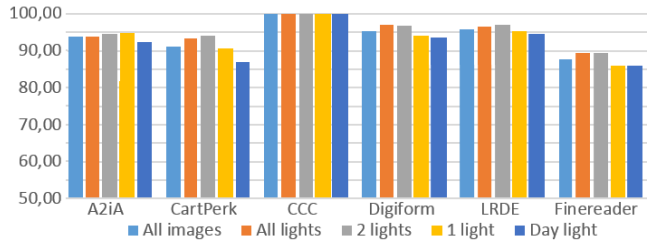


Fig. 3. Character recognition accuracy on different levels of lighting.

Method "CCC" is insensitive to different light conditions, while methods "A2iA" and "LRDE" provide slightly different accuracies with those conditions. "CartPerk" and "Digiform" are more sensitive to light and perspective distortions.

We believe that the "CCC" method had such a highly accurate performance despite distortions due to: training on blurry images, working at gray-scale level for both clustering and LSTM training, sophisticated pre- and post-processing and finally, making use of the nature of the dataset where each document is captured over 240 times.

Most methods – except for "CCC" – have made a minimum of 120 character errors per page. This means that the problem of OCR in mobile captured documents still needs further research and exploration, even for our dataset of contemporary documents with simple layout, Latin-script text, sufficient day light and small degrees of blur in most documents.

## IV. Conclusions and Future Perspectives

This report has presented an overview of the organization and the findings of the ICDAR 2015 smartphone document capture and OCR competition (SmartDoc). There has been 13 participants for the competition in both challenges. This shows an interest of the document analysis and recognition community in this field.

All the details about the competition and the two datasets will be made available on the competition website: http://l3i.univ-larochelle.fr/icdar2015smartdoc at the time of ICDAR 2015. In future work, we will be working on creating more challenging datasets for addressing different research problems in mobile captured documents.

## Acknowledgments

We would like to thank the management team of the DAE platform for hosting the competition datasets. The work of Challenge-1 is partially supported by the People Programme (Marie Curie Actions) of the Seventh Framework Program of the European Union (FP7/2007-2013) under REA grant agreement no. 600388, and by the Agency of Competitiveness for Companies of the Government of Catalonia, ACCIÓ, and the Spanish project TIN2014-52072-P. The work of Challenge-2 is funded by the Conseil General de la Charente Maritime (France), and is supported by the European Commission and the Conseil Rgional de Poitou-Charentes under the FEDER program DATA-PC.

## References

[1] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," *IJDAR*, vol. 7, no. 2, pp. 84–104, 2005.

[2] E. Medvet, A. Bartoli, and G. Davanzo, "A probabilistic approach to printed document understanding," *International Journal of Document Analysis and Recognition*, vol. 14, no. 4, pp. 335–347, December 2011.

[3] G. Ford and G. Thoma, "Ground truth data for document image analysis," in *Proceedings of the Symposium on Document Image Understanding and Technology*, 2003, pp. 199–205.

[4] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher, "A realistic dataset for performance evaluation of document layout analysis," in *Proceedings of the 10th International Conference on Document Analysis and Recognition*, 2009, pp. 296–300.

[5] D. Dimmick, M. Garris, and C. L. Wilson, "Structured forms database," National Institutte of Standards and Technology, Tech. Rep., 1991.

[6] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard, "Building a test collection for complex document information processing," in *Proc. Int. ACM SIGIR Conf.*, 2006, pp. 665–666.

[7] J. Chazalon, M. Rusiñol, J. Ogier, and J. Lladós, "A semi-automatic groundtruthing tool for mobile-captured document segmentation," in *ICDAR*, 2015.

[8] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[9] R. von Gioi, J. Jakubowicz, J. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE Trans. PAMI*, vol. 32, no. 4, pp. 722–732, 2010.

[10] S. Rice, F. Jenkins, and T. Nartker, "The fifth annual test of OCR accuracy," Information Science Research Institute, Tech. Rep., 1996.