SmartDoc 2017 Video Capture: Mobile Document Acquisition in Video Mode

J. Chazalon[⋆†], P. Gomez-Krämer[⋆], J.-C. Burie[⋆], M. Coustaty[⋆], S. Eskenazi[⋆],
M. Luqman[⋆], N. Nayef[⋆], M. Rusiñol[⋄], N. Sidère[⋆] and J.-M. Ogier[⋆]
[⋆]L3i — Univ. La Rochelle, La Rochelle, France
[†]LRDE — EPITA, Paris, France
[⋄]CVC — Univ. Autònoma de Barcelona, Barcelona, Spain
Email: icdar (dot) smartdoc (at) gmail (dot) com

***Abstract*—As mobile document acquisition using smartphones is getting more and more common, along with the continuous improvement of mobile devices (both in terms of computing power and image quality), we can wonder to which extent mobile phones can replace desktop scanners. Modern applications can cope with perspective distortion and normalize the contrast of a document page captured with a smartphone, and in some cases like bottle labels or posters, smartphones even have the advantage of allowing the acquisition of non-flat or large documents. However, several cases remain hard to handle, such as reflective documents (identity cards, badges, glossy magazine cover, etc.) or large documents for which some regions require an important amount of detail. This paper introduces the SmartDoc 2017 benchmark (named "SmartDoc Video Capture"), which aims at assessing whether capturing documents using the video mode of a smartphone could solve those issues. The task under evaluation is both a stitching and a reconstruction problem, as the user can move the device over different parts of the document to capture details or try to erase highlights. The material released consists of a dataset, an evaluation method and the associated tool, a sample method, and the tools required to extend the dataset. All the components are released publicly under very permissive licenses, and we particularly cared about maximizing the ease of understanding, usage and improvement.**

I. Introduction

Nowadays, it is obvious that people use their smartphones to capture documents, edit them, and share them. The following factors were determinant to enable this usage: the imaging performance of smartphones is a very discriminative feature and vendors dedicated much work in this direction; more and more users own a smartphone; the computing power of those devices have increased significantly; their connectivity (LTE deployment in particular) allows to transfer files easily; and finally, users carry their smartphone constantly with them, allowing to capture content at any moment.

Desktop scanners, as available to the large majority of users, are progressively loosing their advantages over smartphones regarding document image acquisition. Indeed, there are many applications which can assist the user during the capture by correcting the perspective and the contrast, and also sometimes by triggering the capture when the document is well framed and when the picture is sharp.

The SmartDoc series is an ongoing effort aiming at evaluating the performance of document imaging solutions using smartphones, and encouraging progress in this direction. The first release, organized as a competition for ICDAR 2015 [1], was composed of two challenges. *SmartDoc 2015 – Challenge 1* was about locating a document page (its corners) within video frames to simulate how some applications could assist the user by detecting the document in real time during the preview phase preceding the capture. *SmartDoc 2015 – Challenge 2* was about character recognition from document images (high resolution) captured with smartphones. The second release, named *SmartDoc QA* [2], is a dataset which extends the dataset of SmartDoc 2015 – Challenge 2 with more variability in capture conditions. For all those releases, we provide a dataset with the associated ground-truth, and the evaluation tools. All the existing content can be found at:

⋆ http://smartdoc.univ-lr.fr ⋆

The latest release of the SmartDoc series is introduced in this paper. It continues the series by tackling a new challenge: mobile document imaging using video capture. As mobile document imaging makes great progress, we identified some cases for which it can provide a better solution than traditional scanners: first large documents like posters which cannot always be captured with a single image due to their size and the desirable level of detail; and second reflective documents, like identity cards, badges, magazine covers, etc. Such a task is therefore both a stitching and a reconstruction challenge as the user can move the camera over the document freely, swiping over it to gather document parts, getting the camera very close the reveal details in some specific regions, or tilting to try to remove highlights. Because of the nature of the problem, a method addressing it could also, to some extent, be capable of handling some corner cases like documents with occlusions in front of them, or even documents displayed on a screen (which produces Moiré patterns).

After a brief review of the related work (Sec. II), this paper introduces the following contributions.

- A definition of the task of mobile document imaging using video capture, in a way which enables the evaluation of competing methods (Sec. III).
- The *SmartDoc Video Capture* dataset (Sec. IV).
- A new evaluation protocol to compare the restored image with the original ground-truth image (Sec. V).

- Open material; containing datasets, evaluation tools and tools to ease the extension of the dataset (Sec. VI).
- Some thoughts about the benefits and the limitations of this way of releasing research material (Sec. VII).

II. Related Work

This SmartDoc release can be compared to different kind of works depending on the aspect we focus on.

First it is worth noting that there are few datasets and even fewer benchmarks targeting the evaluation of the performance of methods supporting mobile document image acquisition. Some works were published in the more general context of camera-based document image acquisition and recognition. Shafait et al. launched in 2007 their *Document Image Dewarping Contest* [3] which is related to our work as the task is a form of restoration of the image. The input was however comprised of isolated images, and the evaluation protocol consisted in computing the edit distance between the original text from the pages and the text extracted by some OCR engine from the dewarped (restored) images. Such evaluation procedure was considered for our benchmark, but it requires the use of documents containing mostly text and therefore we left this option for a later version. Furthermore, it restricts the evaluation to text recognition tasks.

Following this work, Bukhari et al. created in 2012 the *IUPR Dataset of Camera-Captured Document Images* [4] which was used for their evaluation of page dewarping algorithms using SIFT features [5]. Regarding the dataset, it contains camera-captured warped images and flat (scanned) images, along with text ground-truth and pixel-accurate content type classification. The evaluation procedure consists in counting the number of correct (in spatial and descriptor spaces) matches between SIFT features extracted from ground-truth and restored image. While SIFT features provide a very powerful way to register document images, as we will see in Sec. V, the ambiguity in the features extracted from document images (due to high redundancy in textures) introduces a lot of uncertainty regarding the cause for rejecting matches, letting us think that this evaluation method may suffer from reliability issues.

A last dataset we are aware of is the dataset published by Kumar et al. for the evaluation of their DeltaDom method [6]. This dataset targets mobile document imaging and provides a set of document images captured at various focus levels. Authors provided the OCR accuracy of three standard OCR engines for all the images, and the task consisted in predicting OCR accuracy given the original images. Once again, this dataset does not feature videos of documents. Regarding datasets and benchmarks, no existing work features what *SmarDoc 2017 Video Capture* does: documents captures in video mode using smartphones, with the associated evaluation procedure and tools, in a totally open and extensible way.

In the particular case of the later SmartDoc release, focusing on mobile document image acquisition using video, there are only a handful of publications related to the task of reconstructing a document image from a video input. The first notable work is the one of Liang et al. [7], which introduces a framework, composed of a registration phase and a blending phase, for reconstructing a document image from a set of images of various points of views, illumination and focus levels. This publication reports encouraging results but the evaluation in conducted on a private dataset with an OCR metric, calling for the creation of a separate and public benchmark.

More recently, Luqman et al. [8] introduced a mobile application and the associated restoration tool which enable the acquisition of a set of document images using a smartphone so that the perspective transforms between the different images can be estimated precisely and the resulting image exhibits a resolution higher than individual images. Once again, this work presents interesting results but is evaluated on a private dataset.

Finally, regarding the open character of the benchmark we introduce, an important inspiration for our work was the *Robust Reading Competition* series at ICDAR [9]–[12]. This competition targets text detection, language identification and text recognition in natural and digital born images, a task different from the one of SmartDoc 2017 Video Capture. However, we found the open character of the competition very inspiring: a synchronized competition between a first set of methods is organized at ICDAR, then the challenge remains open and new participants can submit results and enter the leader board at any time. This is the first public platform for continuous competition on (document) image processing. The drawbacks of this approach is that the ground-truth remains secret and the evaluation methods are only available through online submission (the implementation not being released yet). For the latest SmartDoc release, we chose to explore a fully open strategy where everything is made public and open.

III. Task Definition

The task of this benchmark is defined to evaluate the possibility of implementing the following use case. (This task is different from the ones of the previous challenges, and extends our previous works.) Let us imagine that some smartphone application enables some user to aim at a document — would it be a business card or a poster — and shows him/her the detected region of this document during the preview. We suppose this detection (based on the performance of methods evaluated during the SmartDoc 2015 competition – Challenge 1) works well. Once the user is satisfied with what the application identifies as the document, he/she triggers the capture *in video mode*. This *video capture* lets the user move the camera around the document to gather as much information as possible, recording the stream of images as a video of approximatively 15 seconds. He/she may want to remove highlights, zoom on a particular part with important details, or ensure that some parts are sharp enough. We make no assumption on how the user feedback should be presented.

What we want to evaluate is how reconstruction methods could produce a high-quality document image, as close as

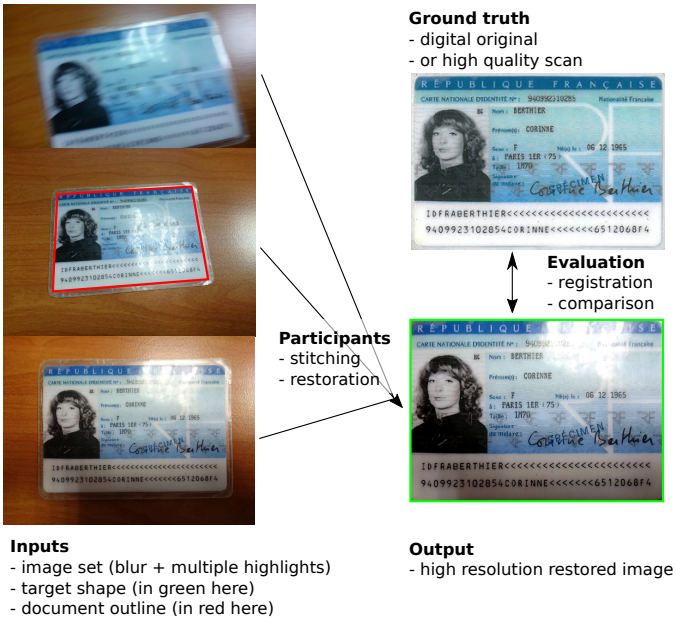

Fig. 1. Overview of the task under evaluation for the SmartDoc 2017 Video Capture benchmark.

possible to what the original document is or what a perfectly scanned version would be, using the following inputs (as summarized in Fig. 1):

1) the video sequence captured using the previously defined protocol;
2) the target resolution and the shape of the image to produce;
3) the coordinates of the region of interest in a reference frame at the beginning of the video sequence.

Target resolution is an arbitrary choice which has little impact on the internals of a solution, while facilitating greatly the evaluation. Target shape and coordinates of the region of interest are a bit more debatable in the sense that a real application would need either to ask the user for those elements, or discover it by itself. However, there already are some existing solutions for detecting the outline [1] of the document under capture and recover its shape [13], which produce good results. We therefore think that we should focus on the subsequent stages of the pipeline, and we believe this legitimates the fact that we do not simply provide the raw video sequence, but also some extra elements which facilitate the evaluation without simplifying the actual problem.

This task is therefore both a stitching problem and a restoration problem.

IV. Dataset

This benchmark introduces a new dataset. The latter contains realistic input data captured by hand from multiple users using various devices. Furthermore, documents exhibit various challenges as large size, reflective surfaces, text and graphic content. Finally, capture conditions are also challenging with blurry frames, various illumination conditions (both in direction and intensity), and variable motions as the user tries to remove highlights, to zoom into details, etc.

The dataset is separated into two subsets:

- a sample set (or training set) composed of 10 documents captures;
- a test set composed of 37 document captures.

Table I describes the documents contained in each subset of the dataset. Figure 2 provides some examples of documents and captured frames. Each document is captured only once, to avoid encouraging the use of multiple acquisitions to improve the restoration.

TABLE I
Overview of the documents contained in the dataset.

Name	Subset	Notes
card01	Train.	A6, business card like
card02	Train.	A5, reflective badge
paper01	Train.	A4 research paper
paper02	Train.	A4 research paper
poster01	Train.	A1 research poster
poster02	Train.	A1 research poster
receipt01	Train.	Receipt with cut ground-truth
receipt02	Train.	Receipt
screen01	Train.	Filmed on computer screen
slide01	Train.	Projected slides
sample01	Test	Poster - A3-landscape
sample02	Test	Poster - A3-landscape
sample03	Test	Paper - A4-portrait
sample04	Test	Paper - A4-portrait
sample05	Test	Screen - A4-portrait
sample06	Test	Screen - A4-portrait
sample07	Test	Screen - A4-portrait
sample08	Test	Poster - A0-portrait
sample09	Test	Poster - A0-portrait
sample10	Test	Poster - A0-portrait
sample11	Test	Poster - A0-portrait
sample12	Test	Poster - A0-portrait
sample13	Test	Poster - A0-portrait
sample14	Test	Poster - A0-portrait
sample15	Test	Poster - A0-portrait
sample16	Test	Poster - A0-portrait
sample17	Test	Poster - A0-landscape
sample18	Test	Poster - A0-portrait
sample19	Test	Poster - A1-portrait
sample20	Test	Booklet - A4-portrait
sample21	Test	Booklet - A4-portrait
sample22	Test	Booklet - A4-portrait
sample23	Test	DoorSign - A6
sample24	Test	Paper - A4-portrait
sample25	Test	Screen - Slide-landscape
sample26	Test	Screen - Slide-landscape
sample27	Test	Screen - Slide-landscape
sample28	Test	Screen - Slide-landscape
sample29	Test	Screen - Slide-landscape
sample30	Test	Screen - Slide-landscape
sample31	Test	Screen - Slide-landscape
sample32	Test	Screen - Slide-landscape
sample33	Test	Screen - Slide-landscape
sample34	Test	TransportMap - A4-landscape
sample35	Test	Poster - A0-portrait
sample36	Test	Brochure - 1/3 A4 vertical
sample37	Test	Brochure - 1/3 A4 vertical

Each *document capture* contains the following files:

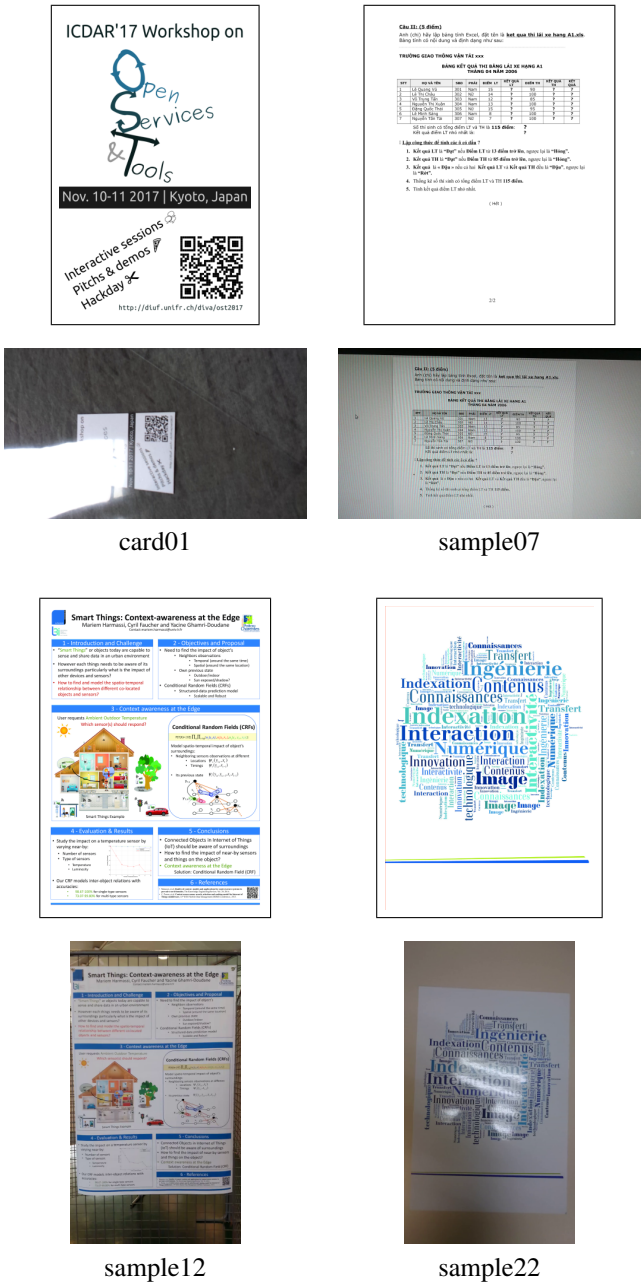

Fig. 2. Some examples of ground truth images (top rows) and the reference frame used in the sample (bottom rows).

- `ground-truth.png`: the image against which restored results are evaluated — an ideal image generated from the source;
- `input.mp4`: the input video sequence;
- `reference_frame_XX_extracted.png`: a copy of the reference frame for which the coordinates of the object to capture are known;
- `reference_frame_XX_extracted_viz.png`: the reference frame where the outline of the object is indicated with a colored line;
- `reference_frame_XX_dewarped.png`: the perspective-corrected version of the reference frame — this image has the same shape as the image to produce;
- `task_data.json`: an easy-to-parse file containing the size of the image to produce, the index of the reference frame, and the coordinates of the object to detect in the reference frame;
- `extra_pictureYY.jpg`: some extra pictures of the scene containing the document to capture — this may enable other uses for the dataset;
- `source.pdf` or `source_bitmap.png`: the source used to generate the ground-truth image and to produce the physical document.

The reference frame is a frame found at the beginning of the video stream inside which the document is fully visible and sharp. The selection of the reference frame was performed manually.

When creating the dataset, the following constraints were put on the documents:

- the source is available in a digital/vector format;
- the document must be rectangular;
- it must be almost flat during the capture (no bottle labels for instance) — this was not strictly observed as several posters exhibited some warping due to their hanging;
- the source must not be available publicly — because the dataset was originally created to be used in a competition, but this constraint is not relevant anymore;
- it must be possible to distribute the original document (no copyright issues, no confidential information).

V. Evaluation Protocol

The evaluation protocol is designed to estimate the quality of the reconstructed images for a very general usage: *how to capture and archive an image which is clear and readable for humans.* OCR and other automated processing should also be performed on such images, but in order to keep evaluation simple and because some documents contain only little text, we chose a metric focusing on perceived visual quality: the Mean Structural Similarity (MSSIM) proposed by Wang et al. [14]. This is a *full-reference image quality assessment measure* (which compares a degraded image against a reference image) which was proved to correlate well with how humans perceive quality variations [15]. This evaluation measure is mostly based on a computation of the correlation between the gradients of the two image signals. In our case, we consider only the *luminance channel* of the images. This evaluation protocol is entirely different from the one of the previous SmartDoc editions which focused on OCR recognition and page outline segmentation.

The drawback of this evaluation method is that it is very sensitive to alignment, making a single pixel shift a major issue as it causes a drop in the evaluation measure even if the image is perfectly restored. In the particular case of this benchmark, small mis-alignments should not be penalized as they would not prevent a human viewer from reading the content, nor an automated system from performing well. To enable the use of MSSIM, we added several alignment stages to cope

with this issue. Experiments showed that alignment using local descriptors could be precise enough to enable the use of the MSSIM evaluation method.

The first alignment stage is performed during dataset creation. Using local descriptors (SIFT in our case), we align the ground-truth image with the reference frame to produce a dewarped reference frame. There is no restriction for the domain of the resulting homography at this first stage. Such image can be used as a coarse alignment indication by reconstruction methods.

The second alignment stage is performed during the evaluation of the quality of the reconstruction of the result image against the ground-truth image. Improving the alignment will improve the MSSIM measure, and therefore the goal is to maximize the MSSIM value within a certain reasonable variation domain for the alignment parameters. For each image pair (reconstructed image R, ground-truth image G), we compute three variants of the MSSIM measure and report the highest one.

1) *No alignment:* R and G are compared using MSSIM without any alignment, to provide a baseline for the improvement of the alignment.
2) *Global alignment:* R and G are aligned globally using local descriptors (SIFT), and the resulting homography is restricted in such way that it cannot "displace" any corner of R from more than 1% of the size of the image in each direction; i.e. the upper left corner at $(0,\, 0)$ of a 512×512 ground-truth image cannot by corrected by a perspective transform which introduces a displacement that is bigger than 5.12 pixel in horizontal and vertical directions. The computation of the MSSIM is then performed on the dewarped image, taking into account that some parts of the image may not be visible: as mask of the visible area is used to restrict the computation to those pixels.
3) *Local alignment:* local patches of R and G are aligned to cope with local deformations of the surface of the document (this is typically the case with printed paper which cannot be exactly flat). The tolerance for the displacement of patch corners is limited to 5% of the size in both directions. As for global alignment, only visible pixels are considered for the computation.

VI. Summary of Available Material

We summarize here the material we released publicly. There are two datasets, released under the Creative Commons Attribution-ShareAlike 4.0 International License:

1) a sample set of 10 document capture samples with the associated ground-truth;
2) a test set of 37 document capture samples with the associated ground-truth and extra scene images.

This license, used by Wikipedia and targeting non-software content, lets others *"remix, tweak, and build upon [our] work even for commercial purposes, as long as they credit [us] and license their new creations under the identical terms"*.

To complete the benchmark, we also released the following tools under the very permissive MIT license:

1) *dataset creation tools:* this repository contains the procedure and the tools used to generate the dataset (reference frame identification in particular);
2) *evaluation tools:* this repository contains our implementation of the evaluation method we proposed;
3) *a sample method:* implemented in Python 2.7+, this naive reconstruction method provides a starting point for any student or researcher who wants to have a first running (and ready-to-evaluate) implementation within a few minutes.

The MIT license grants permissions for private and commercial use, distribution and modification, under the condition of integrating the license and copyright notice. It explicitly state we provide no warranty and cannot be liable for any reason.

VII. Discussion

In this last section, we would like to share our experience regarding the production and release of benchmarks.

First we believe that, while being time-consuming, there are important benefits from creating a benchmark before having any method to test on it:

- it forces a clear problem definition before conducting any actual work on a solution;
- it avoids creating or interpreting experimental data in a way that favors a given method — this is related to the recent "pre-registration" initiative [16] promoting more transparent research;
- it encourages the use of realistic data;
- it facilitates technology transfer by defining a common evaluation "language" between companies and labs;
- it automates experiments and lowers time to publication.

Furthermore, it also gives a good visibility to benchmark creators, which compensates greatly the time invested.

To give our work the best potential, we tried to make it easy to find (website and available on GitHub), easy to use (Python programs with few dependencies and documentation), easy to understand (minimal software architecture) and completely open. This has the disadvantage of not allowing us to control the actual performance of submitted results, like the Robust Reading competition does, but on the other hand this enables everyone to criticize and to improve the evaluation process, without requiring some heavy infrastructure to run an online evaluation platform.

In the future, we will work on releasing our software both as open-source and as online tools thanks to platforms like DIVA-Services [17]. We have also started to report the results published by authors on our website, making the cost for hosting and for managing the online test suite very moderate when compared to online competitions. Finally, we plan to release Docker images for every tool we created, in order to ease as much as possible their use.

ACKNOWLEDGMENT

This work was supported by the Spanish project TIN2014-52072-P, by the CERCA Program / Generalitat de Catalunya, and by the MOBIDEM project, part of the "Systematic Paris-Region" and "Images & Network" Clusters, funded by the French Government and its economic development agencies.

REFERENCES

[1] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. Luqman, M. Mehri, N. Nayef, J. Ogier, S. Prum, and M. Rusiñol, "ICDAR2015 competition on smartphone document capture and OCR (smartdoc)," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2015, pp. 1161–1165.

[2] N. Nayef, M. Luqman, S. Prum, S. Eskenazi, J. Chazalon, and J. Ogier, "SmartDoc-QA: A dataset for quality assessment of smartphone captured document images - single and multiple distortions," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2015.

[3] F. Shafait and T. M. Breuel, "Document image dewarping contest," in *2nd Int. Workshop on Camera-Based Document Analysis and Recognition*, Curitiba, Brazil, 2007, pp. 181–188.

[4] S. S. Bukhari, F. Shafait, and T. M. Breuel, "The IUPR Dataset of Camera-Captured Document Images," in *Camera-Based Document Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Iwamura and F. Shafait, Eds. Springer Berlin Heidelberg, Jan. 2012, no. 7139, pp. 164–171.

[5] ——, "An Image Based Performance Evaluation Method for Page Dewarping Algorithms Using SIFT Features," in *Camera-Based Document Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Iwamura and F. Shafait, Eds. Springer Berlin Heidelberg, Jan. 2012, no. 7139, pp. 138–149.

[6] J. Kumar, P. Ye, and D. Doermann, "A Dataset for Quality Assessment of Camera Captured Document Images," in *International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, 2013, pp. 39–44.

[7] J. Liang, D. DeMenthon, and D. Doermann, "Camera-based document image mosaicing," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 2. IEEE, 2006, pp. 476–479.

[8] M. M. Luqman, P. Gomez-Krmer, and J.-M. Ogier, "Mobile phone camera-based video scanning of paper documents," in *International Workshop on Camera-Based Document Analysis and Recognition*. Springer, 2013, pp. 164–178.

[9] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, "ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email)," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1485–1490.

[10] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1491–1496.

[11] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1484–1493.

[12] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, and others, "ICDAR 2015 competition on robust reading," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 1156–1160.

[13] Z. Zhang and L.-w. He, "Note-taking with a camera: whiteboard scanning and image enhancement," Microsoft Research, Technical Report MSR-TR-2003-39, Jun. 2003.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[15] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *Image Processing, IEEE Transactions on*, vol. 15, no. 11, pp. 3440–3451, 2006.

[16] B. Nosek, C. Ebersole, A. DeHave, and D. Mellor, "The Preregistration Revolution," Open Science Framework, Tech. Rep., Jun. 2017. [Online]. Available: https://osf.io/2dxu5/

[17] M. Würsch, R. Ingold, and M. Liwicki, "Divaservices – a restful web service for document image analysis methods," in *Digital Humanities*, Sydney, Australia, 07/2015 2015.