Mass Digitization of Archival Documents using Mobile Phones

Florian Kleber, Markus Diem, Fabian Hollaus and Stefan Fiel
Computer Vision Lab
TU Wien
Favoritenstrasse 9
1040 Vienna, Austria
Email: florian.kleber@tuwien.ac.at

***Abstract*—Digital copies of historical documents are needed for the Digital Humanities. Currently, cameras of standard mobile phones are able to capture documents with a resolution of about 330 dpi for document sizes up to DIN A4 (German standard, 297 x 210 mm), which allows a digitization of documents using a standard device. Thus, scholars are able to take images of documents in archives themselves without the need of book scanners or other devices. This paper presents a scanning app, which comprises a real time page detection, quality assessment (focus measure) and an automated detection of a page turn over if books are scanned. Additionally, a portable device - the ScanTent - to place the mobile phone during scanning is presented. The page detection is evaluated on the ICDAR2015 SmartDoc competition dataset and shows a reliable page detection with an average Jaccard index of 75%.**

I. Introduction

One of the main prerequisites for research in Digital Humanities is the availability of digital documents. Though, large amounts of library holdings are already digitized, there are many collections in libraries and archives where no digital copy is available. The conference of the head of the archive administration of the German federation (*Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder (KLA)*[1] states in a recommendation letter [1] that archives plan a medium-term digitization of 5-10% of archives holdings. Thus, 90-95% of documents are not digitally available.

In that case, scholars interested in documents have two options: either to order the scans via a digitization-on-demand service (which can be cost-intensive and time consuming) or to use a digital device available at the library to take pictures from the documents of their interest. Challenges if documents are scanned by archive visitors are: (1) binding of documents (e.g. books), (2) the light condition in reading rooms and (3) a prohibition of flash light. The German research foundation - *Deutsche Forschungsgemeinschaft (DFG)* - has recommendation for the digitization of documents[2], which states that at least 300dpi and 8-bit color depth for each channel should be used. Currently, this can be achieved with cameras of standard mobile phones. The Nexus 5X has a camera resolution of 4032 x 3024 (12MP), which results in a resolution of about 330 dpi for DIN A4 (German standard, 297 x 210 mm). Thus, mobile phones can be used as standard devices which are available to almost all archive visitors. Additionally, the use of digital cameras avoids the problems of book bindings compared to scanners, since a complete flat surface is not necessary.

In this paper DocScan[3], a document scan app, is presented, which has a real time page detection and a quality assessment (focus measure) as feedback for the user. Additionally, a background model is learned to detect if a books page is turned. This allows a series mode, which takes an image automatically once a page is turned. The page detection is evaluated on the dataset of the ICDAR2015 SmartDoc competition [2]. DocScan is open source, available on github[4], and in the Google Playstore. Additionally, a device to place the mobile phone, the ScanTent, for the series mode is presented shortly. The ScanTent blocks ambient light, provides a non-destructive led light and works as a mount for mobile phones. Hence, operators just have to flick through a book in order to digitize it but do not need to press any button to take the picture. The DocScan app and the ScanTent are developed in the course of the READ (Recognition and Enrichment of Archival Documents) H2020 project[5]. The goal of the project is to provide services, such as automatic transcriptions, to the main target groups, which are archives, libraries, humanities scholars, public users, and computer scientists. These services will be made available via the Transkribus platform. The basis for the services are digital images of documents which can be done with the DocScan app.

The research contribution is a fast methodology for page detection and a fast in-focus measure for mobile applications. Additionally, it is shown that a standard background detection method can be applied to recognize automatically a page turn. The combination of DocScan with the ScanTent results in a low cost mobile scanning system to digitize documents in archives with your mobile phone.

This paper is organized as follows: Section II shortly

[1] https://www.bundesarchiv.de/fachinformationen/kla/index.html.de, accessed 28.08.2017

[2] http://www.dfg.de/formulare/12_151/12_151_de.pdf, accessed 25.08.2017

[3] https://scantent.caa.tuwien.ac.at

[4] https://github.com/TUWien/DocScan

[5] https://read.transkribus.eu/

presents the ScanTent. In Section III the DocScan app is presented including the page detection, focus measure and series mode, while in Section IV the results are discussed; then follows the conclusion.

II. ScanTent

If a digital camera is used to digitize documents an equipment is needed to fulfill the following constraints: (1) the distance should be kept constant if multiple pages are scanned (same size of the document in the image), (2) the camera position should be fixed to avoid motion blur, and (3) a non-destructive homogenous illumination should be present (since sufficient light is often not given in a usual reading room). Additionally, since documents are often present in a bound form and cannot be opened in the same way as a modern book (due to the condition), the use of two hands make it easier to fix the documents position and flatten the pages. Thus, the ScanTent[6], a cheap (costs will be about 100 Euro) transportable device, has been developed, which fulfills the mentioned constraints. Figure 1 shows a technical sketch of the ScanTent.

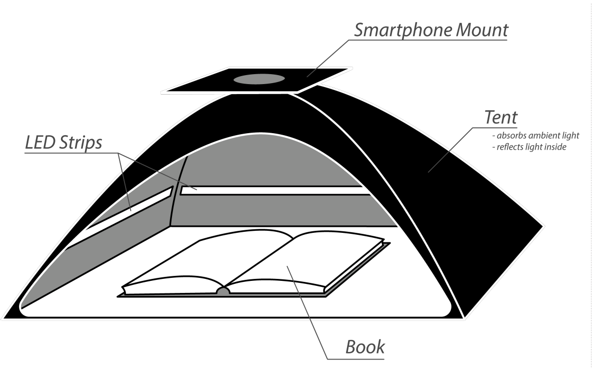


Fig. 1. Technical sketch of the ScanTent.

The main innovative aspect is to take the concept of a dome tent and model it into a ScanTent. The features are (1) portability, (2) fixed distance between book and camera, (3) illumination system with polarized light and polarization filter in front of the camera to avoid specular reflections, (4) a base with a ruling and (5) allows to shoot hands-free. An image of the ScanTent prototype is shown in Figure 2. The bottom part of the ScanTent consists of a ruled base to provide a homogeneous background and to detect the dimensions of a document.

The Depth-of-Field (DoF) is dependent on the mobile phones' camera and the distance of the object to the camera. Exemplarily, the camera of the Samsung Galaxy S6 has a focal length of 4.3 mm, aperture size of F/1.9 (sensor diagonal of about 6.8mm). The Circle-of-Confusion (CoC) is estimated as $sensor-diagonal/1500$. The distance between the document and camera is about 45cm because of the ScanTent. According to Vaquero et al. [3] the depth of field limits D_{near} and D_{far} can be calculated as follows:

$$D_{near} = \frac{sf^2}{f^2 + Nc(s-f)}; D_{far} = \frac{sf^2}{f^2 - Nc(s-f)} \quad (1)$$

where f is the focal length, s is the focused distance, N is the lens' f-number and c is the CoC. This results in depth of field limits of $D_{near} = 37cm$ and $D_{far} = 56cm$ for the given example. Thus, the entire area of a document is in focus, even if the document is curved.

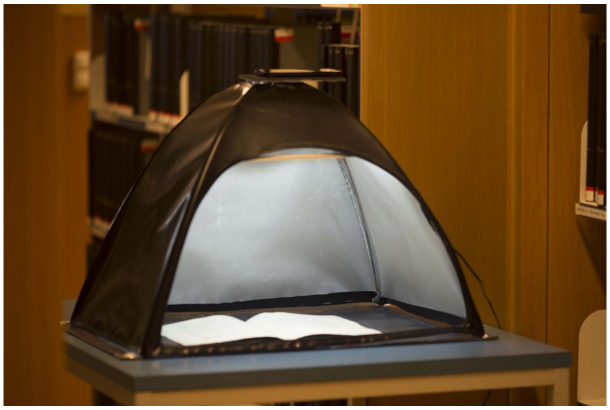
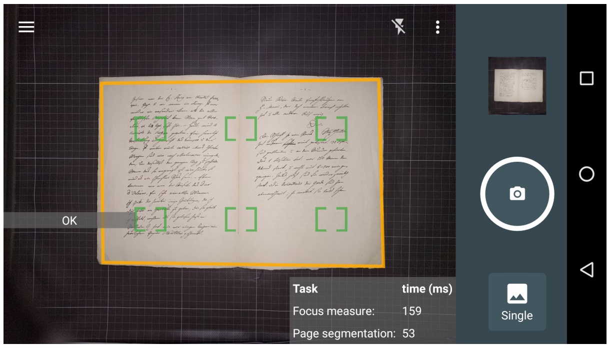


Fig. 2. The ScanTent in the library.

The DocScan app is directly accommodated to the ScanTent. This means that pictures can be taken in an automated mode tailored for scanning bound documents.

III. DocScan

DocScan is a document scanner app, which has a live view and detects in real time the document page. Furthermore, it evaluates if a picture is in focus to give the user feedback and to assure a certain picture quality which will be used for further processing with document image analysis methods. Additionally, it learns a background model to detect page turns and automatically take new pictures. Thus, a book can be scanned by flipping through the pages.



Task	time (ms)
Focus measure:	159
Page segmentation:	53

Fig. 3. Screenshot of the DocScan app with page detection and focus measure.

Figure 3 shows a screenshot of DocScan. The yellow rectangle shows the detected page and the green markers show areas in focus (red indicates out-of-focus sections in the

[6] https://scantent.caa.tuwien.ac.at/en/

document area). The time to calculate the page segmentation is on average 72 ms and for the focus measure 62 ms (on a Samsung S6). The app is developed in Android Studio (Version 2.1). To allow an efficient image processing OpenCV library (Version 3.1.0) is used. The OpenCV functions are called via Java Native Interfaces (JNI). A (Gradle) build script is used to compile the native C++ source files and OpenCV functions with Android SDK. The camera is addressed using the android.hardware.camera class which is backward compatible with Android 4.0. DocScan is open source and available on github[7]. The following sections describe the page detection, the focus measure and the background model in detail.

A. Page Detection

The page segmentation is blob based and is designed in favor of speed rather than accuracy. The polygons of white blobs which are generated using multiple thresholds are analyzed. Quasi rectangular polygons are then compared to find the best hypothesis page region.

In a first step, incoming frames are resized to a maximal width of 900px. This reduces the computation time and false positive rate. The size reduction additionally smooths the image which improves the later polygon estimation. The reduced size image is then converted to the *Lab* color space for blob extraction. Thresholding the color channels *ab* allows for correctly detecting pages in the presence of uneven lighting conditions.

Binary blobs are extracted from each channel using multiple thresholds t_i. In our experiments, we choose t_i to be 10 for the luminance channel *L* and 5 for both color channels *ab*. In addition to blobs generated using these thresholds, a Canny Edge detection with a high false positive rate is performed in order to improve detection in scenes with challenging lighting. This computation strategy returns Maximally Stable Extremal Regions (MSER)-like [4] regions with the exceptions that these regions are not maximally stable (Hence we get multiple similar or even identical regions).

In order to detect quasi rectangular regions in an image, the convex hull of all blobs is approximated using polygons. The conversion to polygons reduces on the one hand memory, on the other hand, properties such as the convexity of a region can be easily computed using polygons. In order to find the most probable page region, only convex polygons with exactly four points are further observed. This strict rule is applicable because the blob detection generated a lot of false positive regions. In addition, taking the convex hull of a book page results in four corner points even if the page is pressed down with a finger.

After filtering the regions, a few dozen potential page regions are left. Two criteria are used in order to find the best hypothesis rectangle. First we compute the "rectangularity (r_i)" of each region by:

$$r_i = \max_{j=1,\ldots,N} abs(cos(\theta_j)) \tag{2}$$

with N being the number of corners ($N = 4$). Then, the region with minimal r_i is chosen as page region. By these means, we account for potential perspective distortions and choose the most page-like (rectangular) region. If there are more regions with the same r_i value, the one with maximal pixel area is returned.

The page detection is evaluated on the ICDAR2015 SmartDoc competition dataset [2]. The detailed evaluation is presented in Section IV.

B. Focus Measure

To determine if a document region is in focus, the image is subdivided into smaller regions and based on a Focus Measure Operator (FMO) [5] the in-focus-value is calculated. The FMO determines for every pixel in the image the level of focus. Pertuz et al. [5] have defined 6 operators for Shape-from-Focus: (1) gradient-based, (2) laplacian-based, (3) wavelet-based, (4) statistics-based, (5) DCT-based and (6) miscellaneous operators. For a detailed description see [5]. Due to computational restrictions on a mobile phone Brenner's FMO [6], [7] (gradient based approach) is used. Brenner is "*based on the second difference of the image gray-levels of an image I*" [5]:

$$FMO = \sum_{i,j} |I(i,j) - I(i+2,j)|^2 \tag{3}$$

Thus, Brenner analyzes the edge sharpness. For autofocusing algorithm the FMO is maximized by changing the position of the lens [8], which is not possible if a single image is analyzed. To get a normalized focus measure in the range between 0 and 1 only edge pixels are taken into account, and the presented FMO is normalized with the number of edge pixels. The edge pixels are defined by a binarized image. A binary image represents the ideal edge image and is calculated using Otsu. The number of edge pixels are used to normalize the FMO measure. Figure 4 shows exemplarily two noisy characters and the gradient values. The binarized image and the relevant edge pixels are also shown.

Thus, only regions where binary egdes are present are taken into account. The proposed approach is similar to Kumar et al. [9] who propose to use the contrast at the edge as normalization of their DoM measure. Based on experiments a threshold for the normalized FMO (0.15) is defined to decide if an image patch is in- or out-of-focus.

C. Series Mode

To detect if a page is turned a background model is learned. This is based on the "*Improved adaptive gaussian mixture model for background subtraction*" from Zivkovic [10], [11]: In this work a background subtraction method is proposed, that makes use of Gaussian mixture models (GMM). Each pixel is modeled as a mixture of Gaussians and the parameters and number of components of the GMM's are constantly adapted in an online procedure. In our implementation two background models are learned: The first background model is used to determine changes stemming from sudden movements - such

[7] https://github.com/TUWien/DocScan

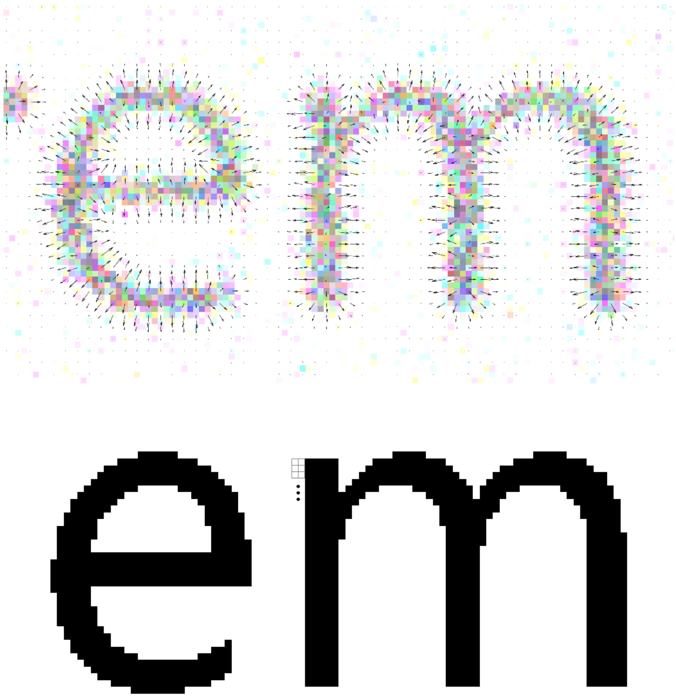


Fig. 4. Image showing the gradients of two noisy characters (for illustration a vector is only assigned to every 2nd pixel) and the binarized reference image. In the binary image the relevant edge pixels are marked exemplarily.

as page flipping. If no change is detected, the current image is added to the second background model. This background model recognizes a page flip if 10% of the image content are foreground pixels. In this case the series mode shoots a new image and the two background models are reinitialized with the new image.

IV. Results

The page segmentation is evaluated on the ICDAR2015 Competition on Smartphone Document Capture and OCR (SmartDoc) dataset [2]. Challenge-1 deals with the detection of page outlines on images acquired with mobile devices. The dataset comprises 6 different document types and videos of the documents are taken with 5 different backgrounds. The videos are Full HD (1920 x 1080) and comprise about 25.000 frames in total. For a detailed description see [2]. As evaluation measure the Jaccard Index (JI) was proposed, which is a measure for the overlapping of the detected quadrilaterals:

$$JI = \frac{area(GT \cap DP)}{area(GT \cup DP)} \tag{4}$$

where GT defines the Ground Truth (GT) polygon of the page and DP defines the Detected Polyon. The JI has a range from 0 to 1, where 1 is the best segmentation possible. The results of the page detection for the proposed method (CVL) and the winner of SmartDoc (LRDE) [2] are shown in Table I.

TABLE I
Jaccard index of the proposed (CVL) page segmentation and the winner of SmartDoc (LRDE).

	CVL JI	LRDE JI
Overall JI	0.7572	0.9716
Background 01	0.8907	0.9869
Background 02	0.8013	0.9775
Background 03	0.8463	0.9889
Background 04	0.8277	0.9837
Background 05	0.0138	0.8613
datasheet	0.8189	0.9758
letter	0.7852	0.9718
magazine	0.7397	0.9707
paper	0.7605	0.9715
patent	0.7227	0.9698
tax	0.7086	0.9696

TABLE II
Ranking of SmartDoc with the proposed (CVL) page segmentation included.

Ranking	Method	JI	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	CVL	0.7572	[0.7526, 0.7617]
8	RPPDI-UPE	0.7408	[0.7359, 0.7456]
8	SEECS-NUST	0.7393	[0.7353, 0.7432]

Table II shows the ranking of challenge 1 of the SmartDoc competition with the results of the proposed method included.

Background 01 to Background 04 have a JI in the range from 80% to 89%. Only Background 05 has a JI from 1.38%. The reason that most images of Background 05 fail is the occlusion of the document with pens and cables. Due to the occlusions no documents are detected in the images. Figure 5 shows an example image of Background 05 where no page has been detected. Figure 6 shows another example of the Background 05 dataset with a wrong detection of the page. Still the occlusion with the cable results in the error. Green rectangles show the GT and white rectangles show the detected document page.

The other backgrounds do not comprise any occlusions of the documents. An example image with a correct page detection is shown in Figure 7 (JI ¿ 0.99). Another example of Background 01 with a wrong page detection is shown in Figure 8

Discussion: The results in Table II show that the method ranks in the lower third of the competition. Compared to the results of LRDE with an average JI of 0.972, there is clearly potential for improvements. An important difference between the presented method and those presented at SmartDoc (e.g. LRDE) is that we detect pages in each frame individually. Hence, there is no performance gained from the fact that

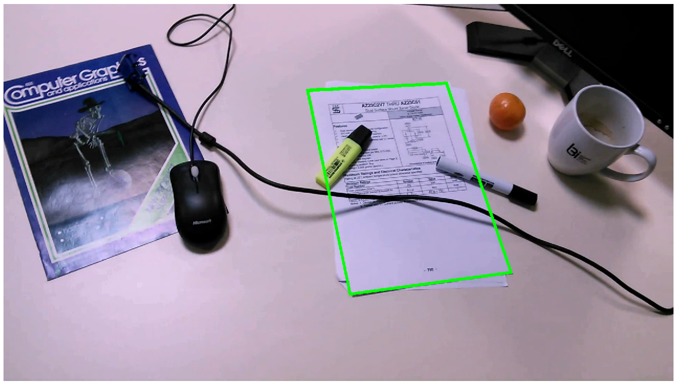


Fig. 5. Example image of dataset *Background 05* where no page is detected.

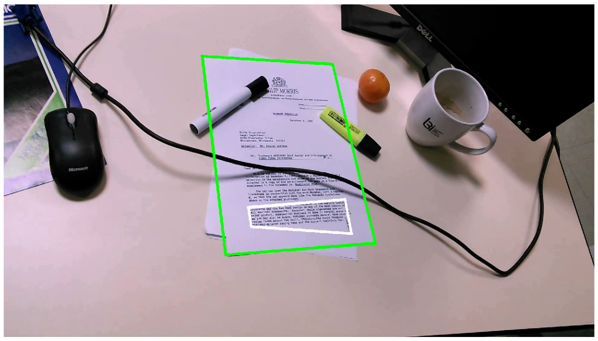


Fig. 6. Example image of dataset *Background 05* with a wrong page detection.

SmartDoc uses videos for evaluation. This is a clear disadvantage if only the page detection quality is considered (as prior detection results could correct wrong outliers). However, it is an advantage in terms of computational speed since we can distribute multiple frames (in a live stream) to different cores and compute them in parallel. The method's low performance of $JI = 0.014$ on Background 05 is by design: Since the presented method aims at giving feedback to users whether a document page is good for digitization or not, pages with occlusions like the cable in Background 05 should not be highlighted. The dataset of SmartDoc comprises also images where two pages are present (e.g. see Figure 5). Due to the ambiguous result no page should be detected in this image. The presented method is unsupervised with the only prerequisite that a document page is a rectangular object with affine transformations. Hence, the method can find pages correctly even if parts of its sides or maximally one corner is occluded. Because of the multiple thresholds, the page detection correctly finds pages in the presence of low contrast between a page and its background. Moreover, it is designed for speed (72ms on average on a Samsung S6) and low memory consumption (8 bit RGB image with max 900 px). It is also planned to create an additional public dataset to show the performance of the page detection on historical documents in combination with the homogeneous background provided by the ScanTent. Experiments have shown, that the well defined settings of the ScanTent lead to a higher performance of the page detection.

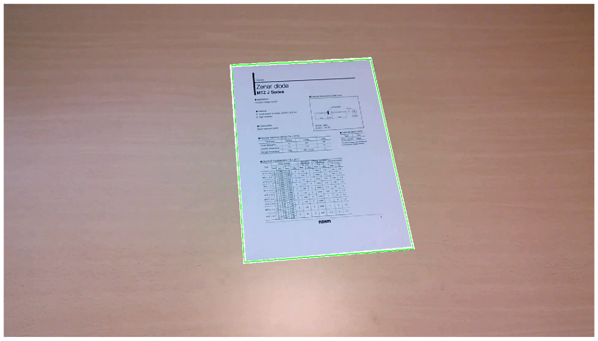


Fig. 7. Example image of dataset *Background 01* with a correct page detection (JI ¿ 0.99).

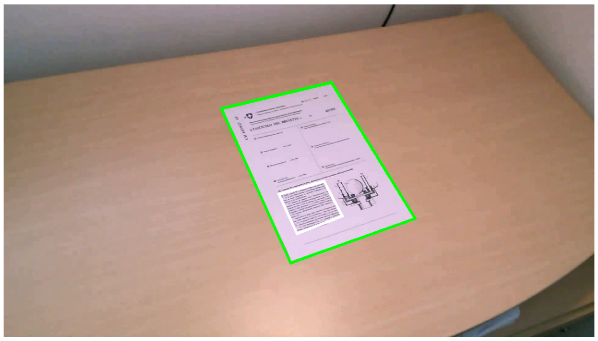


Fig. 8. Example image of dataset *Background 01* with a wrong page detection (JI of 0.099).

V. Conclusion

A scan app, DocScan, with a real time page detection and quality assessment has been presented. The page detection is evaluated on the ICDAR2015 SmartDoc Competition dataset and has an overall Jaccard index of 0.75. This shows a reliable page detection if no occlusions are present in the document. The average time on a Samsung S6 is 72 ms for the page detection and 62 ms for the focus measure. DocScan also automatically takes a picture if the page is turned based on a learned background model. The app is open source and available on github. Additionally, the ScanTent is presented which works as a mount for the mobile phone to assure the best image quality. The ScanTent provides also a lighting system with polarized light. Thus, a low cost mobile scanning system based on mobile phones to digitize documents in archives has been presented. As future work the evaluation of the focus measure and the series mode is planned. Additionally, a public dataset with historical documents to evaluate the page detection will be created.

VI. Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943 and by the Austrian Federal Ministry of Science, Research and Economy.

References

[1] F. A. der KLA, "Wirtschaftliche digitalisierung in archiven," 2016.

[2] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. M. Luqman, M. Mehri, N. Nayef, J. Ogier, S. Prum, and M. Rusiñol, "Icdar2015 competition on smartphone document capture and ocr (smartdoc)," in *13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug 2015, pp. 1161–1165.

[3] D. Vaquero, N. Gelfand, M. Tico, K. Pulli, and M. Turk, "Generalized autofocus," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, Jan 2011, pp. 511–518.

[4] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004, british Machine Vision Computing 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885604000435

[5] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognition*, vol. 46, no. 5, pp. 1415–1432, May 2013. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2012.11.011

[6] J. Brenner, B. Dew, and J. Horton, "An automated microscope for cytologic research," *J Histochen Cytochem*, no. 24, 1976.

[7] L. Firestone, K. Cook, K. Culp, N. Talsania, and K. Preston, "Comparison of autofocus methods for automated microscopy," *Cytometry*, vol. 12, no. 3, pp. 195–206, 1991. [Online]. Available: http://dx.doi.org/10.1002/cyto.990120302

[8] M. Subbarao and J. K. Tyan, "Selecting the optimal focus measure for autofocusing and depth-from-focus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 864–870, Aug 1998.

[9] J. Kumar, F. Chen, and D. Doermann, "Sharpness estimation for document and scene images," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 3292–3295.

[10] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, ser. ICPR '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 28–31. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.479

[11] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, May 2006. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2005.11.005