Combining Focus Measure Operators to Predict OCR Accuracy in Mobile-Captured Document Images

Marçal Rusiñol*†, Joseph Chazalon* and Jean-Marc Ogier*
*L3i Laboratory, Université de La Rochelle
Avenue Michel Crépeau
17042 La Rochelle Cédex 1, France
†Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain

***Abstract*—Mobile document image acquisition is a new trend raising serious issues in business document processing workflows. Such digitization procedure is unreliable, and integrates many distortions which must be detected as soon as possible, on the mobile, to avoid paying data transmission fees, and losing information due to the inability to re-capture later a document with temporary availability. In this context, out-of-focus blur is a major issue: users have no direct control over it, and it seriously degrades OCR recognition. In this paper, we concentrate on the estimation of focus quality, to ensure a sufficient legibility of a document image for OCR processing. We propose two contributions to improve OCR accuracy prediction for mobile-captured document images. First, we present 24 focus measures, never tested on document images, which are fast to compute and require no training. Second, we show that a combination of those measures enables state-of-the art performance regarding the correlation with OCR accuracy. The resulting approach is fast, robust, and easy to implement in a mobile device. Experiments are performed on a public dataset, and precise details about image processing are given.**

I. Introduction

Document image acquisition with mobile devices, especially smartphones, is becoming a essential entry-point in digitization workflows for companies. Despite the evident appeal of on-the-go digitization and near-instant transfer of document images, three major challenges are still to overcome to release the potential of mobile capture for document images.

Digitization distortions are, in the case of mobile-captured images of recent documents, the main cause of perturbation of the image signal. Among the most common distortions, out-of-focus blur is particularly delicate. It seriously alters *legibility*, both for humans and OCR systems, and is linked to camera internal behavior, contrary to motion blur, perspective distortion or lightening conditions over which the user has more control. To prevent bad performance in document processing workflow, the legibility of a mobile-captured document image must be controlled *as early as possible* in the process.

Mobile data transmission fees impose a strict selection on the images to transfer over a network. As a consequence, such legibility control, must be performed *on the mobile device*, to avoid sending unusable data.

Mobility situation changes the way people capture and archive documents: the availability of some document for digitization may be temporary, and depend on the current location of the mobile user. Any control on the captured images must therefore be performed *during or right after the capture*, to avoid missing the opportunity for another capture.

Such an efficient control should enable the notification of a poor shot to the user after the capture, and explanations for such evaluation. In the case of a real-time evaluation, it may be possible to automatically trigger the capture when the conditions are optimal, or even to assist the user during the capture with precise direction like “move closer”, or “light too low”, as suggested in [1].

In this paper, we concentrate on the evaluation of the fitness of a mobile-captured machine-printed document image for later OCR recognition, regarding the amount of out-of-focus blur it contains. We are then interested in OCR accuracy prediction, which is a particular case of *no-reference Document Image Quality Assessment (DIQA)*, as defined in [2]. Here *“no-reference”* means that only the test image is available.

Our claim is that mobile document image acquisition requires *lightweight* and robust methods, which can be implemented using a combination of simple focus measures developed in others communities, and not yet applied to document images.

This paper is organized as follows: Section 2 reviews existing work and shows that interesting focus measures from the *Shape from Focus* and *Autofocus* communities have not been applied on document images yet. Section 3 explains the basic pre-processing we perform on document images before computing those measures. Section 4 presents those measures. Section 5 explains how we normalize and combine them. Section 6 presents our experimental protocol and the results we obtain on a public dataset. Section 7 discusses those results.

II. Related Work

A good introduction to DIQA subcategories is presented in [2]. This study shows that, despite some prior work on OCR accuracy prediction for scanned document images, only a few approaches deal with camera- or mobile-captured images, and even less are considering out-of-focus blur.

In the Document Analysis and Recognition (DAR) community, a major effort to better understand the effects of out-of-focus blur on OCR accuracy in presented in [3]. The authors introduce a dataset of mobile-captured document images with

978-1-4799-3243-6/14 $31.00 © 2014 IEEE
DOI 10.1109/DAS.2014.11

various amounts of out-of-focus blur for which they computed the accuracy of three different OCR systems. They compare the performance of three methods on this dataset: Q (presented in [4]) which relies on singular value decomposition of local image gradient matrix, ΔDoM (presented in [5]) which is based on the detection of edges with gradient analysis, and CORNIA (presented in [6], [7]) which is based on an automatic feature selection, sparse representation, and a regression model to construct generic predictors over images.

We believe that, while the CORNIA method is expected to perform better thanks to its strong machine-learned model, it is less suitable for mobile processing due to its memory or CPU usage, and "lightweight" methods should be considered separately in our case. A recent improvement over this method was recently proposed in [8], with a real-time feature extraction step. However, the regression step still requires important resources, and, more generally, the authors themselves admit that this kind of approach relies heavily on the quality of the training set. Such dataset is hard to produce, and the coverage of OCR accuracy values, as well as the number of elements, may limit the performances of training methods.

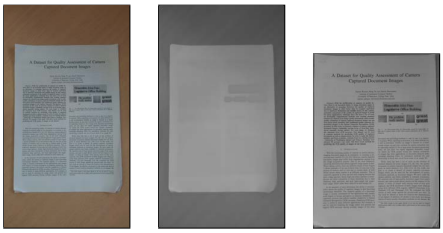
In the field of OCR accuracy prediction for mobile- or camera-captured document images, another method was proposed in [9]. The authors define features based on edge gradient and height-width ratio for words and characters. They use a Support Vector Regression to calculate the word error rate for a given document. While the evaluation method described, focusing on decision making, is highly relevant, the learning step and some dependence to character shapes limits, in our opinion, such method.

An interesting study of available focus measures, for an entirely different purpose, is presented in [10]. In this work, the authors present more than 30 "focus operators" and their application to recover depth information for each pixel in natural images. The authors not only perform a great comparison of run-time properties, weaknesses and strengths of each method: they also propose to group them in 6 main categories which exhibit consistent behavior under the same perturbations like image contrast, image noise, or image saturation.

We therefore propose to investigate the application of those mature focus measures on document images to estimate out-of-focus blur, and its correlation with OCR accuracy. We are particularly interested in evaluating how those measures can be combined: as they are intended to be computationally efficient and were sometimes even used for autofocus systems, it really makes sense to try to form a global method which would be more robust to real capture conditions, and overcome individual weaknesses. To enable the experimental evaluation of our contribution, we used the public dataset presented in [3].

III. Simple Page Segmentation

Since we are dealing with mobile-acquired images, the incoming images do not correspond just to the digitized document page, but usually contain some background that is irrelevant for us. Page segmentation not being the main topic of our research, we just implemented a simple yet effective page detection that performs well enough on the tested images. Color images are first transformed to gray-scale by computing the luminance of the image. By applying a median filter with a large enough structuring element we are able to get rid of the text from the page while keeping the white background of the page (c.f. Fig. 1b)). In our experimental setup, the structuring element was a 31×31 rectangle. By thresholding this image with Otsu's method and getting the largest connected component we obtain the page's bounding-box that is used for segmenting the page (c.f. Fig. 1c)).



a) b) c)

Fig. 1. Simple page segmentation. a) Mobile-acquired document, b) median filtering, c) final page segmentation.

Of course, such simple techniques will fail for severely distorted images presenting highlights and shadows, strong perspective changes, or if the page is not highly contrasted from the background, but provided accurate enough results in our test scenario.

IV. Focus Measures

In [10], the authors distinguished 6 families of focus operators. *Gradient-based* (GRA) and *Laplacian-based* (LAP) operators estimate the sharpness and the amount of edges in an image using gradient or first derivative, and, respectively, second derivative or Laplacian. *Wavelet-based* (WAV) and *DCT-based* operators are based on discrete wavelet transform and discrete cosine transform (DCT) coefficients. *Statistic-based* (STA) operators are based on local descriptors (texture analysis, in particular). Finally, the *miscellaneous* (MIS) operators are the ones which do not fit in the previous categories, due to some dependence on global indicators like histograms or image contrast.

Grounding our work on the focus measure operators proposed in [10], we selected 24 techniques which appeared as most promising. Whereas the techniques used for *shape from focus* in [10] need to produce pixel-level measures, we just need a global estimation of the focus for the whole image. We also discarded the DCT-based operators, due to their specificity to some image and video formats, or their processing time. Finally, some preliminary experiments led us to filter out the operators with bad performances on document images.

The Table I lists the focus measures we considered in our experiment, along with the abbreviations from [10] where their exact formulation can be found.

V. Feature Fusion

Even though the measures listed above are state of the art auto-focus techniques, some of them might not correlate well with the output of an OCR engine. However, since it is difficult to a-priori judge which ones will perform better than the rest and which is the best amount of features to

TABLE I. USED FOCUS MEASURES AND THEIR ABBREVIATIONS FROM [10].

Focus operator	Abbr.	Focus operator	Abbr.
Gaussian derivative	GRA1	Ratio of wavelet coefficients	WAV3
Gradient energy	GRA2	Gray-level variance	STA3
Thresholded absolute gradient	GRA3	Gray-level local variance	STA4
Squared gradient	GRA4	Normalized gray-level variance	STA5
Tenengrad	GRA6	Histogram entropy	STA7
Tenengrad variance	GRA7	Histogram range	STA8
Energy of Laplacian	LAP1	Brenner's measure	MIS2
Modified Laplacian	LAP2	Image curvature	MIS4
Diagonal Laplacian	LAP3	Helmli and Scherer's mean	MIS5
Variance of Laplacian	LAP4	Steerable filters-based	MIS7
Sum of wavelet coefficients	WAV1	Spatial frequency measure	MIS8
Variance of wavelet coefficients	WAV2	Vollath's autocorrelation	MIS9

combine, we have tested all the possible feature combinations. Being $S = \{FM_1, FM_2, \ldots, FM_n\}$ the set of all $n = 24$ focus measures FM applied to an incoming image under test, we tested all the possible combinations T of this set. We denote T_m, with $m \in [1,\ n]$ all the possible subsets of measures formed by taking m elements from S. Particularly, T_m^j, with $j \in \left[1,\ \binom{n}{m}\right]$ will denote the j-th subset having m elements. Such combinations resulted in testing more than 16 million different configurations. Once we have a subset of focus measures to combine, we have to face two different aspects. On the one hand how to normalize the measures so they fall within a similar range and on the other hand, how to combine them into a single number.

A. Measure Normalization

Since each of the different focus measures fall within different numeric ranges, before combining them we must normalize their ranges [11]. In our experiments we have tested four off-the-shelf normalization techniques. Given a set of focus measures T_m^j with their normalized measures ${T_m^j}'$ are obtained by the following normalizations.

- **Min-max**: applies a scaling factor and transforms the measures in a common range $[0, 1]$. Being $min(T_m^j)$ and $max(T_m^j)$ the minimum and maximum of the scores respectively,

$$ {T_m^j}' = \frac{T_m^j - min(T_m^j)}{max(T_m^j) - min(T_m^j)}. $$

  Such method is highly sensitive to outliers in the data used for estimation.

- **Z-score**: is one of the most common normalization techniques. It is computed using the arithmetic mean μ and standard deviation σ of the given data.

$$ {T_m^j}' = \frac{T_m^j - \mu}{\sigma}. $$

  By using the arithmetic mean and standard deviation, the method is also sensitive to outliers.

- **Tanh**: is a more robust and efficient normalization technique that also takes into account the mean and standard deviation.

$$ {T_m^j}' = \frac{1}{2}\left\{\tanh\left[0.01 \cdot \left(\frac{T_m^j - \mu}{\sigma}\right)\right] + 1\right\}. $$

- **MAD**: the median and median absolute deviation are insensitive to outliers and the points in the extreme tails of the distribution.

$$ {T_m^j}' = \frac{T_m^j - median(T_m^j)}{MAD}, $$

  where $MAD = median(|T_m^j - median(T_m^j)|)$.

B. Fusion Strategies

After normalizing the response of the focus measures, a subset of focus operators ${T_m^j}'$ can be easily combined in order to obtain a single indicator by just computing the maximum, minimum, sum, product, average or median values of this subset. Formally, we will denote those fusion strategies as

$$
\begin{aligned}
combMAX &= max\left({T_m^j}'\right), \\
combMIN &= min\left({T_m^j}'\right), \\
combSUM &= \sum_{j=1}^{\binom{n}{m}} \left({T_m^j}'\right), \\
combPROD &= \prod_{j=1}^{\binom{n}{m}} \left({T_m^j}'\right), \\
combAVG &= mean\left({T_m^j}'\right), \\
combMED &= median\left({T_m^j}'\right).
\end{aligned}
$$

We will report in the experimental section the results obtained by each normalization technique and each fusion strategy.

VI. EXPERIMENTAL RESULTS

In order to evaluate our proposal, we have used the publicly available DIQA dataset [3]. This dataset is composed of 25 documents. Each document has been acquired several times by a cellphone camera at different focal lengths: some acquisitions are perfectly focused whereas some others present severe blurring effects. The final dataset is composed of 175 images and each of those have been OCRed by three different engines (ABBYY FineReader, Omnipage and Tesseract). The comparison between the OCR output and a manual transcription of the document permit to obtain the OCR accuracy for each image. Median Pearson (LCC) and Spearman (SROCC) correlation factors are provided in order to assess whether the proposed metrics are in agreement with the obtained OCR accuracies,

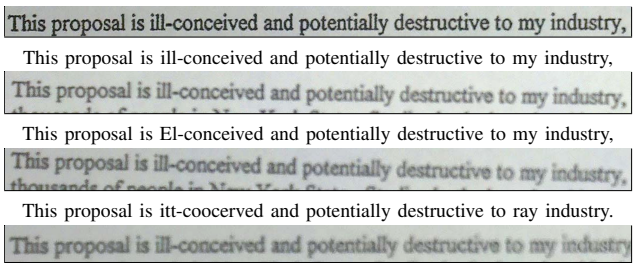


Fig. 2. Example of a portion of a document from the DIQA dataset with different blurring levels and their OCR outputs.

TABLE II. MEDIAN LCC FOR DIFFERENT NORMALIZATIONS AND FUSION STRATEGIES.

Fusion method	Normalization			
	Min-max	Z-score	Tanh	MAD
CombMAX	0.87439	0.86364	0.86363	0.86679
CombMIN	0.89791	**0.9378**	0.93779	0.93092
CombSUM	0.92056	0.92164	0.92165	0.91945
CombPROD	0.87073	0.91372	0.92175	0.86989
CombAVG	0.92056	0.92164	0.92165	0.91945
CombMED	0.92056	0.92164	0.92165	0.91945

and thus be useful to predict the OCR behavior. We can see an example of this dataset in Fig. 2.

We can see in Table II the results that we obtained for each normalization and fusion strategies in terms of the median Pearson correlation coefficient (LCC). Results in Table II just report the best feature combination for each normalization and fusion. In most of the cases, the best performance was reached by just combining between two and four of the 24 possible focus measures. We can appreciate that despite the selected normalization, the fusion strategy that performs best is usually the combMIN. Thus, for a given set of focus measures, just taking the most pessimistic one that considers the document more out of focus, is the strategy correlates the best with the OCR accuracies.

In general, the focus measures that worked the best were the ones from the gradient family. In particular, our best configuration used a subset of the four focus measures $T_4^j = \{GRA1, GRA2, GRA4, STA8\}$. The time taken to compute the four measures and its normalization and combination in a desktop PC under an unoptimized Matlab code was 0.61 secs. in average.

In Table III we compare our results with the state-of-the-art methods presented by Kumar et al. in [3]. We can appreciate that the proposed method outperforms both Q [4] and ΔDOM [5] methods in terms of median LCC and median SROCC. However, the CORNIA [7] method that uses machine learning techniques in order to learn how to predict the OCR accuracy still performs better than the metric-based approaches.

We report in Table IV the obtained median LCCs for the three different OCR engines in the DIQA dataset. The fact that no substantial changes can be appreciated between ABBYY and Tesseract despite their huge difference in accuracy (c.f. Figs 4 and 6 in [3]), might indicate that the evaluation protocol proposed in [3] is biased to produce overoptimistic results.

TABLE III. COMPARISON WITH STATE OF THE ART [3].

Method	Learning	Med. LCC	Med. SROCC
Q	Metric-based	0.8271	0.9370
ΔDOM	Metric-based	0.8488	0.9524
CORNIA	Learning-based	0.9747	0.9286
Proposed	Metric-based	0.9378	0.96429

TABLE IV. MEDIAN LCC WITH DIFFERENT OCR ENGINES.

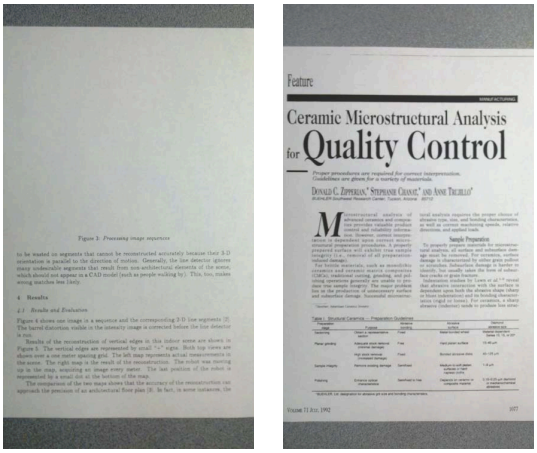
Method	ABBYY	Omnipage	Tesseract
CombMIN+Z-score	0.9378	0.8794	0.9197

LCCs are computed independently document-wise, i.e. just considering the 6 to 8 documents for a given document class. Performance is evaluated then by reporting the median of those 25 LCCs. By reporting the median LCC value, outlier classes in which the methods might not perform well are disregarded. If we compute directly the LCC for all the 175 images we obtain a 0.6467 which compares quite inferiorly to the reached 0.9378.

Finally, we show in Fig. 3 some failure cases. In the image of Fig. 3a), we obtain a rather low focus measure whereas the OCR accuracy is beyond 90%. Such low focus measure is probably provoked by the large white space in this page. Contrarily, in Fig. 3b) the focus measure is rather high but the OCR accuracy is low (25%). In this case, most of the text of the document image is out of focus, but the huge headline provokes that the focus measure is optimistically high.

VII. DISCUSSION AND CONCLUSIONS

In this paper we have presented a metric-based method for quality assessment of mobile-acquired document images able to predict at some extent the accuracy that an OCR engine will yield. Starting with the hypothesis that by combining several focus measures from different families we should reach better performances than just relying on a single metric, we finally found out that gradient-based features are the ones that correlate the best with the response of OCR engines. Although the proposed method outperforms other metrics proposed in the



a) b)

Fig. 3. Failure cases

literature, methods having a single indicator are still far from machine learning-based techniques.

In this paper we have dealt with simple combination of the focus measures, however, maybe better performances could be further reached with more complex combination strategies such as a linear combination of the focus measures with learned weights.

VIII. Appendix: Selected Focus Measures

This appendix summarizes the focus measure operators in which we obtain the best performances. We refer the interested reader to [10] for the details of the rest of the measures.

A. Gaussian Derivative (GRA1)

The Gaussian derivative focus measure is computed by

$$\phi = \sum_{(x,y)} (I \times \Gamma_x)^2 + (I \times \Gamma_y)^2, \tag{1}$$

with Γ_x and Γ_y are the partial derivatives of the gaussian function

$$\Gamma(x, y, \sigma) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right). \tag{2}$$

B. Gradient Energy (GRA2)

The gradient energy is computed as the sum of squares of the first derivative in the x and y directions

$$\phi_{x,y} = \sum_{(i,j)\in\Omega(x,y)} (I_x(i,j)^2 + I_y(i,j)^2), \tag{3}$$

in which $\Omega(x, y)$ defines a local neighborhood of the pixel (x, y). The global measure ϕ for the whole image is obtained by averaging all the $\phi_{x,y}$

C. Squared Gradient (GRA4)

The squared gradient method computes the first derivative of the image in the horizontal dimension, squared in order to increase the influence of larger gradients.

$$\phi_{x,y} = \sum_{(i,j)\in\Omega(x,y)} I_x(i,j)^2, \ |I_x(i,j)| \geq T. \tag{4}$$

D. Histogram range (STA8)

The histogram range is computed as

$$\phi = max(k|H > 0) - min(k|H > 0), \tag{5}$$

begin H the histogram of the image under analysis.

Acknowledgment

This work has been partially funded by the Valconum consortium (http://valconum.fr) and the EU-COFUND project TECNIOSPRING AVAL-DOC TECSPR13-1-0017.

References

[1] F. Chen, S. Carter, L. Denoue, and J. Kumar, "SmartDCap: semi-automatic capture of higher quality document images from a smartphone," in *Proceedings of the 2013 international conference on Intelligent user interfaces (IUI'13)*, Mar. 2013, pp. 287–296.

[2] P. Ye and D. Doermann, "Document image quality assessment: A brief survey," in *12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 723–727.

[3] 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.

[4] X. Zhu and P. Milanfar, "Automatic parameter selection for denoising algorithms using a no-reference measure of image content," *IEEE Transactions on Image Processing*, vol. 19, no. 12, p. 3116–3132, 2010.

[5] J. Kumar, F. Chen, and D. Doermann, "Sharpness estimation for document and scene images," in *21st International Conference on Pattern Recognition (ICPR)*, 2012, p. 3292–3295.

[6] P. Ye and D. Doermann, "Learning features for predicting OCR accuracy," in *21st International Conference on Pattern Recognition (ICPR)*, 2012, p. 3204–3207.

[7] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, p. 1098–1105.

[8] ——, "Real-time no-reference image quality assessment based on filter learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2013.

[9] X. Peng, H. Cao, K. Subramanian, R. Prasad, and P. Natarajan, "Automated image quality assessment for camera-captured OCR," in *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2621–2624.

[10] S. Pertuz, D. Puig, and M. Angel Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognition*, vol. 46, no. 5, pp. 1415–1432, May 2013.

[11] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.